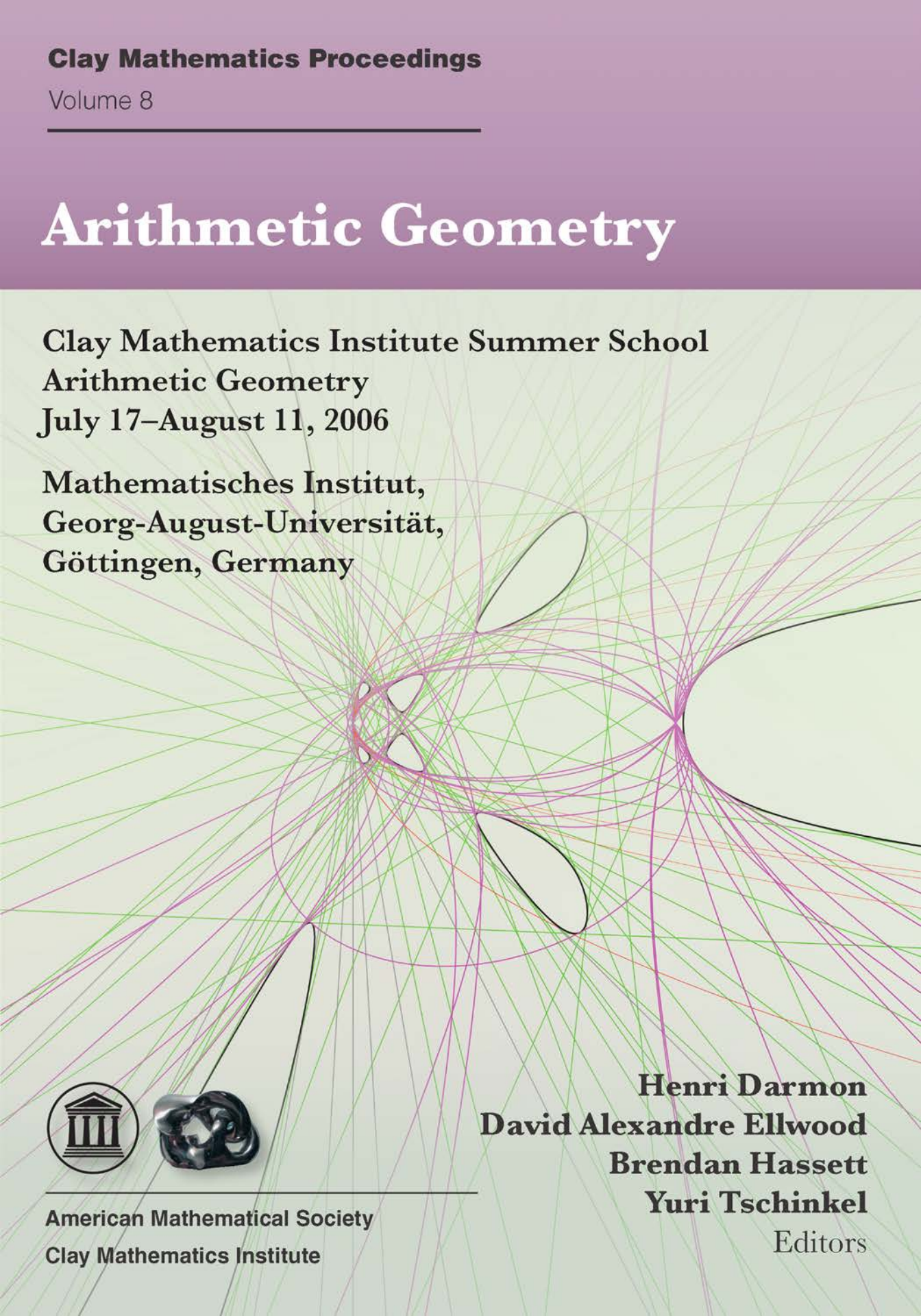
Clay Mathematics Proceedings
Volume 8

Arithmetic Geometry

Clay Mathematics Institute Summer School
Arithmetic Geometry
July 17–August 11, 2006

Mathematisches Institut,
Georg-August-Universität,
Göttingen, Germany



Henri Darmon
David Alexandre Ellwood
Brendan Hassett
Yuri Tschinkel
Editors

American Mathematical Society
Clay Mathematics Institute

Arithmetic Geometry

Clay Mathematics Proceedings
Volume 8

Arithmetic Geometry

Clay Mathematics Institute Summer School
Arithmetic Geometry
July 17–August 11, 2006

Mathematisches Institut,
Georg-August-Universität,
Göttingen, Germany

Henri Darmon
David Alexandre Ellwood
Brendan Hassett
Yuri Tschinkel
Editors

American Mathematical Society
Clay Mathematics Institute

2000 *Mathematics Subject Classification.* Primary 14E30, 14G05, 14D10.

The figure on the front cover, "Rational Curves on a $K3$ Surface," is courtesy of Noam D. Elkies.

Library of Congress Cataloging-in-Publication Data

Arithmetic geometry / Henri Darmon . . . [et al.], editors.
p. cm. — (Clay mathematics proceedings ; v. 8)
Includes bibliographical references.
ISBN 978-0-8218-4476-2 (alk. paper)
1. Arithmetical algebraic geometry. I. Darmon, Henri, 1965–

QA242.5.A754 2009
516.3′5—dc22

2009027374

Published by the American Mathematical Society, Providence, RI,
for the Clay Mathematics Institute, Cambridge, MA.
Printed in the United States of America.

Visit the AMS home page at `http://www.ams.org/`
Visit the Clay Mathematics Institute home page at `http://www.claymath.org/`

10 9 8 7 6 5 4 3 2 1 14 13 12 11 10 09

Contents

Higher-dimensional varieties

Introduction

Classically, arithmetic is the study of rational or integral solutions of Diophantine equations. From a modern standpoint, this is a particular case of the study of schemes over algebraically nonclosed fields and more general commutative rings. The geometric viewpoint, dating back to ancient Greece, has been a source of inspiration to generations of mathematicians. The guiding principle is that

geometry determines arithmetic.

The tremendous power of this principle has been amply demonstrated in the works of Faltings on the Mordell conjecture and Wiles on Fermat's last theorem.

This volume grew out of the 2006 Clay Summer School held at the Mathematisches Institut of the University of Göttingen. The goal of the school was to introduce participants to the wealth of new techniques and results in arithmetic geometry. The first three weeks of the school were devoted to three main courses, covering curves, surfaces and higher-dimensional varieties, respectively; the last week was dedicated to more advanced topics. An important component of the school was a seminar focused on computational and algorithmic aspects of arithmetic geometry. The present proceedings volume reflects this structure.

Curves:

The main *geometric* invariant of a curve is its genus; the *arithmetic* is very different for curves of genus $0, 1$ and ≥ 2 respectively. In genus 0, we can answer, completely and effectively, whether or not a curve contains rational points and how these points are distributed. The theory of genus 1 curves is one of the richest subjects in mathematics, with spectacular recent theorems, e.g., modularity of elliptic curves over the rationals, and with many outstanding open questions, such as the Birch/Swinnerton-Dyer conjecture. In higher genus, the most fundamental result is the proof of the Mordell conjecture by Faltings, and the most challenging open question is to give an effective version of this result.

The lecture notes by Darmon cover the following topics:

- Faltings' proof of the Mordell Conjecture;
- Rational points on modular curves and Mazur's approach to bounding them;
- Rational points on Fermat curves and Wiles' proof of Fermat's Last Theorem;

- Elliptic curves and the Birch and Swinnerton-Dyer conjecture, following Gross-Zagier and Kolyvagin.

Contributions by Chapdelaine, Charollois, Dasgupta, Greenberg, Rebolledo, and Voight discuss more specialised topics that grew out of these lectures, such as

- Generalised Fermat equations (Chapdelaine);
- Merel's extension of Mazur's techniques to study rational points on modular curves over number fields, and the uniform boundedness conjecture for torsion of elliptic curves (Rebolledo);
- Natural generalisations of Fermat's Last Theorem due to Kraus and Halberstadt, building on Frey's approach (Charollois);
- CM points on modular curves and their applications to elliptic curves (Dasgupta, Voight);
- Shimura curves with a focus on computational aspects (Voight, Greenberg);
- Stark-Heegner points (Greenberg).

In addition, a paper by Manin treats modular symbols (which play an important role in Merel's proof of the uniform boundedness conjecture explained in Rebolledo's article) and discusses higher dimensional generalizations.

Surfaces:

The geometry of surfaces over the complex numbers is much more involved, and their birational classification was a milestone in algebraic geometry. Hassett's paper gives a thorough introduction to this classification over nonclosed fields, and its implications for Diophantine questions like the existence of rational points and weak approximation. It also touches on geometric descent constructions generalizing Fermat's descent (universal torsors) and algebraic approaches to these objects (Cox rings).

Harari's paper discusses non-abelian versions of descent, which have yielded new counterexamples to local-global principles for rational points on surfaces over number fields. Once rational points exist, one can ask whether they are Zariski dense and analyze their distribution with respect to heights; these questions are addressed, for both surfaces and higher-dimensional varieties, in Tschinkel's survey.

Vioreanu offers tantalizing computational evidence for conjectures about the algebraic structure of rational points on cubic surfaces. He explores whether all points can be generated from a small number using elementary geometric operations.

Higher-dimensional varieties:

Some of the most interesting higher-dimensional varieties from the arithmetic point of view are low-degree hypersurfaces and varieties closely related to algebraic groups: toric varieties, homogeneous spaces, and equivariant compactifications of groups. Here one is interested in existence questions, density of rational points, and counting points of bounded height. For the last problem, height zeta functions are an important tool and techniques of harmonic analysis can be profitably employed. A selection of recent results in this direction appears in the survey of Tschinkel.

we have a basis (x_0, x_1, x_2) for $\Gamma(D)$ for which S can be given by

$$\begin{aligned} S(x_0,x_1,x_2) \;=\; & 3686400x_0^6 \\ & - 256(33975x_1^2 - 8569x_1x_2 - 45x_2^2)x_0^4 \\ & + (5130225x_1^4 - 1860100x_1^3x_2 + 138414x_1^2x_2^2 - 4180x_1x_2^3 + 9x_2^4)x_0^2 \\ & - 8(5643x_1^5x_2 - 2495x_1^4x_2^2 + 209x_1^3x_2^3 - 5x_1^2x_2^4). \end{aligned}$$

The symmetry takes $(x_0 : x_1 : x_2)$ to $(-x_0 : x_1 : x_2)$, and the node is at $(x_0 : x_1 : x_2) = (0 : 0 : 1)$. Our model of $\mathcal{X}/\mathbf{Q}$ is obtained from the double cover $y^2 = S(x_0, x_1, x_2)$ by blowing up the preimage of this node. In the picture, the real locus of the sextic curve $C : S = 0$ is plotted in black in the $(x_0/x_2, x_1/x_2)$ plane; it consists of nine components, together with an isolated point at the node.

A line $\ell \subset \mathbf{P}^2$ is tritangent to C if and only if it lifts to a pair of smooth rational curves $\ell_\pm$ on $\mathcal{X}$. Then $\ell_+ + \ell_- = D$ in $\mathrm{NS}(\mathcal{X})$. Orthogonal projection to $L \otimes \mathbf{Q}$ then maps $\mathrm{NS}(\mathcal{X})$ to a lattice L' containing L with index 2, taking the curves $\ell_\pm$ to a pair of vectors $\pm v = \ell_\pm - \frac{1}{2}D \in L'$ of norm $5/2$ that are orthogonal to R_L. Conversely, every such pair comes from a tritangent line. There are 43 such lines; one of these is $x_2 = 0$, which is the line at infinity in our picture, and the remaining 42 are plotted in green. (Some of the tangency points are not in the picture because they are either complex conjugate or real but outside the picture frame.)

Each of these lines has the property that the restriction $S|_\ell$ is the square of a cubic polynomial. The same is true if ℓ is a line passing through the node of C and tangent to C at two other points. There are nine such lines, plotted in gray. They correspond to norm-$(5/2)$ vectors in L' not orthogonal to R_L, up to multiplication by -1 and translation by R_L.

A generic line $\lambda \subset \mathbf{P}^2$ meets these $43 + 9$ lines in 52 distinct points that lift to 52 pairs of rational points on the genus-2 curve $y^2 = S|_\lambda$. This already improves on the previous record for an infinite family of genus-2 curves over $\mathbf{Q}$ (which was 24 pairs, due to Mestre). We do better yet by exploiting rational curves of higher degree in $\mathbf{P}^2$ on which S restricts to a perfect square.

There are 1240 conics $c \subset \mathbf{P}^2$ for which $S|_c$ is a square; geometrically these are the conics such that each point in the intersection $c \cap C$ has even multiplicity (either the node of C or a point of tangency). Such a conic lifts to a pair of rational curves $c_\pm$ on $\mathcal{X}$ with $c_+ + c_- = 2D$. These $c_\pm$ come from vectors $c_\pm - D \in L$ of norm 4 up to translation by R_L, except for norm-4 vectors of the form $v - v'$ with $v, v' \in L'$ of norm $5/2$. The conics c are all rational over $\mathbf{Q}$, because for each c we can find c' such that $c_+ \cdot c'_+$ is odd. In general the intersections of c with a generic line $\lambda \subset \mathbf{P}^2$ need not be rational, but we can choose λ so as to gain a few rational points. Most notably, 18 of the conics happen to pass through the point $P_0 : (x_0 : x_1 : x_2) = (0 : 1 : 3)$ on the axis of symmetry $x_0 = 0$ of the sextic C. These conics are plotted in purple on our picture. If λ is a generic line through P_0 then the genus-2 curve $y^2 = S|_\lambda$ gains 18 more pairs of points above the second intersections of λ with the purple conics. We also lose one pair because two of our 52 tritangent lines pass through P_0, but we gain two more pairs by finding two rational cubic curves $\kappa \subset \mathbf{P}^2$ for which $S|_\kappa$ is a square and P_0 is the node of κ. This brings the total to $52+18-1+2 = 71$. If c_1, c_2 are two of the remaining 1222 conics such that $(c_1)_+ \cdot (c_2)_+$ is odd then we have infinitely many choices (parametrized by an elliptic curve of positive rank) of lines $\lambda \ni P_0$ for which each of $\lambda \cap c_1$ and $\lambda \cap c_2$ consists of two further rational points, bringing our total to 75. This is the

current record for the number of pairs of rational points on an infinite family of genus-2 curves over $\mathbf{Q}$; the previous record, due to Mestre, was 24 pairs.

In another direction, C has the rational point $P_1 : (x_0 : x_1 : x_2) = (0 : 1 : 0)$ (in our picture this is the point at infinity in the horizontal direction). If $\lambda \ni P_1$ then the genus-2 curve $y^2 = S|_\lambda$ has a rational Weierstrass point mapping to P_1. The tangent to P_1 is the line of infinity, which is one of our 52 tritangent lines; but this still leaves 51 pairs of rational points. In fact we get 4 more because four of our 1240 conics contain P_1. These are shown in our picture as red horizontal parabolas. As before we can get at least 4 more pairs for infinitely many choices of λ parametrized by an elliptic curve of positive rank. This yields infinitely many genus-2 curves over $\mathbf{Q}$ with a rational Weierstrass point and at least 59 further pairs of rational points.

References

[Sch08] Matthias Schütt, *K3 surfaces with Picard rank 20*, 2008, arXiv:0804.1558.

Department of Mathematics, Harvard University, Cambridge, MA 02138
E-mail address: elkies@math.harvard.edu

Curves

Clay Mathematics Proceedings
Volume 8, 2009

Rational points on curves

Henri Darmon

Abstract. This article surveys a few of the highlights in the arithmetic of curves: the proof of the Mordell Conjecture, and the more detailed theory that has developed around the classes of curves most studied until now by number theorists: modular curves, Fermat curves, and elliptic curves.

Contents

Introduction

Algebraic number theory is first and foremost the study of Diophantine equations. Such a definition is arguably too narrow for a subject whose scope has expanded over the years to encompass an ever-growing list of fundamental notions: number fields and their class groups, abelian varieties, moduli spaces, Galois representations, p-divisible groups, modular forms, Shimura varieties, and L-functions, to name just a few. All of these subjects will be broached (sometimes too briefly, for reasons having less to do with their relative importance than with limitations of time, space, and the author's grasp of the subject) in this survey, which is devoted to the first nontrivial class of Diophantine equations: those associated to varieties of dimension one, or *algebraic curves*.

The term *Diophantine equation* refers to a system of polynomial equations

2000 *Mathematics Subject Classification.* Primary 11G30, Secondary 11G05, 11G18, 11G40, 14G05, 14G35.

$$X : \begin{cases} f_1(x_1, \dots, x_n) = 0 \\ \vdots \\ f_m(x_1, \dots, x_n) = 0 \end{cases} \quad (\text{with } f_i \in \mathbf{Z}[x_1, \dots, x_n]). \tag{1}$$

Given such a system, one wishes to understand (and, if possible, determine completely) its set of integer or rational solutions.

Little of the essential features of the question are lost, and much flexibility is gained, if one replaces the base ring $\mathbf{Z}$ by a more general ring $\mathcal{O}$. The prototypical examples are the ring of integers $\mathcal{O}_K$ of a number field K, or the ring $\mathcal{O}_{K,S}$ of its S-integers, for a suitable finite set S of primes of $\mathcal{O}_K$.

Fix such a base ring $\mathcal{O} = \mathcal{O}_{K,S}$ from now on, and assume that the polynomials in (1) have coefficients in $\mathcal{O}$.

If R is any $\mathcal{O}$-algebra, the set of solutions of (1) with coordinates in R is denoted $X(R)$:

$$X(R) := \{(x_1, \dots, x_n) \in R^n \text{ satisfying (1)}\}.$$

The functor $R \mapsto X(R)$ from the category of $\mathcal{O}$-algebras to the category of sets is *representable*,

$$X(R) = \mathrm{Hom}_{\mathcal{O}}(A_X, R), \quad \text{where } A_X = \mathcal{O}[x_1, \dots, x_n]/(f_1, \dots, f_m). \tag{2}$$

In this way the system (1) determines the *affine scheme* $X := \mathrm{Spec}(A_X)$ over $\mathrm{Spec}(\mathcal{O})$.

When the polynomials in (1) are homogeneous, it is customary to view X as giving rise to a *projective scheme* over $\mathcal{O}$. When R is a principal ideal domain, the set $X(R)$ is a subset of the set $\mathbb{P}_{n-1}(R)$ of n-tuples $(x_1, \dots, x_n) \in R^n$ satisfying $Rx_1 + \cdots + Rx_n = R$, taken modulo the equivalence relation defined by

$$(x_1, \dots, x_n) \sim (x_1', \dots, x_n') \quad \text{if } x_i x_j' - x_j x_i' = 0, \qquad \forall \quad 1 \le i, j \le n.$$

Specifically,

$$X(R) := \{(x_1, \dots, x_n) \in \mathbb{P}_{n-1}(R) \text{ satisfying (1)}\}.$$

In the projective setting, replacing the base ring $\mathcal{O}$ by its fraction field K, and X by its *generic fiber* X_K—a projective variety over K—does not change the Diophantine problem. For instance, the natural map $X(\mathcal{O}) \longrightarrow X_K(K)$ is a bijection. So there is no distinction between the study of integral and rational points on a scheme whose generic fiber is a projective variety.

Here are some of the basic questions that can be asked about the behaviour of $X(\mathcal{O})$.

Question 1. *What is the cardinality of $X(\mathcal{O})$? Is it finite, or infinite?*

Question 2. *If $X(\mathcal{O})$ is finite, can its cardinality be bounded by a quantity depending in a simple way on X and $\mathcal{O}$?*

Question 3. *Can $X(\mathcal{O})$ be effectively determined?*

The arithmetic complexity of a point $P \in X(\mathcal{O})$—roughly speaking, the amount of space that would be required to store the coordinates of P on a computer—is measured by a (logarithmic) *height function*

$$h : X(\mathcal{O}) \longrightarrow \mathbf{R}.$$

The precise definitions and basic properties of heights are discussed elsewhere in this volume. Let us just mention that for any real $B > 0$, the number $N(X; B)$ of $P \in X(\mathcal{O})$ with $h(P) \leq B$ is finite, in any reasonable definition of h.

QUESTION 4. *When $X(\mathcal{O})$ is infinite, what can be said about the asymptotics of the function $N(X; B)$ as $B \longrightarrow \infty$?*

A related question is concerned with the *equidistribution* properties of the points in $X(\mathcal{O})$ (ordered by increasing height), relative to some natural measure on $X(\mathbf{R})$ or $X(\mathbf{C})$.

An *algebraic curve* over $\mathcal{O}$ is a scheme X (either affine or projective) of relative dimension one over $\mathrm{Spec}(\mathcal{O})$. If its generic fiber is smooth, the set $X(\mathbf{C})$ (relative to a chosen embedding of $\mathcal{O}$ into $\mathbf{C}$, through which $\mathbf{C}$ becomes an $\mathcal{O}$-algebra) is a one-dimensional complex manifold. While a curve is often described by equations like (1), it is to be viewed up to isomorphism, as an equivalence class of such equations modulo suitable changes of variables. The main objects we will study are curves X over $\mathrm{Spec}(\mathcal{O})$, and the behaviour of the sets $X(R)$ as R ranges over different $\mathcal{O}$-algebras.

Remark. The term "integral points on elliptic curves" is often used (particularly by number theorists) to refer to the integral solutions of an affine Weierstrass equation:

$$E_0 : y^2 = x^3 + ax + b$$

which describes an *affine curve* over the base ring $\mathbf{Z}[a, b]$. This is an abuse of terminology, since elliptic curves are always defined as projective varieties by passing to the projective equation

$$E : y^2 z = x^3 + axz^2 + bz^3,$$

resulting in the addition of the "point at infinity" $O := (0, 1, 0)$ to E_0. This passage is crucial. Note, for instance, that E has the structure of an algebraic group, while E_0 does not. It should be kept in mind that the common usage "integral points on E" refers to the integral points on the affine curve $E_0 = E - \{O\}$, which is not an elliptic curve at all, and that, according to the definitions in standard usage, $E(\mathcal{O})$ is equal to $E(K)$ because E is projective.

The fundamental trichotomy for curves

Suppose that the curve X is *generically smooth*, i.e., its generic fiber is a nonsingular curve over K, so that $X(\mathbf{C})$ has the structure of a smooth Riemann surface. The set $X(\mathbf{C})$ is (topologically and analytically) identified with

$$X(\mathbf{C}) \simeq S - \{P_1, \ldots, P_s\},$$

where S is a compact Riemann surface (of genus g, say) and $P_1, \ldots, P_s$ are distinct points. The invariants g and s, which completely determine the topological isomorphism class of $X(\mathbf{C})$, can be packaged into the *Euler characteristic*

$$\chi(X) = 2 - 2g - s.$$

The answers to Questions 1—4 above depend on the sign of $\chi(X)$ in an essential way.

I. Positive Euler characteristic. If $\chi(X) > 0$, then $g = 0$ and $s = 0$ or 1. Therefore X is isomorphic over $\bar{K}$ either to the projective line $\mathbb{P}_1$ or the affine line

$\mathbb{A}^1$. Forms of $\mathbb{P}_1$ over K correspond to conics, for which one has the following basic result.

THEOREM 5. *Let X be a smooth conic over K. The following are equivalent.*

(a) *The curve X is isomorphic over K to $\mathbb{P}_1$.*
(b) *The set $X(K)$ is nonempty.*
(c) *The set $X(K_v)$ is nonempty, for all completions K_v of K.*

The equivalence between (a) and (b) follows from the Riemann–Roch theorem: given a rational point $\infty \in X(K)$, there is a rational function with only a simple pole at ∞; such a function gives an isomorphism between X and $\mathbb{P}_1$ over K. The equivalence between (b) and (c) is the Hasse–Minkowski theorem, one of the most basic instances of the so-called *local-global principle* which is discussed at greater length elsewhere in this volume.

REMARK 6. The proof of the Hasse–Minkowski theorem, which relies on the geometry of numbers, leads to an upper bound on the smallest height of a point on $X(K)$, and thus is effective. Attempts to generalise Theorem 5 to higher dimensional varieties have led to a rich theory which forms the basis for some of the articles in this volume.

The case of positive Euler characteristic, for which the basic questions 1—4 are in some sense well-understood thanks to Theorem 5, will not be treated any further in these notes.

II. Euler characteristic zero. There are two types of curve with Euler characteristic zero:

- The affine case: $g = 0$ and $s = 2$.
- The projective case: $g = 1$ and $s = 0$.

The prototypical example of the affine case is when

$$X = \mathbb{P}_1 - \{0, \infty\} = \mathbb{G}_m.$$

The set $X(\mathcal{O}) = \mathcal{O}^\times$ is an abelian group under multiplication, and X is naturally equipped with the structure of a commutative group scheme over $\mathcal{O}$. Something similar happens in the projective case: since X is a curve of genus one, it is isomorphic over K either to an elliptic curve, if $X(K) \neq \emptyset$, or to a principal homogeneous space over such a curve. For the following theorem, suppose that $X(\mathcal{O}) \neq \emptyset$, and that X can be equipped with the structure of a group scheme over $\mathcal{O}$.

THEOREM 7. *The group $X(\mathcal{O})$ is finitely generated.*

In the affine case, Theorem 7 is essentially Dirichlet's S-unit theorem, while in the projective case it corresponds to the Mordell–Weil Theorem that the group of rational points on an elliptic curve over a number field is finitely generated.

III. Negative Euler characteristic. The theory of curves with negative Euler characteristic is dominated by the following basic finiteness result.

THEOREM 8. *If $\chi(X) < 0$, then $X(\mathcal{O})$ is finite.*

In the affine case this is a theorem of Siegel proved in 1929. In the interesting special case where $X = \mathbb{P}_1 - \{0, 1, \infty\}$, the points in $X(\mathcal{O})$ correspond to solutions of the so-called *S-unit equation*

$$u + v = 1 \quad \text{with } u, v \in \mathcal{O}^\times.$$

In the projective case Theorem 8 used to be known as the *Mordell Conjecture*. Its proof by Faltings in 1983 represents a significant achievement in the Diophantine theory of curves.

We now describe the contents of these notes.

Section 1 recalls some preliminary results that are used heavily in later sections: the main finiteness results of algebraic number theory, and the method of descent based on unramified coverings and the Chevalley–Weil theorem. Hugo Chapdelaine's article [**Chaa**] in these proceedings further develops these themes by describing a relatively elementary application of Faltings' theorem to a Diophantine equation—the *generalised Fermat equation* $x^p + y^q + z^r = 0$—that appears to fall somewhat beyond the scope of the study of algebraic curves, but to which, it turns out, the "fundamental trichotomy" described in this introduction can still be applied.

The main goal of Section 2 is to give a survey of Faltings' proof of the Mordell Conjecture. In many ways, this section forms the heart of these notes. The ideas in Section 2 are used to motivate the startlingly diverse array of techniques that arise in the Diophantine study of curves. These techniques are deployed in subsequent sections to study several important and illustrative classes of algebraic curves—specifically, modular curves, Fermat curves, and elliptic curves.

Section 3 focuses on what may appear at first glance to be a rather special collection of algebraic curves, the so-called *modular curves* over $\mathbf{Q}$ classifying isomorphism classes of elliptic curves with extra level structure. Singling out modular curves for careful study can be justified on (at least) two grounds.

(1) They are the simplest examples of *moduli spaces*. Classifying the rational points on modular curves translates into "uniform boundedness" statements for the size of torsion subgroups of elliptic curves over $\mathbf{Q}$, and therefore leads to nontrivial results concerning rational points on curves of genus one.
(2) Modular curves are also the simplest examples of *Shimura varieties*, and their Jacobians and ℓ-adic cohomology are closely tied to spaces of modular forms. (It is from this connection that they derive their name.) This makes it feasible to address finer questions about the rational points on modular curves, following a line of attack that was initiated by Mazur [**Maz77**] in his landmark paper on the Eisenstein ideal.

Section 3 attempts to convey some of the flavour of Mazur's approach by describing a simple but illustrative special case of his general results: namely, his proof of the conjecture, originally due to Ogg, that the size of the torsion subgroup of elliptic curves over $\mathbf{Q}$ is uniformly bounded, by 14. The approach we describe incorporates an important strengthening due to Merel exploiting progress on the Birch and Swinnerton-Dyer conjecture that grew out of later work of Gross–Zagier

and Kolyvagin–Logachev. Marusia Rebolledo's article [**Reb**] in these proceedings takes this development one step further by describing Merel's proof of the *strong uniform boundedness conjecture* over number fields: given $d \geq 1$, the modular curves $Y_1(p)$ contain no points of degree d when p is large enough (relative to d).

Section 4 describes the approach initiated by Frey, Serre, and Ribet for reducing Fermat's Last Theorem to deep questions about the relationship between elliptic curves and modular forms. This subject is only lightly touched upon in these notes. Pierre Charollois's article in this volume [**Chab**] describes a technique of Halberstadt and Kraus that strengthens the "modular approach" to prove a result on the generalised Fermat equation $ax^p + by^p + cz^p = 0$ that is notable for its generality. This result also suggests that it might be profitable to view the modular approach as part of a general method, rather than just a serendipitous "trick" for proving Fermat's Last Theorem.

Section 5 gives a rapid summary of the author's second week of lectures at the Göttingen summer school, devoted largely to curves of genus 1, particularly elliptic curves. This section is less detailed than the others, partly because it covers topics that have already been treated elsewhere, notably in [**Dar04**]. The main topics that are touched upon (albeit briefly) in Section 5 are:

(1) The collection of Heegner points on a modular elliptic curve, and Kolyvagin's use of them to prove essentially all of the Birch and Swinnerton-Dyer conjecture for elliptic curves with analytic rank ≤ 1. Kolyvagin's techniques also supply a crucial ingredient in Merel's proof of the uniform boundedness conjecture, further justifying its inclusion as a topic in the present notes. The article by Samit Dasgupta and John Voight [**DV**] in these proceedings describes an application of the theory of Heegner points to Sylvester's conjecture on the primes that can be expressed as a sum of two rational cubes.
(2) Variants of the modular parametrisation which can be used to produce more general systems of algebraic points on elliptic curves over $\mathbf{Q}$. Such systems are likely to continue to play an important role in further progress on the Birch and Swinnerton-Dyer conjecture. A key example is the fact that many elliptic curves defined over totally real fields are expected to occur as factors of the Jacobians of Shimura curves attached to certain quaternion algebras. The articles by John Voight [**Voi**] and Matthew Greenberg [**Greb**] in these proceedings discuss the problem of calculating with Shimura curves and their associated parametrisations from two different angles: from the point of view of producing explicit equations in [**Voi**], and relying on the Cherednik–Drinfeld p-adic uniformisation in [**Greb**].
(3) The theory of Stark–Heegner points, which is meant to generalise classical Heegner points. Matthew Greenberg's second article [**Grea**] in these proceedings discusses Stark–Heegner points attached to elliptic curves over imaginary quadratic fields. Proving the existence and basic algebraicity properties of the points that Greenberg describes how to calculate numerically would lead to significant progress on the Birch and Swinnerton-Dyer conjecture—at present, there is no elliptic curve that is "genuinely" defined over a quadratic imaginary field for which this conjecture is proved in even its weakest form.

1. Preliminaries

1.1. Zero-dimensional varieties. In order to get a good understanding of algebraic varieties of dimension $d+1$, it is useful to understand the *totality* of algebraic varieties of dimension d. Such a principle is hardly surprising, since a $(d+1)$-dimensional variety can be expressed as a family of d-dimensional varieties, parametrized by a one-dimensional base. Any discussion of the Diophantine properties of curves must therefore necessarily begin with a mention of the zero-dimensional case.

A zero-dimensional variety (of finite type) over a field K is an affine scheme of the form $X = \mathrm{Spec}(R)$, where R is a finite-dimensional commutative K-algebra without nilpotent elements. Let

$$n := \#X(\bar{K}) = \#\mathrm{Hom}(R, \bar{K}) = \dim_K(R),$$

where $\bar{K}$ denotes as usual an algebraic closure of the field K. Finding the rational points on X amounts to solving a degree n polynomial in one variable over K.

An *integral model* of X over $\mathcal{O}$ is an affine scheme of the form $\mathrm{Spec}(R_{\mathcal{O}})$, where $R_{\mathcal{O}} \subset R$ is an $\mathcal{O}$-algebra satisfying $R_{\mathcal{O}} \otimes_{\mathcal{O}} K = R$. Such a model is said to be *smooth* if $R_{\mathcal{O}}$ is finitely generated as an $\mathcal{O}$-module and $R_{\mathcal{O}}/\mathfrak{p}$ is a ring without nilpotent elements for all $\mathfrak{p} \in \mathrm{Spec}(\mathcal{O})$. The reader can check that X has a smooth model over $\mathrm{Spec}(\mathcal{O})$ if and only if $R = \prod_i L_i$ is a product of field extensions L_i/K which are unramified outside of S.

It is of interest to consider the collection of zero-dimensional varieties of fixed cardinality n which possess a smooth model over $\mathrm{Spec}(\mathcal{O})$. The following classical finiteness result is extremely useful in the study of curves.

THEOREM 1.1 (Hermite–Minkowski). *Given n and $\mathcal{O} = \mathcal{O}_{K,S}$, there are finitely many isomorphism classes of varieties of cardinality n over K which possess a smooth model over* $\mathrm{Spec}(\mathcal{O})$. *Equivalently, there are finitely many field extensions of K of degree at most n which are unramified outside of S.*

The proof is explained, for example, in [**Szp85**], p. 91. In the simplest special case where $K = \mathbf{Q}$ and $S = \emptyset$, we mention the following more precise statement:

THEOREM 1.2 (Minkowski). *Any zero–dimensional variety over* $\mathbf{Q}$ *which has a smooth model over* $\mathrm{Spec}(\mathbf{Z})$ *is isomorphic to* $\mathrm{Spec}(\mathbf{Q}^n)$ *for some $n \geq 1$. Equivalently, there are no nontrivial everywhere unramified field extensions of* $\mathbf{Q}$.

1.2. Etale morphisms and the Chevalley–Weil theorem. If $\pi : X \longrightarrow Y$ is a nonconstant, finite morphism of projective curves defined over K (or of affine curves over $\mathcal{O} = \mathcal{O}_{K,S}$), then π induces finite-to-one maps $\pi_K : X(K) \longrightarrow Y(K)$ and $\pi_{\mathcal{O}} : X(\mathcal{O}) \longrightarrow Y(\mathcal{O})$. In particular, if $Y(K)$ is finite, then so is $X(K)$. This simple principle reduces the study of rational points on a curve X to the often simpler study of points on the image curve Y. (For instance, the genus of Y is less than or equal to the genus of X, by the Riemann–Hurwitz formula.) As a historical illustration, Fermat proved that the equation $x^4 + y^4 = z^4$ (which corresponds to a projective curve of genus 3 over $\mathbf{Q}$) has no nontrivial rational points by studying the integer solutions of the auxiliary equation $x^4 + y^4 = z^2$ which are *primitive* in the sense of [**Chaa**]. These primitive solutions correspond to rational points on a curve of genus one (in line with the principles explained in [**Chaa**]), and Fermat was able to dispose of these rational points by his method of descent.

In contrast, the finiteness of $X(K)$ does not imply the finiteness of $Y(K)$ in general, because the maps π_K or $\pi_{\mathcal{O}}$ need not be surjective, and in fact are usually far from being so. The following weakening of the notion of surjectivity is frequently useful in practice.

DEFINITION 1.3. The map $\pi : X \longrightarrow Y$ of curves over $\mathrm{Spec}(\mathcal{O}_{K,S})$ is said to be *almost surjective* if there is a finite extension L of K and a finite set T of primes of L containing the primes above those in S, such that $Y(\mathcal{O}_{K,S})$ is contained in the image of $\pi_{\mathcal{O}_{L,T}}$.

DEFINITION 1.4. A morphism $\pi : X \longrightarrow Y$ of curves over $\mathrm{Spec}(\mathcal{O}_{K,S})$ is said to be *generically étale* if it satisfies any of the following equivalent conditions:

(a) The induced map $\pi_{\mathbf{C}} : X(\mathbf{C}) \longrightarrow Y(\mathbf{C})$ is an unramified covering of Riemann surfaces;
(b) The map $\pi_K : X_K \longrightarrow Y_K$ is an étale morphism of K-varieties on the generic fibers;
(c) There exists a finite set $S' \supset S$ of primes of K such that the map $\pi_{\mathcal{O}_{K,S'}} : X_{\mathcal{O}_{K,S'}} \to Y_{\mathcal{O}_{K,S'}}$ is a finite étale morphism of schemes over $\mathrm{Spec}(\mathcal{O}_{K,S'})$.

The following result, known as the *Chevalley–Weil theorem*, gives a criterion for a map π to be almost surjective.

THEOREM 1.5 (Chevalley–Weil). *If the morphism π is generically étale, then it is almost surjective.*

PROOF. Suppose that π is generically étale. By Property (c) in the definition, we may suitably enlarge S so that the map π becomes étale over $\mathrm{Spec}(\mathcal{O}_{K,S})$. If P belongs to $Y(\mathcal{O}) = \mathrm{Hom}(\mathrm{Spec}(\mathcal{O}), Y)$, let $P^*(X) = \pi^{-1}(P)$ denote the fiber of π above P. This fiber can be described as a scheme over $\mathrm{Spec}(\mathcal{O})$ by viewing P as a morphism $\mathrm{Spec}(\mathcal{O}) \longrightarrow Y$, and $\pi^{-1}(P)$ as the scheme-theoretic pullback of π to $\mathrm{Spec}(\mathcal{O})$ via P, for which the following diagram is cartesian

$$\begin{array}{ccc} P^*(X) & \longrightarrow & X \\ \downarrow & & \downarrow \\ \mathrm{Spec}(\mathcal{O}) & \xrightarrow{P} & Y. \end{array}$$

Note that $\pi^{-1}(P)$ is a zero-dimensional scheme over $\mathrm{Spec}(\mathcal{O})$ of cardinality $n = \deg(\pi)$, which is smooth because π is étale. By the Hermite–Minkowski theorem (Theorem 1.1) there are finitely many possibilities for $\pi^{-1}(P)$, as P ranges over $Y(\mathcal{O})$. Hence the compositum L of their fraction fields is a finite extension of K. Let T denote the set of primes of L above those in S. Then, by construction, $Y(\mathcal{O})$ is contained in $\pi(X(\mathcal{O}_{L,T}))$. It follows that π is almost surjective. □

EXAMPLE 1.6. *The Klein and Fermat curves.* The quartic curve

$$Y : x^3y + y^3z + z^3x = 0 \tag{3}$$

studied by Felix Klein is a curve of genus 3 having an automorphism group $G = \mathbf{PSL}_2(\mathbb{F}_7)$ of order 168. By the Hurwitz bound, this is the largest number of automorphisms a curve of genus 3 may have. (A curve with this property is in fact unique up to $\bar{\mathbf{Q}}$-isomorphism.) The curve Y is also a model for the *modular curve* $X(7)$. (Cf. Section 4.1 for a brief discussion of $X(n)$.) The automorphism group $\mathbf{PSL}_2(7)$ arises from the transformations that preserve the fibers of the natural

projection of $Y(7)$ onto the j-line. In [**Hur08**], Hurwitz proved that Y has no nontrivial rational points, as follows: let (x, y, z) be a point on the Klein quartic with integer coordinates, satisfying $\gcd(x, y, z) = 1$. Although x, y and z have no common factor, they need not be pairwise coprime; setting

$$u = \gcd(x, y), \quad v = \gcd(y, z), \quad w = \gcd(z, x),$$

one sees (after changing the signs of u, v, and/or w if necessary) that

$$(x, y, z) = (u^3 w, v^3 u, w^3 v) =: \pi(u, v, w). \tag{4}$$

Substituting back into the original equation (3) and dividing by $u^3 v^3 w^3$, one finds that (u, v, w) is a rational point on the Fermat curve of degree 7:

$$X : u^7 + v^7 + w^7 = 0.$$

Through this argument, Hurwitz showed that the map $\pi : X \longrightarrow Y$ given by (4), a generically étale map of degree 7, is almost surjective (in fact, surjective) on rational points. This is a simple special case of Theorem 1.5. Hurwitz then applied Lamé's result for the Fermat equation of degree 7 to conclude that the Klein quartic has no integer solutions except the trivial ones.

Note that this example gives a nontrivial Diophantine relation between modular curves and Fermat curves. More sophisticated connections between these two classes of curves are discussed in Section 4.

Example 1.7. *Algebraic groups.* Recall that $\mathcal{O}$ is the ring of S-integers of a number field K. Let G be any commutative group scheme of finite type over $\mathrm{Spec}(\mathcal{O})$. Then for any integer $n \geq 1$, the morphism $[n]$ given by $g \mapsto g^n$ is generically étale (more precisely, étale over $\mathrm{Spec}(\mathcal{O}[1/n])$). Therefore, the Chevalley-Weil theorem implies that there is a *finite* extension L of K for which $G(\mathcal{O})/nG(\mathcal{O})$ maps to the kernel of the natural map $G(K)/nG(K) \longrightarrow G(L)/nG(L)$. A standard construction shows that this kernel injects into the finite group $H^1(\mathrm{Gal}(L/K), G[n](L))$, where $G[n](L)$ is the finite group of n-torsion points on $G(L)$. It follows that $G(\mathcal{O})/nG(\mathcal{O})$ is finite. (When $G = \mathbb{G}_m$, this statement is a weak form of Dirichlet's S-unit theorem, while when $G = A$ is an elliptic curve or an abelian variety, it is the *weak Mordell–Weil theorem* asserting that $A(K)/nA(K)$ is finite.)

Example 1.8. It is not hard to exhibit a projective curve X of genus greater than 1 equipped with a map $\pi : X \longrightarrow \mathbb{P}_1$ which is unramified outside $\{0, 1, \infty\}$. Examples include

(a) The Fermat curve $x^n + y^n = z^n$ with $\pi(x, y, z) = x^n/z^n$;
(b) The modular curves $X_0(n)$ and $X_1(n)$ introduced in Section 3.1, with their natural maps to the j-line.

One can use the map π to show that Theorem 8 for projective curves (Faltings' Theorem) implies the case $X = \mathbb{P}_1 - \{0, 1, \infty\}$ over $\mathrm{Spec}(\mathcal{O})$ of Theorem 8 (Siegel's Theorem).

More generally, a celebrated theorem of Belyi asserts that *any* projective curve X/K can be equipped with a morphism $\pi : X \longrightarrow \mathbb{P}_1$ which is unramified outside $\{0, 1, \infty\}$. (See Hugo Chapdelaine's article in these proceedings.) This fact has been exploited by Elkies [**Elk91**] to prove that the abc conjecture implies Faltings' theorem.

Further topic: Hugo Chapdelaine's article in these proceedings explains how the discussion of unramified coverings and the Chevalley–Weil Theorem can be adapted to treat the primitive solutions of the *generalised Fermat equation* $x^p + y^q + z^r = 0$. The reader who has mastered the ideas in Section 1 may skip directly to this article if so inclined.

2. Faltings' theorem

This section is devoted to explaining the main ideas in Faltings' proof of the Mordell Conjecture (Theorem 8 for projective curves over number fields).

THEOREM 2.1 (Faltings). *Let X be a smooth projective curve of genus ≥ 2 defined over a number field K. Then $X(K)$ is finite.*

The proof will be presented as a series of reductions.

2.1. Prelude: the Shafarevich problem. The first of these reductions, explained in Section 2.2, reduces Theorem 2.1 to a finiteness conjecture of Shafarevich. The Shafarevich problem is concerned with the collection of all arithmetic objects sharing certain common features and having "good reduction" over the ring $\mathcal{O}$ of S-integers of a number field K, taken, of course, up to isomorphism over K. Some key examples are:

(1) the set $\mathcal{F}_d(\mathcal{O})$ of smooth zero-dimensional schemes over $\mathrm{Spec}(\mathcal{O})$ of cardinality d;
(2) the set $\mathcal{M}_g(\mathcal{O})$ of smooth curves of genus g over $\mathrm{Spec}(\mathcal{O})$;
(3) the set $\mathcal{A}_g(\mathcal{O})$ of abelian schemes of dimension g over $\mathrm{Spec}(\mathcal{O})$;
(4) the set $\mathcal{I}_g(\mathcal{O})$ of K-isogeny classes of abelian varieties of dimension g over $\mathrm{Spec}(\mathcal{O})$.

The following question is known as the Shafarevich problem:

QUESTION 2.2. *How large are the sets above? Are they finite?*

One can also ask what happens for specific values of K and S, the most interesting special case being $\mathcal{O} = \mathbf{Z}$ (i.e., $K = \mathbf{Q}$ and $S = \emptyset$).

We now discuss these questions for the various cases listed above:

(1) The set $\mathcal{F}_d(\mathcal{O})$ corresponds to the set of étale K-algebras (i.e., products of separable field extensions) of rank d over K which are unramified outside S. The finiteness of $\mathcal{F}_d(\mathcal{O})$ is just a restatement of the Hermite–Minkowski Theorem (Theorem 1.1).
(2) The set $\mathcal{M}_0(\mathcal{O})$ consists of the set of smooth conics over K which have good reduction outside of S. It admits a cohomological interpretation, via the exact sequence

$$0 \longrightarrow \mathcal{M}_0(\mathcal{O}) \longrightarrow H^2(K, \pm 1) \longrightarrow \bigoplus_{v \notin S} H^2(K_v, \pm 1).$$

The fundamental results of local and global class field theory imply that $\mathcal{M}_0(\mathcal{O})$ is finite, and in fact, its order can be evaluated precisely:

$$\#\mathcal{M}_0(\mathcal{O}) = 2^{\#S + r - 1},$$

where r is the number of real places of K. In particular, when $K = \mathbf{Q}$ and $S = \emptyset$, then $\mathcal{M}_0(\mathbf{Z})$ consists of one element, corresponding to the projective line $\mathbb{P}_1$ over $\mathbf{Q}$.

(3) The set $\mathcal{M}_1(\mathcal{O})$ can be infinite; in fact, an infinite set of curves of genus 1 which are all isomorphic over $\bar{K}$ and have good reduction outside of S can sometimes be found, even if S consists of just one prime of K. (See [**Maz86**], p. 241.) On the other hand, a deep conjecture of Shafarevich and Tate implies that $\mathcal{M}_1(\mathcal{O})$ is finite if S is empty. Also, if one replaces $\mathcal{M}_1$ by the set $\mathcal{E}$ of K-isomorphism classes of elliptic curves, i.e., curves of genus 1 equipped with a K-rational point, then Shafarevich [**Šaf63**] showed that $\mathcal{E}(\mathcal{O})$ is always finite.

When $g > 1$, the following conjecture of Shafarevich can be viewed as a one-dimensional analogue of the Hermite–Minkowski theorem (Theorem 1.1):

CONJECTURE 2.3. *Let $g \geq 2$ be an integer, and let $\mathcal{O}$ be the ring of S-integers of a number field K, for a finite set S of primes of K.*

(a) *(Shafarevich conjecture for curves). The set $\mathcal{M}_g(\mathcal{O})$ is finite, i.e., there are only finitely many K-isomorphism classes of curves of genus g defined over K and having good reduction outside of S.*
(b) *(Shafarevich conjecture for abelian varieties). The set $\mathcal{A}_g(\mathcal{O})$ is finite, i.e., there are only finitely many isomorphism classes of abelian varieties of dimension g defined over K and having good reduction outside of S.*
(c) *(Shafarevich conjecture for isogeny classes). The set $\mathcal{I}_g(\mathcal{O})$ is finite, i.e., there are only finitely many K-isogeny classes of abelian varieties of dimension g with good reduction outside of S.*

REMARK 2.4. It is a deep theorem of Fontaine [**Fon85**] that the sets $\mathcal{A}(\mathbf{Z})$ and $\mathcal{M}_g(\mathbf{Z})$ are *empty* for $g \geq 2$, i.e., there are no abelian varieties, or smooth curves of genus ≥ 2, over $\mathrm{Spec}(\mathbf{Z})$.

2.2. First reduction: the Kodaira–Parshin trick. In [**Par68**], Parshin showed that part (a) of Conjecture 2.3 implies Theorem 2.1.

THEOREM 2.5. *(Kodaira–Parshin). The Shafarevich conjecture for curves implies Mordell's conjecture.*

SKETCH OF PROOF. Let X be a curve of genus $g > 1$ defined over a number field K. To each point $P \in X(K)$ one associates a curve X_P and a covering map $\phi_P : X_P \longrightarrow X$ with the following properties:

(1) The curve X_P and the map ϕ_P can be defined over a finite extension K' of K which does not depend on P.
(2) The genus g' of X_P (and the degree of ϕ_P) is fixed and in particular does not depend on P.
(3) The map ϕ_P is ramified only over the point P.
(4) The curve X_P has good reduction outside a finite set of primes S' of K' which does not depend on P.

For a description of this assignment, see [**Maz86**], p. 243-244, [**FWG+92**], p. 191-197, or [**Par68**]. The reader should note that one has some leeway in constructing it, and that different versions appear in the literature.

We will describe one approach here, which consists in considering the embedding $X \longrightarrow J$ of X into its Jacobian that sends P to the origin of J, and letting $\tilde{X}$ be the pullback to X of the multiplication-by-2 map $[2] : J \longrightarrow J$. This map induces an unramified covering $\pi : \tilde{X} \longrightarrow X$ of degree 2^{2g}, and hence the genus

of $\tilde{X}$ can be calculated explicitly using the Riemann–Hurwitz formula. The fiber $\pi^{-1}(P)$ can be written as

$$\pi^{-1}(P) = \tilde{P} + D,$$

where $\tilde{P}$ corresponds to the identity element of J, and hence belongs to $\tilde{X}(K)$, and D is an effective divisor of degree $2^{2g}-1$ defined over K with support disjoint from $\tilde{P}$. Let J_D be the *generalised Jacobian* attached to $\tilde{X}$ and D: the group $J_D(\bar{K})$ is identified with the group of degree zero divisors on $\tilde{X}$ with support outside D, modulo the subgroup of principal divisors of the form $\mathrm{div}(f)$, as f ranges over the functions satisfying $f(D_0) = 1$, for all degree zero divisors D_0 supported on D. The functor $L \mapsto J_D(\bar{K})^{G_L}$ (where $G_L := \mathrm{Gal}(\bar{L}/L)$) on finite extensions of K is representable by the algebraic group over K denoted J_D, which is an extension of J by a torus T over K of rank $(2^{2g}-2)$. In other words, there is a natural exact sequence

$$1 \longrightarrow T \longrightarrow J_D \longrightarrow J \longrightarrow 1$$

of commutative algebraic groups over K.

One can embed $\tilde{X} - D$ into J_D by sending a point Q to the equivalence class of the divisor $(Q) - (\tilde{P})$. The multiplication-by-2 map $[2]$ on J_D induces a map $X_P^0 \longrightarrow \tilde{X} - D$, as summarised by the following diagram with Cartesian squares in which the vertical maps are induced by multiplication by 2:

$$\begin{array}{ccccc} J_D & \longleftarrow & X_P^0 & & \\ \downarrow & & \downarrow & & \\ J_D & \longleftarrow & \tilde{X} - D & \longrightarrow & J \\ & & \downarrow & & \downarrow \\ & & X & \longrightarrow & J. \end{array} \tag{5}$$

The closure X_P of X_P^0 has the desired properties 1-4: it is defined over K, and it follows directly from the Riemann–Hurwitz formula that its genus g' does not depend on P. Furthermore, the map $X_P^0 \longrightarrow \tilde{X} - D$ is unramified, and hence $X_P \longrightarrow X$ is ramified only over the point P. Finally, if X is smooth over $\mathrm{Spec}(\mathcal{O})$, the curve X_P has a smooth model over $\mathcal{O}' := \mathcal{O}[1/2]$.

The assignment $P \mapsto X_P$ therefore gives rise to a well-defined map

$$R_1 : X(K) \longrightarrow \mathcal{M}_{g'}(\mathcal{O}').$$

But this assignment is finite-to-one; for otherwise there would be a curve Y and infinitely many (by property 3) distinct maps $\phi_P : Y \longrightarrow X$. This would contradict the following geometric finiteness result of De Franchis (cf. [**Maz86**], p. 227).

THEOREM 2.6. *If X and Y are curves over any field K, and Y has genus $g \geq 2$, then the set* $\mathrm{Mor}_K(X, Y)$ *of K-morphisms from X to Y is finite.*

The Shafarevich conjecture for curves, which asserts the finiteness of $\mathcal{M}_{g'}(\mathcal{O}')$, therefore implies the finiteness of $X(K)$. This completes the proof of Theorem 2.5. □

REMARK 2.7. The reader will note that the proof of Theorem 2.5 breaks down (as it should!) when $g = 1$, because the set $\mathrm{Mor}_K(Y, X)$ can be (and in fact, frequently is) infinite when X has genus 1.

2.3. Second reduction: passing to the Jacobian. The second step in the proof of the Mordell conjecture consists in observing that the Shafarevich conjecture for curves would follow from the corresponding statement for abelian varieties.

PROPOSITION 2.8. *The Shafarevich conjecture for curves follows from the Shafarevich conjecture for abelian varieties.*

To prove Proposition 2.8, one studies the map R_2 which associates to a curve X its Jacobian J. If X is smooth over $\mathrm{Spec}(\mathcal{O})$, the same is true of J, and hence R_2 defines a map $\mathcal{M}_g(\mathcal{O}) \longrightarrow \mathcal{A}_g(\mathcal{O})$. Key to Proposition 2.8 is the following corollary of Torelli's theorem:

THEOREM 2.9. *If $g \geq 2$, then the map R_2 is finite-to-one.*

PROOF. Torelli's theorem asserts that a curve X of genus ≥ 2 can be recovered by the data of its Jacobian J together with the principal polarisation associated to the Riemann theta-divisor. But a given abelian variety can carry only finitely many principal polarisations. (See [**CS86**] for a more detailed exposition of the Torelli Theorem and surrounding concepts.) □

2.4. Third reduction: passing to isogeny classes. The third, crucial and more difficult reduction was carried out by Faltings himself.

THEOREM 2.10. *(Faltings). The Shafarevich conjecture for abelian varieties follows from the Shafarevich conjecture for isogeny classes.*

As one would expect, the proof is based on showing that the natural map $R_3 : \mathcal{A}_g(\mathcal{O}) \longrightarrow \mathcal{I}_g(\mathcal{O})$ has finite fibers. This is a consequence of the following key result:

THEOREM 2.11. *(Faltings) There are finitely many isomorphism classes of abelian varieties over K in a given K-isogeny class.*

This result is the technical heart of Faltings' proof, and rests on his theory of heights on moduli spaces of abelian varieties. Things become somewhat simpler if we assume that the abelian varieties in the isogeny class are semistable. This can be assumed without loss of generality because of Grothendieck's semistable reduction theorem which asserts that every abelian variety becomes semistable after a finite extension of the ground field (for instance, one over which the points of order 3 become rational). For a finite extension K'/K, there are finitely many K-isomorphism classes of abelian varieties that are K'-isomorphic to a given abelian variety over K', and hence the finiteness of the K-isogeny class follows from that of any K'-isogeny class.

Faltings defines a height function (now called the Faltings height) of an abelian variety. We will not dwell on the definition, but will content ourselves with stating two of its main finiteness properties:

THEOREM 2.12. *Let K be a number field and H be a positive constant. There are finitely many isomorphism classes of g-dimensional abelian varieties over K with height less than H.*

The second finiteness property concerns the behaviour of the Faltings height on a K-isogeny class. Given a prime ℓ, the ℓ-isogeny class of an abelian variety A is the set of abelian varieties which are isogenous to A via an isogeny of ℓ-power degree.

More generally, if M is any finite set of rational primes, two abelian varieties are said to be M-isogenous if they are related by a K-isogeny whose degree is a product of primes in M.

THEOREM 2.13. *If A is a semistable abelian variety over a number field K, then:*

(1) *There exists a finite set M of rational primes, depending only on the isogeny class of A, such that if $A \longrightarrow B$ is a K-isogeny of degree not divisible by the primes in M, then*
$$h(A) = h(B).$$
(2) *For any finite set S of rational primes, the Faltings height is bounded on S-isogeny classes.*

The proof of this theorem relies on deep results of Tate and Raynaud on group schemes and p-divisible groups; cf. Theorems 2.4 and 2.6 of [**Del85**].

For more details on the proof of theorems 2.12 and 2.13 see the expositions [**CS86**], [**FWG$^+$92**], [**Szp85**], [**Del85**], or [**ZP89**]. Note that these two theorems together imply:

PROPOSITION 2.14. *Let A be a semistable abelian variety over K, and let M be as in part 1 of Theorem 2.13.*

(1) *Up to K-isomorphism, there are finitely many abelian varieties that are K-isogenous to A via an isogeny of degree not divisible by the primes in M.*
(2) *Given any abelian variety B over K and any finite set S of rational primes, there are finitely many abelian varieties in the S-isogeny class of B.*

Proof of Theorem 2.11: Let $\phi : A \longrightarrow B$ be a K-isogeny. We can write ϕ as a composition of isogenies
$$A \xrightarrow{\phi_o} B_0 \xrightarrow{\phi_1} B_1,$$
where ϕ_0 is of degree not divisible by the primes in M, and ϕ_1 is an M-isogeny. By part 1 of Proposition 2.14, there are finitely many possibilities for ϕ_0 and for B_0. By part 2 of this proposition, for each B_0 there are finitely many possibilities for B_1. Theorem 2.11 follows.

2.5. Fourth reduction: from isogeny classes to ℓ-adic representations. To an abelian variety A over K of dimension g and a prime ℓ, one can associate the ℓ-adic Tate module and ℓ-adic representations
$$T_\ell(A) := \varprojlim A[\ell^n], \qquad V_\ell(A) := T_\ell(A) \otimes \mathbf{Q}_\ell,$$
where the inverse limit is taken with respect to the multiplication-by-ℓ maps. The $\mathbf{Q}_\ell$ vector space $V_\ell(A)$ is $2g$-dimensional and is equipped with a $\mathbf{Q}_\ell$-linear action by two *commuting* $\mathbf{Q}_\ell$-algebras E and Π_K defined by
$$E = \mathrm{End}_K(A) \otimes \mathbf{Q}_\ell, \qquad \Pi_K := \mathbf{Z}_\ell[[G_K]] \otimes \mathbf{Q}_\ell.$$
Here $\mathbf{Z}_\ell[[G_K]]$ denotes the profinite group ring $\varprojlim \mathbf{Z}_\ell[\mathrm{Gal}(L/K)]$, where the projective limit is taken over all finite Galois extensions $L \subset \bar{K}$ of K.

If A and B are K-isogenous abelian varieties, they give rise to ℓ-adic representations that are isomorphic as Π_K-modules. In other words, the assignment $A \mapsto V_\ell(A)$ yields a map

$$R_4 : \mathcal{I}_g(\mathcal{O}) \longrightarrow \left\{ \begin{array}{l} \text{Isomorphism classes of} \\ 2g\text{-dimensional } \ell\text{-adic} \\ \text{representations of } \Pi_K \end{array} \right\}.$$

The strategy will now consist in showing that R_4 has finite fibers, and finally in describing the image R_4 precisely enough to show that it is finite.

We begin by introducing some further notations and recalling some background. Given a prime v of K, let $I_v \subset G_v \subset G_K$ be the inertia and decomposition subgroups of G_K attached to v. Note that the groups G_v and I_v are only well-defined up to conjugation in G_K, since they depend on a choice of a prime of $\bar{K}$ above v. The quotient G_v/I_v is procyclic with a canonical generator Frob_v called the *Frobenius element* at v, which induces the automorphism $x \mapsto x^{\mathbf{N}v}$ on the residue field, where $\mathbf{N}v$ denotes the norm of v (the cardinality of the associated residue field).

If V is any finite-dimensional $\mathbf{Q}_\ell$-vector space equipped with a continuous Π_K-action, we say that V is *unramified* at v if I_v acts trivially on V. When this happens, the Frobenius element $\mathrm{Frob}_v \in G_v/I_v$ gives an element of $\mathbf{GL}(V)$ which is well-defined up to conjugation in this group.

The following theorem lists some of the basic properties of $V_\ell(A)$.

THEOREM 2.15. *Let A be an abelian scheme over $\mathrm{Spec}(\mathcal{O}_{K,S})$. The ℓ-adic Galois representation $V_\ell(A)$ satisfies the following properties:*

1. *It is semisimple as a representation of E.*
2. *It is unramified at all $v \notin S' := S \cup \{\lambda|\ell\}$.*
3. *(Rationality) If $v \notin S'$, then the characteristic polynomial of Frob_v has rational integer coefficients. The complex roots of this polynomial have absolute value $\mathbf{N}v^{1/2}$.*
4. *(Tate conjecture) The representation $V_\ell(A)$ is semisimple as a representation of Π_K.*

Property (1) follows from the basic theory of duality for abelian varieties, and properties (2) and (3) were shown by Weil (cf. [**Wei48**]). Property (4), a particular case of the Tate conjecture, is one of Faltings' important contributions. We now explain how Faltings proved the semisimplicity of $V_\ell(A)$ over Π_K, adapting an idea used by Tate to prove the corresponding statement over finite fields.

LEMMA 2.16. *For every Π_K-invariant subspace W in $V_\ell(A)$, there is an element $u \in E$ such that*

$$uV_\ell(A) = W.$$

PROOF. The $\mathbf{Z}_\ell$-module $W_\infty = W \cap T_\ell(A)$ gives rise to a collection of groups $W_n = W_\infty/\ell^n W_\infty \subset A[\ell^n]$ which are defined over K and compatible under the natural maps $A[\ell^{n+1}] \longrightarrow A[\ell^n]$. Let

$$\alpha_n : A \longrightarrow A_n := A/W_n,$$

be the natural isogeny with kernel W_n, and let β_n denote the isogeny characterised by

$$\alpha_n\beta_n = \ell^n, \qquad \beta_n\alpha_n = \ell^n.$$

Note that $\beta_n(A_n[\ell^n]) = W_n$ by construction, in light of the first identity above. By Faltings' finiteness theorem 2.11, there exists an infinite set $I = \{n_0, n_1, \ldots\} \subset \mathbf{Z}^{>0}$ for which there exist isomorphisms

$$\nu_i : A_{n_0} \simeq A_i$$

for all $i \in I$. Now define a sequence of K-endomorphisms of A by the rule

$$u_i := \beta_i \nu_i \alpha_{n_0}.$$

Since $\mathrm{End}_K(A) \otimes \mathbf{Z}_\ell$ is compact in the ℓ-adic topology, the sequence (u_i) has a convergent subsequence $(u_i)_{i \in J}$ in this topology. Let u denote the limit of such a subsequence. After eventually refining J further, we can assume that for each $i \in J$, we have natural maps

$$u(A[\ell^i]) = u_i(A[\ell^i]) \longrightarrow \beta_i(A_i[\ell^i]) = W_i,$$

with kernel and cokernel bounded independently of i, because they arise from α_{n_0}. It follows that

$$u(V_\ell(A)) = W,$$

as was to be shown. □

Corollary 2.17. *The representation $V_\ell(A)$ is a semisimple Π_K-module.*

Proof. Let W be a Π_K-stable subspace of $V_\ell(A)$, and let $u \in E$ be an element constructed in Lemma 2.16, satisfying $u(V_\ell(A)) = W$. Consider the right ideal uE in the algebra E. Because E is semisimple, this ideal is generated by an idempotent u_0. Note that $u_0(V_\ell(A)) = W$. The subspace $\ker(u_0)$ is therefore a Π_K-stable complement of W in $V_\ell(A)$. Hence $V_\ell(A)$ is semisimple over Π_K. □

In conclusion, let $\mathrm{Rep}_S(G_K, 2g)$ be the set of isomorphism classes of rational semisimple ℓ-adic representations of G_K of dimension $2g$ which are unramified outside of S. We have shown that R_4 maps $\mathcal{I}_g(\mathcal{O})$ to $\mathrm{Rep}_S(G_K, 2g)$. To complete the proof of the Mordell conjecture, it remains to show:

(1) The map R_4 is finite-to-one.
(2) The set $\mathrm{Rep}_S(G_K, 2g)$ is finite.

We will prove the first in the next section, and the second in Section 2.7.

2.6. The isogeny conjecture. The proof of the following deep conjecture of Tate is a cornerstone of Faltings' strategy for proving the Mordell conjecture.

Theorem 2.18. *(Isogeny conjecture). Let A and B be abelian varieties defined over a number field K. If $V_\ell(A)$ is isomorphic to $V_\ell(B)$ as a Π_K-module, then the abelian varieties A and B are isogenous.*

In other words, the map R_4 is *injective*.

We first note that Theorem 2.18 can be reduced to the following statement, known as the *Tate conjecture* for abelian varieties.

Theorem 2.19. *(Tate conjecture). Let A and B be abelian varieties defined over K. Then the natural map*

$$\mathrm{Hom}_K(A, B) \otimes \mathbf{Q}_\ell \longrightarrow \mathrm{Hom}_{\Pi_K}(V_\ell(A), V_\ell(B))$$

is surjective.

To see that Theorem 2.19 implies Theorem 2.18, let $j : V_\ell(A) \simeq V_\ell(B)$ be a Π_K-equivariant isomorphism. By Theorem 2.19, this isomorphism comes from an element $u \in \mathrm{Hom}_K(A,B) \otimes \mathbf{Q}_\ell$. After multiplying u by some power of ℓ, we can assume that u belongs to $\mathrm{Hom}_K(A,B) \otimes \mathbf{Z}_\ell$. Note that $\mathrm{Hom}_K(A,B)$ is dense in $\mathrm{Hom}_K(A,B) \otimes \mathbf{Z}_\ell$. Any good enough ℓ-adic approximation to u in $\mathrm{Hom}_K(A,B)$ gives the desired K-isogeny between A and B. Theorem 2.18 follows.

We next observe that Theorem 2.19 can be reduced to the following special case:

THEOREM 2.20. *Let A be an abelian variety over K. The natural map*

$$\mathrm{End}_K(A) \otimes \mathbf{Q}_\ell \longrightarrow \mathrm{End}_{\Pi_K}(V_\ell(A))$$

is surjective.

The fact that Theorem 2.20 implies Theorem 2.19 can be seen by applying Theorem 2.20 to the abelian variety $A \times B$, since

$$\mathrm{End}_K(A \times B) = \mathrm{End}_K(A) \oplus \mathrm{Hom}_K(A,B) \oplus \mathrm{Hom}_K(B,A) \oplus \mathrm{End}_K(B)$$

and likewise for $\mathrm{End}(V_\ell(A \times B)) = \mathrm{End}(V_\ell(A) \times V_\ell(B))$.

Proof of Theorem 2.20: Let ϕ be an element of $\mathrm{End}_{\Pi_K}(V_\ell(A))$, and let

$$W = \{(x, \phi(x)) \in V_\ell(A) \times V_\ell(A)\} \subset V_\ell(A \times A)$$

be the graph of ϕ. Note that W is Π_K-stable. Hence there is an endomorphism $u \in \mathrm{End}_K(A \times A) \otimes \mathbf{Q}_\ell = M_2(E)$ associated to W by Lemma 2.16, satisfying $u(V_\ell(A \times A)) = W$.

Let $E^0 = \mathrm{End}_E(V_\ell(A))$ denote the commutant of E in $\mathrm{End}(V_\ell(A))$. For any $\alpha \in E^0$, the matrix $\left(\begin{smallmatrix} \alpha & 0 \\ 0 & \alpha \end{smallmatrix}\right)$ with entries in $\mathrm{End}(V_\ell(A))$ commutes with $u \in M_2(E) \subset M_2(\mathrm{End}(V_\ell(A))$. It follows that this matrix preserves $W = \mathrm{image}(u)$, and hence α commutes with ϕ. Since this argument is valid for any $\alpha \in E^0$, the endomorphism ϕ belongs to the double commutant E^{00} which is equal to E by the semisimplicity of $V_\ell(A)$ as a module over E.

2.7. The finiteness principle for rational ℓ-adic representations. Now that the map R_4 has been shown to be injective, it remains to prove that the target $\mathrm{Rep}_S(G_K, 2g)$ is finite. The main theorem of this section is:

THEOREM 2.21. *(Finiteness principle for rational semisimple ℓ-adic representations). Let K be a number field and S a finite set of primes of K. Then there are finitely many isomorphism classes of rational, semisimple ℓ-adic representations of G_K of dimension d which are unramified outside of S.*

Remark. The reader will observe that this finiteness principle is close in spirit to the Hermite–Minkowski theorem: it asserts that there are only finitely many extensions of K (albeit, of infinite degree) of a certain special kind with bounded ramification. The proof of Theorem 2.21 will in fact rely crucially on the Hermite–Minkowski theorem, as well as on the Chebotarev density theorem.

We begin by establishing the following key lemma.

LEMMA 2.22. *There exists a finite set T of primes of K (depending on S and d) satisfying the following two properties:*

(1) *T is disjoint from $S_\ell := S \cup \{v|\ell\}$.*

(2) *Two representations* $\rho_1, \rho_2 \in \mathrm{Rep}_S(G_K, d)$ *are isomorphic if and only if*

$$\mathrm{trace}(\rho_1(\mathrm{Frob}_v)) = \mathrm{trace}(\rho_2(\mathrm{Frob}_v)), \quad \textit{for all } v \in T.$$

PROOF. Consider the set of all extensions of K of degree $\leq l^{2d^2}$ which are unramified outside S_ℓ. By Theorem 1.1 (Hermite–Minkowski), there are finitely many such extensions, and hence their compositum L is a finite extension of K. Let $T = \{v_1, \dots v_N\}$ be a set of primes of K which are not in S and such that the Frobenius conjugacy classes Frob_{v_i} generate $\mathrm{Gal}(L/K)$. The existence of such a finite set follows from the Chebotarev density theorem. We claim that this set T satisfies the conclusion of Lemma 2.22. Given $\rho_1, \rho_2 \in \mathrm{Rep}_S(G_K, d)$, a choice of G_K-stable $\mathbf{Z}_\ell$-lattices in the underlying representation spaces makes it possible to view each ρ_i as a homomorphism from $\mathbf{Z}_\ell[[G_K]]$ to $M_d(\mathbf{Z}_\ell)$. Let

$$j = \rho_1 \oplus \rho_2 : \mathbf{Z}_\ell[[G_K]] \longrightarrow M_d(\mathbf{Z}_\ell) \times M_d(\mathbf{Z}_\ell),$$

and let M denote the image of j. The induced homomorphism

$$\bar{j} : G_K \longrightarrow (M/\ell M)^\times$$

factors through $\mathrm{Gal}(L/K)$, since the cardinality of $M/\ell M$ is at most ℓ^{2d^2} and $\bar{j}$ is unramified outside of S_ℓ. It follows that the elements

$$\bar{j}(\mathrm{Frob}_{v_1}), \dots, \bar{j}(\mathrm{Frob}_{v_N})$$

generate $M/\ell M$. By Nakayama's lemma, the elements

$$j(\mathrm{Frob}_{v_1}), \dots, j(\mathrm{Frob}_{v_N})$$

generate M as a $\mathbf{Z}_\ell$-module.

In particular, if

$$\mathrm{trace}(\rho_1(\mathrm{Frob}_{v_j})) = \mathrm{trace}(\rho_2(\mathrm{Frob}_{v_j})), \quad \text{for } j = 1, \dots, N,$$

then

$$M \subseteq \Delta \subset M_d(\mathbf{Z}_\ell) \times M_d(\mathbf{Z}_\ell),$$

where $\Delta = \{(A, B) \text{ such that } \mathrm{trace}(A) = \mathrm{trace}(B)\}$. Therefore one has

$$\mathrm{trace}(\rho_1(\sigma)) = \mathrm{trace}(\rho_2(\sigma)) \quad \text{for all } \sigma \in \Pi_K.$$

Hence ρ_1 and ρ_2 have the same traces. Since they are semisimple, it follows that they are isomorphic as Π_K-representations. □

Proof of Theorem 2.21. Let $T = \{v_1, \dots, v_N\}$ be as in the statement of Lemma 2.22. The assignment

$$\rho \mapsto (\mathrm{Tr}\,(\rho(\mathrm{Frob}_{v_1})), \dots, \mathrm{Tr}\,(\rho(\mathrm{Frob}_{v_N})))$$

is injective on $\mathrm{Rep}_S(G_K, d)$, and can only assume finitely many values, by the rationality of ρ. (More precisely, each $\mathrm{Tr}\,(\mathrm{Frob}_{v_i})$ is a rational integer of absolute value $\leq d\mathbf{N}v_i^{1/2}$.) Theorem 2.21 follows.

2.8. A summary of Faltings' proof. Faltings' proof of Mordell's conjecture is based on a sequence of maps (here X is a curve of genus g defined over K and having good reduction outside of the finite set S of primes of K):

$$\left\{\begin{array}{l}K\text{-rational}\\ \text{points on } X\end{array}\right\} \xrightarrow{R_1} \left\{\begin{array}{l}\text{Curves of genus } g' \text{ over } K'\\ \text{with good reduction outside } S'\end{array}\right\}$$

$$\xrightarrow{R_2} \left\{\begin{array}{l}\text{Isomorphism classes of semistable}\\ \text{abelian varieties of dimension } g'\\ \text{with good reduction outside } S'\end{array}\right\}$$

$$\xrightarrow{R_3} \left\{\begin{array}{l}\text{Isogeny classes of abelian varieties}\\ \text{of dimension } g'\\ \text{with good reduction outside } S'\end{array}\right\}$$

$$\xrightarrow{R_4} \left\{\begin{array}{l}\text{Rational semisimple } \ell\text{-adic representations}\\ \text{of dimension } 2g' \text{ unramified outside } S'_\ell\end{array}\right\}$$

(1) The map R_1 is given by Parshin's construction, and is finite-to-one, by the geometric theorem of De Franchis.
(2) The map R_2 is defined by passing to the Jacobian of a curve, and is finite-to-one by Torelli's theorem.
(3) The map R_3 is the obvious one, and is finite-to-one, by Faltings' fundamental Theorem 2.11 on finiteness of abelian varieties in a given isogeny class.
(4) The map R_4 is defined by passing to the Tate module, and is one-to-one, thanks to the Tate conjectures proved by Faltings. The proof of the Tate conjectures is obtained by combining a strategy of Tate with the finiteness Theorem 2.11. These ideas are also used to show that the Galois representations arising in the image of R_4 are *semisimple*.
(5) The last set in this sequence of maps is finite by the finiteness principle for rational semisimple ℓ-adic representations, which is itself a consequence of the Chebotarev density theorem and the Hermite–Minkowski theorem.

3. Modular curves and Mazur's theorem

The first step in the proof of the Mordell conjecture (the Kodaira–Parshin reduction) consists in transforming a question about rational points on a given curve into the Shafarevich conjecture. This new Diophantine question is concerned with the moduli space of curves themselves, to which an array of techniques (notably, Jacobians, ℓ-adic representations, etc.) can be applied. It is therefore apparent that the extra structures afforded by moduli spaces are of great help in studying the Diophantine questions that are associated to them. So it is natural to examine more closely the simplest class of moduli spaces, which are also curves in their own right: the *modular curves* classifying elliptic curves with extra level structure.

3.1. Modular curves. Let p be a prime ≥ 5, and write Z for the ring $\mathbf{Z}[1/p]$. The functor $Y_1(p)$ which to any Z-algebra R associates the set of R-isomorphism classes of pairs (E, P) where E is an elliptic curve over $\mathrm{Spec}(R)$ and P is a point of order p on E_R is representable by a smooth affine scheme over $\mathrm{Spec}(Z)$ of relative dimension one, denoted $Y_1(p)$.

The group $(\mathbf{Z}/p\mathbf{Z})^\times$ acts on $Y_1(p)$ by the rule $t \cdot (E, P) := (E, tP)$, and the quotient of $Y_1(p)$ by this action is an affine scheme $Y_0(p)$ over $\mathrm{Spec}(Z)$ which is a

coarse moduli scheme classifying pairs (E, C) consisting of an elliptic curve over R and a cyclic subgroup scheme $C \subset E$ of order p defined over R.

These curves admit analytic descriptions as quotients of the Poincaré upper half-plane

$$\mathcal{H} = \{\tau \in \mathbf{C}, \qquad \mathrm{Im}(\tau) > 0\}$$

by the action of the following discrete subgroups of $\mathbf{SL}_2(\mathbf{Z})$:

$$\begin{aligned} \Gamma_1(p) &= \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{with } a-1 \equiv c \equiv d-1 \equiv 0 \pmod{p} \right\}, \\ \Gamma_0(p) &= \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{with } c \equiv 0 \pmod{p} \right\}. \end{aligned}$$

For example, the curve $Y_0(1)$ is identified with $\mathrm{Spec}(Z[j])$, and a birational (singular, even on the generic fiber) model for $Y_0(p)$ over $\mathrm{Spec}(Z)$ is given by

$$\mathrm{Spec}(Z[j, j']/\Phi_p(j, j')),$$

where $\Phi_p(x, y) \in \mathbf{Z}[x, y]$ is the canonical *modular polynomial* of bidegree $p + 1$ satisfying $\Phi_p(j(\tau), j(p\tau)) = 0$, for all $\tau \in \mathcal{H}$.

A rational point on $Y_1(p)$ (resp. on $Y_0(p)$) determines an elliptic curve over $\mathbf{Q}$ with a $\mathbf{Q}$-rational point of order p (resp. a rational subgroup of order p). The main goal of this section is to explain the proof of the following theorem of Mazur.

THEOREM 3.1. *If $p > 13$, then $Y_1(p)(\mathbf{Q}) = \emptyset$.*

REMARK 3.2. Note that Theorem 3.1 can be viewed as a theorem about curves in two different ways. Firstly, it asserts that the collection of modular curves $Y_1(p)$, whose genera grow with p, have no rational points once p is large enough—a type of statement that is similar in flavour to Fermat's Last Theorem. Secondly, it leads to the *uniform boundedness* of the size of the torsion subgroups $E(\mathbf{Q})_{\mathrm{tors}}$ as E ranges over all elliptic curves over $\mathbf{Q}$, and is therefore also a theorem about curves of genus one.

3.2. Mazur's criterion. An important role is played in Mazur's argument by the compactification $X_0(p)$ of the affine curve $Y_0(p)$. As a Riemann surface, $X_0(p)(\mathbf{C})$ is obtained by adjoining to $Y_0(p)$ a finite set of cusps which are in bijection with the orbits of $\Gamma_0(p)$ acting on $\mathbb{P}_1(\mathbf{Q})$ by Möbius transformations. More precisely, letting $\mathcal{H}^* := \mathcal{H} \cup \mathbb{P}_1(\mathbf{Q})$, we have

$$X_0(p)(\mathbf{C}) = \Gamma_0(p)\backslash\mathcal{H}^* = (\Gamma_0(p)\backslash\mathcal{H}) \cup \{0, \infty\} = Y_0(p)(\mathbf{C}) \cup \{0, \infty\}.$$

The complex structure in a neighbourhood of ∞ is defined by letting $q = e^{2\pi i\tau}$ be a local parameter at ∞.

The equation for the universal elliptic curve in a formal punctured neighbourhood of ∞ is given by the Tate curve

$$E_q = Z[[q]]^\times/q^{\mathbf{Z}} : y^2 + xy = x^3 + a(q)x + b(q) \qquad \text{over } Z((q)),$$

where

$$a(q) = -5\sum_{n=1}^{\infty} \sigma_3(n)q^n, \quad b(q) = -\frac{1}{12}\sum_{n=1}^{\infty}(7\sigma_5(n) + 5\sigma_3(n))q^n.$$

(Recall that $\sigma_k(n) = \sum_{d|n} d^k$.) The discriminant of E_q is equal to

$$\Delta(E_q) = q \prod_{n \geq 1} (1 - q^n)^{24},$$

and therefore E_q defines an elliptic curve over $Z((q))$.

The important *q-expansion principle* asserts that the parameter q is also a local parameter for the scheme $X_0(p)_Z$ in a neighborhood of ∞. Thanks to the q-expansion principle, the completion of the local ring of $X_0(p)_Z$ at ∞ is identified with the power series ring $Z[[q]]$:

$$\hat{\mathcal{O}}_{X_0(p),\infty} = Z[[q]].$$

A basic technique in Mazur's proof is to study the behaviour of certain maps on modular curves, via their behaviour in a formal neighbourhood of ∞. The following definition will be useful.

DEFINITION 3.3. A morphism $j : X \longrightarrow Y$ of schemes over Z is a *formal immersion* at $x \in X(Z)$ if the induced map on completed local rings

$$j^* : \hat{\mathcal{O}}_{Y,j(x)} \longrightarrow \hat{\mathcal{O}}_{X,x}$$

is surjective.

Let $J_0(p)$ denote the Jacobian of $X_0(p)$. It is an abelian variety over Z and is equipped with an embedding

$$\Phi : X_0(p) \longrightarrow J_0(p)$$

defined by letting $\Phi(x)$ be the class of the degree zero divisor $(x) - (\infty)$.

If $J_\sharp(p)$ is any quotient of $J_0(p)$, let $j_\sharp : X_0(p) \longrightarrow J_\sharp(p)$ be the map obtained by composing Φ with the projection to $J_\sharp(p)$. The following criterion of Mazur for $Y_1(p)(\mathbf{Q}) = \emptyset$ is the main result of this section.

THEOREM 3.4. *Assume that $p > 7$. Suppose that there is an abelian variety quotient $J_\sharp(p)$ of $J_0(p)$ satisfying the following conditions:*

(a) *The map $j_\sharp : X_0(p) \longrightarrow J_\sharp(p)$ is a formal immersion at ∞.*
(b) *$J_\sharp(p)(\mathbf{Q})$ is finite.*

Then $Y_1(p)(\mathbf{Q}) = \emptyset$.

SKETCH OF PROOF. Let $\tilde{x}$ be a point in $Y_1(p)(\mathbf{Q})$ corresponding to the pair (E, P), where E is an elliptic curve over $\mathbf{Q}$ and $P \in E(\mathbf{Q})$ is of order p. Let $\mathcal{E}$ be the minimal Weierstrass model of E over Z.

The proof is divided into four steps.

Step 1. If E has potentially good reduction at the prime 3, then the special fiber $\mathcal{E}_{\mathbb{F}_3}$ is either an elliptic curve, or an extension of a finite group of connected components of cardinality $2^a 3^b$ by the additive group $\mathbb{G}_{a/\mathbb{F}_3}$. Such a group cannot contain a point of order $p > 7$, by the Hasse bound. Hence E has potentially multiplicative reduction at 3.

Step 2. Let $x \in X_0(p)(\mathbf{Q})$ be the image of $\tilde{x}$ under the natural map: it corresponds to the pair $(E, \langle P \rangle)$ consisting of the curve E and the cyclic subgroup generated by P. By Step 1, the point x reduces to one of the cusps 0 or ∞ of $X_0(p)$ modulo 3. It can be assumed without loss of generality that x reduces to ∞, by replacing $(E, \langle P \rangle)$ by $(E/\langle P \rangle, E[p]/\langle P \rangle)$ otherwise.

Step 3. Consider the element $j_\sharp(x) \in J_\sharp(p)(\mathbf{Q})$. By step 2 this element belongs to the formal group $J^1_\sharp(p)(\mathbf{Q}_3)$, which is torsion-free because $\mathbf{Q}_3$ is absolutely unramified. It also belongs to $J_\sharp(p)(\mathbf{Q})$, which is torsion by assumption. It follows that $j_\sharp(x) = 0$.

Step 4. We now use the fact that $j_\sharp$ is a formal immersion to deduce that $x = \infty$. To see this, let $\mathrm{Spec}(R)$ be an affine neighborhood of ∞ containing x. The point x gives rise to a ring homomorphism $x : R \longrightarrow \mathbf{Z}_3$, which factors through the local ring $\hat{\mathcal{O}}_{X_0(p),\infty} = Z[[q]]$, so that x can be viewed as a map $Z[[q]] \longrightarrow \mathbf{Z}_3$. By step 3, we have

$$x \circ j_\sharp^* = \infty \circ j_\sharp^*.$$

It follows that $x = \infty$, since $j_\sharp^*$ was assumed to be surjective, contradicting the initial assumption that x belongs to $Y_0(p)$. □

Mazur's criterion reduces Theorem 3.1 to the problem of exhibiting a quotient $J_\sharp(p)$ of $J_0(p)$ satisfying the conditions of Theorem 3.4.

3.3. The Jacobian $J_0(p)$. The fact that makes it possible to analyse the Jacobian $J_0(p)$ precisely, and exhibit a nontrivial quotient of it with finite Mordell–Weil group, arises from two related ingredients.

(a) *Hecke operators.* If n is an integer that is not divisible by p, the modular curve $X_0(np)$ is equipped with two maps π_1, π_2 to $X_0(p)$, defined by

$$\pi_1(E, C) = (E, C[p]), \qquad \pi_2(E, C) = (E/C[n], C/C[n]).$$

The pair (π_1, π_2) gives rise to an embedding of $X_0(pn)$ in the product $X_0(p) \times X_0(p)$. The image in this product, denoted T_n, is an algebraic correspondence on $X_0(p)$ defined over $\mathbf{Q}$, which gives rise to an endomorphism of $J_0(p)$ defined over $\mathbf{Q}$. On the level of divisors, T_n is described by

$$T_n(E, C) = \sum_{E \longrightarrow E'} (E', C'), \tag{6}$$

where the sum is taken over the cyclic isogenies $\varphi : E \longrightarrow E'$ of degree n, and $C' = \varphi(C)$. Let $\mathbf{T}$ denote the subring of $\mathrm{End}_{\mathbf{Q}}(J_0(p))$ generated by the Hecke operators T_n. It is finitely generated (as a ring, and even as a module) over $\mathbf{Z}$. Our basic approach to constructing $J_\sharp(p)$ is to use the endomorphisms in $\mathbf{T}$ to decompose the abelian variety $J_0(p)$ (up to $\mathbf{Q}$-isogeny) into smaller pieces which can then be analysed individually. If R is any ring, let $\mathbf{T}_R$ denote the R-algebra $\mathbf{T} \otimes R$.

(b) *Modular forms.* If R is any Z-algebra, let $S_2(p, R)$ denote the space of regular differentials on $X_0(p)_R$. Restriction to the formal neighborhood $\mathrm{Spec}(R[[q]])$ of $\infty \in X_0(p)$ gives rise to a map (called the *q-expansion map*)

$$q\text{-exp} : S_2(p, R) \longrightarrow R[[q]]dq.$$

When $R = \mathbf{C}$, the space $S_2(p, \mathbf{C})$ is identified with the vector space of homomorphic functions $f : \mathcal{H} \longrightarrow \mathbf{C}$ for which

(i) the differential $2\pi i f(\tau)d\tau$ is invariant under $\Gamma_0(p)$, i.e.,

$$f\left(\frac{a\tau + b}{c\tau + d}\right) = (c\tau + d)^2 f(\tau), \qquad \text{for all } \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(p).$$

(ii) $2\pi i f(\tau)d\tau$ extends to a holomorphic differential on the compactified modular curve $X_0(p)$. In particular, it admits a Fourier expansion of the form

$$f(\tau) = \sum_{n=1}^{\infty} a_n e^{2\pi i n\tau},$$

so that $q\text{-exp}(2\pi i f(\tau)d\tau) = \sum_{n=1}^{\infty} a_n q^n \frac{dq}{q}$.

The action of the Hecke operators T_n on $J_0(p)_R$ induces an action on the cotangent space $S_2(p, R)$, which can be described explicitly on the level of the q-expansions. For example, if $\ell \neq p$ is prime,

$$T_\ell \left(\sum_{n=1}^{\infty} a_n q^n \frac{dq}{q} \right) = \left(\sum_{\ell | n} a_n q^{n/\ell} + \ell \sum_{n=1}^{\infty} a_n q^{n\ell} \right) \frac{dq}{q}. \tag{7}$$

There is an extra Hecke operator T_p defined via an algebraic correspondence

$$X_0(p^2) \subset X_0(p) \times X_0(p)$$

which admits the following simpler formula for its action on q-expansions:

$$T_p \left(\sum_{n=1}^{\infty} a_n q^n \frac{dq}{q} \right) = \left(\sum_{p | n} a_n q^{n/p} \right) \frac{dq}{q}. \tag{8}$$

The definition of T_ℓ for ℓ prime can then be extended to all integers n by the multiplicativity relations implicit in the following identity of formal Dirichlet series:

$$\sum_{n \geq 1} T_n n^{-s} = (1 - T_p p^{-s})^{-1} \prod_{\ell \neq p} (1 - T_\ell \ell^{-s} + \ell^{1-2s})^{-1}. \tag{9}$$

In other words,

$$T_{mn} = T_m T_n \text{ if } \gcd(m, n) = 1, \qquad T_{\ell^{n+1}} = a_\ell T_{\ell^n} - \ell T_{\ell^{n-1}}.$$

PROPOSITION 3.5. *The algebra $\mathbf{T_Q}$ is a commutative semisimple algebra of dimension $g := \dim_{\mathbf{Q}} S_2(p, \mathbf{Q}) = \text{genus}(X_0(p))$.*

SKETCH. The fact that $\mathbf{T_Q}$ is commutative follows from the explicit description of the operators T_n as correspondences given in (6) (or, if one prefers, from equation (7) describing its effect on q-expansions). The semisimplicity arises from the fact that the operators T_n are self-adjoint with respect to the Hermitian pairing on $S_2(p, \mathbf{C})$ (Petersson scalar product) defined by

$$\langle \omega_1, \omega_2 \rangle = \frac{1}{2i} \int_{\Gamma_0(p) \backslash \mathcal{H}} \omega_1 \wedge \bar{\omega}_2.$$

(We mention in passing that in general, the operator T_ℓ acting on $S_2(N, \mathbf{C})$ need not be self-adjoint when $\ell | N$, but it is self-adjoint when restricted to the space of so-called *newforms*. We are using implicitly the fact that $S_2(p, \mathbf{C})$ is equal to its subspace of newforms.) One computes the dimension of $\mathbf{T_Q}$ by showing that the bilinear pairing

$$\mathbf{T_Q} \times S_2(p, \mathbf{Q}) \longrightarrow \mathbf{Q}, \qquad (T, f) := a_1(Tf)$$

is left and right nondegenerate, and in fact positive definite. The details are left to the reader (who may also consult Section 2.2 of [**Dar04**] for more details). □

As a consequence of Proposition 3.5 and its proof, one has the decomposition

$$\mathbf{T}_{\mathbf{Q}} = K_1 \times \cdots \times K_t$$

of $\mathbf{T}_{\mathbf{Q}}$ into a product of totally real fields, with $\sum_{j=1}^{t}[K_j : \mathbf{Q}] = n$. The factors K_j are indexed by:

(a) The points $\phi_1, \ldots, \phi_t$ of $\mathrm{Spec}(\mathbf{T}_{\mathbf{Q}})$, viewed as algebra homomorphisms $\phi_j : \mathbf{T}_{\mathbf{Q}} \longrightarrow \bar{\mathbf{Q}}$ (taken modulo the natural action of $G_{\mathbf{Q}} = \mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q})$).
(b) The distinct $G_{\mathbf{Q}}$-equivalences classes $f_1, \ldots, f_t$ of eigenforms for $\mathbf{T}$, normalised so that $a_1(f_j) = 1$. The q-expansions of these eigenforms are described by

$$f_j = \sum_{n=1}^{\infty} \phi_j(T_n) q^n.$$

The quotient A_f attached to f is defined by letting

$$A_f := J_0(p)/I_f, \qquad \text{where } I_f := \ker(\mathbf{T} \longrightarrow K_f).$$

With these notations, the main result of this section is the following *Eichler-Shimura decomposition*, which asserts that $J_0(p)$ is isogenous to a product of $\mathbf{Q}$-simple factors indexed by the ($G_{\mathbf{Q}}$-orbits of) normalised eigenforms f_j ($j = 1, \ldots, t$).

THEOREM 3.6. *The abelian variety $J_0(p)$ is $\mathbf{Q}$-isogenous to the product*

$$\prod_{i=1}^{t} A_{f_i},$$

of $\mathbf{Q}$-simple abelian varieties A_{f_i}. The varieties A_f that occur in this decomposition have the following properties:

(a) $\dim(A_f) = [K_f : \mathbf{Q}]$;
(b) *The natural image of $\mathbf{T}_{\mathbf{Q}}$ in $\mathrm{End}_{\mathbf{Q}}(A_f) \otimes \mathbf{Q}$ is isomorphic to K_f.*

For more details on this decomposition see Chapter 2 of [**Dar04**].

Thanks to Theorem 3.6, we are reduced to the following question:

QUESTION 3.7. *Find a criterion involving the normalised eigenform f for the quotient A_f to have finite Mordell–Weil group.*

3.4. The Birch and Swinnerton-Dyer conjecture. The key to bounding the rank of $A_f(\mathbf{Q})$ (and showing that this rank is zero, for a sufficiently large collection of normalised eigenforms f) lies in studying the so-called *Hasse–Weil L-series* attached to A_f.

Let A be an abelian variety (of dimension d, say) defined over $\mathbf{Q}$. The Hasse–Weil L-series of A is most conveniently defined in terms of the ℓ-adic representation $V_\ell(A)$ that was introduced in Section 2.5. If $p \neq \ell$ is a prime, the Frobenius element acts naturally on on the space $V_\ell(A)^{I_p}$ of vectors in $V_\ell(A)$ that are fixed under the action of the inertia group at p. (Recall that $V_\ell(A)^{I_p} = V_\ell(A)$ if A has good reduction at $p \neq \ell$.) By the rationality of the representation $V_\ell(A)$, the characteristic polynomial

$$F_p(T) := \det(1 - \mathrm{Frob}_p|_{V_\ell(A)^{I_p}} T)$$

has integer coefficients. Furthermore, it does not depend on the choice of ℓ, and can therefore be defined for all p. This makes it possible to define the Hasse–Weil

L-series as a function of the complex variable s, by the infinite product

$$L(A,s) = \prod_p F_p(p^{-s})^{-1}.$$

Using the rationality of the Galois representation $V_\ell(f)$ in the sense of Theorem 2.15, one can show that the infinite product defining $L(A,s)$ converges uniformly on compact subsets of $\{s \in \mathbf{C} | \Re(s) > 3/2\}$, and hence defines an analytic function in this region.

Concerning the behaviour of $L(A,s)$ and its connection to the arithmetic of A over $\mathbf{Q}$, there are the following two fundamental conjectures:

CONJECTURE 3.8. *The L-series $L(A,s)$ has an analytic continuation to the entire complex plane and a functional equation of the form*

$$\Lambda(A,s) := (2\pi)^{-ds}\Gamma(s)^d N^{s/2} L(A,s) = \pm\Lambda(A, 2-s),$$

where N is the conductor *of A.*

In particular, if Conjecture 3.8 is true, the process of analytic continuation gives meaning to the behaviour of $L(A,s)$ in a neighborhood of the central critical point $s = 1$ for the functional equation, and in particular, the order of vanishing of $L(A,s)$ at $s = 1$ is defined. The *Birch and Swinnerton-Dyer conjecture* relates this order of vanishing to the arithmetic of A over $\mathbf{Q}$:

CONJECTURE 3.9. *If A is an abelian variety over $\mathbf{Q}$, then*

$$\mathrm{rank}(A(\mathbf{Q})) = \mathrm{ord}_{s=1}(L(A,s)).$$

In particular, $A(\mathbf{Q})$ is finite if $L(A,1) \neq 0$.

Both Conjectures 3.8 and 3.9 are far from being proved in general. But much more is known when $A = A_f$ occurs in the Eichler–Shimura decomposition of the modular Jacobian $J_0(N)$, as will be explained in the next section.

3.5. Hecke theory. A *newform* of level N is a normalised eigenform $f = \sum_{n\geq 1} a_n q^n \frac{dq}{q}$ on $\Gamma_0(N)$ whose associated sequence $(a_n)_{(n,N)=1}$ of Fourier coefficients is different from that of any eigenform g on $\Gamma_0(d)$ with $d|N$ and $d \neq N$.

To each newform $f = \sum_{n\geq 1} a_n q^n \frac{dq}{q} \in S_2(N, \mathbf{C})$, one can associate an L-series

$$L(f,s) := \sum_{n=1}^{\infty} a_n n^{-s}.$$

This L-series enjoys the following properties, which were established by Hecke:

(a) **Euler product**: It admits the Euler product factorisation given by

$$L(f,s) = \prod_{p\nmid N} (1 - a_p p^{-s} + p^{1-2s})^{-1} \prod_{p|N} (1 - a_p p^{-s})^{-1},$$

as can be seen by applying φ_f to the formal identity (9) expressing the Hecke operators T_n in terms of the operators T_ℓ for ℓ prime.

(b) **Integral representation**: The L-series $L(f,s)$ can be represented as an integral transform of the modular form f, by the formula:

$$\Lambda(f,s) := (2\pi)^{-s}\Gamma(s) N^{s/2} L(f,s) = N^{s/2} \int_0^\infty f(it) t^{s-1} dt, \tag{10}$$

where $\Gamma(s) = \int_0^\infty e^{-t}t^{s-1}dt$ is the Γ-function. In particular, because f is of rapid decay at the cusps, this integral converges absolutely to an analytic function of $s \in \mathbf{C}$.

(c) **Functional equation**: The involution w defined on $S_2(N, \mathbf{C})$ by the rule

$$w(f)(\tau) = \frac{1}{N\tau^2} f\left(\frac{-1}{N\tau}\right) \tag{11}$$

commutes with the Hecke operators and hence preserves its associated eigenspaces. It follows that for the eigenform f,

$$w(f) = \varepsilon f, \quad \text{where } \varepsilon = \pm 1. \tag{12}$$

The L-series $L(f, s)$ satisfies the functional equation

$$\Lambda(f, s) = -\Lambda(w(f), 2 - s) = -\varepsilon\Lambda(f, 2 - s). \tag{13}$$

It is a direct calculation to derive this functional equation from the integral representation of $\Lambda(f, s)$.

For the next result, we view f as an element of $S_2(N, K_f)$. (Recall that K_f is the totally real field generated by the Fourier coefficients of f.) Any complex embedding $\sigma : K_f \hookrightarrow \mathbf{C}$ yields an eigenform f^σ with complex coefficients, to which the Hecke L-function $L(f^\sigma, s)$ may be attached. The following result relates the L-series of Hasse–Weil and of Hecke.

THEOREM 3.10. *Let A_f be the abelian variety associated to the newform $f \in S_2(N, \mathbf{C})$ by the Eichler–Shimura construction. Then*

$$L(A_f, s) = \prod_{\sigma : K_f \longrightarrow \mathbf{C}} L(f^\sigma, s).$$

In particular, Conjecture 3.8 holds for A_f.

The main ingredient in the proof of Theorem 3.10 is the Eichler–Shimura congruence which relates the Hecke correspondence $T_p \subset X_0(N)^2$ in characteristic p to the graph of the Frobenius morphism and its transpose. For more details and references see Chapter 2 of [**Dar04**].

Theorem 3.10 reveals that one has better control of the arithmetic of the abelian varieties A_f—Conjecture 3.8 remains open for the general abelian variety A over $\mathbf{Q}$. In fact, one has the following strong evidence for Conjecture 3.9 for the abelian varieties A_f.

THEOREM 3.11. *If $L(A_f, 1) \neq 0$, then $A_f(\mathbf{Q})$ is finite.*

The main ingredients that go into the proof of Theorem 3.11 are

(1) The theory of Heegner points on modular curves;
(2) The theorem of Gross–Zagier expressing the canonical heights of the images of these points in A_f in terms of special values of L-series closely related to $L(f, s)$;
(3) A theorem of Kolyvagin which relates the system of Heegner points and the arithmetic of A_f over $\mathbf{Q}$.

These ingredients will be discussed in somewhat more detail in Section 5 devoted to elliptic curves and the Birch and Swinnerton-Dyer conjecture.

3.6. The winding quotient. The criterion for the finiteness of $A_f(\mathbf{Q})$ supplied by Theorem 3.11 allows us to construct a quotient $J_\sharp(p)$ which is in some sense the "largest possible" quotient with finite Mordell–Weil group.

We construct $J_\sharp(p)$, following Merel, by letting e_0 be the vertical path from 0 to $i\infty$ on $\mathcal{H}$; its image in $X_0(p)$ gives an element in the relative homology $H_1(X_0(p)(\mathbf{C}), \mathbf{Z}; \{\text{cusps}\})$. By a result of Manin–Drinfeld, the element e_0 gives rise to an element e in the rational homology $\mathbb{H} := H_1(X_0(p)(\mathbf{C}), \mathbf{Q})$. This element is referred to as the *winding element*.

The Hecke algebra $\mathbf{T} = \mathbf{T}_\mathbf{Q}$ acts on $\mathbb{H}$ by functoriality of correspondences. Let e_f denote the image of e in $\mathbb{H}/I_f\mathbb{H}$. The integral formula (10) for $L(f, s)$ shows that

$$e_f \neq 0 \quad \text{if and only if } L(f, 1) \neq 0.$$

Hence it is natural to define

$$J_e(p) := J_0(p)/I_e, \quad \text{where } I_e := \mathrm{Ann}_\mathbf{T}(e).$$

THEOREM 3.12. *The Mordell–Weil group $J_e(p)(\mathbf{Q})$ is finite.*

PROOF. Up to isogeny, $J_e(p)$ decomposes as

$$J_e(p) \sim \prod_{e_f \neq 0} A_f = \prod_{L(f,1)\neq 0} A_f.$$

Theorem 3.11 implies that $A_f(\mathbf{Q})$ is finite for all the f that appear in this decomposition. The theorem follows. □

In order to exploit Mazur's criterion with $J_\sharp(p) = J_e(p)$, and thereby prove Theorem 3.1, it remains to show that the natural map $j_e : X_0(p) \longrightarrow J_e(p)$ is a formal immersion at ∞. (So that in particular $J_e(p)$ is nontrivial, which is not clear *a priori* from its definition!) This is done in the article of Marusia Rebolledo in these proceedings (cf. Theorem 4 of Section 2.3 of [**Reb**]). Rebolledo's article goes significantly further by showing that the natural map from the d-th symmetric power $X_0(p)^{(d)}$ of $X_0(p)$ sending $(P_1, \ldots, P_d)$ to the image in $J_\sharp(p)$ of the divisor class $(P_1) + \cdots + (P_d) - d(\infty)$ is a formal immersion at $(\infty, \ldots, \infty)$, as soon as p is sufficiently large relative to d.

REMARK 3.13. Our presentation of Mazur's argument incorporates an important simplification due to Merel, which consists in working with the winding quotient $J_e(p)$ whose finiteness is known thanks to Theorem 3.11. At the time of Mazur's original proof described in [**Maz77**], Theorems 3.11 and 3.12 were not available, and Mazur's approach worked with the so-called *Eisenstein quotient* $J_{\text{eis}}(p)$. This quotient contains a rational torsion subgroup of order $n = \text{numerator}(\frac{p-1}{12})$, and one of the key results in [**Maz77**] is to establish the finiteness of $J_{\text{eis}}(\mathbf{Q})$ by an n-descent argument. In Merel's approach, Mazur's somewhat delicate "Eisenstein descent" is in effect replaced by Kolyvagin's descent based on Heegner points and the theorem of Gross–Zagier.

3.7. More results and questions. By various refinements of the techniques discussed above, Mazur was able to classify all possible rational torsion subgroups of elliptic curves over $\mathbf{Q}$ and obtained the following results:

THEOREM 3.14. *Let T be the torsion subgroup of the Mordell–Weil group of an elliptic curve E over $\mathbf{Q}$. Then T is isomorphic to one of the following 15 groups:*

$$\begin{array}{cc} \mathbf{Z}/m\mathbf{Z} & \text{for} \quad 1 \leq m \leq 10 \text{ or } m = 12, \\ \mathbf{Z}/2m\mathbf{Z} \times \mathbf{Z}/2\mathbf{Z} & \text{for} \quad 1 \leq m \leq 4. \end{array}$$

For the proof, see [**Maz77**], p. 156. We note in passing that all possibilities for T that are not ruled out by Mazur's theorem do in fact occur infinitely often: the associated modular curves are of genus 0 and have a rational point.

Mazur's theorem implies that rational points of order p on elliptic curves cannot occur for $p > 7$. One can ask similar questions for rational subgroups. In this direction, Mazur proved the following result in [**Maz78**].

THEOREM 3.15. *Suppose that there is an elliptic curve E over $\mathbf{Q}$ with a rational subgroup of prime order p. Then $p \leq 19$ or $p = 37, 43, 67$, or 163.*

The four exceptional values of p in Theorem 3.15 correspond to discriminants of imaginary quadratic fields of class number one. The corresponding elliptic curves with complex multiplication can be defined over $\mathbf{Q}$ and have a rational subgroup of order p.

Theorem 3.15 implies that for large enough p, the Galois representation

$$\rho_{E,p} : G_{\mathbf{Q}} \longrightarrow \mathrm{Aut}(E[p])$$

is always *irreducible*. One can also ask whether, for large enough p, this Galois representation is in fact necessarily *surjective*. The existence of elliptic curves with complex multiplication, for which $\rho_{E,p}$ is *never* surjective when $p \geq 3$, precludes an affirmative answer to this question. Discarding elliptic curves with complex multiplication, the following conjecture (which appears in [**Ser72**], p. 299, §4.3, phrased more prudently as an open question) can be proposed:

CONJECTURE 3.16. *(Surjectivity conjecture) If E is an elliptic curve over $\mathbf{Q}$ without complex multiplication, and $p \geq 19$ is prime, then the Galois representation associated to $E[p]$ is surjective.*

The surjectivity conjecture remains open, more than 30 years after [**Maz77**]. The hypothetical cases that are the most difficult to dispose of are those where the image of $\rho_{E,p}$ is contained in the normalizer of a Cartan subgroup, particularly a nonsplit Cartan subgroup.

It is also natural to search for analogues of Theorem 3.14 over number fields other than $\mathbf{Q}$; a remarkable breakthrough was achieved on this problem by S. Kamienny and Merel around 1992 ([**Kam92**], [**Mer96**]).

THEOREM 3.17. *Let K be a number field. Then the size of $E(K)_{\mathrm{tors}}$ is bounded by a constant $B(K)$ which depends only on K. In fact, this constant can be made to depend only on the degree of K over $\mathbf{Q}$.*

The proof of this theorem is explained in the article by Marusia Rebolledo [**Reb**] in these proceedings.

We finish with a conjecture that can be viewed as a "mod p analogue" of Theorem 2.18 (Tate's isogeny conjecture).

CONJECTURE 3.18. *There exists an integer M such that, for all $p \geq M$, any two elliptic curves E_1 and E_2 over $\mathbf{Q}$ are isogenous if and only if $E_1[p] \simeq E_2[p]$ as $G_{\mathbf{Q}}$-modules.*

This conjecture appears to be difficult. It is not even clear what the best value M might be, assuming it exists. (Calculations of Cremona [**Cre**] based on his complete tables of elliptic curves over $\mathbf{Q}$ of conductor $\leq 30,000$ show that necessarily $M > 13$.) We mention Conjecture 3.18 here because it implies strong results about ternary Diophantine equations analogous to Fermat's Last Theorem, thanks to the methods explained in Section 4.

4. Fermat curves

The purpose of this section is to discuss the Fermat curves

$$F_n : x^n + y^n = z^n,$$

and the proof of Fermat's Last Theorem, that these curves have no nontrivial rational points when $n \geq 3$. Fermat's Last Theorem has the same flavour as Mazur's Theorem 3.1, since it determines all of the rational points in a naturally arising infinite collection of algebraic curves. Although Fermat curves are simpler to write down as explicit equations, they do not admit a direct moduli interpretation, and therefore turn out to be harder to analyse than modular curves. In fact, the eventual solution of Fermat's Last Theorem is based on an elaborate *reduction* of the study of Fermat curves to Diophantine questions about modular curves. In particular, Theorem 3.1—its statement, as well as some of the techniques used in its proof—play an essential role in the proof of Fermat's Last Theorem.

4.1. Motivation for the strategy. Hugo Chapdelaine's article in these proceedings discusses the more general problem of classifying the primitive integer solutions of the generalised Fermat equation

$$x^p + y^q + z^r = 0, \tag{14}$$

and sets up a "dictionary" relating

$$\left\{ \begin{array}{l} \text{Strategies for studying} \\ \text{primitive solutions of} \\ x^p + y^q + z^r = 0 \end{array} \right\} \quad \text{and} \quad \left\{ \begin{array}{l} \text{Unramified coverings} \\ \text{of } \mathbb{P}_1 - \{0, 1, \infty\} \\ \text{of signature } (p, q, r) \end{array} \right\}.$$

The idea explained in [**Chaa**] is that, given an unramified covering

$$\pi : X \longrightarrow \mathbb{P}_1 - \{0, 1, \infty\},$$

one can study (14) by

(1) Attempting to classify the possible fibers of π over the points in

$$\Sigma_{p,q,r} = \left\{ \frac{a^p}{c^r}, \quad \text{with } a^p + b^q = c^r \quad \text{and } (a, b, c) \text{ primitive} \right\} \subset \mathbb{P}_1(\mathbf{Q}).$$

Since the ramification in these fibers is bounded, there can only be finitely many, by the Hermite–Minkowski theorem. In particular, the compositum of these extensions is a finite extension of $\mathbf{Q}$, denoted L.

(2) Understanding the L-rational points on the curve X.

To apply these principles to the classical Fermat equation, one is led to consider unramified coverings of $\mathbb{P}_1 - \{0, 1, \infty\}$ of signature (p, p, p). Among such coverings, one finds:

(1) the Fermat curve $F_p : x^p + y^p = z^p$ itself, equipped with the natural projection $\pi : (x, y, z) \mapsto t = \frac{x^p}{z^p}$ of degree p^2. For this π, it is clear that $\pi(F_p(\mathbf{Q})) \supset \Sigma_{p,p,p}$; but this merely leads to a tautological reformulation of the original question.
(2) There are many coverings of signature (p, p, p) with solvable Galois groups, and studying these leads to classical attempts to prove Fermat's Last Theorem by factoring $x^p + y^p$ over the p-th cyclotomic fields. This circle of ideas led to many interesting questions on cyclotomic fields and their class groups, but has proved unsuccessful (so far) in settling Fermat's Last Theorem.

A third type of covering is obtained from modular curves. These coverings, which are nonsolvable, arise naturally in light of the strong results obtained in Section 3.

More precisely, let $Y(n)$ be the open modular curve that classifies elliptic curves with full level n structure, i.e., pairs

$$(E, \iota : \mathbf{Z}/n\mathbf{Z} \times \mu_n \longrightarrow E[n])$$

where ι is an identification which induces an isomorphism

$$\bigwedge^2 \iota : \bigwedge^2 (\mathbf{Z}/n\mathbf{Z} \times \mu_n) = \mu_n \simeq \bigwedge^2 (E[n]) = \mu_n.$$

Over the base $Z = \mathbf{Z}[1/2]$, the curve $Y(2)_Z$ is identified with

$$\mathrm{Spec}(Z[\lambda, 1/\lambda, 1/(\lambda - 1)]) = (\mathbb{P}_1 - \{0, 1, \infty\})_Z,$$

where λ is the parameter that occurs in the Legendre family

$$E_\lambda : y^2 = x(x - 1)(x - \lambda).$$

The natural covering map $\pi : Y(2p) \longrightarrow Y(2)$ is an unramified covering of signature (p, p, p), with Galois group $\mathbf{SL}_2(\mathbf{Z}/p\mathbf{Z})/\langle \pm 1 \rangle$. Given $\lambda = \frac{a^p}{c^p} \in \Sigma_{p,p,p}$, the fiber $\pi^{-1}(\lambda)$ is contained in the field of definition of the field of p-division points of the elliptic curve

$$y^2 = x(x - 1)(x - a^p/c^p). \tag{15}$$

In practice, it is more convenient to work with the closely related *Frey curve*,

$$E_{a,b,c} : y^2 = x(x - a^p)(x - c^p),$$

which differs from (15) by a quadratic twist, and replace the study of the fiber of π at λ with considerations involving the *mod p Galois representation*

$$\rho_{a,b,c} : G_{\mathbf{Q}} \longrightarrow \mathrm{Aut}(E_{a,b,c}[p]) \simeq \mathbf{GL}_2(\mathbf{Z}/p\mathbf{Z}).$$

We normalise (a, b, c) so that $a \equiv 3 \pmod 4$ and c is even. (This can always be done, by permuting a, b and c and changing their signs if necessary.) With this normalisation, the minimal discriminant, conductor, and j-invariant associated to $E_{a,b,c}$ are

$$\Delta = 2^{-8}(abc)^{2p}, \qquad N = \prod_{\ell | abc} \ell, \qquad j = \frac{2^8(b^{2p} + a^p c^p)^3}{(abc)^{2p}}. \tag{16}$$

In particular, the elliptic curve $E_{a,b,c}$ is *semistable*: it has either good or (split or nonsplit) multiplicative reduction at all primes. (The reader may wish to consult Section 2 of the article by Pierre Charollois in this proceedings volume, which discusses the local invariants of Frey curves in greater detail.)

4.2. Galois representations associated to Frey curves. The following theorem states the main *local properties* of the Galois representation $\rho_{a,b,c}$.

THEOREM 4.1. *The representation $\rho = \rho_{a,b,c}$ has the following properties.*

(a) *It is unramified outside 2 and p;*

(b) *The restriction of ρ to a decomposition group D_2 at 2 is of the form*

$$\rho_{a,b,c}|_{D_2} = \begin{pmatrix} \chi_{\mathrm{cyc}}\psi & \kappa \\ 0 & \psi^{-1} \end{pmatrix},$$

where $\chi_{\mathrm{cyc}} : G_{\mathbf{Q}_2} \longrightarrow (\mathbf{Z}/p\mathbf{Z})^\times$ is the mod p cyclotomic character, and ψ is an unramified character of order 1 or 2.

(c) *The restriction of ρ to D_p comes from the Galois action on the points of a finite flat group scheme over $\mathbf{Z}_p$.*

PROOF. (a) Let $\ell \neq 2, p$ be a prime. The analysis of the restriction of $\rho = \rho_{a,b,c}$ to D_ℓ can be divided into three cases:

Case 1: The prime ℓ does not divide *abc*. In that case, it is a prime of good reduction for $E_{a,b,c}$, and the action of D_ℓ on $E_{a,b,c}[p]$ is therefore unramified, by the criterion of Néron–Ogg–Shafarevich.

Case 2: The prime ℓ divides *abc*. It is therefore a prime of *multiplicative reduction* for $E_{a,b,c}$. Hence the curve $E_{a,b,c}$, or a twist of it over the unramified quadratic extension of $\mathbf{Q}_\ell$, is isomorphic to the Tate curve $\mathbb{G}_m/q_\ell^{\mathbf{Z}}$ over $\mathbf{Q}_\ell$. More precisely, replacing $E_{a,b,c}$ by its twist if necessary, we have an identification which respects the action of $G_{\mathbf{Q}_\ell}$ on both sides:

$$E(\bar{\mathbf{Q}}_\ell) \simeq \bar{\mathbf{Q}}_\ell^\times / \langle q_\ell \rangle, \tag{17}$$

where $q_\ell \in \mathbf{Q}_\ell^\times$ is the ℓ-adic Tate period, which is obtained by inverting the power series with integer coefficients

$$j = \frac{1}{q} + 744 + 196884q + \cdots$$

that expresses j in terms of q, to obtain a power series

$$q = \mathrm{Tate}(1/j) = 1/j + \cdots \quad \in (1/j)\mathbf{Z}[[1/j]]^\times.$$

In particular, note that, by (16),

$$\mathrm{ord}_\ell(q_\ell) = \mathrm{ord}_\ell(1/j) = \mathrm{ord}_\ell(\Delta) \equiv 0 \pmod{p}. \tag{18}$$

The explicit description of the $G_{\mathbf{Q}_\ell}$-module $E(\bar{\mathbf{Q}}_\ell)$ given by (17) implies that

$$E(\bar{\mathbf{Q}}_\ell)[p] \simeq \{\zeta_p^a q_\ell^{b/p}, \quad 0 \leq a, b \leq p-1\},$$

where ζ_p is a primitive pth root of unity in $\bar{\mathbf{Q}}_\ell^\times$. In the basis $(\zeta_p, q_\ell^{1/p})$ for $E[p]$, the restriction of $\rho = \rho_{a,b,c}$ to D_ℓ can be written as

$$\rho(\sigma) = \begin{pmatrix} \chi_{\mathrm{cyc}}(\sigma)\psi(\sigma) & \kappa(\sigma) \\ 0 & \psi^{-1}(\sigma) \end{pmatrix}, \tag{19}$$

where χ_{cyc} is the p-th cyclotomic character giving the action of D_ℓ on the p-th roots of unity, and ψ is an unramified character of order at most 2 (which is trivial precisely when E has split multiplicative reduction at ℓ.) Furthermore, the cocycle κ is unramified, by (18): this is because the extension $\mathbf{Q}_\ell(\zeta_p, q_\ell^{1/p})$ through which $\rho_{a,b,c}|_{G_{\mathbf{Q}_\ell}}$ factors is unramified. Part (a) of Theorem 4.1 follows.

(b) When $\ell = 2$, the elliptic curve $E_{a,b,c}$ has multiplicative reduction at 2, and hence is identified with a Tate curve over $\mathbf{Q}_2$. The result then follows from (19) with $\ell = 2$.

(c) When $\ell = p$ does not divide abc, the Galois representation $\rho_{a,b,c}$ arises from the p-torsion of an elliptic curve with good reduction at p, and hence from a finite flat group scheme over $\mathbf{Z}_p$. In the case where $p|abc$ (which corresponds to what was known classically as the *second case* of Fermat's Last Theorem) one has a similar conclusion: essentially, the condition $\mathrm{ord}_p(q_p) \equiv 0 \pmod{p}$ limits the ramification of $\rho_{a,b,c}$ at p and implies that $E_{a,b,c}[p]$ extends to a finite flat group scheme over $\mathbf{Z}_p$, in spite of the fact that $E_{a,b,c}$ itself does not have a smooth model over $\mathbf{Z}_p$. □

The following theorem gives a *global* property of the representation $\rho_{a,b,c}$.

THEOREM 4.2 (Mazur). *The Galois representation $\rho_{a,b,c}$ is irreducible.*

PROOF. This follows (at least when p is large enough) from Theorem 3.15. We will now give a self-contained proof which rests on the ideas developed in the proof of Theorem 3.4.

Suppose that $\rho_{a,b,c}$ is reducible. Then $E = E_{a,b,c}$ has a rational subgroup C of order p, and the pair (E, C) gives rise to a rational point x on the modular curve $X_0(p)$. Let $\ell \neq p$ be an odd prime that divides abc. Then E has multiplicative reduction at ℓ. Therefore, the point x reduces to one of the cusps 0 or ∞ of $X_0(p)$ modulo ℓ. It can be assumed without loss of generality that x reduces to ∞, as in Step 2 of the proof of Theorem 3.4. Now recall the natural projection $\Phi_e : J_0(p) \longrightarrow J_e(p)$ of $J_0(p)$ to its winding quotient $J_e(p)$, and the resulting map $j_e : X_0(p) \longrightarrow J_e(p)$. The element $j_e(x)$ belongs to the formal group $J_e^1(p)(\mathbf{Q}_\ell)$, which is torsion-free, and to $J_e(p)(\mathbf{Q})$, which is torsion by Theorem 3.12. Hence $j_e(x) = 0$. We now use the fact that j_e is a formal immersion to deduce that $x = \infty$, as in Step 4 of the proof of Theorem 3.4 (with 3 replaced by ℓ). □

REMARK 4.3. The importance of the Diophantine study of modular curves described in Section 3 in the proof of Fermat's Last Theorem, via Theorem 4.2, cannot be overemphasised. It is sometimes underplayed in expositions of Fermat's Last Theorem, which tend to focus on the ingredients that were supplied later.

Thanks to Theorems 4.1 and 4.2, Fermat's Last Theorem is now reduced to the problem of "classifying" the irreducible two-dimensional mod p representations satisfying the strong restrictions on ramification imposed by Theorem 4.1—or in some sense, to make Theorem 1.1 precise for the class of extensions of $\mathbf{Q}$ arising from such representations. The control we have over questions of this type (which in general seem very hard) arises from the deep and largely conjectural connection that is predicted to exist between Galois representations and *modular forms*.

4.3. Modular forms and Galois representations. Let $f = \sum_n a_n q^n$ be a newform in $S_2(N, \mathbf{C})$. Let K_f denote as before the finite extension of $\mathbf{Q}$ generated by the Fourier coefficients of f, so that f belongs to $S_2(N, K_f)$. The Fourier coefficients of f belong to the ring $\mathcal{O}_f$ of integers of K_f. Let $\mathfrak{p}$ be a prime ideal of $\mathcal{O}_f$ and let $K_{f,\mathfrak{p}}$ denote the completion of K_f at $\mathfrak{p}$.

THEOREM 4.4. *There exists a Galois representation*

$$\rho_{f,\mathfrak{p}} : G_{\mathbf{Q}} \longrightarrow \mathbf{GL}_2(K_{f,\mathfrak{p}})$$

such that

(1) *The representation $\rho_{f,\mathfrak{p}}$ is unramified outside Np.*
(2) *The characteristic polynomial of $\rho_{f,\mathfrak{p}}(\mathrm{Frob}_\ell)$ is equal to $x^2 - a_\ell x + \ell$, for all primes ℓ not dividing p.*
(3) *The representation $\rho_{f,\mathfrak{p}}$ is* odd*, i.e., the image of complex conjugation has eigenvalues* 1 *and* -1*.*

Sketch of proof. Let A_f be the abelian variety quotient of $J_0(N)$ associated to f by the Eichler–Shimura construction (Theorem 3.6). Its endomorphism ring $\mathrm{End}_{\mathbf{Q}}(A_f)$ contains $\mathbf{T}/I_f$, which is an order in K_f. In this way, the Galois representation $V_p(A_f)$ is equipped with an action of $K_f \otimes \mathbf{Q}_p$ which commutes with the action of $G_{\mathbf{Q}}$. Let

$$V_{f,\mathfrak{p}} = V_p(A_f) \otimes_{K_f} K_{f,\mathfrak{p}}.$$

It is a two-dimensional $K_{f,\mathfrak{p}}$-vector space, equipped with a continuous linear action of $G_{\mathbf{Q}}$. The fact that it has the desired properties, particularly property (2), is a consequence of the Eichler–Shimura congruence that was used to prove the equality of L-series given in Theorem 3.10. See Chapter 2 of [**Dar04**] for further details and references. □

4.4. Serre's conjecture. Modular forms can also be used to construct two-dimensional representations of $G_{\mathbf{Q}}$ over finite fields. More precisely, let $\mathcal{O}_{f,\mathfrak{p}}$ be the ring of integers of $K_{f,\mathfrak{p}}$. Since $G_{\mathbf{Q}}$ is compact and acts continuously on $V_{f,\mathfrak{p}}$, it preserves an $\mathcal{O}_{f,\mathfrak{p}}$-stable sublattice $V^0_{f,\mathfrak{p}} \subset V_{f,\mathfrak{p}}$ of rank two over $\mathcal{O}_{\mathfrak{p},p}$. Let $\mathbb{F}_{\mathfrak{p}} := \mathcal{O}_{f,\mathfrak{p}}/\mathfrak{p}$ be the residue field of $\mathcal{O}_f$ at $\mathfrak{p}$. The action of $G_{\mathbf{Q}}$ on the two-dimensional $\mathbb{F}_{\mathfrak{p}}$-vector space $W_{f,\mathfrak{p}} := V^0_{f,\mathfrak{p}}/\mathfrak{p}V^0_{f,\mathfrak{p}}$ gives rise to a two-dimensional mod $\mathfrak{p}$ representation

$$\bar{\rho}_{f,\mathfrak{p}} : G_{\mathbf{Q}} \longrightarrow \mathbf{GL}_2(\mathbb{F}_{\mathfrak{p}}).$$

Like its $\mathfrak{p}$-adic counterpart, this representation is unramified outside of pN and also satisfies parts 2 and 3 of Theorem 4.4.

In [**Ser87**], Serre associated to *any* two-dimensional Galois representation

$$\rho : G_{\mathbf{Q}} \longrightarrow \mathbf{GL}_2(\mathbb{F}) \tag{20}$$

with coefficients in a finite field $\mathbb{F}$ two invariants $N(\rho)$ and $k(\rho)$, called the *Serre conductor* and *Serre weight* of ρ, respectively. The Serre conductor $N(\rho)$ is only divisible by primes distinct from the characteristic of $\mathbb{F}$ at which ρ is ramified. When $\rho = \bar{\rho}_{f,\mathfrak{p}}$ arises from a modular form, the Serre conductor $N(\rho)$ always divides (but is not necessarily equal to) the level N of f. In particular, using parts (a) and (b) of Theorem 4.1, one can show that

$$N(\rho_{a,b,c}) = 2. \tag{21}$$

The recipe for defining $k(\rho)$ is somewhat more involved, but depends only on the restriction of ρ to the decomposition group (in fact, the inertia group) at p. It will suffice, for the purposes of this survey, to note that when ρ arises from the p-division points of a finite flat group scheme over $\mathbf{Z}_p$, then Serre's recipe gives $k(\rho) = 2$. Hence, by part (c) of Theorem 4.1,

$$k(\rho_{a,b,c}) = 2. \tag{22}$$

In [**Ser87**], Serre conjectured that *any* odd irreducible two-dimensional mod p Galois representation ρ as in (20) necessarily arises from an appropriate modular form mod p of weight $k(\rho)$ and level $N(\rho)$. This conjecture has recently been proved

by Khare and Wintenberger (cf. Theorem 1.2 of [**KW**]) in the case where $N(\rho_{a,b,c})$ is odd, and follows in the general case from a similar method, using a result of Kisin [**Kis**].

THEOREM 4.5. *Let ρ be an odd, irreducible two-dimensional mod p representation of $G_{\mathbf{Q}}$. Then there exists an eigenform f of weight $k(\rho)$ on $\Gamma_1(N(\rho))$, and a prime $\mathfrak{p}|p$ of the field K_f such that ρ is isomorphic to $\bar{\rho}_{f,\mathfrak{p}}$ as a representation of $G_{\mathbf{Q}}$.*

Proof of Fermat's Last Theorem. Let (a, b, c) be a primitive nontrivial solution of Fermat's equation $x^p + y^p = z^p$, and consider the Galois representation $\rho = \rho_{a,b,c}$ associated to the p-division points of the associated Frey curve. It follows from Theorem 4.2 that ρ is an odd, irreducible mod p representation of $G_{\mathbf{Q}}$. Its Serre conductor and weight are $N(\rho) = 2$ and $k(\rho) = 2$ by (21) and (22). Therefore Theorem 4.5 implies the existence of a nontrivial cusp form in $S_2(2, \mathbf{C})$. This leads to a contradiction, because there are no such cusp forms: the modular curve $X_0(2)$ has genus zero and hence has no regular differentials. This contradiction implies Fermat's Last Theorem.

4.5. The Shimura–Taniyama conjecture. Historically, the proof of Theorem 4.5 by Khare and Wintenberger came almost 10 years after Wiles proved Fermat's Last theorem. In essence, Wiles proved enough of Theorem 4.5 to cover the Galois representations $\rho_{a,b,c}$ arising from hypothetical solutions of Fermat's equation.

More precisely, the articles [**Wil95**] and [**TW95**] proved the following result, known as the Shimura-Taniyama conjecture for semistable elliptic curves:

THEOREM 4.6. *Let E be a semistable elliptic curve over $\mathbf{Q}$ of conductor N. Then there is a normalised eigenform f in $S_2(N, \mathbf{Z})$ such that $V_p(E)$ is isomorphic to $V_p(A_f)$.*

The proof of this theorem—or even an outline of its main ideas—is beyond the scope of this survey. For details the reader is invited to consult [**DDT94**] for example.

Theorem 4.6 implies that $\rho_{a,b,c}$ arises from a modular form in $S_2(N, \mathbf{C})$, where $N = \prod_{\ell|abc} \ell$. The Serre conjecture (Theorem 4.5) for $\rho_{a,b,c}$ then follows from an earlier theorem of Ribet (which also played an important role in Wiles' original approach to proving Theorem Theorem 4.6.)

THEOREM 4.7. *Suppose that p is odd. Let ρ be an irreducible mod p Galois representation which arises from a modular form (of some weight and level). Then it also arises from an eigenform of weight $k(\rho)$ and level $N(\rho)$.*

Aside from the fact that it proves Fermat's Last Theorem, the importance of Theorem 4.6 can be justified on several other levels.

Firstly, the methods used to prove Theorem 4.6 were subsequently refined in [**BCDT01**] to prove the full Shimura–Taniyama conjecture: all elliptic curves over $\mathbf{Q}$ are modular. This result is of great importance in understanding the arithmetic of elliptic curves over $\mathbf{Q}$, as will be explained in more detail in the next section.

Secondly—and this is a theme that we will not begin to do justice to, because it falls outside the scope of this survey—Wiles' method for proving Theorem 4.6 has led to a general, flexible method for establishing relationships between Galois

representations and modular forms. It was by building on these techniques that Khare and Wintenberger proved Serre's conjecture (Theorem 4.5). Over the years, many other conjectures of this type have been proved building on the proof of Theorem 4.6: for instance, special cases of Artin's conjecture relating representations with finite image to modular forms of weight one (cf. for example [**Tay03**] and the references contained therein), and a proof of the Sato–Tate conjecture for elliptic curves over $\mathbf{Q}$ in [**Tay**], [**HSBT**], and [**CHT**].

Closer to the themes that have been developed in this section, we mention a natural generalisation of Theorem 4.6 concerning abelian varieties *of* $\mathbf{GL}_2$*-type*. An abelian variety A over $\mathbf{Q}$ is said to be of $\mathbf{GL}_2$-type if $\mathrm{End}_{\mathbf{Q}}(A) \otimes \mathbf{Q}$ contains a field K with $[K : \mathbf{Q}] = \dim(A)$. The reason for this terminology is that such an A gives rise, for each prime ideal $\mathfrak{p}$ of K, to a two-dimensional Galois representation

$$\rho_{A,\mathfrak{p}} : G_{\mathbf{Q}} \longrightarrow \mathbf{GL}_2(K_{\mathfrak{p}})$$

arising from the action of $G_{\mathbf{Q}}$ on $V_p(A) \otimes_K K_{\mathfrak{p}}$. The abelian varieties A_f arising from the Eichler-Shimura construction are examples of abelian varieties of $\mathbf{GL}_2$-type. A conjecture of Fontaine and Mazur predicts that all abelian varieties of $\mathbf{GL}_2$-type arise as quotients of Jacobians of modular curves. It can be shown that this generalisation of the Shimura-Taniyama conjecture follows from Theorem 4.5. (Cf. for example [**Ser87**] or the introduction of [**Kis**].)

4.6. A summary of Wiles' proof. There are some enlightening parallels to be drawn between the proof of Fermat's Last Theorem and Faltings' proof of the Mordell conjecture as summarised in Section 2.8. Like Faltings' proof, the proof of Fermat's Last theorem is based on a sequence of maps, resulting in a sequence of transformations leading from the original problem to questions about other types of structures, such as Galois representations, and ultimately modular forms. These reductions are summarised in the diagram below.

$$\left\{\begin{array}{l}\text{Integer solutions}\\ (a,b,c)\text{ of}\\ x^p+y^p=z^p\end{array}\right\} \xrightarrow{R_1} \left\{\begin{array}{l}\text{Semistable elliptic curves}\\ \text{of conductor } N=abc\\ \text{and discriminant } 2^{-8}(abc)^{2p}\end{array}\right\}$$
$$\xrightarrow{R_4} \left\{\begin{array}{l}\text{Irreducible galois representations}\\ \rho : G_{\mathbf{Q}} \longrightarrow \mathbf{GL}_2(\mathbb{F}_p)\\ \text{with } N(\rho)=2 \text{ and } k(\rho)=2.\end{array}\right\}$$
$$\xrightarrow{R_5} \left\{\begin{array}{l}\text{Cusp forms in}\\ S_2(2,\mathbf{Z}/p\mathbf{Z}).\end{array}\right\}.$$

(1) The map R_1 is defined via the Frey curve, and is reminiscent of the Kodaira–Parshin construction of Section 2.2. An important difference is that the set of primes of bad reduction of the Frey curve associated to (a, b, c) is *not* bounded independently of (a, b, c). In fact, the set of primes of bad reduction for E consists *exactly* of the primes that divide abc.

(2) The map R_4 plays a role analogous to the passage to the ℓ-adic representations in Faltings' proof. An important difference here is that we consider mod p representations (with coefficients in a finite field) rather than p-adic representations. The justification for doing this is given by Theorem 4.1, which shows that the mod p representation ρ attached to $E_{a,b,c}$ has bounded ramification. Note that the corresponding p-adic representation would be ramified precisely at the primes dividing $pabc$. It is an exercise

to show that the map R_4 is finite-to-one when $p \geq 7$. (Hint: use Faltings' Theorem 2.1, and the fact that $X(p)$ has genus > 1 when $p \geq 7$.) It is even believed that R_4 is injective once p is large enough (cf. Conjecture 3.18), but this assertion is still unproved.

(3) The map R_5 is a new ingredient that has no counterpart in Faltings' proof of Mordell's conjecture, and exploits the deep "dictionary" that is expected to exist between Galois representations and modular forms—in this case, the Serre conjecture proved by Khare and Wintenberger.

(4) The final step in the argument exploits the fact that there are no modular forms of weight two and level two. This last point may seem like a "lucky accident" in the proof of Fermat's Last Theorem. Indeed the presence of modular forms of higher level presents an obstruction for the method based on Frey curves to yield results on more general ternary Diophantine equations of Fermat type. However, see the article by Charollois in this volume [**Chab**], where a refinement of the techniques described in this section leads to a strikingly general result on the generalised Fermat equation $ax^p + by^p + cz^p = 0$.

REMARK 4.8. One of the consequences of Conjecture 3.18 is that the generalised Fermat equation $ax^n + by^n + cz^n = 0$ (with a, b, c fixed) has no primitive integer solutions (x, y, z) with $xyz \neq 0, \pm 1$, once n is large enough. (The reader who masters the ideas in the article by Pierre Charollois in this proceedings volume will be able to prove this assertion.)

5. Elliptic curves

After surveying curves of genus > 1, we turn our attention to curves of genus 1. A projective curve of genus 1 over a field K, equipped with a distinguished K-rational point over that field, is endowed with a natural structure of a commutative algebraic group over K for which the distinguished element becomes the identity. Such a curve is called an *elliptic curve.*

If E is an elliptic curve defined over a number field K, then the Mordell–Weil Theorem (cf. Theorem 7 of the introduction) asserts that the group $E(K)$ of K-rational points on E is *finitely generated.* Let $r(E, K)$ denote the rank of this finitely generated abelian group. Many of the important questions in the theory of elliptic curves revolve around calculating this invariant, and understanding its behaviour as E or K vary.

QUESTION 5.1. *Is there an effective algorithm to calculate $r(E, K)$, given E and K?*

Showing that Fermat's method of descent yields such an effective algorithm is intimately connected to the Shafarevich–Tate conjecture asserting the finiteness of the Shafarevich–Tate group $Ш(E/K)$ of E/K.

One can also fix a base field (the most natural, and interesting, case being the case where $K = \mathbf{Q}$) and ask

QUESTION 5.2. *Is the rank $r(E, K)$ unbounded, as E ranges over all elliptic curves defined over K?*

One can also fix an elliptic curve E and enquire about the variation of $r(E, K)$ as K ranges over different number fields.

The main tool available at present to study $r(E, K)$ is the relationship between the rank and the Hasse–Weil L-series predicted by the Birch and Swinnerton-Dyer conjecture (Conjecture 3.9).

Assume that E is an elliptic curve over $\mathbf{Q}$. Thanks to Theorem 4.6 (and its extension to all elliptic curves over $\mathbf{Q}$ given in [**BCDT01**]), the Hasse–Weil L-series $L(E, s)$ is equal to $L(f, s)$ for some newform f of weight two. In particular, $L(E, s)$ has an analytic continuation to the entire complex plane, and a functional equation.

The main result we will discuss in this section is the following:

THEOREM 5.3. *Let E be an elliptic curve over $\mathbf{Q}$, and let $L(E, s)$ be its Hasse–Weil L-series. If $r := \mathrm{ord}_{s=1} L(E, s) \leq 1$, then $r(E, \mathbf{Q}) = r$ and $Ш(E/\mathbf{Q})$ is finite.*

5.1. Modular parametrisations. Let E be an elliptic curve over $\mathbf{Q}$ of conductor N. Recall the modular curve $X_0(N)$ that was introduced in Section 3.1. The following theorem, which produces a dominant rational map from such a curve to E, plays a crucial role in the proof of Theorem 5.3.

THEOREM 5.4. *There exists a nonconstant map of curves over $\mathbf{Q}$*

$$\varphi : X_0(N) \longrightarrow E.$$

PROOF. By Theorem 4.6, there is a normalised eigenform f in $S_2(N, \mathbf{Z})$ satisfying $L(E, s) = L(f, s)$. Let A_f be the quotient of $J_0(N)$ associated to f via the Eichler–Shimura construction. By assumption, the Galois representations $V_p(E)$ and $V_p(A_f)$ are isomorphic. Hence the isogeny conjecture (Theorem 2.18) implies the existence of an isogeny $\alpha : A_f \longrightarrow E$ defined over $\mathbf{Q}$. Composing such an isogeny with the natural surjective morphism $J_0(N) \longrightarrow A_f$ gives a nonconstant map $\Phi : J_0(N) \longrightarrow E$. The modular parametrisation φ is defined by setting $\varphi(x) := \Phi((x) - (\infty))$. □

It is useful to describe briefly how the modular parametrisation φ can be computed analytically. The pullback $\varphi^*(\omega_E)$ is a nonzero rational multiple of the differential form

$$\omega_f := 2\pi i f(\tau) d\tau = \sum_{n=1}^{\infty} a_n q^n \frac{dq}{q}.$$

Denote by Λ_f the collection of periods of ω_f (integrals of ω_f against smooth closed one-chains C in $X_0(N)(\mathbf{C})$):

$$\Lambda_f := \left\{ \int_C \omega_f, \text{ where } \partial C = 0 \right\}.$$

It is a lattice in $\mathbf{C}$, and $A_f(\mathbf{C}) \simeq \mathbf{C}/\Lambda_f$. Let us replace E by A_f, so that $\alpha = 1$. It is suggestive (for later generalisations) to view φ as a map

$$\varphi : \mathrm{Div}^0(X_0(N)) \longrightarrow E.$$

This map is defined on $\mathrm{Div}^0(X_0(N)(\mathbf{C}))$ by the rule

$$\varphi(\Delta) := \int_C \omega_f \pmod{\Lambda_f}, \tag{23}$$

where the integral is taken over any smooth one-chain C whose boundary is Δ. The invariant $\varphi(\Delta) \in \mathbf{C}/\Lambda_f$ is viewed as a point on $E(\mathbf{C})$ via the Weierstrass uniformisation.

5.2. Heegner points. Perhaps the most important arithmetic application of the modular parametrisation arises from the fact that $X_0(N)$ is endowed with a systematic supply of algebraic points defined over abelian extensions of imaginary quadratic fields—the so-called CM-points. These points correspond, in the moduli interpretation of $X_0(N)$, to pairs (A, C) where A is an elliptic curve whose endomorphism ring $\mathcal{O} = \mathrm{End}(A)$ is an order in a quadratic imaginary field K. Such an elliptic curve is said to have *complex multiplication* by K, and the corresponding points on $X_0(N)$ are called *CM-points* attached to K. Let $\mathrm{CM}(K)$ denote the set of all CM points in $X_0(N)$ attached to K. It satisfies the following properties.

(1) The set $\mathrm{CM}(K)$ is dense in $X_0(N)(\mathbf{C})$ (relative to the Zariski topology, and also the complex topology).
(2) Let K^{ab} denote the maximal abelian extension of K. Then $\mathrm{CM}(K)$ is contained in $X_0(N)(K^{\mathrm{ab}})$.
(3) Analytically, $\mathrm{CM}(K) = \Gamma_0(N)\backslash(\mathcal{H} \cap K)$.

Definition 5.5. The collection of points

$$\mathrm{HP}(K) := \{\varphi(\Delta)\}_{\Delta \in \mathrm{Div}^0(\mathrm{CM}(K))} \subset E(K^{\mathrm{ab}})$$

is called the system of *Heegner points* on E attached to K.

The usefulness of Heegner points arises from two facts:

(1) They can be related to L-series, thanks to the theorem of Gross–Zagier and its generalisations.
(2) They can be used to bound Mordell–Weil groups and Shafarevich–Tate groups of elliptic curves, following a descent method that was discovered by Kolyvagin.

Heegner points and L-series.

For simplicity, suppose that the imaginary quadratic field K satisfies the following so-called *Heegner hypothesis*:

Hypothesis 5.6. *There exists a ideal $\mathcal{N}$ of norm N in $\mathcal{O}_K$ with cyclic quotient.*

This hypothesis is used to construct a distinguished element in $\mathrm{HP}(K)$. More precisely, let h denote the class number of K, and let H be its Hilbert class field. By the theory of complex multiplication, there are precisely h distinct (up to isomorphism over $\mathbf{C}$) elliptic curves $A_1, \ldots, A_h$ having endomorphism ring equal to $\mathcal{O}_K$. The j-invariants of these curves belong to H, and are permuted simply transitively by the action of $\mathrm{Gal}(H/K)$. It is therefore possible to choose $A_1, \ldots, A_h$ in such a way that they are defined over H, and permuted by the action of $\mathrm{Gal}(H/K)$.

The pairs $(A_i, A_i[\mathcal{N}])$ (with $1 \leq i \leq h$) correspond to points P_i in $X_0(N)(H)$. Let

$$P_K := \varphi((P_1) + \cdots + (P_h) - h(\infty)) \in E(K). \tag{24}$$

The fact that the point P_K has an explicit moduli description makes it possible to establish some of its key properties. For example, let $\bar{P}_K$ denote the image of P_K under complex conjugation. Then it can be shown that

$$\bar{P}_K = wP_K \pmod{E(K)_{\mathrm{tors}}}, \tag{25}$$

where $w \in \{\pm 1\}$ is the *negative* of the sign in the functional equation for $L(E,s) = L(f,s)$. (Cf. Chapter 3 of [**Dar04**].) This provides a simple connection between the behaviour of P_K and the L-series $L(E,s)$.

We note that, in many cases where Hypothesis 5.6 is satisfied (for example, when all the primes dividing N are *split* in the quadratic imaginary field K), the sign in the functional equation for the Hasse–Weil L-series $L(E/K,s)$ is -1, so that $L(E/K,1) = 0$. It then becomes natural to consider the first derivative $L'(E/K,1)$ at the central critical point. The following theorem of Gross and Zagier establishes an explicit link between P_K and this quantity.

THEOREM 5.7. *Let $\langle f, f\rangle$ denote the Petersson scalar product of f with itself, and let $h(P_K)$ denote the Néron–Tate canonical height of P_K on $E(K)$. There is an explicit nonzero rational number t such that*

$$L'(E/K,1) = t \cdot \langle f, f\rangle \cdot h(P_K). \tag{26}$$

In particular, the point P_K is of infinite order if and only if $L'(E/K,1) \neq 0$.

The proof of Theorem 5.7 given in [**GZ86**] proceeds by a direct calculation in which both sides of (26) are computed explicitly, compared, and found to be equal.

REMARK 5.8. Let P_n be a point in $\mathrm{CM}(K)$ corresponding to an elliptic curve with endomorphism ring equal to the order $\mathcal{O}_n$ of conductor n in K. Such a point can be defined over the ring class field H_n of K of conductor n, whose Galois group $G_n := \mathrm{Gal}(H_n/K)$ is canonically identified with the class group $\mathrm{Pic}(\mathcal{O}_n)$ by class field theory. If $\chi : G_n \longrightarrow \mathbf{C}^\times$ is a complex character, one can generalise (24) to define

$$P_\chi := \varphi\left(\sum_{\sigma \in G_n} \chi(\sigma) P_n^\sigma\right) \in E(H_n) \otimes \mathbf{C}. \tag{27}$$

A generalisation of Theorem 5.7 due to S. Zhang (cf. for example [**Zha01b**], [**Zha01a**], [**How**] and [**How07**]) relates the height of P_χ to the derivative of the twisted L-series $L(E/K,\chi,s)$ at $s = 1$.

When $L'(E/K,1) \neq 0$, the method of Heegner points gives an efficient method for producing a point of infinite order in $E(K)$. The following proposition asserts the existence of many K for which the L-series does not vanish.

PROPOSITION 5.9. *Suppose that $r := \mathrm{ord}_{s=1} L(E,s) \leq 1$. Then there exist infinitely many imaginary quadratic fields K satisfying Hypothesis 5.6 for which*

$$\mathrm{ord}_{s=1}(L(E/K,s)) = 1.$$

The proof of this proposition is explained in [**MM97**].

Heegner points and arithmetic: Kolyvagin's descent

Theorem 5.7 implies that if $L'(E/K,1) \neq 0$, then P_K is of infinite order and hence $r(E,K) \geq 1$. The following theorem of Kolyvagin gives a bound in the other direction as well.

THEOREM 5.10 (Kolyvagin). *Suppose that P_K is of infinite order in $E(K)$. Then $r(E,K) = 1$, and $Ш(E/K)$ is finite.*

For a proof of this theorem, see [**Gro91**] or Chapter 10 of [**Dar04**]. Let us just mention here that Kolyvagin's proof makes essential use of the fact that the point P_K does not come alone, but rather is part of a norm-compatible system of points in $E(K^{\mathrm{ab}})$ arising from the (infinite) collection of points in $\mathrm{HP}(K)$. These points are used to construct global cohomology classes in $H^1(K, E[p])$ whose local behaviour can be controlled precisely and related to P_K. Under the assumption that P_K is of infinite order, this system of ramified cohomology classes is enough to bound the p-Selmer group of E/K and show that $r(E, K) = 1$.

5.3. Proof of Theorem 5.3. We will now explain how the properties of P_K and $\mathrm{HP}(K)$ described in the previous section can be combined to prove Theorem 5.3:

Proof of Theorem 5.3. Assume that $r \leq 1$. By Proposition 5.9, there is an imaginary quadratic field K satisfying Hypothesis 5.6, for which

$$\mathrm{ord}_{s=1}(L(E/K, s)) = 1.$$

Fix such a K, and consider the point P_K. Since $L'(E/K, 1) \neq 0$, Theorem 5.7 implies that P_K is of infinite order. Theorem 5.10 then shows that

$$r(E, K) = 1, \quad \text{and } Ш(E/K) \text{ is finite.}$$

Let E' denote the quadratic twist of E over K. We then have

$$1 = r(E, K) = r(E, \mathbf{Q}) + r(E', \mathbf{Q}).$$

To be able to ignore finer phenomena associated to torsion in $E(K)$, it is convenient to replace P_K by its image in $E(K) \otimes \mathbf{Q}$. Since $E(K) \otimes \mathbf{Q}$ is generated by P_K, it follows that

$$r(E, \mathbf{Q}) = \begin{cases} 0 & \text{if } \bar{P}_K = -P_K, \\ 1 & \text{if } \bar{P}_K = P_K. \end{cases}$$

Theorem 5.3 now follows from (25). □

Remark 5.11. The proof of Theorem 5.3 carries over with only minor changes when E is replaced by the abelian variety quotient A_f attached to an arbitrary eigenform f of weight 2 on $\Gamma_0(N)$. This is how Theorem 3.11 is proved:

$$L(A_f, 1) \neq 0 \implies A_f(\mathbf{Q}) \text{ is finite.}$$

The reader will recall the key role played by this theorem in the proof of Theorem 3.1 and (even more importantly) in Merel's proof of the uniform boundedness conjecture for elliptic curves explained in Marusia Rebolledo's article in these proceedings.

5.4. Modularity of elliptic curves over totally real fields. Because of the crucial role played by the system $\mathrm{HP}(K)$ in the proof of Theorem 5.3, it is natural to ask whether such structures are present in more general settings. For example, we would like to prove analogues of Theorem 5.3 for elliptic curves defined over number fields other than $\mathbf{Q}$. The class of number fields for which this program is best understood is the class of *totally real fields.*

More precisely, let F be a totally real field of degree n, and let E be an elliptic curve over F, of conductor N. Assume, for simplicity, that F has narrow class number one, so that in particular the conductor can now be viewed as a totally positive element of $\mathcal{O}_F$ rather than just an ideal.

The group $\Gamma_0(N;\mathcal{O}_F) \subset \mathbf{SL}_2(\mathcal{O}_F)$ is defined as the group of matrices that are upper triangular modulo N. The n distinct real embeddings $v_1, \dots, v_n : F \longrightarrow \mathbf{R}$ of F allow us to view $\Gamma_0(N;\mathcal{O}_F)$ as a subgroup of $\mathbf{SL}_2(\mathbf{R})^n$. This subgroup acts discretely on the product $\mathcal{H}^n$, and the analytic quotient $\Gamma_0(N;\mathcal{O}_F)\backslash\mathcal{H}^n$ represents the natural generalisation of modular curves to this setting:

(1) This quotient is identified with the complex points of an n-dimensional algebraic variety $Y_0(N;\mathcal{O}_F)$ defined over F. This variety can be compactified by adjoining a finite set of cusps, much as in the setting $n = 1$ of classical modular curves. A suitable desingularisation of the resulting projective variety is denoted $X_0(N;\mathcal{O}_F)$, and is called a *Hilbert modular variety*. Hilbert modular varieties are basic examples of higher dimensional Shimura varieties.
(2) The variety $X_0(N;\mathcal{O}_F)$ is equipped with natural Hecke correspondences T_λ indexed by the prime ideals of $\mathcal{O}_F$.
(3) These correspondences induce linear actions on the n-th deRham cohomology $H^n_{dR}(X_0(N;\mathcal{O}_F))$, and the eigenvalues of the Hecke operators are expected to encode the same type of arithmetic information as in the case where $F = \mathbf{Q}$.

To amplify this last point and make it more precise, we state the following generalisation of the Shimura–Taniyama conjecture to elliptic curves over F:

Conjecture 5.12. *Let E be an elliptic curve over F of conductor N. There exists a closed (in fact, holomorphic) differential form $\omega \in H^n_{\mathrm{dR}}(X_0(N;\mathcal{O}_F))$ satisfying*

$$T_\lambda(\omega) = a_\lambda(E)\omega,$$

for all primes $\lambda \nmid N$ of $\mathcal{O}_F$.

Remark 5.13. In some cases, the methods of Wiles for proving the modularity of elliptic curves over $\mathbf{Q}$ have been extended to the setting of elliptic curves over totally real fields, and many cases of Conjecture 5.12 can be made unconditional.

5.5. Shimura curves. When $n > 1$, the holomorphic differential form ω whose existence is predicted by Conjecture 5.12 cannot be used to directly produce an analogue of the modular parametrisation. In this sense, there is no immediate generalisation of Theorem 5.4, which plays such a crucial role in the construction of $\mathrm{HP}(K)$ when $n = 1$.

To extend the notion of Heegner points, it is necessary to introduce another generalisation of modular curves: the so-called *Shimura curves* associated to certain quaternion algebras over F.

A quaternion algebra B over F is said to be *almost totally definite* if

$$B \otimes_{v_1} \mathbf{R} \simeq M_2(\mathbf{R}), \qquad B \otimes_{v_j} \mathbf{R} \simeq \mathbb{H}, \text{ for } 2 \leq j \leq n.$$

We can associate to any order R in B a discrete subgroup

$$\Gamma := v_1(R^\times) \subset \mathbf{SL}_2(\mathbf{R}),$$

which acts discretely on $\mathcal{H}$ by Möbius transformations. When $F = \mathbf{Q}$ and $B = M_2(\mathbf{Q})$ is the split quaternion algebra, one recovers the analytic description of the modular curves $X_0(N)$. Otherwise, the analytic quotient $\Gamma \subset \mathcal{H}$ is a *compact* Riemann surface which can be identified with the complex points of an algebraic

curve X possessing a canonical model over F. The curve X can be related (following a construction of Shimura) to the solution of a moduli problem and is also equipped with a supply $\mathrm{CM}(K) \subset X(K^{\mathrm{ab}})$ of CM points, associated this time to any quadratic totally imaginary extension K of F.

An elliptic curve E over F is said to be *arithmetically uniformisable* if there is a nonconstant map defined over F, generalising Theorem 5.4,

$$\varphi : \mathrm{Div}^0(X) \longrightarrow E.$$

The theory of Jacquet–Langlands gives a precise (partly conjectural) understanding of the class of elliptic curves that should be arithmetically uniformisable:

THEOREM 5.14. *Let E be an elliptic curve over F which is not isogenous to any of its Galois conjugates. Then E is arithmetically uniformisable if*

(1) *E is modular in the sense of Conjecture 5.12;*
(2) *E has potentially semistable reduction at a prime of F, or can be defined over a field F of odd degree.*

The collection $\mathrm{HP}(K) := \varphi(\mathrm{Div}^0(CM(K))) \subset E(K^{\mathrm{ab}})$, for suitable totally complex quadratic extensions K/F, can be used to obtain results analogous to Theorem 5.3 for elliptic curves over totally real fields. See [**Zha01b**] where general results in this direction are obtained.

The articles [**Voi**] and [**Greb**] in this volume describe Shimura curves and the associated parametrisations in more detail, from a computational angle. The article [**Voi**] discusses explicit equations for Shimura curves of low degree, and [**Greb**] explains how to approach the numerical calculation of the systems $\mathrm{HP}(K)$ of Heegner points via p-adic integration of the associated modular forms, exploiting the theory of p-adic uniformisation of these curves due to Cherednik and Drinfeld.

5.6. Stark–Heegner points. Heegner points arising from Shimura curve parametrisations do not completely dispel the mystery surrounding the Birch and Swinnerton-Dyer conjecture for (modular) elliptic curves over totally real fields, since (even assuming the modularity Conjecture 5.12) there remain elliptic curves over F that are *not* arithmetically uniformisable.

The simplest example of such an elliptic curve is one that has everywhere good reduction over a totally real field F of even degree, and is not isogenous to any of its Galois conjugates. (More generally, one can also consider any quadratic twist of such a curve.) For these elliptic curves, there is at present very little evidence for the Birch and Swinnerton-Dyer conjecture, and in particular the analogue of Theorem 5.3 is still unproved when $\mathrm{ord}_{s=1} L(E/F, s) = 1$. (In the case where $L(E/F, 1) \neq 0$, see the work of Matteo Longo [**Lon06**].)

The notion of Stark–Heegner points represents an attempt to remedy this situation (albeit conjecturally) by exploiting the holomorphic differential n-form ω whose existence is predicted by Conjecture 5.12 rather than resorting to a Shimura curve parametrisation. We note that the holomorphic form ω can be written

$$\omega = f(\tau_1, \ldots, \tau_n) d\tau_1 \cdots d\tau_n,$$

where f is a (holomorphic) Hilbert modular form of parallel weight 2 on $\Gamma_0(N)$, satisfying, for all matrices $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(N)$,

$$f\left(\frac{a_1\tau_1 + b_1}{c_1\tau_1 + d_1}, \ldots, \frac{a_n\tau_n + b_n}{c_n\tau_n + d_n}\right) = (c_1\tau_1 + d_1)^2 \cdots (c_n\tau_n + d_n)^2 f(\tau_1, \ldots, \tau_n).$$

We let any unit $\epsilon \in \mathcal{O}_F^\times$ act on $\mathcal{H}^n$ by the rule:

$$\epsilon \star \tau_j = \begin{cases} \epsilon_j \tau_j & \text{if } \epsilon_j > 0; \\ \epsilon_j \bar{\tau}_j & \text{if } \epsilon_j < 0. \end{cases}$$

For any subset $S \subset \{2, \ldots, n\}$ of cardinality m, we can then define a closed differential n-form of type $(n-m, m)$ by choosing a unit ϵ of $\mathcal{O}_F^\times$ which is negative at the places of S, and positive at the other embeddings, and setting

$$\omega_S = f(\epsilon \star \tau_1, \ldots, \epsilon \star \tau_n) d(\epsilon \star \tau_1) \ldots d(\epsilon \star \tau_n).$$

Finally we set

$$\omega_E := \sum_{S \subset \{2,\ldots,n\}} \omega_S.$$

The following conjecture is due to Oda [**Oda82**].

Conjecture 5.15. *The set of periods*

$$\Lambda_f := \left\{ \int_C \omega_E \quad \text{for } C \in H_n(X_0(N, F)(\mathbf{C}), \mathbf{Z}) \right\} \subset \mathbf{C}$$

is a lattice which is commensurable with the period lattice of $E_1 := v_1(E)$.

Conjecture 5.15 can be used to define a generalisation of the modular parametrisation of equation (23). This map is defined on homologically trivial $(n-1)$-cycles on $X_0(N; \mathcal{O}_F)(\mathbf{C})$ by the rule

$$\varphi(\Delta) := \int_C \omega_E \pmod{\Lambda_f}, \quad \text{where } \partial C = \Delta. \tag{28}$$

The interest of this generalisation of (23) is that it is possible to define a collection of distinguished topological $(n-1)$-cycles on which φ is conjectured to take algebraic values.

These cycles, which play the same role that Heegner divisors of degree zero played in the case where $n = 1$, are defined in terms of certain quadratic extensions K of F. Such a quadratic extension is said to be *almost totally real* if

$$K \otimes_{v_1} \mathbf{R} \simeq \mathbf{C}, \qquad K \otimes_{v_j} \mathbf{R} \simeq \mathbf{R} \oplus \mathbf{R} \quad \text{for } 2 \leq j \leq n.$$

Let $\iota : K \longrightarrow M_2(F)$ be an F-algebra embedding, and let $K_1^\times$ be the group of elements whose norm to F is equal to 1. The torus $v_1(\iota(K_1^\times))$ acts on $\mathcal{H}$ with a unique fixed point τ_1, and $\iota(K_1^\times)$ acts on the region $\{\tau_1\} \times \mathcal{H}^{n-1}$ without fixed points. The orbit of any point in this region under the action of $\iota((K \otimes_F \mathbf{R})_1^\times)$ is a real $(n-1)$-dimensional manifold $Z_\iota \subset \{\tau_1\} \times \mathcal{H}^{n-1}$ which is homeomorphic to $\mathbf{R}^{n-1}$. The group $G_\iota := \iota(K^\times) \cap \Gamma_0(N, \mathcal{O}_F)$ is an abelian group of rank $n-1$, corresponding to a finite index subgroup of the group of relative units in K/F. Consider a fundamental region for the action of G_ι on Z_ι. The image Δ_ι of such a region in the Hilbert modular variety $X_0(N; \mathcal{O}_F)$ is a closed $(n-1)$-cycle, which is topologically isomorphic to a real $(n-1)$-dimensional torus.

Conjecture 5.16. *Assume that* Δ_ι *is homologically trivial. Then the point* $\varphi(\Delta_\iota) \in E_1(\mathbf{C})$ *is an algebraic point, and is in fact the image of a point in* $E(K^{\mathrm{ab}})$ *under any embedding* $K^{\mathrm{ab}} \longrightarrow \mathbf{C}$ *extending* $v_1 : F \longrightarrow \mathbf{R}$.

REMARK 5.17. The original formulation of Conjecture 5.16 given in [**DL03**] was phrased in terms of group cohomology. The definition of $\varphi(\Delta_\iota)$ used in Conjecture 5.16, which suggests an analogy between φ and higher Abel-Jacobi maps, was formulated only later, in [**CD08**] (in a context where cusp forms are replaced by Eisenstein series; the elements $\varphi(\Delta_\iota)$ can then be related to Stark units). The equivalence between Conjecture 5.16 and the main conjecture of [**DL03**] is explained in [**CD08**].

Conjecture 5.16 can be formulated more precisely, in a way that makes a prediction about the fields of definition of the points $\varphi(\Delta_\iota)$. It is expected that the system of points

$$\mathrm{HP}(K) := \{\varphi(\Delta_\iota)\}_{\iota: K \longrightarrow M_2(F)},$$

as ι ranges over all possible embeddings, gives rise to an infinite collection of algebraic points in $E(K^{\mathrm{ab}})$ with properties similar to those of the system of Heegner points defined in Section 5.2. Such a system of points (if its existence, and basic properties, could be established, a tall order at present!) would lead to a proof of Theorem 5.3 for all (modular) elliptic curves defined over totally real fields, not just those that are arithmetically uniformisable.

For more details on Conjecture 5.16, a more precise formulation, and numerical evidence, see Chapter 8 of [**Dar04**], or [**DL03**]. For an explanation of the relation between Conjecture 5.16 and the conjectures of [**DL03**], see [**CD08**].

The Stark-Heegner points attached to Hilbert modular forms that were defined and studied in [**DL03**] and [**CD08**] can be viewed as the *basic prototype* for the general notion of Stark-Heegner points. Here are some further variants that have been explored so far in the literature:

(1) If E is an elliptic curve over $\mathbf{Q}$ of conductor $N = pM$ with $p \nmid M$, a p-adic analogue of the map φ of equation (28)—described in terms of group cohomology rather than singular cohomology, following the same approach and in [**DL03**]—is defined in [**Dar01**], by viewing E as uniformised by the "mock Hilbert surface"

$$\Gamma_0(M; \mathbf{Z}[1/p]) \backslash (\mathcal{H}_p \times \mathcal{H}),$$

where $\mathcal{H}_p := \mathbf{C}_p - \mathbf{Q}_p$ is the p-adic upper half plane, and $\Gamma_0(M; \mathbf{Z}[1/p])$ is the group of matrices in $\mathbf{SL}_2(\mathbf{Z}[1/p])$ which are upper-triangular modulo M. The resulting map φ associates a point in $P_\iota \in E(\bar{\mathbf{Q}}_p)$ to any embedding $\iota : K \longrightarrow M_2(\mathbf{Q})$ when K is a *real quadratic* field in which p is inert. The system $\{P_\iota\} \subset E(\bar{\mathbf{Q}}_p)$, as ι ranges over all embeddings of K into $M_2(\mathbf{Q})$, is expected to yield a system of points in $E(K^{\mathrm{ab}})$ with the same properties as the Heegner points attached to an imaginary quadratic base field. This construction is not expected to yield new cases of the Birch and Swinnerton-Dyer over the base field $\mathbf{Q}$—this conjecture is completely known when $\mathrm{ord}_{s=1} L(E, s) \leq 1$, thanks to Theorem 5.3. However, it would give new cases of this conjecture over certain abelian extensions of real quadratic fields, and, more importantly perhaps, it suggests an explicit analytic construction of class fields of real quadratic fields. For more details on Stark-Heegner points attached to real quadratic fields, see [**Dar01**] or Chapter 9 of [**Dar04**]. The article [**DP06**] describes efficient

algorithms for calculating the points $\varphi(\Delta_\iota)$, and uses them to gather numerical evidence for the conjectures of [**Dar01**], while [**BD**] provides some theoretical evidence.

(2) The article [**Tri06**] formulates and tests numerically a Stark–Heegner construction that leads to conjectural systems of algebraic points on elliptic curves defined over a quadratic imaginary base field. The details of the construction of [**Tri06**] are explained in the article [**Grea**] by Matt Greenberg in this proceedings volume. We remark that there is not a single example of an elliptic curve E genuinely defined over such a field (i.e., which is not isogenous to its Galois conjugate) for which Theorem 5.3 (or even just the Shafarevich–Tate conjecture) has been proved.

References

[BCDT01] C. Breuil, B. Conrad, F. Diamond, and R. Taylor, *On the modularity of elliptic curves over* $\mathbf{Q}$*: wild 3-adic exercises*, J. Amer. Math. Soc. **14** (2001), no. 4, 843–939 (electronic). MR 1839918 (2002d:11058)

[BD] M. Bertolini and H. Darmon, *The rationality of Stark-Heegner points over genus fields of real quadratic fields*, Ann. of Math. (2), to appear.

[CD08] P. Charollois and H. Darmon, *Arguments des unités de Stark et périodes de séries d'Eisenstein*, Algebra Number Theory **2** (2008), no. 6, 655–688. MR 2448667

[Chaa] H. Chapdelaine, *Non-abelian descent and the generalized Fermat equation*, in this volume.

[Chab] P. Charollois, *Generalized Fermat equations (d'après Halberstadt-Kraus)*, in this volume.

[CHT] L. Clozel, M. Harris, and R. Taylor, *Automorphy for some ℓ-adic lifts of automorphic mod ℓ representations*, to appear.

[Cre] J. Cremona, Private communication.

[CS86] G. Cornell and J. H. Silverman (eds.), *Arithmetic geometry*, Springer-Verlag, New York, 1986, Papers from the conference held at the University of Connecticut, Storrs, Connecticut, July 30–August 10, 1984. MR 861969 (89b:14029)

[Dar01] H. Darmon, *Integration on $\mathcal{H}_p \times \mathcal{H}$ and arithmetic applications*, Ann. of Math. (2) **154** (2001), no. 3, 589–639. MR 1884617 (2003j:11067)

[Dar04] ———, *Rational points on modular elliptic curves*, CBMS Regional Conference Series in Mathematics, vol. 101, Published for the Conference Board of the Mathematical Sciences, Washington, DC, 2004. MR 2020572 (2004k:11103)

[DDT94] H. Darmon, F. Diamond, and R. Taylor, *Fermat's last theorem*, Current developments in mathematics, 1995 (Cambridge, MA), Int. Press, Cambridge, MA, 1994, Reprinted in Elliptic curves, modular forms & Fermat's last theorem (Hong Kong, 1993), 2–140, International Press, Cambridge, MA, 1997, pp. 1–157. MR 1474977 (99d:11067a)

[Del85] P. Deligne, *Preuve des conjectures de Tate et de Shafarevitch (d'après G. Faltings)*, Astérisque (1985), no. 121-122, 25–41, Seminar Bourbaki, Vol. 1983/84. MR 768952 (87c:11026)

[DL03] H. Darmon and A. Logan, *Periods of Hilbert modular forms and rational points on elliptic curves*, Int. Math. Res. Not. (2003), no. 40, 2153–2180. MR 1997296 (2005f:11110)

[DP06] H. Darmon and R. Pollack, *Efficient calculation of Stark-Heegner points via overconvergent modular symbols*, Israel J. Math. **153** (2006), 319–354. MR 2254648 (2007k:11077)

[DV] S. Dasgupta and J. Voight, *Heegner points and Sylvester's conjecture*, in this volume.

[Elk91] N. D. Elkies, *ABC implies Mordell*, Internat. Math. Res. Notices (1991), no. 7, 99–109. MR 1141316 (93d:11064)

[Fon85] J. M. Fontaine, *Il n'y a pas de variété abélienne sur* $\mathbf{Z}$, Invent. Math. **81** (1985), no. 3, 515–538. MR 807070 (87g:11073)

[FWG+92] G. Faltings, G. Wüstholz, F. Grunewald, N. Schappacher, and U. Stuhler, *Rational points*, third ed., Aspects of Mathematics, E6, Friedr. Vieweg & Sohn, Braunschweig, 1992, Papers from the seminar held at the Max-Planck-Institut für Mathematik, Bonn/Wuppertal, 1983/1984, With an appendix by Wüstholz. MR 1175627 (93k:11060)

[Grea] M. Greenberg, *The arithmetic of elliptic curves over imaginary quadratic fields and Stark-Heegner points*, in this volume.

[Greb] ———, *Computing Heegner points arising from Shimura curve parametrizations*, in this volume.

[Gro91] B. H. Gross, *Kolyvagin's work on modular elliptic curves*, L-functions and arithmetic (Durham, 1989), London Math. Soc. Lecture Note Ser., vol. 153, Cambridge Univ. Press, Cambridge, 1991, pp. 235–256. MR 1110395 (93c:11039)

[GZ86] B. H. Gross and D. B. Zagier, *Heegner points and derivatives of L-series*, Invent. Math. **84** (1986), no. 2, 225–320. MR 833192 (87j:11057)

[How] B. Howard, *Twisted Gross-Zagier theorems*, Canad. J. Math., to appear.

[How07] ———, *Central derivatives of L-functions in Hida families*, Math. Ann. **339** (2007), no. 4, 803–818. MR 2341902

[HSBT] M. Harris, N. I. Shepherd-Barron, and R. Taylor, *A family of Calabi-Yau varieties and potential automorphy*, preprint.

[Hur08] A. Hurwitz, *Über die diophantische Gleichung* $x^3y + y^3z + z^3x = 0$, Math. Ann. **65** (1908), no. 3, 428–430. MR 1511476

[Kam92] S. Kamienny, *Torsion points on elliptic curves and q-coefficients of modular forms*, Invent. Math. **109** (1992), no. 2, 221–229. MR 1172689 (93h:11054)

[Kis] M. Kisin, *Modularity of* 2*-adic Barsotti-Tate representations*, preprint, to appear.

[KW] C. Khare and J.-P. Wintenberger, *Serre's modularity conjecture (I)*, preprint, to appear.

[Lon06] M. Longo, *On the Birch and Swinnerton-Dyer conjecture for modular elliptic curves over totally real fields*, Ann. Inst. Fourier (Grenoble) **56** (2006), no. 3, 689–733. MR 2244227 (2008f:11071)

[Maz77] B. Mazur, *Modular curves and the Eisenstein ideal*, Inst. Hautes Études Sci. Publ. Math. (1977), no. 47, 33–186 (1978). MR 488287 (80c:14015)

[Maz78] ———, *Rational isogenies of prime degree (with an appendix by D. Goldfeld)*, Invent. Math. **44** (1978), no. 2, 129–162. MR 482230 (80h:14022)

[Maz86] ———, *Arithmetic on curves*, Bull. Amer. Math. Soc. (N.S.) **14** (1986), no. 2, 207–259. MR 828821 (88e:11050)

[Mer96] L. Merel, *Bornes pour la torsion des courbes elliptiques sur les corps de nombres*, Invent. Math. **124** (1996), no. 1-3, 437–449. MR 1369424 (96i:11057)

[MM97] M. R. Murty and V. K. Murty, *Non-vanishing of L-functions and applications*, Progress in Mathematics, vol. 157, Birkhäuser Verlag, Basel, 1997. MR 1482805 (98h:11106)

[Oda82] T. Oda, *Periods of Hilbert modular surfaces*, Progress in Mathematics, vol. 19, Birkhäuser Boston, Mass., 1982. MR 670069 (83k:10057)

[Par68] A. N. Paršin, *Algebraic curves over function fields. I*, Izv. Akad. Nauk SSSR Ser. Mat. **32** (1968), 1191–1219, English translation in: Math. USSR. Izv. 2 (1968). MR 0257086 (41 #1740)

[Reb] M. Rebolledo, *Merel's theorem on the boundedness of the torsion of elliptic curves*, in this volume.

[Šaf63] I. R. Šafarevič, *Algebraic number fields*, Proc. Internat. Congr. Mathematicians (Stockholm, 1962), Inst. Mittag-Leffler, Djursholm, 1963, English translation in: AMS Transl. (2) 31 (1963), pp. 163–176. MR 0202709 (34 #2569)

[Ser72] J.-P. Serre, *Propriétés galoisiennes des points d'ordre fini des courbes elliptiques*, Invent. Math. **15** (1972), no. 4, 259–331. MR 0387283 (52 #8126)

[Ser87] ———, *Sur les représentations modulaires de degré* 2 *de* $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$, Duke Math. J. **54** (1987), no. 1, 179–230. MR 885783 (88g:11022)

[Szp85] L. Szpiro, *La conjecture de Mordell (d'après G. Faltings)*, Astérisque (1985), no. 121-122, 83–103, Seminar Bourbaki, Vol. 1983/84. MR 768955 (87c:11033)

[Tay] R. Taylor, *Automorphy for some* ℓ*-adic lifts of automorphic mod* ℓ *representations. II*, preprint.

[Tay03] ———, *On icosahedral Artin representations. II*, Amer. J. Math. **125** (2003), no. 3, 549–566. MR 1981033 (2004e:11057)

[Tri06] M. Trifković, *Stark-Heegner points on elliptic curves defined over imaginary quadratic fields*, Duke Math. J. **135** (2006), no. 3, 415–453. MR 2272972 (2008d:11064)

[TW95] R. Taylor and A. Wiles, *Ring-theoretic properties of certain Hecke algebras*, Ann. of Math. (2) **141** (1995), no. 3, 553–572. MR 1333036 (96d:11072)

[Voi] J. Voight, *Shimura curve computations*, in this volume.

[Wei48] A. Weil, *Variétés abéliennes et courbes algébriques*, Actualités Sci. Ind., no. 1064 = Publ. Inst. Math. Univ. Strasbourg 8 (1946), Hermann & Cie., Paris, 1948. MR 0029522 (10,621d)

[Wil95] A. Wiles, *Modular elliptic curves and Fermat's last theorem*, Ann. of Math. (2) **141** (1995), no. 3, 443–551. MR 1333035 (96d:11071)

[Zha01a] S.-W. Zhang, *Gross-Zagier formula for* GL_2, Asian J. Math. **5** (2001), no. 2, 183–290. MR 1868935 (2003k:11101)

[Zha01b] ———, *Heights of Heegner points on Shimura curves*, Ann. of Math. (2) **153** (2001), no. 1, 27–147. MR 1826411 (2002g:11081)

[ZP89] Y. G. Zarkhin and A. N. Parshin, *Finiteness problems in diophantine geometry*, Eight Papers Translated from the Russian, Amer. Math. Soc. Transl. (2), vol. 143, 1989, from the Appendix to the Russian translation of Serge Lang's *Fundamentals of Diophantine Geometry*, "Mir", Moscow, 1986, 369–438, pp. 35–102.

Mc Gill University, The Department of Mathematics and Statistics, 805 Sherbrooke Street West, Montreal QC H3A 2K6, Canada

E-mail address: darmon@math.mcgill.ca

Clay Mathematics Proceedings
Volume 8, 2009

Non-abelian descent and the generalized Fermat equation

Hugo Chapdelaine

Abstract. We prove a finiteness result about the set of primitive solutions of the generalized Fermat equation $x^p + y^q = z^r$ when $1/p + 1/q + 1/r < 1$.

Contents

1. The main result

This Chapter gives some finiteness results for the set of primitive solutions of the generalized Fermat equation

$$x^p + y^q = z^r \tag{1}$$

where the exponents p, q, r satisfy the inequality $1/p + 1/q + 1/r < 1$. The very special 'shape' of the surface defined by (1) allows us to use some geometry to reduce its study to the study of non-abelian unramified covers of $\mathbb{P}^1\backslash\{0, 1, \infty\}$ of *signature* (p, q, r) in the sense of Definition 1.1. Therefore the study of the arithmetic of equation (1) can be transferred to the setting of algebraic curves. The main ingredients in the proof are a variant of the Chevalley-Weil theorem, and the finiteness theorems of Hermite-Minkowski and Faltings. This finiteness result for (1) which was proved in [**DG95**] can be viewed as an illustrative special case of the Campana program which was presented in Dan Abramovich's lecture series at this summer school.

The author would like to thank Henri Darmon for a careful proofreading of this article which led to many improvements.

A solution $(a, b, c) \in \mathbb{Z}^3$ of (1) is called *nontrivial* if $abc \neq 0$ and *primitive* if $\gcd(a, b, c) = 1$. When the exponents p, q and r are pairwise coprime, the following

2000 *Mathematics Subject Classification.* Primary 11D41, Secondary 11G30, 14H30.

exercise shows that (1) has infinitely many nontrivial but not necessarily primitive solutions.

Exercise 1 Let p, q and r be pairwise coprime. Show that the affine surface defined by $x^p + y^q = z^r$ in $\mathbb{A}^3_{\mathbb{Q}}$ is rational, i.e., the quotient field of $\mathbb{Q}[x, y, z]/(x^p + y^q - z^r)$ is purely transcendental of degree 2 over $\mathbb{Q}$.

From now on we are only interested in studying the set of nontrivial primitive solutions of (1). The study of (1) can be split into three cases:

(1) The *spherical case*: $1/p + 1/q + 1/r > 1$. The possibilities are $\{p, q, r\} = \{2, 2, k\}$ with $k \geq 2$, $\{2, 3, 3\}$,$\{2, 3, 4\}$ and $\{2, 3, 5\}$.
(2) The *Euclidean case*: $1/p + 1/q + 1/r = 1$. The possibilities are $\{p, q, r\} = \{3, 3, 3\}$, $\{2, 4, 4\}$ and $\{2, 3, 6\}$.
(3) The *hyperbolic case*: $1/p + 1/q + 1/r < 1$.

This division is reminiscent of the classification of algebraic curves which also falls into 3 cases depending on the genus or the sign of the Euler characteristic. Here is the main theorem that we wish to prove.

THEOREM 1.1. *(Darmon, Granville) If* $1/p + 1/q + 1/r < 1$ *then* (1) *has only finitely many nontrivial primitive solutions.*

Note that the statement of this theorem concerns the existence of integral points on a surface. We would like to reduce the study of integral solutions of (1) to the study of K-rational points on an auxiliary projective curve X/K where K is a suitable number field. We consider the map

$$\{\text{Set of nontrivial primitive solutions of equation (1)}\} \to \mathbb{P}^1(\mathbb{Q}) \subseteq \mathbb{P}^1(\mathbb{C})$$
$$(a, b, c) \mapsto \frac{a^p}{c^r},$$

which allows us to reduce the study of (1) to the study of certain branched coverings of $\mathbb{P}^1(\mathbb{C})$. We define the set

$$\Sigma_{p,q,r} := \left\{ \frac{a^p}{c^r} \in \mathbb{Q} : a^p + b^q = c^r, abc \neq 0, \gcd(a, b, c) = 1 \right\} \subseteq \mathbb{P}^1(\mathbb{Q}).$$

Exercise 2 Show that $\#\Sigma_{p,q,r} < \infty$ if and only if (1) has finitely many primitive solutions.

Now let us explain the main ideas of Theorem 1.1.

Proof of Theorem 1.1 We want to show that the set of nontrivial primitive solutions of (1) is finite. By Exercise 2, it is enough to show that $\Sigma_{p,q,r}$ is finite when $1/p + 1/q + 1/r < 1$. The proof can be broken into four steps.

First step: The existence of a Galois branched covering.

DEFINITION 1.1. A Galois covering $\pi : X \to \mathbb{P}^1$ is said to be of *signature* (p, q, r) if its ramification indices above $0, 1$ and ∞ are equal to p, q and r respectively, and if π is unramified everywhere else.

The first stage of the proof consists in constructing a Galois covering of $\mathbb{P}^1$ of signature (p, q, r) defined over a suitable number field K and Galois over that field (The construction of such a cover will be done in detail in Section 2). The Riemann-Hurwitz formula then determines the genus $g(X)$ of X in terms of the

degree d of π :

$$\begin{aligned} 2g(X)-2 &= d(2g(\mathbb{P}^1(\mathbb{C}))-2)+\frac{d}{p}(p-1)+\frac{d}{q}(q-1)+\frac{d}{r}(r-1) \\ &= d(1-1/p-1/q-1/r). \end{aligned}$$

Since $1-1/p-1/q-1/r>0$ we conclude that $g(X)>1$.

Second step: A Chevalley-Weil theorem for branched coverings.

Given $t\in\mathbb{P}^1(K)$, let L_t be the smallest field of definition of the closed points in $\pi^{-1}(t)$. As is explained in Section 3, the field L_t is a Galois extension of K with Galois group isomorphic (non-canonically) to a subgroup of $\mathrm{Gal}(X/\mathbb{P}^1)$. The Chevalley-Weil theorem for branched coverings (see Theorem 3.2) shows that the ramification of L_t, for $t\in\Sigma_{p,q,r}$, is bounded *independently* of t, in light of the following elementary property of $\Sigma_{p,q,r}$:

LEMMA 1.1. Let $t=\frac{a^p}{c^r}\in\Sigma_{p,q,r}$; then for all prime numbers ℓ we have

(1) $v_\ell(\mathrm{Numerator}(t))\equiv 0 \pmod p$,
(2) $v_\ell(\mathrm{Numerator}(t-1))\equiv 0 \pmod q$,
(3) $v_\ell(\mathrm{Numerator}(\frac{1}{t}))\equiv 0 \pmod r$,

where for $x\in\mathbb{Q}$, $v_\ell(x)$ stands for the valuation of x at the prime ℓ.

Note that the proof of Lemma 1.1 uses in a crucial way the primitivity of the solution (a,b,c) corresponding to $t=\frac{a^p}{c^r}$ and the fact that $t-1=-\frac{b^q}{c^r}$.

Third step: Hermite-Minkowski.

By the Hermite-Minkowski theorem (cf. Theorem 1.1 in Section 1.1 of [**Dar**]) the compositum L of all the number fields L_t, for $t\in\Sigma_{p,q,r}$, is a finite extension of K.

Fourth step: Faltings' Theorem.

By definition of L we have $\pi^{-1}(\Sigma_{p,q,r})\subseteq X(L)$. Since $g(X)>1$, we deduce by Faltings' theorem that $X(L)$ is a finite set and therefore $\pi^{-1}(\Sigma_{p,q,r})$ and $\Sigma_{p,q,r}$ are also finite sets. This concludes the sketch of the proof of Theorem 1.1. □

REMARK 1.1. The conclusion of Theorem 1.1 remains the same if we replace the equation $x^p+y^q=z^r$ by the more general equation $Ax^p+By^q=Cz^r$ for nonzero fixed integers A, B and C. For a further discussion of the equation $Ax^p+By^q=Cz^r$, see [**DG95**].

REMARK 1.2. In some special cases, for example when $(p,q,r)=(n,n,n)$ with $n\geq 3$ we know by the work of Wiles and Taylor (see [**Wil95**] and [**TW95**]) that (1) has no nontrivial solutions. Using similar techniques, Darmon and Merel (see [**Dar00**] and [**DM97**]) could also treat the case (p,p,r) where $r=2$ or 3 and p is a prime number larger than or equal to $6-r$ to conclude that (1) has no nontrivial primitive solutions.

For the rest of the paper, we would like first to explain in detail the construction of the auxiliary branched covering $(X_K,\pi,\mathbb{P}^1_K)$ of signature (p,q,r) above $\{0,1,\infty\}$ which was needed in the first step of the proof of Theorem 1.1. Secondly, we would like to give a more detailed discussion about the variant of the Chevalley-Weil theorem that we have used to control the ramification of the number field L_t over K for the special elements $t\in\Sigma_{p,q,r}$. We won't say anything about Steps 3 and 4, which are discussed in [**Dar**]. Sections 2 and 3 are devoted to a discussion of Steps 1 and 2 respectively.

2. Construction of the branched covering

In this section we will use the theory of Riemann surfaces in order to construct certain *analytic* Galois branched coverings over $\mathbb{P}^1(\mathbb{C})$ unramified outside $\{0,1,\infty\}$.

For every triple of integers (p,q,r) with $p,q,r \geq 2$ we define the Hecke triangle group by the abstract presentation

$$\Gamma_{p,q,r} := \langle \gamma_0, \gamma_1, \gamma_\infty | \gamma_0^p = \gamma_1^q = \gamma_\infty^r = \gamma_0\gamma_1\gamma_\infty = 1 \rangle.$$

It is convenient to allow the exponents p, q and r to be infinite, which will be taken to mean that the order of the corresponding element is infinite.

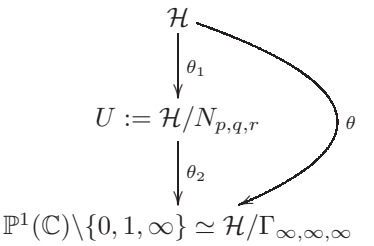
One has that $\pi_1(\mathbb{P}^1(\mathbb{C})\backslash\{0,1,\infty\}) \simeq \Gamma_{\infty,\infty,\infty} = \langle l_0, l_1, l_\infty | l_0 l_1 l_\infty = 1 \rangle$, which is isomorphic to the free group on two generators. We have the short exact sequence

$$1 \to N_{p,q,r} \to \Gamma_{\infty,\infty,\infty} \xrightarrow{\varphi} \Gamma_{p,q,r} \to 1,$$

where $\varphi(l_0) = \gamma_0$, $\varphi(l_1) = \gamma_1$ and $N_{p,q,r} = \ker(\varphi)$. The universal covering space of $\mathbb{P}^1(\mathbb{C})\backslash\{0,1,\infty\})$ is the upper half-plane, see for example Theorem 6.4.3 of [**Ser92**]. Let us denote by

$$\theta : \mathcal{H} \to \mathbb{P}^1(\mathbb{C})\backslash\{0,1,\infty\} \tag{2}$$

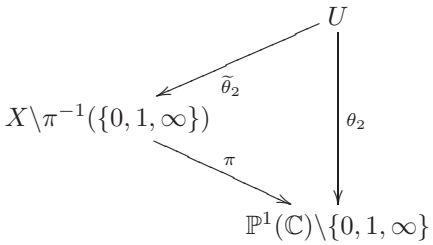
a choice of such a universal covering map. From the theory of covering spaces one has a (non-canonical) isomorphism between the group of deck transformations of (2) and the fundamental group of $\mathbb{P}^1(\mathbb{C})\backslash\{0,1,\infty\})$; see for example §80 of [**Mun00**]. Such an isomorphism allows us to define an action of $\Gamma_{\infty,\infty,\infty} \simeq \pi_1(\mathbb{P}^1(\mathbb{C})\backslash\{0,1,\infty\})$ on $\mathcal{H}$. From this, one may deduce the following diagram:



where θ_1 (resp. θ_2) is the covering map induced by the action of $N_{p,q,r}$ on $\mathcal{H}$ (resp. $\Gamma_{\infty,\infty,\infty}/N_{p,q,r}$ on U).

Note that U is a connected Riemann surface such that $\pi_1(U) \simeq N_{p,q,r}$ and that θ_2 is a Galois covering map with Galois group isomorphic to $\Gamma_{\infty,\infty,\infty}/N_{p,q,r} \simeq \Gamma_{p,q,r}$. One can show that θ_2 is of finite degree if and only if $1/p+1/q+1/r > 1$ (see Exercise 4). Since in our setting we work under the assumption that $1/p + 1/q + 1/r < 1$ we see that in this case the map θ_2 is never of finite degree. The pair (U, θ_2) is universal among all Galois coverings over $\mathbb{P}^1(\mathbb{C})\backslash\{0,1,\infty\}$ of signature (p,q,r) in the following sense: Let $\pi : X \to \mathbb{P}^1(\mathbb{C})$ be a Galois branched covering unramified outside $\{0,1,\infty\}$ with ramification index p above 0, q above 1 and r above ∞. Then π factors through θ_2, i.e., there exists a covering map $\widetilde{\theta_2} : U \to X\backslash\pi^{-1}(\{0,1,\infty\})$

which makes the following diagram commutative:



Note that $\widetilde{\theta}_2$ is onto and unramified since all the ramification happens already in π. Let us assume that π is finite of degree d; then, in this case, X is a compact Riemann surface. Using the Riemann-Hurwitz formula one gets that

$$\begin{aligned} 2g(X) - 2 &= d(2g(\mathbb{P}^1(\mathbb{C})) - 2) + \frac{d}{p}(p-1) + \frac{d}{q}(q-1) + \frac{d}{r}(r-1) \\ &= d(1 - 1/p - 1/q - 1/r). \end{aligned}$$

We thus see that

(1) $g(X) = 0$ if $1/p + 1/q + 1/r > 1$,
(2) $g(X) = 1$ if $1/p + 1/q + 1/r = 1$,
(3) $g(X) \geq 2$ if $1/p + 1/q + 1/r < 1$.

Again using Theorem 6.4.3 of [**Ser92**], one may deduce that the universal covering space of X is $\mathbb{P}^1(\mathbb{C})$ if $1/p + 1/q + 1/r > 1$, $\mathbb{C}$ if $1/p + 1/q + 1/r = 1$, and $\mathcal{H}$ if $1/p + 1/q + 1/r < 1$. This explains the trichotomy for the study of (1).

We would like to give a geometrical realization of the universal pair (U, θ_2) in the case where $1/p + 1/q + 1/r < 1$. This will be used to understand the set of elliptic elements of $\Gamma_{p,q,r}$ (see Exercise 3). Since $1/p + 1/q + 1/r < 1$, there exists a hyperbolic triangle in the Poincaré unit disc with angles $\pi/p, \pi/q, \pi/r$; see Figure 1. Let σ_P be the symmetry with respect to the geodesic passing through

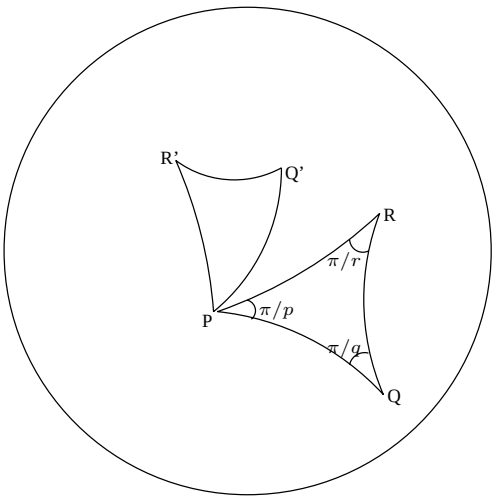


FIGURE 1. Hyperbolic triangle inside the Poincaré disc.

QR, σ_Q the symmetry with respect to the geodesic passing through PR and σ_R the symmetry with respect to the geodesic passing through PQ. Let $\gamma_P = \sigma_Q\sigma_R$ be the rotation around P with angle $\frac{2\pi}{p}$, $\gamma_Q = \sigma_R\sigma_P$ be the rotation around Q with angle $\frac{2\pi}{q}$ and $\gamma_R = \sigma_P\sigma_Q$ be the rotation around R with angle $\frac{2\pi}{r}$. We have drawn the image of the triangle PQR under the rotation γ_P in Figure 1. Since the open unit disc $D(0,1)$ is biholomorphic to $\mathcal{H}$ we can identify the group $\langle\gamma_P, \gamma_Q, \gamma_R\rangle$ as a subgroup of $PSL_2(\mathbb{R}) \simeq \mathrm{Aut}(\mathcal{H})$. We have an isomorphism between $\langle\gamma_P, \gamma_Q, \gamma_R\rangle$ and $\Gamma_{p,q,r}$ given by $\gamma_P \mapsto \gamma_0$, $\gamma_Q \mapsto \gamma_1$ and $\gamma_R \mapsto \gamma_\infty$ (prove it by using Figure 1). In particular, we can think of $\Gamma_{p,q,r}$ as a subgroup of $PSL_2(\mathbb{R})$. The group $\Gamma_{p,q,r}$, when applied to the triangle PQR, gives a 'half tessellation' of $D(0,1)$. A fundamental domain for the action of $\Gamma_{p,q,r}$ on $D(0,1)$ is given for example by the geodesic quadrilateral $PQRQ'$, where the geodesic RQ' is identified with the geodesic RQ and the geodesic PQ with the geodesic PQ'. It thus follows that the quotient $\mathcal{H}/\Gamma_{p,q,r}$ is isomorphic to $\mathbb{P}^1(\mathbb{C})$. Let

$$\widetilde{\pi} : \mathcal{H} \to \mathcal{H}/\Gamma_{p,q,r} \simeq \mathbb{P}^1(\mathbb{C}).$$

Since $PSL_2(\mathbb{C})$ acts triply transitively on $\mathbb{P}^1(\mathbb{C})$ we can assume that $\widetilde{\pi}(P) = 0, \widetilde{\pi}(Q) = 1$ and $\widetilde{\pi}(R) = \infty$. Therefore the Galois branched covering $\widetilde{\pi}$ has signature (p,q,r) above $\{0,1,\infty\}$. Unfortunately $\widetilde{\pi}$ has infinite degree but the next lemma takes care of this difficulty.

Exercise 3 Define $U := \mathcal{H}\backslash\widetilde{\pi}^{-1}\{0,1,\infty\}$. Show that the map

$$\widetilde{\pi}|_U : U \to \mathbb{P}^1(\mathbb{C})\backslash\{0,1,\infty\}$$

corresponds to the universal map associated to Galois branched coverings over $\mathbb{P}^1(\mathbb{C})$ of signature (p,q,r). It thus gives a geometrical realization of U as the unit disc minus the vertices of all the $\Gamma_{p,q,r}$-translates of the triangle PQR. Conclude that an element $\gamma \in \Gamma_{p,q,r}$ is elliptic if and only if it fixes a vertex of a $\Gamma_{p,q,r}$-translate of the triangle PQR. Recall that an elliptic element in $PSL_2(\mathbb{R})$ is by definition a matrix which fixes a point in $\mathcal{H}$.

Exercise 4 Show that $\Gamma_{p,q,r}$ is finite if and only if $1/p + 1/q + 1/r > 1$. Show that $\Gamma_{p,q,r}$ is infinite and nonabelian if and only if $1/p + 1/q + 1/r \leq 1$.

LEMMA 2.1. There exists a normal subgroup $H \leq \Gamma_{p,q,r}$ such that $[\Gamma_{p,q,r} : H] < \infty$ and such that H acts without fixed point, i.e., H contains no elliptic elements.

REMARK 2.1. Note that the set of all elliptic elements of $\Gamma_{p,q,r}$ consists of the union of the conjugacy classes in $\Gamma_{p,q,r}$ of $\gamma_0^{\mathbb{Z}}$, $\gamma_1^{\mathbb{Z}}$ and $\gamma_\infty^{\mathbb{Z}}$. Moreover, if H is as in Lemma 2.1 then the orders of $\overline{\gamma}_0$, $\overline{\gamma}_1$ and $\overline{\gamma}_\infty$ in $\Gamma_{p,q,r}/H$ are equal to p, q and r, respectively.

Proof of lemma 2.1 We follow essentially the proof of Proposition 4.4 of [**Beu04**]. Let us construct an abstract group homomorphism of $\Gamma_{p,q,r}$ onto a certain subgroup of $PSL_2(\mathbb{C})$ for which all its matrices have algebraic entries. Consider the matrices

$$A = \begin{pmatrix} 0 & -\zeta_{2p}^{-1} \\ \zeta_{2p} & \zeta_{2p} + \zeta_{2p}^{-1} \end{pmatrix} \qquad C = \begin{pmatrix} 0 & \zeta_{2p}^{-1}\zeta_{2q}^{-1} \\ -\zeta_{2p}\zeta_{2q} & \zeta_{2r} + \zeta_{2r}^{-1} \end{pmatrix} \qquad B = AC^{-1}$$

where $\zeta_n = e^{2\pi i/n}$. One can verify that the orders of A, B and C in $PSL_2(\mathbb{C})$ are p, q and r, respectively. For example, to show that A has order p one can use the observation that $(-1,1)$ and $(-\zeta_p^{-1}, 1)$ are eigenvectors with eigenvalues ζ_{2p} and

ζ_{2p}^{-1}. A similar argument can be used for B and C. We thus have an onto group homomorphism

$$\rho : \Gamma_{p,q,r} \to \langle A, B, C\rangle =: \mathcal{N} \subseteq PSL_2(R)$$

given by $\rho(\gamma_0) = A^{-1}, \rho(\gamma_1) = B, \rho(\gamma_\infty) = C$, where $R = \mathbb{Z}[\zeta_{2p}, \zeta_{2q}, \zeta_{2r}]$. Note that ρ sends an elliptic element of $\Gamma_{p,q,r}$ to an elliptic element of $\mathcal{N}$ and all elliptic elements of $\mathcal{N}$ are contained in a conjugacy class of $A^{\mathbb{Z}}$, $B^{\mathbb{Z}}$ or $C^{\mathbb{Z}}$. Let π be some prime ideal of R. Note that $A = P \begin{pmatrix} \zeta_{2p} & 0 \\ 0 & \zeta_{2p}^{-1} \end{pmatrix} P^{-1}$ for some matrix $P \in PSL_2(R)$. Therefore if $A \pmod{\pi} \equiv I \pmod{\pi}$ then $\begin{pmatrix} \zeta_{2p} & 0 \\ 0 & \zeta_{2p}^{-1} \end{pmatrix} \equiv I \pmod{\pi}$, where I stands for the identity matrix. This implies that $\pi|(1 - \zeta_{2p})$. We have a similar thing for B and C. Let us choose a prime ideal π such that π does not divide $1-\zeta_n^k$ for $1 \leq k \leq n-1$ and $n \in \{p, q, r\}$. Finally, define the group

$$H := \{g \in \Gamma_{p,q,r} | \rho(g) \equiv I \pmod{\pi}\}. \tag{3}$$

The group H satisfies the property of Lemma 2.1. □

We can finally define the auxiliary curve that was used in the course of the proof of Theorem 1.1. Define

$$X := \mathcal{H}/H,$$

where H is as in Lemma 2.1 and let π be the natural map

$$\pi : X \to \mathcal{H}/\Gamma_{p,q,r} \simeq \mathbb{P}^1(\mathbb{C}). \tag{4}$$

By construction π is a finite complex analytic Galois branched covering over $\mathbb{P}^1(\mathbb{C})$ of signature (p, q, r). Since π has finite degree and $\mathbb{P}^1(\mathbb{C})$ is compact we deduce that X is a compact Riemann surface. Note that the complex structure of X is inherited from the complex structure of $\mathcal{H}$, where some care should be taken in order to define local charts around fixed points of elliptic elements of $\Gamma_{p,q,r}$.

There is a dictionary between non singular projective curves over $\mathbb{C}$ and compact Riemann surfaces:

THEOREM 2.1. *Any compact Riemann surface S is algebraic.*

Let us sketch a proof of this important result in the special case where S is the compact Riemann surface X that was previously constructed as a quotient of the upper half-plane.

Sketch of the proof We will break the proof into three steps.

Step 1: X admits a large supply of non-constant meromorphic functions.

We first show that X admits a large supply of non-constant meromorphic functions in the sense that for every pair of points $P, Q \in X$ with $P \neq Q$ there exists a meromorphic function f on X such that $f(P) \neq f(Q)$ (separates points) and for every $P \in X$ there exists a meromorphic function g on X such that g is a local chart in a small neighborhood of P (separates tangents).

Let G be the preimage of H under the natural projection $SL_2(\mathbb{R}) \to PSL_2(\mathbb{R})$. Note that G is a discrete subgroup of $SL_2(\mathbb{R})$ which contains the element $-I$. For every pair of points $P, Q \in \mathcal{H}$ consider the Poincaré series (modular form)

$$f_m(P, Q, z) = \sum_{g \in G} r_{P,Q}(gz) j(g, z)^{-m} \tag{5}$$

where m is any fixed *even integer* larger or equal to 4, $r_{P,Q}(z) = \frac{z-P}{z-Q}$, $gz = \frac{az+b}{cz+d}$ and $j(g,z) = (cz+d)$ for $g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in G$. The infinite sum (5) converges absolutely since $m \geq 3$. Therefore the function $f_m(P,Q,z)$ is meromorphic on $\mathcal{H}$ and satisfies the important transformation formula

$$f_m(P,Q,gz) = (cz+d)^m f_m(P,Q,z) \quad \forall\, g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in G. \tag{6}$$

Note that when m is odd, the transformation formula (6) applied to the matrix $-I \in G$ implies that $f_m(P,Q,z)$ is identically equal to 0. Let $P, Q \in \mathcal{H}$ be arbitrary points such that GP and GQ are distinct right orbits. Now choose a third point $R \in \mathcal{H}$ such that $f_m(R,Q,z)$ does not vanish at $z = P$ (It is easy to see that such a point R always exists by considering for example the function $\omega \mapsto f_m(\omega, Q, P)$). A simple calculation reveals that the function $f_m(R,Q,z)$ has a pole of order one at every elements of GQ (this uses the fact that m is even). It thus follows that $f_m(R,Q,z)$ is a non-constant meromorphic function on $\mathcal{H}$. Now let us consider the quotient

$$F_m(z) = \frac{f_m(R,P,z)}{f_m(R,Q,z)}.$$

Using (6), one readily sees that $F_m(z)$ descends to a meromorphic function on X. Moreover, by construction, it has a zero of order one at the point $GQ \in X$ and a pole of order one at the point $GP \in X$. This shows that X has a set of meromorphic functions that separates points and tangents.

Step 2: Riemann-Roch.

Let D be a divisor of X and let $\mathcal{L}_D$ be the locally free $\mathcal{O}_X$-module of rank 1 associated to D where for every open set $U \subseteq X$

$$\mathcal{L}_D(U) = \{ f : U \to \mathbb{C} : f \text{ is meromorphic and } \operatorname{div}(f) \geq -D|_U \}.$$

Then the famous theorem of Riemann-Roch says

THEOREM 2.2. *(Riemann-Clebsch-Roch)*

$$\dim_{\mathbb{C}} H^0(X, \mathcal{L}_D) - \dim_{\mathbb{C}} H^1(X, \mathcal{L}_D) = \deg(D) + 1 - g,$$

where g stands for the genus of X and $\deg(D)$ for the degree of the divisor D.

For an elementary proof of Theorem 2.2 which uses only Step 1, see chapter IV of [**Mir95**].

Step 3: Construction of a planar parametrization of X.

PROPOSITION 2.1. Let z be a non-constant meromorphic function on X. Then there exists a meromorphic function f and an irreducible algebraic equation $P(z,f)$ defined over $\mathbb{C}$ such that the map

$$x \mapsto (z(x), f(x))$$

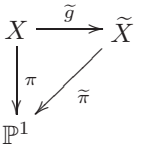
is a conformal bijection of X onto the compact Riemann surface associated to the irreducible equation $P(z,f) = 0$.

Proposition 2.1 is a nice application of Riemann-Roch and the analytic continuation principle for germs of holomorphic functions. For a detailed proof, see for example the discussion on p. 242 of [**Jos06**]. This concludes the sketch of the proof of the algebraicity of X. $\square$

REMARK 2.2. Historically, Riemann proved the inequality $\dim_{\mathbb{C}} H^0(X, \mathcal{L}_D) \leq \deg(D) + 1$ by constructing meromorphic differential forms with prescribed poles at points appearing in D, see [**Rie57**]. His construction appealed to the so-called Dirichlet's principle which back then was not rigorously proved. An inequality going in the other direction was proved by Clebsch (see [**Cle65**]) and then refined by Roch (see [**Roc65**]). In general, the construction of non-constant meromorphic functions (or non-zero meromorphic differential forms) on an abstract compact complex manifold of dimension one (i.e., a compact Riemann surface) is a highly non-trivial fact. When the dimension is higher than one, it is the lack of non-constant meromorphic functions which prevents compact complex manifolds to be algebraic. In dimension one, the construction of such functions can done abstractly by the use of harmonic analysis; see for example Section 5.2 of [**Jos06**]. Note that in Step 1 of the previous argument, we could get around this non-trivial fact by taking advantage of the description of X as a certain quotient of $\mathcal{H}$. This allowed us to define directly Poincaré series which are meromorphic $\frac{m}{2}$-fold differential forms. The idea of constructing meromorphic $\frac{m}{2}$-fold differential forms by averaging over the elements of a Fuchsian group is due to Poincaré. Poincaré was the first one to announce that for every algebraic curve $P(x, y) = 0$ (of genus ≥ 2) there exists two non-constant Fuchsian functions $f(z)$ and $g(z)$ such that $P(f(z), g(z)) \equiv 0$, see [**Poi82**]. Finally, we should mention a more recent way of proving Theorem 2.1 under the *additional assumption* that the compact Riemann surface S admits a single non-constant meromorphic function f, i.e., a non-constant holomorphic function $f : S \to \mathbb{P}^1(\mathbb{C})$. This alternative approach is a special case of a general equivalence between analytic and algebraic coherent sheaves on smooth projective algebraic varieties. Very often, this equivalence is quoted under the acronym 'GAGA principle'; see Section 6.1 of [**Ser92**] and [**Ser56**]. The key point is that this holomorphic function $f : S \to \mathbb{P}^1(\mathbb{C})$ gives rise to a coherent analytic sheaf $\mathcal{F}$ on $\mathbb{P}^1(\mathbb{C})$ (which is an algebraic curve) and therefore, by GAGA, $\mathcal{F}$ is an algebraic sheaf. From this we may conclude that S is algebraic.

Now we recall that we have constructed previously a branched covering $\pi : X \to \mathbb{P}^1(\mathbb{C})$ of signature (p, q, r). Now, armed with Proposition 2.1, we know that there exists a meromorphic function f on X and a polynomial $P(x, y) \in \mathbb{C}[x, y]$ such that $P(\pi, f) = 0$. Let M be the subfield of $\mathbb{C}$ generated by the coefficients of $P(x, y)$. Note that M is a finitely generated field over $\mathbb{Q}$. In general the field M will not be an algebraic extension over $\mathbb{Q}$. Nevertheless we have the following key proposition:

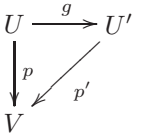
PROPOSITION 2.2. There exists a smooth projective algebraic curve $\widetilde{X}$ defined over a number field K such that the following diagram commutes:



where $\widetilde{g} : X(\mathbb{C}) \to \widetilde{X}(\mathbb{C})$ is an isomorphism defined over $\mathbb{C}$ and where $\widetilde{\pi}$ is a branched covering defined over K.

Proposition 2.2 is a direct application of the following general result:

THEOREM 2.3. *Let V be an algebraic variety defined over an algebraically closed field L of characteristic 0, and let L' be an algebraically closed field extension of L. Then every covering $p : U \to V$ defined over L' comes from a covering $p' : U' \to V$ defined over L in the sense that there exists a commutative diagram*



where g is an isomorphism of varieties defined over L' and p' is a covering defined over L.

Proof See the proof of Theorem 6.3.3 of [**Ser92**]. □

Let us explain how the existence of $\widetilde{f}$ and $\widetilde{\pi}$ follows from Theorem 2.3. Let Y be the algebraic curve over $\mathbb{C}$ defined by $X \backslash \pi^{-1}\{0, 1, \infty\}$. Note that $\pi|_Y : Y \to \mathbb{P}^1 \backslash \{0, 1, \infty\}$ is a covering defined over $\mathbb{C}$ and that $\mathbb{P}^1 \backslash \{0, 1, \infty\}$ is an algebraic curve defined over $\overline{\mathbb{Q}}$ (in fact over $\mathbb{Q}$!). From Theorem 2.3, we know that there exists a covering $\pi' : Y' \to \mathbb{P}^1 \backslash \{0, 1, \infty\}$ defined over $\overline{\mathbb{Q}}$ and an isomorphism $g : Y \to Y'$ defined over $\mathbb{C}$ such that $\pi' \circ g = \pi$. Let K be the field generated by the coefficients of the equations defining the algebraic curve Y'. Note that K is finitely generated over $\mathbb{Q}$ and therefore it is a number field. The open Riemann surfaces $Y(\mathbb{C})$ and $Y'(\mathbb{C})$ admit natural compactifications X and $\widetilde{X}$ (just add the deleted points) where $\widetilde{X}$ can be chosen to be defined over K. Finally, note that the map g (resp. π') extends uniquely to a map $\widetilde{g} : X \to \widetilde{X}$ defined over $\mathbb{C}$ (resp. $\widetilde{\pi} : \widetilde{X} \to \mathbb{P}^1$ defined over K).

REMARK 2.3. Unfortunately, the proof of Theorem 2.3 doesn't give any control on the number field K which appears in Proposition 2.2. For a different proof which gives some control on the number field K, see [**Köc04**].

REMARK 2.4. Note that Proposition 2.2 implies the 'if part' of the famous *Belyi's theorem*, which states that a compact Riemann surface X admits a model over $\overline{\mathbb{Q}}$ if and only if there exists a branched covering $\pi : X \to \mathbb{P}^1(\mathbb{C})$ which is unramified outside $\{0, 1, \infty\}$. Historically, this direction is due to Weil; see [**Wei56**]. The 'only if part' is not really longer to prove, in fact it is shorter. Its proof is completely algorithmic and is due to Belyi; see [**Bel79**].

REMARK 2.5. In general, for higher dimensional complex varieties one has the following criterion which characterizes varieties which admit a model over a number field

THEOREM 2.4. *(Gonzàlez-Diez) An irreducible complex projective variety X can be defined over a number field if and only if the family of all its conjugates X^σ, where σ is any field automorphism of $\mathbb{C}$, contains only finitely many isomorphism classes of complex projective varieties.*

For a proof of this criterion see [**GD06**].

Combining Theorem 2.1, Proposition 2.2 and our discussion on branched coverings we see that every finite index normal subgroup $H \leq \Gamma_{p,q,r}$ which contains no elliptic elements gives rise to an algebraic Galois branched covering over $\mathbb{P}^1$ of signature (p, q, r) defined over a suitable number field K, where the number field

K depends on H. Such covers turn out to be extremely useful since they can be used to study the set of integral solutions of (1). From the previous observation one may deduce the following general principle:

PRINCIPLE 2.1. *There is a dictionary between the distinct strategies for studying $x^p + y^q = z^r$ and the finite quotients of the Hecke triangle group $\Gamma_{p,q,r}$.*

This principle is slightly imprecise but at least, from the author's point of view, has the virtue of being inspiring. We won't say more about it and we encourage the reader to look at [**Dar04**] where Principle 2.1 is explained in greater detail.

3. A Chevalley-Weil theorem for branched coverings

In this section we would like to present a variant of the Chevalley-Weil theorem that allowed us, during the second step of the proof of Theorem 1.1, to control the ramification of the field extension L_t over K for the special elements $t \in \Sigma_{p,q,r}$. Let us first recall the Chevalley-Weil theorem in the context of curves (see also Section 1.2 of [**Dar**]).

THEOREM 3.1. *(Chevalley-Weil) Let X and Y be smooth schemes of relative dimension one defined over the ring of S-integers $\mathcal{O}_{L,S}$ of a number field L, where S is a finite set of places of L. Let $f : X \to Y$ be a morphism of schemes defined over $\mathcal{O}_{L,S}$ which is unramified over the generic fiber. Then there exists a finite extension L'/L such that*

$$f^{-1}(Y(\mathcal{O}_{L,S})) \subseteq X(\mathcal{O}_{L',S'}),$$

where the set of places S' extends the set of places of S.

REMARK 3.1. In the statement it was important to work with integral models of X and Y in order to make sense of the notion of integral points, i.e., $\mathcal{O}_{L,S}$-valued points. In general, the notions of integral points and rational points differ since the set $X(\mathcal{O}_{L,S})$ could be smaller that the set $X(L)$. For example, consider the *affine* curve E defined by the equation $y^2 - x^3 - 73x = 0$. By Siegel's Theorem one has that $\#E(\mathbb{Z}) < \infty$. On the other hand, since the Mordell-Weil group of $E/\mathbb{Q}$ has positive rank, one has that $\#E(\mathbb{Q}) = \infty$. However, there is an important situation where the two notions coincide, namely in the special case where the curve X is projective.

REMARK 3.2. At this point we can't resist giving a nice application of the Chevalley-Weil theorem when combined with Faltings' theorem. Consider the affine complex curve embedded in $\mathbb{A}^4(\mathbb{C})$ defined by the zero locus

$$Z(u+v-1, uw-1, vt-1) = \{(u,v,w,t) \in \mathbb{A}^4(\mathbb{C}) : u+v-1 = uw-1 = vt-1 = 0\}.$$

It is easy to see that the map

$$\begin{aligned} \mathbb{P}^1(\mathbb{C})\backslash\{0,1,\infty\} &\to Z(u+v-1, uw-1, vt-1) \\ [u,1] &\mapsto (u, 1-u, 1/u, 1/(u-1)) \end{aligned}$$

is an isomorphism of complex curves. From this, we deduce that the coordinate ring of $\mathbb{P}^1(\mathbb{C})\backslash\{0,1,\infty\}$ is

$$\mathbb{C}[u, \frac{1}{u}, \frac{1}{u-1}] \simeq \mathbb{C}[u,v,w,t]/(u+v-1, uw-1, vt-1).$$

Now choose a *covering* (so unramified)

$$\pi : Y(\mathbb{C}) \to \mathbb{P}^1(\mathbb{C})\backslash\{0,1,\infty\}$$

where $Y(\mathbb{C})$ is an open Riemann surface of genus larger than or equal to 2 (there are infinitely many possibilities for π). Finally, combining Faltings and Chevalley-Weil we may conclude that the equation

$$u + v = 1$$

has only finitely many solutions in $\mathcal{O}_{L,S}^{\times}$ where L is an arbitrary number field and S is any finite set of places of L. Historically, Siegel was the first to prove this result. Of course, he proved it without appealing to Faltings' theorem.

For the rest of this section we would like to discuss in more detail the variant of the Chevalley-Weil theorem that was used in the proof of Theorem 1.1. Let $(X_K, \pi, \mathbb{P}^1_K)$ be the algebraic Galois branched covering of degree d, with Galois group G and signature (p,q,r) constructed in Section 2. Let us fix an embedding of K into $\mathbb{C}$. Since π is defined over K we have a natural action of $\mathrm{Gal}(\mathbb{C}/K)$ on all the fibers of π above points $t \in \mathbb{P}^1(K)$. Moreover, for every $t \in \mathbb{P}^1(\mathbb{C})\backslash\{0,1,\infty\}$, we have a simply transitive action of G on $\pi^{-1}(t)$ since π is Galois. We thus get two group homomorphisms:

$$\rho_1 : \mathrm{Gal}(\mathbb{C}/K) \to \mathrm{Sym}(\pi^{-1}(t)) \quad \text{and} \quad \rho_2 : G \to \mathrm{Sym}(\pi^{-1}(t)).$$

It is important to know how ρ_1 and ρ_2 are related. Let us choose a complex embedding $\varphi : X(\mathbb{C}) \hookrightarrow \mathbb{P}^N(\mathbb{C})$. For every $P \in X(\mathbb{C})$ let us denote the image of P by φ by $\varphi(P) = [\varphi_0(P), \varphi_1(P), \ldots, \varphi_N(P)] \in \mathbb{P}^N(\mathbb{C})$. For $t \in \mathbb{P}^1(K)\backslash\{0,1,\infty\}$ define the number field L_t to be the field generated over K by all the coordinates of $\varphi(P)$ for all $P \in \pi^{-1}(t)$. Let $\pi^{-1}(t) = \{P_1, \ldots, P_d\}$. The first thing to notice is that the number field $L' := K(\varphi(P_1))$ is equal to L_t. For every $i \in \{1, \ldots, d\}$ there exists an element $g \in G$ such that $g(\varphi(P_1)) = \varphi(P_i)$. Therefore the coordinates of $\varphi(P_i)$ can be expressed algebraically in terms of the coordinates of $\varphi(P_1)$ so $L' = L_t$. It thus follows that the action of an element $\sigma \in \mathrm{Gal}(\overline{K}/K)$ on L_t is completely determined by its action on the coordinates of $\varphi(P_1)$. Since $\sigma(\varphi(P_1)) = \varphi(P_i)$ for a unique i we readily see that every automorphism of L_t/K can be realized 'algebraically' by the action of a unique element $g \in G$ (G acts simply transitively on the fibers). We have the following identification

$$\mathrm{Gal}(L_t/K) = \{g \in G : \exists \sigma \in \mathrm{Gal}(\overline{K}/K) \text{ such that } \sigma(\varphi(P_1)) = g\varphi(P_1)\} \subseteq G.$$

We would now like to understand the ramification of L_t over K when $t \in \mathbb{P}^1(K)$. The morphism $\pi : X_K \to \mathbb{P}^1_K$ induces an inclusion of fields $K(\mathbb{P}^1) \simeq K(x) \hookrightarrow K(X)$, where x is a variable. Note that $K(X)/K(x)$ is Galois. Let $t \in K$. We define the specialization of π at t to be the K-algebra map

$$K \simeq K[x]/(x-t) \hookrightarrow K[X]/(x-t)$$

where $K[X]$ corresponds to the integral closure of $K[x]$ in $K(X)$. Let

$$t \in \mathbb{P}^1(K)\backslash\{0,1,\infty\}.$$

Since π is unramified at all the points above t we have $(x-t)K[X] = \mathfrak{p}_1 \cdots \mathfrak{p}_r$ where the $\mathfrak{p}_i$'s are distinct prime ideals of $K[X]$. We thus find that $K[X]/(x-t) \simeq L_1 \oplus \cdots \oplus L_r$, where $L_i = K[X]/\mathfrak{p}_i$. Note that all L_i's are Galois over K with Galois group $D(\mathfrak{p}_i/(x-t)) = \{g \in G : g(\mathfrak{p}_i) = \mathfrak{p}_i\}$ so that all the L_i's collapse to the same number field in a fixed algebraic closure of K.

Exercise 5 Show that $L_i/K \simeq L_t/K$.

In order to understand the ramification of L_i over K we need to define the arithmetic intersection between two points $a, b \in \mathbb{P}^1(K)$ at a prime ideal $\wp$ of K.

DEFINITION 3.1. Let $\wp$ be a prime ideal of K and $a, b \in K \cup \{\infty\}$. We define

$$I_\wp(a,b) := \begin{cases} \mathrm{ord}_\wp(a-b) & \text{if } \mathrm{ord}_\wp(a) \geq 0, \mathrm{ord}_\wp(b) \geq 0 \\ \mathrm{ord}_\wp(\frac{1}{a} - \frac{1}{b}) & \text{if } \mathrm{ord}_\wp(\frac{1}{a}) \geq 0, \mathrm{ord}_\wp(\frac{1}{b}) \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\mathrm{ord}_\wp(0) = \infty$ and $\mathrm{ord}_\wp(\infty) = -\infty$.

Before stating the Chevalley-Weil theorem for branched coverings we need to make one more definition.

DEFINITION 3.2. Let $X \xrightarrow{G} \mathbb{P}^1$ be a Galois branched covering over $\mathbb{C}$. A Galois branched covering $x : X_K \xrightarrow{G} \mathbb{P}^1_K$ is called a good model for $X \xrightarrow{G} \mathbb{P}^1$ over K if the primes of $\mathcal{O}_K$ (when viewed as primes in $\mathcal{O}_K[x]$) that ramify in $\mathcal{O}_K[X_K]$ are contained in S_{bad}. The ring $\mathcal{O}_K[X_K]$ stands for the integral closure of $\mathcal{O}_K[x]$ in $K(X_K)$ and the set S_{bad} is the union of the set of primes that divide the order of G and the set of primes at which two branch points meet.

We can now state in more detail a result due to Beckmann which implies the 'ramification control' of $L_1/K \simeq L_t/K$ (by Exercise 5), where

$$K[X]/(x-t) \simeq L_1 \oplus \cdots \oplus L_r \tag{7}$$

and L_t for $t \in \Sigma_{p,q,r}$ is defined as in Step 2 of the proof of Theorem 1.1. We have the following theorem, which is a special case of Theorem 1.2 of [**Bec91**].

THEOREM 3.2. *(Chevalley-Weil for branched coverings) Assume that $X_K \xrightarrow{G} \mathbb{P}^1_K$ is a good model where $G = \langle \overline{\gamma}_0, \overline{\gamma}_1, \overline{\gamma}_\infty \rangle$ and let $L = L_1$ be as in* (7)*. Then L is ramified only at the places $S = S_{\mathrm{bad}} \cup S_t$ where*

$$S_{\mathrm{bad}} := \{\wp \text{ is a finite prime of } K : \wp | \#G\}$$

and

$$S_t = \{\wp \text{ is a finite prime of } K : I_\wp(t,j) > 0 \text{ for some } j \in \{0,1,\infty\}\}.$$

Moreover, if t meets $j \in \{0,1,\infty\}$ at $\wp$, i.e., $I_\wp(t,j) > 0$ (note that j can at most meet one of those values) then

$$I(\mathfrak{p}/\wp) = \langle \overline{\gamma}_j^{I_\wp(t,j)} \rangle$$

up to conjugation in G where $\mathfrak{p}$ is some prime ideal of L above $\wp$.

The last part of the theorem says basically that the geometric ramification 'controls' the arithmetic ramification.

REMARK 3.3. In general one cannot always guarantee the existence of a good model but nevertheless, Theorem 3.2 remains valid if we add to the set S_{bad} the finite set of primes that prevent the model to be good.

Using the previous theorem one deduces the following proposition.

PROPOSITION 3.1. Let $t \in \Sigma_{p,q,r}$ and $\wp \nmid S_{\mathrm{bad}}$ then L_t/K is unramified at $\wp$.

Proof Since $t \in \Sigma_{p,q,r} \subseteq \mathbb{P}^1(\mathbb{C})\backslash\{0,1,\infty\}$ we have $t = \frac{a^p}{c^r}$ for coprime integers a, c. Moreover $t - 1 = -\frac{b^q}{c^r}$ for b and c coprime. By Lemma 1.1 we have

$$I_\wp(t,0) \equiv 0 \pmod{p},$$
$$I_\wp(t,1) \equiv 0 \pmod{q},$$
$$I_\wp(t,\infty) \equiv 0 \pmod{r}.$$

Using the last part of Theorem 3.2 we deduce that $I(\mathfrak{p}/\wp) = 1$. Therefore L_t is unramified at $\wp$. □

References

[Bec91] S. Beckmann, *On extensions of number fields obtained by specializing branched coverings*, J. Reine Angew. Math. **419** (1991), 27–53. MR 1116916 (93a:11095)

[Bel79] G. V. Belyi, *Galois extensions of a maximal cyclotomic field*, Izv. Akad. Nauk SSSR Ser. Mat. **43** (1979), no. 2, 267–276, 479. MR 534593 (80f:12008)

[Beu04] F. Beukers, *The generalized Fermat equation*, 2004, Lectures held at Institut Henri Poincaré, available at http://www.math.uu.nl/people/beukers/Fermatlectures.pdf.

[Cle65] A. Clebsch, *Über diejenige ebenen Curven, deren Coordinaten rationale Functionen eines Parameters sind*, J. Reine Angew. Math. **64** (1865), 43–65.

[Dar] H. Darmon, *Rational points on curves*, in this volume.

[Dar00] ———, *Rigid local systems, Hilbert modular forms, and Fermat's last theorem*, Duke Math. J. **102** (2000), no. 3, 413–449. MR 1756104 (2001i:11071)

[Dar04] ———, *A fourteenth lecture on Fermat's last theorem*, Number theory, CRM Proceedings & Lecture Notes, vol. 36, American Mathematical Society, 2004, Papers from the 7th Conference of the Canadian Number Theory Association held at the University of Montreal, Montreal, QC, May 19–25, 2002, pp. 103–115. MR 2070918 (2005b:11004)

[DG95] H. Darmon and A. Granville, *On the equations $z^m = F(x,y)$ and $Ax^p + By^q = Cz^r$*, Bull. London Math. Soc. **27** (1995), no. 6, 513–543. MR 1348707 (96e:11042)

[DM97] H. Darmon and L. Merel, *Winding quotients and some variants of Fermat's last theorem*, J. Reine Angew. Math. **490** (1997), 81–100. MR 1468926 (98h:11076)

[GD06] G. González-Diez, *Variations on Belyi's theorem*, Q. J. Math. **57** (2006), no. 3, 339–354. MR 2253591 (2007g:14031)

[Jos06] J. Jost, *Compact Riemann surfaces*, third ed., Universitext, Springer-Verlag, Berlin, 2006, An introduction to contemporary mathematics. MR 2247485 (2007b:32024)

[Köc04] B. Köck, *Belyi's theorem revisited*, Beiträge Algebra Geom. **45** (2004), no. 1, 253–265. MR 2070647 (2005j:14036)

[Mir95] R. Miranda, *Algebraic curves and Riemann surfaces*, Graduate Studies in Mathematics, vol. 5, American Mathematical Society, Providence, RI, 1995. MR 1326604 (96f:14029)

[Mun00] J. R. Munkres, *Topology*, 2nd ed., Prentice Hall, 2000.

[Poi82] H. Poincaré, *Sur les fonctions fuchsiennes*, C.R. Acad. Sci. Paris **92** (1882), 1038–1040.

[Rie57] B. Riemann, *Theorie der Abel'schen Functionen*, J. Reine und Angew. Math. **54** (1857), 115–155.

[Roc65] G. Roch, *Über die Anzahl der willkürlichen Constanten in algebraischen Functionen*, J. Reine Angew. Math. **64** (1865), 372–376.

[Ser56] J.-P. Serre, *Géométrie algébrique et géométrie analytique*, Ann. Inst. Fourier, Grenoble **6** (1955–1956), 1–42. MR 0082175 (18,511a)

[Ser92] ———, *Topics in Galois theory*, Research Notes in Mathematics, vol. 1, Jones and Bartlett Publishers, Boston, MA, 1992, Lecture notes prepared by Henri Damon [Henri Darmon], With a foreword by Darmon and the author. MR 1162313 (94d:12006)

[TW95] R. Taylor and A. Wiles, *Ring-theoretic properties of certain Hecke algebras*, Ann. of Math. (2) **141** (1995), no. 3, 553–572. MR 1333036 (96d:11072)

[Wei56] A. Weil, *The field of definition of a variety*, Amer. J. Math. **78** (1956), 509–524. MR 0082726 (18,601a)

[Wil95] A. Wiles, *Modular elliptic curves and Fermat's last theorem*, Ann. of Math. (2) **141** (1995), no. 3, 443–551. MR 1333035 (96d:11071)

Département de mathématiques et de statistique, Université Laval, 1045 rue de la médecine, Québec (Québec) Canada G1V 0A6

E-mail address: hugo.chapdelaine@mat.ulaval.ca

Clay Mathematics Proceedings
Volume 8, 2009

Merel's theorem on the boundedness of the torsion of elliptic curves

Marusia Rebolledo

Abstract. In this note, we give the key steps of Merel's proof of the Strong Uniform Boundedness Conjecture. This proof relies on three fundamental ingredients: the geometric approach of Mazur and Kamienny, the innovative introduction of the winding quotient by Merel, and the use of Manin's presentation of the homology group of modular curves.

1. Introduction

Interest in elliptic curves dates back at least to Fermat, who introduced his fundamental method of infinite descent to prove his "Last Theorem" in degree 4. Poincaré seems to have been the first to conjecture, around 1901, the now famous theorem of Mordell asserting that the group of rational points of an elliptic curve over $\mathbb{Q}$ is finitely generated. This result was later generalized by Weil to encompass all abelian varieties over number fields. If E is an elliptic curve over a number field K, it is therefore known that

$$E(K) \cong \mathbb{Z}^r \oplus T$$

as abstract groups, where $T = E(K)_{\text{tors}}$ is the finite *torsion subgroup* of $E(K)$. The integer r, called the rank, is a subtle invariant about which little is known and which can be rather hard to compute given E and K. The torsion subgroup, in contrast, is readily computed in specific instances, and this makes it realistic to ask more ambitious questions about the variation of $E(K)_{\text{tors}}$ with E and K. A fundamental result in this direction is the theorem of Mazur presented in Chapter 3 of Darmon's lecture in this volume, which gives a uniform bound on $E(\mathbb{Q})_{\text{tors}}$ as E varies over all elliptic curves over $\mathbb{Q}$. Kamienny [**Kam92**] was able to extend Mazur's result to quadratic fields, obtaining a bound on $E(K)_{\text{tors}}$ for K quadratic that was even independent of K itself. This led him to formulate the Strong Uniform Boundedness Conjecture, asserting that the cardinality of $E(K)_{\text{tors}}$ can be bounded above by a constant which depends only on the degree of $K/\mathbb{Q}$. (The weaker conjecture asserting that the torsion can be bounded uniformly in the field K is presented as being 'a part of the folklore' by Cassels [**Cas66**] (p. 264).) Actually, according to Demjanenko (see [**Dem72**] and entry MR0302654 in Mathematical Reviews) this

2000 *Mathematics Subject Classification.* Primary 11G05.

conjecture was posed in the 70's by Shafarevich; his paper proved a result in this direction. The Strong Uniform Boundedness Conjecture was proved in 1994 by Merel, building on the methods developed by Mazur and Kamienny.

THEOREM 1 (Merel 1994). *For all $d \in \mathbb{Z}, d \geq 1$ there exists a constant $B(d) \geq 0$ such that for all elliptic curves E over a number field K with $[K : \mathbb{Q}] = d$ then*

$$| E(K)_{\text{tors}} | \leq B(d).$$

Merel actually proved the following bound on the prime numbers dividing $E(K)_{\text{tors}}$:

THEOREM 2 (Merel - 1994). *Let E be an elliptic curve over a number field K such that $[K : \mathbb{Q}] = d > 1$. Let p be a prime number. If $E(K)$ has a p-torsion point then $p < d^{3d^2}$.*

It is then sufficient to conclude for the case $d > 1$. Mazur and Kamienny [**KM95**] have indeed shown that, by work of Faltings and Frey, Theorem 2 implies Theorem 1. The case $d = 1$ of Theorem 1 has been proved by Mazur [**Maz77, Maz78**] in 1976 as explained by Henri Darmon in his lecture. Mazur gives more precisely a list of all possibilities for the torsion group over $\mathbb{Q}$. It was actually a conjecture of Levi formulated around 1908. We can mention also that the cases $2 \leq d \leq 8$ and $9 \leq d \leq 14$ have been treated respectively by Kamienny and Mazur (see [**KM95**]), and Abramovich [**Abr95**].

The goal of this note is to give the key steps of the proof of Theorem 2.

REMARK 1. Oesterlé [**Oes**] later improved the bound of Theorem 2 to $(3^{d/2}+1)^2$ but we will focus on Merel's original proof (see Section 3.6 concerning Oesterlé's trick).

REMARK 2. Unfortunately, the reduction of Theorem 1 to Theorem 2 is not effective; this explains why the global bound $B(d)$ is not explicit. However, in 1999, Parent [**Par99**] gave a bound for the p^r-torsion ($r \geq 1, p$ prime) and thus obtained a global effective bound for the torsion (later improved by Oesterlé). This bound is exponential in d. It is conjectured that $B(d)$ can be made polynomial in d.

We will now give the sketch of the proof of Theorem 2. From now on, we will denote by $d \geq 1$, an integer, by p a prime number and write $Z = \mathbb{Z}[1/p]$. Following the traditional approach, Mazur and Kamienny translated the assertion of the theorem into an assertion about rational points of some modular curves.

2. Mazur's method

2.1. To a problem on modular curves. We briefly recall that there exist smooth schemes $X_0(p)$ and $X_1(p)$ over Z which classify, coarsely and finely respectively, the generalized elliptic curves endowed with a subgroup, respectively a point, of order p. We refer for instance to Chapter 3 of [**Dar**] for more details. We denote by $Y_0(p)$ and $Y_1(p)$ the respective affine parts of $X_0(p)$ and $X_1(p)$. We use the subscript $_{\mathbb{Q}}$ for the algebraic curves over $\mathbb{Q}$ obtained by taking the generic fiber of $X_0(p)$ or $X_1(p)$. We will denote by $J_0(p)$ the Néron model over Z of the Jacobian $J_0(p)_{\mathbb{Q}}$ of $X_0(p)_{\mathbb{Q}}$.

Suppose that E is an elliptic curve over a number field K of degree $d \geq 1$ over $\mathbb{Q}$, endowed with a K-rational p-torsion point P. Then (E, P) defines a point

$\tilde{x} \in Y_1(p)(K)$. We can map this point to a point $x \in Y_0(p)(K)$ through the usual covering $X_1(p) \longrightarrow X_0(p)$.

If we denote by $v_1, \ldots, v_d$ the embeddings of K into $\mathbb{C}$, we then obtain a point $\underline{x} = (v_1(x), \ldots, v_d(x)) \in X_0(p)^{(d)}(\mathbb{Q})$. Here we denote by $X_0(p)^{(d)}$ the d-th symmetric power of $X_0(p)$, that is to say the quotient scheme of $X_0(p)$ by the action of the permutation group Σ_d. It is a smooth scheme over Z.

2.2. The Mazur and Kamienny strategy. The strategy is almost the same as in the case $d = 1$ explained in [**Dar**] Ch.3. Let $A_{\mathbb{Q}}$ denote an abelian variety quotient of $J_0(p)_{\mathbb{Q}}$ and A its Néron model over Z. Kamienny's idea is to approach the Uniform Boundedness Conjecture by studying the natural morphism

$$\phi_A^{(d)} : X_0(p)^{(d)} \xrightarrow{\phi^{(d)}} J_0(p) \longrightarrow A$$

defined as follows. Over $\mathbb{Q}$, this morphism is defined as the composition of the Albanese morphism $(Q_1, \ldots, Q_d) \mapsto [(Q_1) + \ldots (Q_d) - d(\infty)]$ with the surjection of $J_0(p)_{\mathbb{Q}}$ to $A_{\mathbb{Q}}$. It then extends to a morphism from the smooth Z-scheme $X_0(p)^{(d)}$ to A. For any prime number $l \neq p$, we denote by $\phi_{A,\mathbb{F}_l}^{(d)} : X_0(p)_{\mathbb{F}_l}^{(d)} \longrightarrow A_{\mathbb{F}_l}$ the morphism obtained by taking the special fibers at l. Just as in the case $d = 1$, we have

THEOREM 3 (Mazur-Kamienny). *Suppose that*

(1) $A(\mathbb{Q})$ *is finite;*
(2) *there exists a prime number* $l > 2$ *such that* $p > (1 + l^{d/2})^2$ *and* $\phi_{A,\mathbb{F}_l}^{(d)}$ *is a formal immersion at* $\infty_{\mathbb{F}_l}^{(d)}$.

Then $Y_1(p)(K)$ *is empty for all number fields* K *of degree* d *over* $\mathbb{Q}$*, i.e., there does not exist any elliptic curve with a point of order* p *over any number field of degree* d.

PROOF. The proof of this theorem is analogous to the one in the case $d = 1$. The principal ingredients of the proof are explained in [**Dar**] Ch. 3. For a complete proof, the reader can see [**Maz78**], [**Kam92**] or, for a summary, [**Edi95**]. The idea is the following: suppose that there exists a number field K of degree d and a point of $Y_1(p)(K)$ and consider the point $\underline{x} \in X_0(p)^{(d)}(\mathbb{Q})$ obtained as explained in Section 2.1. The condition $p > (1 + l^{d/2})^2$ of Theorem 3 implies that the section s of $X_0(p)^{(d)}$ corresponding to $\underline{x}$ crosses $\infty^{(d)}$ in the fiber at l. Since $s \neq \infty^{(d)}$, the fact that $\phi_{A,\mathbb{F}_l}^{(d)}$ is a formal immersion at $\infty_{\mathbb{F}_l}^{(d)}$ and Condition 1 will then give a contradiction.

□

We now need an abelian variety $A_{\mathbb{Q}}$ quotient of $J_0(p)_{\mathbb{Q}}$ of rank 0 (see section 3.1) and a formal immersion criterion (see below).

2.3. Criterion of formal immersion. Recall first that a morphism $\phi : X \longrightarrow Y$ of noetherian schemes is a *formal immersion* at a point $x \in X$ which maps to $y \in Y$ if the induced morphism on the formal completed local rings $\hat{\phi} : \widehat{\mathcal{O}_{Y,y}} \longrightarrow \widehat{\mathcal{O}_{X,x}}$ is surjective. Equivalently, it follows from Nakayama's lemma that ϕ is a formal immersion at x if the two following conditions hold:

(1) the morphism induced on the residue fields $k(y) \longrightarrow k(x)$ is an isomorphism;

(2) the morphism induced on the cotangent spaces $\phi^* : \mathrm{Cot}_y(Y) \longrightarrow \mathrm{Cot}_x(X)$ is surjective.

The first condition is verified in our situation, so we are now looking for a criterion to have

$$\phi^{(d)}_{A,\mathbb{F}_l}{}^* : \mathrm{Cot}(A_{\mathbb{F}_l}) \longrightarrow \mathrm{Cot}_{\infty^{(d)}_{\mathbb{F}_l}}(X_0(p)^{(d)}_{\mathbb{F}_l})$$

surjective. For this, we will look in more detail at $\phi^{(d)*}_A$.

Let R be a Z-algebra. As in [**Dar**], denote by $S_2(\Gamma_0(p), R)$ the regular differentials on $X_0(p)_R = X_0(p) \times_Z R$. For $R = \mathbb{C}$, we obtain the vector space of classical modular forms $S_2(\Gamma_0(p), \mathbb{C})$. The *q-expansion principle* gives an injective morphism of R-modules

$$S_2(\Gamma_0(p), R) \hookrightarrow R[[q]].$$

Furthermore, we have an isomorphism between $\mathrm{Cot}(J_0(p)(\mathbb{C}))$ and $S_2(\Gamma_0(p), \mathbb{C})$ coming from the composition of

(1) the isomorphism $H^0(J_0(p)(\mathbb{C}), \Omega^1) \longrightarrow \mathrm{Cot}(J_0(p)(\mathbb{C}))$ which maps a differential form to its evaluation at 0 ;
(2) the isomorphism $H^0(J_0(p)(\mathbb{C}), \Omega^1) \xrightarrow{\phi^*} H^0(X_0(p)(\mathbb{C}), \Omega^1) = S_2(\Gamma_0(p), \mathbb{C})$ given by Serre duality.

It is a nontrivial fact that this isomorphism $\mathrm{Cot}(J_0(p)(\mathbb{C})) \cong S_2(\Gamma_0(p), \mathbb{C})$ extends to an isomorphism over Z (and actually even over $\mathbb{Z}$). Indeed, Grothendieck duality can be applied in this setting instead of Serre duality and we then obtain an isomorphism: $\mathrm{Cot}(J_0(p)) \cong S_2(\Gamma_0(p), Z)$ (see [**Maz78**] 2 e)).

Our next task is to analyze the cotangent bundle $\mathrm{Cot}_{\infty^{(d)}}(X_0(p)^{(d)})$. Recall that q is a formal local parameter of $X_0(p)$ at ∞, i.e., $\widehat{\mathcal{O}}_{X_0(p),\infty} \cong Z[[q]]$. We then have

$$\widehat{\mathcal{O}}_{X_0(p)^{(d)},(\infty)^{(d)}} \cong Z[[q_1, \ldots, q_d]]^{\Sigma_d} = Z[[\sigma_1, \ldots, \sigma_d]]$$

where for $i = 1, \ldots, d$, q_i is a local parameter at ∞ on the ith factor of $X_0(p)^d$ and $\sigma_1 = q_1 + \cdots + q_d, \ldots, \sigma_d = q_1 \cdots q_d$ are the symmetric functions in $q_1, \ldots, q_d$. Consequently, $\mathrm{Cot}_{\infty^{(d)}}(X_0(p)^{(d)})$ is a free Z-module of rank d with a basis given by the differential forms $(d\sigma_1, \ldots, d\sigma_d)$.

We obtain the following diagram:

$$\begin{array}{ccccc} \mathrm{Cot}(J_0(p)) & \xrightarrow[\sim]{\phi^*} & S_2(\Gamma_0(p), Z) & \xhookrightarrow{q\text{-exp}} & Z[[q]] \\ \Big\downarrow{\scriptstyle \phi^{(d)*}} & & & & \\ \mathrm{Cot}(X_0(p)^{(d)}) & & & & \end{array}$$

Lemma 1. *Let $\omega \in \mathrm{Cot}(J_0(p))$ be such that $\phi^*(\omega)$ has a q-expansion equal to $\sum_{m\geq 1} a_m q^m \frac{dq}{q}$. Then we have*

$$\phi^{(d)*}(\omega) = a_1 d\sigma_1 - a_2 d\sigma_2 + \cdots + (-1)^{d-1} a_d d\sigma_d.$$

Proof. Denote by $\pi : X_0(p)^d \longrightarrow X_0(p)^{(d)}$ the canonical map. We have

$$\pi^* \phi^{(d)*}(\omega) = \sum_{i=1}^{d} \sum_{m\geq 1} a_m q_i^m \frac{dq_i}{q_i} = \sum_{m\geq 1} a_m m^{-1} ds_m$$

where $s_m = \sum_{i=1}^{d} q_i^m$. Then Newton's formula

$$s_m - \sigma_1 s_{m-1} + \cdots + (-1)^m m\sigma_m = 0$$

gives $m^{-1} ds_m = (-1)^m d\sigma_m$ for $m \in \{1, \ldots, d\}$. □

We suppose in the sequel that $A_{\mathbb{Q}}$ is the quotient of $J_0(p)_{\mathbb{Q}}$ by an ideal I of the *Hecke algebra* $\mathbb{T} \subset \mathrm{End}(J_0(p)_{\mathbb{Q}})$, so that there is an induced action of $\mathbb{T}$ on A. The exact sequence

$$0 \to IJ_0(p)_{\mathbb{Q}} \to J_0(p)_{\mathbb{Q}} \to A_{\mathbb{Q}} \to 0$$

induces a reverse exact sequence for the cotangent bundles after scalar extension by $Z[1/2]$

$$0 \to \mathrm{Cot}(A_{Z[1/2]}) \to \mathrm{Cot}(J_0(p)_{Z[1/2]}) \to \mathrm{Cot}(J_0(p)_{Z[1/2]})[I] \to 0$$

where we denote by $\mathrm{Cot}(J_0(p)_{Z[1/2]})[I]$ the differential forms annihilated by I. This is due to a *specialization lemma* of Raynaud (see [**Maz78**] Proposition 1.1 and Corollary 1.1).

Let $l \neq 2, p$ be a prime number. We finally have the following diagram in characteristic l:

$$\begin{array}{ccccccc} \mathrm{Cot}(A_{\mathbb{F}_l}) & \hookrightarrow & \mathrm{Cot}(J_0(p)_{\mathbb{F}_l}) & \xrightarrow[\sim]{\phi_{\mathbb{F}_l}^*} & S_2(\Gamma_0(p), \mathbb{F}_l) & \xhookrightarrow{q\text{-exp}} & \mathbb{F}_l[[q]] \\ & \searrow^{\phi_{A,\mathbb{F}_l}^{(d)\,*}} & \downarrow{\scriptstyle \phi_{\mathbb{F}_l}^{(d)\,*}} & & & & \\ & & \mathrm{Cot}_{\infty_{\mathbb{F}_l}^{(d)}}(X_0(p)_{\mathbb{F}_l}^{(d)}) & & & & \end{array}$$

This diagram and Lemma 1 give a criterion for $\phi_{A,\mathbb{F}_l}^{(d)}$ to be a formal immersion at $\infty_{\mathbb{F}_l}^{(d)}$ (see Theorem 5 below). Historically, Mazur first showed the following result which completes the proof of Mazur's theorem sketched in Section 4 of [**Dar**] using for $A_{\mathbb{Q}}$ the *Eisenstein quotient*.

THEOREM 4. *The morphism $\phi_{A,\mathbb{F}_l}$ is a formal immersion at $\infty_{\mathbb{F}_l}$ for all prime numbers $l \neq 2, p$.*

PROOF. There is a nonzero $\omega \in \mathrm{Cot}(A_{\mathbb{F}_l})$ such that $\phi_{\mathbb{F}_l}^*(\omega) \in S_2(\Gamma_0(p), \mathbb{F}_l)$ is an eigenform (under the action of the Hecke algebra $\mathbb{T}$). Then by the q-expansion principle and the injectivities in the above diagram, its q-expansion is not identically zero (because if it were, $\phi_{\mathbb{F}_l}^*(\omega)$ itself would be zero). We deduce that $a_1(\omega) \neq 0$: indeed, if it were, since ω is an eigenform, we should have $a_m(\omega) = a_1(T_m\omega) = \lambda_m(\omega)a_1(\omega) = 0$ for all $m \geq 1$, so $\omega = 0$, which is impossible. It follows that $a_1(\omega)$ spans $\mathrm{Cot}_{\infty_{\mathbb{F}_l}}(X_0(p)_{\mathbb{F}_l}) \cong \mathbb{F}_l$ and, by Lemma 1, that $\phi_{A,\mathbb{F}_l}$ is a formal immersion at $\infty_{\mathbb{F}_l}$. □

THEOREM 5 (Kamienny). *The following assertions are equivalent:*

1. *$\phi_{A,\mathbb{F}_l}^{(d)}$ is a formal immersion at $\infty_{\mathbb{F}_l}^{(d)}$;*
2. *there exist d weight-two cusp forms $f_1, \ldots, f_d$ annihilated by I such that the vectors $(a_1(f_i), \ldots, a_d(f_i))_{i=1,\ldots,d}$ are linearly independent mod l;*
3. *the images of $T_1, \ldots, T_d$ in $\mathbb{T}/(l\mathbb{T} + I)$ are $\mathbb{F}_l$-linearly independent.*

PROOF. The equivalence of (1) and (2) follows directly from Lemma 1 since $\mathrm{Cot}(A)$ maps to the forms annihilated by I via the isomorphism ϕ^*. Condition (3) is dual to Condition (2) Indeed, the multiplicity one theorem implies that the pairing

$$\begin{array}{rccc} \langle\,,\,\rangle\ : & S_2(\Gamma_0(p),\mathbb{Z})\times\mathbb{T} & \longrightarrow & \mathbb{Z} \\ & (f,t) & \longmapsto & a_1(tf) \end{array}$$

is perfect and then induces an isomorphism of $\mathbb{T}$-modules between $S_2(\Gamma_0(p),\mathbb{Z})$ and the $\mathbb{Z}$-dual of $\mathbb{T}$. For a more detailed proof of this theorem, see [**Kam92**] or [**Oes**] Sections 3, 4 and 6. □

3. Merel's proof

3.1. The Winding Quotient. Denote by $J_{e,\mathbb{Q}}$ the *winding quotient* (see [**Dar**] Ch. 3) and J_e its Néron model over Z. We just recall that $J_{e,\mathbb{Q}}$ is the abelian variety quotient of $J_0(p)_\mathbb{Q}$ by the *winding ideal* I_e of $\mathbb{T}$.

Considering Theorem 3, we are now looking for a quotient $A_\mathbb{Q}$ of $J_0(p)_\mathbb{Q}$ by an ideal $I\subset\mathbb{T}$ such that $A(\mathbb{Q})$ is finite. Mazur and Kamienny have used the *Eisenstein quotient*, which has this property (see [**Maz77, Kam92**]). Merel's fundamental innovation was to use the winding quotient; this quotient is larger and easier to exploit than the Eisenstein quotient. This was made possible after the works of Kolyvagin on the Birch and Swinnerton-Dyer conjecture; indeed, it then turned out that $J_e(\mathbb{Q})$ is finite by construction (see [**Mer96**] or [**Dar**] for a summary). Actually, the Birch and Swinnerton-Dyer conjecture predicts that the winding quotient is the largest quotient of $J_0(p)_\mathbb{Q}$ of rank zero.

Finally, to prove Theorem 2, thanks to Theorems 3 and 5, it suffices to determine for which prime numbers p the following is true for a prime number $l\neq 2$ such that $p>(1+l^{d/2})^2$:

($\star_l$) the images of $T_1,\dots,T_d$ in $\mathbb{T}/(l\mathbb{T}+I_e)$ are $\mathbb{F}_l$-linearly independent.

3.2. Merel's strategy. Suppose now that $d\geq 3$. Recall that the Hecke algebra $\mathbb{T}\subset\mathrm{End}(J_0(p))$ also acts on the first group of absolute singular homology $H_1(X;\mathbb{Z})$ of the compact Riemann surface $X=X_0(p)(\mathbb{C})$ and that I_e is the annihilator of the *winding element* $e\in H_1(X;\mathbb{Q})$ (see the article of Darmon in this volume). Then $\mathbb{T}\cdot e$ is a free $\mathbb{T}/I_e$-module of rank 1. It follows that ($\star_l$) is equivalent to

($\star_l$) the images of $T_1e,\dots,T_de$ in $\mathbb{T}e/l\mathbb{T}e$ are $\mathbb{F}_l$-linearly independent.

As before, the characteristic zero analogous condition

($\star$) $T_1e,\dots,T_de$ are Z-linearly independent in $\mathbb{T}\cdot e$.

is equivalent to $\phi_{I_e}^{(d)}$ being a formal immersion at $\infty_\mathbb{Q}^{(d)}$. If ($\star_l$) is true for a prime number l then ($\star$) is true, while the condition ($\star$) implies ($\star_l$) for almost all prime numbers l. Kamienny showed that if ($\star$) is true then there exists a prime number $l<2(d!)^{5/2}$ (depending on p) such that ($\star_l$) is true (see [**Kam92**] Corollary 3.4 and [**Edi95**] 4.3 for the precise bound). The heart of Merel's proof for the boundedness of the torsion of elliptic curves is then to prove ($\star$) for $p>d^{3d^2}>2^{d+1}(d!)^{5d/2}\geq(1+(2(d!)^{5/2})^{d/2})^2$.

We will now explain the key steps of this proof omitting the details of the calculations. For a completed proof, we will refer to [**Mer96**].

Consider a fixed prime number $p > d^{3d^2}$ for $d \geq 3$ an integer. To prove that $e, T_2 e, \ldots, T_d e$ are linearly independent, it suffices to prove that so are $e, t_2 e, \ldots, t_d e$ where $t_r = T_r - \sigma'(r)$ with $\sigma'(r)$ the sum of divisors of r coprime to p. These slightly different Hecke operators t_r are more pleasant to work with because they annihilate the "Eisenstein part" of e and we can then work as if e were equal to the *modular symbol* $\{0, \infty\}$ (see section 3.3 for a definition)[1].

The idea of the proof is to use the intersection product

$$\bullet : H_1(X; \mathbb{Z}) \times H_1(X; \mathbb{Z}) \longrightarrow \mathbb{Z}.$$

Suppose indeed that $\lambda_1 e + \lambda_1 t_2 e + \cdots + \lambda_c t_c e = 0$ for $1 \leq c \leq d$ and some $\lambda_1, \ldots, \lambda_c$ in $\mathbb{Z}$ with $\lambda_c \neq 0$. The strategy is then to find $x_c \in H_1(X; \mathbb{Z})$ such that

$$i)\ t_c e \bullet x_c \neq 0 \qquad \text{and} \qquad ii)\ t_r e \bullet x_c = 0 \quad (1 \leq r \leq c-1).$$

This will give a contradiction.[2]

Two key facts make it possible to follow this strategy: first, there is a presentation of $H_1(X; \mathbb{Z})$ by generators and relations due to Manin [**Man72**] (see the section 3.3); secondly, a lemma called *lemme des cordes* by Merel (Proposition 1 below) enables us to compute the intersection product of two such generators. It suffices then to express $t_r e$ in terms of Manin's generators (see 3.4).

3.3. Manin's symbols. Denote by $\mathfrak{H}$ the Poincaré upper half-plane. For $\alpha, \beta \in \mathbb{P}^1(\mathbb{Q})$, consider the image in $\Gamma_0(p)\backslash\mathfrak{H}$ of the geodesic path from α to β in $\mathfrak{H}$. Denote by $\{\alpha, \beta\}$ its homology class in the homology group $H_1(X, cusps; \mathbb{Z})$ relative to the set *cusps* of the cusps of X.

EXERCISE 1. (1) *Show that* $\{\alpha, \beta\}$ *is the sum of classes of type* $\{b/d, a/c\}$ *with* $a, b, c, d \in \mathbb{Z}$ *such that* $ad - bc = 1$ *(hint: use continued fractions).*
(2) *Show that* $\{b/d, a/c\}$ *depends only on the coset* $\Gamma_0(p) \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$.

For a solution of this exercise, see [**Man72**] for instance.

The preceding results imply that there is a surjective map

$$\begin{array}{rccl} \xi\ : & \mathbb{Z}[\Gamma_0(p)\backslash \mathrm{SL}_2(\mathbb{Z})] & \longrightarrow & H_1(X, cusps; \mathbb{Z}) \\ & \Gamma_0(p) \cdot g & \longmapsto & \{g \cdot 0, g \cdot \infty\} = \{\frac{b}{d}, \frac{a}{c}\} \quad g = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{Z}). \end{array}$$

Since there is moreover an isomorphism

$$\begin{array}{ccc} \Gamma_0(p)\backslash \mathrm{SL}_2(\mathbb{Z}) & \longrightarrow & \mathbb{P}^1(\mathbb{F}_p) \\ \Gamma_0(p) \cdot \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) & \mapsto & [c : d], \end{array}$$

we will simply write $\xi(c/d) := \xi(\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right))$.

For $k \in \mathbb{F}_p^\times$ we obtain $\xi(k) = \{0, 1/k\}$ which is an element of $H_1(X; \mathbb{Z})$ (seen as a submodule of $H_1(X, cusps; \mathbb{Z})$) because 0 and $1/k$ are conjugate modulo $\Gamma_0(p)$. These elements are generators of $H_1(X; \mathbb{Z})$. The other generators of $H_1(X, cusps; \mathbb{Z})$ are $\xi(0)$ and $\xi(\infty)$ and they verify $\xi(0) = -\xi(\infty) = \{0, \infty\}$.

The following proposition, called *lemme des cordes* by Merel, gives a method to compute the intersection product of two Manin symbols in the absolute homology group. For $k \in \{1, \ldots, p-1\}$, denote by k_* the element of $\{1, \ldots, p-1\}$ such that $kk_* \equiv -1 \pmod{p}$.

[1] In the relative homology group, the winding element e differs from $\{0, \infty\}$ by an element which is an eigenvector for all T_n with system of eigenvalues $\{\sigma'(n)\}_{n \geq 1}$ (up to a constant): this is what I called the *Eisenstein part*.

[2] Actually, for $c = 1$ the situation will be slightly different because of the Eisenstein part of e.

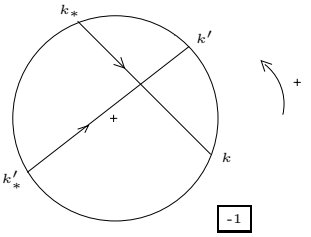


FIGURE 1. Lemme des cordes. Here $\xi(k) \bullet \xi(k') = -1$.

PROPOSITION 1 (Merel). *Let $k, k' \in \{1, \dots, p-1\}$. Denote by C_k the chord of the unit circle from $e^{2i\pi k_*/p}$ to $e^{2i\pi k/p}$ and similarly for k'. Then*

$$\xi(k) \bullet \xi(k') = C'_k \wedge C_k$$

where $C_{k'} \wedge C_k$ is the number of intersections of $C_{k'}$ by C_k (equal to $1, 0$ or -1 according to the trigonometric orientation of the unit circle).

PROOF. See [**Mer96**] Lemma 4. □

3.4. Two useful formulas. Because of their technical aspect, we will not reproduce the proofs of the following formulas which appear in Lemmas 2 and 3 of [**Mer96**].

We have first a formula for $t_r e$ $(r > 1)$ in terms of the Manin symbols $\xi(k)$:

PROPOSITION 2 (Merel). *Let $r < p$ be a positive integer. Then*

$$t_r e = - \sum_{\left(\begin{smallmatrix} u & v \\ w & t \end{smallmatrix}\right) \in X_r} \xi(w/t)$$

where X_r is the set of matrices $\left(\begin{smallmatrix} u & v \\ w & t \end{smallmatrix}\right)$ of determinant r such that $0 < w < t$ and $u > v \geq 0$.

For $r = 1$, we can compute directly the intersection of e with a Manin generator:

PROPOSITION 3 (Merel). *For any $k \in \{1, \dots, p-1\}$ we have*

$$(p-1)e \bullet \xi(k) = \frac{k_* - k}{p}(p-1) - 12S(k,p),$$

where $S(k,p) = \sum_{h=0}^{p-1} \bar{B}_1(\frac{h}{p})\bar{B}_1(\frac{kh}{p})$ is the Dedekind sum *and $\bar{B}_1$ the first Bernoulli polynomial made 1-periodic.*

REMARK 3. Note that in Proposition 2 the $\xi(0)$ and $\xi(\infty)$ terms vanish. This is not surprising since $t_r e$ lies in the absolute homology group.

3.5. Conclusion of the proof. We will now explain how Merel put all the previous ingredients together to obtain the proof of $(\star)$ for p large enough.

Suppose that there are integers $\lambda_1, \dots, \lambda_d$ such that

$$\lambda_1 e + \lambda_2 t_2 e + \cdots + \lambda_d t_d e = 0.$$

We will show successively that $\lambda_i = 0$ for all $i \in \{1, \dots, d\}$, treating the case of λ_1 independently.

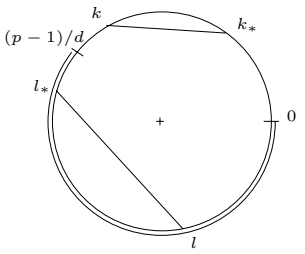


FIGURE 2. Case $i = 1$.

Case $i = 1$. We look for x_1 of the form $x_1 = \xi(k)$ for some k such that

$$i)\ e \bullet \xi(k) \neq 0 \qquad \text{and} \qquad ii)\ t_r e \bullet \xi(k) = 0 \quad (1 < r \leq d).$$

Suppose that $p > d$. By Proposition 2, the condition $ii)$ is equivalent to

$$\sum_{\left(\begin{smallmatrix} u & v \\ w & t \end{smallmatrix}\right) \in X_r} \xi(w/t) \bullet \xi(k) = 0 \qquad (1 \leq r \leq d).$$

It suffices to find k such that $\xi(w/t) \bullet \xi(k) = 0$ for all $\left(\begin{smallmatrix} u & v \\ w & t \end{smallmatrix}\right) \in X_r$. That is what Merel does. Let $l \in \{1, \ldots, p-1\}$ such that $l \equiv wt^{-1} \pmod{p}$ for some $\left(\begin{smallmatrix} u & v \\ w & t \end{smallmatrix}\right) \in X_r$. Then $l_* \equiv -tw^{-1} \pmod{p}$. By Remark 3, we can suppose that neither t nor w are divisible by p.

EXERCISE 2. *Show that l and l_* are larger than $\frac{p-1}{d}$.*

Applying the *lemme des cordes* it suffices to find k such that both the complex numbers $e^{2i\pi k/p}$ and $e^{2i\pi k_*/p}$ are in a portion of the circle where $e^{2i\pi l/p}$ cannot be, so for instance, by the exercise, such that both k and k_* lie in $[0, \frac{p-1}{d}[$. Merel uses then the following analytic lemma ([**Mer96**] Lemma 5) to ensure that, provided $p > d^{3d^2}$ and $k \in \mathbb{Z} \cap]\frac{p}{10d}, \frac{p}{5d} + 1[$ then $k_* \in \mathbb{Z} \cap]\frac{p}{2d} - 1 - \frac{1}{d}, \frac{p-1}{d}[$. (More precisely, this is already true when $p/\log^4(p) > d^4$.)

LEMMA 2. *Let p be a prime number and $a, b \geq 1$ two real numbers. Let $A, B \subset \{1, \ldots, p-1\}$ be two intervals of cardinalities p/a and p/b respectively. If $p > a^2 b^2 \log^4(p)$ then there exists $k \in A$ such that $k_* \in B$.*

We deduce from the following exercise that condition $i)$ above is also verified assuming that $p > d^{3d^2}$.

EXERCISE 3. *Using the Dedekind's reciprocity formula*

$$12(S(k,p) + S(p,k)) = -3 + \frac{p}{k} + \frac{k}{p} + \frac{1}{pk}$$

and the inequality $|12S(p,k)| \leq k$, show that

$$e \bullet \xi(k) \geq \frac{p}{10d} - 10d - 2$$

for all k as before.

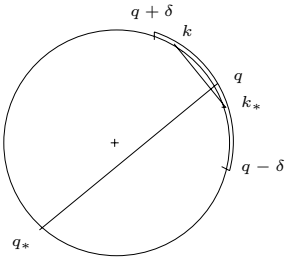


FIGURE 3. Case $i > 1$.

Case $i > 1$. Suppose now that

$$\lambda_2 t_2 e + \cdots + \lambda_c t_c e = 0$$

for some $c \leq d$. The method is almost the same as before: we look for $x_c = \xi(k)$ such that

$$i)\ t_c e \bullet \xi(k) \neq 0 \qquad \text{and} \qquad ii)\ t_r e \bullet \xi(k) = 0 \quad (2 \leq r < c).$$

We remark that in the formulas for $t_r e, r = 2, \ldots, c$, of Proposition 2, the Manin symbol $\xi(1/c)$ occurs only in $t_c e$ and not in $t_r e$ for $r < c$. So we will look for k such that $\xi(1/c) \bullet \xi(k) = \pm 1$ and $\xi(w/t) \bullet \xi(k) = 0$ for all $\left(\begin{smallmatrix} u & v \\ w & t \end{smallmatrix}\right) \in X_r$ $(r \leq c)$ such that $w/t \neq 1/c$.

Let q and l in $\{1, \ldots, p-1\}$ such that $q \equiv 1/c \pmod{p}$ and $l \equiv w/t \neq 1/c \pmod{p}$ for some $\left(\begin{smallmatrix} u & v \\ w & t \end{smallmatrix}\right) \in X_r$ $(r \leq c)$.

EXERCISE 4. *Show that* $|l - q| \geq \delta$, *where* $\delta = \frac{p-d^2}{d(d-1)}$.

By the same analytic lemma as before, it is possible to find $k \in]q, q+\delta]$ such that $k_* \in [q-\delta, q[$ and $q_* \notin [q-\delta, q+\delta]$ when p is large enough, more precisely when $p/\log^4(p) > \mathrm{Sup}(d^8, 400d^4)$. By the lemme des cordes, this then forces λ_c to be zero.

This finishes the proof of Theorem 2.

3.6. Oesterlé's variant. As we said in Remark 1, Oesterlé improved Merel's bound for the torsion of elliptic curves. For this, Oesterlé proved directly the formal immersion in positive characteristic:

PROPOSITION 4. *Suppose that* $p/\log^4 p \geq (2d)^6$. *Then for all* $l \geq 3$, *the condition* $(\star_l)$ *is true, that is to say* $\phi^{(d)}_{A,\mathbb{F}_l}$ *is a formal immersion at* $\infty^{(d)}_{\mathbb{F}_l}$.

For $d \geq 33$, Theorem 2 with the bound $(3^{d/2}+1)^2$ then follows directly from Theorem 4, since $p > (3^{d/2}+1)^2$ implies $p/\log^4 p \geq (2d)^6$ in that case. Oesterlé studied the cases $d < 37$ by computations.

Let us give a sketch of proof of Proposition 4. Let T'_s be defined by $T_r = \sum_{s|r} T'_s$ for all $r \geq 1$ and, instead of $t_r = T_r - \sigma'(r)$ $(r \geq 1)$, consider the following generators of the Eisenstein ideal I:

$$I_1 = n_p \quad \text{and} \quad I_r = \begin{cases} T'_r - r & \text{if } p \nmid r \\ T'_r & \text{if } p | r \end{cases} (r \geq 2),$$

where we denote by n_p the numerator of $(p-1)/12$. We have $t_r = \sum_{s|r, s\neq 1} I_s$ for all $r > 1$.

PROPOSITION 5. *If the images of $I_2 e, \ldots, I_{2d} e$ in Ie/lIe are $\mathbb{F}_l$-linearly independent, then $T_1 e, \ldots, T_d e$ are $\mathbb{F}_l$-linearly independent in $\mathbb{T}e/l\mathbb{T}e$; that is to say $(\star_l)$ is true.*

PROOF. We have

$$T'_2 T'_r = \begin{cases} I_{2r} - 2I_r & \text{if } r \text{ is odd} \\ I_{2r} - 3I_r + 2I_{r/2} & \text{if } r \text{ is even.} \end{cases}$$

So if $I_2 e, \ldots, I_{2r} e$ are linearly independent in Ie/lIe, so are $T'_2 e, \ldots, T'_2 T'_{2r} e$ and, since $T'_2 e = (T_2 - 3)e \in Ie$, we obtain that $T'_1 e, \ldots, T'_d e$ are linearly independent in $\mathbb{T}e/l\mathbb{T}e$. But $T_r = T'_r + \sum_{s|r, s<r} T'_s$ so $T_1 e, \ldots, T_d e$ are linearly independent in $\mathbb{T}e/l\mathbb{T}e$. □

Moreover, Oesterlé used Proposition 2 and the *lemme des cordes* to give an explicit formula for $t_r e \bullet \xi(k)$ and then for $I_r e \bullet \xi(k)$ (which is the unique "r-th term" of $t_r e \bullet \xi(k)$):

$$(1) \qquad I_r e \bullet \xi(k) = \left[\frac{rk}{p}\right] - \left[\frac{rk_*}{p}\right] + v_r(k) - v_r(k_*) \quad (r \geq 2, k \in \{1, \ldots, p-1\}),$$

where $v_r(k) = \#\{(a, a', b, b') \in \mathbb{Z}, a, a', b, b' \geq 1, aa' + bb' = r, (a, b) = 1, bk \equiv a \pmod{p}\}$. The end of the proof is then *mutatis mutandis* the same as Merel's: using Lemma 2, Oesterlé showed that, when $p/\log^4(p) > d^6$, it is possible for each $r \geq 2$ to find k such that $I_r e \bullet \xi(k) = 1$ and $I_s e \bullet \xi(k) = 0$ for $s < r$. He deduced that for $p/\log^4(p) > d^6$, $I_2 e, \ldots, I_d e$ are linearly independent. Applying this for $2d$ instead of d and using Proposition 5 gives Proposition 4.

This is how one can obtain Oesterlé's bound. As we said in Remark 2, the question of finding a bound growing polynomially in d remains open.

REMARK 4. As Merel pointed out to me, the result of Proposition 5 is still true replacing I_r by t_r, $(2 \leq r \leq 2d)$. Indeed, a calculation proves that $t_2 T_i \in t_{2i} + \sum_{1 \leq j \leq i} \mathbb{Z} T_j$. Using the results of the section 3.5 case $i > 1$, it follows that when $p/\log^4(p) > \mathrm{Sup}(d^8, 400d^4)$, $(\star_l)$ is true for all $l \geq 3$. Since $p > (3^{d/2} + 1)^2$ implies $p/\log^4(p) > \mathrm{Sup}(d^8, 400d^4)$ provided that $d \geq 37$, it gives Oesterlé's bound in that case. The other cases have been studied by Oesterlé.

References

[Abr95] D. Abramovich, *Formal finiteness and the torsion conjecture on elliptic curves. A footnote to a paper: "Rational torsion of prime order in elliptic curves over number fields" [Astérisque No. 228 (1995), 3, 81–100; MR1330929 (96c:11058)] by S. Kamienny and B. Mazur*, Astérisque (1995), no. 228, 3, 5–17, Columbia University Number Theory Seminar (New York, 1992). MR 1330925 (96c:11059)

[Cas66] J. W. S. Cassels, *Diophantine equations with special reference to elliptic curves*, J. London Math. Soc. **41** (1966), 193–291. MR 0199150 (33 #7299)

[Dar] H. Darmon, *Rational points on curves*, in this volume.

[Dem72] V. A. Dem′janenko, *The boundedness of the torsion of elliptic curves*, Mat. Zametki **12** (1972), 53–58. MR 0447260 (56 #5575)

[Edi95] B. Edixhoven, *Rational torsion points on elliptic curves over number fields (after Kamienny and Mazur)*, Astérisque (1995), no. 227, Exp. No. 782, 4, 209–227, Séminaire Bourbaki, Vol. 1993/94. MR 1321648 (96c:11056)

[Kam92] S. Kamienny, *Torsion points on elliptic curves over fields of higher degree*, Internat. Math. Res. Notices (1992), no. 6, 129–133. MR 1167117 (93e:11072)

[KM95] S. Kamienny and B. Mazur, *Rational torsion of prime order in elliptic curves over number fields*, Astérisque (1995), no. 228, 3, 81–100, With an appendix by A. Granville, Columbia University Number Theory Seminar (New York, 1992). MR 1330929 (96c:11058)

[Man72] Ju. I. Manin, *Parabolic points and zeta functions of modular curves*, Izv. Akad. Nauk SSSR Ser. Mat. **36** (1972), 19–66. MR 0314846 (47 #3396)

[Maz77] B. Mazur, *Modular curves and the Eisenstein ideal*, Inst. Hautes Études Sci. Publ. Math. (1977), no. 47, 33–186 (1978). MR 488287 (80c:14015)

[Maz78] ———, *Rational isogenies of prime degree (with an appendix by D. Goldfeld)*, Invent. Math. **44** (1978), no. 2, 129–162. MR 482230 (80h:14022)

[Mer96] L. Merel, *Bornes pour la torsion des courbes elliptiques sur les corps de nombres*, Invent. Math. **124** (1996), no. 1-3, 437–449. MR 1369424 (96i:11057)

[Oes] J. Oesterlé, *Torsion des courbes elliptiques sur les corps de nombres*, unpublished.

[Par99] P. Parent, *Bornes effectives pour la torsion des courbes elliptiques sur les corps de nombres*, J. Reine Angew. Math. **506** (1999), 85–116. MR 1665681 (99k:11080)

Laboratoire de mathématiques, Université Blaise Pascal Clermont-Ferrand 2, Campus universitaire des Cézeaux, 63177 Aubière FRANCE

E-mail address: Marusia.Rebolledo@math.univ-bpclermont.fr

Clay Mathematics Proceedings
Volume 8, 2009

Generalized Fermat equations (d'après Halberstadt-Kraus)

Pierre Charollois

Abstract. In this paper, we summarize the work of Halberstadt and Kraus on generalized Fermat equations of the shape $ax^n + by^n = cz^n$. In particular, we sketch the proof that, for fixed odd coprime integer coefficients a, b, c, there is a set of primes n of positive density for which only trivial solutions (x, y, z) occur.

Contents

1. Introduction

Our purpose is to publicize the statement and the proof of the following theorem [**HK02**, théorème 2.1]:

Theorem 1.1 (Halberstadt-Kraus (2002)). *Let a, b, c be odd pairwise coprime integers. Then there is a set of primes $\mathcal{P} = \mathcal{P}(a, b, c)$ of positive density such that if $p \in \mathcal{P}$, then the equation*

$$ax^p + by^p + cz^p = 0 \tag{1}$$

has only trivial rational solutions $(x, y, z) \in \mathbb{Q}^3$.

A solution (x, y, z) is called trivial in our context if $xyz = 0$.

One must point out that before Wiles's work, even the case $a = b = c = 1$ was unknown. Theorem 1.1 exhibits the first infinite family of generalized Fermat equations having only trivial solutions.

Note that the set of primes $\mathcal{P}$ will be given by congruence conditions. These can be made more precise and explicit for particular choices of triples (a, b, c). For instance, the

2000 *Mathematics Subject Classification.* Primary 11D41, Secondary 11F11, 11G05.

proof of Theorem 1.1 yields the following, providing a partial answer to a question raised by Serre [**Ser87**, p.204]:

THEOREM 1.2. *If $p \geq 7$ is a prime number satisfying $p \not\equiv 1 \mod 12$, the equation*

$$x^p + 3y^p + 5z^p = 0$$

has only trivial solutions over $\mathbb{Q}$. So does the equation

$$x^p + y^p + 15z^p = 0.$$

The proof of these theorems relies crucially on the modularity theorem for elliptic curves from Wiles and his followers, as well as Ribet's level-lowering theorem. Another expository paper on the application of these modular techniques to Diophantine equations can be found in [**Sik07**].

It is a pleasure to thank Henri Darmon and Alain Kraus for their help and their support.

2. Preliminary section

In this section, we give some classical necessary preparation for the theorems. Namely, following the lines of the exposition in section 4 of [**Dar**], we attach successively three objects to a hypothetical solution (x, y, z) of (1):

1. A Frey curve E_0 whose invariants can be computed.
2. A representation ρ describing the action of $\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ on the p-division points of E_0.
3. Corresponding to ρ is a cusp form f of weight 2 for $\Gamma_0(N)$, where N divides the conductor of E_0. We then reduce to the case where f has integer coefficients.

After this preparation, the point is to be able to discard all such modular forms. Halberstadt and Kraus manage to do so using their so-called "symplectic criterion" which will be explained in detail in the last section.

We proceed by contradiction and start from a hypothetical non-trivial solution

$$(x, y, z) \in \mathbb{Q}^3$$

of equation (1). Adjusting p^{th}-powers and clearing denominators, we can assume without loss of generality that x, y, z are coprime integers, and that a, b and c do not contain any p^{th}-powers.

One can reorder and label the three integers ax^p, by^p and cz^p by A, B and C respectively so that B is the only even integer among them, and $A \equiv \pm 1 \mod 4$. By adjusting the signs of our solution, we are reduced to the case where $A \equiv -1 \mod 4$. To this data $A + B + C = 0$ we attach the Frey curve over $\mathbb{Q}$

$$E_0 : \qquad Y^2 = X(X - A)(X + B).$$

The computation of its invariants on this model using classical formulae [**Sil86**, p.46] leads to:

$$\tilde{c}_4 = 16(A^2 + AB + B^2) \qquad \text{and} \qquad \tilde{\Delta} = 16(abc)^2(xyz)^{2p} = 16(ABC)^2.$$

If $\ell \neq 2$ is a prime dividing $\tilde{\Delta}$, it cannot divide $\tilde{c}_4$. Hence E_0 is semi-stable outside 2.

To study the reduction of E_0 at $\ell = 2$, let us change the variables to $X' = 4X$ and $Y' = 8Y + 4X$. Assuming that 16 divides B (since we will assume that $p \geq 5$, even 32 divides B), one obtains a global minimal Weierstrass equation for E_0 over $\mathbb{Q}$. At this point the minimal discriminant turns out to be

$$\Delta(E_0) = 2^{-8}(ABC)^2, \tag{2}$$

and $c_4(E_0)$ is odd. Finally, E_0 is also semi-stable at $\ell = 2$, and thus is semi-stable. Its conductor is the radical of the discriminant, that is (because 32 divides B)

$$N_{E_0} = \prod_{\ell \text{ prime}, \ell | ABC} \ell.$$

Key observation: Notice the factor 2^{-8} involved in formula (2) for the minimal discriminant. The "minus sign" of the exponent turns out to be crucial in the proof of Theorem 1.1.

The set of p-torsion points $E_0[p]$ of $E_0(\bar{\mathbb{Q}})$ forms a $\mathbb{F}_p$-vector space of dimension 2. The absolute Galois group $G_{\mathbb{Q}} = \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ acts naturally on $E_0[p]$. Thus we obtain a representation

$$\rho : G_{\mathbb{Q}} \to \mathrm{Aut}(E_0[p]) \simeq GL_2(\mathbb{F}_p).$$

If ρ is reducible, then E_0 contains a rational subgroup of order p. This cannot be the case if $p \geq 17$ because of the boundedness result of Mazur [**Maz77**, Th. 8] for the torsion of elliptic curves over $\mathbb{Q}$. Hence ρ is irreducible if p is large enough. Notice how our original Diophantine question has been transferred to this new Diophantine problem solved by Mazur. For more on this result, see [**Reb**] in this volume. This bound $p \geq 17$ is sufficient for us to prove Theorem 1.1. Nevertheless, a more precise result is given in [**Kra97**, Lemma 4] showing that ρ is irreducible as soon as $p \geq 5$.

Serre [**Ser87**] associates to such a representation a conductor $N | N_{E_0}$. In our context we have

$$N = 2\,\mathrm{rad}(abc) := 2 \prod_{\ell \text{ prime}, \ell | abc} \ell.$$

By the result of Wiles [**Wil95**], the semi-stable elliptic curve E_0 is modular: the function on the upper half-plane $\tau \mapsto \sum_{n \geq 1} a_n(E_0) q^n$ belongs to the space $S_2(\Gamma_0(N_{E_0}))$ of cuspidal modular forms of weight 2 on $\Gamma_0(N_{E_0})$.

The "lowering the level" Theorem of Ribet [**Rib90**] ensures that the representation ρ is then modular: there exists a newform $f = q + \sum_{n \geq 2} a_n q^n$ of weight 2 on $\Gamma_0(N)$ (where N now depends only on abc and not on (x, y, z) or p) and a place $\mathfrak{p}$ of $K_f = \mathbb{Q}(a_2, \ldots, a_n, \ldots)$ above p such that

$$\begin{array}{lll} i) & a_\ell \equiv a_\ell(E_0) & \mod \mathfrak{p} \text{ if } \ell \nmid N_{E_0} p \\ ii) & a_\ell \equiv \pm(\ell + 1) & \mod \mathfrak{p} \text{ if } \ell \mid N_{E_0} \text{ and } \ell \nmid pN. \end{array} \tag{3}$$

In the case of Fermat's last Theorem, one could show that $N = 2$ and the previous results were enough (!) to derive a contradiction since there is no cusp form of weight 2 and this level. In proving Theorem 1.1 and 1.2, Halberstadt and Kraus needed to refute the existence of such a form using an additional argument.

3. Proof of Theorems 1.1 and 1.2

Let f be the modular form of level N given by the previous construction. Both f and N do not depend on the solution (x, y, z) nor on p. We first reduce to the case where the modular form f has coefficients in $\mathbb{Z}$. Otherwise, the finite extension $K = K_f$ of $\mathbb{Q}$ has degree bounded by $g = \dim_{\mathbb{Q}}(S_2^{\mathrm{new}}(\Gamma_0(N))$. Let $a_\ell \notin \mathbb{Z}$ for the smallest possible prime ℓ. Both g and ℓ do not depend on p. We can assume that ℓ does not divide pN because a_ℓ would be $0, \pm 1$. Thus in the previous case $i)$ p divides $N_{K/\mathbb{Q}}(a_\ell - a_\ell(E_0))$, while in case $ii)$ p divides $N_{K/\mathbb{Q}}(a_\ell \pm (\ell + 1))$. The Hasse bound gives $|a_\ell(E_0)| \leq 2\sqrt{\ell}$, while Weil-Deligne's bound shows that $|\sigma(a_\ell)| \leq 2\sqrt{\ell}$ for each real embedding σ of K.

In any case p is bounded by a number depending only on a, b, c. Therefore, choosing large enough p we can make sure that f has integer coefficients. Under this hypothesis, the Eichler-Shimura theory provides an elliptic curve E' over $\mathbb{Q}$ of conductor N such that the Hasse-Weil function of E' is $\sum a_n n^{-s}$.

For almost all primes ℓ, the congruence relations (3) impose that $a_\ell \equiv a_\ell(E_0) \bmod p$. This is enough to show that the Galois modules $E[p]$ and $E'[p]$ are isomorphic. For, if $\ell \nmid pN_{E_0}$ the Frobenius element Frob_ℓ in $\mathrm{Aut}(E[p])$ has trace (resp. determinant) a_ℓ mod p (resp. ℓ mod p). The same occurs with E'. By the Chebotarev density theorem, this implies that an element $g \in G_\mathbb{Q}$ has the same characteristic polynomial when it acts on $E[p]$ or $E'[p]$. Thus the two representations of $G_\mathbb{Q}$ in the p-division points of E and E' have isomorphic semi-simplifications. Our assertion follows since $E[p]$ is irreducible.

At this point, the following key proposition is in order:

Proposition 3.1 ([**KO92**], Prop. 2). *Let E and E' be two elliptic curves over $\mathbb{Q}$ with minimal discriminants Δ and Δ', and let p be a prime number.*

Assume that the groups of p-torsion points $E[p]$ and $E'[p]$ are isomorphic as $G_\mathbb{Q}$-modules. Assume also that E and E' have multiplicative reduction at a common prime $\ell \neq p$ such that p does not divide the valuation $v_\ell(\Delta)$. Then we have

a) *The prime p does not divide $v_\ell(\Delta')$.*
b) *The following conditions are equivalent:*
   (i) there is a symplectic (viz. compatible with the Weil pairing on $E[p]$ and $E'[p]$) isomorphism between these two representations.
   (ii) the quotient $v_\ell(\Delta)/v_\ell(\Delta')$ is a square in $(\mathbb{Z}/p\mathbb{Z})^*$.

We postpone the proof of this "symplectic criterion" to the last section. The way it implies Theorems 1.1 and 1.2 is a bit tricky. Up to isogeny, there is only a finite number of elliptic curves over $\mathbb{Q}$ of conductor N, say $E_1, \ldots, E_h$. We label our previous curve $E' = E_j$ among them, and we want to apply the criterion to the pair (E_0, E_j).

Recall that E_0 has minimal discriminant

$$\Delta(E_0) = 2^{-8}(abc)^2(xyz)^{2p}.$$

We can assume that $|abc| > 2$ by Fermat's last theorem. Now we choose a first prime ℓ_1 dividing the odd integer abc, and $\ell_2 = 2$. If p is large enough, p divides neither $v_{\ell_1}(\Delta(E_0))$ nor $v_2(\Delta(E_0))$.

Let us emphasise that we are not going to decide whether or not E_j and E_0 are symplectically isomorphic. But in both cases, Proposition 3.1.b implies that *the product of the two terms*

$$\frac{v_{\ell_1}(\Delta(E_0))}{v_{\ell_1}(\Delta(E_j))} \bmod p \quad \text{and} \quad \frac{v_2(\Delta(E_0))}{v_2(\Delta(E_j))} \bmod p$$

is a square mod p because both terms are simultaneously squares or non-squares.

Equality (2) shows that the numerator of this product is

$$\begin{aligned} v_{\ell_1}(\Delta(E_0))v_2(\Delta(E_0)) &\equiv 2\, v_{\ell_1}(abc)(-8) \mod p \\ &\equiv -16\, v_{\ell_1}(abc) \mod p. \end{aligned}$$

Therefore the symplectic criterion implies that the integer n_j defined by

$$n_j = -v_{\ell_1}(abc)v_{\ell_1}(\Delta(E_j))v_2(\Delta(E_j))$$

has to be a square mod p.

Hence if $p \gg_{a,b,c} 0$ is a prime satisfying

$$\left(\frac{n_j}{p}\right) = -1 \quad \text{for all } j = 1, \dots, h, \tag{4}$$

the equation $ax^p + by^p + cz^p = 0$ has no non-trivial solution. It remains to show that these conditions are simultaneously satisfied on a set of positive density. To do this, let p be a prime such that

i) -1 is a non-square mod p;
ii) each prime divisor of n_j ($j = 1, \dots, h$) is a square mod p.

The previous two conditions define a subset of $\mathcal{P}$ which has positive density by Chebotarev's Theorem. Theorem 1.1 follows. □

Proof of Theorem 1.2 (sketch):

Both equations $x^p + 3y^p + 5z^p = 0$ and $x^p + y^p + 15z^p = 0$ have coefficients a, b, c satisfying $abc = 15$. The existence of a putative non-trivial rational solution with $p \geq 7$ leads to cusp forms of level $N = 30$. There is only one such newform of weight 2. Thus the Galois module $E_0[p]$ has to be isomorphic to $E_1[p]$, where E_1 is an elliptic curve of conductor 30, say 30A1 in Cremona's tables. The minimal discriminant of E_1 is

$$\Delta(E_1) = -2^4 3^3 5.$$

Then we choose $\ell_1 = 2$ and $\ell_2 = 5$ to deduce that $n_1 = -1$ must be a square mod p. But we could also use $\ell_2 = 3$ and obtain that -3 must be a square mod p.

The only primes p satisfying both conditions are those congruent to 1 mod 12. If we avoid such primes, there can be no non-trivial solutions. Therefore we obtain the conclusion of Theorem 1.2, at least for p large enough. The lower bound for p can be made precise using the explicit formulations of [**Kra97**]. □

4. Proof of the symplectic criterion

We conclude this paper by proving the key Proposition 3.1, following closely the lines of [**KO92**]. The proof consists of a local study of E and E' at the place ℓ, for which the Tate curve model can be used to make the computations explicit.

Let $K = \mathbb{Q}_\ell^{\mathrm{nr}}$ denote the maximal unramified extension of $\mathbb{Q}_\ell$. Both E and E' having multiplicative reduction over $\mathbb{Q}$ at ℓ, their j-invariant is not an integer in K. We deduce from [**Sil94**, Th. V.5.3] that E is uniformized over K by a Tate curve $\mathbb{G}_m/q^{\mathbb{Z}}$, where q in K has valuation $e = -v_\ell(j(E)) = v_\ell(\Delta)$. The same is true for E', with a $q' \in K$ of valuation $e' = v_\ell(\Delta')$.

The given isomorphism and the previous uniformizations combine to provide a $\mathrm{Gal}(\bar{K}/K)$-module isomorphism Ψ between the p-division points $E[p]$ of $\bar{K}^*/q^{\mathbb{Z}}$ and those $E'[p]$ of $\bar{K}^*/q'^{\mathbb{Z}}$.

Let us describe the effect of $G_K = \mathrm{Gal}(\bar{K}/K)$ and Ψ on a basis of $E[p]$, following [**Sil94**, Prop. 5.6.1]. First note that K contains the p^{th}-roots of unity, and let us fix ζ a primitive one. Fix also $\gamma \in \bar{K}$, a p^{th}-root of q. Then $\{\zeta q^{\mathbb{Z}}, \gamma q^{\mathbb{Z}}\}$ forms a basis for $E[p]$. The Galois group G_K acts transitively on the p conjugates $\{\zeta^j \gamma,\ 1 \leq j \leq p\}$. Hence there is a distinguished element $\sigma \in G_K$ which satisfies $\sigma(\zeta) = \zeta\gamma$, i.e. whose matrix is $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$.

As G_K fixes ζ, it acts trivially on $E[p]$ iff γ is in K, that is, iff p divides $e = v_\ell(q)$. The same assertion holds for $E'[p]$ and e'. These two Galois modules are isomorphic and p

does not divide e by assumption, so we conclude that p cannot divide e', which is assertion a). Hence there are integers m and n such that

$$e' = ne + mp.$$

We detail how Ψ acts on our basis. Since $q'/(q^n l^{mp})$ is a unit in K, it has a p^{th}-root $\alpha \in K$. We obtain a p^{th}-root of q' by setting $\gamma' = \gamma^n l^m \alpha$, completing a basis $\{\zeta q'^{\mathbb{Z}}, \gamma' q'^{\mathbb{Z}}\}$ of $E'[p]$.

Observe that for all $g \in G_K$, we have $\Psi(\zeta q^{\mathbb{Z}})^g = \Psi((\zeta q^{\mathbb{Z}})^g) = \Psi(\zeta q^{\mathbb{Z}})$ because Ψ is compatible with G_K. Therefore the matrix of Ψ with respect to the previous basis is upper triangular, say of the form $\left(\begin{smallmatrix} a & b \\ 0 & d \end{smallmatrix}\right)$.

The very definitions of σ and γ' lead to the identity $\sigma(\gamma') = \sigma(\gamma)^n \sigma(l^m \alpha) = \zeta^n \gamma'$. Compatibility between Ψ and σ can be written in matrix terms as follows:

$$\begin{pmatrix} a & b \\ 0 & d \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}\begin{pmatrix} a & b \\ 0 & d \end{pmatrix}.$$

Identification of upper right entries shows the intermediate identity

$$a = nd. \tag{5}$$

Now we turn to the Weil pairing. It is a bilinear alternate pairing satisfying the following identities on $E[p]$ and $E'[p]$ respectively:

$$B(\gamma q^{\mathbb{Z}}, \zeta q^{\mathbb{Z}}) = \zeta, \quad B'(\gamma q'^{\mathbb{Z}}, \zeta q'^{\mathbb{Z}}) = \zeta.$$

Assuming that Ψ is a symplectic isomorphism, we obtain

$$\zeta = B(\gamma q^{\mathbb{Z}}, \zeta q^{\mathbb{Z}}) = B'(\Psi(\gamma q^{\mathbb{Z}}), \Psi(\zeta q^{\mathbb{Z}})) = B'(\zeta^b \gamma^d q'^{\mathbb{Z}}, \zeta^a q'^{\mathbb{Z}}) = \zeta^{ad}.$$

It follows that $ad \equiv 1 \mod p$, or $nd^2 \equiv 1 \mod p$ by (5) and n is a square modulo p.

Reciprocally, if n is a square modulo p, there is an integer r such that $r^2 nd^2 = 1 \mod p$. It can be easily checked that the r^{th}-power Ψ^r defines the required *symplectic* isomorphism between the Tate curves, hence E and E' are symplectically isomorphic. □

5. Limitations of the method

The paper [**HK02**] presents the symplectic method and two others (called the reduction method and the decomposition method) to handle the case of different generalized Fermat equations. Even if Theorem 1.1 is successful, as it solves an infinite family of Fermat equations, many questions are still open.

For instance, the remaining case $p = 1 \mod 12$ in Theorem 1.2 cannot be settled using the methods of Halberstadt and Kraus. This would provide a complete answer to the question raised by Serre.

The authors also mention (Exemple 2.12) the case of the curve

$$16x^7 + 87y^7 + 625z^7 = 0.$$

Denote by E_0 the corresponding Frey curve and by E_1 the elliptic curve 435C2. The symplectic criterion cannot ensure that $\rho_7^{E_0}$ and $\rho_7^{E_1}$ are not isomorphic since E_0 and E_1 have discriminants $3^2 5^8 29^2 (xyz)^{14}$ and $3^4 5^2 29^2$ respectively.

Moreover, the aim of their three methods is to show that the set of solutions of some generalized Fermat equation is trivial. Thus the case of the Diophantine equation $ax^p + by^p + cz^p = 0$ with $a + b + c = 0$ falls out of their scope because the non-trivial solution $(1, 1, 1)$ has to be considered.

Nevertheless, a result in this setting has been obtained in [**DM97**], providing an optimistic conclusion to this section and to this note:

THEOREM 5.1 (Darmon-Merel (1997) [**DM97**]). *Let $n \geq 3$ be an arbitrary integer. Then the equation*

$$x^n + y^n - 2z^n = 0$$

has no integer solutions $(x, y, z) \in \mathbb{Z}^3$ with $|xyz| > 1$.

References

[Dar] H. Darmon, *Rational points on curves*, in this volume.

[DM97] H. Darmon and L. Merel, *Winding quotients and some variants of Fermat's last theorem*, J. Reine Angew. Math. **490** (1997), 81–100. MR 1468926 (98h:11076)

[HK02] E. Halberstadt and A. Kraus, *Courbes de Fermat: résultats et problèmes*, J. Reine Angew. Math. **548** (2002), 167–234. MR 1915212 (2003h:11068)

[KO92] A. Kraus and J. Oesterlé, *Sur une question de B. Mazur*, Math. Ann. **293** (1992), no. 2, 259–275. MR 1166121 (93e:11074)

[Kra97] A. Kraus, *Majorations effectives pour l'équation de Fermat généralisée*, Canad. J. Math. **49** (1997), no. 6, 1139–1161. MR 1611640 (99g:11039)

[Maz77] B. Mazur, *Modular curves and the Eisenstein ideal*, Inst. Hautes Études Sci. Publ. Math. (1977), no. 47, 33–186 (1978). MR 488287 (80c:14015)

[Reb] M. Rebolledo, *Merel's theorem for the boundedness of the torsion of elliptic curves*, in this volume.

[Rib90] K. A. Ribet, *On modular representations of* $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ *arising from modular forms*, Invent. Math. **100** (1990), no. 2, 431–476. MR 1047143 (91g:11066)

[Ser87] J.-P. Serre, *Sur les représentations modulaires de degré* 2 *de* $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$, Duke Math. J. **54** (1987), no. 1, 179–230. MR 885783 (88g:11022)

[Sik07] S. Siksek, *The modular approach to Diophantine equations*, Graduate Texts in Mathematics, vol. 240, ch. 15, pp. xxiv+596, Springer, New York, 2007, a book by H. Cohen. MR 2312338 (2008e:11002)

[Sil86] J. H. Silverman, *The arithmetic of elliptic curves*, Graduate Texts in Mathematics, vol. 106, Springer-Verlag, New York, 1986. MR 817210 (87g:11070)

[Sil94] J.H. Silverman, *Advanced topics in the arithmetic of elliptic curves*, Graduate Texts in Mathematics, vol. 151, Springer-Verlag, New York, 1994. MR 1312368 (96b:11074)

[Wil95] A. Wiles, *Modular elliptic curves and Fermat's last theorem*, Ann. of Math. (2) **141** (1995), no. 3, 443–551. MR 1333035 (96d:11071)

INSTITUT DE MATHÉMATIQUES DE JUSSIEU, UNIVERSITÉ PARIS 6, EQUIPE DE THÉORIE DES NOMBRES CASE 247 - 4, PLACE JUSSIEU - 75252 PARIS CEDEX FRANCE

E-mail address: charollois@math.jussieu.fr

Clay Mathematics Proceedings
Volume 8, 2009

Heegner points and Sylvester's conjecture

Samit Dasgupta and John Voight

Abstract. We consider the classical Diophantine problem of writing positive integers n as the sum of two rational cubes, i.e. $n = x^3 + y^3$ for $x, y \in \mathbb{Q}$. A conjecture attributed to Sylvester asserts that a rational prime $p > 3$ can be so expressed if $p \equiv 4, 7, 8 \pmod 9$. The theory of mock Heegner points gives a method for exhibiting such a pair (x, y) in certain cases. In this article, we give an expository treatment of this theory, focusing on two main examples: a theorem of Satgé, which asserts that $x^3 + y^3 = 2p$ has a solution if $p \equiv 2 \pmod 9$, and a proof sketch that Sylvester's conjecture is true if $p \equiv 4, 7 \pmod 9$ and 3 is not a cube modulo p.

1. A Diophantine problem

1.1. Sums of rational cubes. We begin with the following simple Diophantine question.

Question. Which positive integers n can be written as the sum of two cubes of rational numbers?

For $n \in \mathbb{Z}_{>0}$, let E_n denote the (projective nonsingular) curve defined by the equation $x^3 + y^3 = nz^3$. This curve has the obvious rational point $\infty = (1 : -1 : 0)$, and equipped with this point the curve E_n has the structure of an elliptic curve over $\mathbb{Q}$. The equation for E_n can be transformed via the change of variables

$$X = 12n\frac{z}{x+y}, \quad Y = 36n\frac{x-y}{x+y} \tag{1}$$

to yield the affine Weierstrass equation $Y^2 = X^3 - 432n^2$.

We then have the equivalent question: Which curves E_n have a nontrivial rational point? For n not a cube or twice a cube, $E_n(\mathbb{Q})_{\text{tors}} = \{\infty\}$ (see [**Sil86**, Exercise 10.19]), so also equivalently, which curves E_n have positive rank $\text{rk}(E_n(\mathbb{Q})) > 0$?

Examples. Famously, $1729 = 1^3 + 12^3 = 9^3 + 10^3$; also,

$$\left(\frac{15642626656646177}{590736058375050}\right)^3 + \left(\frac{-15616184186396177}{590736058375050}\right)^3 = 94.$$

2000 *Mathematics Subject Classification.* Primary 11G05; Secondary 11F11, 11D25.

Key words and phrases. Modular forms, elliptic curves, Heegner points, Diophantine equations.

In each case, these solutions yield generators for the group $E_n(\mathbb{Q})$. (Note $n = 94 = 2 \cdot 47$ is a case covered by Satgé's theorem below, cf. §3.1.)

1.2. Sylvester's conjecture. We now consider the case $n = p \geq 5$ is prime.

Conjecture (Sylvester, Selmer [**Sel51**]). *If $p \equiv 4, 7, 8 \pmod 9$, then p is the sum of two rational cubes.*

Although this conjecture is traditionally attributed to Sylvester (see [**Syl79b**, §2] where he considers "classes of numbers that cannot be resolved into the sum or difference of two rational cubes"), we cannot find a specific reference in his work to the above statement or one of its kind (see also [**Syl79a, Syl80a, Syl80b**]).

An explicit 3-descent (as in [**Sel51**], see also [**Sat86**]) shows that

$$\operatorname{rk}(E_p(\mathbb{Q})) \leq \begin{cases} 0, & \text{if } p \equiv 2, 5 \pmod 9; \\ 1, & \text{if } p \equiv 4, 7, 8 \pmod 9; \\ 2, & \text{if } p \equiv 1 \pmod 9. \end{cases}$$

Hence $\operatorname{rk}(E_p(\mathbb{Q})) = 0$ for $p \equiv 2, 5 \pmod 9$, a statement which can be traced back to Pépin, Lucas, and Sylvester [**Syl79b**, Section 2, Title 1].

The sign of the functional equation for the L-series of E_p is

$$\operatorname{sign}(L(E_p/\mathbb{Q}, s)) = \begin{cases} -1, & \text{if } p \equiv 4, 7, 8 \pmod 9; \\ +1, & \text{otherwise.} \end{cases}$$

(See [**Kob02**]; this can be derived from the determination of the local root numbers $w_p(E_p) = (-3/p)$ and $w_3(E_p) = 1$ if and only if $p \equiv \pm 1 \pmod 9$.)

Putting these together, for $p \equiv 4, 7, 8 \pmod 9$, the Birch–Swinnerton-Dyer (BSD) conjecture predicts that $\operatorname{rk}(E_p(\mathbb{Q})) = 1$.

1.3. A few words on the case $p \equiv 1 \pmod 9$. For $p \equiv 1 \pmod 9$, the BSD conjecture predicts that $\operatorname{rk}(E_p(\mathbb{Q})) = 0$ or 2, depending on p. This case was investigated by Rodriguez-Villegas and Zagier [**RVZ95**].

Define $S_p \in \mathbb{R}$ by

$$L(E_p/\mathbb{Q}, 1) = \frac{\Gamma(\frac{1}{3})^3\sqrt{3}}{2\pi\sqrt[3]{p}} S_p;$$

then in fact $S_p \in \mathbb{Z}$, and conjecturally (BSD) we have $S_p = 0$ if $\#E_p(\mathbb{Q}) = \infty$ and $S_p = \#Ш(E_p)$ otherwise. Rodriguez-Villegas and Zagier give two formulas for S_p, one of which proves that S_p is a square. They also give an efficient method to determine whether $S_p = 0$.

1.4. The case $p \equiv 4, 7, 8 \pmod 9$: an overview. Assume from now on that $p \equiv 4, 7, 8 \pmod 9$. We can easily verify Sylvester's conjecture for small primes p.

$$\begin{aligned} 7 &= 2^3 + (-1)^3 \\ 13 &= (7/3)^3 + (2/3)^3 \\ 17 &= (18/7)^3 + (-1/7)^3 \\ 31 &= (137/42)^3 + (-65/42)^3 \\ 43 &= (7/2)^3 + (1/2)^3 \\ &\vdots \end{aligned}$$

Again, the BSD conjecture predicts that we should always have that p is the sum of two cubes. General philosophy predicts that in this situation where E_p has expected rank 1, one should be able to construct rational nontorsion points on E_p using the theory of complex multiplication (CM).

In §2, we introduce the construction of *Heegner points*, which uses the canonical modular parametrization $\Phi : X_0(N) \to E_p$ where N is the conductor of E_p; this strategy requires a choice of imaginary quadratic extension K and is therefore not entirely "natural". If instead we try to involve the field $K = \mathbb{Q}(\omega)$, we arrive at a theory of *mock Heegner points*. We then choose a fixed modular parametrization $X_0(N) \to E$ where E is a designated *twist* of E_p for each prime p.

In §3, we illustrate one such example, originally due to Satgé. We look at the parametrization $X_0(36) \to E$ where $E : y^2 = x^3 + 1$ is a twist of the curve E_{2p}. We show that when $p \equiv 2 \pmod 9$, the equation $x^3 + y^3 = 2p$ has a solution; the proof involves a careful analysis of the relevant Galois action using the Shimura reciprocity law and explicit recognition of modular automorphisms.

In §4, we return to Sylvester's conjecture, and we sketch a proof that the conjecture is true if $p \equiv 4, 7 \pmod 9$ and 3 is not a cube modulo p; here, we employ the parametrization $X_0(243) \to E_9$. We close with some open questions.

2. Heegner and Mock Heegner points

2.1. Heegner points. The curve E_p has conductor $N = 9p^2$ if $p \equiv 7 \pmod 9$ and conductor $N = 27p^2$ if $p \equiv 4, 8 \pmod 9$. We have the modular parametrization

$$\Phi : X_0(N) \to E_p,$$

from which we may define Heegner points as follows.

Let $K = \mathbb{Q}(\sqrt{D})$ be a imaginary quadratic field of discriminant D such that 3 and p split in K; the pair (E_p, K) then satisfies the *Heegner hypothesis*. Let $\mathcal{O}_K$ denote the ring of integers of K, and let $\mathfrak{N} \subset \mathcal{O}_K$ be a cyclic ideal of norm N. Then the cyclic N-isogeny

$$\mathbb{C}/\mathcal{O}_K \to \mathbb{C}/\mathfrak{N}^{-1}$$

defines a *CM point* $P \in X_0(N)(H)$, where H is the Hilbert class field of K.

Let $Y = \mathrm{Tr}_{H/K} \Phi(P) \in E_p(K)$ denote the trace, known as a *Heegner point*. After adding a torsion point if necessary, we may assume $Y \in E_p(\mathbb{Q})$ (see [**Dar04**, §3.4], and note $E_p(K)_{\mathrm{tors}} = E_p[3](K) \cong \mathbb{Z}/3\mathbb{Z}$.)

2.2. Gross-Zagier formula. The Gross-Zagier formula indicates when we expect the point $Y \in E_p(\mathbb{Q})$ to be nontorsion, i.e. when its canonical height $\hat{h}(Y)$ is nonzero.

Theorem (Gross-Zagier formula [**Dar04**, Theorem 3.20]). *We have*

$$\hat{h}(Y) \doteq L'(E_p/K, 1) = L'(E_p/\mathbb{Q}, 1)L(E_p/\mathbb{Q}, \chi_K, 1).$$

Here the symbol $\doteq$ denotes equality up to an explicit nonzero "fudge factor." Thus if we choose K such that $L(E_p/\mathbb{Q}, \chi_K, 1) \neq 0$, the BSD conjecture implies that $\hat{h}(Y) \neq 0$ and hence Y will be nontorsion. Working algebraically, without making any reference to L-functions, one might hope to prove that Y is nontorsion directly and unconditionally. But this strategy seems tricky—in particular, no natural candidate for K presents itself. In the next section we discuss a more "natural" approach to constructing a nontorsion point on E_p.

2.3. Mock Heegner points. We consider a variation of the above method where we construct what are known as *mock Heegner points*; this terminology is due to Monsky [**Mon90**, p. 46], although Heegner's original construction can be described as an example of such "mock" Heegner points.

Consider the field $K = \mathbb{Q}(\sqrt{-3}) = \mathbb{Q}(\omega)$, where ω is a primitive cube root of unity. Note that the elliptic curve $E_n : x^3 + y^3 = nz^3$ has CM by $\mathcal{O}_K$, given by

$$[\omega](x, y) = (\omega x, \omega y).$$

The prime 3 is ramified in K, so the Heegner hypothesis is not satisfied for the pair (E_p, K). Nevertheless, Heegner-like constructions of points defined by CM theory may still produce nontorsion points in certain situations.

2.4. Twisting. Notice that

$$(r/\sqrt[3]{p})^3 + (s/\sqrt[3]{p})^3 = 1 \quad \Longleftrightarrow \quad r^3 + s^3 = p. \tag{2}$$

The obvious equivalence (2) suggests that to find points on $E_p(K)$, we may identify E_p as the *cubic twist* of E_1 by $\sqrt[3]{p}$. More precisely, let $L = K(\sqrt[3]{p})$, and let σ be the generator of $\mathrm{Gal}(L/K)$ satisfying $\sigma(\sqrt[3]{p}) = \omega\sqrt[3]{p}$. The Galois group $\mathrm{Gal}(K/\mathbb{Q})$ is generated by complex conjugation, which we denote by $\overline{\ }$. We have an isomorphism of groups

$$\begin{aligned} E_p(\mathbb{Q}) &\cong \{(r/\sqrt[3]{p}, s/\sqrt[3]{p}) \in E_1(L) : r, s \in \mathbb{Q}\} \\ &= \{Y \in E_1(L) : Y^\sigma = \omega^2 Y,\ \overline{Y} = Y\}. \end{aligned}$$

In other words, we look for points on $E_1(L)$ with specified behavior under $\mathrm{Gal}(L/\mathbb{Q})$.

More generally (see [**Sil86**, §X.5]), if $E/\mathbb{Q}$ is an elliptic curve, then one defines the set of *twists* of E to be the set of elliptic curves over $\mathbb{Q}$ that become isomorphic to E over $\overline{\mathbb{Q}}$, modulo isomorphism over $\mathbb{Q}$. There is a natural bijection between the set of twists of E and the Galois cohomology group

$$H^1(\mathbb{Q}, \mathrm{Aut}(E)) := H^1(\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}), \mathrm{Aut}(E_{\overline{\mathbb{Q}}})).$$

In our setting,

$$\begin{aligned} E_p(\mathbb{Q}) &\cong \{Y \in E_1(L) : Y^\sigma = \omega^2 Y,\ \overline{Y} = Y\} \\ &= \{Y \in E_1(L) : Y^\tau = c_\tau Y \text{ for all } \tau \in \mathrm{Gal}(L/\mathbb{Q})\} \end{aligned} \tag{3}$$

where $[c_\tau] \in H^1(\mathbb{Q}, \mathrm{Aut}(E_1))$ is the cohomology class represented by the cocycle $c_\tau := \sqrt[3]{p}/\tau(\sqrt[3]{p})$. To find a point Y in the set (3), we may take any $Q \in E_1(L)$ and consider the *twisted trace*

$$Q' = Q + \omega Q^\sigma + \omega^2 Q^{\sigma^2} \in E_1(L).$$

The point Q' has the property that $(Q')^\sigma = \omega^2(Q')$.

Now suppose that Q' is nontorsion. Consider then the point $Y = Q' + \overline{Q'}$ in the set (3); either it will be nontorsion, or else it will be trivial and then instead $\sqrt{-3}Q'$ is a nontorsion point in the set (3). Thus, in any case, a nontorsion Q' will yield a nontorsion Y.

2.5. Mock Heegner points on $X_0(27)$. To summarize, if we can construct a point $Q \in E_1(L)$, then by taking a twisted trace we can construct a (hopefully nontorsion) point $Y \in E_p(\mathbb{Q})$. We look to CM theory to construct the point Q.

We have a modular parametrization

$$\begin{aligned}\Phi : X_0(27) &\xrightarrow{\sim} E_1 : Y^2 + 9Y = X^3 - 27\\ z &\mapsto (X, Y) = \left(\frac{\eta(9z)^4}{\eta(3z)\eta(27z)^3}, \frac{\eta(3z)^3}{\eta(27z)^3}\right),\end{aligned}$$

where $\eta(z) = q^{1/24}\prod_{n=1}^{\infty}(1-q^n)$ is the Dedekind eta-function and $q = \exp(2\pi i z)$. In this case, the map Φ is an isomorphism of curves.

The field $L = K(\sqrt[3]{p})$ is a cyclic extension of K with conductor

$$\mathfrak{f}(L/K) = f = \begin{cases} 3p, & \text{if } p \equiv 4, 7 \pmod 9;\\ p, & \text{if } p \equiv 8 \pmod 9.\end{cases}$$

As L is of dihedral type over $\mathbb{Q}$, it is contained in the ring class field of K of conductor f, denoted H_f. Let $\mathcal{O}_{K,f} = \mathbb{Z} + f\mathcal{O}_K$ denote the order of $\mathcal{O}_K$ of conductor f, and let $P \in X_0(27)(H_f)$ be defined by a cyclic 27-isogeny between elliptic curves with CM by $\mathcal{O}_{K,f}$. We define the point $Q = \mathrm{Tr}_{H_f/L}\,\Phi(P) \in E_1(L)$ and ask: Is the point Q nontorsion?

Let us compute an example with $p = 7$. For an element z in the complex upper half plane $\mathfrak{H}$, denote by $\langle z\rangle$ the elliptic curve $\mathbb{C}/\langle 1, z\rangle$. We have a cyclic 27-isogeny, obtained as a chain of 3-isogenies, given by

$$\langle \omega p/3\rangle \to \langle \omega p\rangle \to \langle (\omega p + 2)/3\rangle \to \langle (\omega p + 2)/9\rangle; \tag{4}$$

this isogeny has conductor $3p$. Under the identification $\Gamma_0(N)\backslash\mathfrak{H} \cong Y_0(N)$, an element $z \in \mathfrak{H}$ represents the isogeny $\langle z\rangle \to \langle Nz\rangle$. The isogeny in (4) is represented by the point $z = M(\omega p/3)$, where $M = \left(\begin{smallmatrix} 2 & 1\\ 3 & 2\end{smallmatrix}\right) \in SL_2(\mathbb{Z})$. In this case, we have $H_f = H_{3p} = K(\alpha)$ with $\alpha = \sqrt[6]{-7} = \sqrt[6]{7}\exp(\pi i/6)$. One computes that the point $\Phi(z) = P = (X, Y) \in E_1(H_{3p})$, in Weiestrass coordinates as above, agrees with the point

$$\begin{aligned}X = {} & (-180\omega - 90)\alpha^5 + (-216\omega - 216)\alpha^4 + \tfrac{1}{2}(-345\omega - 690)\alpha^3\\ & - 414\alpha^2 + (330\omega - 330)\alpha + \tfrac{1}{2}(1581\omega),\\ Y = {} & (-6210\omega + 6210)\alpha^5 - 14877\omega\alpha^4 + (-23760\omega - 11880)\alpha^3\\ & + (-28458\omega - 28458)\alpha^2 + (-22725\omega - 45450)\alpha - 54441\end{aligned}$$

to the precision computed. One can then verify computationally that

$$Q = \mathrm{Tr}_{H_f/L}(P) = (3\omega, 0) \in E_1(L)$$

is torsion!

The method we have outlined thus fails in this case; we see similar behavior for the eight other distinguished cyclic 27-isogenies of conductor $3p$, as well as for other values of p.

3. Satgé's construction

3.1. Satgé's construction. Our first attempt at constructing a mock Heegner point using the parametrization $X_0(27) \to E_1$ (in §2.5) yielded only torsion

points on $E_p(\mathbb{Q})$. We now exhibit a similar construction which *does* work, but not one which addresses Sylvester's conjecture.

THEOREM (Satgé [**Sat87**]). *If* $p \equiv 2 \pmod 9$, *then* $\#E_{2p}(\mathbb{Q}) = \infty$. *If* $p \equiv 5 \pmod 9$, *then* $\#E_{2p^2}(\mathbb{Q}) = \infty$.

Another result in the same vein is the following.

THEOREM (Coward [**Cow00**]). *If* $p \equiv 2 \pmod 9$, *then* $\#E_{25p}(\mathbb{Q}) = \infty$. *If* $p \equiv 5 \pmod 9$, *then* $\#E_{25p^2}(\mathbb{Q}) = \infty$.

Our expository treatment of Satgé's theorem will treat the first case, where $p \equiv 2 \pmod 9$; see also the undergraduate thesis of Balakrishnan [**Bal06**]. The second statement follows similarly. Our proof proceeds different than that of Satgé; his original proof is phrased instead in the language of modular forms.

3.2. Twisting. Instead of the parametrization $X_0(27) \to E_1$, we use

$$\Phi : X_0(36) \xrightarrow{\sim} E : y^2 = x^3 + 1.$$

Over K, the cubic twist of E by $\sqrt[3]{p}$ is isomorphic to E_{2p}. (Over $\mathbb{Q}$, it is the *sextic twist* of E by $\sqrt[6]{-27p^2}$, given by $y^2 = x^3 - 27p^2$, which is isomorphic to E_{2p}; the quadratic twist by $\sqrt{-3}$ yields a curve which is isomorphic over K, as well as 3-isogenous over $\mathbb{Q}$.) The twisting is then given by the group isomorphism

$$\begin{aligned} E_{2p}(\mathbb{Q}) &\cong \{P = (r\sqrt[3]{p}, s\sqrt{-3}) \in E(L) : r, s \in \mathbb{Q}\} \\ &= \{P \in E(L) : P^\sigma = c_\tau P \text{ for all } \tau \in \mathrm{Gal}(L/\mathbb{Q})\} \end{aligned}$$

where $[c_\tau] \in H^1(\mathrm{Gal}(L/\mathbb{Q}), \mathrm{Aut}(E))$ is represented by the cocycle

$$c_\tau := \frac{\tau(\beta)}{\beta}, \text{ where } \beta = \sqrt[6]{-27p^2}.$$

3.3. From H_{6p} to H_{3p}. From the cyclic 36-isogeny $\langle \omega p/6\rangle \to \langle 6\omega p\rangle$ of conductor $6p$, we obtain a point $P \in E(H_{6p})$, where $E : y^2 = x^3 + 1$.

We have the following diagram of fields.

$$\begin{array}{c} H_{6p} = H_{3p}(\sqrt[3]{2}) \\ \Big| 3 \\ H_{3p} \\ \Big| (p+1)/3 \\ L = K(\sqrt[3]{p}) \\ \Big| 3 \\ K \\ \Big| 2 \\ \mathbb{Q} \end{array}$$

As we now describe, it turns out that the trace from H_{6p} to H_{3p} is unnecessary in the trace from H_{6p} to L. Let

$$\rho \in \mathrm{Gal}(H_{6p}/H_{3p}) \subset \mathrm{Gal}(H_{6p}/K)$$

satisfy $\rho(\sqrt[3]{2}) = \omega\sqrt[3]{2}$.

PROPOSITION. *For $P \in E(H_{6p})$ as defined above, we have*

$$P^\rho = P + (0,1),$$

where $(0,1)$ is a 3-torsion point.

This proposition can be proved using the methods we introduce below, and so is left to the reader. It follows from this proposition that $\mathrm{Tr}_{H_{6p}/H_{3p}} P = 3P$. To eliminate this factor of 3, we introduce the point

$$T = (-\sqrt[3]{4}, -\sqrt{-3}) \in E[3](H_6)$$

and note that it also satisfies $T^\rho = T + (0,1)$. Thus letting

$$P_T := P - T, \tag{5}$$

we find $(P_T)^\rho = P_T$, so $P_T \in E(H_{3p})$.

3.4. From H_{3p} to $\mathbb{Q}$. Define

$$Q = \mathrm{Tr}_{H_{3p}/L} P_T \in E(L). \tag{6}$$

We now claim that the following equation holds.

PROPOSITION. *Let $\sigma \in \mathrm{Gal}(L/K)$ satisfy $\sigma(\sqrt[3]{p}) = \omega\sqrt[3]{p}$. Then we have*

$$Q^\sigma = \omega Q + (0,-1). \tag{7}$$

The point $(0,-1)$ is a 3-torsion point. It follows from equation (7) that the twisted trace is just

$$Y := Q + \omega^2 Q^\sigma + \omega Q^{\sigma^2} = 3Q \in E(L),$$

which via twisting corresponds to a point $Y' \in E_{2p}(K)$.

To conclude the proof of Theorem 3.1, assuming that equation (7) holds, we need to prove that Y, and hence Y', is nontorsion. It suffices to prove that Q is nontorsion. But $E_{\mathrm{tors}}(L) = \{O, (0,\pm 1)\}$, and no S in this set satisfies equation (7): indeed, $S^\sigma = S = \omega S$, so equation (7) for S would yield the contradiction $S = S + (0,-1)$. Note that this argument proves not only that the point Y' is nontorsion, but that it is not divisible by 3 in the group $E_{2p}(K)/E_{2p}(K)_{\mathrm{tors}}$.

3.5. The $\mathrm{Gal}(L/K)$-action. We now prove the equation (7). We will in fact prove an equation for $P \in E(H_{6p})$. We choose a lift of $\sigma \in \mathrm{Gal}(L/K)$ to $\mathrm{Gal}(H_{6p}/K)$. Namely, we let $\alpha_\sigma = 1 + 2p\omega$ and let $I_\sigma = \alpha\mathcal{O}_K \cap \mathcal{O}_{K,6p}$. One can show directly that under the Artin map

$$\mathrm{Frob} : I_{K,6p}/P_{\mathbb{Z},6p} \xrightarrow{\sim} \mathrm{Gal}(H_{6p}/K), \tag{8}$$

the ideal I_σ corresponds to an element $\sigma \in \mathrm{Gal}(H_{3p}/K)$ such that $\sigma(\sqrt[3]{p}) = \omega\sqrt[3]{p}$. In (8), $I_{K,6p}$ denotes the group of fractional ideals of K that are relatively prime to $6p$, and $P_{\mathbb{Z},6p}$ denotes the subgroup generated by principal ideals (α) where $\alpha \in \mathcal{O}_K$ satisfies $\alpha \equiv a \pmod{6p}$ for some $a \in (\mathbb{Z}/6p\mathbb{Z})^\times$.

The equation we will prove is

$$P^\sigma = \omega P + (-1,0), \tag{9}$$

from which one can deduce equation (7) using equations (5) and (6). The proof uses two ingredients: an explicit calculation with the *Shimura reciprocity law*, and an explicit identification of this action with a *modular automorphism*.

We begin with the first of these two steps in the following lemma.

LEMMA. *We have* $P^\sigma = \langle 3\omega p/2\rangle \to \langle (2\omega p+1)/3\rangle$.

PROOF. The point P is given by the isogeny $\langle \omega p/6\rangle \to \langle 6\omega p\rangle$. The Shimura reciprocity law ([**Shi71**, §6.8]) implies that P^σ is given by the isogeny

$$I_\sigma^{-1}\cdot\langle \omega p/6\rangle \to I_\sigma^{-1}\cdot\langle 6\omega p\rangle.$$

An explicit calculation shows that $I_\sigma^{-1}\cdot\langle \omega p/6\rangle \sim \langle 3\omega p/2\rangle$, where $\sim$ denotes homothety equivalence. Similarly, we find that $I_\sigma^{-1}\langle 6\omega p\rangle \sim \langle (2\omega p+1)/3\rangle$, thus concluding the proof. □

We now proceed with the second step. Any element of the normalizer of $\Gamma_0(36)$ in the group $\mathrm{PSL}_2(\mathbb{R})$ provides by linear fractional transformations an automorphism of $\Gamma_0(36)\backslash\mathfrak{H}^* = X_0(36)$. The group of such *modular automorphisms* is denoted $N(\Gamma_0(36))$. In the second step of the proof of (9), we find a modular automorphism M such that $M(P) = P^\sigma$. Moreover, since $X_0(36)$ is a curve of genus one, it is easy to determine its automorphism group; we may then identify M explicitly as an element of this automorphism group to obtain the relation (9). For more details concerning the results on modular automorphisms used in this section, see [**Ogg80**].

We now look for a matrix M in $N(\Gamma_0(36))$ such that $M(P) = P^\sigma$. Let H be the subgroup of $N(\Gamma_0(36))$ generated by the Atkin-Lehner involutions $w_4 = \left(\begin{smallmatrix} 4 & -1 \\ 36 & -8\end{smallmatrix}\right)$ and $w_9 = \left(\begin{smallmatrix} 9 & 2 \\ 36 & 9\end{smallmatrix}\right)$, together with the *exotic automorphism* $e = \left(\begin{smallmatrix} 1 & 0 \\ 6 & 1\end{smallmatrix}\right)$ of order 6—there exists such an exotic automorphism $\left(\begin{smallmatrix} 1 & 0 \\ N/t & 1\end{smallmatrix}\right)$ normalizing $\Gamma_0(N)$ whenever $t \in \mathbb{Z}_{>0}$ satisfies $t \mid 24$ and $t^2 \mid N$ (see [**Ogg80**]). The group H is a solvable group of order $\#H = 72$. One computes directly that $M = \left(\begin{smallmatrix} 9 & -4 \\ 36 & -15\end{smallmatrix}\right) \in H$ satisfies $M(P) = P^\sigma$, using the previous lemma.

Now the matrix M corresponds to an element of $\mathrm{Aut}(X_0(36))$, the automorphism group of $X_0(36)$ as an abstract curve. Via the isomorphism Φ, we may view $X_0(36)$ as the elliptic curve E and hence write $M(Z) = aZ + b$ for some $a \in \mathrm{Aut}(E) \cong \mu_6$ and some $b \in E(K)$. To determine a and b, we evaluate M on the cusps. The point $\infty \in X_0(36)$ corresponds under Φ to the origin of the elliptic curve. We find that $M(\infty) = 1/4$, which corresponds to the point $\Phi(1/4) = (-1, 0)$. Thus $b = (-1, 0)$. Similarly, evaluating at the cusp 0, we find that $a = \omega$. Putting these pieces together, we have $P^\sigma = M(P) = \omega P + (-1, 0)$ as claimed.

3.6. An example with $p = 11$. We illustrate the method of the preceding section with $p = 11$. Beginning with $z = \omega p/6$, we compute $P \in E(H_{6p})$ with x-coordinate which satisfies

$$x^{36} + 462331656\omega x^{35} + 11767817160\omega^2 x^{34} + 179182057872x^{33} + 543458657808\omega x^{32}$$
$$+ \cdots + 50331648x^3 + 1939159514087424\omega x^2 + 16777216 = 0$$

to the precision computed.

We next compute $P_T = P - T \in E(H_{3p})$, where $T = (-\sqrt[3]{4}, -\sqrt{-3})$ as above. The point P_T has x-coordinate which satisfies

$$25x^{12} + (354\omega - 270)x^{11} + (-5313\omega - 3432)x^{10} + (2376\omega + 17578)x^9$$
$$+ (21879\omega - 297)x^8 + (-6732\omega - 24552)x^7 + (-16632\omega + 61116)x^6$$
$$+ (3168\omega - 9504)x^5 + (-12672\omega - 45936)x^4 + (-19008\omega - 2816)x^3$$
$$+ (10560\omega)x^2 + (17664\omega - 5376)x + 10240 = 0.$$

The trace $Q = \mathrm{Tr}_{H_{3p}/L} P_T \in E(L)$, again to the precision calculated, is the point

$$Q = \Big(-\frac{1849}{5776}\sqrt[3]{11}^2 + \frac{645}{5776}\omega\sqrt[3]{11} + \frac{225\omega+225}{5776}, \\ \frac{27735\omega+55470}{438976}\sqrt[3]{11}^2 + \frac{-9675\omega+9675}{438976}\sqrt[3]{11} + \frac{871202\omega+435601}{438976}\Big).$$

We indeed find that the equation $Q^\sigma = \omega Q + (0,1)$ holds as in (7). Finally, the twisted trace is

$$Y = 3Q = \left(-\frac{767848016929}{79297693200}\omega\sqrt[3]{11}, \frac{672808015029320783}{11661518761992000}\sqrt{-3} \right).$$

The point Y gives rise to the solution (as in (1))

$$\left(\frac{684469533791312783}{112919729369578740}\right)^3 + \left(-\frac{661146496267328783}{112919729369578740}\right)^3 = 22,$$

which is twice a Mordell-Weil generator $(17299/9954, 25469/9954)$.

4. Sylvester's conjecture, revisited

4.1. A theorem of Elkies: A breakthrough. We now return to the original question of Sylvester's conjecture. In 1994, Elkies announced the following result [**Elk94**], which remains unpublished.

THEOREM (Elkies). *If $p \equiv 4, 7 \pmod 9$, then $\#E_p(\mathbb{Q}) = \#E_{p^2}(\mathbb{Q}) = \infty$.*

The method of Elkies can be sketched as follows. Write $p = \pi\overline{\pi} \in \mathbb{Z}[\omega]$, where $\pi, \overline{\pi} \equiv 1 \pmod 3$. Elkies defines a modular curve X defined over K, and constructs an explicit modular parametrization

$$\Phi : X \to E_\pi : x^3 + y^3 = \pi$$

defined over K. He uses the map Φ to define a point on E_π over $K(\sqrt[3]{\overline{\pi}})$, and twists to get a point in $E_p(K)$.

4.2. Mock Heegner points, revisited. Using the strategy of mock Heegner points, we have re-proved the theorem under a further hypothesis on p.

THEOREM. *If $p \equiv 4, 7 \pmod 9$ and 3 is not a cube modulo p, then $\#E_p(\mathbb{Q}) = \#E_{p^2}(\mathbb{Q}) = \infty$.*

We remark that two-thirds of primes $p \equiv 4, 7 \pmod 9$ have the property that 3 is not a cube modulo p.

We only provide a sketch of the proof. Consider the modular parametrization $\Phi : X_0(243) \to E_9 : x^3 + y^3 = 9$; the curve $X_0(243) = X_0(3^5)$ has genus 19. The modular automorphism group of $X_0(243)$ is isomorphic to $\mathbb{Z}/3\mathbb{Z} \times S_3$, where the S_3 factor is generated by $\left(\begin{smallmatrix} 28 & 1/3 \\ -81 & 1 \end{smallmatrix}\right)$ and the Atkin-Lehner involution $w_{243} = \left(\begin{smallmatrix} 0 & -1 \\ 243 & 0 \end{smallmatrix}\right)$. The modular parametrization Φ is exactly the quotient of $X_0(243)$ by this S_3.

We start with a cyclic 243-isogeny of conductor $9p$, which yields a point $P \in E_9(H_{9p})$. One can descend the point $P \in E_9(H_{9p})$ with a twist by $\sqrt[3]{3}$ to a point $Q \in E_1(H_{3p})$. We next consider the trace $R = \mathrm{Tr}_{H_{3p}/L} Q \in E_1(L)$. We show that $R^\sigma = \omega R + T$ where $\sigma(\sqrt[3]{p}) = \sqrt[3]{p}$ and T is a 3-torsion point. Thus R yields a point

$Y \in E_{p^2}(K)$ by twisting. (This depends on the choice of P; another choice yields a point on $E_p(K)$.)

Unfortunately, there exist points $S \in E_1(K)_{\mathrm{tors}}$ that satisfy the equation $S^\sigma = S = \omega S + T$! Indeed, in certain cases the point R (equivalently, Y) is torsion; see section 4.4 below for a discussion of when we expect R to be torsion. To prove that the point R is nontorsion when 3 is not a cube modulo p, we instead consider the reduction of R modulo p. The prime p factors as $(\mathfrak{p}\overline{\mathfrak{p}})^3$ in L, so we consider the pair

$$(R \bmod \mathfrak{p}, R \bmod \overline{\mathfrak{p}}) \in (E_1)_{\mathbb{F}_\mathfrak{p}} \times (E_1)_{\mathbb{F}_{\overline{\mathfrak{p}}}} \cong E_1(\mathbb{F}_p)^2.$$

By an explicit computation with η-products, we are able to show that when 3 is not a cube modulo p, this reduction is not the image of any torsion point $S \in E_1(L)_{\mathrm{tors}}$.

4.3. Example. We illustrate our method with $p = 7$.

The isogeny $\langle 7\omega/9\rangle \to \langle (7\omega - 1)/27\rangle$ is a cyclic 243-isogeny with conductor 63, which yields a point $P = (x, y) \in E_9(H_{63})$ with

$$x^6 - 81x^3 + 5184 = 0, \quad y^6 + 63y^3 + 4536 = 0.$$

The twist $Q = (x, y) \in E_1(H_{21})$ has

$$x^2 + 3\omega^2 x + 4\omega = 0, \quad y^6 + 7y^3 + 56 = 0.$$

We again have $H_{21} = K(\alpha)$ where $\alpha^6 + 7 = 0$; we then recognize

$$Q = (\tfrac{1}{2}\omega^2\alpha^3 - \tfrac{3}{2}\omega^2, -\tfrac{1}{2}\alpha^4 + \tfrac{1}{2}\alpha)$$

to the precision computed. The trace $R = \mathrm{Tr}_{H_{21}/L} Q \in E_1(L)$ is then simply

$$R = (-\tfrac{3}{2}\sqrt[3]{7^2}, \tfrac{11}{2}\omega^2),$$

which yields the solution $Y = (11/3, -2/3)$, i.e.

$$\left(\frac{11}{3}\right)^3 + \left(\frac{-2}{3}\right)^3 = 7^2.$$

4.4. A Gross-Zagier formula. A direct naïve analogue of the Gross-Zagier formula in this case would state that

$$\hat{h}(Y) \doteq L'(E_9/K, \chi_{3p}, 1),$$

where $\chi_{3p} : \mathrm{Gal}(H_{3p}/K) \to \mu_3$ is the cubic character associated to the field $K(\sqrt[3]{3p})$. Since formally

$$L(E_9/K, \chi_{3p}, s) = L(E_p/\mathbb{Q}, s)L(E_{3p^2}/\mathbb{Q}, s),$$

this formula becomes

$$\hat{h}(Y) \doteq L'(E_p/\mathbb{Q}, 1)L(E_{3p^2}/\mathbb{Q}, 1).$$

When 3 is not a cube modulo p, one can prove that $\mathrm{rk}(E_{3p^2}(\mathbb{Q})) = 0$ (see [**Sat86**]), which motivates the fact that the point Y in our construction is nontorsion in this case. Furthermore, one can show that 3 is a cube modulo p if and only if either 3 divides $\#Ш(E_{3p^2}/\mathbb{Q})$ or $\mathrm{rk}(E_{3p^2}/\mathbb{Q}) > 0$; the order of this Tate-Shafarevich group is conjecturally the "algebraic part" of $L(E_{3p^2}/\mathbb{Q}, 1)$ when this value is non-zero. Thus the "naïve analogue of Gross-Zagier" combined with the BSD conjecture suggest the equivalence

$$Y \text{ is divisible by 3 in } E_p(K)/E_p(K)_{\mathrm{tors}} \iff 3 \text{ is a cube modulo } p.$$

The proof sketched in §4.2 yields the forward direction of this implication unconditionally. It should be possible to prove the converse as well, though we have not yet attempted to do so.

In our description of Satgé's construction with $p \equiv 2 \pmod 9$, we constructed a point on the cubic twist of E_2 by $\sqrt[3]{p}$, so a direct analogue of Gross-Zagier would yield

$$\hat{h}(Y) \doteq L'(E_{2p}/\mathbb{Q}, 1)L(E_{2p^2}/\mathbb{Q}, 1).$$

In this case one can prove that $\mathrm{rk}(E_{2p^2}(\mathbb{Q})) = 0$ and $3 \nmid \#Ш(E_{2p^2}/\mathbb{Q})$ without extra conditions. This provides intuition for why Satgé's construction produces points that are provably not divisible by 3 (in particular nontorsion) without any extra condition, whereas our result for $p \equiv 4, 7 \pmod 9$ requires an extra condition.

QUESTION. What is the precise statement of the Gross-Zagier formula in the cases when the Heegner hypothesis does not hold?

This is the subject of current research by Ben Howard at Boston College. Some aspect of this new formula (perhaps some extra Euler factors which sometimes trivially vanish) would have to account for various cases when the mock Heegner point is torsion even when the derivative of the L-function is not zero. Also, this formula would have to exhibit a dependence on the choice of CM point—the formula will in general not depend only on the conductor as in the classical Heegner case.

4.5. The case $p \equiv 8 \pmod 9$. What remains untouched by our discussion so far is the case $p \equiv 8 \pmod 9$ in Sylvester's conjecture. In this case, we may use the parametrization $\Phi : X_0(243) \to E_3$ and a cyclic isogeny of conductor $9p$, corresponding to a point $P \in E_3(H_{9p})$.

Adding a torsion point, the point P descends with a twist to a point $Q \in E_1(H_{3p})$, and a twisted trace $Y \in E_p(\mathbb{Q})$. Here, Gross-Zagier would imply that

$$\hat{h}(Y) \doteq L'(E_3/K, \chi_{9p}, 1) = L'(E_p/\mathbb{Q}, 1)L(E_{9p^2}/\mathbb{Q}, 1).$$

There seems to be no simple criterion for $L(E_{9p^2}/\mathbb{Q}, 1) \neq 0$, though one could hope to prove an analogue of the formulas of Rodriguez-Villegas and Zagier [**RVZ95**].

QUESTION. When $p \equiv 8 \pmod 9$, can one prove that the point Y is nontorsion when $L(E_{9p^2}/\mathbb{Q}, 1) \neq 0$, or perhaps at least when 3 does not divide the algebraic part of $L(E_{9p^2}/\mathbb{Q}, 1)$?

References

[Bal06] J. Balakrishnan, *CM constructions for elliptic curves*, 2006, Senior thesis, Harvard University.

[BS83] B. J. Birch and N. M. Stephens, *Heegner's construction of points on the curve $y^2 = x^3 - 1728e^3$*, Seminar on number theory, Paris 1981–82 (Paris, 1981/1982), Progr. Math., vol. 38, Birkhäuser Boston, Boston, MA, 1983, pp. 1–19. MR 729156 (85j:11062)

[Cow00] D. R. Coward, *Some sums of two rational cubes*, Q. J. Math. **51** (2000), no. 4, 451–464. MR 1806452 (2001k:11096)

[Dar04] H. Darmon, *Rational points on modular elliptic curves*, CBMS Regional Conference Series in Mathematics, vol. 101, Published for the Conference Board of the Mathematical Sciences, Washington, DC, 2004. MR 2020572 (2004k:11103)

[Elk94] N. D. Elkies, *Heegner point computations*, Algorithmic number theory (Ithaca, NY, 1994), Lecture Notes in Comput. Sci., vol. 877, Springer, Berlin, 1994, pp. 122–133. MR 1322717 (96f:11080)

[Kob02] S.-I. Kobayashi, *The local root number of elliptic curves*, Currents trends in number theory (Allahabad, 2000), Hindustan Book Agency, New Delhi, 2002, pp. 73–83. MR 1925642 (2004e:11060)

[Mon90] P. Monsky, *Mock Heegner points and congruent numbers*, Math. Z. **204** (1990), no. 1, 45–67. MR 1048066 (91e:11059)

[Ogg80] A. P. Ogg, *Modular functions*, The Santa Cruz Conference on Finite Groups (Univ. California, Santa Cruz, Calif., 1979), Proc. Sympos. Pure Math., vol. 37, Amer. Math. Soc., Providence, R.I., 1980, pp. 521–532. MR 604631 (82h:10037)

[RVZ95] F. Rodríguez Villegas and D. Zagier, *Which primes are sums of two cubes?*, Number theory (Halifax, NS, 1994), CMS Conf. Proc., vol. 15, Amer. Math. Soc., Providence, RI, 1995, pp. 295–306. MR 1353940 (96g:11049)

[Sat86] P. Satgé, *Groupes de Selmer et corps cubiques*, J. Number Theory **23** (1986), no. 3, 294–317. MR 846960 (87i:11070)

[Sat87] ———, *Un analogue du calcul de Heegner*, Invent. Math. **87** (1987), no. 2, 425–439. MR 870738 (88d:11057)

[Sel51] E. S. Selmer, *The Diophantine equation* $ax^3 + by^3 + cz^3 = 0$, Acta Math. **85** (1951), 203–362 (1 plate). MR 0041871 (13,13i)

[Shi71] G. Shimura, *Introduction to the arithmetic theory of automorphic functions*, Publications of the Mathematical Society of Japan, No. 11. Iwanami Shoten, Publishers, Tokyo, 1971, Kanô Memorial Lectures, No. 1. MR 0314766 (47 #3318)

[Sil86] J. H. Silverman, *The arithmetic of elliptic curves*, Graduate Texts in Mathematics, vol. 106, Springer-Verlag, New York, 1986. MR 817210 (87g:11070)

[Syl79a] J. J. Sylvester, *On Certain Ternary Cubic-Form Equations*, Amer. J. Math. **2** (1879), no. 3, 280–285. MR 1505225

[Syl79b] ———, *On Certain Ternary Cubic-Form Equations*, Amer. J. Math. **2** (1879), no. 4, 357–393. MR 1505237

[Syl80a] ———, *On Certain Ternary Cubic-Form Equations*, Amer. J. Math. **3** (1880), no. 1, 58–88. MR 1505246

[Syl80b] ———, *On Certain Ternary Cubic-Form Equations*, Amer. J. Math. **3** (1880), no. 2, 179–189. MR 1505257

Department of Mathematics, Harvard University, Cambridge, MA 02138

Current address: Mathematics Department, University of California, Santa Cruz, 194 Baskin Engineering, Santa Cruz, CA 95064

E-mail address: sdasgup2@ucsc.edu

Department of Mathematics and Statistics, University of Vermont, Burlington, VT 05401

E-mail address: jvoight@gmail.com

Clay Mathematics Proceedings
Volume 8, 2009

Shimura curve computations

John Voight

Abstract. We introduce Shimura curves first as Riemann surfaces and then as moduli spaces for certain abelian varieties. We give concrete examples of these curves and do some explicit computations with them.

1. Introduction: modular curves

We motivate the introduction of Shimura curves by first recalling the definition of modular curves.

For each $N \in \mathbb{Z}_{>0}$, we define the subgroup

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z}) : c \equiv 0 \pmod{N} \right\} \subset SL_2(\mathbb{Z}).$$

The group $\Gamma_0(N)$ acts on the completed upper half-plane $\mathfrak{H}^* = \mathfrak{H} \cup \mathbb{P}^1(\mathbb{R})$ by linear fractional transformations, and the quotient $X_0(N)_{\mathbb{C}} = \Gamma_0(N) \backslash \mathfrak{H}^*$ can be given the structure of a compact Riemann surface. The curve $X_0(N)_{\mathbb{C}}$ parametrizes cyclic N-isogenies between (generalized) elliptic curves and therefore has a model $X_0(N)_{\mathbb{Q}}$ defined over $\mathbb{Q}$. On $X_0(N)_{\mathbb{Q}}$, we also have *CM points*, which correspond to isogenies between elliptic curves which have complex multiplication (CM) by an imaginary quadratic field K.

Shimura curves arise in generalizing this construction from the matrix ring $M_2(\mathbb{Q})$ to certain quaternion algebras over totally real fields F. A Shimura curve is the quotient of the upper half-plane $\mathfrak{H}$ by a discrete, "arithmetic" subgroup of $\mathrm{Aut}(\mathfrak{H}) = PSL_2(\mathbb{R})$. Such a curve also admits a description as a moduli space, yielding a model defined over a number field, and similarly comes equipped with CM points.

The study of the classical modular curves has long proved rewarding for mathematicians both theoretically and computationally, and an expanding list of conjectures have been naturally generalized to the setting of Shimura curves. These curves, which although at first are only abstractly defined, can also be made very concrete.

In §2, we briefly review the relevant theory of quaternion algebras and then define Shimura curves as Riemann surfaces. In §3, we provide a detailed example

2000 *Mathematics Subject Classification.* Primary 11G18, Secondary 14G35.
Key words and phrases. Shimura curves, moduli spaces, triangle groups.

of a Shimura curve over $\mathbb{Q}$. In §4, we discuss the arithmetic of Shimura curves: we explain their interpretation as moduli spaces, and define CM points, Atkin-Lehner quotients, and level structure. Finally, in §5, we illustrate these concepts by considering the case of Shimura curves arising from triangle groups, in some sense the "simplest" class, and do some explicit computations with them.

2. Quaternion algebras and complex Shimura curves

2.1. Quaternion algebras. We refer to [**Vig80**] as a reference for this section.

As in the introduction, we look again at $SL_2(\mathbb{Z}) \subset M_2(\mathbb{Q})$: we have taken the group of elements of determinant 1 with integral entries in the $\mathbb{Q}$-algebra $M_2(\mathbb{Q})$. The algebras akin to $M_2(\mathbb{Q})$ are quaternion algebras.

Let F be a field with $\operatorname{char} F \neq 2$. A *quaternion algebra* over F is a central simple F-algebra of dimension 4. Equivalently, an F-algebra B is a quaternion algebra if and only if there exist $\alpha, \beta \in B$ which generate B as an F-algebra such that

$$\alpha^2 = a, \quad \beta^2 = b, \quad \beta\alpha = -\alpha\beta$$

for some $a, b \in F^*$. We denote this algebra by $B = \left(\frac{a,b}{F}\right)$.

Example. As examples of quaternion algebras, we have the ring of 2×2-matrices over F, or $M_2(F) \cong \left(\frac{1,1}{F}\right)$, and the division ring $\mathbb{H} = \left(\frac{-1,-1}{\mathbb{R}}\right)$ of Hamiltonians.

From now on, let B denote a quaternion algebra over F. There is a unique anti-involution $\overline{\ } : B \to B$, called *conjugation*, with the property that $\alpha\overline{\alpha} \in F$ for all $\alpha \in B$. The map $\operatorname{nrd}(\alpha) = \alpha\overline{\alpha}$ is known as the *reduced norm*.

Example. If $B = \left(\frac{a,b}{F}\right)$, and $\theta = x + y\alpha + z\beta + w\alpha\beta$, then

$$\overline{\theta} = x - y\alpha - z\beta - w\alpha\beta, \text{ and } \operatorname{nrd}(\theta) = x^2 - ay^2 - bz^2 + abw^2.$$

From now on, let F be a number field. Let v be a noncomplex place of F, and let F_v denote the completion of F at v. If $B_v = B \otimes_F F_v$ is a division ring, we say that B is *ramified* at v; otherwise $B_v \cong M_2(F_v)$ and we say B is *split* at v. The number of places v where B is ramified is finite and of even cardinality; their product is the *discriminant* $\operatorname{disc}(B)$ of B. Two quaternion algebras B, B' over F are isomorphic (as F-algebras) if and only if $\operatorname{disc}(B) = \operatorname{disc}(B')$.

Let $\mathbb{Z}_F$ denote the ring of integers of F. An *order* of B is a subring $\mathcal{O} \subset B$ (containing 1) which is a $\mathbb{Z}_F$-submodule satisfying $F\mathcal{O} = B$. A *maximal order* is an order which is maximal under inclusion. Maximal orders are not unique—but we mention that in our situation (where B has at least one unramified infinite place, see the next section), a maximal order in B is unique up to conjugation.

2.2. Shimura curves as Riemann surfaces. Let $\mathcal{O} \subset B$ be a maximal order. We then define the group analogous to $SL_2(\mathbb{Z})$, namely the group of units of $\mathcal{O}$ of norm 1:

$$\mathcal{O}_1^* = \{\gamma \in \mathcal{O} : \operatorname{nrd}(\gamma) = 1\}.$$

In order to obtain a discrete subgroup of $PSL_2(\mathbb{R})$ (see [**Kat92**, Theorem 5.3.4]), we insist that F is a totally real (number) field and that B is split at exactly one real place, so that

$$B \hookrightarrow B \otimes_{\mathbb{Q}} \mathbb{R} \cong M_2(\mathbb{R}) \times \mathbb{H}^{[F:\mathbb{Q}]-1}.$$

We denote by $\iota_\infty : B \hookrightarrow M_2(\mathbb{R})$ the projection onto the first factor.

We then define the group

$$\Gamma^B(1) = \iota_\infty(\mathcal{O}_1^*/\{\pm 1\}) \subset PSL_2(\mathbb{R}).$$

The quotient $X^B(1)_{\mathbb{C}} = \Gamma^B(1)\backslash\mathfrak{H}$ can be given the structure of a Riemann surface [**Kat92**, §5.2] and is known as a *Shimura curve*.

From now on, we assume that $B \ncong M_2(\mathbb{Q})$, so that we avoid the (classical) case of modular curves; it then follows that B is a division ring and, unlike the case for modular curves, the Riemann surface $X^B(1)_{\mathbb{C}}$ is already compact [**Kat92**, Theorem 5.4.1].

3. Example

We now make this theory concrete by considering an extended example.

We take $F = \mathbb{Q}$ and the quaternion algebra B over $\mathbb{Q}$ with $\mathrm{disc}(B) = 6$, i.e. B is ramified at the primes 2 and 3, and unramified at all other places, including ∞. Explicitly, we may take $B = \left(\frac{-1,3}{\mathbb{Q}}\right)$, so that $\alpha, \beta \in B$ satisfy

$$\alpha^2 = -1, \quad \beta^2 = 3, \quad \beta\alpha = -\alpha\beta.$$

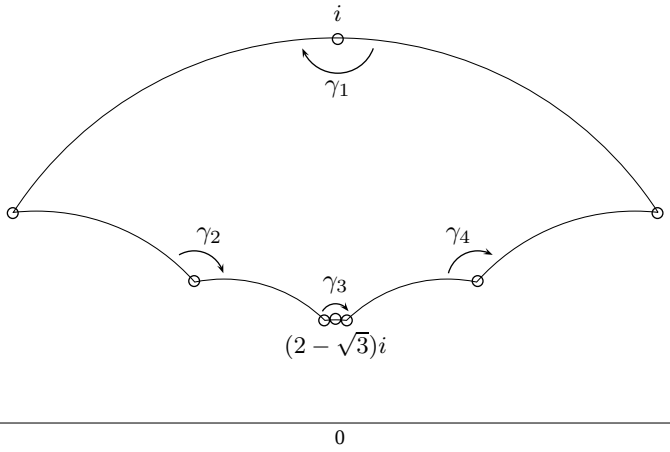
We find the maximal order

$$\mathcal{O} = \mathbb{Z} \oplus \mathbb{Z}\alpha \oplus \mathbb{Z}\beta \oplus \mathbb{Z}\delta \text{ where } \delta = (1 + \alpha + \beta + \alpha\beta)/2,$$

and we have an embedding

$$\begin{aligned} \iota_\infty : B &\to M_2(\mathbb{R}) \\ \alpha, \beta &\mapsto \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} \sqrt{3} & 0 \\ 0 & -\sqrt{3} \end{pmatrix}. \end{aligned}$$

With respect to this embedding, we compute a fundamental domain D for the action of $\Gamma^B(1) = \iota_\infty(\mathcal{O}_1^*/\{\pm 1\})$ as follows. (For an alternate presentation, see [**AB04**, §5.5.2] or [**KV03**, §5.1].)



The elements

$$\gamma_1 = \alpha,\ \gamma_2 = \alpha + \delta,\ \gamma_3 = 2\alpha + \alpha\beta,\ \gamma_4 = 1 + \alpha - \beta + \delta$$

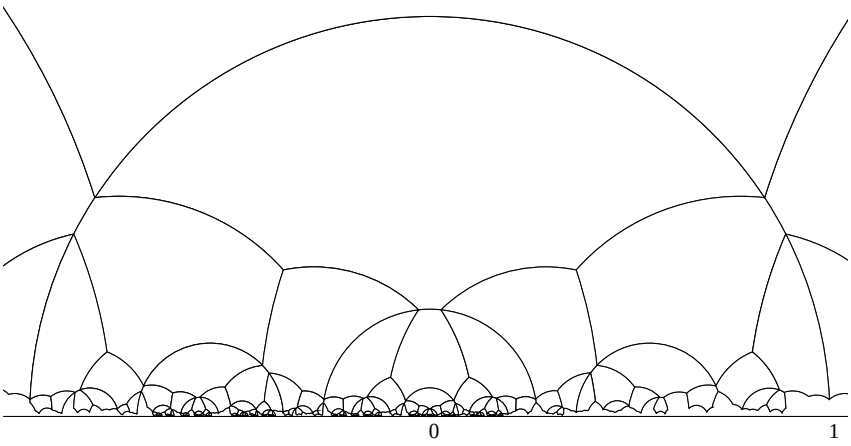
are known as *side-pairing elements*; they yield the presentation

$$\Gamma^B(1) \cong \langle \gamma_1, \dots, \gamma_4 \mid \gamma_1^2 = \gamma_2^3 = \gamma_3^2 = \gamma_4^3 = \gamma_4\gamma_3\gamma_2\gamma_1 = 1 \rangle.$$

One can compute the area $\mu(D)$ of the above fundamental domain D by triangulation, but we also have the formula (see [**Elk98**, §2.2])

$$\mu(D) = \mu(X^B(1)) = \frac{\pi}{3} \prod_{p \mid \mathrm{disc}(B)} (p-1) = \frac{2\pi}{3}.$$

The group $\Gamma^B(1)$ then tessellates $\mathfrak{H}$ as follows.



(The algorithm for drawing hyperbolic polygons is due to Verrill [**Ver06**].)

The genus g of X can be computed by the Riemann-Hurwitz formula as

$$2g - 2 = \frac{\mu(X^B(1))}{2\pi} - \sum_q e_q \left(1 - \frac{1}{q}\right),$$

where e_q is the number of (conjugacy classes of) elliptic points of order q. From the presentation for $\Gamma^B(1)$ above, we can see directly that $e_2 = e_3 = 2$ and hence

$$2g - 2 = 1/3 - 2(1 - 1/2) - 2(1 - 1/3) = -2$$

so $g = 0$. Alternatively, we can compute the number of these elements by the formulas

$$e_2 = \prod_{p \mid \mathrm{disc}(B)} \left(1 - \left(\frac{-4}{p}\right)\right) = 2, \quad e_3 = \prod_{p \mid \mathrm{disc}(B)} \left(1 - \left(\frac{-3}{p}\right)\right) = 2.$$

Since the genus of X is zero, we have a map $X^B(1)_{\mathbb{C}} \to \mathbb{P}^1_{\mathbb{C}}$.

4. Arithmetic of Shimura curves

4.1. Shimura curves as moduli spaces. Just as with modular curves, Shimura curves are in fact moduli spaces, and this moduli description yields a model for $X^B(1)_{\mathbb{C}}$ which is defined over a number field.

In the case $F = \mathbb{Q}$, the curve $X^B(1)$ is a coarse moduli space for pairs (A, ι), where:

- A is an abelian surface, and
- $\iota : \mathcal{O} \hookrightarrow \mathrm{End}(A)$ is an embedding.

We say that such an A has *quaternionic multiplication* (QM) *by* $\mathcal{O}$. The involution $\bar{}$ on $\mathcal{O}$ induces via ι an involution on $\mathrm{End}(A)$, and there is a unique principal polarization on A which is compatible with this involution, then identified with the Rosati involution.

If $F \neq \mathbb{Q}$, the moduli description is more complicated: since B is then neither totally definite nor totally indefinite, it follows from the classification of endomorphism algebras of abelian varieties over $\mathbb{C}$ (see [**Mum70**, Theorem 21.3]) that we cannot have $\mathrm{End}(A) \otimes_{\mathbb{Z}} \mathbb{Q} \cong B$. Instead, one must choose an imaginary quadratic extension K of F, as in [**Zha01**, §1.1.2], and consider a moduli problem over K. For simplicity, we assume from now on that F has narrow class number 1: under this hypothesis, we have a natural choice, namely $K = F(\sqrt{-d})$, where d is a totally positive generator for the discriminant $\mathrm{disc}(B)$. One may then think of the objects parametrized by a Shimura curve $X^B(1)_F$ as "abelian varieties with QM by $\mathcal{O}$"—the precise meaning of this phrase will be neglected here.

It then follows from this moduli description that there exists a *canonical model* $X^B(1)_F$ for $X^B(1)_{\mathbb{C}}$ defined over F, a theorem due to Shimura [**Shi67**] and Deligne [**Del71**].

4.2. Example: Models. The model $X^B(1)_{\mathbb{Q}}$ over $\mathbb{Q}$ for our Shimura curve with $\mathrm{disc}(B) = 6$ is given by the conic

$$X^B(1)_{\mathbb{Q}} : x^2 + y^2 + 3z^2 = 0,$$

a result attributed to Ihara [**Kur79**, p. 279].

This identification can be made quite explicit, a computation due to Baba-Granath [**BG08**]. For $k \in \mathbb{Z}_{\geq 0}$, we denote by $M_k(\Gamma)$ the space of holomorphic weight k *modular forms* for the group $\Gamma = \Gamma^B(1)$, namely, the space of holomorphic maps $f : \mathfrak{H} \to \mathbb{C}$ such that

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^k f(z)$$

for all $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$. Using an elementary formula due to Shimura, we compute the dimension of $M_k(\Gamma)$:

$$\dim_{\mathbb{C}} M_4(\Gamma) = \dim_{\mathbb{C}} M_6(\Gamma) = 1, \quad \dim_{\mathbb{C}} M_{12}(\Gamma) = 3.$$

From this, one can show that there exist normalized $h_k \in M_k(\Gamma)$ for $k = 4, 6, 12$ such that

$$h_{12}^2 + 3h_6^4 + h_4^6 = 0,$$

which realizes the map $X^B(1)_{\mathbb{C}} \to X^B(1)_{\mathbb{Q}}$.

4.3. CM points. On the modular curves $X_0(N)$, we have CM points arising from elliptic curves with extra endomorphisms. These points are defined over ring class extensions H of an imaginary quadratic field K, and the Shimura reciprocity law describes explicitly the action of $\mathrm{Gal}(H/K)$ on them. In a similar way, on the Shimura curve $X^B(1)$ we have *CM points* which correspond to abelian varieties with extra endomorphisms. Let $K \supset F$ be a totally imaginary quadratic extension which *splits* B, i.e. $B \otimes_F K \cong M_2(K)$; the field K splits B if and only if there exists an embedding $\iota_K : K \hookrightarrow B$, and the map ι_K is concretely given by an element $\mu \in \mathcal{O}$ such that $\mathbb{Z}_F[\mu] = \mathbb{Z}_K$. Let $z = z_D$ be the fixed point of $\iota_\infty(\mu)$ in $\mathfrak{H}$; we then say z is a *CM point* on $X^B(1)_{\mathbb{C}}$. When $F = \mathbb{Q}$, CM points on $X^B(1)$ correspond to abelian surfaces A with endomorphism algebra $\mathrm{End}(A) \otimes_{\mathbb{Z}} \mathbb{Q} \cong M_2(K)$; the interpretation is again more subtle when $F \neq \mathbb{Q}$, but there one may think of these points as similarly having "extra endomorphisms".

On the model $X^B(1)_F$, these points are defined over the Hilbert class field H of K (or more generally, ring class extensions), and one has also a Shimura reciprocity law; see [**Shi67**] for a discussion and proof.

4.4. Example: CM points. The following computation can be found in Elkies [**Elk98**, §3.4] and Baba-Granath [**BG08**, §3.3].

We return to the example from §2, with $F = \mathbb{Q}$. Let $K = \mathbb{Q}(\sqrt{-19})$, and $\mathbb{Z}_K = \mathbb{Z}[(1+\sqrt{-19})/2]$. We have $\#\,\mathrm{Cl}(\mathbb{Z}_K) = 1$, and the elliptic curve $E = \mathbb{C}/\mathbb{Z}_K$ with CM by $\mathbb{Z}_K$ has j-invariant -96^3.

The genus 2 curve C defined by

$$C : y^2 = 2t^6 - 3(1+9\sqrt{-19})t^4 - 3(1-9\sqrt{-19})t^2 + 2$$

has Jacobian $J(C) \cong E \times E$, and $\mathrm{End}(J(C)) \cong M_2(\mathbb{Z}_K)$. This curve C "corresponds" to the moduli point $[C] = (32 : 27 : 13\sqrt{-19})$ on the Shimura curve $X^B(1) : x^2+3y^2+z^2 = 0$. (The field of moduli of the point $[C]$ is $\mathbb{Q}$, but $\mathbb{Q}$ is not a field of definition for C; the automorphism group of C is $\mathrm{Aut}(C) \cong \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$.)

4.5. Atkin-Lehner involutions. Shimura curves also possess natural involutions, just like modular curves. The normalizer

$$N(\mathcal{O}) = \{\alpha \in B^*/F^* : \alpha\mathcal{O} = \mathcal{O}\alpha,\ \mathrm{nrd}(\alpha) \text{ is totally positive}\}$$

acts via ι_∞ as automorphisms of $X^B(1)_F$, and generates a subgroup

$$W \cong \prod_{\mathfrak{p}|\mathrm{disc}(B)} \mathbb{Z}/2\mathbb{Z} = (\mathbb{Z}/2\mathbb{Z})^e.$$

The elements of W are known as *Atkin-Lehner involutions*. Letting $\Gamma^{B*}(1) = \iota_\infty(N(\mathcal{O}))$, we see that the curve $X^{B*}(1) = \Gamma^{B*}(1)\backslash\mathfrak{H}$ is the quotient of $X^B(1)$ by W.

When $F = \mathbb{Q}$, these involutions have a natural moduli interpretation. Recall that the curve $X^B(1)$ parametrizes pairs (A, ι), where A is an abelian surface (over $\mathbb{C}$, say) with QM by $\mathcal{O}$ specified by an embedding $\iota : \mathcal{O} \hookrightarrow \mathrm{End}(A)$. But there may be more than one such embedding ι for a given A, even up to isomorphism: for each divisor $\mathfrak{m} \mid \mathrm{disc}(B)$, we can "twist" ι by $\mathfrak{m}$ to obtain a new pair $(A, \iota^{\mathfrak{m}})$. All such twists arise in this way (see [**Rot04**, §3]), and therefore the quotient $X^{B*}(1)$ of $X^B(1)$ by W parametrizes abelian surfaces A which can be given the structure ι of QM by $\mathcal{O}$, without a particular choice of ι.

4.6. Example: Atkin-Lehner quotient. The two Atkin-Lehner involutions w_2, w_3 act on $X^B(1)_{\mathbb{Q}} : x^2 + y^2 + 3z^2 = 0$ by

$$w_2(x : y : z) = (x : -y : z), \quad w_3(x : y : z) = (-x : y : z).$$

The quotients are therefore

$$\begin{array}{ccc} X & \longrightarrow & X^{\langle w_2\rangle} = \mathbb{P}^1 \\ (x : y : z) & \longmapsto & (x : z) \end{array} \qquad \begin{array}{ccc} X & \longrightarrow & X^{\langle w_3\rangle} = \mathbb{P}^1 \\ (x : y : z) & \longmapsto & (y : z). \end{array}$$

and the quotient by the full group $W = \langle w_2, w_3\rangle$ can be given by

$$\begin{array}{ccc} j : X & \longrightarrow & X^W = \mathbb{P}^1 \\ (x : y : z) & \longmapsto & (16y^2 : 9x^2), \end{array}$$

under our normalization. Our moduli point $[C]$ corresponding to K with discriminant -19 was $[C] = (32 : 27 : 13\sqrt{-19})$, and so we find $j([C]) = 81/64 = 3^4/2^6$.

4.7. Level structure: congruence subgroups. Having defined the group $\Gamma^B(1)$ which replaces $PSL_2(\mathbb{Z})$, we now introduce the curves analogous to the modular curves. Let $\mathfrak{N}$ be an ideal of $\mathbb{Z}_F$ that is coprime to the discriminant of B, and let $\mathbb{Z}_{F,\mathfrak{N}}$ be the completion of $\mathbb{Z}_F$ at $\mathfrak{N}$; then there exists an embedding

$$\iota_{\mathfrak{N}} : \mathcal{O} \hookrightarrow \mathcal{O} \otimes_{\mathbb{Z}_F} \mathbb{Z}_{F,\mathfrak{N}} \cong M_2(\mathbb{Z}_{F,\mathfrak{N}}).$$

We define

$$\Gamma_0^B(\mathfrak{N}) = \{\iota_\infty(\gamma) : \gamma \in \mathcal{O}_1^*,\ \iota_{\mathfrak{N}}(\gamma) \text{ is upper triangular modulo } \mathfrak{N}\}/\{\pm 1\}$$

and we again obtain a Riemann surface $X_0^B(\mathfrak{N})_{\mathbb{C}} = \Gamma_0^B(\mathfrak{N})\backslash\mathfrak{H}$.

In a similar way, for $F = \mathbb{Q}$, the curves $X_0^B(N)_{\mathbb{C}}$ parametrize cyclic N-isogenies between abelian surfaces with QM by $\mathcal{O}$. For any F, one can also show that the curve $X_0^B(\mathfrak{N})_{\mathbb{C}}$ admits a model over a number field.

5. Triangle groups

5.1. The $(2, 4, 6)$-triangle group. Recall from §4.5 that the group

$$\Gamma^{B*}(1) = \{\iota_\infty(\alpha) : \alpha \in B^*/F^*,\ \alpha\mathcal{O} = \mathcal{O}\alpha,\ \mathrm{nrd}(\alpha) \text{ is totally positive}\}$$

realizes the space $X^{B*}(1) = \Gamma^{B*}(1)\backslash\mathfrak{H}$. The quotient

$$\frac{\Gamma^{B*}(1)}{\Gamma^B(1)} \cong \prod_{p|\mathrm{disc}(B)} \mathbb{Z}/2\mathbb{Z},$$

arises from elements whose reduced norm divides $\mathrm{disc}(B) = 6$.

We can see the group $\Gamma^{B*}(1)$ again explicitly: it has a presentation

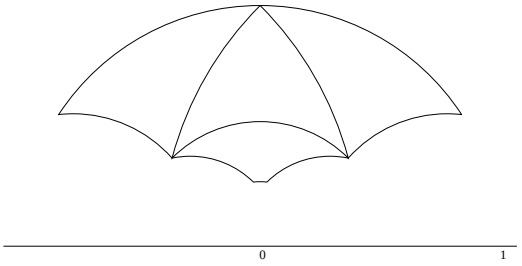
$$\Gamma^{B*}(1) \cong \langle s_2, s_4, s_6 \mid s_2^2 = s_4^4 = s_6^6 = s_6 s_4 s_2 = 1\rangle$$

where

$$s_2 = -1 + 2\alpha - \beta + 2\delta,\ s_4 = -1 + \alpha,\ s_6 = -2 + \alpha + \delta$$

have $\mathrm{nrd}(s_2) = 6$, $\mathrm{nrd}(s_4) = 2$, $\mathrm{nrd}(s_6) = 3$, respectively. This group $\Gamma^{B*}(1)$ is known as a $(2, 4, 6)$-*triangle group*; a fundamental domain D for $\Gamma^{B*}(1)$ is the union of a *fundamental triangle*, a hyperbolic triangle with angles $\pi/2, \pi/4, \pi/6$ with vertices at the fixed points of s_2, s_4, s_6, respectively, together with its image in the reflection in the geodesic connecting any two of the vertices.

We can visualize the $(2, 4, 6)$-triangle group $\Gamma^{B*}(1)$ inside $\Gamma^B(1)$ as follows.



5.2. Cocompact arithmetic triangle groups. More generally, for $p, q, r \in \mathbb{Z}_{\geq 2}$ with $1/p + 1/q + 1/r < 1$, we may define the (p, q, r)-triangle group similarly as the group with presentation

$$\langle s_p, s_q, s_r \mid s_p^p = s_q^q = s_r^r = s_r s_q s_p = 1 \rangle.$$

By work of Takeuchi [**Tak77**], there are exactly 18 quaternion algebras B (up to isomorphism), defined over one of 13 totally real fields F, that give rise to such a *cocompact arithmetic triangle group* $\Gamma^{B*}(1)$. Already these contain a number of curves worthwhile of study. (In this light, we could consider the classical $SL_2(\mathbb{Z})$ to be a $(2, 3, \infty)$-triangle group, though we still exclude this case in our discussion.)

Each of these "simplest" Shimura curves has genus zero, so we have a map $j : X^{B*}(1) \to \mathbb{P}^1_{\mathbb{C}}$. (In fact, one can show that the canonical model provided by Shimura and Deligne for $X^{B*}(1)_{\mathbb{C}}$ over F is already $\mathbb{P}^1_F$.) We normalize this map by taking the images of the elliptic fixed points z_p, z_q, z_r of s_p, s_q, s_r, respectively, to be $0, 1, \infty$.

5.3. Explicit computation of CM points. To summarize, from cocompact arithmetic triangle groups associated with certain quaternion algebras B over totally real fields F we obtain Riemann surfaces $X^{B*}(1)$ of genus 0 together with a map $j : X^{B*}(1) \to \mathbb{P}^1_{\mathbb{C}}$. There are CM points of arithmetic interest which we would like to compute.

THEOREM ([**Voi06**]). *There exists an algorithm that, given a totally imaginary quadratic field $K \supset F$, computes the CM point $j(z) \in \mathbb{P}^1(\mathbb{C})$ associated to K to arbitrary precision, as well as all of its conjugates by the group* $\mathrm{Gal}(H/K)$.

One can then recognize the value j as an algebraic number by considering the polynomial defined by its conjugates.

5.4. Second example. We now give an example where $F \neq \mathbb{Q}$. Let F be the totally real subfield of $\mathbb{Q}(\zeta_9)$, where ζ_9 is a primitive ninth root of unity. We have $\mathbb{Z}_F = \mathbb{Z}[b]$, where $b = -(\zeta_9 + 1/\zeta_9)$. We take $B = \left(\frac{-3, b}{F}\right)$, i.e. B is generated by α, β with

$$\alpha^2 = -3, \quad \beta^2 = b, \quad \beta\alpha = -\alpha\beta.$$

Here, we have $\mathrm{disc}(B) = \mathbb{Z}_F$, i.e. B is ramified at no finite place and at exactly two of the three real places. We fix the isomorphism $\iota_\infty : B \otimes_F \mathbb{R} \xrightarrow{\sim} M_2(\mathbb{R})$, given explicitly as

$$\alpha \mapsto \begin{pmatrix} 0 & 3 \\ -1 & 0 \end{pmatrix}, \qquad \beta \mapsto \begin{pmatrix} \sqrt{b} & 0 \\ 0 & -\sqrt{b} \end{pmatrix}.$$

We next compute a maximal order $\mathcal{O} = \mathbb{Z}_F \oplus \mathbb{Z}_F\zeta \oplus \mathbb{Z}_F\eta \oplus \mathbb{Z}_F\omega$, where

$$\begin{aligned}\zeta &= -\tfrac{1}{2}b + \tfrac{1}{6}(2b^2 - b - 4)\alpha \\ \eta &= -\tfrac{1}{2}b\beta + \tfrac{1}{6}(2b^2 - b - 4)\alpha\beta \\ \omega &= -b + \tfrac{1}{3}(b^2 - 1)\alpha - b\beta + \tfrac{1}{3}(b^2 - 1)\alpha\beta.\end{aligned}$$

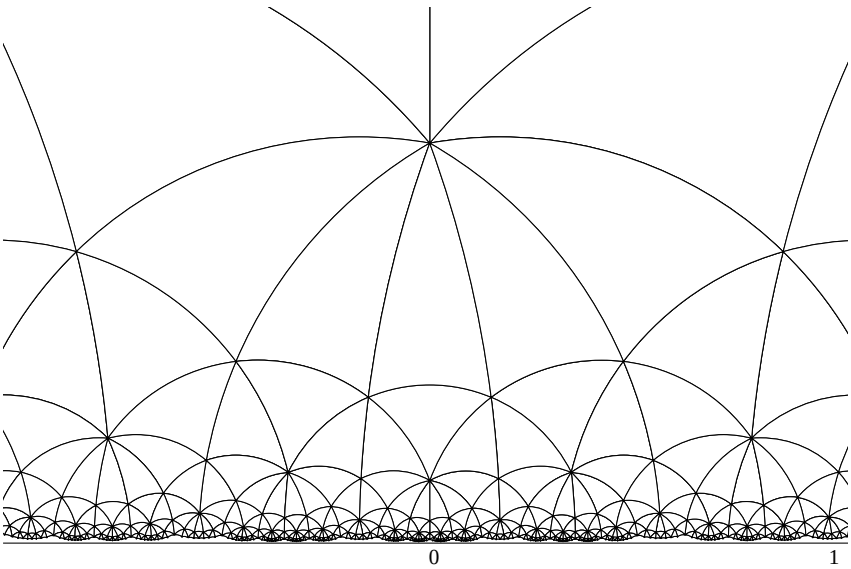
By work of Takeuchi [**Tak77**], we know that $\Gamma^B(1) = \Gamma^{B*}(1)$ is a triangle group with signature $(p, q, r) = (2, 3, 9)$. Explicitly, we find the generators

$$s_2 = b + \omega - 2\eta, s_3 = -1 + (b^2 - 3)\zeta + (-2b^2 + 6)\omega + (b^2 + b - 3)\eta, s_9 = -\zeta$$

which satisfy the relations $s_2^2 = s_3^3 = s_9^9 = s_2s_3s_9 = 1$. The fixed points of these elements are

$$z_2 = 0.395526\ldots i,\ z_3 = -0.153515\ldots + 0.364518\ldots i,\ z_9 = i,$$

and they form the vertices of a fundamental triangle.



Each triangle in the above figure is a fundamental domain formed by the union of two such fundamental triangles.

5.5. CM points. As an example, we first take $K = F(\sqrt{-2})$ with class number 3. We find $\mu \in \mathcal{O}$ satisfying $\mu^2 + 2 = 0$, so $\mathbb{Z}_F[\mu] = \mathbb{Z}_K$ has discriminant -8; explicitly,

$$\mu = (-b^2 - b + 1) + (-2b^2 + 2)\zeta + (2b^2 - b - 5)\omega + (-b^2 + b + 1)\eta.$$

We obtain the CM point $j(z) = 17137.9737\ldots$ as well as its Galois conjugates $0.5834\ldots \pm 0.4516\ldots i$, which yields the minimal polynomial for $j = j(z)$

$$j^3 - \tfrac{1096905}{64}j^2 + \tfrac{41938476081}{2097152}j - \tfrac{9781803409}{1048576} = 0$$

to the precision computed (300 digits). Note that

$$\tfrac{9781803409}{1048576} = \tfrac{7^2 71^2 199^2}{2^{20}}.$$

We verify that $K(j) = H = K(c)$, where $c^3 - 3c + 10 = 0$.

Larger examples can be computed, including over ring class extensions. Consider the field $K = F(\sqrt{-5})$ with discriminant $\operatorname{disc}(K/F) = -20$. We consider the order $\mathbb{Z}_{K,f} \subset K$ of conductor $f = b - 1$; note that $N_{F/\mathbb{Q}}(b-1) = 3$.

The CM point z has $j = j(z)$ which satisfies a polynomial of degree $14 = \#\operatorname{Cl}(\mathbb{Z}_{K,f})$, with $N(j)$ equal to

$$-\frac{71^8 127^8 163^4 179^2 487^4 971^2 1619^2 2591^2 2699^2 7451^2 10079^2 13859^2 17099^2}{2^{84} 5^9 89^9 269^9 719^9}.$$

The extension $K(j) = K(c)$ is generated by an element c which satisfies

$$c^{14} - c^{13} - 2c^{12} + 19c^{11} - 37c^{10} - 122c^9 + 251c^8 + 211c^7 - 589c^6 + 470c^5 - 41c^4 - 73c^3 + 22c^2 + 11c + 1 = 0.$$

References

[AB04] M. Alsina and P. Bayer, *Quaternion orders, quadratic forms, and Shimura curves*, CRM Monograph Series, vol. 22, American Mathematical Society, Providence, RI, 2004. MR 2038122 (2005k:11226)

[BG08] S. Baba and H. Granath, *Genus 2 curves with quaternionic multiplication*, Canad. J. Math. **60** (2008), no. 4, 734–757. MR 2423455

[Del71] P. Deligne, *Travaux de Shimura*, Séminaire Bourbaki, 23ème année (1970/71), Exp. No. 389, Lecture Notes in Math., vol. 244, Springer, Berlin, 1971, pp. 123–165. MR 0498581 (58 #16675)

[Elk98] N. D. Elkies, *Shimura curve computations*, Algorithmic number theory (Portland, OR, 1998), Lecture Notes in Comput. Sci., vol. 1423, Springer, Berlin, 1998, proceedings of ANTS-III, 1998, pp. 1–47. MR 1726059 (2001a:11099)

[Kat92] S. Katok, *Fuchsian groups*, Chicago Lectures in Mathematics, University of Chicago Press, Chicago, IL, 1992. MR 1177168 (93d:20088)

[Kur79] A. Kurihara, *On some examples of equations defining Shimura curves and the Mumford uniformization*, J. Fac. Sci. Univ. Tokyo Sect. IA Math. **25** (1979), no. 3, 277–300. MR 523989 (80e:14010)

[KV03] D. R. Kohel and H. A. Verrill, *Fundamental domains for Shimura curves*, J. Théor. Nombres Bordeaux **15** (2003), no. 1, 205–222, Les XXIIèmes Journées Arithmetiques (Lille, 2001). MR 2019012 (2004k:11096)

[Mum70] D. Mumford, *Abelian varieties*, Tata Institute of Fundamental Research Studies in Mathematics, No. 5, Published for the Tata Institute of Fundamental Research, Bombay, 1970. MR 0282985 (44 #219)

[Rot04] V. Rotger, *Modular Shimura varieties and forgetful maps*, Trans. Amer. Math. Soc. **356** (2004), no. 4, 1535–1550 (electronic). MR 2034317 (2005a:11086)

[Shi67] G. Shimura, *Construction of class fields and zeta functions of algebraic curves*, Ann. of Math. (2) **85** (1967), 58–159. MR 0204426 (34 #4268)

[Tak77] K. Takeuchi, *Arithmetic triangle groups*, J. Math. Soc. Japan **29** (1977), no. 1, 91–106. MR 0429744 (55 #2754)

[Ver06] H. Verrill, *Subgroups of $PSL_2(\mathbb{R})$*, Handbook of Magma functions, vol. V, Sydney, July 2006, edited by J. Cannon and W. Bosma, pp. 1117–1138.

[Vig80] M.-F. Vignéras, *Arithmétique des algèbres de quaternions*, Lecture Notes in Mathematics, vol. 800, Springer, Berlin, 1980. MR 580949 (82i:12016)

[Voi06] J. Voight, *Computing CM points on Shimura curves arising from cocompact arithmetic triangle groups*, Algorithmic number theory, Lecture Notes in Comput. Sci., vol. 4076, Springer, Berlin, 2006, proceedings of ANTS-VII, Berlin, 2006, pp. 406–420. MR 2282939 (2008g:11104)

[Zha01] S.-W. Zhang, *Heights of Heegner points on Shimura curves*, Ann. of Math. (2) **153** (2001), no. 1, 27–147. MR MR1826411 (2002g:11081)

Department of Mathematics and Statistics, University of Vermont, Burlington, VT 05401

E-mail address: jvoight@gmail.com

Clay Mathematics Proceedings
Volume 8, 2009

Computing Heegner points arising from Shimura curve parametrizations

Matthew Greenberg

Abstract. Let E be an elliptic curve defined over $\mathbb{Q}$ or over a real quadratic field which is uniformized by the Jacobian of a Shimura curve X. We discuss a p-adic analytic algorithm for computing certain *Heegner points* on E – images under the above uniformization of degree zero CM-divisors on X.

1. Heegner points

1.1. Modular parametrizations. Let $E/\mathbb{Q}$ be an elliptic curve of conductor N. By the modularity theorem of Wiles et. al., we have a holomorphic *modular parametrization*

$$\Phi_N : X_0(N)(\mathbb{C}) \longrightarrow E(\mathbb{C}),$$

where the Riemann surface $X_0(N)(\mathbb{C})$ is the quotient of the extended complex upper half-plane $\mathcal{H}$ by the standard congruence subgroup $\Gamma_0(N)$ of level N. Assume that $\Phi_N(\infty)$ is the zero element of $E(\mathbb{C})$. Let $P \in X_0(N)(\mathbb{C})$ and let $\tau \in \mathcal{H}$ be any lift of P. Then

$$\Phi_N(P) = W\Big(\int_\infty^\tau 2\pi i f_E(z)dz\Big) = W\Big(\sum_{n\geq 1} \frac{a_n(f_E)}{n} e^{2\pi i n\tau}\Big) \tag{1.1}$$

where W is the Weierstrass parametrization of E, $f_E \in \mathcal{S}_2(N)$ is the normalized newform attached to E and $a_n(f_E)$ is the n-th Fourier coefficient of f_E.

1.2. CM-points. For the purposes of this talk, a *quadratic order* (resp. an *imaginary quadratic order*) $\mathcal{O}$ is a subring of a quadratic number field (resp. an imaginary quadratic number field) K such that $K = \mathbb{Q}\mathcal{O}$.

The Riemann surface $X_0(N)(\mathbb{C})$ may be identified with the complex-valued points of a curve $X_0(N)$ defined over $\mathbb{Q}$. This curve is in fact a moduli space — $X_0(N)$ classifies isogenies $P = (A \to A')$ of elliptic curves whose kernel is cyclic of order N. We will call a point $P \in X_0(N)(\mathbb{C})$ a *CM-point*, and say that P has CM

2000 *Mathematics Subject Classification.* Primary 11G05, Secondary 11F11, 11F85, 11G15, 11G18.

Key words and phrases. elliptic curves, modular forms, Heegner points, Shimura curves, p-adic uniformization.

by the quadratic order $\mathcal{O}$, if both A and A' have CM by $\mathcal{O}$. In this case, the theory of complex multiplication says that

$$P \in X_0(N)(H_{\mathcal{O}}), \quad \text{where } H_{\mathcal{O}} = \text{ring class field attached to } \mathcal{O}.$$

1.3. The classical Heegner hypothesis. Let $\mathcal{O} \subset K$ be an imaginary quadratic order of discriminant prime to N.

LEMMA 1.1 ([**Dar04**, Proposition 3.8]). *The following are equivalent:*

(1) *There exists a point on $X_0(N)$ with CM by $\mathcal{O}$.*
(2) *All primes ℓ dividing N split in K.*

Conditions (1) and (2) are know as the *Heegner hypothesis*. Thus, when the Heegner hypothesis is satisfied, the above construction yields a systematic supply of algebraic points on E defined over specific class fields of K. Due to the importance of Heegner points to the arithmetic theory of elliptic curves (see [**Dar**] or [**Dar04**, Chapter 3]), it is natural to desire an analogous construction of algebraic points defined over class fields of imaginary quadratic fields which do not necessarily satisfy this stringent hypothesis, as well as methods to compute such points in practice. Such a generalization requires admitting uniformizations of E by certain *Shimura curves*.

2. Shimura-Heegner points

2.1. Shimura curve parametrizations (over $\mathbb{Q}$). Assume that N is square-free and let $N = N^+N^-$ be a factorization of N such that N^- has an even number of prime factors. Let C be the unique quaternion algebra over $\mathbb{Q}$ ramified precisely at N^-. (For basic definitions related to quaternion algebras, see [**Voi**, §1.2] or the comprehensive [**Vig80**].) Fix an identification ι_∞ of $C \otimes_{\mathbb{Q}} \mathbb{R}$ with $M_2(\mathbb{R})$. Let S be an Eichler order in C of level N^+ and set

$$\Gamma^C(S) = \{\iota_\infty(s) : s \in S, \det \iota_\infty(s) = 1\}/\{\pm 1\} \subset \mathrm{PSL}_2(\mathbb{R}).$$

The group $\Gamma^C(S)$ acts discontinuously on $\mathcal{H}$ with quotient denoted $X^C(S)(\mathbb{C})$.

EXAMPLE 2.1. If $N^- = 1$, then $C \cong M_2(\mathbb{Q})$ and S may be taken to be

$$R_0(N) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbb{Z}) : N^+ \text{ divides } c \right\}.$$

In this case, the group $\Gamma^C(S)$ is the usual congruence subgroup $\Gamma_0(N)$. It is known that $X^C(S)(\mathbb{C})$ is compact if and only if $N^- \neq 1$.

EXAMPLE 2.2. Suppose $p \equiv 3 \pmod 4$, $N^+ = 1$ and $N^- = 2p$. Then the quaternion algebra C is that which Voight denotes $\left(\dfrac{-1, p}{\mathbb{Q}}\right)$ in [**Voi**]. The Eichler order S is simply a maximal order in C and is unique up to conjugation by C^*.

A space of modular forms $\mathcal{S}_2(\Gamma^C(S))$, complete with Hecke action, can be defined as in the classical case $N^- = 1$. By the modularity of E and the Jacquet-Langlands correspondence, there exists an eigenform $g_E \in S_2(\Gamma^C(S))$ with system of Hecke eigenvalues $\{a_p(g_E)\} = \{a_p(E)\}$, as well as a map

$$\Phi_{N^+,N^-} : \mathrm{Div}^0 \mathcal{H} \to \mathrm{Jac}\, X^C(S)(\mathbb{C}) \to E(\mathbb{C})$$

given by

$$(\tau') - (\tau) \mapsto W\Big(\int_{\tau}^{\tau'} g_E(z)dz\Big),$$

where W is the Weierstrass parametrization of $E(\mathbb{C})$. If $N^- = 1$, then we are in the situation of §1.1 and Φ_{N^+,N^-} is induced by the map Φ_N.

2.2. CM-points.

THEOREM 2.3 (Shimura). *$X^C(S)(\mathbb{C})$ is the set of complex points of a curve $X^C(S)$ defined over $\mathbb{Q}$. This curve classifies abelian surfaces with "level N^+-structure" whose endomorphism rings contain S.*

(For a discussion of this moduli problem, see [**Zha01**, Chapter 1].)

Let $\mathcal{O} \subset K$ be an imaginary quadratic order. We say a point P in $X^C(S)(\mathbb{C})$ has *CM by* $\mathcal{O}$ if it corresponds to an abelian surface whose endomorphism ring contains $\mathcal{O}$ as a subring commuting with S. Let $\mathrm{CM}(\mathcal{O})$ denote the set of such points P. The map from $\mathrm{Jac}\, X^C(S)(\mathbb{C})$ to $E(\mathbb{C})$ induced by Φ_{N^+,N^-} is also defined over $\mathbb{Q}$, so

$$\Phi_{N^+,N^-}(\mathrm{Div}^0\, \mathrm{CM}(\mathcal{O})) \subset E(H_{\mathcal{O}}).$$

We will call these points on E *Shimura-Heegner points*.

2.3. The Shimura-Heegner hypothesis. Let $\mathcal{O} \subset K$ an imaginary quadratic order whose discriminant is prime to N.

LEMMA 2.4. *The following are equivalent*

(1) *The set* $\mathrm{CM}(\mathcal{O})$ *is nonempty.*
(2) *All primes ℓ dividing N^+ (resp. N^-) are split (resp. inert) in K.*

Call conditions (1) and (2) are the *Shimura-Heegner hypothesis*. If the Shimura-Heegner hypothesis is satisfied for the maximal order $\mathcal{O}$ of K, call (N^+, N^-, K) a *Shimura-Heegner triple*. (This is not standard terminology and is in force in this paragraph only.) For a given imaginary quadratic field K of discriminant prime to N, there exists a factorization $N = N^+N^-$ such that (N^+, N^-, K) is a Shimura-Heegner triple if and only if the sign in the functional equation of $L(E/K, s)$ is -1. Thus, we have a Heegner-point type construction available exactly when the Birch and Swinnerton-Dyer conjecture predicts that the rank of $E(K)$ is positive for reasons of parity.

2.4. Elliptic curves over real quadratic fields. The phenomenon of elliptic curves being parametrized by Shimura curves generalizes to certain elliptic curves defined over totally real fields. For simplicity, let F be a real quadratic field with infinite places σ_1 and σ_2, and let $\mathfrak{p}$ be a finite prime of F. (The much more mysterious case of imaginary quadratic base fields will be discussed in [**Gre**].) Let C be the quaternion F-algebra ramified at $\mathfrak{p}$ and σ_1 and let S be a maximal order of C. Fix an isomorphism

$$\iota_{\sigma_2} : C \otimes_{\sigma_2} \mathbb{R} \to M_2(\mathbb{R}),$$

and let

$$\Gamma^C(S) = \{\iota_{\sigma_2}(s) : s \in S,\, \det \iota_{\sigma_2}(s) = 1\}/\{\pm 1\} \subset \mathrm{PSL}_2(\mathbb{R}).$$

As before, $\Gamma^C(S)$ acts discontinuously on $\mathcal{H}$. The quotient $\Gamma\backslash\mathcal{H}$ is a compact Riemann surface which admits a description as the complex points of a Shimura curve X, as well as a corresponding CM-theory.

Let $f \in \mathcal{S}_2(\mathfrak{p})$ be a Hilbert modular newform with rational Hecke eigenvalues. Then the Jacquet-Langlands correspondence together with an Eichler-Shimura construction implies the existence of an elliptic curve E/F parametrized by the Jacobian variety J of X whose L-function matches that of f. Again, we want to compute the images on E of degree zero CM divisors on J, which we also call Shimura-Heegner points.

3. Computing Heegner and Shimura-Heegner points

The classical Heegner points may be efficiently computed using formula (1.1). The quantities $a_n(f_E)$ can be computed using the formula

$$a_p(f_E) = p + 1 - \#\tilde{E}(\mathbb{F}_p),$$

(where p is a prime and $\tilde{E}$ is the reduction of E modulo p) in conjunction with the Euler product for $L(f_E, s)$. For details and a complexity analysis, see [**Elk94**].

The following questions remain: How do we efficiently compute Φ_{N^+,N^-}, and hence Shimura-Heegner points, when $N^- \neq 1$? The computability of the modular parametrization Φ_N of (1.1) relies on the Fourier expansion of f_E. When $N^- \neq 1$, such an expansion is not available. How about when E is defined over a real quadratic field?

In his article in this volume [**Voi**], John Voight discussed efficient methods for computing CM-points on certain Shimura curves. Unfortunately (for our purposes), the curves that he discussed were all of genus zero, and hence cannot parametrize elliptic curves.

N. Elkies [**Elk98**] has also developed methods for performing these computations in certain cases using archimedean analysis and explicit presentations of the groups $\Gamma^C(S)$. His methods are in fact related to those of Voight.

We present an approach based on p-adic analysis. Our main tools are the Cherednik-Drinfeld theorem, p-adic integration and the theory of rigid-analytic automorphic forms on definite quaternion algebras.

4. p-adic integration and uniformization

4.1. The Cherednik-Drinfeld interchange of invariants. Let $E/\mathbb{Q}$ be an elliptic curve of conductor $N = N^+N^-$ and suppose that p is a prime dividing N^-. (In particular, $N^- \neq 1$.) Let B be the quaternion algebra ramified at the primes dividing N^-/p, together with the place at infinity. (We interchange the roles of the places p and infinity — hence the title of this subsection.) Let R be an Eichler $\mathbb{Z}$-order in B of level N^+p.

Example 4.1. In the situation of Example 2.2, B is the algebra of Hamilton's quaternions, denoted $\left(\frac{-1,-1}{\mathbb{Q}}\right)$ in [**Voi**].

Since B is split at p, we may choose an isomorphism

$$\iota_p : B_p := B \otimes_{\mathbb{Q}} \mathbb{Q}_p \longrightarrow M_2(\mathbb{Q}_p)$$

such that ι_p induces an isomorphism of $R_p := R \otimes_{\mathbb{Z}} \mathbb{Z}_p$ with

$$R_0(p\mathbb{Z}_p) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathbb{Z}_p) : p \text{ divides } c \right\}.$$

4.2. The p-adic uniformization theorem.

DEFINITION 4.2 (Multiplicative integral). Let

- $\mathcal{B}_n$ be the standard decomposition of $\mathbb{P}^1(\mathbb{Q}_p)$ into p^n+p^{n-1} balls of radius p^{-n},
- μ be a $\mathbb{Z}$-valued measure on $\mathbb{P}^1(\mathbb{Q}_p)$, and
- f be a continuous, nonvanishing function on $\mathbb{P}^1(\mathbb{Q}_p)$.

Define

$$\int\!\!\!\!\!\!\diagup_{\mathbb{P}^1(\mathbb{Q}_p)} f(x)d\mu(x) = \lim_{n\to\infty}\prod_{U\in\mathcal{B}_n} f(t_U)^{\mu(U)},$$

where t_U is any point of U.

In his lecture on p-adic uniformization [**Dar**], Darmon constructs a $\mathbb{Z}$-valued distribution μ_E on $\mathbb{P}^1(\mathbb{Q}_p)$ (i.e. a finitely additive, $\mathbb{Z}$-valued function on the compact-open subsets of $\mathbb{P}^1(\mathbb{Q}_p)$) which is invariant under the group $R[1/p]_1^*$ of units in $R[1/p]$ of reduced norm 1. (The group B_p^* acts via ι_p on the projective line $\mathbb{P}^1(\mathbb{Q}_p)$ and hence on its compact-open subsets.) Let

$$\mathcal{H}_p := \mathbb{P}^1(\mathbb{C}_p) - \mathbb{P}^1(\mathbb{Q}_p)$$

be the p-adic upper half-plane and let

$$\text{Tate} : \mathbb{C}_p^* \to E(\mathbb{C}_p)$$

be the Tate parametrization of E.

THEOREM 4.3 (p-adic uniformization of E).

(1) *(Cherednik-Drinfeld) There is a canonical surjective map*

$$\text{CD} : \mathcal{H}_p \longrightarrow X^C(S)(\mathbb{C}_p).$$

(2) *(Bertolini-Darmon) The map* CD *satisfies*

$$\Phi_{N^+,N^-}\big((\text{CD}(\tau')) - (\text{CD}(\tau))\big) = \text{Tate}\left(\int\!\!\!\!\!\!\diagup_{\mathbb{P}^1(\mathbb{Q}_p)} \Big(\frac{x-\tau'}{x-\tau}\Big)d\mu_E(x)\right).$$

Furthermore, one may explicitly describe $\text{CD}^{-1}(\text{CM}(\mathcal{O})) \subset \mathcal{H}_p$.

4.3. Some details. In this subsection, we briefly indicate how the map CD is constructed and we identify the set $\text{CD}^{-1}(\text{CM}(\mathcal{O})) \subset \mathcal{H}_p$. Let

$$\Gamma^B(R[1/p]) = \{\iota_p(r) : r \in R[1/p],\, \det \iota_p(r) = 1\}/\{\pm 1\} \subset \text{PSL}_2(\mathbb{Q}_p).$$

The group $\Gamma^B(R[1/p])$ acts discontinuously on $\mathcal{H}_p$ and the quotient $\Gamma^B(R[1/p])\backslash\mathcal{H}_p$, has the structure of a rigid-analytic curve. To prove (1) of Theorem 4.3, Cherednik and Drinfeld show that there is a canonical rigid-analytic isomorphism

$$\text{CD} : \Gamma^B(R[1/p])\backslash\mathcal{H}_p \longrightarrow X^C(S)_{\mathbb{C}_p}.$$

Let $\mathcal{O} \subset K$ be an imaginary quadratic order satisfying the Shimura-Heegner hypothesis. Call an embedding ψ of $\mathcal{O}[1/p]$ into $R[1/p]$ *optimal* if it does not extend to an embedding of a larger $\mathbb{Z}[1/p]$-order in K and denote by $\mathcal{E}_p(\mathcal{O})$ the set of all such optimal embeddings. The Shimura-Heegner hypothesis guarantees that $\mathcal{E}_p(\mathcal{O})$ is nonempty. For each $\psi \in \mathcal{E}_p(\mathcal{O})$, the group $\mathcal{O}[1/p]^*$ acts on $\mathcal{H}_p$ via the composite $\iota_p \circ \psi$ with a unique fixed point $\tau_\psi \in \mathcal{H}_p$ satisfying

$$\alpha\begin{pmatrix}\tau_\psi\\1\end{pmatrix} = \psi(\alpha)\begin{pmatrix}\tau_\psi\\1\end{pmatrix}$$

for all $\alpha \in \mathcal{O}[1/p]^*$. Let $\mathcal{H}_p(\mathcal{O})$ be the set of all such τ_ψ. It can be shown (see [**BD96**]) that

$$\mathcal{H}_p(\mathcal{O}) = \mathrm{CD}^{-1}(\mathrm{CM}(\mathcal{O})).$$

Set

$$J(\tau,\tau') = \prod_{\mathbb{P}^1(\mathbb{Q}_p)} \left(\frac{x-\tau'}{x-\tau}\right) d\mu_E(x). \tag{4.1}$$

By statement (2) of Theorem 4.3, the points $\mathrm{Tate}(J(\tau,\tau'))$ for $\tau,\tau' \in \mathcal{H}_p(\mathcal{O})$ are Shimura-Heegner points on E defined over the ring class field $H_\mathcal{O}$ attached to $\mathcal{O}$. Slightly more generally, we are interested in the image of an arbitrary element $\mathfrak{d} \in \mathrm{Div}^0 \mathcal{H}_p(\mathcal{O})$ in $E(H_\mathcal{O})$. Suppose $\mathfrak{d}$ has the form

$$\mathfrak{d} = \sum_{i=1}^n ((\tau_i') - (\tau_i)).$$

For later use, we introduce the notation

$$\prod_{\mathfrak{d}} \omega_{\mu_E} = \prod_{i=1}^n J(\tau_i,\tau_i').$$

Thus we have

$$\Phi_{N^+,N^-}(\mathfrak{d}) = \mathrm{Tate}\left(\prod_{\mathfrak{d}} \omega_{\mu_E}\right).$$

Not only can we describe the Shimura-Heegner points analytically, but also the action of $\mathrm{Gal}\, H_\mathcal{O}/K$ on them: One can show (see [**Gro87**, §3.2]) that the class group $\mathrm{Pic}\,\mathcal{O}$ acts freely on the set $\mathcal{H}_p(\mathcal{O})$. Let

$$\mathrm{rec} : \mathrm{Pic}\,\mathcal{O} \longrightarrow \mathrm{Gal}\, H_\mathcal{O}/K.$$

be the map induced by the reciprocity homomorphism of class field theory.

THEOREM 4.4 (Shimura's reciprocity law). *Let $\tau,\tau' \in \mathcal{H}_p(\mathcal{O})$. Then for all $\alpha \in \mathrm{Pic}\,\mathcal{O}$, we have*

$$\Phi_{N^+,N^-}((\tau'^\alpha) - (\tau^\alpha)) = \Phi_{N^+,N^-}((\tau') - (\tau))^{\mathrm{rec}\,\alpha}.$$

We utilize Shimura's reciprocity extensively in performing our computations.

Summing up this section, we have seen that to compute Shimura-Heegner points p-adically, it suffices to be able to compute p-adic integrals of the form (4.1).

5. Computing p-adic integrals

5.1. The naive approach. It is natural to attempt to evaluate $J(\tau,\tau')$ from the definition, i.e. by evaluating the "Riemann products" defining the multiplicative integral (see Definition 4.2). One can show that we do not lose generality by assuming that the reductions of the points τ and τ' lie in $\mathbb{P}^1(\overline{\mathbb{F}}_p) - \mathbb{P}^1(\mathbb{F}_p)$, and for the rest of the talk we shall work under this assumption. In this case, it is not hard to show that

$$J(\tau,\tau') \equiv^* \prod_{U\in\mathcal{B}_N} \left(\frac{t_U-\tau'}{t_U-\tau}\right)^{\mu(U)} \pmod{p^N},$$

where $x \equiv^* y \pmod{p^N}$ means $x/y - 1 \equiv 0 \pmod{p^N}$. Unfortunately, the size of $\mathcal{B}_N$ is $p^N + p^{N-1}$ — exponential in N. Thus, the naive approach does not facilitate the calculation of (4.1) to high accuracy.

5.2. Outline of the method. In this subsection, we give a sketch of our alternate method for computing (4.1), and hence Shimura-Heegner points. For complete details, see [**Gre06**].

First, we observe that the Teichmüller representative of $J(\tau,\tau')$ is the same as that of

$$\prod_{a=0}^{p-1}\left(\frac{a-\tau'}{a-\tau}\right)^{\mu(a+p\mathbb{Z}_p)},$$

an easily computed quantity. Consequently, it is sufficient to compute $\log J(\tau,\tau')$, where "log" denotes the (standard) branch of the p-adic logarithm satisfying $\log p = 0$.

For simplicity, we assume that there is some $i \in R[1/p]_1^*$ such that $\iota_p(i) = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. This is easy to arrange if B is the algebra of Hamilton's quaternions, for instance. Write

$$\log J(\tau,\tau') = \sum_{a\in\mathbb{P}^1(\mathbb{F}_p)} \log J_a(\tau,\tau'), \quad \text{where}$$
$$J_a(\tau,\tau') = \oint_{\mathbf{b}_a}\left(\frac{x-\tau'}{x-\tau}\right)d\mu_E(x),$$

and $\mathbf{b}_a$ is the standard residue disk around a. Let

$$J_\infty(\tau) = \oint_{\mathbf{b}_0}(1+\tau x)d\mu_E(x),$$
$$J_a(\tau) = \oint_{\mathbf{b}_a}(x-\tau)d\mu_E(x), \quad 0 \le a \le p-1.$$

Then for each $a \in \mathbb{P}^1(\mathbb{F}_p)$, we have

$$J_a(\tau,\tau') = J_a(\tau')/J(\tau).$$

(To prove the above for $a=\infty$, we use the above assumption on the existence of i.)

Straightforward manipulations (see [**DP06**, §1.3]) show that the expansions

$$\log J_\infty(\tau) = \sum_{n\ge 1}\frac{(-1)^n}{n}\tau^n\omega(0,n), \tag{5.1}$$

$$\log J_a(\tau) = \sum_{n\ge 1}\frac{1}{n(a-\tau)^n}\omega(a,n), \quad 0 \le a \le p-1. \tag{5.2}$$

are valid, where (following the notation of [**DP06**]),

$$\omega(a,n) = \int_{\mathbf{b}_a}(x-a)^n d\mu_E(x), \quad 0 \le a \le p-1.$$

Let

$$M' = \max\{n : \mathrm{ord}_p(p^n/n) < M\}, \quad M'' = M + \left\lfloor \frac{\log M'}{\log p}\right\rfloor. \tag{5.3}$$

Examining formulas (5.1), (5.2), and (5.3), it is easy to deduce the following:

PROPOSITION 5.1. *To compute* $\log J(\tau,\tau')$ *to a precision of* p^{-M}*, it suffices to compute the data*

$$\omega(a,n) \pmod{p^{M''}}, \quad 0 \le a \le p-1, \quad 0 \le n \le M'. \tag{5.4}$$

THEOREM 5.2. *The data (5.4) may be computed in $O(M^3p^3\log M)$ operations on integers of size on the order of p^M.*

Theorem 5.2 is proved in detail in [**Gre06**, Proposition 7]. The idea is to relate the moments $\omega(a,n)$ to certain automorphic forms on the definite quaternion algebra B. We show that the data (5.4) is encoded in a natural way in an automorphic form $\Phi^{M''}$ taking values in a module of "approximate distributions" on $\mathbb{Z}_p$. $\Phi^{M''}$ is characterized by a finite amount of data and can be represented nicely on a computer. The form $\Phi^{M''}$ is the M''-th term in a sequence of approximations Φ_n. The transition from the n-th approximation Φ_n to the $(n+1)$-st Φ_{n+1} proceeds by an application of the Hecke operator U_p, a process which can be carried out algorithmically. Each of the M'' required applications of the U_p operator requires $O(M^2p^3\log M)$ operations on elements of $\mathbb{Z}/p^{M''}\mathbb{Z}$. The running-time estimate of Theorem 5.2 follows from the fact that $M''\approx M$.

6. Sample computations

EXAMPLE 6.1. Let $E/\mathbb{Q}$ be the curve

$$E: y^2+xy+y=x^3+x^2-70x-279 \quad (38B2)$$

of conductor $N=38=2\cdot 19$. Taking $N^-=N$ and $p=19$ as in Example 4.1, we have that B is the algebra of Hamilton's quaternions. Consider the maximal order $\mathcal{O}=\mathbb{Z}[\xi]\subset K=\mathbb{Q}(\xi)$, where

$$\xi=\frac{1+\sqrt{-195}}{2}.$$

Both 2 and 19 are inert in K, so the Shimura-Heegner hypothesis is satisfied. One may compute that

$$\operatorname{Pic}\mathcal{O}=\mathbb{Z}/2\mathbb{Z}\oplus\mathbb{Z}/2\mathbb{Z},\quad H_{\mathcal{O}}=K(\sqrt{-3},\sqrt{5})$$

Let χ_1, χ_2, χ_3 be the characters of $\operatorname{Pic}\mathcal{O}$ of exact order 2. Assume these characters are indexed so that the fields corresponding to χ_1, χ_2, and χ_3 are $K(u)$, $K(v)$, and $K(w)$, respectively, where

$$u=\frac{1+\sqrt{-15}}{2},\quad v=\frac{1+\sqrt{5}}{2}\quad w=\frac{1+\sqrt{65}}{2}.$$

We remark that $K(u,v,w)$ is the Hilbert class field of K. Choose an optimal embedding of $\mathcal{O}$ into the maximal order of B and let $\tau\in\mathcal{H}_p(\mathcal{O})$ be its fixed point. Define degree 0 CM divisors

$$\mathfrak{d}_i=\sum_{\alpha\in\operatorname{Pic}\mathcal{O}}\chi_i(\alpha)\tau^\alpha,\quad i=1,2,3.$$

(Note that this makes sense as χ takes values in $\{1,-1\}$.) Define a degree 0 divisor corresponding to the trivial character by

$$\mathfrak{d}_0=\sum_{\alpha\in\operatorname{Pic}\mathcal{O}}((3+1-T_3)\tau)^\alpha,$$

where T_3 is the usual Hecke operator. Set

$$P_i=\operatorname{Tate}\left(\int_{\mathfrak{d}_i}\omega_{\mu_E}\right),\quad i=0,1,2,3.$$

We computed 19-adic approximations to the $P_i \in E(K_{19})$ modulo 19^{40}. These approximations were recognized as the global points

$$\begin{aligned} P_0 &= (-4610/39, (-277799\xi + 228034)/1521), \\ P_1 &= (25/12, -94/9u + 265/72), \\ P_2 &= (10, -11v), \\ P_3 &= (1928695/2548, (-2397574904w + 1023044339)/463736). \end{aligned}$$

But how do we recognize 19-adic approximations as points with algebraic coordinates? We represent a generic element $19^a u + 19^b v\xi \in K_{19}$ with $u, v \in \mathbb{Z}_{19}^*$ as the quadruple $(a, u \pmod{19^{40}}, b, v \pmod{19^{40}})$. Thus, to recognize such an approximation as an element of K, it suffices to be able to recognize an approximation to an element of $\mathbb{Z}_{19}^*$ as a rational number. This is accomplished using lattice reduction techniques as in [**DP06**]. These ideas allowed us to recognize the coordinates of P_0 as elements of K. The coordinates $x(P_1)$ and $y(P_1)$ should be rational over $K(u)$, not over K itself. Let σ be a generator of $\operatorname{Gal} K(u)/K$. Using Shimura reciprocity, we can compute approximations to $x(P_1)^\sigma$ and $y(P_1)^\sigma$ in K_{19}. If $x(P_1) = u + v\sqrt{-15}$ with $u, v \in K$, then

$$u = \frac{1}{2}(x(P_1) + x(P_1)^\sigma), \quad v = \frac{1}{2\sqrt{-15}}(x(P_1) - x(P_1)^\sigma).$$

Fixing an embedding of $K(u)$ into K_{19}, we may compute approximations to u and v as elements of K_{19} and then attempt to recognize them as elements of K as described above. The coordinate $y(P_1)$, as well as the coordinates of P_2 and P_3, were identified in the same way.

We remark that, for this example, the computation of the data (5.4) to a precision of 40 19-adic digits took approximately one minute.

Example 6.2. Let

$$\omega = \frac{1 + \sqrt{5}}{2}, \quad F = \mathbb{Q}(\omega),$$

and consider the elliptic curve

$$E: \ y^2 + xy + \omega y = x^3 - (\omega + 1)x^2 - (30\omega + 45)x - (11\omega + 117).$$

of conductor $(3 - 5\omega) =: \mathfrak{p}$. (We have $(31) = \mathfrak{p}\bar{\mathfrak{p}}$.) In this case, the definite quaternion algebra B which comes into play is the base change to F of the $\mathbb{Q}$-algebra of Hamilton's quaternions.

Consider the CM-field $K = F(\sqrt{2\omega - 15})$ with maximal order $\mathcal{O}$. $\operatorname{Pic}\mathcal{O} \cong \mathbb{Z}/8\mathbb{Z}$ and thus has a unique character χ of exact order 2 with corresponding field $K(\sqrt{-13\omega + 2})$. Choose a base point $\tau \in \mathcal{H}_p(\mathcal{O})$ and a define a divisor

$$\mathfrak{d}_\chi = \sum_{\alpha \in \operatorname{Pic}\mathcal{O}} \chi(\alpha)\tau^\alpha$$

and a point

$$P_\chi = \operatorname{Tate}\left(\int\!\!\!\!\!\!\diagup_{\mathfrak{d}_\chi} \omega_{\mu_E} \right)$$

associated to χ. (Again, this makes sense as χ takes values in $\{1, -1\}$.)

Using the techniques described above, the point P_χ was recognized as the global point

$$(x, y) \in E(F(\sqrt{-13\omega + 2})), \quad \text{where}$$

$$\begin{aligned}
x &= 1/501689727224078580\times \\
&\quad (-2048932971295530218 1\omega + \\
&\qquad 1590697243182535465) \\
y &= 1/7945803389515397981338566 00\times \\
&\qquad (-2430756213639475197971343 8023\omega \\
&\qquad - 5224406254275398040668003 6861) \\
&\qquad \times \sqrt{-13\omega + 2} \\
&\qquad + 1/1003379454448157160\times \\
&\qquad (19987639985731223601\omega \\
&\qquad - 1590697243182535465).
\end{aligned}$$

Our computations were all carried out using the Magma computer algebra system.

References

[BD96] M. Bertolini and H. Darmon, *Heegner points on Mumford-Tate curves*, Invent. Math. **126** (1996), no. 3, 413–456. MR 1419003 (97k:11100)

[Dar] H. Darmon, *Rational points on curves*, in this volume.

[Dar04] H. Darmon, *Rational points on modular elliptic curves*, CBMS Regional Conference Series in Mathematics, vol. 101, Published for the Conference Board of the Mathematical Sciences, Washington, DC, 2004. MR 2020572 (2004k:11103)

[DP06] H. Darmon and R. Pollack, *Efficient calculation of Stark-Heegner points via overconvergent modular symbols*, Israel J. Math. **153** (2006), 319–354. MR 2254648 (2007k:11077)

[Elk94] N. D. Elkies, *Heegner point computations*, Algorithmic number theory (Ithaca, NY, 1994) (L. M. Adleman and M.-D. Huang, eds.), Lecture Notes in Comput. Sci., vol. 877, Springer, Berlin, 1994, proceedings of ANTS-1, 1994, pp. 122–133. MR 1322717 (96f:11080)

[Elk98] ———, *Shimura curve computations*, Algorithmic number theory (Portland, OR, 1998) (J.P. Buhler, ed.), Lecture Notes in Comput. Sci., vol. 1423, Springer, Berlin, 1998, proceedings of ANTS-3, 1998, pp. 1–47. MR 1726059 (2001a:11099)

[Gre] M. Greenberg, *The arithmetic of elliptic curves over imaginary quadratic fields and Stark-Heegner points*, in this volume.

[Gre06] ———, *Heegner point computations via numerical p-adic integration*, Algorithmic number theory (F. Hess, S. Pauli, and M. Pohst, eds.), Lecture Notes in Comput. Sci., vol. 4076, Springer, Berlin, 2006, proceedings of ANTS-7, 2006, pp. 361–376. MR 2282936 (2008a:11069)

[Gro87] B. H. Gross, *Heights and the special values of L-series*, Number theory (Montreal, Que., 1985) (H. Kisilevsky and J. Labute, eds.), CMS Conf. Proc., vol. 7, Amer. Math. Soc., Providence, RI, 1987, pp. 115–187. MR 894322 (89c:11082)

[Vig80] M.-F. Vignéras, *Arithmétique des algèbres de quaternions*, Lecture Notes in Mathematics, vol. 800, Springer, Berlin, 1980. MR 580949 (82i:12016)

[Voi] J. Voight, *Shimura curve computations*, in this volume.

[Zha01] S.-W. Zhang, *Heights of Heegner points on Shimura curves*, Ann. of Math. (2) **153** (2001), no. 1, 27–147. MR 1826411 (2002g:11081)

Department of Mathematics, McGill University, Montreal, Quebec, Canada
Current address: Max-Planck-Institut für Mathematik, 53111 Bonn, Germany
E-mail address: fmgreenberg@gmail.com

Clay Mathematics Proceedings
Volume 8, 2009

The arithmetic of elliptic curves over imaginary quadratic fields and Stark-Heegner points

Matthew Greenberg

Abstract. Heegner points are crucial to our understanding of the arithmetic of elliptic curves over $\mathbb{Q}$ as well as over totally real fields. In this note, we describe a conjectural construction due to Trifković of analogues of Heegner points for elliptic curves defined over imaginary quadratic fields. We expect these points to enrich our understanding of the arithmetic of such curves.

1. Introduction

A large proportion of research into the arithmetic of elliptic curves is devoted to the understanding of *Mordell-Weil groups* — groups of points on elliptic curves rational over number fields. Many questions regarding the structure of Mordell-Weil groups, most famously the conjecture of Birch and Swinnerton-Dyer (BSD), remain open. Much of what we *do* know about these groups (e.g. BSD for elliptic curves over $\mathbb{Q}$ of analytic rank at most one) is due to the existence of a systematic construction of points — so-called *Heegner points* — of Mordell-Weil groups in towers of number fields. In appropriate situations, these Heegner points govern the behaviour of Mordell-Weil groups in a very strong way.

Heegner points on an elliptic curve E are, by definition, the images of CM points under *modular parametrizations* of E: dominant morphisms from modular or Shimura curves to E. In particular, for Heegner points to exist, E needs to admit a modular parametrization in the first place, a condition only reasonable to expect in any kind of generality if E can be defined over a totally real field. Due to the absolutely crucial role played by Heegner points in the study of Mordell-Weil groups, it is extremely natural to desire a generalization of the Heegner point construction to elliptic curves which do not necessarily admit modular parametrizations. In this article, we present such a generalization, due to Trifković [**Tri06**], in the case of elliptic curves defined over imaginary quadratic fields.

Trifković's work is based on Darmon's construction of *Stark-Heegner points* on elliptic curves defined over $\mathbb{Q}$ — analogues of Heegner points which are conjectured

2000 *Mathematics Subject Classification.* Primary 11G05, Secondary 11F11, 11F67, 11G40.
Key words and phrases. elliptic curves, modular forms, imaginary quadratic fields, Stark-Heegner points.

to be rational over ring class fields of real quadratic fields. Although Darmon's construction makes essential use of the modular forms attached to elliptic curves over $\mathbb{Q}$, the modular parametrizations are not explicitly involved.[1] It is this characteristic which raises the prospect of generalizing the Stark-Heegner point construction to base fields other than totally real ones where, although modular parametrizations are not expected to be available, the elliptic curves in question are still expected to be "modular."

The central role played by rational points in the arithmetic of elliptic curves is summed up beautifully by the following lines from the abstract of [**BMSW07**]:

> "Rational points on elliptic curves are the gems of arithmetic: they are, to Diophantine geometry, what units in rings of integers are to algebraic number theory, what algebraic cycles are to algebraic geometry. A rational point in just the right context, at one place in the theory, can inhibit and control — thanks to the ideas of Kolyvagin — the existence of rational points and other mathematical structures elsewhere."

This article is divided into three main parts. First, we will define modular forms and modular symbols relative to an imaginary quadratic base field and state some fundamental results concerning these. Armed with these notions, we will describe Trifković's Stark-Heegner point construction and state his conjectures concerning their algebricity. In the last part, we shall discuss issues related to the computation of these points in practice.

The author would like to sincerely thank the anonymous referee for numerous insightful suggestions which led to significant improvements in this article.

2. Modular forms for imaginary quadratic fields

2.1. Upper half-space. In addition to [**Tri06**], some good references for this section are [**Byg98, Cre84, Cre, CW94, Lin05**]. Reference [**Byg98**] in particular is extremely detailed and contains a wealth of background material. Let F be an imaginary quadratic field of discriminant D with maximal order $\mathcal{O}_F$, and assume that $\mathcal{O}_F$ is a principal ideal domain. Fix an ideal $\mathcal{N}$ of $\mathcal{O}_F$. In analogy with the classical situation, define

$$\mathcal{H} = \mathrm{GL}_2(\mathbb{C})/\mathbb{C}^* \cdot \mathrm{SU}_2$$

and call $\mathcal{H}$ the *upper half-space*. The group $\mathrm{GL}_2(\mathbb{C})$ admits a decomposition $\mathrm{GL}_2(\mathbb{C}) = BKZ$, where

$$B = \left\{ \begin{pmatrix} t & z \\ 0 & 1 \end{pmatrix} : \quad \begin{matrix} z \in \mathbb{C} \\ t \in \mathbb{R}_{>0} \end{matrix} \right\}, \quad K = \mathrm{SU}_2, \quad \text{and} \quad Z = \mathbb{C}^*,$$

mirroring the analogous decomposition of $\mathrm{GL}_2^+(\mathbb{R})$ where

$$B = \left\{ \begin{pmatrix} y & x \\ 0 & 1 \end{pmatrix} : \quad \begin{matrix} x \in \mathbb{R} \\ y \in \mathbb{R}_{>0} \end{matrix} \right\}, \quad K = \mathrm{SO}_2, \quad \text{and} \quad Z = \mathbb{R}^*,$$

Projecting onto the B-coordinate, we have an identification

$$\mathcal{H} \cong \{(z,t) : z \in \mathbb{C},\, t \in \mathbb{R}_{>0}\}.$$

[1] S. Dasgupta [**Das05**] has shown how to explicitly lift the Stark-Heegner points on E to an appropriate modular Jacobian.

The action of $\mathrm{GL}_2(\mathbb{C})$ on $\mathcal{H}$ takes the form

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}(z,t) = \frac{1}{|cz+d|^2+|ct|^2}\big((az+b)\overline{(cz+d)} + a\bar{c}t, |ad-bc|t\big) \tag{2.1}$$

The upper half-space $\mathcal{H}$ is equipped with a $\mathrm{GL}_2(\mathbb{C})$-invariant Euclidean metric given by

$$ds^2 = \frac{dzd\bar{z}+dt^2}{t^2}.$$

Let $\mathcal{H}^*$ be the disjoint union of $\mathcal{H}$ with $\mathbb{P}^1(F)$. (Note that, although this is not reflected in the notation, the set $\mathcal{H}^*$ depends on the field F.) Extend the topology of $\mathcal{H}$ to $\mathcal{H}^*$ by declaring sets of the form

$$U_h = \{(z,t) \in \mathcal{H} : t > h\} \cup \{\infty\},$$

as well as their translates by elements of $\mathrm{GL}_2(F)$, to be open. The action of

$$\Gamma_0(\mathcal{N}) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}_2(\mathcal{O}_F) : c \in \mathcal{N} \right\}.$$

extends naturally to $\mathcal{H}^*$, so we may consider the quotient

$$X_0(\mathcal{N}) := \Gamma_0(\mathcal{N})\backslash\mathcal{H}^*.$$

We assume that $\Gamma_0(\mathcal{N})$ has no elements of finite order, in which case $X_0(\mathcal{N})$ is a smooth 3-manifold. (See [**Kur78**] for details on dealing with the situation where $\Gamma_0(\mathcal{N})$ contains elements of finite order.) The points $\Gamma_0(\mathcal{N})\backslash\mathbb{P}^1(F)$ are called the *cusps* of $X_0(\mathcal{N})$.

2.2. Modular forms on the upper half space.

Definition 2.1. A *modular form of weight 2 for* $\Gamma_0(\mathcal{N})$ is a $\Gamma_0(\mathcal{N})$-invariant harmonic differential form on $\mathcal{H}$. If it descends to a harmonic differential form on $X_0(\mathcal{N})$, then we call it a *cusp form.*

We denote the set of modular (resp. cusp) forms of weight two for $\Gamma_0(\mathcal{N})$ by $\mathcal{M}_2(\mathcal{N})$ (resp. $\mathcal{S}_2(\mathcal{N})$). Consider the basis of smooth differential 1-forms on $\mathcal{H}$ given by

$$\omega = (\omega_1, \omega_2, \omega_3)^t = (-dz/t, dt/t, d\bar{z}/t)^t.$$

and let $f = (f_1, f_2, f_3)^t$ be a vector of smooth functions on $\mathcal{H}$.

Lemma 2.2.

(1) *The differential form $f \cdot \omega$ is $\Gamma_0(\mathcal{N})$-invariant if and only if*

$$f(z,t) = (f|\gamma)(z,t) := J(\gamma,(z,t))f(\gamma(z,t))$$

for all $\gamma \in \Gamma_0(\mathcal{N})$, where

$$J(\gamma,(z,t)) = \frac{1}{|r|^2+|s|^2}\begin{pmatrix} r^2\Delta & -2rs\Delta & s^2\Delta \\ r\bar{s} & |r|^2-|s|^2 & -\bar{r}s \\ \overline{s^2\Delta} & \overline{2rs\Delta} & \overline{r^2\Delta} \end{pmatrix},$$

$$\Delta = \det\gamma, \quad r = \overline{cz+d} \quad s = \bar{c}t.$$

(2) *The differential form $f\cdot\omega$ is harmonic if and only if the following partial differential equations are satisfied:*

$$\begin{aligned}\frac{\partial f_1}{\partial\bar z}+\frac{\partial f_3}{\partial z}&=0\\ \frac{\partial f_2}{\partial z}+\frac{\partial f_1}{\partial t}-t^{-1}f_1&=0\\ \frac{\partial f_2}{\partial\bar z}-\frac{\partial f_3}{\partial t}+t^{-1}f_3&=0\\ \frac{t}{2}\frac{\partial f_2}{\partial t}-f_2-2t\frac{\partial f_1}{\partial\bar z}&=0\end{aligned}$$

If $f\cdot\omega$ is a modular (resp. cusp) form, then we shall call f a modular (resp. cusp) form too.

2.3. Fourier expansions and cusp forms. Suppose that f satisfies $f|\gamma=f$ for all $\gamma\in\Gamma_0(\mathcal{N})$. This implies that $J(\left(\begin{smallmatrix}1&\alpha\\0&1\end{smallmatrix}\right),(z,t))$ is the identity matrix for all $\alpha\in\mathcal{O}_F$, so for each fixed t, the function $f_i(z,t)$ is periodic with respect to the lattice $\mathcal{O}_F\subset\mathbb{C}$. Let $g(z)$ be any function with this property and let

$$\psi:\mathbb{C}\longrightarrow S^1,\quad \psi(z)=e^{2\pi i(z+\bar z)}$$

be the standard unitary character of the additive group of $\mathbb{C}$. Then g admits a Fourier expansion of the form

$$g(z)=\sum_{\chi}b_g(\chi)\chi(z),$$

where χ varies over the unitary characters of $\mathbb{C}$ which are trivial on $\mathcal{O}_F$. But each of these characters has the form

$$z\mapsto\psi(\alpha z)\quad\text{for some}\quad\alpha\in\mathfrak{d}_F^{-1}=\frac{1}{\sqrt{D}}\mathcal{O}_F.$$

Thus, the expansion of g takes the form

$$g(z)=\sum_{\alpha\in\mathcal{O}_F}c_g(\alpha)\psi\left(\frac{\alpha z}{\sqrt{D}}\right).$$

It follows that for each $\alpha\in\mathcal{O}_F$ there is a vector-valued function

$$c_f(\alpha,t)=(c_1(\alpha,t),c_2(\alpha,t),c_3(\alpha,t))$$

of t such that

$$f(z,t)=\sum_{\alpha\in\mathcal{O}_F}c_f(\alpha,t)\psi\left(\frac{\alpha z}{\sqrt{D}}\right).$$

One may verify that if $\varepsilon\in\mathcal{O}_F^*$ and $\gamma=\left(\begin{smallmatrix}\varepsilon&0\\0&1\end{smallmatrix}\right)$, then $c_{f|\gamma}(\alpha,t)=c_f(\varepsilon\alpha,t)$. Therefore, as $f|\gamma=\gamma$, we have $c_f(\varepsilon\alpha,t)=c_f(\alpha,t)$ for all $\varepsilon\in\mathcal{O}_F^*$ and all $\alpha\in\mathcal{O}$. Consequently, recalling that we assume $\mathcal{O}_F$ to be a PID, we may rewrite the above sum as a sum over ideals of $\mathcal{O}_F$:

$$f(z,t)=c_f(0,t)+\sum_{0\subsetneq(\alpha)\subset\mathcal{O}_F}c_f(\alpha,t)\sum_{\varepsilon\in\mathcal{O}_F^*}\psi\left(\frac{\varepsilon\alpha z}{\sqrt{D}}\right).$$

LEMMA 2.3. *If $c_{f|\gamma}(0,t)=0$ for each $\gamma\in\mathrm{GL}_2(\mathcal{O}_F)$, then f is a cusp form on $\Gamma_0(\mathcal{N})$.*

The harmonicity of $f \cdot \omega$ implies that the components of $c_f(\alpha, t)$ are of a special form.

DEFINITION 2.4. For $i = 0, 1$, let $K_i(t)$ denote the solution to the differential equation

$$\frac{d^2K_i}{dt^2} + \frac{1}{t}\frac{dK_i}{dt} - \left(1 + \frac{1}{t^{2i}}\right) K_i = 0$$

which decreases rapidly at infinity (see [**Byg98**, Ch. 4]). The functions K_i are called *Bessel functions*.

Set

$$K(t) = \left(-\frac{i}{2}K_1(t), K_0(t), \frac{i}{2}K_1(t)\right).$$

It can be shown that for each $\alpha \in \mathcal{O}_F$ there is a constant $c_f(\alpha)$ such that

$$c_f(\alpha, t) = c_f(\alpha) t^2 K\left(\frac{4\pi|\alpha|t}{\sqrt{|D|}}\right),$$

so the Fourier expansion of f takes the form

$$f(z, t) = \sum_{\alpha \in \mathcal{O}_F} c_f(\alpha) t^2 K\left(\frac{4\pi|\alpha|t}{\sqrt{|D|}}\right) \psi\left(\frac{\alpha z}{\sqrt{D}}\right).$$

2.4. Hecke operators. The vector space $\mathcal{M}_2(\mathcal{N})$ admits an action of certain Hecke operators. Let λ (resp. π) be a prime element of $\mathcal{O}_F$ prime to (resp. dividing) $\mathcal{N}$. Then operators T_λ and U_π are defined by the "usual" formulas:

$$f|T_\lambda = \sum_{\alpha \bmod \lambda} f\left|\begin{pmatrix} 1 & \alpha \\ 0 & \lambda \end{pmatrix} + f\right|\begin{pmatrix} \lambda & 0 \\ 0 & 1 \end{pmatrix},$$
$$f|U_\pi = \sum_{\alpha \bmod \pi} f\left|\begin{pmatrix} 1 & \alpha \\ 0 & \pi \end{pmatrix}\right..$$

The effect of the Hecke operators on Fourier coefficients is given by the familiar formulas:

$$c_{f|T_\lambda}(\alpha) = \begin{cases} c_f(\lambda\alpha) + \mathrm{Norm}(\lambda) c_f(\alpha/\lambda) & \text{if } \lambda|\alpha, \\ c_f(\lambda\alpha) & \text{if } \lambda \nmid \alpha \end{cases},$$
$$c_{f|U_\pi}(\alpha) = c_f(\pi\alpha).$$

It follows that the operators T_λ and U_π depend only on the ideals (λ) and (π), respectively.

The Hecke operators generate a commutative subalgebra of $\mathrm{End}\,\mathcal{M}_2(\mathcal{N})$ which preserves $\mathcal{S}_2(\mathcal{N})$. If $f \in \mathcal{S}_2(\mathcal{N})$ is an eigenform for all the Hecke operators and is normalized so that $c_f(1) = 1$, then $f|T_\lambda = c_f(\lambda)f$ and $f|U_\pi = c_f(\pi)f$. A notion of newform may be defined, and an Atkin-Lehner theory developed, in a manner analogous to that employed in the classical case (i.e. over $\mathbb{Q}$).

2.5. L-functions and Shimura-Taniyama. Let $f \in \mathcal{S}_2(\mathcal{N})$ be a normalized newform and define

$$L(f,s) = \sum_{(0)\subsetneq \mathfrak{a}\subset \mathcal{O}_F} c_f(\mathfrak{a})\,\mathrm{Norm}(\mathfrak{a})^{-s},$$

where the sum is over nonzero ideals of $\mathcal{O}_F$.

THEOREM 2.5. *The series defining $L(f,s)$ converges in the right half-plane $\Re s > 3/2$. It admits an Euler product, analytic continuation to the whole complex plane, and satisfies a functional equation relating its values at s and $2-s$.*

The analogue of the Shimura-Taniyama conjecture in this context is:

CONJECTURE 2.6. *There is a one-to-one correspondence between normalized cuspidal newforms $f \in \mathcal{S}_2(\mathcal{N})$ with rational Hecke-eigenvalues and isogeny classes of elliptic curves $E_{/F}$ which do not have complex multiplication by an order in F. If f corresponds to E, then*

$$L(f,s) = L(E,s).$$

REMARK 2.7. If case that E has CM by an order in F, then E corresponds to an Eisensein series on $\Gamma_0(\mathcal{N})$ rather than to a cusp form.

REMARK 2.8. In the classical case (i.e. over $\mathbb{Q}$), the Eichler-Shimura construction attaches both a Galois representation and an elliptic curve to a newform g. In many cases, Richard Taylor has succeeded in constructing Galois representations attached to modular forms over imaginary quadratic fields by relating them to *holomorphic* Siegel modular forms. This allows him to use algebro-geometric methods to locate the desired Galois representations in the ℓ-adic cohomology of these varieties. His construction does not, however, give a construction of an elliptic curve associated to the form g. If one has a prospective elliptic curve E in mind, though, one can use the Faltings-Serre method to show that the Galois representation attached to g is that arising from the Galois action on the Tate module of E.

REMARK 2.9. In [**Tri06**, p. 432], the analogue of the Shimura-Taniyama conjecture is phrased in terms of plusforms. Trifković's plusform condition is always satisfied for the modular forms in this paper since we require them to be invariant with respect to a congruence subgroup of $\mathrm{GL}_2(\mathcal{O}_F)$ whereas Trifković asks only for invariance with respect to a conguence subgroup of $\mathrm{SL}_2(\mathcal{O}_F)$.

3. Modular symbols and mixed period integrals

Let $f \in S_2(\mathcal{N})$ be a Hecke-eigenform where $\mathcal{N} \in \mathcal{O}_F$. Suppose further that $\mathcal{N}$ has the form $\pi\mathcal{M}$, where π is a prime element of $\mathcal{O}_F$ (lying over the rational prime p, say) and $\pi \nmid \mathcal{M}$. Let F_π be the completion of F at the ideal (π) and let $\mathcal{O}_{F,\pi}$ be its ring of integers.

PROPOSITION 3.1 ([**Kur78**]). *There exists a unique positive real number $\Omega_f \in \mathbb{R}$ such that*

$$\left\{ \int_r^s f\cdot\omega : r,s \ \textit{cusps} \right\} = \Omega_f\mathbb{Z}.$$

We call the quantity Ω_f the *period of* f. Using this definition of the period, Darmon's mixed period integral formalism (see [**Dar01**] or [**Dar**]) extends easily to our setting: Let

$$\tilde{\Gamma} = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}_2(\mathcal{O}_F[1/\pi]) : c \in \mathcal{M} \right\}.$$

The group $\mathrm{GL}_2(F_\pi)$ acts from the left on $\mathbb{P}^1(F_\pi)$ by fractional linear transformations. We call a subset B of $\mathbb{P}^1(F_\pi)$ a *ball* if it is of the form $\sigma\mathcal{O}_{F,\pi}$ for some $\sigma \in \mathrm{GL}_2(F_\pi)$. By the strong approximation theorem, the group $\tilde{\Gamma}$ acts transitively on the set $\mathcal{B}$ of these balls.

A $\mathbb{Z}$-*valued distribution* on $\mathbb{P}^1(F_\pi)$ is, by definition, a finitely additive, $\mathbb{Z}$-valued function on $\mathcal{B}$. We shall denote the set of such by $\mathcal{D}_{\mathbb{Z}}(\mathbb{P}^1(F_\pi))$. If $\mu \in \mathcal{D}_{\mathbb{Z}}(\mathbb{P}^1(F_\pi))$ and φ is a nonvanishing, continuous, $\mathbb{C}_p$-valued function on $\mathbb{P}^1(F_\pi)$, we define the *multiplicative integral*

$$\int\!\!\!\!\!\!\diagup_{\mathbb{P}^1(F_\pi)} \varphi(x) d\mu(x) = \lim_{\mathcal{U}} \prod_{U \in \mathcal{U}} \varphi(x_U)^{\mu(U)} \in \mathbb{C}_p^*,$$

where $\mathcal{U}$ varies over increasingly fine covers of $\mathbb{P}^1(F_\pi)$ by pairwise disjoint balls, and x_U is any point in U. (Note that since we are exponentiating by the values of μ, it is essential that μ is $\mathbb{Z}$-valued.)

Let $c_f(\pi) = \pm 1$ be the U_π-eigenvalue of f and define a $\mathcal{D}_{\mathbb{Z}}(\mathbb{P}^1(F_\pi))$-*valued modular symbol*

$$F : \mathbb{P}^1(F) \times \mathbb{P}^1(F) \to \mathcal{D}_{\mathbb{Z}}(\mathbb{P}^1(F_\pi))$$

by the rule

$$F\{r \to s\}(\sigma\mathcal{O}_{F,\pi}) = \frac{c_f(\pi)^{\mathrm{ord}_\pi \det \sigma}}{\Omega_f} \int_r^s (f|\sigma^{-1}) \cdot \omega.$$

That $F\{r \to s\}$ is finitely additive follows from the fact that f is a U_π-eigenform. By Proposition 3.1, the distributions $F\{r \to s\}$ are all $\mathbb{Z}$-valued.

Let $\mathcal{H}_\pi$ denote the π-adic upper half-plane $\mathbb{P}^1(\mathbb{C}_p) - \mathbb{P}^1(F_\pi)$. For cusps r, s and points $\tau, \tau' \in \mathcal{H}_\pi$, define the *mixed period integral*

$$\int\!\!\!\!\!\!\diagup_\tau^{\tau'} \int_r^s f = \int\!\!\!\!\!\!\diagup_{\mathbb{P}^1(F_p)} \left(\frac{x - \tau'}{x - \tau} \right) dF\{r \to s\}(x) \in \mathbb{C}_p^*.$$

Trifković conjectures that, up to certain π-adic periods, the above mixed period integral map be written as a quotient of two indefinite integrals (defined below). Let $E_{/F}$ be a representative of the isogeny class of elliptic curves associated to f by Conjecture 2.6. Then E has multiplicative reduction over F_π and therefore a Tate uniformization over $\mathbb{C}_p$, where p is the rational prime below π.

Conjecture 3.2. *There exists a lattice $\Lambda \subset \mathbb{C}_p^*$ commensurable with the Tate lattice of E and a function*

$$\mathcal{H}_\pi \times \mathbb{P}^1(F) \times \mathbb{P}^1(F) \to \mathbb{C}_p, \quad \textit{written} \quad (\tau, r, s) \mapsto \int\!\!\!\!\!\!\diagup_r^\tau \int_r^s f, \tag{3.1}$$

such that

(1) $\int\!\!\!\!\!\!\diagup_{\gamma r}^{\gamma\tau} \int_{\gamma r}^{\gamma s} f = \int\!\!\!\!\!\!\diagup_r^\tau \int_r^s f$ *for all* $\gamma \in \tilde{\Gamma}$ *and all cusps* r, s,

(2) $\int\!\!\!\!\!\!\diagup_r^\tau \int_r^s f \times \int\!\!\!\!\!\!\diagup_s^\tau \int_s^t f = \int\!\!\!\!\!\!\diagup_r^\tau \int_r^t f$ *for all cusps* r, s, t,

$$(3)\quad \not\int_{\tau}^{\tau'}\int_{r}^{s} f = \not\int_{r}^{\tau'}\int_{r}^{s} f \Big/ \not\int_{r}^{\tau}\int_{r}^{s} f \text{ in } \mathbb{C}_p^*/\Lambda.$$

Since the Tate lattices of isogenous elliptic curves are commensurable, the above conjecture does not depend on our choice of representative E of the isogeny class of elliptic curves associated to f. We refer to the function in (3.1) as an *indefinite integral*.

4. Stark-Heegner points

Let K/F be a quadratic extension in which (π) is inert and all prime ideals dividing $\mathcal{M}$ are split (this is the analogue of the Heegner hypothesis in our situation) and let $\mathcal{O}$ be a $\mathcal{O}_F[1/\pi]$-order in K of conductor prime to $\mathcal{M}$. Let

$$R = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M_2(\mathcal{O}_F) : \mathcal{M} \text{ divides } c \right\}.$$

We say that an embedding $\psi : K \to M_2(F)$ of F-algebras is $(\mathcal{O}, R)$-*optimal* if $\psi(K) \cap R = \psi(\mathcal{O})$. Let $\mathcal{E}(\mathcal{O}, R)$ be the set of such embeddings. The conditions that the primes dividing $\mathcal{M}$ split in K and that the conductor of $\mathcal{O}$ is prime to $\mathcal{M}$ guarantee that $\mathcal{E}(\mathcal{O}, R)$ is nonempty. The group $\tilde{\Gamma}$ of units in R acts naturally on $\mathcal{E}(\mathcal{O}, R)$ by conjugation. Moreover, there is a natural free action of $\operatorname{Pic}\mathcal{O}$ on $\mathcal{E}(\mathcal{O}, R)/\tilde{\Gamma}$ which partitions $\mathcal{E}(\mathcal{O}, R)/\tilde{\Gamma}$ into $2^{\omega(\mathcal{M})}\#\operatorname{Pic}\mathcal{O}$ orbits, where $\omega(\mathcal{M})$ denotes the number of prime factors of $\mathcal{M}$. (For details, see [**Tri06**, §3.2].)

For each $\psi \in \mathcal{E}(\mathcal{O}, R)$, there is a unique $\tau_\psi \in \mathcal{H}_\pi$ such that

$$\psi(\alpha)\begin{pmatrix} \tau_\psi \\ 1 \end{pmatrix} = \alpha \begin{pmatrix} \tau_\psi \\ 1 \end{pmatrix}$$

for all $\alpha \in K^*$. As (π) is inert in K/F, the point τ_ψ actually lies in $\mathcal{H}_\pi$. Note that if ψ and ψ' are $\tilde{\Gamma}$-conjugate, then the corresponding fixed points τ and τ' in $\mathcal{H}_\pi$ are in the same $\tilde{\Gamma}$-orbit.

Fix a generator γ of $\mathcal{O}_K^*$, a cusp r, and a positive integer t such that Λ^t is contained in the Tate lattice of E. To an optimal embedding $\psi \in \mathcal{E}(\mathcal{O}, R)$, we associate the period $J_\psi \in \mathbb{C}_p^*/\Lambda$ defined by

$$J_\psi = \not\int_{r}^{\tau_\psi}\int_{r}^{\psi(\gamma)r} f \tag{4.1}$$

and the point

$$P_\psi = \operatorname{Tate}(J_\psi^t) \in E(\mathbb{C}_p).$$

By the $\tilde{\Gamma}$-invariance property of the indefinite integral (property (1) of Conjecture 3.2), the period J_ψ and the point P_ψ depends only on the $\tilde{\Gamma}$-conjugacy class of ψ. We call P_ψ the *Stark-Heegner point* attached to the optimal embedding ψ. Let $H_{\mathcal{O}}$ be the ring class field associated to the order $\mathcal{O}$ and let

$$\operatorname{rec} : \operatorname{Pic}\mathcal{O} \to \operatorname{Gal} H_{\mathcal{O}}/K$$

be the isomorphism induced by the reciprocity map of class field theory.

Conjecture 4.1 (Trifković). *The point P_ψ belongs to $E(H_{\mathcal{O}})$. The Galois action on P_ψ is described by*

$$(P_\psi)^{\operatorname{rec}(\mathfrak{a})} = P_{\psi^{\mathfrak{a}}}.$$

We expect the analogues of the formula of Gross-Zagier and the theorem of Kolyvagin to hold in this context: Assuming Conjecture 4.1, we may let

$$P_K = \mathrm{Trace}_{H_{\mathcal{O}}/K} P_\psi = \sum_{\mathfrak{a} \in \mathrm{Pic}\,\mathcal{O}} P_{\psi^{\mathfrak{a}}} \in E(K).$$

Let $\langle \cdot, \cdot \rangle$ denote the canonical height pairing on $E(K)$.

CONJECTURE 4.2. *There is an explicit nonzero fudge factor α such that*

$$L'(E/K, 1) = \alpha \langle P_K, P_K \rangle.$$

In particular, $L'(E/K, 1) \neq 0$ if and only if P_K is nontorsion.

CONJECTURE 4.3. *Suppose that the point P_K has infinite order in $E(K)$. Then* $\operatorname{rank} E(K) = 1$.

Armed with suitable nonvanishing results for twists of the L-function of E, Conjectures 4.2 and 4.3 should imply BSD for elliptic curves over F of analytic rank at most one. For a sketch of this argument, see [**Dar04**, §3.9].

5. Computing Stark-Heegner points

In the absence of proofs for the above conjectures, some numerical evidence supporting them is most desirable. Trifković provides this in abundance in the case where $\mathcal{O}_F$ is a Euclidean domain and the conductor of E is prime. In order to compute J_ψ in this case, Trifković, following Darmon and Green [**DG02**], begins by producing a candidate for the indefinite integral. Since $\mathcal{O}_F$ is a Euclidean domain, we may use Manin's continued fraction algorithm to write an arbitrary degree zero divisor $(s) - (r)$ on $\mathbb{P}^1(F)$ as

$$(s) - (r) = [(s) - (t_1)] + [(t_1) - (t_2)] + \cdots + [(t_{n-1}) - (t_n)] + [(t_n) - (r)],$$

where each pair in square brackets is a pair of adjacent cusps. (Two elements $(a : b)$ and $(c : d)$ of $\mathbb{P}^1(F)$ are called *adjacent* if $ad - bc = 1$.) Therefore, by the "path-multiplicativity" of the indefinite integral (property (2) of Conjecture 3.2), we may assume that r and s are adjacent cusps. All adjacent pairs of cusps are $\tilde{\Gamma} = \mathrm{PSL}_2(\mathcal{O}_F[1/\pi])$-equivalent. (Note that $\tilde{\Gamma} = \mathrm{PSL}_2(\mathcal{O}_F[1/\pi])$ because the conductor of E is assumed to be prime.) Therefore, by property (1) of Conjecture 3.2, we have reduced the computation of J_ψ to that of integrals of the form

$$\oint_0^\tau \int_0^\infty f.$$

Similar manipulations using the properties of the indefinite integral give the identity

$$\oint_0^\tau \int_0^\infty f = \oint_{1-\frac{1}{\tau}}^{\tau-1} \int_0^\infty f = \oint_{\mathbb{P}^1(\mathbb{Q}_p)} \left(\frac{x - (\tau - 1)}{x - (1 - 1/\tau)} \right) dF\{0 \to \infty\}(x).$$

Thus, we have reduced the calculation to that of multiplicative integrals of the form

$$\oint_{\mathbb{P}^1(\mathbb{Q}_p)} \left(\frac{x - \tau}{x - \tau'} \right) d\mu(x),$$

where $\tau, \tau' \in \mathcal{H}_\pi$ and μ is a measure on $\mathbb{P}^1(F_\pi)$. The ideas used in the computation of such integrals are the same as those discussed in [**Gre**].

We conclude with a numerical example taken from [**Tri06**, §1.3.2]. Let $F = \mathbb{Q}(\alpha)$ where $\alpha = (1+\sqrt{-3})/2$ and consider the elliptic curve

$$E : y^2 + xy = x^3 + (\alpha+1)x^2 + \alpha x.$$

The curve E has prime conductor (π) of norm 73, where $\pi = \alpha + 8$. The Mordell-Weil group $E(F)$ is cyclic of order 6 generated by the point $(-1,1)$. Let $K = F(\beta)$ where $\beta^2 = 2\alpha + 21$. Then (π) is inert in the quadratic extension K of F. Let ψ be an $(\mathcal{O}_K, M_2(\mathcal{O}_F[1/\pi]))$-optimal embedding of K into $M_2(F)$. As the class number of K is one, we expect the Stark-Heegner point P_ψ to be rational over $K = H_{\mathcal{O}_K}$. An approximation to P_ψ module π^{30} was recognized as the global point $(x,y) \in E(K)$, where

$$\begin{aligned} x &= \frac{1259988}{127165927}\alpha + \frac{126090782}{127165927} \\ y &= \left(\frac{2903147975024}{31646131095439}\alpha + \frac{11037094266063}{31646131095439}\right)\beta \\ &\quad + \frac{629994}{127165927}\alpha + \frac{63045391}{127165927}. \end{aligned}$$

References

[BMSW07] B. Bektemirov, B. Mazur, W. Stein, and M. Watkins, *Average ranks of elliptic curves: tension between data and conjecture*, Bull. Amer. Math. Soc. (N.S.) **44** (2007), no. 2, 233–254 (electronic). MR 2291676

[Byg98] J.S. Bygott, *Modular forms and modular symbols over imaginary quadratic fields*, Ph.D. thesis, University of Exeter, 1998, http://www.warwick.ac.uk/staff/J.E.Cremona/theses/bygott.pdf.

[Cre] J. E. Cremona, *Modular forms and elliptic curves over imaginary quadratic fields*, www.exp-math.uni-essen.de/zahlentheorie/preprints/PS/cremona.ps.

[Cre84] J. E. Cremona, *Hyperbolic tessellations, modular symbols, and elliptic curves over complex quadratic fields*, Compositio Math. **51** (1984), no. 3, 275–324. MR 743014 (85j:11063)

[CW94] J. E. Cremona and E. Whitley, *Periods of cusp forms and elliptic curves over imaginary quadratic fields*, Math. Comp. **62** (1994), no. 205, 407–429. MR 1185241 (94c:11046)

[Dar] H. Darmon, *Rational points on curves*, in this volume.

[Dar01] ———, *Integration on $\mathcal{H}_p \times \mathcal{H}$ and arithmetic applications*, Ann. of Math. (2) **154** (2001), no. 3, 589–639. MR 1884617 (2003j:11067)

[Dar04] ———, *Rational points on modular elliptic curves*, CBMS Regional Conference Series in Mathematics, vol. 101, Published for the Conference Board of the Mathematical Sciences, Washington, DC, 2004. MR 2020572 (2004k:11103)

[Das05] S. Dasgupta, *Stark-Heegner points on modular Jacobians*, Ann. Sci. École Norm. Sup. (4) **38** (2005), no. 3, 427–469. MR 2166341 (2006e:11080)

[DG02] H. Darmon and P. Green, *Elliptic curves and class fields of real quadratic fields: algorithms and evidence*, Experiment. Math. **11** (2002), no. 1, 37–55. MR 1960299 (2004c:11112)

[Gre] M. Greenberg, *Computing Heegner points arising from Shimura curve parametrizations*, in this volume.

[Kur78] P. F. Kurčanov, *The cohomology of discrete groups and Dirichlet series that are related to Jacquet-Langlands cusp forms*, Izv. Akad. Nauk SSSR Ser. Mat. **42** (1978), no. 3, 588–601, English translation in: Math. USSR-Izv. 12 (1978), no. 3, 543–555. MR 503433 (80b:10038)

[Lin05] M. Lingham, *Modular forms and elliptic curves over imaginary quadratic fields*, Ph.D. thesis, University of Nottingham, 2005, http://etheses.nottingham.ac.uk/archive/00000138.

[Tri06] M. Trifković, *Stark-Heegner points on elliptic curves defined over imaginary quadratic fields*, Duke Math. J. **135** (2006), no. 3, 415–453. MR 2272972 (2008d:11064)

Department of Mathematics, McGill University, Montreal, Quebec, Canada
Current address: Max-Planck-Institut für Mathematik, 53111 Bonn, Germany
E-mail address: fmgreenberg@gmail.com

Clay Mathematics Proceedings
Volume 8, 2009

Lectures on Modular Symbols

Yuri I. Manin

Abstract. In these lecture notes, written for the Clay Mathematics Institute Summer School "Arithmetic Geometry", Göttingen 2006, I review some classical and more recent results about modular symbols for $SL(2)$, including arithmetic motivations and applications, an iterated version of modular symbols, and relations with the "non–commutative boundary" of the modular tower for elliptic curves.

1. Introduction: arithmetic functions and Dirichlet series

1.1. Arithmetic functions. Many basic questions of number theory involve the behavior of *arithmetic functions*, i.e. sequences of integers $\{a_n \mid n \geq 1\}$ defined in terms of divisors of n, or numbers of solutions of a congruence modulo n, etc. After having chosen such a function, one might ask for example:

(i) Is $\{a_n \mid n \geq 1\}$ multiplicative, that is, does $a_{mn} = a_m a_n$ for $(m, n) = 1$?

(ii) What is the asymptotic behavior of $\sum_{n \leq N} a_n$ as $N \to \infty$?

(iii) Can one give a "formula" for a_n if initially it was introduced only by a computational prescription, such as $a_n :=$ *the number of representations of* n *as a sum of four squares*?

A very universal machinery for studying such questions consists in introducing *a generating series* for a_n depending on a complex parameter, and studying the analytic and algebraic properties of this series.

Two classes of series that are used most often are the Fourier series

$$f(z) := \sum_{n=1}^{\infty} a_n e^{2\pi i n z} \tag{1.1}$$

and the Dirichlet series

$$L_f(s) = \sum_{n=1}^{\infty} a_n n^{-s}. \tag{1.2}$$

In full generality, they must be considered as formal series; however, if a_n does not grow too fast, e.g. is bounded by a polynomial in n, then (1.1) converges in the

2000 *Mathematics Subject Classification.* Primary 11F67, Secondary 11F11, 11F75, 11M41.

upper half-plane $H := \{z \in \mathbf{C} \,|\, \operatorname{Im} z > 0\}$, whereas (1.2) converges in some right half plane $\operatorname{Re} s > D$.

1.2. Mellin transform and modularity. Some of the properties of $\{a_n\}$ are directly encoded in the generating Dirichlet series. For example, multiplicativity of $\{a_n\}$ translates into the existence of an Euler product over primes p:

$$L_f(s) = \prod_p L_{f,p}(s), \qquad L_{f,p}(s) := \sum_{n=1}^{\infty} a_{p^n} p^{-ns}. \tag{1.3}$$

Hence the Dirichlet series for the logarithmic derivative of such a function carries information about the values of a_n restricted to powers of primes. This idea leads to famous "explicit formulas" expressing partial sums of a_{p^n}'s via poles of the logarithmic derivative of $L_f(s)$ i.e. essentially zeroes of $L_f(s)$. Applied to the simplest multiplicative sequence $a_n = 1$ for all n, this formalism produces the classical relationship between primes and zeroes of Riemann's zeta.

It turns out, however, that to establish the necessary analytic properties of $L_f(s)$ such as the analytic continuation in s and a functional equation, and generally even the existence of an Euler product, one should focus first upon the Fourier series $f(z)$. The main reason for this is that interesting functions $f(z)$ more often than not possess, besides the obvious periodicity under $z \mapsto z+1$, additional symmetries, for example, a simple behavior with respect to the substitution $z \mapsto -z^{-1}$. This is the case for $f(z) = \sum_{n\geq 1} e^{2\pi i n^2 z}$ (or the more symmetric $\sum_{n\in\mathbf{Z}} e^{2\pi i n^2 z}$) corresponding to $L_f(s) = \zeta(2s)$.

The transformations $z \mapsto z+1$ and $z \mapsto -z^{-1}$ together generate the full modular group $PSL(2,\mathbf{Z})$ of fractional linear transformations of H, and Fourier series of various *modular forms* with respect to this group and its subgroups of finite index generate a vast supply of most interesting arithmetic functions.

The basic relation between $f(z)$ and $L_f(s)$ allowing one to translate analytic properties of $f(z)$ into those of $L_f(s)$ is the integral *Mellin transform*

$$\Lambda_f(s) := \int_0^{i\infty} f(z) \left(\frac{z}{i}\right)^s \frac{dz}{z}. \tag{1.4}$$

Here the s–th power in the integrand is interpreted as the branch of the exponential function which takes real values for real s and imaginary z. Convergence at $i\infty$ is usually automatic whereas convergence at 0 is justified by a functional equation (possibly after disposing of a controlled singularity).

Whenever we can integrate termwise using (1.1) (for large $\operatorname{Re} s$), an easy calculation shows that

$$\Lambda_f(s) = (2\pi)^{-s}\Gamma(s)L_f(s). \tag{1.5}$$

A functional equation for $f(z)$ with respect to $z \mapsto -z^{-1}$ (or more generally, $z \mapsto -(Nz)^{-1}$ for some N) then leads formally to a functional equation of Riemann type connecting $\Lambda_f(s)$ with $\Lambda_f(1-s)$ or $\Lambda_f(D-s)$ for an appropriate D defining *the critical strip* $0 \leq \operatorname{Re} s \leq D$ for $L_f(s)$.

This is a very classical story, which acquired its final shape in the work of Hecke in the 1920's and 30's. More modern insights concern the role of Γ–factors as Euler factors at *arithmetic infinity*, and most important, the universality of this

picture and the existence of its vast generalizations crystallized in the *Taniyama–Weil conjecture* and the so-called *Langlands program*. This involves, in particular, consideration of much more general arithmetic groups than $PSL(2)$ as modular groups.

We will not discuss this vast development in these lectures and focus upon the classical modular group and related modular symbols. For some generalizations, see [**AB90**], [**AR79**].

2. Classical modular symbols and Shimura integrals

2.1. Modular symbols as integrals. Since we are interested in Mellin transforms of the form (1.4) where $f(z)$ has an appropriate modular behavior with respect to a subgroup of $PSL(2,\mathbf{Z})$, we must keep track of similar integrals taken over $PSL(2,\mathbf{Z})$–images of the upper semi–axis as well. The latter are geodesics connecting two *cusps* in the partial compactification $\overline{H} := H \cup \mathbf{P}^1(\mathbf{Q})$.

Roughly speaking, the *classical modular symbols* are linear functionals (spanned by)

$$\{\alpha,\beta\} \,:\, f \mapsto \int_\alpha^\beta f(z) z^{s-1} dz, \qquad \alpha,\beta \in \mathbf{P}^1(\mathbf{Q})$$

on appropriate spaces of 1–forms $f(z)z^{s-1}dz$. To be more precise, we must recall the following definitions.

The group of real matrices with positive determinant $GL^+(2,\mathbf{R})$ acts on H by fractional linear transformations $z \mapsto [g]z$. Let $j(g,z) := cz+d$ where (c,d) is the lower row of g. Then we have, for any function f on H and homogeneous polynomial $P(X,Y)$ of degree $k-2$,

$$\begin{aligned} g^*[f(z)\,P(z,1)\,dz] &:= f([g]z)\,P([g]z,1)\,d([g]z) \\ &= f([g]z)\,(j(g,z))^{-k} P(az+b,cz+d)\det g\,dz \end{aligned} \tag{2.1}$$

where (a,b) is the upper row of g. From the definition it is clear that the diagonal matrices act identically so that we have in fact an action of $PGL^+(2,\mathbf{R})$.

This action induces for any integer $k \geq 2$ the weight k action of $GL^+(2,\mathbf{R})$ on functions on H. In the literature one finds two different normalizations of such an action. They differ by a determinantal twist and therefore coincide on $SL(2,\mathbf{R})$ and the modular group. For example, in [**Mer94**] and [**Man06**] the action

$$f\,|[g]_k(z) := f([g]z)\,j(g,z)^{-k}\,(\det g)^{k/2} \tag{2.2}$$

is used.

A holomorphic function $f(z)$ on H is a modular form of weight k for a group $\Gamma \subset SL(2,\mathbf{R})$ if $f|[\gamma]_k(z) = f(z)$ for all $\gamma \in \Gamma$ and $f(z)$ is finite at cusps.

Such a form is called a cusp form if it vanishes at cusps.

Let $S_k(\Gamma)$ be the space of cusp forms of weight k. Denote by $Sh_k(\Gamma)$ the space of 1–forms on the complex upper half plane H of the form $f(z)\,P(z,1)\,dz$ where $f \in S_k(\Gamma)$, and $P = P(X,Y)$ runs over homogeneous polynomials of degree $k-2$ in two variables. Thus, the space $Sh_k(\Gamma)$ is spanned by 1–forms *of cusp modular type with integral Mellin arguments in the critical strip* in the terminology of [**Man06**], Def. 2.1.1, and 3.3 below.

We will now describe the space of classical modular symbols $MS_k(\Gamma)$ as the space of *linear functionals* on $S_k(\Gamma)$ spanned by the Shimura integrals

$$f(z) \mapsto \int_\alpha^\beta f(z) z^{m-1} dz; \quad 1 \le m \le k-1; \quad \alpha, \beta \in \mathbf{P}^1(\mathbf{Q}). \tag{2.3}$$

Three descriptions of $MS_k(\Gamma)$ are known:

(i) *Combinatorial (Shimura-Eichler-Manin)*: generators and relations.

(ii) *Geometric (Shokurov)*: $MS_k(\Gamma)$ can be identified with a (part of) the middle homology of the Kuga–Sato variety $M^{(k)}$.

(iii) *Cohomological (Shimura)*: The dual space to $MS_k(\Gamma)$ can be identified with the cuspidal group cohomology $H^1(\Gamma, W_{k-2})_{cusp}$, with coefficients in the $(k-2)$-nd symmetric power of the basic representation of $SL(2)$.

We give some details below.

2.2. Combinatorial modular symbols. In this description, $MS_k(\Gamma)$ appears as an explicit subquotient of the space $W_{k-2} \otimes \overline{C}$, where W_{k-2} consists of polynomial forms $P(X,Y)$ of degree $k-2$ of two variables, and $\overline{C}$ is the space of formal linear combinations of pairs of cusps $\{\alpha, \beta\} \in \mathbf{P}^1(\mathbf{Q})$. Coefficients of these linear combinations can be $\mathbf{Q}$, $\mathbf{R}$ or $\mathbf{C}$, as in the theory of Hodge structures.

Each element of the form $P \otimes \{\alpha, \beta\}$ produces a linear functional

$$f \mapsto \int_\beta^\alpha f(z)\, P(z,1) dz.$$

This is extended to all of $W_{k-2} \otimes \overline{C}$ by linearity.

Denote by C the quotient of $\overline{C}$ by the subspace generated by sums $\{\alpha, \beta\} + \{\beta, \gamma\} + \{\gamma, \alpha\}$. Since $\int_\beta^\alpha + \int_\gamma^\beta + \int_\alpha^\gamma = 0$, our linear functional (Shimura integral) descends to $W_{k-2} \otimes C$. We will still denote by $P \otimes \{\alpha, \beta\}$ the class of this element in C.

The group $GL^+(2, \mathbf{Q})$ acts from the left on W_{k-2} by (notation as in (2.1))

$$(gP)(X,Y) := P(bX - dY, -cX + aY),$$

and on C by $g\{\alpha, \beta\} := \{g\alpha, g\beta\}$. Hence it acts on the tensor product. A change of variable formula then shows that the Shimura integral restricted to $S_k(\Gamma)$ vanishes on the subspace of $W_{k-2} \otimes C$ spanned by $P \otimes \{\alpha, \beta\} - gP \otimes \{g\alpha, g\beta\}$ for all $P \in W_{k-2}$, $g \in \Gamma$.

Denote by $MS_k(\Gamma)$ the quotient of $W_{k-2} \otimes C$ by the latter subspace.

The subspace of cuspidal modular symbols $MS_k(\Gamma)_{cusp}$ is defined by the following construction. Consider the space B freely spanned by $\mathbf{P}^1(\mathbf{Q})$. Define the space $B_k(\Gamma)$ as the quotient of $W_{k-2} \otimes B$ by the subspace generated by $P \otimes \{\alpha\} - gP \otimes \{g\alpha\}$ for all $g \in \Gamma$. There is a well-defined boundary map $MS_k(\Gamma) \to B_k(\Gamma)$ induced by $P \otimes \{\alpha, \beta\} \mapsto P \otimes \{\alpha\} - P \otimes \{\beta\}$. Its kernel is denoted $MS_k(\Gamma)_{cusp}$.

By construction, any (real) modular symbol in $MS_k(\Gamma)_{cusp}$ defines a $\mathbf{C}$–valued functional $\int$ on $S_k(\Gamma)$ and in fact even on $S_k(\Gamma) \oplus \overline{S}_k(\Gamma)$.

The first result of the theory is:

Theorem (Shimura). *$\int$ is an isomorphism of $MS_k(\Gamma)_{cusp}$ with the dual space of $S_k(\Gamma) \oplus \overline{S}_k(\Gamma)$.*

2.3. Geometric modular symbols. Let $\Gamma^{(k)}$ be the semidirect product $\Gamma \ltimes (\mathbf{Z}^{k-2} \times \mathbf{Z}^{k-2})$ acting on $H \times \mathbf{C}^{k-2}$ via

$$(\gamma;\, n, m)\,(z, \zeta) := ([\gamma]z;\, j(\gamma, z)^{-1}(\zeta + zn + m))$$

where $n = (n_1, \ldots, n_{k-2})$, $m = (m_1, \ldots, m_{k-2})$, $\zeta = (\zeta_1, \ldots, \zeta_{k-2})$, and $nz = (n_1 z, \ldots, n_{k-2} z)$.

If $f(z)$ is a Γ–invariant cusp form of weight k, then

$$f(z) dz \wedge d\zeta_1 \wedge \cdots \wedge d\zeta_{k-2}$$

is a $\Gamma^{(k)}$–invariant holomorphic volume form on $H \times \mathbf{C}^{k-2}$. Hence one can push it down to a Zariski open smooth subset of the quotient $\Gamma^{(k)} \setminus (H \times \mathbf{C}^{k-2})$. An appropriate smooth compactification $M^{(k)}$ of this subset is called *a Kuga–Sato variety,* cf. [**Sho76**],[**Sho80b**],[**Sho80a**].

Denote by ω_f the image of this form on $M^{(k)}$. Notice that it depends only on f, not on any Mellin argument. The latter can be accommodated in the structure of (relative) cycles in $M^{(k)}$, so that integrating ω_f over such cycles we obtain the respective Shimura integrals.

Concretely, let $\alpha, \beta \in \mathbf{P}^1(\mathbf{Q})$ be two cusps in $\overline{H}$ and let p be a geodesic joining α to β. Fix (n_i) and (m_i) as above. Construct a cubic singular cell $p \times (0,1)^{k-2} \to H \times \mathbf{C}^{k-2}$: $(z, (t_i)) \mapsto (z, (t_i(zn_i + m_i)))$. Take the S_{k-2}–symmetrization of this cell and push down the result to the Kuga–Sato variety. We will get a relative (modulo fibers of $M^{(k)}$ over cusps) cycle whose homology class is Shokurov's higher modular symbol $\{\alpha, \beta;\, n, m\}_\Gamma$. One easily sees that

$$\int_\alpha^\beta f(z) \prod_{i=1}^{k-2} (n_i z + m_i)\, dz = \int_{\{\alpha, \beta;\, n, m\}_\Gamma} \omega_f\,.$$

The singular cube $(0,1)^{k-2}$ may also be replaced by an evident singular simplex.

Theorem (Shokurov). *(i) The map $f \mapsto \omega_f$ is an isomorphism $S_k(\Gamma) \to H^0(M^{(k)}, \Omega^{k-1}_{M^{(k)}})$.*

(ii) The homology subspace spanned by Shokurov modular symbols with vanishing boundary is canonically isomorphic to the space of cuspidal combinatorial modular symbols.

2.4. Cohomological modular symbols. In this description, the space dual to $MS_k(\Gamma)$ is identified with the group cohomology $H^1(\Gamma, W_{k-2})$.

A bridge between the geometric and the cohomological descriptions is furnished by the identification of $H^1(\Gamma, W_{k-2})_{cusp}$ with the cohomology of a local system on $M_{1,1}$, namely $H^1_!(M_{1,1}, \mathrm{Sym}^{k-2} R^1\pi_* \mathbf{Q})$.

2.5. Some arithmetic applications. The formalism sketched above allows one to get some quite precise information about two classes of number–theoretic objects: *coefficients* of modular forms and *their periods*, which are essentially values of their Mellin transforms at integer points of the critical strip. For illustration, we give two examples taken from [**Man72**] and [**Man73**].

Example 1. Let

$$\Phi(z) := e^{2\pi iz}\prod_{n=1}^{\infty}(1-e^{2\pi niz})^{24} = \sum_{n=1}^{\infty}\tau(n)e^{2\pi niz}.$$

The coefficients $\tau(n)$ form a multiplicative sequence. This follows from the fact that $\Phi(z)$ is the (essentially unique) cusp form of weight 12 with respect to the full modular group; hence in particular it is an eigenform for all Hecke operators, which ensures multiplicativity.

The formalism of modular symbols leads to an expression for $\tau(n)$ through representations of n by an indefinite quadratic form. Namely, we have

$$\tau(n) = \sum_{d/n} d^{11} + \sum_{n=\Delta\Delta'+\delta\delta'} \frac{691}{18}(\Delta^8\delta^2 - \Delta^2\delta^8) + \frac{691}{6}(\Delta^6\delta^4 - \Delta^4\delta^6). \qquad (2.4)$$

The second summation is taken over the following set of solutions: we require that $\Delta > \delta > 0$ and either $\Delta' > \delta' > 0$, or $\Delta/n, \Delta' = n/\Delta, \delta' = 0, 0 < \delta/\Delta \le 1/2$.

Periods of $\Phi(z)$ are Shimura integrals

$$r_k(\Phi) := \int_0^{i\infty} \Phi(z)z^k dz, \qquad 0 \le k \le 10 - w$$

that is, via Mellin transform,

$$r_k(\Phi) = \frac{k!i^{k+1}}{(2\pi)^{k+1}} L_\Phi(k+1).$$

The invariance of $\Phi(z)(dz)^6$ with respect to $z \mapsto -z^{-1}$ shows that

$$r_k(\Phi) = r_k(\Phi)(-1)^{k+1}r_{10-k}(\Phi).$$

Finally, the formalism of modular symbols allows one to establish that the $\mathbf{Q}$–space spanned by periods is at most two–dimensional. More precisely,

$$(r_0 : r_2 : r_4) = (1 : -\frac{691}{2^2\cdot 3^4\cdot 5} : \frac{691}{2^3\cdot 3^2\cdot 5\cdot 7}),\ (r_1 : r_3 : r_5) = (1 : -\frac{5^2}{2^4\cdot 3} : \frac{5}{2^2\cdot 3}).$$

Example 2: a non–commutative reciprocity law. Here we start with a cusp form of weight two

$$F(z) := e^{2\pi iz}\prod_{n=1}^{\infty}(1-e^{2\pi niz})^2(1-e^{22\pi niz})^2 = \sum_{n=1}^{\infty}\lambda_n e^{2\pi niz}$$

with respect to the subgroup $\Gamma_0(11)$ of Γ.

The Mellin transform of this form can be identified with the Weil zeta function of the elliptic modular curve $\Gamma_0(11)\setminus\overline{H}$ defined over $\mathbf{Q}$. From this it follows that for any prime $p \ne 2, 11$, we can characterize $1 - \lambda_p + p$ as the number of solutions of the congruence

$$y^2 + y \equiv x^3 - x^2 - 10x - 20 \bmod p \qquad (2.5)$$

(including the infinite solution).

On the other hand, the formalism of modular symbols allows one to write for this number an expression having the same structure as (2.4):

$$1 - \lambda_p + p = \sum_{p=\Delta\Delta'+\delta\delta'} y_{11}(\Delta, \delta). \tag{2.6}$$

This time, however, $y_{11}(\Delta, \delta)$ is not a polynomial: it depends only on $(\Delta : \delta) \bmod 11$: for the values of the latter $0, \infty, \pm 1, \pm 2, \pm 3, \pm 4, \pm 5 \bmod 11$, the values of y_{11} are respectively $2, -2, 0, 10, 5, -5, -10$.

Thus, we have connected solutions modulo p of the equation (2.5) "depending on 11" as its conductor with solutions modulo 11 of the equation $p = \Delta\Delta' + \delta\delta'$ depending on p. This justifies the name "non–commutative reciprocity law" suggested for (2.6) and its generalizations in [**Man72**].

Such formulas can be used to make more explicit the exact arithmetic content of special cases of the very general and therefore somewhat abstruse Langlands formalism.

Proofs of formulas for coefficients such as (2.4), (2.6) consist of two steps. For simplicity, we will illustrate this for the case of weight two cusp form $f(z)$ which is an eigenform with respect to a Hecke operator T_n so that $T_n f = a_n f$. We integrate this identity, say, from 0 to $i\infty$ and get

$$\int_0^\infty T_n f dz = a_n \int_0^{i\infty} f dz.$$

Now, use the explicit definition of the Hecke operator T_n on the left hand side and make a change of variables. We will get a sum of modular symbols. Using a continued fraction trick and a lemma initially proved by Heilbronn, we finally reduce the left hand side to a sum over solutions of $n = \Delta\Delta' + \delta\delta'$.

2.6. Relations with noncommutative geometry and a real analog of p–adic integration. The role of the upper half-plane in our constructions is of course explained by the fact that it parametrizes elliptic curves: complex tori $\mathbf{C}/\langle 1, \tau\rangle$, $\tau \in H$. The action of the modular group extends to this family, and the respective quotient is a non–complete algebraic variety. The cusps $\tau \in \mathbf{P}^1(\mathbf{Q})$ can be added to compactify this quotient by degenerate elliptic curves. However, for irrational values $\theta \in \mathbf{R} \setminus \mathbf{Q}$, the quotient $\mathbf{C}/\langle 1, \theta\rangle = \mathbf{C}^*/(e^{2\pi i\theta})$ is a "bad" topological group, and the common wisdom is that it is best represented by a non–commutative space, (a version of) the quantum torus T_θ.

Tori T_θ are parametrized by $\theta \in \mathbf{R}$. However, if one considers only tori modulo Morita equivalence, then they are parametrized by $PGL(2, \mathbf{Z}) \setminus \mathbf{P}^1(\mathbf{R})$. Set–theoretically, $PGL(2, \mathbf{Z}) \setminus \mathbf{P}^1(\mathbf{R})$ = the set of equivalence classes of $\alpha \in [0, 1)$ modulo the relation

$$\alpha \equiv \beta \;\Leftrightarrow\; \exists\, n_0, n_1\ \forall n > 0, \quad k_{n+n_0}(\alpha) = k_{n+n_1}(\beta).$$

Here $k_n(\alpha)$ are successive components of the continued fraction of α.

Thus, we can imagine an "invisible boundary" of the modular tower supporting a family of non–commutative spaces, the phantom of the classical modular family.

This viewpoint was discussed in [**MM02**], see also [**Mar05**], and in particular the Gauss problem on the distribution of continued fractions and its generalizations were treated as a measure theory on the "non–commutative modular curves".

We will describe here one result of this study, which produces an "∞–adic analogue" of the theory of p–adic integration used to construct p–adic Mellin transforms of cusp forms in [**Man73**].

Fix a prime number $N > 0$ and put $G_0 = \Gamma_0(N)$. We will assume that the genus of $X_{G_0} = X_0(N)$ is ≥ 1. Consider a $\Gamma_0(N)$–invariant differential $\omega = f(z)dz$ on H such that $f(z)$ is a cusp eigenform of weight two for all Hecke operators and denote by $L_f^{(N)}(s)$ (resp. $\zeta^{(N)}(s)$) its Mellin transform (resp. Riemann's zeta) with omitted Euler N–factor. More precisely, the coefficients of $L_f^{(N)}(s)$ are Hecke eigenvalues of f.

For $\alpha \in (0,1)$, denote by $p_n(\alpha)/q_n(\alpha)$ the n–th convergent of α.

Theorem. *We have for* $\operatorname{Re} t > 0$*:*

$$\int_0^1 d\alpha \sum_{n=0}^{\infty} \frac{q_{n+1}(\alpha) + q_n(\alpha)}{q_{n+1}(\alpha)^{1+t}} \int_0^{q_n(\alpha)/q_{n+1}(\alpha)} f(z)\, dz =$$

$$\left[\frac{\zeta(1+t)}{\zeta(2+t)} - \frac{L_f^{(N)}(2+t)}{\zeta^{(N)}(2+t)^2}\right] \int_0^{i\infty} f(z)\, dz. \tag{2.7}$$

If $\int_0^{i\infty} f(z)\, dz \neq 0$, we can read (2.7) as an expression for $L_f^{(N)}(s)$ which has striking structural similarities to the p–adic Mellin integral. In particular, both formulas involve a construction of a measure out of modular symbols, on $(0,1)$ and on $\mathbf{Z}_p^*$ respectively.

The proof of (2.7) given in [**MM02**] combines an old lemma by P. Lévy with the continued fractions trick alluded to above.

The Theorem above does not involve directly the non–commutative geometry of the invisible boundary. However, it was shown in [**MM02**], Sec. 4, and [**Mar05**], Sec. 6 of Ch. 4, that modular symbols themselves can be identified with specific elements in the K–theory of this space, giving additional weight to the geometric intuition behind this picture.

3. Iterated modular symbols

3.1. Multiple zeta values and iterated integrals. The theory of iterated modular symbols (cf. [**Man06**], [**Man05**]) is a simultaneous generalization of two constructions–of classical modular symbols and of multiple zeta values–and is an elaboration of a special case of Chen's iterated integrals theory ([**Che77**]) in a holomorphic setting.

Multiple zeta values are the numbers given by the k–multiple Dirichlet series

$$\zeta(m_1, \dots, m_k) = \sum_{0<n_1<\dots<n_k} \frac{1}{n_1^{m_1} \cdots n_k^{m_k}} \tag{3.1}$$

which converge for all integer $m_i \geq 1$ and $m_k > 1$, or equivalently by the m–multiple iterated integrals, $m = m_1 + \cdots + m_k$,

$$\zeta(m_1, \ldots, m_k) = \int_0^1 \frac{dz_1}{z_1} \int_0^{z_1} \frac{dz_2}{z_2} \int_0^{z_2} \cdots \int_0^{z_{m_{k-1}}} \frac{dz_{m_k}}{1 - z_{m_k}} \cdots \tag{3.2}$$

where the sequence of differential forms in the iterated integral consists of consecutive subsequences of the form $\frac{dz}{z}, \ldots, \frac{dz}{z}, \frac{dz}{1-z}$ of lengths $m_k, m_{k-1}, \ldots, m_1$.

Easy combinatorial considerations allow one to express in two different ways products $\zeta(l_1, \ldots, l_j) \cdot \zeta(m_1, \ldots, m_k)$ as linear combinations of multiple zeta values.

If one uses for this the integral representation (3.2), one gets a sum over shuffles which enumerate the simplices of highest dimension occurring in the natural simplicial decomposition of the product of two integration simplices.

If one uses instead (3.1), one gets sums over shuffles with repetitions which enumerate some simplices of lower dimension as well.

These relations and their consequences are called double shuffle relations. Both types of relations can be succinctly written down in terms of formal series on free noncommuting generators. One can include in these relations regularized multiple zeta values for arguments where the convergence of (3.1), (3.2) fails. A clear and systematic exposition of these results can be found in [**Del01**] and [**Rac00**], [**Rac02**].

In fact, the formal generating series for (regularized) iterated integrals (3.2) appeared in the famous Drinfeld paper [**Dri90**], essentially as *the Drinfeld associator*, and more relations for multiple zeta values were implicitly deduced there. The question about interdependence of (double) shuffle and associator relations does not seem to be settled at the moment of writing this: cf. [**Rac04**]. The problem of completeness of these systems of relations is equivalent to some difficult transcendence questions.

Multiple zeta values are interesting, because they and their generalizations appear in many different contexts involving mixed Tate motives ([**DG05**], [**Ter02**]), deformation quantization ([**Kon99**]), knot invariants, etc.

In order to make contact with modular symbols, notice first that the differentials $\frac{dz}{z}, \frac{dz}{1-z}$ span the space of meromorphic differential forms with no more than logarithmic singularities at points $\{0, 1, \infty\}$ of $\mathbf{P}^1(\mathbf{C})$. We can identify

$$(\mathbf{P}^1(\mathbf{C}), \{0, 1, \infty\}) \cong \Gamma_0(4) \setminus (\overline{H}, cusps).$$

Then $\frac{dz}{z}, \frac{dz}{1-z}$ lift to Eisenstein series of weight two for $\Gamma_0(4) \subset SL(2, \mathbf{Z})$.

In the general theory sketched below, $\Gamma_0(4)$ is replaced by an arbitrary (congruence) subgroup Γ of $SL(2, \mathbf{Z})$, Eisenstein series of weight two are replaced by (cusp form + Eisenstein series) with respect to Γ, multiplied by $z^{s-1}dz$ for appropriate s. (We mostly focus on cusp forms; in the presence of logarithmic singularities, the necessary regularization procedure is described for weight two in Sec. 3.6.)

Finally, ordinary integrals along geodesics connecting two cusps are replaced by iterated integrals.

3.2. Formalism of iterated integrals. We will work on a Riemann surface, and study general iterated integrals of holomorphic 1–forms. We will show that if one replaces a simple integral not by an individual iterated integral but by a generating series of all such integrals, then the usual properties like additivity and variable change formula reappear in a multiplicative/noncommutative version.

Let X be a connected complex Riemann surface, and $\omega_V := (\omega_v \,|\, v \in V)$ a family of holomorphic 1–forms indexed by a finite set V. Denote by $A_V := (A_v \,|\, v \in V)$ free associative formal variables, commuting with complex numbers, functions, and differentials on X, and put

$$\Omega := \sum_{v\in V} A_v\omega_v.$$

Consider the total iterated integral of Ω along a piecewise smooth path $\gamma : [0,1] \to U \subset X$:

$$J_\gamma(\Omega) := 1 + \sum_{n=1}^{\infty} \int_0^1 \gamma^*(\Omega)(t_1) \int_0^{t_1} \gamma^*(\Omega)(t_2) \cdots \int_0^{t_{n-1}} \gamma^*(\Omega)(t_n) \in \mathbf{C}\langle\langle A_V\rangle\rangle$$

taken over the simplex $0 < t_n < \cdots < t_1 < 1$. If γ, γ' with the same ends are homotopic then $J_\gamma(\Omega) = J_{\gamma'}(\Omega)$. Fixing implicitly such a homotopy class, we can use another notation: $z_i = \gamma(t_i) \in X$, $a = \gamma(0)$, $z = \gamma(1)$,

$$J_a^z(\Omega) := 1 + \sum_{n=1}^{\infty} \int_a^z \Omega(z_1) \int_a^{z_1} \Omega(z_2) \cdots \int_a^{z_{n-1}} \Omega(z_n)\,.$$

If $U \subset X$ is connected and simply connected, this is an unambiguously defined element of $\mathcal{O}_X(U)\langle\langle A_V\rangle\rangle$. Otherwise it is a multivalued function of z in this domain.

Proposition. *(i) $J_a^z(\Omega)$ as a function of z satisfies the equation*

$$dJ_a^z(\Omega) = \Omega(z)\, J_a^z(\Omega).$$

In other words, $J_a^z(\Omega)$ is a horizontal (multi)section of the flat connection $\nabla_\Omega := d - l_\Omega$ on $\mathcal{O}_X\langle\langle A_V\rangle\rangle$, where l_Ω is the operator of left multiplication by Ω.

(ii) If U is a simply connected neighborhood of a, $J_a^z(\Omega)$ is the only horizontal section with initial condition $J_a^a = 1$. Any other horizontal section K^z can be uniquely written in the form $CJ_a^z(\Omega)$, $C \in \mathbf{C}\langle\langle A_V\rangle\rangle$. In particular, for any $b \in U$,

$$J_b^z(\Omega) = J_a^z(\Omega) J_b^a(\Omega).$$

Corollary. *Let γ be a closed oriented contractible contour in U, $a_1, \ldots, a_n$ points along this contour (cyclically) ordered compatibly with orientation. Then*

$$J_{a_2}^{a_1}(\Omega) J_{a_3}^{a_2}(\Omega) \cdots J_{a_n}^{a_{n-1}}(\Omega) J_{a_1}^{a_n}(\Omega) = 1. \tag{3.3}$$

Formula (3.3) is the multiplicative version of the additivity of simple integrals with respect to the join of integration paths.

Proposition. *Consider the comultiplication*

$$\Delta : \mathbf{C}\langle\langle A_V\rangle\rangle \to \mathbf{C}\langle\langle A_V\rangle\rangle \widehat{\otimes}_{\mathbf{C}} \mathbf{C}\langle\langle A_V\rangle\rangle,\ \Delta(A_v) = A_v \otimes 1 + 1 \otimes A_v$$

and extend it to the series with coefficients $\mathbf{C}(X)$ and $\Omega^1(X)$. Then

$$\Delta\left(J_a^z(\omega_V)\right) = J_a^z(\omega_V) \widehat{\otimes}_{\mathcal{O}_X} J_a^z(\omega_V)\,. \tag{3.4}$$

Claim 1. *The identity (3.4) encodes all shuffle relations between the iterated integrals of the forms* ω_v.

Claim 2. *The identity (3.4) is equivalent to the fact that* $\log J_a^z(\omega_V)$ *can be expressed as a series in commutators (of arbitrary length) of the variables* A_v.

Formula (3.4) expresses the group–like property of $J_a^z(\Omega)$. It is a multiplicative version of the additivity of a simple integral as a functional of the integration form.

Functorality. Let $g : X \to X$ be an automorphism such that g^* maps into itself the linear space spanned by ω_v: $g^*(\omega_v) = \sum_u g_{vu}\omega_u$. Define $g_*(A_u) = \sum_v A_v g_{vu}$. Then we have

$$J_{ga}^{gz}(\omega_V) = g_*(J_a^z(\omega_V))\,. \tag{3.5}$$

Formula (3.5) is a multiplicative version of the variable change formula.

3.3. Iterated integrals on the upper half–plane and total Mellin transform. A 1–form ω on H will be called a form of modular type if it can be represented as $f(z)z^{s-1}dz$, where s is a complex number and $f(z)$ is a modular form of some weight with respect to a finite index subgroup Γ of the modular group $SL(2,\mathbf{Z})$.

The modular form $f(z)$ is then well defined and called the associated modular form (to ω), and the number s is called the Mellin argument of ω.

ω is called a form of cusp modular type if the associated $f(z)$ is a cusp form.

Let $f_1,\dots,f_k$ be a finite sequence of cusp forms with respect to Γ, $\omega_j(z) := f_j(z)\, z^{s_j-1}dz$. The iterated Mellin transform of (f_j) is

$$M(f_1,\dots,f_k;s_1,\dots,s_k) := I_{i\infty}^0(\omega_1,\dots,\omega_k) =$$

$$\int_{i\infty}^0 \omega_1(z_1)\int_{i\infty}^{z_1}\omega_2(z_2)\cdots\int_{i\infty}^{z_{n-1}}\omega_n(z_n).$$

Let $f_V = (f_v \,|\, v \in V)$ be a finite family of cusp forms with respect to Γ, $s_V = (s_v \,|\, v \in V)$ a finite family of complex numbers, $\omega_V = (\omega_v)$, where $\omega_v(z) := f_v(z)\, z^{s_v-1}dz$. The total Mellin transform of f_V is

$$TM(f_V;s_V) := J_{i\infty}^0(\omega_V) =$$

$$1+\sum_{n=1}^{\infty}\sum_{(v_1,\dots,v_n)\in V^n} A_{v_1}\dots A_{v_n}\, M(f_{v_1},\dots,f_{v_n};s_{v_1},\dots,s_{v_n}).$$

Theorem. *Assume that the space spanned by* $f_v(z)$ *is stable with respect to* $g_N : z \mapsto -1/Nz$. *Let* k_v *be the weight of* $f_v(z)$, *and* $k_V = (k_v)$. *Then*

$$TM(f_V;s_V) = g_{N*}(TM(f_V;k_V - s_V))^{-1}$$

for an appropriate linear transformation g_{N*} *of the formal variables* A_v.

3.4. Iterated Shimura integrals and non–commutative cohomology. Let G be a group, N a group with left action of G by group automorphisms: $(g,n) \mapsto gn$. Cocycles with coefficients in N are defined as $Z^1(G,N) := \{\, u : G \to N \,|\, u(g_1g_2) = u(g_1)\, g_1u(g_2)\,\}$. Two cocycles are cohomologous, $u' \sim u$, iff for some $n \in N$ and all $g \in G$, we have $u'(g) = n\, u(g)\, (gn)^{-1}$. The cohomology set is $H^1(G,N) := Z^1(G,N)/(\sim)$. It is endowed with a marked point: the class of trivial cocycles $u(g) = n^{-1} \cdot gn$.

We will apply this formalism to iterated Shimura integrals. The role of G will be played by a group $G = P\Gamma \subset PSL(2,\mathbf{Z})$ where $\Gamma \subset SL(2,\mathbf{Z})$.

To define coefficients, choose as above a family of Shimura differentials $\omega_v = f_v(z)z^{m_v-1}dz$, where f_v form a basis of $\oplus_i S(k_i,\Gamma)$, and for a fixed weight, m_v runs over all critical integers for this weight. The forms ω_v span a $P\Gamma$–invariant space. Put $\Omega := \sum_{v\in V} A_v\omega_v$. The role of N will be played by $\Pi :=$ the group of group–like elements of $(1 + \sum_{v\in V} A_v\mathbf{C}\langle\langle A_v\rangle\rangle)^*$. The left action of $P\Gamma$ on Π is the functorality action g_*.

Theorem. *(i) For any $a \in \overline{H}$, the map $P\Gamma \to \Pi : \gamma \mapsto J^a_{\gamma a}(\Omega)$ is a noncommutative 1–cocycle ζ_a in $Z^1(P\Gamma,\Pi)$.*

(ii) The cohomology class of ζ_a in $H^1(P\Gamma,\Pi)$ does not depend on the choice of a and is called the noncommutative modular symbol.

(iii) This cohomology class belongs to the cuspidal subset $H^1(P\Gamma,\Pi)_{cusp}$ consisting of those cohomology classes whose restriction on all stabilizers of Γ–cusps is trivial.

Using the non–commutative Shapiro Lemma, we can reduce the general case to that of $PSL(2,\mathbf{Z})$.

Shapiro Lemma. *Let $G \subset H$ be a subgroup, N a left G–group, $N_H :=$ $\mathrm{Map}_G(N,H)$ with pointwise multiplication and left action of G, $(g_*\phi)(h) := \phi(hg)$. There is a canonical isomorphism of pointed sets:*

$$H^1(G,N) = H^1(H,N_H).$$

In the notation as above, we apply it to the case

$$G := P\Gamma, \quad H := PSL(2,\mathbf{Z}), \quad N := \Pi, \quad \Pi^0 := N_H.$$

It is well known that $H = PSL(2,\mathbf{Z})$ is a free product of two subgroups $\mathbf{Z}_2$ and $\mathbf{Z}_3$ generated respectively by

$$\sigma = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \tau = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}.$$

Theorem. *(i) An iterated Shimura cocycle restricted to (σ,τ) belongs to the set*

$$\{\, (X,Y) \in \Pi^0 \times \Pi^0 \,|\, X \cdot \sigma_*X = 1,\, Y \cdot \tau_*Y \cdot \tau_*^2Y = 1\,\}.$$

(ii) The cohomology relation between cocycles translates as

$$(X,Y) \sim (m^{-1}X\sigma_*(m), m^{-1}Y\tau_*(m)).$$

(iii) The cuspidal part of the cohomology is generated by the pairs

$$\{\, (X,Y) \,|\, \exists Z,\, X \cdot \sigma_*Y = Z^{-1}(\sigma\tau)_*Z\,\}.$$

3.5. Iterated Shimura integrals as multiple Dirichlet series. Start with the family of 1–forms on H:

$$\omega_v(z) = \sum_{n=1}^{\infty} c_{v,n} e^{2\pi i n z} z^{m_v - 1} dz, \quad c_{v,n} \in \mathbf{C}\,, \quad m_v \in \mathbf{Z},\, m_v \geq 1;\; c_{v,n} = O(n^C).$$

Put

$$L(z; \omega_{v_k}, \dots, \omega_{v_1}; j_k, \dots, j_1) :=$$
$$(2\pi i z)^{j_k} \sum_{n_1,\dots,n_k \geq 1} \frac{c_{v_1,n_1} \cdots c_{v_k,n_k} e^{2\pi i (n_1 + \cdots + n_k) z}}{n_1^{m_{v_1} + j_0 - j_1} (n_1 + n_2)^{m_{v_2} + j_1 - j_2} \cdots (n_1 + \cdots + n_k)^{m_{v_k} + j_{k-1} - j_k}}\,.$$

Exponentials ensure absolute convergence for any z with $\operatorname{Im} z > 0$. Formal substitution $z = 0$ may lead to divergence.

Theorem. *For any $k \geq 1$, $(v_1, \dots, v_k) \in V^k$, and $\operatorname{Im} z > 0$ we have*

$$(2\pi i)^{m_{v_1} + \cdots + m_{v_k}} I^z_{i\infty}(\omega_{v_k}, \dots, \omega_{v_1}) =$$
$$(-1)^{\sum_{i=1}^k (m_{v_i} - 1)} \sum_{j_1 = 0}^{m_{v_1} - 1} \sum_{j_2 = 0}^{m_{v_2} - 1 + j_1} \cdots \sum_{j_k = 0}^{m_{v_k} - 1 + j_{k-1}} (-1)^{j_k} \times$$
$$\frac{(m_{v_1} - 1)!(m_{v_2} - 1 + j_1)! \cdots (m_{v_k} - 1 + j_{k-1})!}{j_1! j_2! \cdots j_k!}\, L(z; \omega_{v_k}, \dots, \omega_{v_1}; j_k, \dots, j_1)\,.$$

Proposition. *Assume that ω_V as above is a basis of a space of 1-forms invariant with respect to g_N. Then*

$$J^0_{i\infty}(\omega_V) = (g_{N*}(J^{\frac{i}{\sqrt{N}}}_{i\infty}(\omega_V)))^{-1} J^{\frac{i}{\sqrt{N}}}_{i\infty}(\omega_V)\,. \tag{3.6}$$

Replacing the coefficients of the formal series in the r.h.s of (3.6) by their (convergent) representations via multiple Dirichlet series with exponents we get such representations for $I^0_{i\infty}(\omega_{v_k}, \dots, \omega_{v_1})$ and avoid divergences at $z = 0$.

The multiple Dirichlet series generated by Shimura integrals as above do not form, however, a closed system with respect to multiplication, so that we cannot deduce an analog of shuffle relations with repetitions valid for multiple zeta values. If we complete the family of such series using a combinatorial trick described in [**Man06**], then representation of such series as iterated integrals will involve more general 1–forms than we have been considering up to now. This subject deserves a further study.

3.6. Differentials with logarithmic singularities at the endpoints of integration. We will now assume, as in the initial Drinfeld setting, that the integration limits of the iterated integral are logarithmic singularities of the form Ω. Generally, they diverge and must be regularized. The dependence on the regularization can be described as a version of Deligne's choice of the "base point at infinity".

Let $a =$ a fixed point of the Riemann surface, z a variable point. Put $r_{v,a} := \operatorname{res}_a \omega_v$, $\quad R_a := \operatorname{res}_a \Omega = \sum_v r_{v,a} A_v$. Denote by t_a a local parameter at a, and by $\log t_a$ a local branch of the logarithm real on $t_a \in \mathbf{R}_+$. Finally, put $t_a^{R_a} := e^{R_a \log t_a}$.

Definition. *A local solution to $dJ^z = \Omega(z)J^z$ is called normalized at a (with respect to a choice of t_a) if it is of the form $J = K \cdot t_a^{R_a}$, where K is a holomorphic section in a neighborhood of a and $K(a) = 1$.*

Claim. *(i) The normalized solution exists and is unique.*

(ii) It depends only on the tangent vector $\partial/\partial t_a\,|_a$.

(iii) If $J'_a = K'(t'_a)^{R_a}$ is normalized with respect to t'_a, and $\tau_a := dt'_a/dt_a\,|_a$, then $J'_a = J_a \cdot \tau_a^{R_a}$.

Now, having chosen $(a, t_a), (b, t_b)$, a 1–form $\Omega = \sum A_v\omega_v$ with at most logarithmic singularities at a, b, and a (homotopy class of) path(s) from a to b avoiding other singularities of Ω, we construct the normalized solutions J_a, J_b analytically continued along γ and *the scattering operator*

$$\widetilde{J}_b^a = J_a^{-1}J_b \in \mathbf{C}\langle\langle A_V\rangle\rangle\,.$$

Its coefficients (as power series in (A_v)), by definition, are *regularized iterated integrals* of (ω_v). It turns out that $\widetilde{J}_b^a$ satisfy the general properties of the iterated integrals summarized in 3.2.

Example: Drinfeld's associator. Let $X = \mathbf{P}^1(\mathbf{C})$, $V = \{0, 1\}$,

$$\omega_0 = \frac{1}{2\pi i}\,\frac{dz}{z}, \quad \omega_1 = \frac{1}{2\pi i}\,\frac{dz}{z-1}\,.$$

Then

$$\Omega = A_0\omega_0 + A_1\omega_1$$

has poles at $0, 1, \infty$ with residues $A_0/2\pi i$, $A_1/2\pi i$, $-(A_0+A_1)/2\pi i$ respectively. Put $t_0 = z, t_1 = 1 - z$. Then $\widetilde{J}_0^1$ in our notation is the Drinfeld associator $\phi_{KZ}(A_0, A_1)$.

Example: modular generalization of multiple zeta values. Let Γ be a congruence subgroup of the modular group, (f_v) := a basis of Eisenstein series of weight 2 wrt Γ, $\{\omega_v$ = push forward of $f_v(z)dz\}$: 1–forms with logarithmic singularities at cusps on X_Γ. The space of such forms has the maximal possible dimension, because the difference of any two cusps has finite order in the Jacobian (cf. [**Elk90**]).

Regularized iterated integrals of Eisenstein series of weight two along geodesics between cusps provide a modular generalization of multiple zeta values.

References

[AB90] A. Ash and A. Borel, *Generalized modular symbols*, Cohomology of arithmetic groups and automorphic forms (Luminy-Marseille, 1989), Lecture Notes in Math., vol. 1447, Springer, Berlin, 1990, pp. 57–75. MR 1082962 (92e:11058)

[AR79] A. Ash and L. Rudolph, *The modular symbol and continued fractions in higher dimensions*, Invent. Math. **55** (1979), no. 3, 241–250. MR 553998 (82g:12011)

[Che77] K.-T. Chen, *Iterated path integrals*, Bull. Amer. Math. Soc. **83** (1977), no. 5, 831–879. MR 0454968 (56 #13210)

[CM04] A. Connes and M. Marcolli, *Renormalization and motivic Galois theory*, Int. Math. Res. Not. (2004), no. 76, 4073–4091. MR 2109986 (2006b:81173)

[Del01] P. Deligne, *Multizeta values*, 2001, Notes d'exposés, IAS, Princeton.

[DG05] P. Deligne and A. Goncharov, *Groupes fondamentaux motiviques de Tate mixte*, Ann. Sci. École Norm. Sup. (4) **38** (2005), no. 1, 1–56, math.NT/0302267. MR 2136480 (2006b:11066)

[Dri73] V. G. Drinfeld, *Two theorems on modular curves*, Funkcional. Anal. i Priložen. **7** (1973), no. 2, 83–84, English translation in: Functional Anal. Appl. 7 (1973), 155–156. MR 0318157 (47 #6705)

[Dri90] ———, *On quasitriangular quasi-Hopf algebras and on a group that is closely connected with* $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$, Algebra i Analiz **2** (1990), no. 4, 149–181, English translation in: Leningrad Math. J. 2 (1991), no. 4, 829–860. MR 1080203 (92f:16047)

[Elk90] R. Elkik, *Le théorème de Manin-Drinfel′d*, Astérisque (1990), no. 183, 59–67, Séminaire sur les Pinceaux de Courbes Elliptiques (Paris, 1988). MR 1065155 (92e:14018)

[GM04] A. B. Goncharov and Yu. I. Manin, *Multiple ζ-motives and moduli spaces $\overline{\mathcal{M}}_{0,n}$*, Compos. Math. **140** (2004), no. 1, 1–14, math.AG/0204102. MR 2004120 (2005c:11090)

[Gona] A. B. Goncharov, *Multiple polylogarithms and mixed Tate motives*, math.AG/0103059.

[Gonb] ———, *Periods and mixed motives*, math.AG/0202154.

[Gon95] ———, *Polylogarithms in arithmetic and geometry*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Zürich, 1994) (Basel), Birkhäuser, 1995, pp. 374–387. MR 1403938 (97h:19010)

[Gon97] ———, *The double logarithm and Manin's complex for modular curves*, Math. Res. Lett. **4** (1997), no. 5, 617–636. MR 1484694 (99e:11086)

[Gon98] ———, *Multiple polylogarithms, cyclotomy and modular complexes*, Math. Res. Lett. **5** (1998), no. 4, 497–516. MR 1653320 (2000c:11108)

[Gon01] ———, *Multiple ζ-values, Galois groups, and geometry of modular varieties*, European Congress of Mathematics, Vol. I (Barcelona, 2000), Progr. Math., vol. 201, Birkhäuser, Basel, 2001, math.AG/0005069, pp. 361–392. MR 1905330 (2003g:11073)

[Hai02] R. Hain, *Iterated integrals and algebraic cycles: examples and prospects*, Contemporary trends in algebraic geometry and algebraic topology (Tianjin, 2000), Nankai Tracts Math., vol. 5, World Sci. Publ., River Edge, NJ, 2002, math.AG/0109204, pp. 55–118. MR 1945356 (2004a:14009)

[Her01] A. Herremans, *A combinatorial interpretation of Serre's conjecture on modular Galois representations*, Ph.D. thesis, Katholieke Universiteit Leuven, 2001.

[Her03] ———, *A combinatorial interpretation of Serre's conjecture on modular Galois representations*, Ann. Inst. Fourier (Grenoble) **53** (2003), no. 5, 1287–1321. MR 2032935 (2004i:11053)

[Kon99] M. Kontsevich, *Operads and motives in deformation quantization*, Lett. Math. Phys. **48** (1999), no. 1, 35–72. MR 1718044 (2000j:53119)

[KZ01] M. Kontsevich and D. Zagier, *Periods*, Mathematics unlimited—2001 and beyond, Springer, Berlin, 2001, pp. 771–808. MR 1852188 (2002i:11002)

[Man72] Yu. I. Manin, *Parabolic points and zeta functions of modular curves*, Izv. Akad. Nauk SSSR Ser. Mat. **36** (1972), 19–66, English translation in: Math. USSR Izvestija, publ. by AMS, vol. 6, No. 1 (1972), 19–64, and Selected papers, World Scientific, 1996, 202–247. MR 0314846 (47 #3396)

[Man73] ———, *Periods of cusp forms, and p-adic Hecke series*, Mat. Sb. (N.S.) **92(134)** (1973), 378–401, 503, English translation in: Math. USSR Sbornik, 21:3 (1973), 371–393 and Selected papers, World Scientific, 1996, 268–290. MR 0345909 (49 #10638)

[Man05] ———, *Iterated Shimura integrals*, Mosc. Math. J. **5** (2005), no. 4, 869–881, 973, math.AG/0507438. MR 2266463 (2007m:11069)

[Man06] ———, *Iterated integrals of modular forms and noncommutative modular symbols*, Algebraic geometry and number theory, Progr. Math., vol. 253, Birkhäuser Boston, Boston, MA, 2006, math.NT/0502576, pp. 565–597. MR 2263200 (2008a:11062)

[Mar05] M. Marcolli, *Arithmetic noncommutative geometry*, University Lecture Series, vol. 36, American Mathematical Society, Providence, RI, 2005, With a foreword by Yuri Manin. MR 2159918 (2006g:58018)

[Mer93] L. Merel, *Quelques aspects arithmétiques et géométriques de la théorie des symboles modulaires*, Ph.D. thesis, Université Paris VI, 1993.

[Mer94] ———, *Universal Fourier expansions of modular forms*, On Artin's conjecture for odd 2-dimensional representations, Lecture Notes in Math., vol. 1585, Springer, Berlin, 1994, pp. 59–94. MR 1322319 (96h:11032)

[MM02] Yu. I. Manin and M. Marcolli, *Continued fractions, modular symbols, and noncommutative geometry*, Selecta Math. (N.S.) **8** (2002), no. 3, 475–521, math.NT/0102006. MR 1931172 (2004a:11039)

[Rac00] G. Racinet, *Séries génératrices non–commutatives de polyzêtas et associateurs de Drinfeld*, Ph.D. thesis, Université de Picardie-Jules-Verne, 2000.

[Rac02] ———, *Doubles mélanges des polylogarithmes multiples aux racines de l'unité*, Publ. Math. Inst. Hautes Études Sci. (2002), no. 95, 185–231, math.QA/0202142. MR 1953193 (2004c:11117)

[Rac04] ———, *Summary of algebraic relations between multiple zeta values*, 2004, Notes of the talk at MPIM, Bonn, Aug. 17.

[Sho76] V. V. Shokurov, *Modular symbols of arbitrary weight*, Funkcional. Anal. i Priložen. **10** (1976), no. 1, 95–96, English translation in: Functional Anal. Appl. 10 (1976), no. 1, 85–86. MR 0427234 (55 #269)

[Sho80a] ———, *Shimura integrals of cusp forms*, Izv. Akad. Nauk SSSR Ser. Mat. **44** (1980), no. 3, 670–718, 720, English translation in: Math. USSR Izvestiya, 16:3 (1981), 603–646. MR 582162 (82b:10029)

[Sho80b] ———, *A study of the homology of Kuga varieties*, Izv. Akad. Nauk SSSR Ser. Mat. **44** (1980), no. 2, 443–464, 480, English translation in: Math. USSR-Izv. 16 (1981), no. 2, 399–418. MR 571104 (82f:14023)

[Ter02] T. Terasoma, *Mixed Tate motives and multiple zeta values*, Invent. Math. **149** (2002), no. 2, 339–369. MR 1918675 (2003h:11073)

[Zag90] D. Zagier, *Hecke operators and periods of modular forms*, Festschrift in honor of I. I. Piatetski-Shapiro on the occasion of his sixtieth birthday, Part II (Ramat Aviv, 1989), Israel Math. Conf. Proc., vol. 3, Weizmann, Jerusalem, 1990, pp. 321–336. MR 1159123 (93d:11053)

[Zag91] ———, *Periods of modular forms and Jacobi theta functions*, Invent. Math. **104** (1991), no. 3, 449–465. MR 1106744 (92e:11052)

[Zag94] ———, *Values of zeta functions and their applications*, First European Congress of Mathematics, Vol. II (Paris, 1992), Progr. Math., vol. 120, Birkhäuser, Basel, 1994, pp. 497–512. MR 1341859 (96k:11110)

Max–Planck–Institut für Mathematik, Vivatsgasse 7 53111 Bonn, Germany and Department of Mathematics, Northwestern University 2033 Sheridan Road Evanston, IL 60208-2730, USA

E-mail address: manin@math.northwestern.edu

Surfaces

Clay Mathematics Proceedings
Volume 8, 2009

Rational surfaces over nonclosed fields

Brendan Hassett

Abstract. This paper is based on lectures given at the Clay Summer School on Arithmetic Geometry in July 2006.

These notes offer an introduction to the birational geometry of algebraic surfaces, emphasizing the aspects useful for arithmetic. The first three sections are explicitly devoted to birational questions, with a special focus on rational surfaces. We explain the special rôle these play in the larger classification theory. The geometry of rational ruled surfaces and Del Pezzo surfaces is studied in substantial detail. We extend this theory to geometrically rational surfaces over non-closed fields, enumerating the minimal surfaces and describing their geometric properties. This gives essentially the complete classification of rational surfaces up to birational equivalence.

The final two sections focus on singular Del Pezzo surfaces, universal torsors, and their algebraic realizations through Cox rings. Current techniques for counting rational points (on rational surfaces over number fields) often work better for singular surfaces than for smooth surfaces. The actual enumeration of the rational points often boils down to counting integral points on the universal torsor. Universal torsors were first employed in the (ongoing) search for effective criteria for when rational surfaces over number fields admit rational points.

It might seem that these last two topics are far removed from birational geometry, at least the classical formulation for surfaces. However, singularities and finite-generation questions play a central rôle in the minimal model program. And the challenges arising from working over non-closed fields help highlight structural characteristics of this program that usually are only apparent over $\mathbb{C}$ in higher dimensions. Indeed, these notes may be regarded as an arithmetically motivated introduction to modern birational geometry.

In general, the prerequisites for these notes are a good understanding of algebraic geometry at the level of Hartshorne [**Har77**]. Some general understanding of descent is needed to appreciate the applications to non-closed fields. Readers

2000 *Mathematics Subject Classification.* Primary 14J26; Secondary 14G05, 14J20.

Key words and phrases. rational surfaces, Del Pezzo surfaces, minimal models.

Supported in part by NSF Grants #0134259 and #0554491, the Clay Mathematics Institute, and the University of Göttingen.

interested in applications to positive characteristic would benefit from some exposure to étale cohomology at the level of Milne [**Mil80**]. There is one place where we do not fully observe these prerequisites: The discussion of the Cone Theorem is not self-contained although we do sketch the main ideas. Thankfully, a number of books ([**CKM88**], [**KM98**], [**Rei97**], [**Kol96**],[**Mat02**]) give good introductory accounts of this important topic.

Finally, we should indicate how this account relates to others in the literature. The general approach taken to the geometry of surfaces over algebraically closed fields owes much to Beauville's book [**Bea96**] and Reid's lecture notes [**Rei97**]. The extensions to non-closed fields draw from Kollár's book [**Kol96**]. Readers interested in details of the Galois action on the lines of a Del Pezzo surface and its implications for arithmetic should consult Manin's classic book [**Man74**] and the more recent survey [**MT86**]. The books [**CKM88, KM98, Mat02**] offer a good introduction to modern birational geometry; [**Laz04**] has a comprehensive account of linear series. We have made no effort to explain how universal torsors and Cox rings are used for the descent of rational points; the recent book of Skorobogatov [**Sko01**] does a fine job covering this material.

I am grateful to Anthony Várilly-Alvarado, Michael Joyce, Ambrus Pál, and other members of the summer school for helpful comments.

1. Rational surfaces over algebraically closed fields

Let k be an algebraically closed field. Throughout, a *variety* will designate an integral separated scheme of finite type over k.

1.1. Classical example: Cubic surfaces. Here a *cubic surface* means a smooth cubic hypersurface $X \subset \mathbb{P}^3$. We recall a well-known construction for such surfaces:

Let $p_1, \dots, p_6 \in \mathbb{P}^2$ be points in the projective plane in *general position*, i.e.,

- the points are distinct;
- no three of the points are collinear;
- the six points do not lie on a plane conic.

Consider the vector space of homogeneous cubics vanishing at these points; it is an exercise to show this has dimension four

$$I_{p_1,\dots,p_6}(3) = \langle F_0, F_1, F_2, F_3 \rangle$$

and has no additional basepoints.

The resulting linear series gives a rational map

$$\begin{array}{rcl} \rho : \mathbb{P}^2 & \dashrightarrow & \mathbb{P}^3 \\ [x_0, x_1, x_2] & \mapsto & [F_0, F_1, F_2, F_3] \end{array}$$

that is not well-defined at $p_1, \dots, p_6$. Consider the blow-up

$$\beta : X := \mathrm{Bl}_{p_1,\dots,p_6}\mathbb{P}^2 \to \mathbb{P}^2$$

with exceptional divisors

$$E_1, \dots, E_6.$$

Blowing up the base scheme of a linear series resolves its indeterminacy, so we obtain a morphism

$$j : X \to \mathbb{P}^3$$

with $j = \rho \circ \beta$.

PROPOSITION 1.1. *The morphism j gives a closed embedding of X in $\mathbb{P}^3$.*

We leave the proof as an exercise.

Given this, we may describe the image of j quite easily. The first step is to analyze the *Picard group* $\mathrm{Pic}(X)$ and its associated intersection form

$$\begin{array}{ccc} \mathrm{Pic}(X) \times \mathrm{Pic}(X) & \to & \mathbb{Z} \\ (D_1, D_2) & \mapsto & D_1 \cdot D_2 \end{array}.$$

We recall what happens to the intersection form under blow-ups. Let $\beta : Y \to P$ be the blow-up of a smooth surface at a point, with exceptional divisor E. Then we have an orthogonal direct-sum decomposition

$$\mathrm{Pic}(Y) = \mathrm{Pic}(P) \oplus_\perp \mathbb{Z}E, \quad E \cdot E = E^2 = -1,$$

where the inclusion $\mathrm{Pic}(P) \hookrightarrow \mathrm{Pic}(Y)$ is induced by β^*.

Returning to our particular situation, we have

$$\mathrm{Pic}(X) = \mathbb{Z}L \oplus \mathbb{Z}E_1 \oplus \cdots \oplus \mathbb{Z}E_6.$$

Here L is the pullback of the hyperplane from $\mathbb{P}^2$ with $L^2 = L \cdot L = 1$ and $L \cdot E_a = 0$ for each a. We also have $E_a \cdot E_b = 0$ for $a \neq b$.

Since j is induced by the linear series of cubics with simple basepoints at $p_1, \ldots, p_6$, we have

$$j^* \mathcal{O}_{\mathbb{P}^3}(1) = \mathcal{O}_X(3L - E_1 - \cdots - E_6)$$

so that

$$\deg(j(X)) = (3L - E_1 - \cdots - E_6)^2 = 9 - 6 = 3.$$

This proves that the image is a smooth cubic surface. The images of the exceptional divisors $E_1, \ldots, E_6$ have degree

$$E_i \cdot (3L - E_1 - \cdots - E_6) = 1$$

and thus are lines on our cubic surface.

PROPOSITION 1.2. *The cubic surface $j(X) \subset \mathbb{P}^3$ contains the following 27 lines:*

- *the exceptional curves E_a;*
- *proper transforms of lines through p_a and p_b, with class $L - E_a - E_b$;*
- *proper transforms of conics through five basepoints p_a, p_b, p_c, p_d, p_e, with class $2L - E_a - E_b - E_c - E_d - E_e$.*

This beautiful analysis leaves open a number of classification questions:

(1) Does every cubic surface arise as the blow-up of $\mathbb{P}^2$ in six points in general linear position?
(2) Are there exactly 27 lines on a cubic surface?

To address these, we introduce some general geometric definitions:

DEFINITION 1.3. Let Y be a smooth projective surface with canonical class K_Y, i.e., the divisor class associated with the differential two-forms $\Omega^2_Y = \bigwedge^2 \Omega^1_Y$. We say that Y is a *Del Pezzo surface* if $-K_Y$ is ample, i.e., there exists an embedding $Y \subset \mathbb{P}^N$ such that $\mathcal{O}_{\mathbb{P}^N}(1)|Y = \mathcal{O}_Y(-rK_Y)$ for some $r > 0$.

Note that if Y is Del Pezzo then $K_Y^2 > 0$.

REMARK 1.4. Let $X \subset \mathbb{P}^3$ be a cubic surface and H the hyperplane class on $\mathbb{P}^3$. Adjunction

$$K_X = (K_{\mathbb{P}^3} + X)|_X = (-4H + 3H)|_X = -H|_X$$

implies that any cubic surface is a Del Pezzo surface.

DEFINITION 1.5. Let Y be a smooth projective surface. A (-1)-curve is a smooth rational curve $E \subset Y$ with $E^2 = -1$.

Of course, exceptional divisors are the main examples. We also have the following characterization.

PROPOSITION 1.6. *Let Y be a smooth projective surface. Let $E \subset Y$ be an irreducible curve with*

$$E^2 < 0, \quad K_Y \cdot E < 0.$$

Then E is a (-1)-curve. In particular, on a Del Pezzo surface every irreducible curve with $E^2 < 0$ is a (-1)-curve.

PROOF. Let $p_a(E)$ denote the arithmetic genus of E. Since E is an irreducible curve we know that $p_a(E) \geq 0$ with equality if and only if $E \simeq \mathbb{P}^1$. Combining this with the adjunction formula, we obtain

$$-2 \leq 2p_a(E) - 2 = E \cdot (K_Y + E).$$

Thus $E^2 = -1$, $K_Y \cdot E = -1$, and E is a smooth rational curve. □

REMARK 1.7. The lines on a cubic surface are precisely its (-1)-curves. Indeed, if $\ell \subset X$ is a line then the genus formula gives

$$-2 = 2g(\ell) - 2 = \ell^2 + K_X \cdot \ell = \ell^2 - 1.$$

Suppose then that

$$\ell = aL - b_1E_1 - \cdots - b_6E_6$$

is a line on a cubic surface. Then the following equations must be satisfied

$$\begin{aligned} 1 &= -K_X \cdot \ell = 3a - b_1 - b_2 - b_3 - b_4 - b_5 - b_6 \\ -1 &= \ell^2 = a^2 - b_1^2 - b_2^2 - b_3^2 - b_4^2 - b_5^2 - b_6^2 \end{aligned}$$

and these can be solved explicitly. There are precisely 27 solutions; see Exercise 1.1.6 and [**Man74**, 26.2], especially for the connection with root systems. Thus the cubic surfaces arising as blow-ups of $\mathbb{P}^2$ in six points in general position have precisely 27 lines.

We extend this analysis to all smooth cubic surfaces:

THEOREM 1.8. *Let $\mathbb{P}^{19} = \mathbb{P}(\mathrm{Sym}^3(k^4))$ parametrize all cubic surfaces and let*

$$Z = \{(X, \ell) : X \text{ cubic surface, } \ell \subset X \text{ line } \} \subset \mathbb{P}^{19} \times \mathbb{G}(1, 3)$$

denote the incidence correspondence. Let $U \subset \mathbb{P}^{19}$ denote the locus of smooth cubic surfaces and

$$\pi_1 : Z_U := Z \times_{\mathbb{P}^{19}} U \to U$$

the projection. Then π_1 is a finite étale morphism.

Since U is connected the degree of π_1 is constant, and we conclude

COROLLARY 1.9. *Every smooth cubic surface has 27 lines.*

PROOF. (cf. [**Mum95**, p. 173 ff.]) We claim that Z is proper and irreducible of dimension 19: Each line ℓ is contained in a $20-4=16$-dimensional linear series of cubic surfaces, so the projection $Z \to \mathbb{G}(1,3)$ is a $\mathbb{P}^{15}$ bundle. Consequently, π_1 is a proper morphism. In particular, for each one-parameter family of lines in cubic surfaces (X_t, ℓ_t), the flat limit

$$\lim_{t\to 0}(X_t, \ell_t)$$

is also a line in a cubic surface.

Let $\mathcal{N}_{\ell/X}$ denote the normal bundle of a line ℓ in a smooth cubic surface X. We have $\mathcal{N}_{\ell/X} \simeq \mathcal{O}_{\mathbb{P}^1}(-1)$ so that

$$h^0(\mathcal{N}_{\ell/X}) = h^1(\mathcal{N}_{\ell/X}) = 0.$$

Recall that $H^0(\mathcal{N}_{\ell/X})$ (resp. $H^1(\mathcal{N}_{\ell/X})$) is the tangent space (resp. obstruction space) of the scheme of lines on X at ℓ. It follows then that Z_U is smooth of relative dimension zero over U, i.e., π_1 is étale. Furthermore, proper étale morphisms are finite. □

One further piece of information can be extracted from this result: The intersections of the 27 lines are constant over all the cubic surfaces. This means that every cubic surface X contains a pair (and even a sextuple!) of pairwise disjoint lines (cf. Exercise 1.1.3).

PROPOSITION 1.10. *Let X be a smooth cubic surface containing disjoint lines E_1 and E_2. Let $\ell_1, \ldots, \ell_5$ denote the lines in X meeting E_1 and E_2. There is a birational morphism*

$$\begin{array}{rcl} \varphi : X & \to & \mathbb{P}^1 \times \mathbb{P}^1 \\ x & \mapsto & (p_{E_1}(x), p_{E_2}(x)) \end{array}$$

where $p_{E_i} : \mathbb{P}^3 \dashrightarrow \mathbb{P}^1$ is projection from E_i. This contracts $\ell_1, \ldots, \ell_5$ to distinct points $q_1, \ldots, q_5 \in \mathbb{P}^1 \times \mathbb{P}^1$ satisfying the following genericity conditions:

- *no pair of them lie on a ruling of $\mathbb{P}^1 \times \mathbb{P}^1$;*
- *no four of them lie on a curve of bidegree $(1,1)$;*

The inverse φ^{-1} is given by the linear system of forms of bidegree $(2,2)$ through $q_1, \ldots, q_5$.

We leave this as an exercise.

COROLLARY 1.11. *Every smooth cubic surface is isomorphic to $\mathbb{P}^2$ blown up at six points.*

PROOF. We first verify that $\mathrm{Bl}_{q_1,\ldots,q_5}\mathbb{P}^1 \times \mathbb{P}^1$ is isomorphic to $\mathbb{P}^2$ blown up at six points. Indeed, we can realize $\mathbb{P}^1 \times \mathbb{P}^1$ as a smooth quadric $Q \subset \mathbb{P}^3$, so that the fibers of each projection are lines on Q. Let $q \in Q$ be any point and R_1 and R_2 the two rulings passing though q. Projection from q

$$p_q : Q \dashrightarrow \mathbb{P}^2$$

lifts to a morphism

$$\mathrm{Bl}_q Q \to \mathbb{P}^2$$

contracting the proper transforms of R_1 and R_2. □

Before concluding, we draw two morals from this story:

- (-1)-curves govern much of the geometry of a Del Pezzo surface;
- classifying (-1)-curves is a crucial step in classifying the surfaces.

Exercises.

EXERCISE 1.1.1. Show that six distinct points on the plane impose independent conditions on cubics if no four of the points are collinear. Show that the resulting linear system has base scheme equal to these six points.

EXERCISE 1.1.2. Give a careful proof of Proposition 1.1.

EXERCISE 1.1.3. Verify that the 27 curves described in Proposition 1.2 are in fact lines on the cubic surface. Check that each of these has self-intersection -1. Show that

(1) each line is intersected by ten other lines;
(2) any pair of disjoint lines is intersected by five lines;
(3) each line is contained in a collection of six pairwise disjoint lines.

EXERCISE 1.1.4. Prove Proposition 1.10.

EXERCISE 1.1.5. Let X be a smooth cubic surface. Show that the intersection form on $K_X^{\perp} \subset \mathrm{Pic}(X)$ is isomorphic to

	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6
ρ_1	-2	1	0	0	0	0
ρ_2	1	-2	1	0	0	0
ρ_3	0	1	-2	1	0	1
ρ_4	0	0	1	-2	1	0
ρ_5	0	0	0	1	-2	0
ρ_6	0	0	1	0	0	-2

.

Up to sign, this is the Cartan matrix associated to the root system $\mathbf{E}_6$.

EXERCISE 1.1.6. Consider a line on a cubic surface $\ell \subset X$, and the associated class $\lambda = 3\ell + K_X \in K_X^{\perp}$. Verify that $\lambda^2 = -12$ and $\lambda \cdot \eta \equiv 0 \pmod 3$ for each $\eta \in K_X^{\perp} \subset \mathrm{Pic}(X)$. Deduce that there are a finite number of lines on a cubic surface.

1.2. The structure of birational morphisms of surfaces. Our first task is to show that all (-1)-curves arise as exceptional curves of blow-ups:

THEOREM 1.12 (Castelnuovo contraction criterion). [**Har77**, V.5.7] *Let X be a smooth projective surface and $E \subset X$ a (-1)-curve. Then there exists a smooth projective surface Y and a morphism $\beta : X \to Y$ contracting E to a point $y \in Y$, so that X is isomorphic to $\mathrm{Bl}_y Y$. Each morphism $\psi : X \to Z$ contracting E admits a factorization*

$$\psi : X \xrightarrow{\beta} Y \to Z.$$

PROOF. (Sketch) Let H be a very ample divisor on X such that

$$H^1(X, \mathcal{O}_X(H)) = 0.$$

Set $\mathcal{L} = \mathcal{O}_X(H + (H \cdot E)E)$ so that $\mathcal{L}|E \simeq \mathcal{O}_E$. For each $n > 0$ we have the inclusion

$$\mathcal{O}_X(nH) \hookrightarrow \mathcal{L}^n = \mathcal{O}_X(nH + n(H \cdot E)E)$$

which is an isomorphism away from E. Thus the sections in the image of

$$\Gamma(X, \mathcal{O}_X(nH)) \hookrightarrow \Gamma(X, \mathcal{L}^n)$$

induce an embedding of $X \setminus E$.

We claim that $\mathcal{L}$ is globally generated, so we have a morphism

$$\beta : X \to Y := \mathrm{Proj}(\bigoplus_{n\geq 0} \Gamma(X, \mathcal{L}^n)).$$

Since $\mathcal{L}|E$ is trivial, β necessarily contracts E to a point; β is an isomorphism away from E.

Here is the idea: Since $\mathcal{L}|E$ is globally generated, it suffices to show that the restriction

$$\Gamma(X, \mathcal{L}) \to \Gamma(E, \mathcal{L}|E) \simeq \Gamma(\mathbb{P}^1, \mathcal{O}_{\mathbb{P}^1})$$

is surjective. Taking the long exact sequence associated to

$$0 \to \mathcal{L}(-E) \to \mathcal{L} \to \mathcal{L}|E \to 0,$$

we are reduced to showing that $H^1(X, \mathcal{L}(-E)) = 0$. Indeed, we can show inductively that $H^1(X, \mathcal{O}_X(H + aE)) = 0$ for $a = 1, \ldots, H \cdot E - 1$: The exact sequence

$$0 \to \mathcal{O}_X(H + (a-1)E) \to \mathcal{O}_X(H + aE) \to \mathcal{O}_E(H \cdot E - a) \to 0$$

expresses $\mathcal{O}_X(H + aE)$ as an extension of sheaves with vanishing H^1.

The trickiest bit is to check that Y is smooth and β is the blow-up of a point of Y. The necessary local computation can be found in [**Bea96**, II.17] or [**Har77**, pp. 415].

For the factorization step, the standard isomorphism

$$\mathrm{Pic}(X) = \beta^*\mathrm{Pic}(Y) \oplus \mathbb{Z}E$$

identifies $\beta^*\mathrm{Pic}(Y)$ with line bundles on X restricting to zero along E. Moreover, the induced map

$$\Gamma(Y, \mathcal{M}') \to \Gamma(X, \beta^*\mathcal{M}')$$

is an isomorphism. Suppose that $\mathcal{M}$ is very ample on Z so that ψ is induced by certain sections of $\psi^*\mathcal{M}$. However, $\mathcal{M} = \beta^*\mathcal{M}'$ for some $\mathcal{M}'$ on Y and the relevant sections of $\psi^*\mathcal{M}$ come from sections of $\mathcal{M}'$. □

THEOREM 1.13. *Let $\phi : X \to Y$ be a birational morphism of smooth projective surfaces. Then there exists a factorization*

$$X = X_0 \overset{\beta_1}{\to} X_1 \to \cdots \to X_{r-1} \overset{\beta_r}{\to} X_r = Y$$

where each β_j is a blow-up of a point on X_j. (If ϕ is an isomorphism we take $X_0 = X_r$.)

PROOF. We assume ϕ is not an isomorphism. Hence it is ramified and the induced map

$$\phi^*\Omega^2_Y \to \Omega^2_X$$

is not an isomorphism. Since these sheaves are invertible, we can therefore write

$$\phi^*\Omega^2_Y = \Omega^2_X(-(m_1E_1 + \cdots + m_rE_r)),$$

where the E_i are irreducible *ϕ-exceptional curves*, i.e., ϕ contracts E_i to a point in Y. Since $\phi^*K_Y|_{E_i}$ is trivial we have $\phi^*K_Y \cdot E_i = 0$. The multiplicity m_i is positive because ϕ is ramified along E_i. In divisorial notation, we obtain the *discrepancy formula*:

$$K_X = \phi^*K_Y + \sum m_iE_i, \quad m_i > 0. \tag{1.1}$$

By the Hodge index theorem [**Har77**, V.1.9], the intersection form on

$$\Lambda = \mathbb{Z}E_1 + \cdots + \mathbb{Z}E_r$$

is negative definite, so in particular each $E_i^2 < 0$. We claim that $K_X \cdot E_i < 0$ for some i; then Proposition 1.6 guarantees that E_i is a (-1)-curve.

We know that

$$(\sum m_i E_i)^2 = \sum_i m_i E_j \cdot (\sum_j m_j E_j) = \sum_i m_i E_i \cdot K_X$$

is negative, because the intersection form on Λ is negative definite. Hence some $E_i \cdot K_X$ must be negative.

Using the Castelnuovo criterion we contract E_i

$$X = X_0 \to X_1,$$

so that X_0 is the blow-up of X_1 at a point. Moreover, ϕ factors through X_1. This factorization process terminates because the exceptional locus of ϕ has a finite number of irreducible components. □

DEFINITION 1.14. A smooth projective surface X is *minimal* if every birational morphism $\phi : X \to Y$ to a smooth variety is an isomorphism.

Theorem 1.13 says that X is minimal if and only if it has no (-1)-curves.

Exercises.

EXERCISE 1.2.1. Let X be the blow-up of $\mathbb{P}^2$ at $[0,0,1], [0,1,0], [1,0,0]$. Realize X in $\mathbb{P}^2 \times \mathbb{P}^2 \subset \mathbb{P}^8$ using the bihomogeneous equations

$$x_0 y_0 = x_1 y_1 = x_2 y_2.$$

Verify that the proper transforms of the lines $x_0 = 0, x_1 = 0, x_2 = 0$ are (-1)-curves and write down explicit linear series contracting each one individually.

EXERCISE 1.2.2. Let X be a cubic surface, realized as $\mathbb{P}^2$ blown up at six points. Describe a basepoint-free linear series on X contracting the six curves

$$2L - E_a - E_b - E_c - E_d - E_e.$$

What is the image of the corresponding morphism $X \to Y$?

1.3. Relative minimality and ruled surfaces. Let $f : X \to B$ denote a dominant morphism from a smooth projective surface to a variety. We say that X is *minimal relative to* f if there exists no commutative diagram

$$\begin{array}{ccc} X & \xrightarrow{\phi} & Y \\ & \searrow \quad \swarrow & \\ & B & \end{array}$$

where ϕ is birational and Y is smooth. In analogy to Theorem 1.13, X is minimal relative to f if and only if there are no (-1)-curves in the fibers of f.

A *ruled surface* is a morphism $f : X \to B$ from a smooth projective surface to a smooth curve whose generic fiber is rational; it is *minimal* if it is minimal relative to f. If f is smooth then each fiber is isomorphic to $\mathbb{P}^1$; in this case, $f : X \to B$ is called a $\mathbb{P}^1$*-bundle*.

PROPOSITION 1.15. *Let X be a smooth projective surface and $f : X \to B$ a $\mathbb{P}^1$-bundle. Then each $b \in B$ admits an étale-open neighborhood $U \to B$ and an isomorphism:*

$$\begin{array}{ccc} X \times_B U & \stackrel{\sim}{\to} & \mathbb{P}^1 \times U \\ & \searrow \quad \swarrow & \\ & U & \end{array}$$

PROOF. Since f is smooth it admits a multisection $M \subset X$ with $f|M$ unbranched over b; let $U \subset M$ denote the open set where f is unramified. The pullback $g : X' := X \times_B M \to M$ admits the canonical diagonal section Σ. Consider the direct images of $\mathcal{O}_{X'}(\Sigma)$. Cohomology and base change implies that $\mathbb{R}^1 g_* \mathcal{O}_{X'}(\Sigma)$ is trivial and $\mathcal{E} := g_* \mathcal{O}_{X'}(\Sigma)$ has rank two. Under these conditions cohomology commutes with base change, so a fiber-by-fiber analysis shows that

$$g^* g_* \mathcal{O}_{X'}(\Sigma) \to \mathcal{O}_{X'}(\Sigma)$$

is surjective and the induced morphism

$$X' \to \mathbb{P}(\mathcal{E})$$

is an isomorphism over M. □

THEOREM 1.16. *Let $f : X \to B$ be a minimal ruled surface. Then X is a $\mathbb{P}^1$-bundle over B.*

Before proving this, we'll require a preliminary result.

LEMMA 1.17. Let F denote the class of a fiber of f. Consider a fiber of f with irreducible components $E_1, \ldots, E_r$. Then we have $E_i^2 < 0$ and $F \cdot E_i = 0$ for each i and $K_X \cdot E_i < 0$ for some i. In particular, each reducible fiber contains a (-1)-curve.

PROOF. Each fiber of f is numerically equivalent to F, i.e., has the same intersection numbers with curves in X. Since these fibers are generally disjoint from the E_i, we have $F \cdot E_i = 0$ for each i and $F \cdot F = 0$.

Express $F = \sum_{i=1}^r m_i E_i$ where $m_i > 0$ is the multiplicity of the fiber along E_i. Note that F is connected, e.g., by Stein factorization. Thus each E_i meets some E_j and

$$E_i \cdot \sum_{j \neq i} m_j E_j = E_i \cdot (F - m_i E_i) > 0.$$

It follows that $E_i \cdot E_i < 0$.

Finally, $K_X \cdot F = -2$ by adjunction, so $K_X \cdot E_i < 0$ for some index. □

PROOF. (Theorem 1.16)

The key point is to show that the fibers of f are all isomorphic to $\mathbb{P}^1$. Since f is a dominant morphism from a nonsingular surface to a nonsingular curve, it is flat with fibers of arithmetic genus zero. Each fiber is a Cartier divisor on X and thus has no embedded points. Under the assumptions of Theorem 1.16, each fiber of f is irreducible. We also have that each fiber has multiplicity one. Indeed, writing $F = mE$ we have

$$-2 = K_X \cdot F = m K_X \cdot E$$

so $m = 1, 2$. However, if $m = 2$ then adjunction yields $2g(E) - 2 = E^2 + K_X E = -1$, which is absurd. Thus each fiber of f is isomorphic to $\mathbb{P}^1$, and in particular f is smooth. □

We record one additional fact for future reference, whose proof is left as an exercise:

PROPOSITION 1.18. *Let $E_1, \ldots, E_r$ be the components of a fiber of a ruled surface; let F denote the class of the fiber. The induced intersection form on*

$$(\mathbb{Z}E_1 + \cdots + \mathbb{Z}E_r)/\mathbb{Z}F$$

is negative definite and unimodular.

We now pursue a finer analysis of the structure of ruled surfaces.

PROPOSITION 1.19. *Let C be a variety defined over a field K such that $C_{\bar{K}} \simeq \mathbb{P}^1_{\bar{K}}$. Then there exists a closed embedding $C \hookrightarrow \mathbb{P}^2$ as a plane conic. There is a quadratic extension K'/K such that $C_{K'} \simeq \mathbb{P}^1_{K'}$.*

We leave the proof as an exercise.

We apply Proposition 1.19 in the case where K is the function field of the base B and C is the generic fiber f:

$$\begin{array}{ccc} C & \subset & X \\ {\scriptstyle f_\circ}\downarrow & & \downarrow{\scriptstyle f} \\ \mathrm{Spec}(k(B)) & \to & B \end{array}$$

The Tsen-Lang theorem says that every quadratic form in ≥ 3 variables over $k(B)$ represents zero, so $C(k(B)) \neq \emptyset$. Each rational point corresponds to a section $\mathrm{Spec}(k(B)) \to C$ of $f_\circ$, and thus to a rational map from B to X. Since X is proper, this extends uniquely to a section $s : B \to X$ of f. We have proven the following:

PROPOSITION 1.20. *Let $f : X \to B$ be a ruled surface. There exists a section $s : B \to X$ of f.*

Combining this with the argument for Proposition 1.15, we obtain

COROLLARY 1.21 (Classification of ruled surfaces). *Every minimal ruled surface $f : X \to B$ is isomorphic to $\mathbb{P}(\mathcal{E})$ for some rank-two vector bundle $\mathcal{E}$ on B.*

Combining this with Grothendieck's classification of vector bundles on $\mathbb{P}^1$ gives:

COROLLARY 1.22. *Every ruled surface $f : X \to \mathbb{P}^1$, minimal relative to f, is isomorphic to a* Hirzebruch surface

$$\mathbb{F}_d := \mathbb{P}(\mathcal{O}_{\mathbb{P}^1} \oplus \mathcal{O}_{\mathbb{P}^1}(-d)), \quad d \geq 0.$$

In particular, ruled surfaces over $\mathbb{P}^1$ are rational.

See Exercise 1.3.3 for more details of the argument.

Exercises.

EXERCISE 1.3.1. Prove Proposition 1.19. *Hint:* Note that Ω^1_C is an invertible sheaf on C defined over K and coincides with $\mathcal{O}_{\mathbb{P}^1}(-2)$ over $C_{\bar{K}}$. Use the sections of the dual $(\Omega^1_C)^*$ to embed C.

EXERCISE 1.3.2. Prove Proposition 1.18. *Hint:* Use the mechanism of the proof of Theorem 1.16.

EXERCISE 1.3.3. Give a detailed proof for Corollary 1.22. For the classification assertion, show that each vector bundle $\mathcal{E}$ on $\mathbb{P}^1$ decomposes as

$$\mathcal{E} \simeq \bigoplus_{i=1}^{r} \mathcal{O}_{\mathbb{P}^1}(a_i), \quad a_1 \leq a_2 \leq \cdots \leq a_r.$$

To establish rationality, exhibit a nonempty open subset $U \subset \mathbb{P}^1$ such that $\mathcal{E}|U \simeq \mathcal{O}_U^{\oplus r}$.

EXERCISE 1.3.4. For each $d \geq 0$, show there exists a diagram

$$\begin{array}{ccc} & Y & \\ {\scriptstyle\beta}\swarrow & & \searrow{\scriptstyle\gamma} \\ \mathbb{F}_d & & \mathbb{F}_{d+1} \end{array}$$

where β (resp. γ) is the blow-up of $\mathbb{F}_d$ (resp. $\mathbb{F}_{d+1}$) at a suitable point.

2. Effective cones and classification

From the modern point of view, the presence of (-1)-curves is controlled by how the effective cone and the canonical class interact. In this section, we develop technical tools for analyzing this interaction. We continue to work over an algebraically closed field k.

2.1. Cones of curves and divisors. Let X be a smooth projective complex variety, $N_1(X,\mathbb{Z}) \subset H_2(X,\mathbb{Z})$ the sublattice generated by homology classes of algebraic curves in X, and $N^1(X,\mathbb{Z}) \subset H^2(X,\mathbb{Z})$ the *Néron-Severi group* parametrizing homology classes of divisors in X.

We can extend these definitions to fields of positive characteristic: Consider the Chow group of dimension (resp. codimension) one cycles in X; two cycles are *numerically equivalent* if their intersections with any divisor (resp. curve) are equal. We define $N_1(X,\mathbb{Z})$ (resp. $N^1(X,\mathbb{Z})$) as the quotient of the corresponding Chow group by the cycles numerically equivalent to zero. The rank of $N^1(X,\mathbb{Z})$ is bounded by the second (étale) Betti number of X; see [**Mil80**, V.3.28] for the surface case.

DEFINITION 2.1. A Cartier divisor D on a variety X is *nef* (numerically eventually free or numerically effective) if $D \cdot C \geq 0$ for each curve $C \subset X$.

Here is the main example: A Cartier divisor D is *semiample* if ND is basepoint-free for some $N \in \mathbb{N}$. Since ND remains basepoint-free when restricted to curves $C \subset X$, we have $D \cdot C = \deg D|C \geq 0$.

We have the monoid of effective curves

$$\mathrm{NE}_1(X,\mathbb{Z}) = \{[D] \in N_1(X,\mathbb{Z}) : D \text{ effective sum of curves } \}$$

and the associated closed cone

$$\overline{\mathrm{NE}}_1(X) = \text{ smallest closed cone containing } \mathrm{NE}_1(X,\mathbb{Z}) \subset N_1(X,\mathbb{R}),$$

as well as the monoid of effective divisors

$$\mathrm{NE}^1(X,\mathbb{Z}) = \{[D] \in N^1(X,\mathbb{Z}) : D \text{ effective divisor } \}$$

and the associated cone of *pseudo-effective divisors*

$$\overline{\mathrm{NE}}^1(X) = \text{ smallest closed cone containing } \mathrm{NE}^1(X,\mathbb{Z}) \subset N^1(X,\mathbb{R}).$$

We also have the *nef cone* $\overline{\mathrm{NM}}^1(X) \subset N^1(X,\mathbb{R})$ and $\mathrm{NM}^1_\circ(X)$ its interior. Note that $\overline{\mathrm{NM}}^1(X)$ and $\overline{\mathrm{NE}}_1(X)$ are dual in the sense that

$$\overline{\mathrm{NM}}^1(X) = \{D \in N^1(X,\mathbb{R}) : D \cdot C \geq 0 \text{ for each } C \in \overline{\mathrm{NE}}_1(X)\}.$$

These cones are governed by the following general results:

THEOREM 2.2. *Let X be a proper variety and D a Cartier divisor on X.*

(1) *(Nakai criterion) D is ample if and only if $D^{\dim(Z)} \cdot Z > 0$ for each closed subvariety $Z \subset X$.*
(2) *(Kleiman criterion) Assume X is projective. Then D is ample if and only if $D \in \mathrm{NM}^1_\circ(X)$, i.e., $D \cdot C > 0$ for each nonzero class C in the closure of the cone of curves on X.*

It is not difficult to verify that an ample divisor necessarily satisfies these conditions; we leave this as an exercise. We refer the reader to [**KM98**, §1.5] for proofs that these conditions are sufficient in general and to [**Har77**, §V.1] for Nakai's criterion in the special case of smooth projective surfaces.

THEOREM 2.3. *Let X be a smooth projective variety. A divisor $D \in N^1(X,\mathbb{Z})$ lies in the pseudoeffective cone $\overline{\mathrm{NE}}^1(X)$ if and only if, for each ample H and rational $\epsilon > 0$, some multiple of $D + \epsilon H$ is effective. It lies in the interior*

$$\mathrm{NE}^1_\circ(X) \subset \overline{\mathrm{NE}}^1(X)$$

if and only if there exists an $N \gg 0$ so that

$$ND = A + E$$

where A is ample and E is effective. Such divisors are said to be big.

PROOF. First, any divisor of the form $H + E$ is in the interior of the pseudoeffective cone. If B is an arbitrary divisor then $nH + B$ is very ample for some $n > 0$, and $n(H + E) + B$ is effective.

Conversely, let D lie in the interior of $\overline{\mathrm{NE}}^1(X)$. Consider

$$D - \mathrm{NM}^1_\circ(X) \subset N^1(X,\mathbb{R}),$$

i.e., the cone of anti-ample divisor classes translated so that the vertex is at D. Note that the ample cone of X is open, so we can pick a

$$B \in N^1(X,\mathbb{Q}) \cap (D - \mathrm{NM}^1_\circ(X))$$

and $m > 0$ so that $E := mB$ is an effective divisor. Express

$$B = D - tA$$

for A ample and $t \in \mathbb{Q}_{>0}$. Thus

$$D = \frac{1}{m}E + tA$$

and clearing denominators gives the desired result.

If D is not in $\overline{\mathrm{NE}}^1(X)$ then, for each H ample, there exists an $\epsilon > 0$ so that $D + \epsilon H$ is not effective. Conversely, if $D \in \overline{\mathrm{NE}}^1(X)$ then $D + \epsilon H \in \mathrm{NE}^1_\circ(X)$ and we can write

$$N(D + \epsilon H) = A + E$$

where $N \gg 0$, A is ample, and E is effective. □

COROLLARY 2.4. *Let X be a smooth projective surface. A nef divisor D is big if and only if $D^2 > 0$. Indeed, any divisor D (not necessarily nef) such that $D^2 > 0$ and $D \cdot H > 0$ for some ample divisor H is big.*

In fact, the analogous statement is true in all dimensions [**KM98**, 2.61].

PROOF. If D is big then we can express $D = A + E$, where A is an ample $\mathbb{Q}$-divisor and E is an effective $\mathbb{Q}$-divisor. We expand

$$D^2 = D \cdot (A + E) \geq D \cdot A = A \cdot A + A \cdot E > 0.$$

Conversely, if $D^2 > 0$ then the Riemann-Roch formula implies either

$$h^0(\mathcal{O}_X(mD)) \geq \frac{D^2}{2}m^2, \quad m \gg 0,$$

or

$$h^2(\mathcal{O}_X(mD)) \geq \frac{D^2}{2}m^2, \quad m \gg 0.$$

The latter possibility would imply $K_X - mD$ is effective for $m \gg 0$, which is incompatible with D being nef. (Actually, we only need that $D \cdot H > 0$ for some ample divisor H.) Given A very ample, a straightforward dimension count shows that $h^0(\mathcal{O}_X(mD - A))$ remains positive for $m \gg 0$, i.e., that D can be expressed as a sum of an ample and an effective divisor. □

Exercises.

EXERCISE 2.1.1. Let X denote the blow-up of $\mathbb{P}^2$ at a point. Give examples of big divisors D on X with $D^2 < 0$.

EXERCISE 2.1.2. Let X be a smooth projective variety. Verify that the conditions of the Nakai and Kleiman criteria are necessary for a divisor to be ample. When X is a surface, deduce the sufficiency of the Kleiman criterion from the Nakai criterion.

EXERCISE 2.1.3. The *volume* of a Cartier divisor D on an n-dimensional projective variety X is defined [**Laz04**, 2.2.31] as

$$\mathrm{vol}(D) = \limsup_{m\to\infty} h^0(X, \mathcal{O}_X(mD))/(m^n/n!).$$

When X is a smooth surface, show that D is big if and only if $\mathrm{vol}(D) > 0$.

2.2. Examples of effective cones of surfaces. For surfaces that are obtained by blowing up the plane $\beta : X \to \mathbb{P}^2$, we write L for the pullback of the line class on $\mathbb{P}^2$ and $E_1, E_2, \ldots$ for the exceptional curves.

(1) Let f_1 and f_2 denote the rulings of $X = \mathbb{P}^1 \times \mathbb{P}^1$, so that $\mathrm{Pic}(X) = \mathbb{Z}f_1 + \mathbb{Z}f_2$. Then we have

$$\overline{\mathrm{NE}}_1(X) = \{a_1 f_1 + a_2 f_2 : a_1, a_2 \geq 0\}$$

and

$$\mathrm{NE}_1^\circ(X) = \{a_1 f_1 + a_2 f_2 : a_1, a_2 > 0\}.$$

(2) If X is $\mathbb{P}^2$ blown up at one point then

$$\overline{\mathrm{NE}}_1(X) = \{aE + b(L - E) : a, b \geq 0\}.$$

(3) If X is $\mathbb{P}^2$ blown up at two points then

$$\overline{\mathrm{NE}}_1(X) = \{a_1E_1 + a_2E_2 + b(L - E_1 - E_2) : a_1, a_2, b \geq 0\}.$$

(4) If X is $\mathbb{P}^2$ blown up at three non-collinear points then

$$\overline{\mathrm{NE}}_1(X) = \langle L - E_1 - E_2, L - E_1 - E_3, L - E_2 - E_3, E_1, E_2, E_3 \rangle,$$

i.e., the cone generated by the designated divisors.

(5) If X is a cubic surface with lines $\ell_1, \ldots, \ell_{27}$ then

$$\overline{\mathrm{NE}}_1(X) = \langle \ell_1, \ldots, \ell_{27} \rangle.$$

We describe a technique for verifying these claims, using the crucial fact that curves and divisors coincide on surfaces, i.e., $\overline{\mathrm{NE}}_1(X) = \overline{\mathrm{NE}}^1(X)$. To decide whether a collection of irreducible curves $\Gamma = \{C_1, \ldots, C_N\}$ generates $\overline{\mathrm{NE}}_1(X)$, it suffices to

- Compute a set of generators Ξ for the dual cone $\langle \Gamma \rangle^*$; there are computer programs like PORTA [**CL97**] and Polymake [**GJ00**] which can extract Ξ from Γ.
- Check that each $A_i \in \Xi$ can be written $A_i = \sum m_{ij} C_j, m_{ij} \geq 0$.

Here is why this works: If D is effective then we can write

$$D = M + F, \quad F = \sum_j n_j C_j, n_j \geq 0,$$

where M is effective with no support at $C_1, \ldots, C_N$. (Here F is the portion of the fixed part of D supported in Γ.) In particular, $M \cdot C_j \geq 0$ for each j, i.e., $M \in \langle \Gamma \rangle^*$. But then $M = \sum_i a_i A_i$ with $a_i \geq 0$. Thus we have

$$M = \sum_{ij} a_i m_{ij} C_j$$

and D is an effective sum of the C_j.

Example 2.5. For $X = \mathrm{Bl}_{p_1, p_2} \mathbb{P}^2$ take $\Gamma = \{E_1, E_2, L - E_1 - E_2\}$, which generates a simplicial cone. The dual generators are

$$\Xi = \{L - E_1, L - E_2, L\}$$

and we can write

$$L - E_1 = (L - E_1 - E_2) + E_2, \quad L - E_2 = (L - E_1 - E_2) + E_1,$$
$$L = (L - E_1 - E_2) + E_1 + E_2.$$

It follows that Γ generates $\overline{\mathrm{NE}}_1(X)$.

Exercises.

Exercise 2.2.1. Verify each of our claims about the generators of the effective cone.

2.3. Extremal rays. We'll need the following general notion from convex geometry:

Definition 2.6. Given a closed cone $\mathcal{C} \subset \mathbb{R}^n$, an element $R \in \mathcal{C}$ generates an *extremal ray* if for each representation

$$R = D_1 + D_2, \quad D_1, D_2 \in \mathcal{C}$$

we have $D_1, D_2 \in \mathbb{R}_{\geq 0} R$.

We will conflate the element R and the ray $\mathbb{R}_{\geq 0}R$. For a polyhedral cone, i.e., one generated by a finite number of elements

$$\mathcal{C} = \langle C_1, \ldots, C_N \rangle = \mathbb{R}_{\geq 0}C_1 + \cdots + \mathbb{R}_{\geq 0}C_N,$$

the extremal rays correspond to the irredundant generators. On the other hand, for the cone over the unit circle

$$\{(x, y, z) : x^2 + y^2 - z^2 \leq 0\} \subset \mathbb{R}^3$$

each point of the circle yields an extremal ray.

Our main examples of extremal rays are (-1)-curves:

PROPOSITION 2.7. *Let X be a smooth projective surface and E a (-1)-curve. Then E is extremal in $\overline{\mathrm{NE}}_1(X)$.*

If $\beta : X \to Y$ is the blow-down of E then

$$\beta_* \overline{\mathrm{NE}}_1(X) = \overline{\mathrm{NE}}_1(Y),$$

hence faces of $\overline{\mathrm{NE}}_1(Y)$ correspond to faces of $\overline{\mathrm{NE}}_1(X)$ containing E.

PROOF. If we could express $E = D_1 + D_2$ with $D_1, D_2 \in \overline{\mathrm{NE}}_1(X)$ not in $\mathbb{R}_{\geq 0}E$, then

$$0 = \beta_* D_1 + \beta_* D_2$$

for nonzero $\beta_* D_i \in \overline{\mathrm{NE}}_1(Y)$. This contradicts the fact that $\overline{\mathrm{NE}}_1(Y)$ is strongly convex, i.e., that the origin is extremal.

The inclusion

$$\beta_* \overline{\mathrm{NE}}_1(X) \subset \overline{\mathrm{NE}}_1(Y)$$

is clear because the image of an effective divisor is effective. On the other hand, suppose that D is effective on Y. Since Y is nonsingular, D is a Cartier divisor and $\beta^* D$ is a well-defined effective Cartier divisor. The projection formula $\beta_* \beta^* D = D$ shows that $D \in \beta_* \overline{\mathrm{NE}}_1(X)$. □

COROLLARY 2.8. *Let X be Del Pezzo and $\beta : X \to Y$ a blow-down morphism. Then Y is Del Pezzo.*

PROOF. Indeed, we have the discrepancy formula

$$\beta^* K_Y = K_X - E,$$

where E is the exceptional curve. Since $-K_X$ is positive on $\overline{\mathrm{NE}}_1(X) \setminus \{0\}$, $-K_Y$ is positive on $\overline{\mathrm{NE}}_1(Y) \setminus \{0\}$. A direct argument that $-K_Y$ is ample can be extracted from the proof of the Castelnuovo Contraction Criterion (Theorem 1.12). □

DEFINITION 2.9. Let X be a smooth projective surface. The *positive cone* $\mathcal{C}$ denotes the component of

$$\{D : D^2 > 0\} \subset N^1(X, \mathbb{R})$$

containing the hyperplane class. (The Hodge index theorem implies this has two connected components.) Let $\overline{\mathcal{C}}$ denote its closure.

We can formulate a more general version of Proposition 2.7, which complements Theorem 2.3 and Corollary 2.4:

PROPOSITION 2.10. [**Kol96**, II.4] *Let X be a smooth projective surface. Then each irreducible curve D with $D^2 \leq 0$ lies in the boundary $\partial\overline{\mathrm{NE}}_1(X)$. Furthermore,*

$$\overline{\mathrm{NE}}_1(X) = \overline{\mathcal{C}} + \sum_D \mathbb{R}_{\geq 0} D \tag{2.1}$$

where the sum is taken over irreducible curves D with $D^2 < 0$.

PROOF. Corollary 2.4 implies $\overline{\mathrm{NE}}_1(X) \supset \overline{\mathcal{C}}$, and $\overline{\mathrm{NE}}_1(X) \supset \overline{\mathcal{C}} + \sum_D \mathbb{R}_{\geq 0} D$ follows immediately. It remains to establish the reverse inclusion.

First, suppose that D is irreducible with $D^2 \leq 0$. Let H be a very ample divisor of X. Then for each rational $\epsilon > 0$ we claim that $D - \epsilon H$ fails to be in the effective cone. Indeed, if $D - \epsilon H$ were effective then it could be expressed as a nonnegative linear combination of irreducible curves, some different from D,

$$D - \epsilon H \equiv c_0 D + \sum_j c_j D_j, \quad c_0 \in [0, 1), c_j > 0.$$

Regrouping terms, for some $\epsilon' > 0$ we obtain

$$D - \epsilon' H \equiv \sum_j c'_j D_j, \quad c'_j > 0.$$

However, this contradicts the fact that

$$D \cdot (D - \epsilon' H) < 0.$$

For D irreducible with $D^2 < 0$ consider the closed cone

$$V = \left\langle z \in \overline{\mathrm{NE}}_1(X) : z \cdot D \geq 0 \right\rangle \subset \overline{\mathrm{NE}}_1(X),$$

which contains all effective divisors without support along D. Thus $\overline{\mathrm{NE}}_1(X)$ is the smallest convex cone containing V and D. Since $D \notin V$, it follows that D is extremal in $\overline{\mathrm{NE}}_1(X)$.

Conversely, suppose that Z is extremal with $Z^2 < 0$. There is necessarily some irreducible curve C such that $C \cdot Z < 0$. Let Z_i denote a sequence of effective $\mathbb{Q}$-divisors approaching Z. Since $Z_i \cdot C < 0$ for $i \gg 0$, we must have that $C^2 < 0$. Moreover C appears in Z_i with coefficient c_i and $c = \lim c_i > 0$. Thus $Z - cC$ is pseudoeffective and $C \in \mathbb{R}_{\geq 0} Z$ by extremality. □

For special classes of surfaces, the negative extremal rays are necessarily (-1)-curves or (-2)-curves, i.e., smooth rational curves with self-intersection -2:

COROLLARY 2.11. [**Kol96**, II.4.14] *Suppose X is a smooth projective surface with $-K_X$ nef. Then the sum in expression (2.1) can be taken over D with $D^2 = -1$ or -2 and $D \simeq \mathbb{P}^1$.*

PROOF. Since $K_X \cdot D \leq 0$ and $D^2 < 0$ then the adjunction formula

$$2g(D) - 2 = D^2 + K_X \cdot D$$

allows only the possibilities listed. □

Exercises.

EXERCISE 2.3.1. Classify extremal rays and describe decomposition (2.1) for:

- $\mathbb{P}^2$ blown up at two points or three non-collinear points;
- the Hirzebruch surfaces $\mathbb{F}_d$.

2.4. Structural results on the cone of curves I. The closed cone of effective curves has a very nice structure in the region where the canonical class fails to be nef. There are two different approaches to these structural results. The first emphasizes vanishing theorems (for higher cohomology) on line bundles and the resulting implications for linear series, e.g., basepoint-freeness. You can find details of this approach in references such as [**Rei97**, D] and [**CKM88**]. One significant disadvantage is that the reliance on Kodaira-type vanishing makes generalization to positive characteristic problematic. The second approach emphasizes the geometric properties of the curves themselves, especially the bend-and-break technique of Mori. This approach is taken in Mori's original papers, as well as in [**Kol96**, II.4,III.1].

Since both approaches are important for applications, we will sketch the key elements of each, referring to the literature for complete arguments.

THEOREM 2.12 (Cone Theorem). [**Rei97**, D.3.2] [**KM98**, Thm. 3.7] [**CKM88**, 4.7] *Let X be a smooth projective surface with canonical class K_X. There exists a countable collection of $R_i \in \overline{\mathrm{NE}}_1(X) \cap N_1(X,\mathbb{Z})$ with $K_X \cdot R_i < 0$ such that*

$$\overline{\mathrm{NE}}_1(X) = \overline{\mathrm{NE}}_1(X)_{K_X \geq 0} + \sum_i \mathbb{R}_{\geq 0} R_i,$$

where the first term is the intersection of $\overline{\mathrm{NE}}_1(X)$ with the halfplane $\{v \in N_1(X,\mathbb{R}) : v \cdot K_X \geq 0\}$. Given any ample divisor H and $\epsilon > 0$, there exists a finite number of R_i satisfying $(K_X + \epsilon H) \cdot R_i \leq 0$.

COROLLARY 2.13. *Let X be a Del Pezzo surface. Then $\overline{\mathrm{NE}}_1(X)$ is a finite rational polyhedral cone.*

What's even more remarkable is that the extremal rays can be interpreted geometrically. The following theorem should be understood as a far-reaching extension of the Castelnuovo contraction criterion (Theorem 1.12):

THEOREM 2.14 (Contraction Theorem). [**Rei97**, D.4] [**KM98**, Theorem 3.7] *Let X be a smooth projective surface and R a generator of an extremal ray with $K_X \cdot R < 0$. There exists a morphism $\phi : X \to Y$ to a smooth projective variety, with the following properties:*

1. *$\phi_* R = 0$ and ϕ contracts those curves with classes in the ray $\mathbb{R}_{\geq 0} R$;*
2. *ϕ has relative Picard rank one and $\mathrm{Pic}(Y)$ can be identified with $R^\perp \subset \mathrm{Pic}(X)$.*

PROOF. The proofs of Theorems 2.12 and 2.14 are intertwined. We can only offer a sketch of the arguments required. Some of these work only in characteristic zero, but we will make clear which ones.

Suppose we want to analyze the part of the effective cone along which K_X is negative. Fix an ample divisor H, which is necessarily positive along $\overline{\mathrm{NE}}_1(X)$. Which divisors $\tau K_X + H, \tau \in [0,1]$, are nef? Consider the *nef threshold*

$$t = \sup\{\tau \in \mathbb{R} : \tau K_X + H \text{ nef }\},$$

i.e., $(tK_X + H)^\perp$ is a supporting hyperplane of $\overline{\mathrm{NE}}_1(X)$ provided $tK_X + H$ is nonzero. If we choose H suitably general, we can assume this hyperplane meets $\overline{\mathrm{NE}}_1(X)$ in an extremal ray. (Of course, for special H it might cut out a higher-dimensional face.)

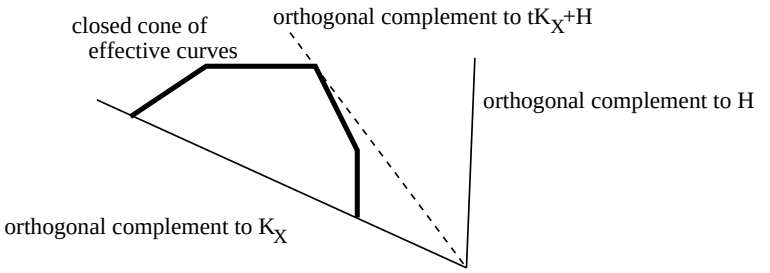


FIGURE 1. Finding a supporting hyperplane of the cone of curves (drawn in the projectivization of $N_1(X, \mathbb{R})$)

The first step is the rationality of the nef threshold. The most straightforward proof [**Rei97**, D.3.1] uses the Riemann-Roch formula and Kodaira vanishing, and thus is valid only in characteristic zero:

LEMMA 2.15 (Rationality). The nef threshold is rational.

Thus the K_X-negative extremal rays of the cone of effective curves are determined by linear inequalities with rational coefficients. These rays here can be chosen to be integral.

The second step is to show that the ($\mathbb{Q}$-)divisor $D := tK_X + H$ is semiample. Then the resulting morphism $\phi : X \to Y$ will contract precisely the extremal rays in the face supported by the hyperplane $(tK_X + H)^{\perp}$, which gives the contraction theorem.

LEMMA 2.16 (Basepoint-freeness). Let D be a nef $\mathbb{Q}$-divisor such that $D = tK_X + H$ for H ample and $t > 0$. Then D is semiample.

PROOF. Since D is nef we have $D^2 \geq 0$. If $D^2 > 0$ then the Nakai criterion (Theorem 2.2) implies D is ample unless there exists an irreducible curve E with $D \cdot E = 0$. The Hodge index theorem implies $E^2 < 0$. Since $K_X \cdot E < 0$, Proposition 1.6 implies that E is a (-1)-curve. The desired contraction exists by the Castelnuovo Criterion (Theorem 1.12).

Now suppose $D^2 = 0$. If D is numerically equivalent to zero then $-K_X$ is ample. In this situation, D is semiample if and only if it is torsion, which is a consequence of the following lemma.

LEMMA 2.17. Suppose that X is a smooth projective surface and $-K_X$ is nef and big, e.g., X is a Del Pezzo surface. Then

- $H^2(X, \mathcal{O}_X) = 0$ and $\mathrm{Pic}(X)$ is smooth;
- $H^1(X, \mathcal{O}_X) = 0$ and the identity component of $\mathrm{Pic}(X)$ is trivial.

PROOF. Since some positive multiple of $-K_X$ is effective, no positive multiple of K_X is effective. Thus

$$h^2(X, \mathcal{O}_X) = h^0(X, \mathcal{O}_X(K_X)) = 0$$

and $\mathrm{Pic}(X)$ is smooth. The identity component has dimension $q = h^1(X, \mathcal{O}_X)$.

Let $b_i(X)$ denote the ith Betti number of X; in positive characteristic, we define these using étale cohomology [**Mil80**]. Recall the formulas [**Mil80**, III.4.18,V.3.12]

$b_1(X) = 2q$ and

$$c_2(X) = \chi(X) = b_0(X) - b_1(X) + b_2(X) - b_3(X) + b_4(X) = 2 - 4q + b_2(X).$$

Noether's formula

$$12\chi(\mathcal{O}_X) = c_1(X)^2 + c_2(X)$$

and the fact that $c_1(X)^2 = K_X^2 > 0$ imply

$$12(1-q) > 2 - 4q + b_2(X).$$

Consequently

$$10 > 8q + b_2(X)$$

and thus $q = 0$ or $q = b_2(X) = 1$. To exclude the last case, observe that if the identity component of the Picard group is positive dimensional then so is the Albanese variety. (Indeed, these abelian varieties are dual to each other.) Furthermore, the Albanese map $X \to \mathrm{Alb}(X)$ [**Lan59**, II.3] is a dominating morphism to an elliptic curve. The classes of a fiber and the pullback of an ample divisor from the Albanese are necessarily independent; thus $b_2(X) \geq 2$. □

We return to the proof of Lemma 2.16. If D is not numerically trivial then $K_X \cdot D < 0$ and Riemann-Roch imply that $h^0(X, \mathcal{O}_X(mD))$ grows at least linearly in m. And since Corollary 2.4 ensures that D is not big, $h^0(\mathcal{O}_X(mD))$ cannot be a quadratic function of m. Decompose D into a moving and a fixed part

$$D = M + F, \quad M^2 \geq 0, \ M \cdot F \geq 0.$$

Note that $D \cdot F = M \cdot F + F^2 \geq 0$ (since D is nef), $M^2 = 0$ (as M is not big), and $F^2 \leq 0$ (since F is not big). On the other hand,

$$0 = D^2 = 2M \cdot F + F^2 \geq M \cdot F$$

so $M \cdot F = 0$ and $F^2 = 0$ as well. The Hodge index theorem implies that M and F are proportional in the Néron-Severi group, provided they are numerically nontrivial. In particular, if $F \neq 0$ then $K_X \cdot F < 0$ and $h^0(F, \mathcal{O}_X(mF))$ grows linearly in m, contradicting the fact that F is fixed. Thus $D = M$ is moving with perhaps isolated basepoints, the number of which is bounded by $M^2 = 0$. We conclude that D is basepoint-free. □

This completes the proof of Theorem 2.14.

Remark 2.18. This argument yields another result we shall use later: Let X be a smooth projective surface with $-K_X$ nef and big. Assume that D is a nef line bundle on X with $D^2 = 0$. Then D is semiample.

The third step is to bound the denominator of the nef threshold (cf. [**Rei97**, D.3.1] and [**CKM88**, 12.12]):

Lemma 2.19 (Bounding denominators). Assume the nef threshold is rational. Then its denominator is ≤ 3.

Proof. Again, the argument is a case-by-case analysis of $D = tK_X + H$. If $D^2 > 0$ then D is orthogonal to a (-1)-curve and $t \in \mathbb{Z}$. If D is numerically trivial then $-K_X$ is ample and Lemma 2.17 implies $\chi(\mathcal{O}_X) = 1$. Express $-K_X = rL$, where L is a primitive ample divisor and $r \in \mathbb{N}$. It suffices to show that $r \leq 3$. Noether's formula and the argument for Lemma 2.17 give

$$12 = r^2L^2 + c_2(X) = r^2L^2 + 2 + b_2(X)$$

so the only possibilities are $r = 1, 2, 3$.

It remains to consider the situation where $D^2 = 0$ but $D \not\equiv 0$. As we've seen, $D \cdot K_X < 0$ in this case. Furthermore, $t'K_X + H$ is never effective for $t' > t$. Here we claim $2t \in \mathbb{N}$. Otherwise, there exist $m, n \in \mathbb{N}$ with $m \gg 0$ such that

$$mt = n + \alpha, \quad 1/2 < \alpha < 1.$$

Thus $nK_X + mH$ is ample, $\Gamma(\mathcal{O}_X(-nK_X - mH)) = 0$, and

$$H^2(\mathcal{O}_X((n+1)K_X + mH)) = 0$$

by Serre duality. We deduce

$$\begin{aligned} h^0(\mathcal{O}_X((n+1)K_X + mH)) &= \chi(\mathcal{O}_X) + \tfrac{1}{2}((n+1)K_X + mH) \cdot (nK_X + mH) \\ &= \chi(\mathcal{O}_X) + \tfrac{1}{2}(-\alpha(1-\alpha)K_X^2 + m(1-2\alpha)D \cdot K_X + m^2D^2) \\ &= \text{constant} + m \cdot \text{positive number} \end{aligned}$$

which is positive for m sufficiently large. Thus $mH + (n+1)K_X$ is effective, a contradiction. □

To complete the proof of Theorem 2.12, we show that the K_X-negative extremal rays have no accumulation points and for any ample H there are finitely many such rays in the region $(K_X + \epsilon H) \leq 0$. Let $H_1, \ldots, H_d$ denote ample divisors forming a basis for the Néron-Severi group such that

$$H = a_1H_1 + \cdots + a_dH_d, \quad a_1, \ldots, a_d \in \mathbb{Q}_{>0}.$$

Let t_j denote the nef threshold of H_j. Consider the local coordinate functions

$$b_j(\gamma) = \frac{H_j \cdot \gamma}{-K_X \cdot \gamma}$$

on the open subset of $\mathbb{P}(N_1(X, \mathbb{R}))$ where $K_X \neq 0$. For K_X-negative extremal rays, $b_j \geq t_j$; Lemmas 2.15 and 2.19 imply these are rational numbers with denominators dividing six. It follows that these rays have no accumulation points. The extremal rays with $(NK_X + H) \cdot R_i \leq 0$ for some $N \in \mathbb{N}$ have coordinates satisfying

$$a_1b_1 + \cdots + a_db_d \leq N.$$

Since the a_j and b_j are positive rational numbers with bounded denominators, there are at most finitely many possibilities. □

The structure of the contraction morphism $\phi : X \to Y$ depends on the intersection properties of irreducible curves E generating our extremal ray:
Case $E^2 < 0$: Proposition 1.6 ensures E is a (-1)-curve and ϕ is the blow-down of E.
Case $E^2 = 0$: By adjunction, $E \simeq \mathbb{P}^1$ and $\phi : X \to Y$ fibers X over a curve with generic fiber $\mathbb{P}^1$. The extremality implies all fibers are irreducible and reduced, thus every fiber of ϕ is a projective line and we have a minimal ruled surface.
Case $E^2 > 0$: Corollary 2.4 implies E is big. Since f contracts E and all its deformations, ϕ is constant. Thus $\mathrm{Pic}(X) = \mathbb{Z}$ and X is Del Pezzo.

Since the first case cannot occur when X is minimal, we obtain:

Corollary 2.20. *Let X be a minimal smooth projective surface. Then one of the following conditions holds:*

- *K_X is nef;*
- *X is a $\mathbb{P}^1$-bundle over a curve B;*
- *X is Del Pezzo with $\mathrm{Pic}(X) = \mathbb{Z}$.*

In the first instance, X is the unique minimal smooth projective surface in its birational equivalence class.

PROOF. We only have to establish uniqueness: Let X' be another minimal smooth projective surface birational to X. Choose a factorization

$$\begin{array}{ccc} & Y & \\ {\scriptstyle\phi'}\swarrow & & \searrow{\scriptstyle\phi} \\ X' & & X \end{array}$$

where Y is smooth projective and the morphisms are birational. Indeed, ϕ and ϕ' are sequences of contractions of (-1)-curves (see Theorem 1.13). Express $K_Y = \phi^* K_X + F$ where F is an effective divisor with support equal to the exceptional locus of ϕ. If $E \subset Y$ is a (-1)-curve contracted by ϕ' then

$$-1 = K_Y \cdot E \geq F \cdot E$$

as K_X is nef. Thus E is contained in the support of F and is contracted by ϕ'. Since each ϕ'-exceptional divisor is ϕ-exceptional, we have a factorization

$$\phi : Y \to X' \to X.$$

Since X' is minimal, it must equal X. □

REMARK 2.21 (Relative version). Given a morphism $f : X \to B$ to a variety, we can also consider the relative cone of effective curves

$$\overline{\mathrm{NE}}_1(f : X \to B) = \{D \in \overline{\mathrm{NE}}_1(X) : f_* D = 0\}.$$

When B is smooth, this is the intersection of $\overline{\mathrm{NE}}_1(X)$ with the orthogonal complement to $f^*\mathrm{Pic}(B)$. The Cone Theorem 2.12 describes its structure in the region where the canonical divisor K_X is negative. There is a relative version of the Contraction Theorem giving contractions over B

$$\begin{array}{ccc} X & \xrightarrow{\phi} & Y \\ & \searrow \quad \swarrow & \\ & B & \end{array}.$$

The classification of ruled surfaces (Theorem 1.16) is a prime example.

Exercises.

EXERCISE 2.4.1. Let X be a smooth projective surface with K_X nef. Show that X is not rational.

EXERCISE 2.4.2. Let $p_1, \ldots, p_8 \in \mathbb{P}^2$ be general points. Let p_0 be the last basepoint of the pencil of cubic curves containing these points. Show that $X = \mathrm{Bl}_{p_0, p_1, \ldots, p_8}\mathbb{P}^2$ has an infinite number of (-1)-curves.

Hints: Let $E_0, \ldots, E_8$ denote the exceptional divisors. Consider the elliptic fibration

$$\eta : X \to \mathbb{P}^1$$

induced by the linear series $|f|$ with $f = -K_X = 3L - E_0 - \cdots - E_8$. Verify that sections of η are all (-1)-curves. Designate E_0 as the zero section of η and let $\sigma_i : \mathbb{P}^1 \to X, i = 1, \ldots, 8$ denote the sections associated with $E_1, \ldots, E_8$. Given sections $\sigma, \sigma' : \mathbb{P}^1 \to X$ we have

$$\begin{aligned} [(\sigma + \sigma')(\mathbb{P}^1)] &= [\sigma(\mathbb{P}^1)] + [\sigma'(\mathbb{P}^1)] - [E_0] + w(\sigma, \sigma')[f] \\ & \quad w(\sigma, \sigma') = -[\sigma(\mathbb{P}^1) - E_0] \cdot [\sigma'(\mathbb{P}^1) - E_0]. \end{aligned}$$

Use this to analyze $m_1\sigma_1 + \cdots + m_8\sigma_8$.

2.5. Structural results on the cone of curves II. Our discussion of the Cone and Contraction Theorems is missing one crucial element: We have not shown that the extremal rays are generated by classes of *rational* curves on X. Another issue is that we cited arguments for the rationality of extremal rays relying on vanishing theorems; these do not readily extend to positive characteristic. These gaps can be filled using Mori's 'bend-and-break' technique:

THEOREM 2.22 (Bend-and-break). [**Kol96**, II.5.14] *Let X be a smooth projective variety, C a smooth projective curve, and $f : C \to X$ a morphism. Let M be a nef $\mathbb{R}$-divisor. Assume that $-K_X \cdot C > 0$. Then for each $x \in f(C)$ there is a rational curve $x \in D_x \subset X$ such that*

$$M \cdot D_x \leq 2\dim(X)\frac{M \cdot C}{-K_X \cdot C}, \quad -K_X \cdot D_x \leq \dim(X) + 1.$$

This is a deep result that we will not prove here. The main idea is to use the fact that the anticanonical class is negative to show that f admits deformations $f_t : C \to X$ whose images still contain x. This strategy works beautifully provided C has genus zero, but in higher genus it is necessary to reduce modulo p and precompose f with the Frobenius map. Then we consider limits of $f_t(C) \subset X$ as $t \to 0$; these necessarily contain rational curves $x \in D' \subset X$. We can iterate this strategy until we obtain a rational curve $D_x \ni x$ with fairly small anticanonical degree, i.e., $-K_X \cdot D_x \leq \dim(X) + 1$.

We still have not mentioned the rôle of the divisor M. This is crucial in applications to the cone of curves:

THEOREM 2.23 (Cone Theorem bis). [**Kol96**, III.1.2] *Let X be a smooth projective surface with canonical class K_X. There exists a countable collection of rational curves $L_i \subset X$ with $0 < -K_X \cdot L_i \leq 3$ such that*

$$\overline{\mathrm{NE}}_1(X) = \overline{\mathrm{NE}}_1(X)_{K_X \geq 0} + \sum_i \mathbb{R}_{\geq 0}[L_i]. \tag{2.2}$$

Given any ample divisor H and $\epsilon > 0$, there exists a finite number of L_i satisfying $(K_X + \epsilon H) \cdot L_i \leq 0$.

PROOF. We offer a sketch proof following [**Kol96**]: Let M be an $\mathbb{R}$-divisor corresponding to a supporting hyperplane of a K_X-negative extremal ray $R \in \overline{\mathrm{NE}}_1(X)$. Thus $M \cdot R = 0$ and $M \cdot \gamma > 0$ for $\gamma \in \overline{\mathrm{NE}}_1(X)$ with $\gamma \notin \mathbb{R}_{\geq 0}R$; in particular, M is a nef $\mathbb{R}$-divisor. We show that M is a supporting hyperplane of the closure W of the cone associated to the right-hand side of (2.2). (We refer the reader to [**Kol96**, III.1] for the argument that the right-hand side of (2.2) defines a closed cone.)

Assume this is not the case. Rescaling M if necessary, we may assume that $M \cdot D \geq 1$. for each irreducible curve $D \subset X$ with $[D] \in W$. Consider the functional

$$\begin{array}{rccc} \phi : N_1(X, \mathbb{R})_{K_X < 0} & \to & \mathbb{R} \\ \gamma & \mapsto & -M \cdot \gamma / K_X \cdot \gamma \end{array}$$

which is nonnegative on $\overline{\mathrm{NE}}_1(X)_{K_X<0}$ and positive away from $\mathbb{R}_{\geq 0}R$. Choose a sequence of effective curves with real coefficients approaching R

$$C_i = \sum_j a_{ij}C_{ij}, a_{ij} > 0, \quad \lim_{i\to\infty} C_i = R.$$

For each i, there exists an index j such that $K_X \cdot C_{ij} < 0$ and $\phi(C_{ij}) \leq \phi(C_i)$. And we have $\lim_{i\to\infty} \phi(C_i) = \phi(R) = 0$.

On the other hand, bend-and-break yields rational curves L_i such that $-K_X \cdot L_i \leq 3$ and

$$M \cdot L_i \leq 4\phi(C_{ij}) < 4\phi(C_i).$$

The left-hand side is bounded from below by 1 while the right-hand side approaches zero, so we obtain a contradiction. □

2.6. Classification of surfaces.

THEOREM 2.24. *If X is a Del Pezzo surface with* $\mathrm{Pic}(X) = \mathbb{Z}$ *then* $X \simeq \mathbb{P}^2$.

PROOF. Our argument follows [**Kol96**, III.3.7]. We first offer a short proof in characteristic zero: Let L be a line bundle generating $\mathrm{Pic}(X)$ with $L \cdot K_X < 0$. Lemma 2.17 ensures that $H^1(\mathcal{O}_X) = H^2(\mathcal{O}_X) = 0$, so by Hodge theory we have $b_1(X) = 0$ and $b_2(X) = 1$. Poincaré duality then implies $L \cdot L = 1$. Noether's formula

$$12\chi(\mathcal{O}_X) = c_1^2(X) + c_2(X)$$

implies $c_1^2(X) = K_X^2 = 9$. We conclude that $K_X = -3L$ and $\chi(L) = 3$. Since $h^2(X, L) = h^0(X, K_X - L) = 0$, we have $h^0(X, L) \geq 3$. Moreover, the members of the corresponding linear series are integral curves of genus zero, i.e., $\mathbb{P}^1$'s. A straightforward inductive argument shows that L is basepoint-free and thus induces a degree-one morphism $X \to \mathbb{P}^2$, i.e., $X \simeq \mathbb{P}^2$.

We only used characteristic zero to show that $K_X = -3L$. Suppose then that $K_X = -rL$ for some $r \in \mathbb{N}$, where L is a generator of $\mathrm{Pic}(X)$. We have already seen in the proof of Lemma 2.19 that $r = 1, 2, 3$. If $r = 2$ then

$$2g(L) - 2 = L^2 + K_X L = -L^2$$

so $L^2 = 2$ and $g(L) = 0$; Riemann-Roch then gives $\chi(X, L) = 4$. Arguing as above, L is basepoint-free and defines a morphism $\phi : X \to \mathbb{P}^n$ for $n \geq 3$. The image is a quadric surface or a plane, and the latter possibility would contradict nondegeneracy. However, a quadric surface cannot be a rank-one Del Pezzo surface.

Finally, suppose that $r = 1$. The Cone Theorem (Theorem 2.23) implies the existence of a rational curve $f : \mathbb{P}^1 \to X$ with $\deg f^*(-K_X) = \deg f^* L \leq 3$. We have $f_*[\mathbb{P}^1] = mL$ for some $m \in \mathbb{N}$ with $mL^2 \leq 3$, and consequently $K_X^2 \leq 3$. Since every curve in X has positive self-intersection, we can deduce a contradiction from the following fact:

LEMMA 2.25. Let Y be a Del Pezzo surface with $K_Y^2 \leq 4$. Then Y contains a (-1)-curve.

Such curves C are called *lines* because $-K_Y \cdot C = 1$.

There are two general approaches to this. The most direct (see [**Kol96**, III.3.6] or Exercise 2.6.2) is to express Y as a hypersurface in a suitable weighted projective space, i.e., as a cubic surface in $\mathbb{P}^3$ (when $K_Y^2 = 3$), a quartic surface in $\mathbb{P}(1, 1, 1, 2)$ (when $K_Y^2 = 2$), or as a sextic surface in $\mathbb{P}(1, 1, 2, 3)$ (when $K_Y^2 = 1$). Proving this entails a fair amount of *ad hoc* analysis of linear series. Another approach (cf. [**Isk79**]) involves showing that Y lifts to characteristic zero and using the classification tools available there. □

THEOREM 2.26 (Castelnuovo's Criterion). [**Bea96**, V.6] [**Kol96**, III.2.4] *Let X be a smooth projective minimal surface. Then X is rational if and only if*

$$q(X) = h^1(\mathcal{O}_X) = 0, \quad P_2(X) := h^0(X, \mathcal{O}_X(2K_X)) = 0.$$

PROOF. The necessity of the numerical conditions is clear, as $P_2(X)$ and $q(X)$ are birational invariants of smooth projective varieties. For sufficiency, we may assume that X is minimal and falls into one of the three categories of Proposition 2.20. The third case (where X is Del Pezzo with $\mathrm{Pic}(X) = \mathbb{Z}$) is rational by Theorem 2.24. In the second case (where X is ruled over a curve B), the assumption $q(X) = 0$ implies that B has genus zero. Corollary 1.22 yields that X is rational. Finally, suppose that K_X is nef, so in particular $K_X^2 \geq 0$. We know that K_X is not effective; if $\Gamma(X, \mathcal{O}_X(K_X)) \neq 0$ then $\Gamma(X, \mathcal{O}_X(2K_X)) \neq 0$. Thus

$$\chi(\mathcal{O}_X) = h^0(\mathcal{O}_X) - h^1(\mathcal{O}_X) + h^2(\mathcal{O}_X) = 1$$

and

$$\chi(\mathcal{O}_X(-K_X)) = K_X^2 + 1 \geq 1.$$

Since $h^2(\mathcal{O}_X(-K_X)) = h^0(\mathcal{O}_X(2K_X)) = 0$ we conclude $h^0(\mathcal{O}_X(-K_X)) > 0$, i.e., $-K_X$ is effective. As K_X is nef, the only possibility is K_X trivial, a contradiction. □

COROLLARY 2.27. *Del Pezzo surfaces are rational.*

PROOF. Let X be a Del Pezzo surface. Since $-K_X$ is ample we have that $P_2(X) = 0$. Lemma 2.17 gives $h^1(\mathcal{O}_X) = 0$. □

COROLLARY 2.28. *Each Del Pezzo surface X is isomorphic to one of the following:*

- $\mathbb{P}^1 \times \mathbb{P}^1$*;*
- *a blow-up of $\mathbb{P}^2$ at eight or fewer points.*

PROOF. By Corollary 2.8, we just need to show that minimal Del Pezzo surfaces X are either $\mathbb{P}^2$ or $\mathbb{P}^1 \times \mathbb{P}^1$. Our previous analysis implies X is $\mathbb{P}^2$ or a Hirzebruch surface $\mathbb{F}_d$. But then X contains a rational curve of self-intersection $-d$, so $d = 0, 1$ by Proposition 1.6. □

REMARK 2.29. The classification of complex surfaces goes back to the work of Castelnuovo and Enriques in the late 19th and early 20th centuries. The extension to positive characteristic is largely due to Zariski, who first proved the Castelnuovo rationality criterion in this context [**Zar58a, Zar58b**].

Exercises.

EXERCISE 2.6.1. Suppose that X is a surface such that K_X is not nef and $\mathrm{Pic}(X)$ has rank at least three. Then X contains a (-1)-curve.

EXERCISE 2.6.2. Let Y be a Del Pezzo surface with $K_Y^2 = 3$ (resp. $K_Y^2 = 4$). Show that $-K_Y$ is very ample and the image under $|-K_Y|$ is a cubic surface (resp. complete intersection of two quadric hypersurfaces.) Conclude that Y contains a line (cf. Corollary 1.9).

3. Classifying surfaces over non-closed fields

Let k be a perfect field with algebraic closure $\bar{k}$ and Galois group $G = \mathrm{Gal}(\bar{k}/k)$. Let X be a smooth projective surface over k so that

$$\bar{X} = X_{\bar{k}} = X \times_{\mathrm{Spec}(k)} \mathrm{Spec}(\bar{k})$$

is connected. We use $\mathrm{Pic}(X)$ to denote line bundles on X defined over k.

3.1. Minimal surfaces.

DEFINITION 3.1. A smooth projective surface X over k is *minimal* if any birational morphism over k to a smooth surface

$$\phi : X \to Y$$

is an isomorphism.

THEOREM 3.2. *X is minimal if and only if $\bar{X}$ admits no Galois-invariant collection of pairwise disjoint (-1)-curves.*

PROOF. Suppose X is not minimal and admits a birational morphism $\phi : X \to Y$. By Theorem 1.13, $\bar{X}$ admits a (-1)-curve E contracted by ϕ. Since ϕ is birational there are only a finite number of such curves, so let $E_1, \ldots, E_r$ denote the curves in the Galois orbit of E. As we saw in the proof of Theorem 1.13, the intersection form on $\mathbb{Z}E_1 + \cdots + \mathbb{Z}E_r$ is negative definite, thus the matrix

$$\begin{pmatrix} E_i^2 & E_iE_j \\ E_iE_j & E_j^2 \end{pmatrix}, \quad i \neq j,$$

has positive determinant. It follows that $E_i \cdot E_j < 1$, which gives the disjointness.

Conversely, let $E_1, \ldots, E_r$ denote a Galois-invariant collection of pairwise disjoint (-1)-curves. Let H be an ample divisor on X. Since $H \cdot E_i = H \cdot E_j$ for each i, j, the divisor

$$H' = H + \sum_{j=1}^{r} (H \cdot E_i)E_i$$

is also Galois-invariant. We just take $Y = \mathrm{Proj}(\bigoplus_{n \geq 0} \Gamma(X, nH'))$, as in the proof of Theorem 1.12. □

REMARK 3.3 (Galois-invariant classes versus divisors defined over k). Not every element $L \in \mathrm{Pic}(\bar{X})^G$ comes from a line bundle defined over k. Applying the Hochschild-Serre spectral sequence [**Mil80**, III.2.20], we find

$$H^1(X, \mathcal{O}_X^*) = \ker\left(H_G^0 H^1(\bar{X}, \mathcal{O}_{\bar{X}}^*) \xrightarrow{d_2^{01}} H_G^2 H^0(\bar{X}, \mathcal{O}_{\bar{X}}^*) \right)$$

which yields

$$\mathrm{Pic}(X) = \ker\left(\mathrm{Pic}(\bar{X})^G \xrightarrow{d_2^{01}} \mathrm{Br}(k) \right).$$

Since $\mathrm{Br}(k)$ is torsion, some power NL with $N > 0$ is defined.

On the other hand, when $X(k) \neq \emptyset$ the homomorphism d_2^{01} is trivial. Indeed, the spectral sequence shows that the image of d_2^{01} lies in the kernel of the homomorphism

$$s^* : \mathrm{Br}(k) \to \mathrm{Br}(X) = H^2(X, \mathcal{O}_X^*)$$

induced by the structure map $s : X \to \mathrm{Spec}(k)$. Each rational point $x : \mathrm{Spec}(k) \to X$ induces $x^* : \mathrm{Br}(X) \to \mathrm{Br}(k)$, a left-inverse of s^*. Thus s^* is injective and d_2^{01} is trivial. (See [**CTS87**] for a comprehensive discussion.)

EXAMPLE 3.4. Suppose we have a cubic surface X with six conjugate disjoint lines $E_1, \ldots, E_6$. Does it follow that X is the blow-up of $\mathbb{P}^2$ at six conjugate points?

The divisor class $-K_X + E_1 + \cdots + E_6 = 3L$ is definitely defined over k. The corresponding linear series gives a morphism

$$X \to Y \subset \mathbb{P}^9$$

blowing down $E_1, \ldots, E_6$; here $\bar{Y} \simeq \mathbb{P}^2_{\bar{k}}$ is embedded via the cubic Veronese embedding. This is an example of a *Brauer-Severi variety*, i.e., a variety Y such that $\bar{Y} \simeq \mathbb{P}^{\dim(Y)}_{\bar{k}}$. Moreover, the invariant class

$$L \in \mathrm{Pic}(\bar{X})^G$$

comes from $\mathrm{Pic}(X)$ if and only if $Y \simeq \mathbb{P}^2_k$. A diagram-chase shows that $d_2^{01}(L) \in \mathrm{Br}(k)$ vanishes if and only if $[Y] \in \mathrm{Br}(k)$ is trivial.

Exercises.

EXERCISE 3.1.1. Let Y be a Brauer-Severi surface. Show there exists a smooth cubic surface X admitting a birational morphism $\phi : X \to Y$. *Hint:* A generic vector field on Y vanishes at three Galois-conjugate points. Blow up along two such collections of points.

EXERCISE 3.1.2 (Degree seven Del Pezzo surfaces). Let X be a surface such that $\bar{X} \simeq \mathrm{Bl}_{p_1,p_2}(\mathbb{P}^2)$. Show there exists a birational morphism $X \to \mathbb{P}^2$, obtained by blowing up a pair of Galois-conjugate points.

EXERCISE 3.1.3 (Some degree eight Del Pezzo surfaces). Let X be a surface such that $\bar{X} \simeq \mathrm{Bl}_p(\mathbb{P}^2)$. Show that X is isomorphic to $\mathrm{Bl}_p(\mathbb{P}^2)$ over k.

EXERCISE 3.1.4 (Degree five Del Pezzo surfaces). [**Sko93**] [**SD72**] Let X be a surface such that $\bar{X} \simeq \mathrm{Bl}_{p_1,p_2,p_3,p_4}(\mathbb{P}^2)$, where the points are distinct and no three are collinear.

(1) Show that the four points are projectively equivalent to
$$[1,0,0], [0,1,0], [0,0,1], [1,1,1]$$
over $\bar{k}$.
(2) Show that sections of $-K_X$ embed X as a quintic surface in $\mathbb{P}^5$.
(3) Show that this surface is cut out by five quadrics. *Hint:* It suffices to verify this on passage to $\bar{k}$.
(4) Choose generic $Q_0, Q_1, Q_2 \in I_X(2)$. Verify that
$$Q_0 \cap Q_1 \cap Q_2 = X \cup W,$$
where $\bar{W}$ is isomorphic to $\mathrm{Bl}_p\mathbb{P}^2$.
(5) Using Exercise 3.1.3, show that the exceptional divisor $E \subset W$ is defined over k and intersects X in one point.

Conclude that $X(k) \neq \emptyset$.

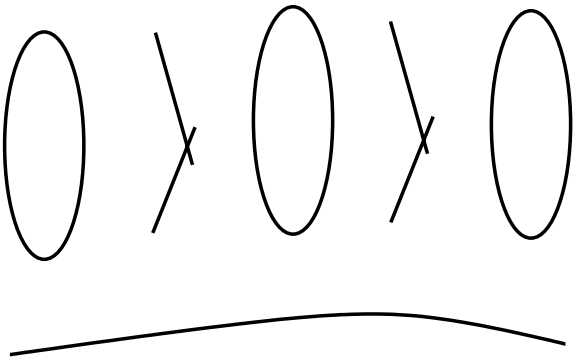


FIGURE 2. Degenerate fibers of a conic bundle

3.2. Conic bundles. Our treatment owes a great deal to Iskovskikh [**Isk79**].

DEFINITION 3.5. A *conic bundle* is a dominant morphism $f : X \to C$ from a smooth projective minimal surface X to a smooth curve, so that the generic fiber is a smooth curve of genus zero.

Of course, over an algebraically closed field this is the same as a minimal ruled surface. However, without the Tsen-Lang Theorem we cannot construct a section of f defined over k.

Proposition 1.19 does still apply: It guarantees that each smooth fiber of f is a plane conic and splits over a quadratic extension. It follows that there exists a *bisection* of f, i.e., an irreducible curve $D \subset C$ so that $f|_D : D \to C$ has degree two. Indeed, intersect the generic fiber (realized as a plane conic) with a line and take the closure in X.

THEOREM 3.6. *Let $f : X \to C$ be a conic bundle. Then any reducible fibers of $\bar{X} \to \bar{C}$ consist of two (-1)-curves intersecting in one point, conjugate under the Galois action.*

PROOF. Suppose F is a reducible fiber of $\bar{X}$; designate the field of definition of $f(F) \in C$ by $k_1 \supset k$. The existence of a reducible fiber guarantees that $\bar{X}$ is not relatively minimal and F contains a (-1)-curve (Theorem 1.16). Let $E_1, \dots, E_r$ be the Galois-orbit under the action of $\mathrm{Gal}(\bar{k_1}/k_1)$; the E_i are *not* pairwise disjoint by minimality (Theorem 3.2).

We claim that the only combinatorial possibility is $r = 2$, $E_1 \cdot E_2 = 1$, and $E_1^2 = E_2^2 = -1$. Write $T = E_1 \cup \cdots \cup E_r$ and set $n = E_i \cdot (\sum_{j \neq i} E_j)$, i.e., the number of points of intersection of each component with the other components. We can compute the arithmetic genus using

$$2p_a(T) - 2 = -2r + rn.$$

Since F has arithmetic genus zero $p_a(T) \leq 0$ and $n = 1$, i.e., each connected component of T consists of two (-1)-curves meeting at one point. Reordering indices if needed, let $E_1 \cup E_2$ denote one of these components.

By Proposition 1.18, if $E_1, \ldots, E_{r+s}$ are the irreducible components of F then the intersection form on

$$(\bigoplus_{j=1}^{r+s} \mathbb{Z}E_j)/\mathbb{Z}F$$

is negative definite. However, we have $(E_1 + E_2)^2 = -1 + 2 - 1 = 0$ so necessarily $F = E_1 + E_2$. This proves the claim and the result. □

We have seen (Proposition 1.19) that the generic fiber of $f : X \to C$ admits a natural realization as a smooth plane conic. This is obtained using sections of the dual to the differential one-forms. We can extend this over all of C using the *relative dualizing sheaf*

$$\omega_f = \Omega^2_X \otimes (f^*\Omega^1_C)^{-1}.$$

We have natural homomorphisms

$$\Omega^1_{f^{-1}(p)} \to \omega_{f^{-1}(p)} = \omega_f|_{f^{-1}(p)},$$

where the first arrow is an isomorphism wherever f is smooth.

COROLLARY 3.7 (Conic bundles really are conic bundles). *Let $f : X \to C$ be a conic bundle with relative dualizing sheaf ω_f. Then we have an embedding over C*

$$\begin{array}{ccc} X & \stackrel{j}{\hookrightarrow} & \mathbb{P}(f_*\omega_f^{-1}) \\ & \searrow \quad \swarrow & \\ & C & \end{array}$$

realizing each fiber of X as a plane conic.

PROOF. We use the classification of fibers in Theorem 3.6. For the smooth fibers, the anticanonical embedding has already been discussed in Proposition 1.19. For the reducible fibers, the anticanonical sheaf is very ample, realizing the fiber as a union of two distinct lines in $\mathbb{P}^2$.

Thus for each $p \in C$, $\omega_f^{-1}|_{f^{-1}(p)}$ is very ample and has no higher cohomology. Cohomology and base change gives that $f_*\omega_f^{-1}$ is locally free of rank three and has cohomology commuting with base extension. Thus we obtain a closed embedding over C

$$j : X \hookrightarrow \mathbb{P}(f_*\omega_f^{-1})$$

in a $\mathbb{P}^2$-bundle over C. □

DEFINITION 3.8. A *rational conic bundle* is a conic bundle $f : X \to C$ over a curve of genus zero.

3.3. Analysis of Néron-Severi lattices. We analyze the Néron-Severi group of rational conic bundles $f : X \to \mathbb{P}^1$. Note that K_X is defined over the base field.

Theorem 3.6 and Corollary 1.22 imply that $\bar{X}$ is a blow-up of a Hirzebruch surface at r points in distinct fibers:

$$\begin{array}{ccc} \bar{X} & \longrightarrow & \mathbb{F}_d \\ & \searrow \quad \swarrow & \\ & \mathbb{P}^1 & \end{array} .$$

The corresponding reducible fibers of $\bar{X} \to \mathbb{P}^1$ are denoted

$$E_1 \cup E_1', E_2 \cup E_2', \ldots, E_r \cup E_r',$$

so that $E_i + E_i' = F$ for each i.

There are a number of natural lattices to consider. We have the relative Néron-Severi lattice

$$N^1(\bar{X} \to \mathbb{P}^1, \mathbb{Z}) = \{D \in N^1(\bar{X}, \mathbb{Z}) : f_* D = 0\} = \mathbb{Z}F + \mathbb{Z}E_1 + \cdots + \mathbb{Z}E_r,$$

the quotient lattice

$$N^1(\bar{X} \to \mathbb{P}^1, \mathbb{Z})/\mathbb{Z}F = (\mathbb{Z}E_1 + \mathbb{Z}E_2 + \cdots + \mathbb{Z}E_r + \mathbb{Z}F)/\mathbb{Z}F$$

with matrix

	E_1	E_2	$\dots$	E_r
E_1	-1	0	$\dots$	0
E_2	0	-1	$\dots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
E_r	0	0	$\dots$	-1

,

and the image Λ of the orthogonal complement $K_{\bar{X}}^{\perp}$. This is generated by

$$\rho_1 = E_1' - E_2, \rho_2 = E_1 - E_2, \rho_3 = E_2 - E_3, \dots, \rho_r = E_{r-1} - E_r$$

with intersection matrix

	ρ_1	ρ_2	ρ_3	ρ_4	$\cdots$	ρ_{r-1}	ρ_r
ρ_1	-2	0	1	0	$\dots$	0	0
ρ_2	0	-2	1	0	$\ddots$	$\ddots$	$\vdots$
ρ_3	1	1	-2	1	$\ddots$	$\ddots$	$\vdots$
ρ_4	0	0	1	-2	$\ddots$	$\ddots$	0
$\vdots$	$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	1	0
ρ_{r-1}	0	$\ddots$	$\ddots$	$\ddots$	1	-2	1
ρ_r	0	$\dots$	$\dots$	0	0	1	-2

Up to sign, this is the Cartan matrix associated to the root system $\mathbf{D}_r$.

Recall the traditional description of $\mathbf{D}_r$: Consider

$$\mathbb{Z}^r = \mathbb{Z}e_1 + \cdots + \mathbb{Z}e_r$$

with the standard pairing $e_i \cdot e_j = \delta_{ij}$. Consider the index two sublattice

$$M = \{m_1 e_1 + \cdots + m_r e_r : m_1 + \ldots + m_r \equiv 0 \pmod 2\} \subset \mathbb{Z}^r$$

with generators

$$\{-e_1 - e_2, e_1 - e_2, e_2 - e_3, \dots, e_{r-1} - e_r\}.$$

The Weyl group $W(\mathbf{D}_r)$ acts on M via reflections associated to the roots $\{\pm e_i \pm e_j\}$. It can be identified with signed $r \times r$ permutation matrices with determinant equal to the sign of the permutation. It is thus a semidirect product

$$W(\mathbf{D}_r) = (\mathbb{Z}/2\mathbb{Z})^{r-1} \rtimes \mathfrak{S}_r,$$

where the first group should be interpreted as the diagonal matrices with entries ± 1 and determinant 1 and the second group as the permutation matrices. Each element of $W(\mathbf{D}_r)$ is thus classified by the induced permutation of signed coordinate vectors

$$\{e_1, e_1' = -e_1, \dots, e_r, e_r' = -e_r\}.$$

Identifying E_i with e_i and E_i' with e_i', we obtain isomorphisms of lattices

$$\begin{array}{ccc} M & \simeq & -\Lambda \\ \downarrow & & \downarrow \\ \mathbb{Z}^r & \simeq & -N^1(f : \bar{X} \to \mathbb{P}^1, \mathbb{Z})/\mathbb{Z}F \end{array}$$

where the vertical arrows are inclusions of index-two subgroups. The Galois action of $G = \mathrm{Gal}(\bar{k}/k)$ on $\mathrm{Pic}(\bar{X})$ induces actions on both Λ and $N^1(\bar{X} \to \mathbb{P}^1, \mathbb{Z})/\mathbb{Z}F$. It is worthwhile to compare these to the action of $W(\mathbf{D}_r)$ on M and $\mathbb{Z}^r$.

Exercises.

EXERCISE 3.3.1. Let $\phi : X \to Y$ be a birational extremal contraction of smooth projective surfaces, i.e., a contraction of a collection of pairwise disjoint (-1)-curves. Show that

$$\Lambda = K_{\bar{X}}^{\perp} \cap N^1(\phi : \bar{X} \to \bar{Y}, \mathbb{Z})$$

is isomorphic to the lattice

$$\mathbb{Z}\rho_1 + \cdots + \mathbb{Z}\rho_{r-1}$$

with intersections

$$\rho_i \cdot \rho_j = \begin{cases} -2 & \text{if } i = j \\ 1 & \text{if } |i-j| = 1 \\ 0 & \text{if } |i-j| > 1 \end{cases}.$$

This is the Cartan matrix for $\mathbf{A}_{r-1}$. Interpret the action of the Weyl group $W(\mathbf{A}_{r-1}) \simeq \mathfrak{S}_r$ in terms of the geometry of ϕ.

3.4. Classification of minimal rational surfaces over general fields. This is due to Manin [**Man66**] and Iskovskikh [**Isk79**]; another proof can be found in [**Kol96**, III.2].

THEOREM 3.9. *Let X be a smooth projective minimal surface with $\bar{X}$ rational. Then X is one of the following:*

- $\mathbb{P}^2$*;*
- $X \subset \mathbb{P}^3$ *a smooth quadric with* $\mathrm{Pic}(X) = \mathbb{Z}$*;*
- *a Del Pezzo surface with* $\mathrm{Pic}(X) = \mathbb{Z}K_X$*;*
- *a conic bundle $f : X \to C$ over a rational curve, with* $\mathrm{Pic}(X) \simeq \mathbb{Z} \oplus \mathbb{Z}$*.*

Notice that the third case includes Brauer-Severi surfaces.

Thus if Y is a smooth projective rational surface over k then there exists a birational morphism $\phi : Y \to X$ defined over k, where X is one of the surfaces listed in Theorem 3.9.

PROOF. Since $\bar{X}$ is rational $K_{\bar{X}}$ cannot be nef (Exercise 2.4.1), and there exists an irreducible curve $L \subset \bar{X}$ such that $K_{\bar{X}} \cdot L < 0$. In particular, $\overline{\mathrm{NE}}_1(\bar{X})$ admits $K_{\bar{X}}$-negative extremal rays. By the Cone Theorem 2.23, elements of $\overline{\mathrm{NE}}_1(\bar{X})$ can be expressed as

$$C + \sum a_i[L_i], \quad a_i > 0, \tag{3.1}$$

where $C \in \overline{\mathrm{NE}}_1(\bar{X})$ satisfies $C \cdot K_{\bar{X}} \geq 0$ and the L_i are rational curves generating $K_{\bar{X}}$-negative extremal rays. Of course, the Galois group G acts on $\mathrm{Pic}(\bar{X})$ and on the $K_{\bar{X}}$-negative extremal rays. Thus for elements of $\overline{\mathrm{NE}}_1(\bar{X})$ the two parts of (3.1) can be taken to be G-invariant.

Let $\overline{\mathrm{NE}}_1(\bar{X})^G$ denote the closure of the Galois-invariant effective cone in the real vector space spanned by Galois-invariant curve classes. Since $\overline{\mathrm{NE}}_1(\bar{X})^G$ has K_X-negative curves, it necessarily admits a K_X-negative extremal ray Z. This need not be extremal in $\overline{\mathrm{NE}}_1(\bar{X})$, but it does lie in some face of that cone, which we analyze. Since Z is extremal and K_X-negative, it must be proportional to the average over the orbit of a single extremal ray of $\bar{X}$:

$$Z = aE, \quad E = \sum_{j=1}^{n} L_j, \quad L_j = g_j L, g_j \in G.$$

In other words, the minimal face of $\overline{\mathrm{NE}}_1(\bar{X})$ containing Z is spanned by the Galois orbit of one extremal ray.

Assume first that $\mathrm{Pic}(X) \simeq \mathbb{Z}$, generated by some ample divisor H defined over k. Then $-K_X = rH$ for some positive integer r and X is Del Pezzo. When $r > 1$ we necessarily have $\bar{X} \simeq \mathbb{P}^2$ or $\mathbb{P}^1 \times \mathbb{P}^1$ by Corollary 2.28. In the first instance, X is a Brauer-Severi surface with a line H defined over the ground field; thus $\Gamma(\mathcal{O}_X(H))$ gives an isomorphism $X \simeq \mathbb{P}^2$. This is the first case of the theorem. In the second instance, the line bundle $\mathcal{O}_{\mathbb{P}^1 \times \mathbb{P}^1}(1,1)$ is defined over the ground field. Its global sections give an embedding $X \hookrightarrow \mathbb{P}^3$ whose image is a quadric surface. This is the second case of the theorem. Finally, if $-K_X$ generates $\mathrm{Pic}(X)$ then we are in the third case of the theorem.

Now assume that $\mathrm{Pic}(X)$ has higher rank, so in particular $E \in \partial\overline{\mathrm{NE}}_1(\bar{X})^G$. It follows that $E^2 \leq 0$. Indeed, if $E^2 > 0$ then E is big by Corollary 2.4 and thus lies in the interior of the effective cone by Theorem 2.3. (And there are some extremal L whose Galois orbits do not lie in any proper face of the cone of curves.)

Suppose now that $E^2 < 0$, which implies that $L^2 < 0$. As before, Proposition 1.6 implies L is a (-1)-curve. Furthermore, $L \cap L_j = \emptyset$ when $L \neq L_j$; indeed, if the Galois conjugates were nondisjoint then their sum would have nonnegative self-intersection. Theorem 3.2 implies that X is not minimal, a contradiction.

Suppose next that $E^2 = 0$. If $L^2 < 0$ then we would still have that L is a (-1)-curve. Since $E^2 = 0$ each curve meets precisely one of its Galois conjugates, transversely at one point. Thus the orbit of L decomposes as

$$\{L_1, L_1'\}, \{L_2, L_2'\}, \ldots, \{L_r, L_r'\},$$

where $L_i \cdot L_i' = 1$ and all other pairs of (-1)-curves are disjoint. Write $F_i = L_i + L_i'$ so that $F_i \cdot F_m = 0$ for each $i, m = 1, \ldots, r$; the Hodge index theorem implies that $F_1 = F_2 = \ldots = F_r$ and $E = rF_i$ for each i. Contracting E (or equivalently, $L_1, L_1', \ldots, L_r'$) gives a morphism

$$f : X \to C$$

whose generic fibers are smooth conics and with $r > 0$ degenerate fibers consisting of reducible singular conics. This is the conic bundle case of the theorem.

Finally, suppose that $E^2 = 0$ and $L^2 = 0$. Then each Galois conjugate of L is necessarily disjoint from L, so the Hodge index theorem argument above shows that $[L]$ is Galois-invariant. Contracting L gives a conic bundle $f : X \to C$ without degenerate fibers. □

Remark 3.10. This almost completes the birational classification of rational surfaces. It remains to enumerate birational equivalences among the surfaces listed in Theorem 3.9. This enumeration can be found in [**MT86**, 3.1.1, 3.3.2]

Exercises.

EXERCISE 3.4.1. To get a feeling for the difficulties involved, show that if X is minimal then $\bar{X} \neq \mathrm{Bl}_{p_1,\dots,p_9}\mathbb{P}^2$. Specify a Galois action on $\mathrm{Pic}(\bar{X})$, in particular, a finite group acting linearly, preserving the intersection form, and fixing K_X. Consider the orbits of the (-1)-curves under this action. Convince yourself there is an orbit consisting of either

- disjoint (-1)-curves; or
- r disjoint pairs of (-1)-curves, with each pair meeting transversely at one point.

3.5. An application: Rational points over function fields. Our next result is due to Manin and Colliot-Thélène [**CT87**]. For more context and discussion, see [**Kol96**, IV.6]:

THEOREM 3.11. *Let B be a smooth curve over $\mathbb{C}$ with function field $k = \mathbb{C}(B)$. Suppose that X is a smooth projective surface over k with $\bar{X}$ rational. Then $X(k) \neq \emptyset$.*

Of course, this result can be obtained from the Graber-Harris-Starr Theorem [**GHS03**]. However, we will present it using our classification techniques.

PROOF. We first reduce to the case where X is minimal. Suppose we have a birational morphism $\phi : X \to Y$ to a smooth projective surface. We can factor ϕ as a sequence

$$X = X_0 \to X_1 \to X_2 \to \cdots \to X_r = Y$$

where each intermediate morphism is the blow-up of a Galois-orbit of points.

Suppose $x \in X_i(k)$ is a rational point. If x is contained in the center of the blow-up $\beta_i : X_{i-1} \to X_i$ then the exceptional divisor $E \subset X_{i-1}$ is rational over k and isomorphic to $\mathbb{P}^1$. It follows that $E(k) \neq \emptyset$ and $X_{i-1}(k) \neq \emptyset$. If x is disjoint from the center of $X_{i-1} \to X_i$ then x lies in the open subset $U \subset X_{i-1}$ over which β_i is an isomorphism. Thus $\beta_i^{-1}(x)$ is a rational point of X_i.

We consider the minimal cases one by one. The case $X = \mathbb{P}^2$ is straightforward. The case of a quadric surface $Q \subset \mathbb{P}^3$ follows from the Tsen-Lang Theorem.

We address the cases of Del Pezzo surfaces of degree $d = K_X^2$. Del Pezzo surfaces with certain degrees *always* have rational points. We assume without proof standard results on anticanonical linear series $|-K_X|$ and embeddings of X in projective space:

$d = 7$ There is no minimal Del Pezzo surface in this degree—see Exercise 3.1.2.

$d = 8$ ($\bar{X} \simeq \mathrm{Bl}_p\mathbb{P}^2$) There is no minimal Del Pezzo surface of the type—see Exercise 3.1.3.

$d = 5$ X always has a rational point—see Exercise 3.1.4.

$d = 1$ $\bar{X} \simeq \mathrm{Bl}_{p_1,\dots,p_8}\mathbb{P}^2$ in this case and

$$-K_{\bar{X}} = 3L - E_1 - \cdots - E_8.$$

In this situation, $|-K_X|$ is the pencil of cubics $C_t, t \in \mathbb{P}^1$, passing through $p_1, \dots, p_8$. The base locus of this pencil on $\mathbb{P}^2$ consists of nine points, i.e., $p_1, \dots, p_8$ and one additional point p_0. The basis locus of $|-K_X|$ on X is just the point p_0. Since $-K_X$ is defined over k, the unique basepoint $p_0 \in X(k)$.

We address the remaining cases using the classification results of §2.6. Since $k = \mathbb{C}(B)$, we can use the following variant of the Tsen-Lang Theorem:

THEOREM 3.12. *Let k be the function field of a curve defined over an algebraically closed field. Let $F_1, \dots, F_r \in k[x_0, \dots, x_n]$ be nonconstant weighted homogeneous polynomials, with weighted degrees satisfying*

$$\deg(F_1) + \cdots + \deg(F_r) \leq n.$$

Then the system $F_1 = \cdots = F_r = 0$ admits a nontrivial solution over k.

$d = 3$ Here X is a cubic surface in $\mathbb{P}^3$ and the result follows from Tsen-Lang.

$d = 9$ X is a Brauer-Severi surface. However, Exercise 3.1.1 allows us to blow up X to obtain a cubic surface, which has rational points by the previous case. (The reader knowledgeable in central simple algebras can prove $\mathrm{Br}(\mathbb{C}(B)) = 0$ using properties of the reduced norm.)

$d = 4$ Here X is a complete intersection of two quadrics in $\mathbb{P}^4$ and our variant of the Tsen-Lang Theorem applies.

$d = 2$ Here $|-K_X|$ induces a morphism

$$X \to \mathbb{P}^2$$

of degree two, branched over a quartic plane curve. It follows that X is a hypersurface of degree four in the weighted projective space $\mathbb{P}(2, 1, 1, 1)$ of the form $w^2 = f(x, y, z)$. An application of the Tsen-Lang Theorem gives our result.

$d = 6$ It suffices to show there exists a quadratic extension k'/k over which rational points are dense on X. Then after blowing up two suitable conjugate points we obtain a degree-four Del Pezzo surface, which has a rational point.

From our analysis of the effective cone of X in §2.2, there are two nef divisors $L, L' \in \mathrm{Pic}(\bar{X})$ so that

$$L^2 = (L')^2 = 1, -K_X \cdot L = -K_X \cdot L' = 3.$$

Indeed, we take $L' = 2L - E_1 - E_2 - E_3$. Their sections induce morphisms

$$\phi, \phi' : \bar{X} \to \mathbb{P}^2$$

blowing down triples of disjoint (-1)-curves. Let k'/k be a quadratic extension over which L and L' are $\mathrm{Gal}(\bar{k}'/k')$ invariant. Then $X_{k'}$ is a blow-up of a Brauer-Severi variety Y over k' at three conjugate points. The $d = 9$ case shows that Y (and hence $X_{k'}$) has lots of k'-rational points.

For the conic bundle case we apply the Tsen-Lang theorem twice. First, we show that $C(k) \neq \emptyset$ so $C \simeq \mathbb{P}^1$. Taking a generic $t \in \mathbb{P}^1(k)$ so that $X_t := f^{-1}(t)$ is a smooth conic, a second application gives $X_t(k) \neq \emptyset$. □

Exercises.

EXERCISE 3.5.1. Let X be a degree-one Del Pezzo surface over an arbitrary field k. Give a complete proof that $X(k) \neq \emptyset$, based on the sketch above. **Challenge:** When can you show that $|X(k)| > 1$?

4. Singular surfaces

In this section, we work over an algebraically closed field.

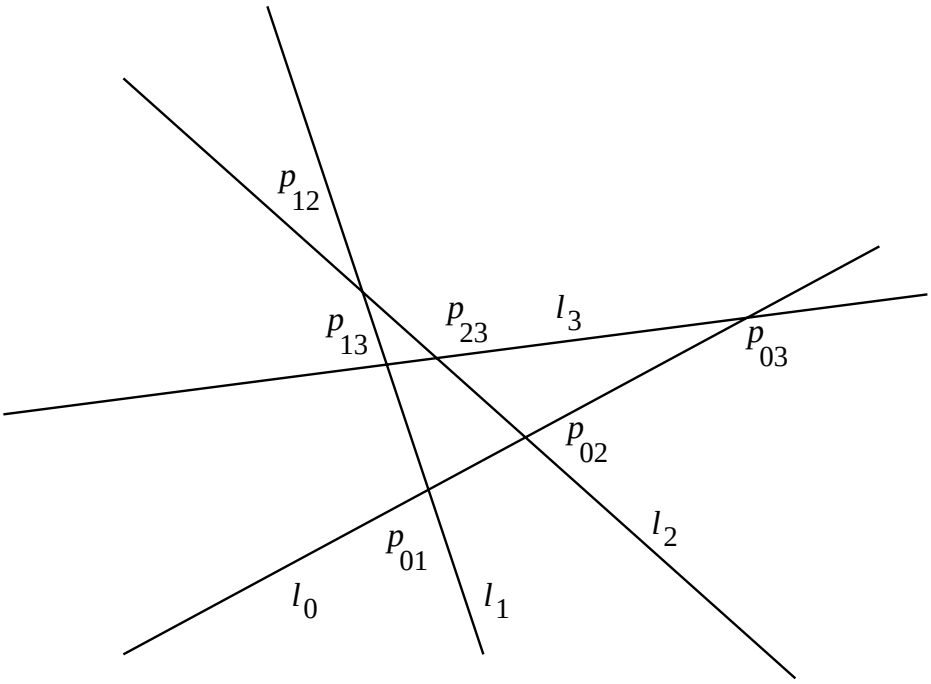


FIGURE 3. Four general lines in the plane

4.1. Cubic surfaces revisited: the Cayley cubic. In §1.1, we constructed smooth cubic surfaces by blowing up six points in general position on the plane. What happens when we relax this assumption?

Consider configurations of six points obtained as pairwise intersections of four general lines in the plane. Given four lines in general position, we can choose coordinates to put them in the standard form:

$$\ell_0 = \{x_0 = 0\},\ \ell_1 = \{x_1 = 0\},\ \ell_2 = \{x_2 = 0\},\ \ell_3 = \{x_0 + x_1 + x_2 = 0\}.$$

The intersection points are denoted $p_{ij} = \ell_i \cap \ell_j$ for $0 \le i < j \le 3$.

The points $p_{01}, \dots, p_{23}$ still impose independent conditions on homogeneous cubics in $x_0, \dots, x_3$, i.e.,

$$I_{p_{01},\dots,p_{23}} = \langle y_0, y_1, y_2, y_3 \rangle$$

where

$$\begin{array}{ll} y_0 = x_1x_2(x_0 + x_1 + x_2) & y_1 = x_0x_2(x_0 + x_1 + x_2) \\ y_2 = x_0x_1(x_0 + x_1 + x_2) & y_3 = -x_0x_1x_2 \end{array}.$$

These satisfy the relation

$$y_0y_1y_2 + y_1y_2y_3 + y_2y_3y_0 + y_3y_0y_1 = 0;$$

the resulting cubic surface $S \subset \mathbb{P}^3$ is called the *Cayley cubic surface* in honor of Arthur Cayley, who classified singular cubic surfaces [**Cay69**].

Here are some of its geometric properties:

- S has ordinary double points at

$$s_0 = [1, 0, 0, 0],\ s_1 = [0, 1, 0, 0],\ s_2 = [0, 0, 1, 0],\ s_3 = [0, 0, 0, 1].$$

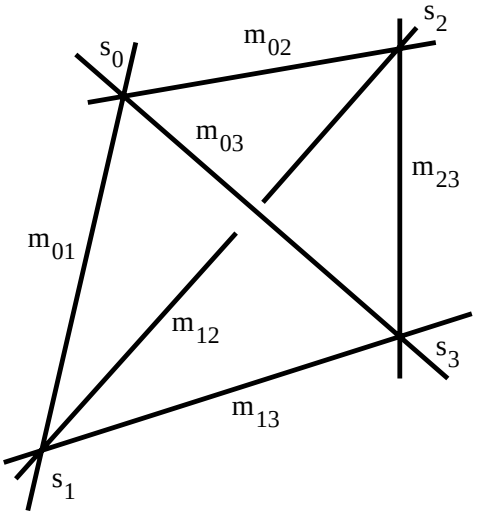


FIGURE 4. Some lines on the Cayley cubic

It is the unique cubic surface with this configuration of singularities, up to projective equivalence.

- S contains nine lines, i.e., the lines $m_{ij}, i, j = 0, \ldots, 3$ spanned by s_i and s_j, as well as the lines

$$\{y_0 + y_3 = y_1 + y_2 = 0\}, \quad \{y_0 + y_1 = y_2 + y_3 = 0\}, \quad \{y_0 + y_2 = y_1 + y_3 = 0\}.$$

- The birational map

$$[y_0, y_1, y_2, y_3] : \mathbb{P}^2 \xrightarrow{\sim} S$$

factors as

$$\begin{array}{ccc} & X & \\ {\scriptstyle\beta}\swarrow & & \searrow{\scriptstyle\sigma} \\ \mathbb{P}^2 & & S \end{array}$$

where β is the blow-up of $p_{01}, \ldots, p_{23}$ and σ is the blow-up of $s_0, \ldots, s_3$. The exceptional divisors of β are the proper transforms E_{ij} of the m_{ij}; the exceptional divisors of σ are the proper transforms ℓ_i' of the ℓ_i.

- Express

$$\mathrm{Pic}(X) = \mathbb{Z}L \oplus \mathbb{Z}E_{01} \oplus \cdots \oplus \mathbb{Z}E_{23}$$

where L is the pullback of the hyperplane class of $\mathbb{P}^2$ via β. The canonical class is

$$K_X = -3L + E_{01} + E_{02} + E_{03} + E_{12} + E_{13} + E_{23}$$

and the proper transforms of the lines are

$$\ell_0' = L - E_{01} - E_{02} - E_{03}, \ \ell_1' = L - E_{01} - E_{12} - E_{13}, \ldots$$

We have $K_X \cdot \ell_j' = 0$ and $(\ell_j')^2 = -2$ for each j, i.e., the exceptional divisors of the resolution σ are (-2)-curves.

4.2. Why consider singular cubic surfaces?

Reason 1: Reduction modulo primes

Let $X = \{F(y_0, y_1, y_2, y_3) = 0\} \subset \mathbb{P}^3$ be a smooth cubic surface defined over $\mathbb{Q}$; we may assume that $F \in \mathbb{Z}[y_0, y_1, y_2, y_3]$ and the greatest common divisor of the coefficients of F is 1. Consider the integral model

$$\pi : \mathcal{X} = \{F = 0\} \subset \mathbb{P}^3_{\mathbb{Z}} \to \mathrm{Spec}(\mathbb{Z}),$$

which is flat and projective over $\mathrm{Spec}(\mathbb{Z})$. For each prime p, we have $\mathcal{X}_p = \mathcal{X} \pmod p$, i.e., the fiber of $\mathcal{X}$ over $p \in \mathrm{Spec}(\mathbb{Z})$. If p divides the discriminant of F then $\mathcal{X}_p$ will have singularities. These singular fibers have a strong influence on the rational points of X.

Reason 1′: Degenerate fibers of families

This is the function-field analog of the previous situation. Let B be a complex curve and

$$\pi : \mathcal{X} \to B$$

a family of cubic surfaces, e.g., a pencil

$$\{sF(y_0, y_1, y_2, y_3) + tG(y_0, y_1, y_2, y_3) = 0\} \subset \mathbb{P}^3_{y_0,y_1,y_2,y_3} \times \mathbb{P}^1_{s,t}$$

with π being projection onto the second factor. At least some of the fibers $\mathcal{X}_b = \pi^{-1}(b), b \in B$ must be singular.

Reason 2: Counting rational points

Proving asymptotics for the number of rational points of bounded heights on *singular* cubic surfaces is often easier than the case of smooth cubic surfaces. Examples include toric cubic surfaces [**dlB98, Fou98, HBM99, Sal98**]

$$y_0^3 = y_1 y_2 y_3,$$

the Cayley cubic surface [**HB03**], the '$\mathbf{E}_6$ cubic surface' [**dlBBD07**]

$$y_1 y_2^2 + y_2 y_0^2 + y_3^3 = 0;$$

and a '$\mathbf{D}_4$ cubic surface' [**Bro06**]

$$y_1 y_2 y_3 = y_4 (y_1 + y_2 + y_3)^2.$$

4.3. What are 'good' singularities? Let S be a normal surface. A *resolution of singularities* $\sigma : X \to S$ is a birational proper morphism from a smooth surface. Abhyankar proved the existence of resolutions of surface singularities in arbitrary characteristic [**Abh56**]. A resolution $\sigma : X \to S$ is *minimal* if there exists no nontrivial factorization

$$X \stackrel{\phi}{\to} Y \to S$$

with Y smooth. This is equivalent to

- there are no (-1)-curves in the fibers of σ; or
- K_X is nef relative to σ.

A relative analog of Corollary 2.20 (see Remark 2.21) implies that minimal resolutions of singularities are unique, in the case of surfaces.

Recall that if $\phi : X \to Y$ is a birational morphism of smooth projective surfaces then (cf. Equation 1.1):

$$K_X = \phi^* K_Y + \sum_i m_i E_i, \quad m_i > 0.$$

The following definition represents a weakening of this condition:

DEFINITION 4.1. Suppose that S is a normal surface with a unique singularity p; assume that K_S is a $\mathbb{Q}$-Cartier divisor at p. (This is the case when S is a complete intersection in some neighborhood of p). Then $p \in S$ is a *canonical singularity* if, for each resolution of singularities $\sigma : X \to S$ we have

$$K_X = \sigma^* K_S + \sum_i m_i E_i, \quad m_i \geq 0,$$

where the E_i are exceptional divisors of σ.

Note here that *a priori* the $m_i \in \mathbb{Q}$; however, the classification of these singularities shows *a posteriori* that the $m_i \in \mathbb{Z}$.

PROPOSITION 4.2. *Suppose that (S, p) is a canonical singularity. A resolution of singularities $\sigma : X \to S$ is minimal if and only if each $m_i = 0$, i.e., $K_X = \sigma^* K_S$. In this case, each σ-exceptional curve is a (-2)-curve, i.e., a nonsingular rational curve E with $E^2 = -2$.*

PROOF. ($\Leftarrow$) Suppose that $m_i = 0$ for each i. Then $K_X \cdot E_i = 0$ for each σ-exceptional divisor. The Hodge index theorem implies that the σ-exceptional divisors have negative self-intersection, i.e., $E_i^2 < 0$. The adjunction formula implies that $E_i^2 = -2$ and E_i is a nonsingular curve of arithmetic genus zero.

($\Rightarrow$) Assume that σ is minimal, i.e., the fibers of σ contain no (-1)-curves. Suppose that $K_X \neq \sigma^* K_S$ so that some $m_i \neq 0$. It follows that $(\sum_i m_i E_i)^2 < 0$ and thus $(\sum_i m_i E_i) \cdot E_j < 0$ for some σ-exceptional curve E_j. Consequently $E_j^2 < 0$ and $K_X E_j < 0$, so E_j is a (-1)-curve by Proposition 1.6. □

Proposition 4.2 suggests the following variation on this definition

DEFINITION 4.3. A normal surface S has *Du Val singularities* if it admits a resolution $\sigma : X \to S$ such that $K_X \cdot E = 0$ for each σ-exceptional divisor E.

Patrick Du Val first classified surface singularities in terms of their discrepancies (or in his terminology, the 'conditions they impose on adjunction') in [**DV34**]. This definition is *a priori* more general than the class of canonical singularities: We do not insist that K_S is $\mathbb{Q}$-Cartier. However, we shall see later (Remark 4.7) that Du Val singularities are canonical.

Exercises.

EXERCISE 4.3.1. We give an example of a surface with 'bad' singularities. Suppose that $p_1, \ldots, p_4 \in \ell \subset \mathbb{P}^2$ are distinct points lying on a line ℓ. Consider

$$\beta : X := \mathrm{Bl}_{p_1,\ldots,p_4} \mathbb{P}^2 \to \mathbb{P}^2$$

and let $\tilde{\ell}$ denote the proper transform of ℓ, L the pullback of the hyperplane class via β, and $E_1, \ldots, E_4$ the exceptional divisors. Verify that

a. The divisor $4L - E_1 - E_2 - E_3 - E_4$ is basepoint-free and yields a morphism $\phi : X \to \mathbb{P}^{10}$.
b. If Y is the image of X under ϕ, show that $\phi : X \to Y$ is an isomorphism over $X \setminus \tilde{\ell}$ and contracts $\tilde{\ell}$ to a point $y \in Y$.
c. Show that Y is normal at y and the canonical class K_Y is $\mathbb{Q}$-Cartier. Compute the divisor $\phi^* K_Y$.
e. Show that $y \in Y$ is not a Du Val singularity.

4.4. Singular Del Pezzo surfaces.

DEFINITION 4.4. A *singular Del Pezzo surface* is a projective surface S with Du Val singularities such that $-K_S$ is ample.

If $\sigma : X \to S$ is a minimal resolution of a singular Del Pezzo surface then $\sigma^* K_S = K_X$, i.e., $-K_X$ is semiample.

Here is one good source of singular Del Pezzo surfaces. Suppose X is a smooth projective surface with $-K_X$ nef and big. It has the following properties:

- $(-K_X)^2 > 0$;
- Any irreducible curve E with $K_X \cdot E = 0$ is a (-2)-curve.
- There are a finite number of (-2)-curves on X.

The first statement is a particular case of Corollary 2.4. The second is contained in the proof of Proposition 4.2. The third follows from the fact that $K_X^{\perp}$ is negative definite, and thus has a finite number of vectors of self-intersection -2.

THEOREM 4.5. *Let X be a smooth projective surface with $-K_X$ nef and big. Then each nef divisor D on X is semiample.*

COROLLARY 4.6. *Let X be a smooth projective surface with $-K_X$ nef and big. Then $-K_X$ is semiample. In particular, there exists a birational morphism $\sigma : X \to S$ to a singular Del Pezzo surface with $\sigma^* K_S = K_X$.*

PROOF. Remark 2.18 addresses this in the special case where D is not big, i.e., when $D^2 = 0$. Thus we may assume that $D^2 > 0$.

The Nakai criterion (Theorem 2.2) implies that D is ample unless $D \cdot E = 0$ for some irreducible curve $E \subset X$. The Hodge index theorem implies that each such curve satisfies $E^2 < 0$. Since $-K_X \cdot E \geq 0$, the only possibilities are (-1)-curves (see Proposition 1.6) or (-2)-curves (see Proposition 4.2). In either case $E \simeq \mathbb{P}^1$.

Suppose X admits (-1)-curves as above. We can apply the Castelnuovo contraction criterion (Theorem 1.12) to obtain a birational morphism $\beta : X \to Y$ such that Y admits a big and nef divisor M on Y with $\beta^* M = D$ and the only curves orthogonal to M are (-2)-curves. Furthermore, $-K_Y$ remains nef and big (cf. Corollary 2.8).

Let $E_1, \ldots, E_r$ denote the (-2)-curves orthogonal to M. We exhibit a birational morphism to a singular projective variety $\sigma : Y \to S$ contracting precisely these curves. Such a contraction exists for more general reasons [**Rei97**, 4.15] [**Art62**, 2.3] but we will sketch an argument in our situation.

We essentially copy the proof of the Castelnuovo Criterion. Let H be a very ample line bundle on Y such that each positive multiple nH has no higher cohomology. Write $d_i = H \cdot E_i$ for $i = 1, \ldots, r$. Since the intersection matrix of $\mathbb{Z}E_1 + \cdots + \mathbb{Z}E_r$ is negative definite, there exist positive integers n and $b_1, \ldots, b_r$ such that

$$nH \cdot E_i = -(b_1 E_1 + \cdots + b_r E_r) \cdot E_i$$

for each i. Let $B = b_1 E_1 + \cdots + b_r E_r$ so that $L := nH + B$ is orthogonal to each E_i.

The adjunction formula implies that each effective divisor A supported on $E_1 \cup \cdots \cup E_r$ has nonpositive arithmetic genus; a straightforward induction gives $H^1(\mathcal{O}_A) = 0$ as well. Here it is crucial that $K_Y \cdot E_i = 0$ for each i; it is not enough to assume that each component of the exceptional locus is rational. Thus we have

$\mathcal{O}_Y(L)|B \simeq \mathcal{O}_B$, i.e., an isomorphism of invertible sheaves, not just an equality of degrees. We obtain the exact sequence

$$0 \to \mathcal{O}_Y(nH) \to \mathcal{O}_Y(L) \to \mathcal{O}_B \to 0.$$

Our vanishing assumption show that

$$\Gamma(Y, \mathcal{O}_Y(L)) \twoheadrightarrow \Gamma(Y, \mathcal{O}_B),$$

i.e., for each point of B there is a section of $\mathcal{O}_Y(L)$ nonvanishing at that point. The sections of $\mathcal{O}_Y(L)$ induce an embedding away from $E_1 \cup \cdots \cup E_r$, so $\mathcal{O}_Y(L)$ is globally generated and induces a morphism $\sigma : Y \to S$ contracting precisely $E_1, \ldots, E_r$. In particular, $\mathcal{O}_Y(L)$ is the pullback of an ample line bundle on S via σ.

To complete the argument, we show there exists a Cartier divisor N on S such that $\sigma^* N$ is a positive multiple of M. Repeating the previous argument for $M + mL$ with $m \gg 0$, we get the same contraction $\sigma : Y \to S$. Here the argument shows that $mL + M$ is the pullback of an ample line bundle from S. It follows that $M = \sigma^* N$ for some Cartier divisor N on X.

Finally, N is ample on S by the Nakai criterion, as we have contracted all the curves along which it is nonpositive. □

Remark 4.7. A variation on this argument shows that Du Val singularities are canonical. Suppose that $\sigma : Y \to S$ is a minimal resolution of Du Val singularities. The canonical class K_Y is nef relative to σ and thus globally generated relative to σ. We obtain a factorization

$$Y \to \operatorname{Proj}_S \left(\bigoplus_{n \geq 0} \sigma_* \mathcal{O}_Y(nK_Y) \right) \xrightarrow{\varpi} S.$$

Since ϖ is a bijective morphism of normal surfaces, it is an isomorphism. However, the canonical divisor of the intermediate surface is $\mathbb{Q}$-Cartier by construction.

Remark 4.8. Suppose the base field is algebraically closed of characteristic zero. There do exist smooth projective rational surfaces admitting nef divisors that are not semiample [**Zar62**, §2]. Thus the assumption that $-K_X$ be nef and big in Theorem 4.5 is necessary. (See Exercise 4.4.1 below and [**Laz04**, 2.3] for more discussion.)

We record one last consequence of Theorem 4.5, an extension of Corollary 2.13:

Proposition 4.9. *Let X be a smooth projective surface with $-K_X$ nef and big. Then $\overline{\mathrm{NE}}_1(X)$ is a finite rational polyhedral cone, generated by (-2)-curves and K_X-negative extremal rational curves.*

Proof. Apply Proposition 2.10 and Corollary 2.11: $\overline{\mathrm{NE}}_1(X)$ is generated by the nonnegative cone $\overline{\mathcal{C}}$, along with the (-1)-curves and (-2)-curves. The Hodge index theorem implies that K_X is negative on $\overline{\mathcal{C}} \setminus \{0\}$, so any extremal rays of $\overline{\mathrm{NE}}_1(X)$ arising from $\overline{\mathcal{C}}$ are necessarily K_X-negative.

The Cone Theorem 2.23 implies that the K_X-negative part of the effective cone is generated by curves L_i with $-K_X \cdot L_i \leq 3$. Theorem 4.5 gives that $-K_X$ is semiample and induces $\sigma : X \to S$. Thus there are at most a finite number of classes $[L_i]$ arising as K_X-negative extremal rays. Indeed, the curves in S with anticanonical degree ≤ 3 are parametrized by a scheme of finite type, as the curves

in projective space of bounded degree are parameterized by a Hilbert scheme of finite type. We see in particular that X admits a finite number of (-1)-curves.

Clearly, there are a finite number of (-2)-curves, as these are all σ-exceptional. Thus $\overline{\mathrm{NE}}_1(X)$ admits a finite number of extremal rays, with the desired interpretations. □

Exercises.

EXERCISE 4.4.1. Assume the base field is of characteristic zero.

Let $C \subset \mathbb{P}^2$ denote a smooth cubic plane curve and H the hyperplane class on $\mathbb{P}^2$. Choose points $p_1, \ldots, p_9 \in C$ such that the divisors $p_1 + \cdots + p_9$ and $H|C$ are linearly independent over $\mathbb{Q}$. Consider the blow-up

$$X := \mathrm{Bl}_{p_1,\ldots,p_9}\mathbb{P}^2 \xrightarrow{\beta} \mathbb{P}^2$$

with exceptional curves $E_1, \ldots, E_9$. Show that $D = -K_X = 3\beta^*H - E_1 - \cdots - E_9$ is nef but not semiample.

Now choose points $q_1, \ldots, q_{12} \in C$ such that $q_1 + \cdots + q_{12}$ and $H|C$ are linearly independent. Consider the blow-up

$$Y := \mathrm{Bl}_{q_1,\ldots,q_{12}}\mathbb{P}^2 \xrightarrow{\gamma} \mathbb{P}^2$$

with exceptional curves $F_1, \ldots, F_{12}$. Show that $D' = 4\gamma^*H - F_1 - \cdots - F_{12}$ is nef but not semiample. Indeed, demonstrate that for each $n > 0$ the divisor nD' has the proper transform of C as a fixed component.

4.5. Classification of Du Val singularities. Suppose that $\sigma : X \to S$ is a minimal resolution of a Du Val surface singularity $p \in S$. Consider the intersection numbers of the irreducible components $E_1, \ldots, E_r$ of $\sigma^{-1}(p)$, which we put into a symmetric matrix $(E_i \cdot E_j)_{i,j=1,\ldots,r}$. This has the following properties:

- $(E_i \cdot E_j)$ is negative definite, by the Hodge index theorem;
- $E_i^2 = -2$ for each i, by Proposition 4.2;
- $E_i \cdot E_j = 0, 1$ for each $i \neq j$; indeed, if $E_i \cdot E_j > 1$ then $(E_i + E_j)^2 > 0$;
- we cannot express

$$\{E_1, \ldots, E_r\} = \{E_{a_1}, \ldots, E_{a_s}\} \cup \{E_{b_1}, \ldots, E_{b_{r-s}}\}$$

with $E_{a_l} \cdot E_{b_m} = 0$ for each l, m; this is because $\sigma^{-1}(p)$ is connected.

Matrices of this type occur throughout mathematics, especially in the classification of the simple root systems via Dynkin diagrams/Cartan matrices in Lie theory. We cannot dwell too much on these interactions, except to refer the reader to some of the literature on this beautiful theory [**Bri71, Dur79, SB01**]. We list the possible matrices that can arise [**FH91**, 21.2]. First, we have the infinite series

$$\begin{matrix}\mathbf{A}_r \\ r \geq 1\end{matrix} \quad \begin{pmatrix} -2 & 1 & 0 & \ldots & \ldots & 0 \\ 1 & -2 & 1 & \ddots & \ddots & \vdots \\ 0 & 1 & -2 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 1 & 0 \\ \vdots & \ddots & \ddots & 1 & -2 & 1 \\ 0 & \ldots & \ldots & 0 & 1 & -2 \end{pmatrix} \quad E_i \cdot E_i = \begin{cases} -2 & \text{if } i = j \\ 1 & \text{if } |i-j| = 1 \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{D}_r \quad r \geq 4 \quad \begin{pmatrix} -2 & 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & -2 & 1 & 0 & \ddots & \ddots & \vdots \\ 1 & 1 & -2 & 1 & \ddots & \ddots & \vdots \\ 0 & 0 & 1 & -2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & 1 & 0 \\ 0 & \ddots & \ddots & \ddots & 1 & -2 & 1 \\ 0 & \dots & \dots & 0 & 0 & 1 & -2 \end{pmatrix} \qquad E_i \cdot E_i = \begin{cases} -2 & \text{if } i = j \\ 1 & \text{if } |i-j| = 1, \\ & \quad i, j \geq 3 \\ & \text{or if } \{i,j\} \\ & = \{1,3\}, \{2,3\} \\ 0 & \text{otherwise} \end{cases}$$

and then the exceptional lattices

$$\mathbf{E}_6 \quad \begin{pmatrix} -2 & 1 & 0 & 0 & 0 & 0 \\ 1 & -2 & 0 & 0 & 0 & 1 \\ 0 & 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 0 & 1 \\ 0 & 0 & 0 & 0 & -2 & 1 \\ 0 & 1 & 0 & 1 & 1 & -2 \end{pmatrix}$$

$$\mathbf{E}_7 \quad \begin{pmatrix} -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & -2 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & -2 \end{pmatrix}$$

$$\mathbf{E}_8 \quad \begin{pmatrix} -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -2 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & -2 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & -2 \end{pmatrix}$$

Remarkably, in characteristic zero there is a *unique* singularity associated to each of these matrices.

PROPOSITION 4.10. *Assume that the base field is algebraically closed of characteristic zero. Then, up to analytic isomorphism, there is a unique Du Val surface singularity associated to each Cartan matrix enumerated above:*

$\mathbf{A}_r, r \geq 1$	$z^2 = x^2 + y^{r+1}$
$\mathbf{D}_r, r \geq 4$	$z^2 = y(x^2 + y^{r-2})$
$\mathbf{E}_6$	$z^2 = x^3 + y^4$
$\mathbf{E}_7$	$z^2 = y(x^3 + y^2)$
$\mathbf{E}_8$	$z^2 = x^3 + y^5$

For a modern proof of this, we refer the reader to [**KM98**, §4.2]. It turns out that these singularities are also related to the class of 'simple' hypersurface singularities, which can be independently classified [**AGZV85**]. There are a multitude of classical characterizations of Du Val singularities [**Dur79**].

EXAMPLE 4.11. The Cayley cubic surface has four $\mathbf{A}_1$ singularities. The toric cubic surface

$$y_0^3 = y_1 y_2 y_3$$

has three $\mathbf{A}_2$ singularities at $[0,1,0,0]$,$[0,0,1,0]$, and $[0,0,0,1]$.

5. Cox rings and universal torsors

We work over an algebraically closed field k unless specified otherwise.

5.1. Universal torsors. Universal torsors are an important tool in higher-dimensional arithmetic geometry. They play a fundamental rôle in the modern theory of descent for rational varieties [**CTS87**]. They are also an important technique and conceptual tool for counting rational points of bounded height [**Sal98**].

Let X be a smooth projective variety. Assume that $\mathrm{Pic}(X)$ is a free abelian group of rank r, generated by the line bundles $L_1, \dots, L_r$ on X. Let $T_X = \mathrm{Hom}(\mathrm{Pic}(X), \mathbb{G}_m)$ denote the *Néron-Severi torus* of X, i.e., the torus with character group $\mathrm{Hom}(T_X, \mathbb{G}_m) = \mathrm{Pic}(X)$.

DEFINITION 5.1. [**CTS87**] The *universal torsor* over X

$$\begin{array}{ccc} T_X & \to & U \\ & & \downarrow \\ & & X \end{array}$$

is a principal homogeneous space over X with structure group T_X with the following universal property: Given a line bundle L on X, if $\lambda_L : T_X \to \mathbb{G}_m = \mathrm{GL}_1$ denotes the corresponding character then the line bundle V_{λ_L} associated to U equals L. In other words, if U is given by a cocycle $\{\tau_{ij}\} \in H^1(X, T_X)$ then L is given by the cocycle $\{\lambda_L(\tau_{ij})\} \in H^1(X, \mathbb{G}_m)$.

Constructing U is straightforward in some sense: Choose $L_1, \dots, L_r$ freely generating $\mathrm{Pic}(X)$ and write

$$P_i = L_i^{-1} \setminus \mathbf{0}_X \subset L_i^{-1}$$

for the complement of the zero-section. This is a $\mathbb{G}_m$-principal bundle arising from L_i^{-1}. Then we can take

$$U = P_1 \times_X \cdots \times_X P_r$$

and T_X-action

$$\begin{array}{ccc} T_X \times U & \to & U \\ (t; s_1, \dots, s_r) & \mapsto & (\lambda_{-L_1}(t)s_1, \dots, \lambda_{-L_r}(t)s_r) \end{array}$$

where s_i is a local section of P_i and λ_L is the character associated with L.

However, for arithmetic applications it is important to have a more concrete presentation of the universal torsor.

EXAMPLE 5.2. Consider the case $X = \mathbb{P}^n$. The standard quotient presentation

$$\mathbb{P}^n = (\mathbb{A}^{n+1} \setminus 0)/\mathbb{G}_m$$

can be interpreted as an identification:

$$\begin{array}{ccccc} \mathcal{O}_{\mathbb{P}^n}(-1) \setminus \mathbf{0}_{\mathbb{P}^n} & & \stackrel{\sim}{\to} & & \mathbb{A}^{n+1} \setminus 0 \\ & \searrow & & \swarrow & \\ & & \mathbb{P}^n & & \end{array}$$

In other words, we regard $\mathcal{O}_{\mathbb{P}^n}(-1)$ as the 'universal line' over $\mathbb{P}^n$. Since $\mathcal{O}_{\mathbb{P}^n}(-1)$ generates $\mathrm{Pic}(\mathbb{P}^n)$, we have

$$U = \mathcal{O}_{\mathbb{P}^n}(-1) \setminus \mathbf{0}_{\mathbb{P}^n} = \mathbb{A}^{n+1} \setminus 0,$$

equivariant with respect to the action of $\mathbb{G}_m = T_{\mathbb{P}^n}$. Note that we can regard

$$\mathbb{A}^{n+1} = \mathrm{Spec}\left(\bigoplus_{N\in\mathbb{Z}} \Gamma(\mathbb{P}^n, \mathcal{O}_{\mathbb{P}^n}(N))\right).$$

More generally, the universal torsor

$$\begin{array}{ccc} T_{\mathbb{P}^m\times\mathbb{P}^n} & \to & U \\ & & \downarrow \\ & & \mathbb{P}^m \times \mathbb{P}^n \end{array}$$

can be identified with

$$\mathbb{A}^{m+n+2}_{x_0,\dots,x_m,y_0,\dots,y_n} \setminus (\{x_0 = \cdots = x_m = 0\} \cup \{y_0 = \cdots = y_n = 0\}).$$

Here the torus acts by the rule

$$(t_1, t_2) \cdot (x_0, \dots, x_m, y_0, \dots, y_n) = (t_1x_0, \dots, t_1x_m, t_2y_0, \dots, t_2y_n).$$

Decomposing the polynomial ring under this action, we can regard

$$\mathbb{A}^{m+n+2} = \mathrm{Spec}\left(\bigoplus_{N_1,N_2\in\mathbb{Z}} \Gamma(\mathbb{P}^m \times \mathbb{P}^n, \mathcal{O}_{\mathbb{P}^m\times\mathbb{P}^n}(N_1, N_2))\right).$$

Exercises.

Exercise 5.1.1. Realize the universal torsor over $X = \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1$ as an explicit open subset $U \subset \mathbb{A}^8$. Describe the action of T_X on U.

5.2. Universal torsors over nonclosed fields. We can only offer a brief summary here; we refer the reader to [**CTS87**] and [**Sko01**] for details and arithmetic applications.

Let k be a perfect field with absolute Galois group G. Suppose that X is defined over k and $\bar{X}$ satisfies the assumptions made in §5.1. The Galois action on $\mathrm{Pic}(\bar{X})$ allows us to define the torus T_X over k. Precisely, the action induces a representation on the character group

$$\varrho : G \to \mathrm{Aut}(\mathrm{Hom}(T_{\bar{X}}, \mathbb{G}_m)) = \mathrm{Aut}(\mathrm{Pic}(\bar{X})),$$

which gives the descent data for T_X. A *universal torsor* over X is a principal homogeneous space

$$\begin{array}{ccc} T_X & \to & U \\ & & \downarrow \\ & & X \end{array}$$

defined over k, such that the universal property is satisfied on passage to the algebraic closure.

Note our use of the indefinite article: Over a nonclosed field, a variety may have more than one universal torsor. Indeed, since any universal torsor U comes with a T_X-action over X, given a cocycle $\eta \in H^1_G(T_{\bar{X}})$ we can twist to obtain

$$\begin{array}{ccc} T_X & \to & U^\eta \\ & & \downarrow \\ & & X, \end{array}$$

another universal torsor over k. However, if the Galois action on $\mathrm{Pic}(\bar{X})$ is trivial then

$$H^1_G(T_{\bar{X}}) = H^1_G(\mathbb{G}_m^r) = 0$$

by Hilbert's Theorem 90. Here the universal torsor is unique whenever it exists.

On the other hand, there may be obstructions to descending a universal torsor over $\bar{X}$ to the field k. These reside in $H^2_G(T_{\bar{X}})$; indeed, the situation is analogous to the descent obstruction for line bundles discussed in Remark 3.3. Whenever $X(k) \neq \emptyset$ this obstruction vanishes [**CTS87**, 2.2.8], which makes universal torsors an important tool for deciding whether X has rational points.

5.3. Cox rings. Let X be a normal projective variety such that the Weil divisor class group is freely generated by $D_1, \dots, D_r$.

Definition 5.3. The *Cox ring* of X is defined as

$$\mathrm{Cox}(X) = \bigoplus_{(n_1,\dots,n_r)\in\mathbb{Z}^r} \Gamma(X, \mathcal{O}_X(n_1D_1 + \cdots + n_rD_r))$$

with multiplication

$$\begin{aligned}\Gamma(X, \mathcal{O}_X(m_1D_1 + \cdots + m_rD_r)) \times \Gamma(X, \mathcal{O}_X(n_1D_1 + \cdots + n_rD_r)) \quad &\to \\ \Gamma(X, \mathcal{O}_X((m_1+n_1)D_1 + \cdots + (m_r+n_r)D_r))&\end{aligned}$$

defined by $(s,t) \mapsto st$.

Example 5.4. We start with the eponymous example [**Cox95**]: Let X be a projective toric variety of dimension n

$$\mathbb{G}_m^n \times X \to X.$$

Let $D_1, \dots, D_d$ denote the boundary divisors, i.e., the irreducible components of the complement of the dense open torus orbit. Let $s_i \in \Gamma(X, \mathcal{O}_X(D_i))$ denote the canonical section, i.e., the one associated with the inclusion

$$\mathcal{O}_X \hookrightarrow \mathcal{O}_X(D_i).$$

(Actually, s_i is canonical up to a nonzero scalar.) Recall that

- each effective divisor D on X can be expressed as a nonnegative linear combination

$$D \equiv n_1D_1 + \cdots + n_dD_d, \quad n_1, \dots, n_d \geq 0;$$

- the canonical section s of $\mathcal{O}_X(D)$ admits a unique expression

$$s = f(s_1, \dots, s_d)$$

where f is a polynomial over k in d variables.

Thus the Cox ring of X is a polynomial ring

$$\mathrm{Cox}(X) \simeq k[s_1, \dots, s_d]$$

with generators indexed by the boundary divisors.

We list some basic properties of the Cox ring of a *smooth* projective variety. We continue to assume that $D_1, \dots, D_r$ are divisors freely generating $\mathrm{Pic}(X)$.

- $\mathrm{Cox}(X)$ is graded by $\mathrm{Pic}(X)$, i.e.,
$$\mathrm{Cox}(X) \simeq \bigoplus_{\mathcal{L}\in \mathrm{Pic}(X)} \mathrm{Cox}(X)_{\mathcal{L}}, \quad \mathrm{Cox}(X)_{\mathcal{L}} \simeq \Gamma(X, \mathcal{L}).$$
Indeed, for a unique choice of $(n_1, \dots, n_r) \in \mathbb{Z}^r$ we have an isomorphism
$$\mathcal{L} \simeq \mathcal{O}_X(n_1 D_1 + \cdots + n_r D_r).$$
- $\mathrm{Cox}(X)$ has a natural action by T_X via the rule
$$(t_1, \dots, t_r) \cdot s = t_1^{n_1} \cdots t_r^{n_r} s$$
when $s \in \Gamma(X, \mathcal{O}_X(n_1 D_1 + \cdots + n_r D_r))$.
- The nonzero graded pieces of $\mathrm{Cox}(X)$ are indexed by $\mathrm{NE}^1(X, \mathbb{Z})$. If $\mathrm{Cox}(X)$ is finitely generated then $\overline{\mathrm{NE}}^1(X)$ is a finitely generated rational polyhedral cone.

5.4. Two theorems. We start with a general result:

PROPOSITION 5.5. *Let X be a projective variety and $A_1, \dots, A_r$ semiample Cartier divisors on X. Then the ring*

$$\bigoplus_{n_1,\dots,n_r \geq 0} \Gamma(\mathcal{O}_X(n_1 A_1 + \cdots + n_r A_r)) \tag{5.1}$$

is finitely generated.

PROOF. (based on [**HK00**, 2.8], with suggestions from A. Várilly-Alvarado) It suffices to show that for some positive $N \in \mathbb{N}$ the ring

$$\bigoplus_{n_1,\dots,n_r \geq 0} \Gamma(\mathcal{O}_X(N(n_1 A_1 + \cdots + n_r A_r)))$$

is finitely generated. Indeed, the full ring is integral over this subring, so our result follows by finiteness of integral closure. Since $A_1, \dots, A_r$ are semiample, there exists an $N > 0$ such that $NA_1, \dots, NA_r$ are globally generated. Thus we may assume that $A_1, \dots, A_r$ are globally generated.

We first consider the special case $r = 1$. We obtain a morphism

$$\phi : X \to \mathbb{P}^m := \mathbb{P}(\Gamma(\mathcal{O}_X(A_1))^*),$$

with $\phi^* \mathcal{O}_{\mathbb{P}^m}(1) = \mathcal{O}_X(A_1)$. This admits a Stein factorization

$$X \xrightarrow{f} Y \xrightarrow{g} \mathbb{P}^m$$

with g finite and f having connected fibers, so in particular $f_* \mathcal{O}_X = \mathcal{O}_Y$. Furthermore, $g^* \mathcal{O}_{\mathbb{P}^m}(1)$ is ample on Y and thus

$$\bigoplus_{n \geq 0} \Gamma(Y, g^* \mathcal{O}_{\mathbb{P}^m}(n))$$

is finitely generated. The projection formula gives

$$g^* \Gamma(Y, g^* \mathcal{O}_{\mathbb{P}^m}(n)) \xrightarrow{\sim} \Gamma(X, \mathcal{O}_X(nA_1))$$

for each $n \in \mathbb{N}$, so

$$\bigoplus_{n \geq 0} \Gamma(X, \mathcal{O}_X(nA_1))$$

is also finitely generated.

Now suppose r is arbitrary. Consider the vector bundle

$$V = A_1 \oplus \cdots \oplus A_r$$

and the associated projective bundle

$$\pi : \mathbb{P}(V^*) \to X.$$

We have the tautological quotient bundle

$$\pi^* V \twoheadrightarrow \mathcal{O}_{\mathbb{P}(V^*)}(1);$$

since V is globally generated (being the direct sum of globally generated line bundles), $\mathcal{O}_{\mathbb{P}(V^*)}(1)$ is semiample. In particular, the ring

$$\bigoplus_{n \geq 0} \Gamma(\mathbb{P}(V^*), \mathcal{O}_{\mathbb{P}(V^*)}(n)) \tag{5.2}$$

is finitely generated.

The tautological quotient induces

$$\mathrm{Sym}^n \pi^* V \twoheadrightarrow \mathcal{O}_{\mathbb{P}(V^*)}(n),$$

and taking direct images via the projection formula we obtain

$$\mathrm{Sym}^n V \xrightarrow{\sim} \pi_* \mathcal{O}_{\mathbb{P}(V^*)}(n)$$

and

$$\Gamma(X, \mathrm{Sym}^n V) = \Gamma(\mathbb{P}(V^*), \mathcal{O}_{\mathbb{P}(V^*)}(n)).$$

Since (5.2) is finitely generated, the algebra

$$\bigoplus_{n \geq 0} \Gamma(X, \mathrm{Sym}^n V)$$

is finitely generated as well. Using the decomposition

$$\mathrm{Sym}^n V = \bigoplus_{\substack{n_1 + \cdots + n_r = n \\ n_1, \ldots, n_r \geq 0}} \mathcal{O}_X(n_1 A_1 + \cdots + n_r A_r),$$

we conclude that (5.1) is finitely generated. □

Remark 5.6 (due to A. Várilly-Alvarado). If A_1 and A_2 are ample then there exists an $N \in \mathbb{N}$ such that the multiplication maps

$$\Gamma(X, \mathcal{O}_X(N m_1 A_1)) \otimes \Gamma(X, \mathcal{O}_X(N m_2 A_2)) \to \Gamma(X, \mathcal{O}_X(N(m_1 A_1 + m_2 A_2)))$$

are surjective for each $m_1, m_2 \geq 0$. However, this fails for *semiample* divisors.

Let $h : X \to \mathbb{P}^1 \times \mathbb{P}^1$ be a double cover branched over a smooth curve of bidegree $(2d, 2d)$; composing with the projections yield morphisms $g_i : X \to \mathbb{P}^1, i = 1, 2$, with connected fibers. Let f_1 and f_2 be the fibers of $\mathbb{P}^1 \times \mathbb{P}^1$; take A_1 and A_2 to be their preimages on X. Then we have

$$\Gamma(X, \mathcal{O}_X(m A_1)) = \Gamma(\mathbb{P}^1, \mathcal{O}_{\mathbb{P}^1}(m))$$

for each $m \geq 0$, i.e., sections of

$$\Gamma(X, \mathcal{O}_X(m_1 A_1)) \otimes \Gamma(X, \mathcal{O}_X(m_2 A_2))$$

are obtained via pullback from sections of $\Gamma(\mathbb{P}^1 \times \mathbb{P}^1, \mathcal{O}_{\mathbb{P}^1 \times \mathbb{P}^1}(m_1, m_2))$. Since $m_1 A_1 + m_2 A_2$ is very ample on X for suitable $m_1, m_2 \gg 0$, we conclude that

$$\Gamma(X, \mathcal{O}_X(m_1 A_1)) \otimes \Gamma(X, \mathcal{O}_X(m_2 A_2)) \to \Gamma(X, \mathcal{O}_X(m_1 A_1 + m_2 A_2))$$

cannot be surjective. Indeed, the decomposable sections cannot separate points in a fiber of h.

THEOREM 5.7. *Suppose X is a smooth projective variety with $\mathrm{Pic}(X)$ free of finite rank. Assume that $\mathrm{Cox}(X)$ is finitely generated. Then the universal torsor admits an embedding*

$$\iota : U \hookrightarrow \mathrm{Spec}(\mathrm{Cox}(X))$$

that is equivariant under the action of the Néron-Severi torus T_X.

PROOF. First, we construct the morphism ι. Again, let $D_1, \ldots, D_r$ denote divisors freely generating the divisor class group of X. The cone of effective divisors of X is finite rational polyhedral and strictly convex, so we can choose $D_1, \ldots, D_r$ such that each effective divisor on X can be written as a nonnegative linear combination of $D_1, \ldots, D_r$. (Of course, the D_i themselves need not be effective.)

Let $L_1, \ldots, L_r$ designate the line bundles associated to the invertible sheaves $\mathcal{O}_X(D_1), \ldots, \mathcal{O}_X(D_r)$. Writing $P_i = L_i^{-1} \setminus \mathbf{0}_X$ we have

$$U = P_1 \times_X \cdots \times_X P_r \subset L_1^{-1} \times_X \cdots \times_X L_r^{-1}$$

which we interpret as the natural inclusion of

$$\mathrm{Spec}_X \left(\bigoplus_{(n_1,\ldots,n_r)\in\mathbb{Z}^r} \mathcal{O}_X(n_1D_1 + \cdots + n_rD_r) \right)$$

into

$$\mathrm{Spec}_X \left(\bigoplus_{n_1,\ldots,n_r\geq 0} \mathcal{O}_X(n_1D_1 + \cdots + n_rD_r) \right).$$

For each $(n_1, \ldots, n_r) \in \mathbb{Z}^r_{\geq 0}$, we have

$$\Gamma(X, \mathcal{O}_X(n_1D_1 + \cdots + n_rD_r)) \otimes \mathcal{O}_X \to \mathcal{O}_X(n_1D_1 + \cdots + n_rD_r)$$

which induces

$$\mathrm{Spec}_X \left(\textstyle\bigoplus_{(n_1,\ldots,n_r)\in\mathbb{Z}^r_{\geq 0}} \mathcal{O}_X(n_1D_1 + \cdots + n_rD_r) \right) \longrightarrow \mathrm{Spec}_X \left(\textstyle\bigoplus_{(n_1,\ldots,n_r)\in\mathbb{Z}^r_{\geq 0}} \Gamma(\mathcal{O}_X(n_1D_1 + \cdots + n_rD_r)) \otimes \mathcal{O}_X \right).$$

Since each effective divisor is a nonnegative sum of the D_i, the target is isomorphic to $X \times \mathrm{Spec}(\mathrm{Cox}(X))$. Thus we get a morphism

$$\begin{array}{ccc} U & \to & X \times \mathrm{Spec}(\mathrm{Cox}(X)) \\ & \searrow \quad \swarrow & \\ & X & \end{array}$$

and composing with the projection yields

$$\iota : U \to \mathrm{Spec}(\mathrm{Cox}(X)).$$

Our construction is clearly equivariant with respect to the actions of T_X.

We prove ι is an open embedding. First, observe that $\mathrm{Spec}(\mathrm{Cox}(X))$ is normal, i.e., $\mathrm{Cox}(X)$ is integrally closed in its fraction field. Since X is normal,

$$\bigoplus_{n_1,\ldots,n_r\geq 0} \mathcal{O}_X(n_1D_1 + \cdots + n_rD_r)$$

is a sheaf of integrally-closed domains, whose global sections form an integrally closed domain (cf, [**Har77**, Ex. 5.14(a)]). Furthermore, $\mathrm{Cox}(X)$ is even a UFD

[**EKW04**, Cor. 1.2]; this should not be surprising, as every effective divisor D on X naturally yields a principal divisor on $\mathrm{Spec}(\mathrm{Cox}(X))$, namely, the locus where the associated section $s \in \Gamma(X, \mathcal{O}_X(D)) \subset \mathrm{Cox}(X)$ vanishes.

We next exhibit a finitely-generated T_X-invariant subalgebra

$$R \subset \mathrm{Cox}(X)$$

such that the induced morphism

$$j : U \xrightarrow{\iota} \mathrm{Spec}(\mathrm{Cox}(X)) \to \mathrm{Spec}(R)$$

is an open embedding. Choose *ample* divisors $A_1, \ldots, A_r$ freely generating $\mathrm{Pic}(X)$. (Since being ample is an open condition, we can certainly produce these.) For each ample A_i, we obtain an embedding

$$X \hookrightarrow \mathbb{P}(w_i)$$

into a weighted projective space, where the weights

$$w_i = (w_{i1}, \ldots, w_{ij(i)})$$

index the degrees of a minimal set of homogeneous generators for the graded ring

$$x_{i1}, \ldots, x_{ij(i)} \in \bigoplus_{N \geq 0} \Gamma(X, \mathcal{O}_X(NA_i)).$$

Take products to obtain

$$X \hookrightarrow \prod_{i=1}^{r} \mathbb{P}(w_i)$$

and let R denote the multihomogeneous coordinate ring of X, i.e., the quotient of the polynomial ring in the x_{ij} by the multihomogeneous polynomials cutting out X. We can then identify

$$U = \mathrm{Spec}(R) - \bigcup_{i=1}^{r} \{x_{i1} = \cdots = x_{ij(i)} = 0\}.$$

Thus we have a diagram

$$\begin{array}{ccc} U & \xrightarrow{\iota} & V \quad := \mathrm{Spec}(\mathrm{Cox}(X)) \\ {\scriptstyle j}\downarrow & & \downarrow \\ j(U) & \subset & W \quad := \mathrm{Spec}(R) \end{array}$$

with V normal and j an open embedding. Let $U' \subset V$ denote the pre-image of $j(U)$ in V. The induced morphism

$$\beta : U' \to j(U) \simeq U$$

is a birational morphism from a normal variety with a section, induced by $\iota \circ j^{-1}$. Any such morphism is an isomorphism. Indeed, the composed morphism

$$U' \xrightarrow{\beta} U \xrightarrow{\iota} U'$$

agrees with the identity on a dense subset of U', hence is the identity. Thus β and ι are inverses of each other. □

THEOREM 5.8. *Let X be a smooth projective surface with $-K_X$ nef and big. Then $\mathrm{Cox}(X)$ is finitely generated.*

PROOF. Proposition 4.9 implies that $\overline{\mathrm{NE}}_1(X)$ is a finite rational polyhedral cone admitting a finite number of (-1)- and (-2)-curves. Thus the nef cone of X takes the form

$$\langle A_1, \ldots, A_r \rangle$$

where the A_i are nef divisors. Theorem 4.5 guarantees that each A_i is semiample. Consider the subring of the Cox ring

$$\mathrm{Cox}'(X) := \bigoplus_{D \in \langle A_1, \ldots, A_r \rangle} \Gamma(X, \mathcal{O}_X(D))$$

which is finitely generated by Proposition 5.5.

We next set up some notation, relying on the fact that $-K_X$ is semiample with associated contraction $\sigma : X \to S$ (Corollary 4.6). Let $E_1, \ldots, E_r$ denote the (-2)-curves on X, i.e., the curves contracted by σ. Let $F_1, \ldots, F_s$ denote the (-1)-curves on X. Choose generators $\eta_i \in \Gamma(\mathcal{O}_X(E_i))$ and $\xi_i \in \Gamma(\mathcal{O}_X(F_i))$, which are unique up to scalars. We regard these as elements of $\mathrm{Cox}(X)$.

LEMMA 5.9. Let D be an effective divisor on X. Express

$$D = M + F \tag{5.3}$$

where F is the fixed part and M is the moving part. Then the support of F consists of (-1)- and (-2)-curves.

PROOF. Suppose that the fixed part of F contains an irreducible component C that is not a (-1)- or (-2)-curve. It follows that $C^2 \geq 0$. Since C is effective, we have

$$h^2(\mathcal{O}_X(C)) = h^0(\mathcal{O}_X(K_X - C)) = 0.$$

Otherwise, $n(K_X - C)$ would be effective for each $n \geq 0$, which contradicts our assumption that $-K_X$ is big. The Hodge index theorem implies $-K_X \cdot C > 0$, so Riemann-Roch implies $h^0(\mathcal{O}_X(C)) > 1$, which means that C is not fixed. □

We interpret this via the Cox ring: Each homogeneous element $t \in \mathrm{Cox}(X)$ can be identified with an effective divisor $D = \{t = 0\}$. Expression (5.3) translates into $t = m \cdot f$, where $m \in \mathrm{Cox}'(X)$ and

$$f = \eta_1^{a_1} \cdots \eta_r^{a_r} \xi_1^{b_1} \cdots \xi_s^{b_s}, \quad a_1, \ldots, a_r, b_1, \ldots, b_s \in \mathbb{N}.$$

It follows then that

$$\mathrm{Cox}(X) = \mathrm{Cox}'(X)[\eta_1, \ldots, \eta_r, \xi_1, \ldots, \xi_s]$$

which completes our proof. □

REMARK 5.10. We make a few observations on the significance of Theorem 5.8 and recent generalizations.

- Hu and Keel [**HK00**] showed that smooth projective varieties with finitely generated Cox rings behave extremely well from the standpoint of birational geometry. Indeed, they designate such varieties *Mori Dream Spaces*.
- Shokurov [**Sho96**, §6] demonstrated how a robust version of the log minimal model program would imply that many classes of varieties have finitely generated Cox rings. For example, he established that log Fano threefolds over fields of characteristic zero have this property. These are a natural generalization of the singular Del Pezzo surfaces discussed here.

- As an application of their proof of the existence of minimal models for varieties of log general type (over fields of characteristic zero), Birkar, Cascini, Hacon, and M$^{\text{c}}$Kernan proved that log Fano varieties of arbitrary dimension have finitely generated Cox rings [**BCHM06**, 1.3.1].

Exercises. Suppose D is an effective divisor on a smooth projective surface X. Consider the graded ring

$$R(D) := \bigoplus_{m \geq 0} \Gamma(X, \mathcal{O}_X(mD)).$$

In the classic paper [**Zar62**], Zariski analyzed when this ring is finitely generated.

EXERCISE 5.4.1. Recall the notation of the second half of Exercise 4.4.1. Show that $R(D')$ is not finitely generated.

EXERCISE 5.4.2. Assume D admits a *Zariski decomposition* [**Zar62**, 7.7] [**Laz04**, 2.3.19], i.e.,

$$D = P + N \tag{5.4}$$

where P and N are $\mathbb{Q}$-divisors with the following properties:

- P is nef;
- N is effective with support
$$\operatorname{supp}(N) = \{C_i\}$$
generating a negative definite (or trivial) sublattice of the Néron-Severi group;
- $P \cdot C_i = 0$ for each $C_i \in \operatorname{supp}(N)$.

Deduce that

- for each $n \geq 0$ the map
$$\Gamma(X, \mathcal{O}_X(nD - \lceil nN \rceil)) \hookrightarrow \Gamma(\mathcal{O}_X(nD))$$
is an isomorphism;
- $\Gamma(X, \mathcal{O}_X(nP)) \simeq \Gamma(\mathcal{O}_X(nD))$ for $n \geq 0$ such that nN is integral.

If $-K_X$ is nef and big, deduce also that

- P is semiample;
- $\operatorname{supp}(N) \subset \{E_1, \ldots, E_r, F_1, \ldots, F_s\}$, the union of the (-1)- and (-2)-curves on X.

Hint: The second assertion is a corollary of the first. To prove this, note that any divisor A with

$$nD - \lceil nN \rceil \prec A \preceq nD$$

intersects some component in $\operatorname{supp}(N)$ negatively, and thus has that component in its fixed part.

EXERCISE 5.4.3. Let X be the Hirzebruch surface $\mathbb{F}_2$, Σ the class of a section at infinity, f the class of a fiber:

	Σ	f
Σ	2	1
f	1	0

This admits a unique (-2)-curve $E = \Sigma - 2f$.

- Show that $\mathrm{Cox}(X) \simeq k[\eta, f_0, f_\infty, t]$ where
$$\Gamma(\mathcal{O}_X(E)) = k\eta, \quad \Gamma(\mathcal{O}_X(f)) = kf_0 + kf_\infty,$$
and
$$\Gamma(\mathcal{O}_X(\Sigma)) = k\eta f_0^2 + k\eta f_0 f_\infty + k\eta f_\infty^2 + kt.$$
- Show that the Zariski decomposition of the divisor $D = \Sigma - f$ is
$$D = P + N, \quad P = \frac{1}{2}\Sigma, N = \frac{1}{2}E.$$
Verify that the fixed part of nD for $n \geq 0$ is $\lceil nN \rceil = \lceil n/2 \rceil E$.

5.5. More Cox rings of Del Pezzo surfaces. For blow-ups $\beta : X \to \mathbb{P}^2$, we write L for the pullback of the line class on $\mathbb{P}^2$ and $E_1, E_2, \ldots$ for the exceptional curves.

Example 5.11 (Degree Six Del Pezzo Surfaces). Let X be isomorphic to $\mathbb{P}^2$ blown up at three non-collinear points, which can be taken to be $p_1 = [1,0,0]$, $p_2 = [0,1,0]$, and $p_3 = [0,0,1]$. This is a toric variety under the action of the diagonal torus. We have seen in §2.2 that $\overline{\mathrm{NE}}_1(X)$ is generated by the (-1)-curves:
$$\{E_1, E_2, E_3, E_{12}, E_{13}, E_{23}\}$$
where E_{ij} is the proper transform of the line joining p_i and p_j with class $L - E_i - E_j$. Here we have (cf. [**BP04**, 3.1]):
$$\mathrm{Cox}(X) = k[\eta_1, \eta_2, \eta_3, \eta_{12}, \eta_{13}, \eta_{23}].$$

Example 5.12 (Degree Five Del Pezzo Surfaces). This example is due to Skorobogatov [**Sko93**] (see also [**BP04**, 4.1]). Suppose that X is isomorphic to $\mathbb{P}^2$ blown up at four points in linear general position, which can be taken to be $p_1 = [1,0,0]$, $p_2 = [0,1,0]$, $p_3 = [0,0,1]$, and $p_4 = [1,1,1]$. Let $E_i, i = 1, \ldots, 4$ denote the exceptional curves and E_{ij} the proper transforms of the lines joining p_i, with class $E_{ij} = L - E_i - E_j$. Skorobogatov shows there exist normalizations of the generators $\eta_{i5} \in \Gamma(\mathcal{O}_X(E_i))$ and $\eta_{ij} \in \Gamma(\mathcal{O}_X(E_{ij}))$ such that
$$\mathrm{Cox}(X) = k[\eta_{12}, \ldots, \eta_{45}] / \langle P_1, P_2, P_3, P_4, P_5 \rangle$$
where each P_i is a *Plücker relation*
$$P_i = \eta_{jk}\eta_{lm} - \eta_{jl}\eta_{km} + \eta_{jm}\eta_{kl}, \quad \{i,j,k,l,m\} = \{1,2,3,4,5\}, j < k < l < m.$$
More geometrically, $\mathrm{Cox}(X)$ is the projective coordinate ring of the Grassmannian $\mathbb{G}(1,4) \subset \mathbb{P}^9$.

Example 5.13 ($\mathbf{E}_6$ cubic surface). See [**HT04**, §3] for more details. Let $S \subset \mathbb{P}^3$ denote the (unique) cubic surface with a singularity of type $\mathbf{E}_6$
$$S = \{(w,x,y,z) : xy^2 + yw^2 + z^3 = 0\} \subset \mathbb{P}^3$$
and $\sigma : X \to S$ its minimal resolution of singularities. Let $E_1, \ldots, E_6$ denote the exceptional curves of σ and $\ell \subset X$ the proper transform of the unique line $\{y = z = 0\} \subset S$. The effective cone here is simplicially generated by (-1)- and (-2)-curves
$$\overline{\mathrm{NE}}_1(X) = \langle \ell, E_1, E_2, E_3, E_4, E_5, E_6 \rangle$$
but the corresponding elements $\xi_\ell, \xi_1, \ldots, \xi_6 \in \mathrm{Cox}(X)$ do not suffice to generate it. In this case, for a suitable ordering of the E_i we have
$$\mathrm{Cox}(X) \simeq k[\xi_1, ..., \xi_6, \xi_\ell, \tau_1, \tau_2, \tau_\ell] / \langle \tau_\ell \xi_\ell^3 \xi_4^2 \xi_5 + \tau_2^2 \xi_2 + \tau_1^3 \xi_1^2 \xi_3 \rangle.$$

We mention some other significant results:

- Batyrev and Popov [**BP04**] showed that the Cox ring of a Del Pezzo surface X of degree $d = 2, 3, 4, 5, 6$ is generated by sections associated with (-1)-curves on X. They show the relations (up to radical) are given by quadratic expressions analogous to the Plücker-type relations above. Furthermore, they conjectured that these quadratic relations actually generate the ideal of all relations.
- The Batyrev-Popov conjecture was proven for Del Pezzo surfaces of degree $d \geq 4$ and cubic surfaces without Eckardt points by Stillman, Testa, and Velasco [**STV07**]. Derenthal [**Der06a**] has also made significant contributions to our understanding of the relations in the Cox ring.
- Laface and Velasco [**LV07**] established the Batyrev-Popov conjecture when $d \geq 2$. Sturmfels and Xu [**SX08**] and Testa, Várilly-Alvarado, and Velasco [**TVAV08**] address Del Pezzo surfaces of degree one.
- For $d = 2, 3, 4, 5$ the affine variety $\mathrm{Spec}(\mathrm{Cox}(X))$ can be related to homogeneous spaces G/P, where G is a simply-connected algebraic group associated to the root system arising from $K_X^{\perp} \subset N^1(S, \mathbb{Z})$ (cf. §3.3.) Here P is the maximal parabolic subgroup associated to a representation of G naturally connected with the (-1)-curves on X. (This generalized the relation discussed between Grassmannians and Cox rings of degree-five Del Pezzos.) See [**SS07**] and [**Der07**] for details, as well as [**Pop01**] for the case of degree four.
- There are numerous examples of singular Del Pezzo surfaces (like the $\mathbf{E}_6$ cubic surface) whose Cox rings admit a single relation. These are classified in [**Der06b**].

Exercises.

EXERCISE 5.5.1. Let X be the blow-up of $\mathbb{P}^2$ at three *collinear* points. Compute generators and relations for $\mathrm{Cox}(X)$. *Hint:* You can find the answer in [**Has04**].

References

[Abh56] S. Abhyankar, *Local uniformization on algebraic surfaces over ground fields of characteristic $p \neq 0$*, Ann. of Math. (2) **63** (1956), 491–526. MR 0078017 (17,1134d)

[AGZV85] V. I. Arnol′d, S. M. Guseĭn-Zade, and A. N. Varchenko, *Singularities of differentiable maps. Vol. I*, Monographs in Mathematics, vol. 82, Birkhäuser Boston Inc., Boston, MA, 1985, The classification of critical points, caustics and wave fronts, Translated from the Russian by I. Porteous and M. Reynolds. MR 777682 (86f:58018)

[Art62] M. Artin, *Some numerical criteria for contractability of curves on algebraic surfaces*, Amer. J. Math. **84** (1962), 485–496. MR 0146182 (26 #3704)

[BCHM06] C. Birkar, P. Cascini, C. D. Hacon, and J. M$^{\mathrm{c}}$Kernan, *Existence of minimal models for varieties of log general type*, 2006, arXiv:math/0610203.

[Bea96] A. Beauville, *Complex algebraic surfaces*, second ed., London Mathematical Society Student Texts, vol. 34, Cambridge University Press, Cambridge, 1996, Translated from the 1978 French original by R. Barlow, with assistance from N. I. Shepherd-Barron and M. Reid. MR 1406314 (97e:14045)

[BP04] V. V. Batyrev and O. N. Popov, *The Cox ring of a del Pezzo surface*, Arithmetic of higher-dimensional algebraic varieties (Palo Alto, CA, 2002), Progr. Math., vol. 226, Birkhäuser Boston, Boston, MA, 2004, pp. 85–103. MR 2029863 (2005h:14091)

[Bri71] E. Brieskorn, *Singular elements of semi-simple algebraic groups*, Actes du Congrès International des Mathématiciens (Nice, 1970), Tome 2, Gauthier-Villars, Paris, 1971, pp. 279–284. MR 0437798 (55 #10720)

[Bro06] T. D. Browning, *The density of rational points on a certain singular cubic surface*, J. Number Theory **119** (2006), no. 2, 242–283. MR 2250046 (2007d:14046)

[Cay69] A. Cayley, *A memoir on cubic surfaces*, Philosophical Transactions of the Royal Society of London **159** (1869), 231–326.

[CKM88] H. Clemens, J. Kollár, and S. Mori, *Higher-dimensional complex geometry*, Astérisque (1988), no. 166, 144 pp. (1989). MR 1004926 (90j:14046)

[CL97] T. Christof and A. Löbel, *Polyhedron representation transformation algorithm (PORTA)*, 1997, software available at `http://www.iwr.uni-heidelberg.de/groups/comopt/software/PORTA`.

[Cox95] D. A. Cox, *The homogeneous coordinate ring of a toric variety*, J. Algebraic Geom. **4** (1995), no. 1, 17–50. MR 1299003 (95i:14046)

[CT87] J.-L. Colliot-Thélène, *Arithmétique des variétés rationnelles et problèmes birationnels*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Berkeley, Calif., 1986) (Providence, RI), Amer. Math. Soc., 1987, pp. 641–653. MR 934267 (89d:11051)

[CTS87] J.-L. Colliot-Thélène and J.-J. Sansuc, *La descente sur les variétés rationnelles. II*, Duke Math. J. **54** (1987), no. 2, 375–492. MR 899402 (89f:11082)

[Der06a] U. Derenthal, *On the Cox ring of Del Pezzo surfaces*, 2006, arXiv:math/0603111v1.

[Der06b] ———, *Singular Del Pezzo surfaces whose universal torsors are hypersurfaces*, 2006, arXiv:math.AG/0604194.

[Der07] ———, *Universal torsors of del Pezzo surfaces and homogeneous spaces*, Adv. Math. **213** (2007), no. 2, 849–864. MR 2332612 (2008i:14052)

[dlB98] R. de la Bretèche, *Sur le nombre de points de hauteur bornée d'une certaine surface cubique singulière*, Astérisque (1998), no. 251, 51–77, Nombre et répartition de points de hauteur bornée (Paris, 1996). MR 1679839 (2000b:11074)

[dlBBD07] R. de la Bretèche, T. D. Browning, and U. Derenthal, *On Manin's conjecture for a certain singular cubic surface*, Ann. Sci. École Norm. Sup. (4) **40** (2007), no. 1, 1–50. MR 2332351 (2008e:11038)

[Dur79] A. H. Durfee, *Fifteen characterizations of rational double points and simple critical points*, Enseign. Math. (2) **25** (1979), no. 1-2, 131–163. MR 543555 (80m:14003)

[DV34] P. Du Val, *On isolated singularities of surfaces which do not affect the conditions of adjunction. I.*, Proc. Camb. Philos. Soc. **30** (1934), 453–459 (English).

[EKW04] E. J. Elizondo, K. Kurano, and K. Watanabe, *The total coordinate ring of a normal projective variety*, J. Algebra **276** (2004), no. 2, 625–637. MR 2058459 (2005b:14013)

[FH91] W. Fulton and J. Harris, *Representation theory*, Graduate Texts in Mathematics, vol. 129, Springer-Verlag, New York, 1991, A first course, Readings in Mathematics. MR 1153249 (93a:20069)

[Fou98] É. Fouvry, *Sur la hauteur des points d'une certaine surface cubique singulière*, Astérisque (1998), no. 251, 31–49, Nombre et répartition de points de hauteur bornée (Paris, 1996). MR 1679838 (2000b:11075)

[GHS03] T. Graber, J. Harris, and J. Starr, *Families of rationally connected varieties*, J. Amer. Math. Soc. **16** (2003), no. 1, 57–67 (electronic). MR 1937199 (2003m:14081)

[GJ00] E. Gawrilow and M. Joswig, *polymake: a framework for analyzing convex polytopes*, Polytopes—combinatorics and computation (Oberwolfach, 1997) (G. Kalai and G.M. Ziegler, eds.), DMV Sem., vol. 29, Birkhäuser, Basel, 2000, software available at `http://www.math.tu-berlin.de/polymake/`, pp. 43–73. MR 1785292 (2001f:52033)

[Har77] R. Hartshorne, *Algebraic geometry*, Springer-Verlag, New York, 1977, Graduate Texts in Mathematics, No. 52. MR 0463157 (57 #3116)

[Has04] B. Hassett, *Equations of universal torsors and Cox rings*, Mathematisches Institut, Georg-August-Universität Göttingen: Seminars Summer Term 2004, Universitätsdrucke Göttingen, Göttingen, 2004, pp. 135–143. MR 2183138 (2007a:14046)

[HB03] D. R. Heath-Brown, *The density of rational points on Cayley's cubic surface*, Proceedings of the Session in Analytic Number Theory and Diophantine Equations (Bonn), Bonner Math. Schriften, vol. 360, Univ. Bonn, 2003, p. 33. MR 2075628 (2005d:14033)

[HBM99] D. R. Heath-Brown and B. Z. Moroz, *The density of rational points on the cubic surface* $X_0^3 = X_1X_2X_3$, Math. Proc. Cambridge Philos. Soc. **125** (1999), no. 3, 385–395. MR 1656797 (2000f:11080)

[HK00] Y. Hu and S. Keel, *Mori dream spaces and GIT*, Michigan Math. J. **48** (2000), 331–348, Dedicated to William Fulton on the occasion of his 60th birthday. MR 1786494 (2001i:14059)

[HT04] B. Hassett and Y. Tschinkel, *Universal torsors and Cox rings*, Arithmetic of higher-dimensional algebraic varieties (Palo Alto, CA, 2002), Progr. Math., vol. 226, Birkhäuser Boston, Boston, MA, 2004, pp. 149–173. MR 2029868 (2005a:14049)

[Isk79] V. A. Iskovskih, *Minimal models of rational surfaces over arbitrary fields*, Izv. Akad. Nauk SSSR Ser. Mat. **43** (1979), no. 1, 19–43, 237, English translation: Math. USSR-Izv. 14 (1980), no. 1, 17–39. MR 525940 (80m:14021)

[KM98] J. Kollár and S. Mori, *Birational geometry of algebraic varieties*, Cambridge Tracts in Mathematics, vol. 134, Cambridge University Press, Cambridge, 1998, With the collaboration of C. H. Clemens and A. Corti, Translated from the 1998 Japanese original. MR 1658959 (2000b:14018)

[Kol96] J. Kollár, *Rational curves on algebraic varieties*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics, vol. 32, Springer-Verlag, Berlin, 1996. MR 1440180 (98c:14001)

[Lan59] S. Lang, *Abelian varieties*, Interscience Tracts in Pure and Applied Mathematics. No. 7, Interscience Publishers, Inc., New York, 1959. MR 0106225 (21 #4959)

[Laz04] R. Lazarsfeld, *Positivity in algebraic geometry. I*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics, vol. 48, Springer-Verlag, Berlin, 2004, Classical setting: line bundles and linear series. MR 2095471 (2005k:14001a)

[LV07] A. Laface and M. Velasco, *Picard-graded Betti numbers and the defining ideals of Cox rings*, 2007, arXiv:0707.3251.

[Man66] Yu. I. Manin, *Rational surfaces over perfect fields*, Inst. Hautes Études Sci. Publ. Math. (1966), no. 30, 55–113, English translation: American Mathematical Society Translations. Series 2, Vol. 84 (1969): Twelve papers on algebra, algebraic geometry and topology. MR 0225780 (37 #1373)

[Man74] ———, *Cubic forms: algebra, geometry, arithmetic*, North-Holland Publishing Co., Amsterdam, 1974, Translated from Russian by M. Hazewinkel, North-Holland Mathematical Library, Vol. 4. MR 0460349 (57 #343)

[Mat02] K. Matsuki, *Introduction to the Mori program*, Universitext, Springer-Verlag, New York, 2002. MR 1875410 (2002m:14011)

[Mil80] J. S. Milne, *Étale cohomology*, Princeton Mathematical Series, vol. 33, Princeton University Press, Princeton, N.J., 1980. MR 559531 (81j:14002)

[MT86] Yu. I. Manin and M. A. Tsfasman, *Rational varieties: algebra, geometry, arithmetic*, Uspekhi Mat. Nauk **41** (1986), no. 2(248), 43–94, English translation: Russian Math. Surveys 41 (1986), no. 2, 51–116. MR 842161 (87k:11065)

[Mum95] D. Mumford, *Algebraic geometry. I*, Classics in Mathematics, Springer-Verlag, Berlin, 1995, Complex projective varieties, Reprint of the 1976 edition. MR 1344216 (96d:14001)

[Pop01] O. N. Popov, *Del Pezzo surfaces and algebraic groups*, 2001, Diplomarbeit, Universität Tübingen.

[Rei97] M. Reid, *Chapters on algebraic surfaces*, Complex algebraic geometry (Park City, UT, 1993), IAS/Park City Math. Ser., vol. 3, Amer. Math. Soc., Providence, RI, 1997, pp. 3–159. MR 1442522 (98d:14049)

[Sal98] P. Salberger, *Tamagawa measures on universal torsors and points of bounded height on Fano varieties*, Astérisque (1998), no. 251, 91–258, Nombre et répartition de points de hauteur bornée (Paris, 1996). MR 1679841 (2000d:11091)

[SB01] N. I. Shepherd-Barron, *On simple groups and simple singularities*, Israel J. Math. **123** (2001), 179–188. MR 1835294 (2002c:14076)

[SD72] H. P. F. Swinnerton-Dyer, *Rational points on del Pezzo surfaces of degree* 5, Algebraic geometry, Oslo 1970 (Proc. Fifth Nordic Summer School in Math.), Wolters-Noordhoff, Groningen, 1972, pp. 287–290. MR 0376684 (51 #12859)

[Sho96] V. V. Shokurov, *3-fold log models*, J. Math. Sci. **81** (1996), no. 3, 2667–2699, Algebraic geometry, 4. MR 1420223 (97i:14015)

[Sko93] A. N. Skorobogatov, *On a theorem of Enriques-Swinnerton-Dyer*, Ann. Fac. Sci. Toulouse Math. (6) **2** (1993), no. 3, 429–440. MR 1260765 (95b:14018)

[Sko01] ———, *Torsors and rational points*, Cambridge Tracts in Mathematics, vol. 144, Cambridge University Press, Cambridge, 2001. MR 1845760 (2002d:14032)

[SS07] V. V. Serganova and A. N. Skorobogatov, *Del Pezzo surfaces and representation theory*, Algebra Number Theory **1** (2007), no. 4, 393–419. MR 2368955

[STV07] M. Stillman, D. Testa, and M. Velasco, *Gröbner bases, monomial group actions, and the Cox rings of del Pezzo surfaces*, J. Algebra **316** (2007), no. 2, 777–801. MR 2358614 (2008i:14054)

[SX08] B. Sturmfels and Z. Xu, *Sagbi bases of Cox-Nagata rings*, 2008, arXiv:0803.0892.

[TVAV08] D. Testa, A. Várilly-Alvarado, and M. Velasco, *Cox rings of degree one del Pezzo surfaces*, 2008, arXiv:0803.0353.

[Zar58a] O. Zariski, *On Castelnuovo's criterion of rationality $p_a = P_2 = 0$ of an algebraic surface*, Illinois J. Math. **2** (1958), 303–315. MR 0099990 (20 #6426)

[Zar58b] ———, *The problem of minimal models in the theory of algebraic surfaces*, Amer. J. Math. **80** (1958), 146–184. MR 0097404 (20 #3873)

[Zar62] ———, *The theorem of Riemann-Roch for high multiples of an effective divisor on an algebraic surface*, Ann. of Math. (2) **76** (1962), 560–615. MR 0141668 (25 #5065)

Department of Mathematics, Rice University, Houston, Texas 77005, USA
E-mail address: hassett@math.rice.edu

Clay Mathematics Proceedings
Volume 8, 2009

Non-abelian descent

David Harari

Abstract. These notes are the written version of three one hour talks presented at the 2006 Clay summer school in Goettingen. They address the application of the technique of non-abelian descent for rational points to bielliptic and Enriques surfaces.

For any field k of characteristic zero, we fix an algebraic closure $\bar{k}$ of k and we set $\Gamma := \mathrm{Gal}\,(\bar{k}/k)$ (we will sometimes write Γ_k for Γ if several fields are involved). The group Γ is the inverse limit of the groups $\mathrm{Gal}\,(L/k)$ when L runs over all finite Galois extensions of k. If k is a number field, we let Ω_k denote the set of all places of k, and k_v the completion of k at v.

1. Review of non-abelian cohomology

In this section k is any field of characteristic zero. The main reference for the non-abelian cohomology of groups is Serre's book [**Ser94**], chapter I.5.

Let G be an algebraic group over k (all k-groups are assumed to be linear, but not necessarily connected), and set $\overline{G} = G \times_k \bar{k}$.

Examples :

- G finite (defining G is the same as giving the abstract finite group $G(\bar{k})$, equipped with a continuous action of Γ for the profinite topology on Γ and the discrete topology on $G(\bar{k})$), e.g., $\mathbf{Z}/n$ (cyclic group of order n with trivial Galois action), μ_n (group of nth-roots of unity in $\bar{k}$ with the natural Galois action).
- G can be a k-torus (this means that $\overline{G}$ is isomorphic to some power of the multiplicative group $\mathbf{G}_m$), e.g., the 1-dimensional torus $R^1_{K/k}\mathbf{G}_m$ defined by the affine equation $x^2 - ay^2 = 1$, where $a \in k^*$ is a constant and $K := k(\sqrt{a})$. More generally, if L is a finite extension of k with k-basis $(\omega_1, ..., \omega_r)$, the $(r-1)$ dimensional torus $R^1_{L/k}\mathbf{G}_m$ is defined by the affine equation
$$N_{L/k}(x_1\omega_1 + \cdots + x_r\omega_r) = 1$$
where $x_1, ..., x_r$ are the variables.

2000 *Mathematics Subject Classification.* Primary 11E72, Secondary 14G05, 11G35 .

- $G = \mathrm{PGL_n}$ (it is semi-simple and *adjoint*, that is, the center is trivial), $G = \mathrm{SL_n}$ (it is semi-simple and simply connected).
- $G = \mathrm{O(q)}$ (orthogonal group of a quadratic form q); this group is not connected, there is an exact sequence of k-groups

$$1 \to \mathrm{SO(q)} \to \mathrm{O(q)} \to \mathbf{Z}/2 \to 0$$

If the rank of q is at least 3, then SO(q) is semi-simple (but not simply connected : its universal covering is $\mathrm{Spin}(q)$); if $q = \langle 1, -a \rangle$ is of rank 2, then SO(q) is just the torus $R^1_{K/k}\mathbf{G}_m$ with $K = k(\sqrt{a})$.

We define the group $H^0(k, G) = H^0(\Gamma, G(\bar{k})) = G(k)$. For example $H^0(\mathbf{Q}, \mu_n)$ is trivial if n is odd. The *Galois cohomology set* $H^1(k, G) = H^1(\Gamma, G(\bar{k}))$ is the quotient of the set of 1-*cocycles* $Z^1(k, G)$ by an equivalence relation defined as follows. The set $Z^1(k, G)$ consists of continuous maps $f : \Gamma \to G(\bar{k})$ satisfying the cocycle condition

$$f(\gamma_1\gamma_2) = f(\gamma_1) \cdot {}^{\gamma_1}f(\gamma_2)$$

for each $\gamma_1, \gamma_2 \in \Gamma$. Two cocycles f, g are equivalent if there exists $b \in G(\bar{k})$ such that $f(\gamma) = b^{-1}g(\gamma)^{\gamma}b$ for every $\gamma \in \Gamma$. There is no canonical group structure on $H^1(k, G)$ if G is not commutative, but there is a distinguished element (denoted 0), namely, the class of the trivial cocycle. Therefore $H^1(k, G)$ is a pointed set.

Remark: The continuity assumption implies that

$$H^1(k, G) = \varinjlim_{L} H^1(\mathrm{Gal}\,(L/K), G(L))$$

where L runs over the finite Galois extensions of k.

Other definition of $H^1(k, G)$. It is also possible to define $H^1(k, G)$ as the set of isomorphism classes of *principal homogeneous spaces* (p.h.s.) of G over k. By definition such a p.h.s. is a non-empty set A, equipped with a left action of Γ and a simply-transitive right action of $G(\bar{k})$, such that the compatibility formula

$$\gamma(x.g) = \gamma(x) \cdot \gamma(g)$$

holds for every $\gamma \in \Gamma$, $x \in A$, $g \in G(\bar{k})$.

The correspondence between the two definitions goes as follows :

Let $\gamma \mapsto c_\gamma$ be a cocycle in $Z^1(k, G)$. Then define A as the p.h.s. with underlying set $G(\bar{k})$, but the *twisted* action of Γ defined by $\gamma(x) = c_\gamma \cdot {}^{\gamma}x$ (and $G(\bar{k})$ acts on the right on A). One checks that cohomologous cocycles give isomorphic p.h.s.

Conversely if A is a p.h.s. of G over k, choose a point $x_0 \in A$; then for each $\gamma \in \Gamma$, there exists a unique $c_\gamma \in G(\bar{k})$ such that $\gamma(x_0) = x_0 \cdot c_\gamma$. This defines a cocycle in $Z^1(k, G)$, and the cohomology class of this cocycle does not depend on x_0; moreover isomorphic p.h.s. also give cohomologous cocycles.

Remark: In the case we consider, any p.h.s. A is *representable* by the k-variety X defined as the quotient of $G \times_k \bar{k}$ by the action of Γ corresponding to A (the quotient exists because a group variety is quasi-projective). The k-variety X is a k-*form* of $\overline{G} := G \times_k \bar{k}$ (that is $\overline{X} \simeq \overline{G}$), and the p.h.s. A is trivial iff $X(k) \neq \emptyset$; the latter is also equivalent to the existence of $x_0 \in A$ such that $\gamma(x_0) = x_0$ for all $\gamma \in \Gamma$.

Properties of $H^1(k, G)$.

- The set $H^1(k,G)$ is covariant in G (easy with the cocycle definition), and in k (it is contravariant in $\operatorname{Spec} k$): if $k \subset L$ is an inclusion of fields, then there is a map $H^1(k,G) \to H^1(L,G)$, induced by the map $X \mapsto X \times_k L$ from isomorphism classes of k-p.h.s. to isomorphism classes of L-p.h.s.
- If

$$1 \to G_1 \to G_2 \to G_3 \to 1$$

is an exact sequence of k-groups (this means that the sequence of groups $1 \to G_1(\bar{k}) \to G_2(\bar{k}) \to G_3(\bar{k}) \to 1$ is exact), then there is an exact sequence of pointed sets

$$1 \to G_1(k) \to G_2(k) \to G_3(k) \to H^1(k,G_1) \to H^1(k,G_2) \to H^1(k,G_3).$$

In the special case when G_1 is central in G_2, this sequence can be extended with a map $H^1(k,G_3) \to H^2(k,G_1)$, but this map is not a morphism of groups in general, even if G_1 and G_3 are abelian.

Remark: "Exact sequence" of pointed sets means that the image of a map is the kernel of the following map; it can happen that a map has trivial kernel but is not injective.

Examples.

- By Hilbert's Theorem 90, we have $H^1(k,\mathrm{GL}_n) = H^1(k,\mathrm{SL}_n) = 0$.
- If T is a non-split torus, it can happen that $H^1(k,T) \neq 0$. For example if $T = R^1_{K/k}\mathbf{G}_m$, we have $H^1(k,T) = k^*/NK^*$; to see this, write T as the kernel of the norm map $R_{K/k}G_m \to \mathbf{G}_m$ (where $R_{K/k}$ stands for Weil's restriction), and use Hilbert's Theorem 90 (by Shapiro's lemma, the cohomology group $H^1(k,R_{K/k}\mathbf{G}_m)$ is isomorphic to $H^1(K,\mathbf{G}_m)$ (hence it is zero) because $(R_{K/k}G_m)(\bar{k})$ is the Galois module induced by $\bar{k}^*$ and the inclusion $\Gamma_K \to \Gamma_k$).
- Suppose G is a semi-simple, connected and simply connected group. Then $H^1(k,G) = 0$ when k is a p-adic field. For a number field k, the natural map

$$H^1(k,G) \to \bigoplus_{v \in \Omega_{\mathbf{R}}} H^1(k_v,G)$$

is an isomorphism (Kneser/Harder/Chernousov). These are special cases of "Serre's conjecture II" (see [**Ser94**], III.3).
- The exact sequence $1 \to \mathbf{G}_m \to \mathrm{GL}_n \to \mathrm{PGL}_n \to 1$ is central. It induces an exact sequence

$$1 \to H^1(k,\mathrm{PGL}_n) \to H^2(k,\mathbf{G}_m) = \operatorname{Br} k$$

Actually the theory of central simple algebras implies that the map from $H^1(k,\mathrm{PGL}_n)$ to the Brauer group $\operatorname{Br} k$ is injective, its image is a subset of the n-torsion $(\operatorname{Br} k)[n]$, and the union of the images of $H^1(k,\mathrm{PGL}_n)$ in $\operatorname{Br} k$ is the whole $\operatorname{Br} k$. By class field theory, the image of $H^1(k,\mathrm{PGL}_n)$ is the whole $(\operatorname{Br} k)[n]$ when k is a p-adic field or a number field, but not in general.

2. Extension to étale cohomology

A reference for this section is Skorobogatov's book [**Sko01**], I.5. See also [**HS02**], section 4.

Let X be an algebraic k-variety. The cohomology set $H^1(X,G)$ is defined using Čech cocycles for the étale topology. As in the case $X = \operatorname{Spec} k$, the pointed set $H^1(X,G)$ classifies isomorphism classes of (right) X-torsors (i.e. p.h.s.) under G. Namely, such a torsor is a k-variety Y equipped with a faithfully flat morphism $f : Y \to X$ and a right action of G on Y, such that $G(\bar{k})$ acts simply transitively on $Y_{\bar{x}} := f^{-1}(\bar{x})$ for each geometric point $\bar{x} \in X(\bar{k})$.

The functorial properties of $H^1(X,G)$ are as in the case $X = \operatorname{Spec} k$, and there is also the same behaviour relative to short exact sequences of k-groups (simply replacing k by X). In particular the class $[Y]$ of a torsor Y in $H^1(X,G)$ is zero iff Y is isomorphic to the trivial torsor $X \times_k G$ iff the morphism $f : Y \to X$ has an X-section. If $X' \to X$ is a morphism of k-varieties, it induces a map $H^1(X,G) \to H^1(X',G)$, which maps $[Y]$ to $[Y \times_X X']$. A morphism of k-groups $G \to H$ induces a map $H^1(X,G) \to H^1(X,H)$, such that the image of $[Y]$ is the class of the *contracted product* $Y \times^G H$, which is defined as the quotient of $Y \times G$ by the diagonal action

$$(y,g) \cdot h := (y \cdot h, h^{-1}g)$$

of H.

Let $m \in X(k)$ and $[Y] \in H^1(X,G)$. The k-morphism $\operatorname{Spec} k \to X$ corresponding to m induces an *evaluation map* $[Y] \mapsto [Y](m) \in H^1(k,G)$, and we have $[Y](m) = 0$ iff the fibre Y_m of the torsor $Y \to X$ has a k-point. More generally, for every cocycle $c \in Z^1(k,G)$, the equality $[Y](m) = [c]$ holds iff $[Y^c](m) = 0$, where Y^c is the *twisted torsor* of Y by c: it is an X-torsor under the *twisted group* G^c. The group G^c is an inner form of G: namely, $\overline{G^c} = \overline{G}$ and the new Galois action on G^c is given by $\gamma(g) = c_\gamma \, {}^\gamma g c_\gamma^{-1}$ for every $\gamma \in \Gamma$ and $g \in G^c(\bar{k}) = G(\bar{k})$; the torsor Y^c is isomorphic to Y over $\bar{k}$, but the Galois action on Y^c is twisted via the formula

$$\gamma(y) = {}^\gamma y \cdot c_\gamma^{-1}$$

If G is abelian, then $G^c = G$ and $[Y^c] = [Y] - [c]$ in the abelian group $H^1(X,G)$. We obtain the obvious (albeit important) *descent statement*:

Proposition 2.1. *Let $f : Y \to X$ be a torsor under a k-group G. For each $c \in Z^1(k,G)$, let $f^c : Y^c \to X$ be the corresponding twisted torsor. Then*

$$X(k) = \bigcup_{[c] \in H^1(k,G)} f^c(Y^c(k))$$

From now on we assume that k is a number field. Let X be a smooth variety such that $X(k_v) \neq \emptyset$ for every completion k_v of k. Let $X(\mathbf{A}_k)$ be the set of adelic points of X; if X is projective this set is just $\prod_{v \in \Omega_k} X(k_v)$. Let $f : Y \to X$ be a torsor under a k-algebraic group G, and define

$$X(\mathbf{A}_k))^f = \bigcup_{[c] \in H^1(k,G)} f^c(Y^c(\mathbf{A}_k))$$

In other words $X(\mathbf{A}_k)^f$ is the subset of $X(\mathbf{A}_k)$ consisting of those points (P_v) such that the evaluation $[Y](P_v) \in \prod_{v \in \Omega_k} H^1(k_v,G)$ belongs to the diagonal image of $H^1(k,G)$. In particular $X(k) \subset X(\mathbf{A}_k)^f$, hence the condition $X(\mathbf{A}_k)^f = \emptyset$ is an

obstruction to the Hasse principle, the *descent obstruction* associated to the torsor $f : Y \to X$ (or to the cohomology class $[Y] \in H^1(X, G)$).

Remark: This construction is not interesting if G is semi-simple and simply connected, or if G is a split torus. Indeed, in these cases we have $H^1(k, G) = 0$ for every field k, hence $X(\mathbf{A}_k)^f = X(\mathbf{A}_k)$.

THEOREM 2.2 ([**HS02**], Th. 4.7). *Assume further that X is projective. Then $X(\mathbf{A}_k)^f$ contains the closure $\overline{X(k)}$ of $X(k)$ in $X(\mathbf{A}_k)$.*

This theorem is a consequence of the so-called Borel-Serre theorem in Galois cohomology ([**Ser94**], III.4). If X is projective and $X(\mathbf{A}_k)^f \neq X(\mathbf{A}_k)$, we obtain a *descent obstruction to weak approximation*.

A natural question is to compare these descent obstructions to the so-called *Brauer-Manin obstruction*. Let X be a smooth and geometrically integral k-variety and $\operatorname{Br} X = H^2(X, \mathbf{G}_m)$ its Brauer group (if X is the spectrum of a field F, then $\operatorname{Br} X$ is just the classical Brauer group $\operatorname{Br} F$ of the field F). The reciprocity law in global class field theory yields an exact sequence

$$0 \longrightarrow \operatorname{Br} k \longrightarrow \bigoplus_{v \in \Omega_k} \operatorname{Br} k_v \stackrel{\sum_v j_v}{\longrightarrow} \mathbf{Q}/\mathbf{Z} \longrightarrow 0$$

where $j_v : \operatorname{Br} k_v \to \mathbf{Q}/\mathbf{Z}$ is the local invariant. Therefore, the set $X(k)$ is a subset of the subset

$$X(\mathbf{A}_k)^{\mathrm{Br}} := \{(P_v) \in X(\mathbf{A}_k), \forall \alpha \in \operatorname{Br} X, \sum_{v \in \Omega_k} j_v(\alpha(P_v)) = 0\}$$

In particular the condition $X(\mathbf{A}_k)^{\mathrm{Br}} = \emptyset$ implies that $X(k) = \emptyset$. This is the Brauer-Manin obstruction to the Hasse principle. If X is assumed to be projective, then the set $X(\mathbf{A}_k)^{\mathrm{Br}}$ contains $\overline{X(k)}$ ([**CTS87**], III.1) and the condition $X(\mathbf{A}_k)^{\mathrm{Br}} \neq X(\mathbf{A}_k)$ is the Brauer-Manin obstruction to weak approximation.

Special cases. a) The theory of descent developed by Colliot-Thélène and Sansuc [**CTS87**] (refined by Skorobogatov) implies that the Brauer-Manin obstruction associated to $\operatorname{Br}_1 X := \ker[\operatorname{Br} X \to \operatorname{Br} \overline{X}]$ corresponds to considering all descent obstructions associated to groups G *of multiplicative type* (i.e. commutative linear groups whose connected component of 1 is a torus), see [**HS02**], Theorem 4.9.

b) There are examples of Brauer-Manin obstructions associated to "transcendental" elements (that is, elements that do not vanish in $\operatorname{Br} \overline{X}$) of $\operatorname{Br} X$ ([**Har96**], [**Wit04**]); they correspond to descent obstructions related to $G = \mathrm{PGL}_n$ ([**HS02**], Th. 4.10). This uses the exact sequence $H^1(X, \mathrm{GL}_n) \to H^1(X, \mathrm{PGL}_n) \to \operatorname{Br} X$, and a theorem of Gabber (cf. [**dJ05**]) saying that $\operatorname{Br} X$ is the union of the images of $H^1(X, \mathrm{PGL}_n)$ in $\operatorname{Br} X$.

c) For G finite and non-commutative, the descent obstruction can refine the Brauer-Manin obstruction, that is, the set $X(\mathbf{A}_k)^{\mathrm{Br}}$ can be strictly bigger than $X(\mathbf{A}_k)^f$. An example of this situation will be explained in the next section.

3. Bielliptic surfaces

3.1. First properties of bielliptic surfaces. Geometrically (that is, over $\bar{k}$), a *bielliptic surface* is the quotient of the product $E_1 \times E_2$ of two elliptic curves by the free action of a finite group F (there are 7 possibilities for F, see for example

[**Bea78**], VI.20). We shall say that a k-variety X is a bielliptic surface if $\overline{X} := X \times_k \bar{k}$ is a bielliptic surface. Then the geometric invariants of X are $H^2(X, \mathcal{O}_X) = 0$ and $\dim H^1(X, \mathcal{O}_X) = 1$. In particular the geometric Brauer group $\mathrm{Br}\,\overline{X}$ is finite by Grothendieck's results ([**Gro68**]).

In these notes, we will restrict ourselves to the case $F = \mathbf{Z}/2$. We consider a bielliptic surface X over k, equipped with an étale covering Y with group $\mathbf{Z}/2$, such that $\overline{Y}$ is the product of two elliptic curves. In particular there is an exact sequence associated to the geometric étale fundamental groups

$$1 \to \pi_1(\overline{Y}) \to \pi_1(\overline{X}) \to \mathbf{Z}/2 \to 1.$$

Unlike $\pi_1(\overline{Y})$, $\pi_1(\overline{X})$ is not abelian. Indeed $\pi_1(\overline{Y})$ is isomorphic to $\widehat{\mathbf{Z}}^4$ and $\pi_1(\overline{X})^{\mathrm{ab}}$ is of rank 2 because $\dim H^1(X, \mathcal{O}_X) = 1$.

Bielliptic surfaces were used by Colliot-Thélène, Skorobogatov and Swinnerton-Dyer ([**CTSSD97**]) to disprove a conjecture of Mazur. Then Skorobogatov exploited the properties of these surfaces to give the first counterexample to the Hasse principle not accounted for by the Brauer-Manin obstruction. In the next subsection, we will summarize his construction.

3.2. Skorobogatov's construction. The reference for this subsection is the paper [**Sko99**].

THEOREM 3.1 (Skorobogatov, 1997). *There exists a bielliptic surface X over $k = \mathbf{Q}$ such that $X(\mathbf{Q}) = \emptyset$ but $X(\mathbf{A_Q})^{\mathrm{Br}} \neq \emptyset$.*

The idea is as follows. Skorobogatov constructs a tower of coverings

$$Y' = C' \times D \to Y = C \times D \xrightarrow{f} X$$

where C and D are curves of genus one with $D(\mathbf{Q}) \neq \emptyset$ (but $C(\mathbf{Q}) = \emptyset$), with the following properties. The map $C' \to C$ makes C' into a torsor under the finite k-group $E[2]$ consisting of 2-torsion points of an elliptic curve E, such that C' itself is a k-torsor under E. The class $[C'] \in H^1(k, E)$ is an element of order exactly 4 in the Tate-Shafarevich group $Ш(E)$. Recall that by definition $Ш(E)$ is the subgroup of $H^1(k, E)$ corresponding to elements whose restriction to $H^1(k_v, E)$ is zero for every place v of k. In particular C' has points in every completion of k but $C'(k) = \emptyset$.

Now the proof of Theorem 3.1 essentially breaks into two steps.

a) Under some assumptions (mainly the fact that $E(k)$ has no points of order exactly 2), prove that $(f')^*(\mathrm{Br}\,X) \subset \pi^*(\mathrm{Br}\,D)$, where f' is the map $Y' \to X$ and π the projection $Y' \to D$. This relies on careful computations of $\mathrm{Br}\,\overline{X} \simeq E[2]$ (hence $(\mathrm{Br}\,\overline{X})^{\Gamma} = 0$) and of $\mathrm{NS}\,\overline{X}$. Then it is very easy to construct points in $X(\mathbf{A}_k)^{\mathrm{Br}}$: it is sufficient to take the projection (Q_v) of $((P_v), R) \in Y'(\mathbf{A}_k)$, where $R \in D(k)$ and $(P_v) \in C'(\mathbf{A}_k)$); indeed for $\alpha \in \mathrm{Br}\,X$ such that $(f')^*(\alpha) = \pi^*(\beta)$ with $\beta \in \mathrm{Br}\,D$, we have

$$\sum_{v \in \Omega_k} j_v(\alpha(Q_v)) = \sum_{v \in \Omega_k} \beta(D)$$

(by functoriality) and $\beta(D) = 0$ because D is a rational point.

b) Prove that $X(k) \neq \emptyset$. This uses a descent argument. Only Y and the twist Y^- of Y by $(-1) \in H^1(\mathbf{Q}, \mathbf{Z}/2) = \mathbf{Q}^*/\mathbf{Q}^{*2}$ have points everywhere locally. Then one shows by a direct computation that $Y(\mathbf{Q}) = Y^-(\mathbf{Q}) = \emptyset$.

3.3. Interpretation in terms of non-abelian torsors. In [**Sko99**] Skorobogatov explains his counterexample by an "iterated version" of the Brauer-Manin obstruction. Namely, he shows that all twisted torsors Y^c of $Y \to X$ satisfy $Y^c(\mathbf{A}_k)^{\mathrm{Br}} = \emptyset$. This implies $Y^c(k) = \emptyset$, hence $X(k) = \emptyset$ by Proposition 2.1.

Actually (see [**HS02**], subsection 5.1 for a complete description of the situation) the emptyness of $Y(\mathbf{A}_k)^{\mathrm{Br}}$ corresponds to a descent obstruction associated to a torsor $g : Z \to Y$ under a finite abelian k-group (which is a k-form of $E[4]$). The composite map $h = f \circ g$ makes Z a torsor over X, but its structural group G is not abelian ($\overline{G}$ is a semi-direct product $E[4] \rtimes \mathbf{Z}/2$). We have $X(\mathbf{A}_k)^h = \emptyset$, which shows that the descent obstruction associated to a finite and non-abelian group can refine the Brauer-Manin obstruction. The situation is different for commutative groups or linear connected groups (see [**Har02**], Th. 2).

More generally, the fact that the geometric étale fundamental group $\pi_1(\overline{X})$ is not abelian is often crucial to construct counterexamples as above. Here is a general statement about weak approximation:

THEOREM 3.2 ([**Har00**]). *Let X be a smooth, projective and geometrically integral k-variety with $X(k) \neq \emptyset$. Assume that $H^2(X, \mathcal{O}_X) = 0$ and that $\pi_1(\overline{X})$ is not abelian. Assume further that the Albanese map (over $\bar{k}$) is flat with connected and reduced fibres. Then the closure $\overline{X(k)}$ of $X(k)$ in $X(\mathbf{A}_k)$ is strictly smaller than $X(\mathbf{A}_k)^{\mathrm{Br}}$.*

The condition on the Albanese map is technical (anyway it holds as soon as $H^1(X, \mathcal{O}_X) = 0$, or $\dim H^1(X, \mathcal{O}_X) = 1$ and $\dim X \geq 2$), the important point here being $\dim X > \dim H^1(X, \mathcal{O}_X)$.

For example, the theorem applies to any bielliptic surface. It works also for some étale quotients of abelian varieties (in higher dimension), and for some elliptic surfaces, as well as for certain general type surfaces. Nevertheless, constructing a similar counterexample to the Hasse principle for a variety of general type remains an open problem.

The idea to prove Theorem 3.2 is that the conditions on H^1 and H^2 mean that the set $X(\mathbf{A}_k)^{\mathrm{Br}}$ is sufficiently big. Then the condition on $\pi_1(\overline{X})$ yields a descent obstruction (associated to a finite and non-abelian group) for some points in $X(\mathbf{A}_k)^{\mathrm{Br}}$.

The theorem does not apply to Enriques surfaces (the geometric fundamental group is $\mathbf{Z}/2$). However we will see in the next sections that using torsors under an extension of $\mathbf{Z}/2$ by a torus, it is still possible to refine the Brauer-Manin obstruction for such surfaces.

4. Composition of two torsors

From now on we follow the paper [**HS05**]. Our goal is to construct an Enriques surface X over k and a torsor $f : Z \to X$ under a linear algebraic group G such that $X(\mathbf{A}_k)^{\mathrm{Br}}$ is not a subset of $X(\mathbf{A}_k)^f$ (in particular the Brauer-Manin obstruction to weak approximation is not the only one). As mentioned before, the group G has to be non-connected and non-commutative. Since we are going to define G as an extension, it is necessary to know that under certain conditions, the composition of two torsors is still a torsor. That is the aim of this section.

Let $Z \to Y$ be a torsor under a k-torus T. Colliot-Thélène and Sansuc defined the notion of *type* of the torsor Y: it is an element of $\mathrm{Hom}_\Gamma(\widehat{T}, \mathrm{Pic}\,\overline{Y})$, where $\widehat{T}$ is the Galois module of characters of $\overline{T} = T \times_k \bar{k}$. To define the type, observe that each element χ of $\widehat{T} = \mathrm{Hom}(\overline{T}, \mathbf{G}_m)$ induces a pushout $\chi_*([Z]) \in H^1(\overline{Y}, \mathbf{G}_m) = \mathrm{Pic}\,\overline{Y}$ of the class $[\overline{Z}] \in H^1(\overline{Y}, \overline{T})$; we obtain a homomorphism $\widehat{T} \to \mathrm{Pic}\,\overline{Y}$, which is clearly Γ-equivariant: this is the type of the torsor Z. When $\mathrm{Pic}\,\overline{Y}$ is torsion-free and T is the *Néron-Severi torus* of Y (that is, $\widehat{T}$ is isomorphic to $\mathrm{Pic}\,\overline{Y}$), Colliot-Thélène and Sansuc also defined *universal torsors* as torsors whose type λ is an isomorphism $\mathrm{Pic}\,\overline{Y} \to \mathrm{Pic}\,\overline{Y}$ (see for example [**Sko01**], (2.22) for more details).

Proposition 4.1 ([**HS05**]). *Let X be a smooth, projective, geometrically integral k-variety. Let $f : Y \to X$ be a torsor under a finite k-group H, and let $p : Z \to Y$ be a torsor under a k-torus T. Assume that the image $\mathrm{Im}\,\lambda \subset \mathrm{Pic}\,\overline{Y}$ of the type λ of Z is $H(\bar{k})$-invariant (e.g. Z universal). Then there exist a k-group G (extension of H by T) such that $f \circ p : Z \to X$ makes Z an X-torsor under G.*

The special case of this proposition we are interested in is when X is an Enriques surface. In this case (assuming $X(k) \neq \emptyset$), we have a $\mathbf{Z}/2$-torsor $f : Y \to X$, where Y is a $K3$ surface, and a universal torsor $Z \to Y$ under the Néron-Severi torus of Y. We obtain a torsor $g : Z \to X$ under a linear k-group G and an exact sequence

$$1 \to T \to G \to \mathbf{Z}/2 \to 1$$

It can be shown ([**HS05**], page 9, example 3) that the group G is commutative if and only if the map $f^* : \mathrm{Pic}\,\overline{X} \to \mathrm{Pic}\,\overline{Y}$ is surjective; this is the "generic" situation, but not the one we are going to consider for our construction.

5. A family of Enriques surfaces of Kummer type

The main theorem is the following.

Theorem 5.1 ([**HS05**]). *There exist an Enriques surface X over $k = \mathbf{Q}$, a torsor $g : Z \to X$ under a linear group G, and an adelic point $(P_v) \in X(\mathbf{A}_k)$ such that $(P_v) \in X(\mathbf{A}_k)^{\mathrm{Br}}$ but $(P_v) \notin X(\mathbf{A}_k)^g$. In particular the Brauer-Manin obstruction to weak approximation is not the only one for X.*

It is likely that there exists an Enriques surface X such that $X(\mathbf{A}_k)^{\mathrm{Br}} \neq \emptyset$ and $X(k) = \emptyset$ (via a descent obstruction associated to a torsor as in Theorem 5.1), but no such example is known.

Let us explain briefly the construction leading to Theorem 5.1. We start with genus one projective curves D_1, D_2 given by affine equations

$$y_1^2 = d_1(x^2 - a)(x^2 - ab^2)$$
$$y_2^2 = d_2(t^2 - a)(t^2 - ac^2)$$

where b, c, d_1, d_2 are constant elements of k^*, and a is a non-square element of k^*. We also demand that b, c are not ± 1. Note that the Jacobian varieties E_1, E_2 of D_1, D_2 have all 2-torsion points defined over k. Let Y be the *Kummer surface* defined as the minimal desingularization of $(D_1 \times D_2)/(-1)$, where (-1) is the involution induced by multiplication by -1 on D_1 and D_2. Namely, the $K3$ surface Y is a minimal smooth and projective model of the affine variety

$$y^2 = d(x^2 - a)(x^2 - ab^2)(t^2 - a)(t^2 - ac^2)$$

where $d = d_1 d_2$. It is equipped with the fixed-point-free involution $\sigma : (x, t, y) \mapsto (-x, -t, -y)$, and the quotient $X = Y/\sigma$ is an Enriques surface (the associated morphism $Y \to X$ will be denoted f).

Under very mild conditions on the constants a, b, c, d_1, d_2, we obtain that the elliptic curves $\overline{E}_1$ and $\overline{E}_2$ are not $\bar{k}$-isogenous (one just has to check that the modular invariant j_1 of E_1 is not integral over $\mathbf{Z}[j_2]$, where j_2 is the modular invariant of E_2). From this we deduce the following important fact:

PROPOSITION 5.2. *There exist 24 lines on $\overline{Y}$, defined over $L = k(\sqrt{a})$, such that:*

a) $\operatorname{Pic}\overline{Y}$ *is generated by the classes of these 24 lines.*

b) The action of the Enriques involution σ on the 24 lines coincides with the action of $\operatorname{Gal}(L/k)$.

The property b) is especially interesting, because it simplifies computations of group cohomology related to X. For example we can now show the following result:

PROPOSITION 5.3. *Let* $\operatorname{Br}_1 X = \ker[\operatorname{Br} X \to \operatorname{Br}\overline{X}]$. *Then* $f^*(\operatorname{Br}_1 X)$ *consists of constants (i.e. elements of* $\operatorname{Im}[\operatorname{Br} k \to \operatorname{Br} Y]$*).*

Proof: We have $\operatorname{Br}_1 X/\operatorname{Br} k = H^1(k, \operatorname{Pic}\overline{X})$ (cf. [**Sko01**], Corollary 2.3.9). The image of this group in $H^1(k, \operatorname{Pic}\overline{Y}) = \operatorname{Br}_1 Y/\operatorname{Br} k$ factors through $H^1(k, \operatorname{Pic}\overline{X}/\mathrm{tors})$ because $\operatorname{Pic}\overline{Y}$ is torsion-free. Thus it is sufficient to prove that $H^1(k, \operatorname{Pic}\overline{X}/\mathrm{tors}) = 0$. The Hochschild-Serre spectral sequence associated to $\bar{f} : \overline{Y} \to \overline{X}$ yields an exact sequence

$$0 \to \mathbf{Z}/2 \to \operatorname{Pic}\overline{X} \to (\operatorname{Pic}\overline{Y})^{\sigma} \to 0$$

(here we are using $H^2(\mathbf{Z}/2, \bar{k}^*) = \widehat{H}^0(\mathbf{Z}/2, \bar{k}^*) = 0$). Therefore $\operatorname{Pic}\overline{X}/\mathrm{tors} = (\operatorname{Pic}\overline{Y})^{\sigma}$ is a lattice with trivial Galois action because the Galois action on $\operatorname{Pic}\overline{Y}$ coincides with the action of σ thanks to Proposition 5.2. It follows that

$$H^1(k, \operatorname{Pic}\overline{X}/\mathrm{tors}) = 0.$$

□

Since X is a projective surface satisfying $H^2(X, \mathcal{O}_X) = 0$ and $\operatorname{NS}\overline{X} = \mathbf{Z}/2$, Grothendieck's results [**Gro68**] imply that $\operatorname{Br}\overline{X} = \mathbf{Z}/2$. The most difficult part in [**HS05**] consists of proving that the non-trivial element of $\operatorname{Br}\overline{X}$ does not come from an element of $\operatorname{Br} X$, which means $\operatorname{Br} X = \operatorname{Br}_1 X$. This holds as soon as neither $-d$ nor $-ad$ is a square in k^*. Using Proposition 5.3 and functoriality, we obtain

PROPOSITION 5.4. *Assume that neither $-d$ nor $-ad$ is a square in k^*. Then the projection on X of every adelic point $(N_v) \in Y(\mathbf{A}_k)$ belongs to $X(\mathbf{A}_k)^{\mathrm{Br}}$.*

The end of the proof of Theorem 5.1 consists of finding an adelic point (N_v) on Y such that $(N_v) \notin Y(\mathbf{A}_k)^p$, where $p : Z \to Y$ is a universal torsor. This is possible for example for $k = \mathbf{Q}$, $a = 5$, $b = 13$, $c = 2$, $d = 1$. Using Prop 4.1, we obtain a torsor $g : Z \to X$ under a group G by composing p with $f : Y \to X$. The group G is an extension of $\mathbf{Z}/2$ by a torus, but it is not commutative. Finally a Galois cohomology computation (sort of non-commutative "diagram-chasing") shows that the property $(N_v) \notin Y(\mathbf{A}_k)^p$ implies that $(M_v) := f(N_v)$ does not belong to $X(\mathbf{A}_k)^g$, although it is an element of $X(\mathbf{A}_k)^{\mathrm{Br}}$ by Proposition 5.4.

Remark: Actually, instead of working with a universal torsor it is easier to work with a torsor of another type (satisfying the assumptions of Proposition 4.1), which is associated to the 1-dimensional torus $R^1_{L/k}\mathbf{G}_m$. Then G is a k-form of an orthogonal group O_2.

6. A summary of results, conjectures, and questions

The following summarizes what is known, what should be true, and what is completely unknown about the Hasse principle and weak approximation on surfaces. Notice that for geometrically simply connected varieties, descent obstructions associated to linear groups cannot refine the Brauer-Manin obstruction because of [**Har02**], Th.2.

- Rational surfaces: it has been conjectured by Colliot-Thélène and Sansuc that the Brauer-Manin obstruction to the Hasse principle and weak approximation is the only one. Several significant cases are known (Châtelet surfaces, conic bundles with at most 5 degenerate fibres [**CTSSD87a, CTSSD87b**], [**CT90**], [**SS91**]).
- Abelian surfaces (with finite Tate-Shafarevich group): The Brauer-Manin obstruction to the Hasse principle is the only one, and the same results holds for weak approximation if archimedean places are not taken into account ([**Man71**], [**Wan96**]).
- Bielliptic surfaces: The Brauer-Manin obstruction to the Hasse principle is not the only one ([**Sko99**]), and similarly for weak approximation ([**Har00**]). The descent obstruction (associated to a finite non-commutative group) can refine the Brauer-Manin obstruction.
- $K3$ surfaces: since a $K3$ surface is geometrically simply connected, descent obstructions do not refine Brauer-Manin obstruction according to [**Har02**], Th. 2. (but "transcendental" obstructions can play a role, see [**Wit04**]). I have no clear idea whether the Brauer-Manin obstruction should be the only one (neither for Hasse principle nor for weak approximation).
- Enriques surfaces: The descent obstruction (associated to a non-connected linear group) can refine the Brauer-Manin obstruction, which is not the only one for weak approximation ([**HS05**]). It is likely (but not known) that the same should hold for the Hasse principle.
- Elliptic surfaces with Kodaira dimension 1: the Brauer-Manin obstruction is not the only one for weak approximation, because of descent obstructions associated to finite non-commutative groups ([**Har00**]). The same should hold for the Hasse principle.
- General type surfaces: the situation is the same as for elliptic surfaces with Kodaira dimension 1.

References

[Bea78] A. Beauville, *Surfaces algébriques complexes*, Société Mathématique de France, Paris, 1978, Avec une sommaire en anglais, Astérisque, No. 54. MR 0485887 (58 #5686)

[CT90] J.-L. Colliot-Thélène, *Surfaces rationnelles fibrées en coniques de degré* 4, Séminaire de Théorie des Nombres, Paris 1988–1989, Progr. Math., vol. 91, Birkhäuser Boston, Boston, MA, 1990, pp. 43–55. MR 1104699 (92j:14027)

[CTS87] J.-L. Colliot-Thélène and J.-J. Sansuc, *La descente sur les variétés rationnelles. II*, Duke Math. J. **54** (1987), no. 2, 375–492. MR 899402 (89f:11082)

[CTSSD87a] J.-L. Colliot-Thélène, J.-J. Sansuc, and P. Swinnerton-Dyer, *Intersections of two quadrics and Châtelet surfaces. I*, J. Reine Angew. Math. **373** (1987), 37–107. MR 870307 (88m:11045a)

[CTSSD87b] ———, *Intersections of two quadrics and Châtelet surfaces. II*, J. Reine Angew. Math. **374** (1987), 72–168. MR 876222 (88m:11045b)

[CTSSD97] J.-L. Colliot-Thélène, A. N. Skorobogatov, and P. Swinnerton-Dyer, *Double fibres and double covers: paucity of rational points*, Acta Arith. **79** (1997), no. 2, 113–135. MR 1438597 (98a:11081)

[dJ05] A. J. de Jong, *A result of gabber*, 2005, unpublished manuscript.

[Gro68] A. Grothendieck, *Le groupe de Brauer. II. Théorie cohomologique*, Dix Exposés sur la Cohomologie des Schémas, North-Holland, Amsterdam, 1968, pp. 67–87. MR 0244270 (39 #5586b)

[Har96] D. Harari, *Obstructions de Manin transcendantes*, Number theory (Paris, 1993–1994), London Math. Soc. Lecture Note Ser., vol. 235, Cambridge Univ. Press, Cambridge, 1996, pp. 75–87. MR 1628794 (99e:14025)

[Har00] ———, *Weak approximation and non-abelian fundamental groups*, Ann. Sci. École Norm. Sup. (4) **33** (2000), no. 4, 467–484. MR 1832820 (2002e:14034)

[Har02] David Harari, *Groupes algébriques et points rationnels*, Math. Ann. **322** (2002), no. 4, 811–826. MR MR1905103 (2003e:14038)

[HS02] D. Harari and A. N. Skorobogatov, *Non-abelian cohomology and rational points*, Compositio Math. **130** (2002), no. 3, 241–273. MR 1887115 (2003b:11056)

[HS05] ———, *Non-abelian descent and the arithmetic of Enriques surfaces*, Int. Math. Res. Not. (2005), no. 52, 3203–3228. MR 2186792 (2006m:14031)

[Man71] Yu. I. Manin, *Le groupe de Brauer-Grothendieck en géométrie diophantienne*, Actes du Congrès International des Mathématiciens (Nice, 1970), Tome 1, Gauthier-Villars, Paris, 1971, pp. 401–411. MR 0427322 (55 #356)

[Ser94] J.-P. Serre, *Cohomologie galoisienne*, fifth ed., Lecture Notes in Mathematics, vol. 5, Springer-Verlag, Berlin, 1994. MR 1324577 (96b:12010)

[Sko99] A. N. Skorobogatov, *Beyond the Manin obstruction*, Invent. Math. **135** (1999), no. 2, 399–424. MR 1666779 (2000c:14022)

[Sko01] ———, *Torsors and rational points*, Cambridge Tracts in Mathematics, vol. 144, Cambridge University Press, Cambridge, 2001. MR 1845760 (2002d:14032)

[SS91] P. Salberger and A. N. Skorobogatov, *Weak approximation for surfaces defined by two quadratic forms*, Duke Math. J. **63** (1991), no. 2, 517–536. MR 1115119 (93e:11079)

[Wan96] L. Wang, *Brauer-Manin obstruction to weak approximation on abelian varieties*, Israel J. Math. **94** (1996), 189–200. MR 1394574 (97e:11069)

[Wit04] O. Wittenberg, *Transcendental Brauer-Manin obstruction on a pencil of elliptic curves*, Arithmetic of higher-dimensional algebraic varieties (Palo Alto, CA, 2002), Progr. Math., vol. 226, Birkhäuser Boston, Boston, MA, 2004, pp. 259–267. MR 2029873 (2005c:11082)

Université Paris-Sud (Orsay), Bâtiment 425, 91405 Orsay Cedex, France
E-mail address: David.Harari@math.u-psud.fr

Clay Mathematics Proceedings
Volume 8, 2009

Mordell-Weil Problem for Cubic Surfaces, Numerical Evidence

Bogdan G. Vioreanu

Abstract. Let V be a plane smooth cubic curve over a finitely generated field k. The Mordell-Weil theorem for V states that there is a finite subset $P \subset V(k)$ such that the whole $V(k)$ can be obtained from P by drawing secants and tangents through pairs of previously constructed points and consecutively adding their new intersection points with V. In this paper we present numerical data regarding the analogous statement for cubic surfaces. For the surfaces examined, we also test Manin's conjecture relating the asymptotics of rational points of bounded height on a Fano variety with the rank of the Picard group of the surface.

1. Introduction

Let V be a smooth cubic surface over a field k in $\mathbb{P}^3$. If $x, y, z \in V(k)$ are three points (with multiplicities) lying on a line in $\mathbb{P}^3$ not belonging to V, we write $x = y \circ z$. Thus $\circ$ is a partial and multivalued composition law on $V(k)$. Note that $x \circ x$ is defined as the set of points in the intersection of $V(k)$ with the tangent plane at x. If x does not lie on a line, this is a cubic curve $C(x)$ with double point $x \in V(k)$. This whole set must be considered as the domain of the multivalued expression $x \circ x$, because geometrically all its points can be obtained by drawing tangents with k-rational direction to x. This means that an important source for generating new rational points on the cubic surface will be doubling the points that were already generated. The analogue of the Mordell-Weil theorem for cubic surfaces states that $(V(k), \circ)$ is finitely generated, i.e., there is a finite subset $P \subset V(k)$ such that the whole $V(k)$ can be obtained from P by drawing secants and tangent planes through pairs of (not necessarily distinct) previously constructed points, and consecutively adding their new intersection points with V. By drawing secants we can add only one rational point to P, while tangent sections give us an infinite number of points that can be generated, by the note above. For a more thorough discussion of various versions of finite generation cf. [**KM01**]. Note that, by Theorem 11.7 of [**Man86**], finite generation of $(V(k), \circ)$ implies that the universal quasi-group of $(V(k), \circ)$, as defined in [**Man86**], chapter II, is finite and has $2^n 3^m$ elements for some $n, m \in \mathbb{Z}_{\geq 0}$.

2000 *Mathematics Subject Classification.* Primary 11G35, Secondary 11G50, 14J26.

In the following, we present the procedure we used to test whether $(V(\mathbb{Q}), \circ)$ is finitely generated, and the results we obtained for thirteen diagonal cubic surfaces, six of them having the rank of their Picard group equal to 1, and seven of them mentioned in [**PT01**], illustrating the cases of surfaces with ranks 2 and 3 of the Picard group. We also bring numerical evidence supporting Manin's conjecture for the asymptotics of rational points of bounded height on a Fano variety. Note that John Slater and Sir Peter Swinnerton-Dyer have proved in [**SP98**] a one-sided estimate for the conjecture in the case when V contains two rational skew lines. All the computations were done using the Magma computer algebra system (cf. [**BCP97**].)

2. Description of the procedure

Let $ax^3+by^3+cz^3+du^3 = 0$, where a, b, c, d are nonzero integers, be a diagonal cubic surface. Using a program due to Dan Bernstein (see [**Ber01**]), we find all rational points on this surface up to height $H = 10^5$ or $H = 1.5 \cdot 10^5$, where the height of a rational point $P = (x : y : z : u)$, with $x, y, z, u \in \mathbb{Z}$ and $\gcd(x, y, z, u) = 1$ is defined as

$$h_{\max}(P) := \max\{|x|, |y|, |z|, |u|\}.$$

We consider also another height function $h_{\text{sum}} : V(\mathbb{Q}) \longrightarrow \mathbb{R}_+$ defined by

$$h_{\text{sum}}(P) := |x| + |y| + |z| + |u| .$$

Note that a rational point P can be uniquely written in the above form up to a sign change of the coordinates. So, if we assume, in addition, that the first nonzero coordinate of P is positive, then there is a unique such 'canonical' form corresponding to each point P. We order the rational points by increasing h_{sum}. If there are two or more points having the same height h_{sum}, then we order them lexicographically according to their coordinates in the canonical form. This defines a total order on the set of rational points. We will write $P < Q$ if P precedes Q in the sorted list, and use the number of a point in this list as its name. We will also refer to this number as the *index* of a rational point.

We will use the $h_{\max}$ height function only to study the asymptotics of the number of rational points on a cubic surface, while for the ordering of the points and in the implementation of the main function we will use h_{sum}.

For testing whether a given set of rational points is generating, we use the procedure *Test Generating Set (TGS)*, which is described below.

The procedure implements essentially a descent method. Given an index bound n and a set of points *GeneratedSet* that is presumably generating, we perform the following iterative process. In one iteration of loop, we consider all points in the range $\{1, \ldots, n\}$ that are not in *GeneratedSet* and test whether they can be decomposed as $x \circ y$, with $x, y \in$ *GeneratedSet*. Every point that can be decomposed in such a way is added to the *GeneratedSet* and at the end of the loop, the procedure is reiterated. As now *GeneratedSet* is bigger, there may be additional points in the range $\{1, \ldots, n\}$ that can be generated because we can choose the points x, y for a possible decomposition from a bigger set. The procedure is repeated until *GeneratedSet* stabilizes, i.e., until some iteration of the loop does not add any new points to the *GeneratedSet*.

In order to avoid repeating some operations of composing points, we use the additional variables *OldGeneratedSet*, *JustAdded* and *Decomp*. *OldGeneratedSet*

stores the value of *GeneratedSet* at the beginning of the iteration of the loop. At the end of the preceding loop, a number of points will have been added to *GeneratedSet*. These points are stored in the set variable *JustAdded*. During an iteration of the loop, we store in *Decomp* decompositions of the type $i = j \circ k$, with $i, j, k \leq n$, $i, k \notin GeneratedSet$ and $j \in GeneratedSet$. These are the only decompositions that we could further use. Indeed, if, at some point, k was added to *GeneratedSet*, then by searching in *Decomp*, we would find the decomposition $j \circ k$ of i and we would add i to *GeneratedSet* without performing any composition of points (which requires multiplications, so is computationally expensive) because we know, by the way we constructed *Decomp*, that $j \in GeneratedSet$ already.

Receiving as input the parameters *GeneratedSet* (a set of points in $V(\mathbb{Q})$ that is assumed to be generating), and n (the index bound for the points used in the decompositions), the *TGS* procedure does the following:

(1) Set $Decomp = \emptyset$, $OldGeneratedSet = \emptyset$.
(2) Set $JustAdded = GeneratedSet \setminus OldGeneratedSet$,
$OldGeneratedSet = GeneratedSet$.
(3) If $JustAdded = \emptyset$, return *GeneratedSet*.
(4) For every point $i \in \{1, 2, \ldots, n\} \setminus GeneratedSet$ do:
search in *Decomp* for decompositions of i as $x \circ y$ with $y \in JustAdded$
if such a decomposition exists, add i to *GeneratedSet*
else for every point j in *JustAdded* do:
$k = i \circ j$
if $k \in JustAdded$
add i to *GeneratedSet*
break
else if $k \leq n$ add the decomposition $(j \circ k)$ of i to *Decomp*
end for
end for.
(5) Go to step 2.

Let us explain in more detail the way the algorithm works. Suppose that an iteration of the outer loop has just finished, and we are in step 2. We set $JustAdded = GeneratedSet \setminus OldGeneratedSet$ and test whether this is the empty set. If this is so, then during the last iteration we could not generate any new points, so the maximum set of points that can be generated is the current *GeneratedSet*. If *JustAdded* is not empty, then during the last iteration we found a number of new points that could be generated and added them to *GeneratedSet* (these are the elements of *JustAdded*), so there is hope of generating other points. We consider a point $i \notin GeneratedSet$. Since we have already tested during the previous iteration whether we could decompose i as $x \circ y$, with $x, y \in OldGeneratedSet$, all we have to check now is whether we can write $i = x \circ y$ for $x \in JustAdded$ and either $y \in OldGeneratedSet$ or $y \in JustAdded$. At the previous iterations of the loop all compositions of i with points in *OldGeneratedSet* that could further be used (i.e., compositions whose result is not bigger than n) were stored in *Decomp*, so we can check for the first possibility by searching in the vector *Decomp*. Since by construction we only store in *Decomp* decompositions of the type $x \circ y$, with $x \in GeneratedSet$, all we have to check in the beginning of step 4 is whether $y \in JustAdded$–we are sure that $x \in GeneratedSet$. In order to check for the second possibility, we have to compose i with every point $j \in JustAdded$. If the result k of

the composition is in *JustAdded*, then we can write x as a composition of two points in *JustAdded*, so we add i to *GeneratedSet*. If the result $k \notin JustAdded$, but could be further used (i.e., $k \leq n$), then we store the corresponding decomposition $j \circ k$ of i in *Decomp*. The 'out of bounds' compositions, i.e., such that $i \circ j > n$, are implicitly remembered in the process (in the sense that they are done only once.)

Using the vector *Decomp* of course implies a trade off between space and speed, but we considered the latter to be more important. Even with *Decomp*, the computations for TGS for bounds n in the range of 10^5 last for several days and sometimes even weeks on an Intel Pentium IV processor with 2.26 GHz.

Before we proceed with the presentation of the results, let us provide an estimate of the height of the composition of two rational points. Here by h we mean either $h_{\max}$ or h_{sum} since the estimation of the asymptotics does not depend on the choice of the height function.

LEMMA 2.1. *Let $V : ax^3 + by^3 + cz^3 + du^3 = 0$ be a diagonal cubic surface, where a, b, c, d are nonzero integers, and let $K := \max\{|a|, |b|, |c|, |d|\}$. If A_1 and A_2 are two distinct points in $V(\mathbb{Q})$ that do not lie on a line in V, then*

$$h(A_1 \circ A_2) = O(K \cdot h(A_1)^2 \cdot h(A_2)^2)\,.$$

Proof: Let $A_1 = (x_1 : y_1 : z_1 : u_1)$, $A_2 = (x_2 : y_2 : z_2 : u_2)$ be in canonical form. Then one can check that

$$A_1 \circ A_2 = (\alpha x_1 - \beta x_2 : \alpha y_1 - \beta y_2 : \alpha z_1 - \beta z_2 : \alpha u_1 - \beta u_2)\,,$$

where

$$\alpha = ax_1x_2^2 + by_1y_2^2 + cz_1z_2^2 + du_1u_2^2 \in \mathbb{Z}\,,$$
$$\beta = ax_1^2x_2 + by_1^2y_2 + cz_1^2z_2 + du_1^2u_2 \in \mathbb{Z}\,.$$

Since the above coordinates of $A_1 \circ A_2$ are integers, the conclusion follows. This upper bound cannot be improved because, in most cases, the formula given represents $A_1 \circ A_2$ in canonical form (up to a sign change of the coordinates).

Concerning the doubling of points, if $A \in V(Q)$ is a rational point not lying on a line in V, then there is no upper bound for the height of the points in $A \circ A$ (since there are infinitely many such points). On the other hand, there can be many points of small height in $A \circ A$, especially if A has small height.

3. Results

Listed below are the thirteen diagonal cubic surfaces that were tested for finite generation, ordered according to the ranks of their Picard groups:

Rank 1 Picard group:

(1) $x^3 + 2y^3 + 3z^3 + 4u^3 = 0$.
(2) $x^3 + 2y^3 + 3z^3 + 5u^3 = 0$.
(3) $17x^3 + 18y^3 + 19z^3 + 20u^3 = 0$.
(4) $4x^3 + 5y^3 + 6z^3 + 7u^3 = 0$.
(5) $9x^3 + 10y^3 + 11z^3 + 12u^3 = 0$.
(6) $x^3 + 5y^3 + 6z^3 + 10u^3 = 0$.

Rank 2 Picard group:

(7) $x^3 + y^3 + 2z^3 + 4u^3 = 0$.
(8) $x^3 + y^3 + 5z^3 + 25u^3 = 0$.
(9) $x^3 + y^3 + 3z^3 + 9u^3 = 0$.

Rank 3 Picard group:

(10) $x^3 + y^3 + 2z^3 + 2u^3 = 0$.
(11) $x^3 + y^3 + 5z^3 + 5u^3 = 0$.
(12) $x^3 + y^3 + 7z^3 + 7u^3 = 0$.
(13) $2x^3 + 2y^3 + 3z^3 + 3u^3 = 0$.

The first six cubic surfaces illustrate the case of Picard group of rank 1. The third surface was considered as an example of a diagonal cubic surface with bigger coefficients. The lack of success in finding a generating set for this surface (as opposed to all the other surfaces examined by that point) motivated the study of the surfaces 4–5, which have coefficients of intermediate value between the coefficients of the first, successful surface, and the third, problematic one. Surface 6 is aimed to illustrate the case of surfaces with 'random' coefficients. The remaining seven surfaces were taken from [**PT01**] as examples of cubic surfaces with the ranks of the Picard group 2 and 3.

In order to find a suitable generating set G to begin with, we tested several small sets for finite generation up to a small index n ($n = 100$, or $n = 1000$). We observed that, if the set G generates more than $80\% - 90\%$ of the first n points for a small n, then this is a good indicator that the set G will generate roughly the same percentage of all points up to a much bigger index bound N (which we took to be either $5 \cdot 10^4$ or 10^5). We chose the initial small sets to be the set of points of indices $\{1, 2, 3, 4\}$. If this did not yield a large enough percentage of points generated, we would enlarge the initial set to $G = \{1, 2, 3, 4, 5\}$, and continue this way. Generally, we were 'lucky', in the sense that a few tries would provide us with a good generating set G (a set G that generates most of the first n points.) Then we would eliminate from G the 'superfluous' points, i.e., the points that could be obtained by composing other points in G. This is the reason for which, for example, the first surface has $G = \{3\}$ instead of $G = \{1, 2, 3, 4\}$: the points of indices 1, 2 and 4 lie in the tangent plane at the point of index 3.

At first, the only exception was the surface 3, which represents, at least computationally, a problem. Having added the surfaces 4–5, we noticed that it is hard to find a generating set using this naive method for these surfaces as well.

The generating sets we found are given in Table 1; these are listed both as sets of indices and as sets of rational points. Here, and in all subsequent tables, the label '$\mathcal{S}$' stands for 'surface'.

Before we go on and list the results we obtained using the *TestGeneratingSet* procedure, let us provide an indication of the asymptotics of the number of points on each cubic surface up to some height H. Note that, as we used Dan Bernstein's program to find rational points on the diagonal cubic surfaces, here 'height' refers to $h_{\max}$. The asymptotics of the number of points seems to be related to the percentage of points that can be generated up to some height. For the last seven surfaces, we did not take into consideration the points on the trivial rational lines, i.e., points of the type $(x : -x : y : -y)$, except for the point $(1 : -1 : 0 : 0)$ on the surfaces 7–9 and the points $(1 : -1 : 0 : 0)$, $(0 : 0 : 1 : -1)$, $(1 : -1 : 1 : -1)$ and $(1 : -1 : -1 : 1)$ on the surfaces 10–13, which we need for finite generation.

TABLE 1. Generating sets

S	G as set of indices	G as set of points
1	$\{3\}$	$\{(1:-1:-1:1)\}$
2	$\{1,2,4\}$	$\{(0:1:1:-1),(1:1:-1:0),(2:-2:1:1)\}$
6	$\{2\}$	$\{(1:-1:-1:1)\}$
7	$\{3\}$	$\{(1:-1:-1:1)\}$
8	$\{1,2\}$	$\{(1:-1:0:0),(1:4:-2:-1)\}$
9	$\{1,2,4\}$	$\{(1:-1:0:0),(1:2:0:-1),(1:2:-3:2)\}$
10	$\{5,6\}$	$\{(1:-1:-1:1),(1:-1:1:-1)\}$
11	$\{3,4\}$	$\{(1:-1:-1:1),(1:-1:1:-1)\}$
12	$\{1,2,5,6\}$	$\{(0:0:1:-1),(1:-1:0:0),(1:-1:-1:1),(1:-1:1:-1)\}$
13	$\{1,2,3,4,5\}$	$\{(0:0:1:-1),(1:-1:0:0),(1:-1:-1:1),(1:-1:1:-1),(3:-6:1:5)\}$

TABLE 2. Data on points of bounded height

S	Number of points up to height									
	100	200	500	1000	2000	5000	10000	20000	50000	100000
1	77	163	436	906	1827	4408	8754	17332	43280	86329
2	180	358	855	1683	3244	8097	16436	32704	82581	166825
3	16	25	62	117	204	502	1055	2084	5479	10840
4	37	78	206	414	778	1937	3877	7756	19701	39433
5	37	67	165	310	595	1580	3148	6257	15499	31134
6	55	120	316	646	1285	3131	6397	12753	32072	64102
7	196	458	1308	2746	6004	16758	35958	75984	205284	433526
8	142	292	766	1734	3872	10892	23338	49608	135128	286040
9	200	438	1270	2768	6200	17434	37018	78980	215626	455164
10	666	1630	5410	12870	29926	89218	205198	465226	1364810	3051198
11	412	1012	3328	7964	18676	56412	131512	299776	881774	1976482
12	702	1870	6010	14130	33156	100580	228696	520700	1526532	3420784
13	384	1052	3196	7752	18400	56348	130476	298860	876776	1966160

Table 2 includes intermediate results of the number of points up to different height limits. These results seem to confirm Manin's conjecture relating the asymptotics of rational points of bounded height on a Fano variety with the rank of the Picard group of the surface (see [**FMT89**]:)

$$\#\{P \in V(\mathbb{Q}) : h(P) < H\} \ \sim \ CH \log^{\mathrm{rkPic}(V)-1} H$$

for $H \longrightarrow \infty$, where h is an anticanonical height on V.

For the surfaces with rank of the Picard group equal to 1 we computed, additionally, the number of rational points up to slightly greater height limits, as summarized in Table 3 ('-' means 'not computed'.)

Relevant to our claim that these results seem to confirm Manin's conjecture are the graphs (Figures 1, 2, and 3) based on the tables above. In all graphs, we plotted the number of points up to height H divided by $H \log^{\mathrm{rkPic}(V)-1} H$ for various values of H. The conjecture would be verified if the plotted points would become arbitrarily close, in the limit, to a line parallel to the Ox axis, of equation $y = C$, where C is the constant predicted by Manin's conjecture. For a conjecture about the value of this constant, see [**PT01**].

TABLE 3. Additional data for surfaces with Picard group of rank one

S	Number of points up to height			
	150000	200000	250000	300000
1	129473	–	–	–
2	250286	–	–	–
3	16123	21627	27026	32507
4	59100	78498	–	–
5	46436	61958	77518	93079
6	96065	–	–	–

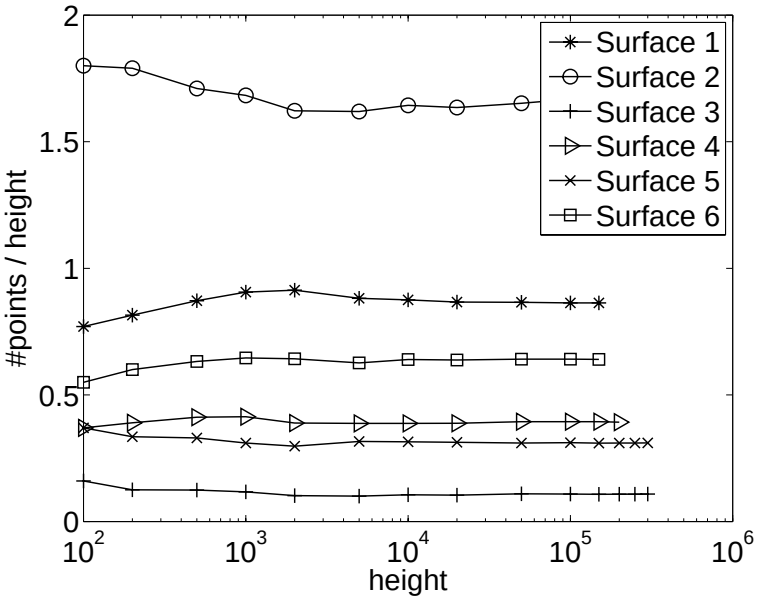


FIGURE 1. Surfaces with Picard group of rank 1

In the remaining, by 'height' we mean h_{sum}.

Note that for the surfaces with rank of the Picard group equal to two, most of the points are 'doubled', i.e., if $(x : y : z : u)$ is a point on the cubic surface, then so is $(y : x : z : u)$, while for the surfaces with rank of the Picard group equal to three, most of the points are 'quadrupled', i.e., if $(x : y : z : u)$ is a point on the cubic surface, then so are $(y : x : z : u)$, $(x : y : u : z)$ and $(y : x : u : z)$. In the following we list the results which were obtained using the *TestGeneratingSet* procedure. The generating sets used are the ones enumerated above, while the index bounds and the corresponding height bounds are given in the third and second columns of Table 4. '# iter' is the number of iterations of the outer loop of the procedure, and the 'first bad point' refers to the point of smallest index that could not be generated by the procedure. For example, the first line in the table reads "The procedure *TestGeneratingSet* called for surface 1, with index bound 100 corresponding to the

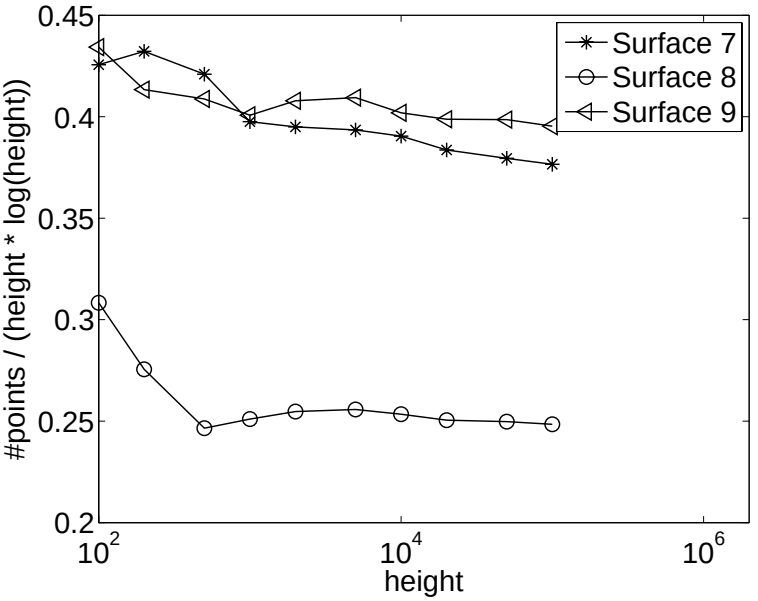


FIGURE 2. Surfaces with Picard group of rank 2

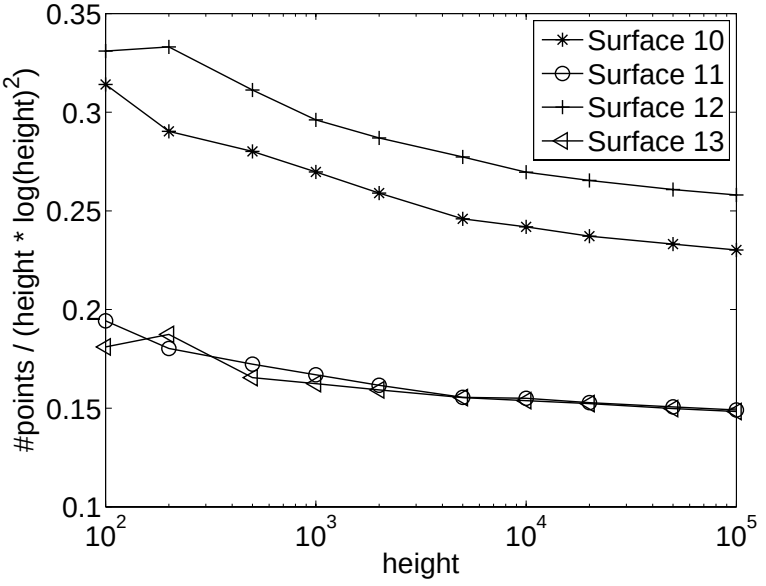


FIGURE 3. Surfaces with Picard group of rank 3

height bound 317, and initial generating set $G = \{3\}$ (or $G = \{\,(1 : -1 : -1 : 1)\,\}$), generates 74 rational points, which represents 74.0% of the first 100 points, in 4 iterations of the outer loop. The point of smallest height which could not be generated has index 30 and height 86."

TABLE 4. Statistics on the points generated by our potential generating set

Surface	Height bound	Index bound	# points generated	% points generated	# iter	First bad point Index	First bad point Height
1	317	100	74	74.0	4	30	86
1	617	200	160	80.0	9	30	86
1	1,443	500	463	92.6	16	42	130
1	2,788	1,000	923	92.3	15	255	788
1	5,574	2,000	1,859	93.0	14	543	1,541
1	14,456	5,000	4,747	94.9	15	1,145	3,192
1	29,074	10,000	9,462	94.6	14	1,593	4,423
1	58,775	20,000	18,957	94.8	14	3,633	10,322
1	147,343	50,000	47,418	94.8	13	8,522	24,677
1	296,822	100,000	94,910	94.9	13	8,522	24,677
2	150	100	97	97.0	7	85	124
2	282	200	196	98.0	9	90	134
2	703	500	483	96.6	8	258	364
2	1,477	1,000	973	97.3	9	358	511
2	3,020	2,000	1931	96.6	9	625	943
2	7,663	5,000	4,813	96.3	10	1,040	1,542
2	15,405	10,000	9,659	96.6	11	1,775	2,656
2	30,651	20,000	19,259	96.3	11	4,262	6,539
2	75,845	50,000	48,181	96.3	11	10,073	15,539
2	151,171	100,000	96,477	96.5	12	15,223	23,243
6	388	100	86	86.0	5	49	209
6	762	200	176	88.0	5	49	209
6	1,864	500	468	93.6	10	169	641
6	3,687	1,000	937	93.7	11	181	688
6	7,557	2,000	1,867	93.3	11	513	1,926
6	18,976	5,000	4,677	93.6	11	1,078	3,984
6	37,612	10,000	9,410	94.1	11	2,271	8,661
6	74,617	20,000	18,963	94.8	11	2,662	10,125
6	186,532	50,000	47,436	94.9	12	6,373	24,068
7	129	100	100	100.0	6	–	–
7	245	200	194	97.0	6	127	167
7	538	500	490	98.0	8	304	376
7	980	1,000	990	99.0	7	550	612
7	1,889	2,000	1,984	99.2	7	1,022	992
7	4,230	5,000	4,974	99.5	7	2,620	2,401
7	7,974	10,000	9,934	99.3	8	5,610	4,707
7	14,775	20,000	19,934	99.7	8	7,512	6,222
7	34,339	50,000	49,880	99.8	7	19,666	14,554
7	64,682	100,000	99,812	99.8	8	38,212	26,672
7	94,215	150,000	149,744	99.8	9	38,212	26,672

TABLE 5. Statistics on the points generated by our potential generating set, contd.

Surface	Height bound	Index bound	# points generated	% points generated	# iter	First bad point Index	First bad point Height
8	172	100	81	81.0	6	42	78
8	316	200	170	85.0	8	56	104
8	750	500	488	97.6	9	152	234
8	1,412	1,000	988	98.8	8	516	774
8	2,484	2,000	1,960	98.0	8	516	774
8	5,632	5,000	4,922	98.4	9	1,855	2,322
8	10,354	10,000	9,874	98.7	8	3,708	4,296
8	19,444	20,000	19,836	99.2	8	6,852	7,538
8	44,750	50,000	49,720	99.4	8	16,058	15,812
8	84,436	100,000	99,626	99.6	9	32,420	30,072
9	114	100	48	48.0	4	8	24
9	242	200	198	99.0	9	126	146
9	522	500	484	96.8	9	318	346
9	978	1,000	956	95.6	11	379	414
9	1,822	2,000	1,968	98.4	9	781	770
9	3,878	5,000	4,954	99.1	9	1,602	1,472
9	7,254	10,000	9,936	99.4	9	3,728	3,046
9	13,610	20,000	19,908	99.5	9	10,420	7,522
9	31,320	50,000	49,806	99.6	8	21,142	14,342
9	58,852	100,000	99,778	99.8	9	32,036	20,884
10	61	100	92	92.0	3	79	51
10	91	200	200	100.0	3	–	–
10	214	500	496	99.2	5	419	184
10	358	1,000	980	98.0	6	651	255
10	612	2,000	1,996	99.8	5	1,791	554
10	1,225	5,000	4,940	98.8	5	2,259	674
10	2,143	10,000	9,916	99.2	6	3,675	976
10	3,806	20,000	19,852	99.3	6	5,779	1,396
10	8,020	50,000	49,732	99.5	7	20,870	3,949
11	94	100	89	89.0	5	61	56
11	144	200	184	92.0	5	61	56
11	274	500	492	98.4	6	257	174
11	474	1,000	988	98.8	6	757	382
11	802	2,000	1,960	98.0	6	1,177	528
11	1,688	5,000	4,924	98.5	7	1,495	642
11	2,882	10,000	9,888	98.9	8	3,873	1,386
11	5,100	20,000	19,732	98.7	9	6,207	2,004
11	10,880	50,000	49,544	99.1	9	11,737	3,308

TABLE 6. Statistics on the points generated by our potential generating set, contd.

Surface	Height bound	Index bound	# points generated	% points generated	# iter	First bad point	
						Index	Height
12	48	100	96	96.0	4	95	46
12	92	200	192	96.0	4	95	46
12	186	500	476	95.2	5	223	106
12	286	1,000	956	95.6	6	223	106
12	484	2,000	1,969	98.5	5	964	284
12	1,014	5,000	4,911	98.2	6	2,315	548
12	1,740	10,000	9,880	98.8	6	3,486	764
12	3,066	20,000	19,832	99.2	7	4,030	856
12	6,514	50,000	49,532	99.1	7	16,064	2,578
13	106	100	96	96.0	4	41	75
13	167	200	196	98.0	5	169	153
13	316	500	484	96.8	6	169	153
13	515	1,000	980	98.0	6	572	360
13	910	2,000	1,944	97.2	6	860	465
13	1,885	5,000	4,896	97.9	7	1,937	897
13	3,310	10,000	9,780	97.8	7	3,102	1,323
13	5,727	20,000	19,672	98.4	7	4,785	1,816
13	12,139	50,000	48,256	96.5	8	8,202	2,805

Note that, in general, when using a greater index bound we found that the 'first bad point' changed (i.e., another point of greater height and index became the 'first bad point'), meaning that using stepping stones of bigger height typically fills up the gaps obtained when using a lower index bound. This is a good indicator that if we continue increasing the index (and thus the height) bounds, we will gradually generate *all* the points up to bigger and bigger heights.

Let us see now what happens with the 'problematic surfaces' 3–5. The data is displayed in Table 7. Unfortunately, any try of finding a generating set to begin with, that finds 'first bad points' of increasing height, and that generates a percentage of points similar to the ones obtained for the 'good' surfaces was not successful. Not even a 'brute force' approach like considering the initial *GeneratedSet* to be, say, the first 100 or 1000 points does not yield satisfactory results. The results are better for the surfaces 4–5 than for the surface 3, with the biggest coefficients, but still very 'bad'. In Table 7 we provide an illustration of the behavior of these surfaces when starting with the $GeneratedSet = \{1, 2, \ldots, 10\}$.

These results seem to support either that $\{1, 2, \ldots, 10\}$ is not a generating set for any of the three surfaces, or that the stepping stones needed to fill up the gaps (i.e., the rational points needed to decompose the 'first bad points') have very big heights. Although the percentages of generated points obtained for the surfaces 4–5 are slightly better than the percentages for the surface 3, they still become smaller and smaller as the index bound limit (and so also the height) grow. But the most important negative indicator is that 'the first bad point' never changes.

TABLE 7. Statistics for the 'problematic surfaces'

Surface	Height bound	Index bound	# points generated	% points generated	# iter	First bad point	
						Index	Height
3	2,161	100	17	17.0	2	13	203
3	5,495	200	24	12.0	2	13	203
3	13,429	500	35	7.0	2	13	203
3	25,874	1,000	49	4.9	2	13	203
3	51,663	2,000	81	4.1	2	13	203
3	124,062	5,000	154	3.1	2	13	203
3	251,103	10,000	274	2.7	2	13	203
3	505,619	20,000	429	2.1	2	13	203
4	658	100	26	26.0	1	12	50
4	1,345	200	50	25.0	2	12	50
4	3,307	500	102	20.4	2	12	50
4	6,774	1,000	172	17.2	3	12	50
4	13,772	2,000	284	14.2	3	12	50
4	34,552	5,000	487	9.7	3	12	50
4	68,425	10,000	781	7.8	3	12	50
4	135,691	20,000	1,222	6.1	4	12	50
5	844	100	19	19.0	1	13	103
5	1,691	200	26	13.0	2	13	103
5	4,394	500	51	10.2	2	13	103
5	8,780	1,000	80	8.0	2	13	103
5	16,962	2,000	119	6.0	2	13	103
5	43,224	5,000	216	4.3	2	13	103
5	87,176	10,000	338	3.4	3	13	103
5	174,128	20,000	538	2.7	3	13	103

In order to make progress, we introduced another approach to finding a generating set for the surfaces 3–5, based on the idea of 'throwing in' (adding to the *Generated Set*) the first bad points if they cannot be generated by decomposition. Our aim is to obtain, after adding sufficiently many 'first bad points', a set of points that generates a stable (or even better, increasing) percentage of points for increasing index bounds, and a 'changing first bad point' behavior, i.e., applying the *TGS* procedure to increasing index bounds would result in finding 'first bad points' of increasing heights.

We implement this new approach in the following way. We apply the *TGS* procedure to a (small) generating set and an index bound of 1000. We obtain a 'first bad point' that unfortunately stays the same when increasing the index bound (as observed when using our first approach). We apply again the *TGS* procedure to the initial generating set *and* this first bad point, with an index bound of 1000. We obtain another 'first bad point', of bigger index and height than the initial one. We add this point to our generating set (which now contains also the initial 'first bad point') and continue this way. We stop when we have added sufficiently many 'first bad points' to our initial set so that this new, bigger generating set fulfills the

two objectives mentioned above. Once we have obtained such a set, we stop adding points to our generating set and just increase the index bounds to make sure the percentage of generated points is indeed stable or increasing, and that the height of the 'first bad point' grows as the index bound is increased.

For example, for surface 4, we start with *Generated Set* $= \{1, 2, \ldots, 10\}$. We obtain the first bad point 12, which is stable–stays the same even if we increase the index bound. We add it to the *Generated Set* and call again the *TGS* procedure. We obtain more points, and another first bad point. We add this new bad point to the *Generated Set* and continue this way, gradually filling the holes. At first we kept the index bound constant, until we obtained a reasonable percentage of generated points. Then we tested whether the 'first bad point' changes when increasing the index bound and keeping the initial *Generated Set* constant (i.e., we stopped filling the holes, and just increased the index bound.) For surfaces 4 and 5 this approach seems reasonably successful, as reflected in Tables 8 and 9.

Unfortunately, for surface 3 this approach does not seem to work. After adding many more 'first bad points' to the initial generating set than for the surfaces 4–5, we still did not obtain a 'good' generating set, as illustrated in Table 10.

Since this process was too slow, we tried 'throwing' in our *Generated Set* not only the first bad point, but the first 10 bad points at every step (see Table 11).

This was again too slow, so we started inserting the first 20 bad points to our *Generated Set* (see Table 12).

Next we present other statistical data.

It seems that the percentage of points on a surface that can be strongly decomposed (a point x is *strongly decomposable* if it has a decomposition $x = y \circ z$ with $y, z < x$) up to some index N is approximately constant for various values of N. This suggests that this percentage may be an invariant for the surface.

It seems likely that if this percentage is bigger then *TestGeneratingSet* will generate more points (up to some index), using a suitable *GeneratedSet*. This is confirmed if we study the first two surfaces. Surface 1 has roughly $\frac{N}{8}$ points that are not strongly decomposable up to the index N (for $N \geq 1000$), while the surface 2 has only $\sim \frac{N}{11}$ such points; and indeed, if we compare the results of TGS for the two surfaces, we notice that TGS for the surface 2 generates more points (up to the same index) than TGS for the surface 1. Also, note that the percentage of points that are strongly decomposable for the surface 3 is very small (approximately 10%.) This may be one of the explanations for our lack of success with this surface.

4. Conclusion

The theory surrounding the Mordell-Weil problem for cubic surfaces seems not very well developed, mainly because of the difficulties caused by the lack of a group structure on the operation of composing points. In this paper we presented numerical data for thirteen diagonal cubic surfaces, in the hope of developing some intuition on a possible finiteness conjecture (first mentioned by Manin, cf. [**Man86**] and [**Man97**]). For each of the surfaces, we tried to find a generating set. A naive method gave positive results for ten of the surfaces, while a more rigorous method was needed to obtain similar (but not as positive) results for two of the other surfaces. For these surfaces, the numerical data suggest that they might be indeed finitely generated. The remaining surface resisted both methods. We cannot say,

TABLE 8. Analysis of surface 4 using 'TGS' procedure

Surface	Height bound	Index bound	# points generated	% points generated	# iter	First bad point	
						Index	Height
4	6,774	1,000	172	17.2	3	12	50
4	6,774	1,000	177	17.7	3	13	55
4	6,774	1,000	194	19.4	4	14	63
4	6,774	1,000	210	21.0	4	15	73
4	6,774	1,000	218	21.8	4	20	107
4	6,774	1,000	230	23.0	4	21	108
4	6,774	1,000	237	23.7	4	22	110
4	6,774	1,000	249	24.9	5	23	125
4	6,774	1,000	268	26.8	6	25	179
4	6,774	1,000	282	28.2	6	27	193
4	6,774	1,000	296	29.6	6	28	199
4	6,774	1,000	325	32.5	13	32	215
4	6,774	1,000	328	32.8	13	35	249
4	6,774	1,000	335	33.5	13	37	262
4	6,774	1,000	338	33.8	13	43	297
4	6,774	1,000	342	34.2	13	49	317
4	6,774	1,000	349	34.9	13	52	329
4	6,774	1,000	351	35.1	13	58	370
4	6,774	1,000	353	35.3	13	62	396
4	6,774	1,000	360	36.0	13	66	413
4	6,774	1,000	372	37.2	13	69	438
4	6,774	1,000	394	39.4	18	73	467
4	6,774	1,000	400	40.0	18	76	487
4	34,552	5,000	1,331	26.6	38	89	570
4	68,425	10,000	2,769	27.7	50	92	611
4	135,691	20,000	6,365	31.8	53	189	1,230
4	204,042	30,000	10,142	33.8	50	233	1,605
4	271,092	40,000	14,403	36.0	45	324	2,115
4	339,994	50,000	18,409	36.8	51	352	2,387

however, whether this means that the surface is not finitely generated, or that this is just a sign of the limits of the methods used.

Acknowledgement: The results of this paper arose as part of the author's research project at the Max Planck Institute for Mathematics in Bonn, under the guidance of Yu. I. Manin. The author would like to thank Yu. I. Manin for providing this very enjoyable opportunity.

The author is grateful to Michael Stoll for many useful discussions which contributed significantly to the improvement of the contents of this paper.

References

[BCP97] W. Bosma, J. Cannon, and C. Playoust, *The Magma algebra system. I. The user language*, J. Symbolic Comput. **24** (1997), no. 3-4, 235–265, Computational algebra and number theory (London, 1993), also see the Magma home page at http://www.maths.usyd.edu.au:8000/u/magma/. MR 1484478

[Ber01] D. J. Bernstein, *Enumerating solutions to* $p(a) + q(b) = r(c) + s(d)$, Math. Comp. **70** (2001), no. 233, 389–394. MR 1709145 (2001f:11203)

[FMT89] J. Franke, Yu. I. Manin, and Y. Tschinkel, *Rational points of bounded height on Fano varieties*, Invent. Math. **95** (1989), no. 2, 421–435. MR 974910 (89m:11060)

[KM01] D. Kanevsky and Yu. Manin, *Composition of points and the Mordell-Weil problem for cubic surfaces*, Rational points on algebraic varieties, Progr. Math., vol. 199, Birkhäuser, Basel, 2001, math.AG/0011198, pp. 199–219. MR 1875175 (2002m:14018)

[Man86] Yu. I. Manin, *Cubic forms*, second ed., North-Holland Mathematical Library, vol. 4, North-Holland Publishing Co., Amsterdam, 1986, Algebra, geometry, arithmetic, Translated from the Russian by M. Hazewinkel. MR 833513 (87d:11037)

[Man97] ———, *Mordell-Weil problem for cubic surfaces*, Advances in mathematical sciences: CRM's 25 years (Montreal, PQ, 1994) (L. Vinet, ed.), CRM Proc. Lecture Notes, vol. 11, Amer. Math. Soc., Providence, RI, 1997, math.AG/9407009, pp. 313–318. MR 1479681 (99a:14029)

[PT01] E. Peyre and Y. Tschinkel, *Tamagawa numbers of diagonal cubic surfaces of higher rank*, Rational points on algebraic varieties, Progr. Math., vol. 199, Birkhäuser, Basel, 2001, math.AG/9809054, pp. 275–305. MR 1875177 (2003a:11076)

[SP98] J. B. Slater and Swinnerton-Dyer P., *Counting points on cubic surfaces. I*, Astérisque (1998), no. 251, 1–12, Nombre et répartition de points de hauteur bornée (Paris, 1996). MR 1679836 (2000d:11087)

Jacobs University Bremen, College Ring 7, 28759 Bremen, Germany

Current address: Yale University Mathematics Department, PO Box 208283, New Haven, CT 06520-8283, USA

E-mail address: bogdan.vioreanu@yale.edu

TABLE 9. Analysis of surface 5 using 'TGS' procedure

Surface	Height bound	Index bound	# points generated	% points generated	# iter	First bad point	
						Index	Height
5	8,780	1,000	80	8.0	2	13	103
5	8,780	1,000	87	8.7	3	14	111
5	8,780	1,000	100	10.0	3	15	112
5	8,780	1,000	114	11.4	6	16	122
5	8,780	1,000	142	14.2	8	17	125
5	8,780	1,000	149	14.9	8	18	126
5	8,780	1,000	157	15.7	8	19	127
5	8,780	1,000	170	17.0	8	21	150
5	8,780	1,000	175	17.5	8	23	168
5	8,780	1,000	177	17.7	8	25	177
5	8,780	1,000	207	20.7	16	27	188
5	8,780	1,000	211	21.1	16	28	190
5	8,780	1,000	219	21.9	16	32	211
5	8,780	1,000	223	22.3	16	37	276
5	8,780	1,000	230	23.0	16	39	298
5	8,780	1,000	232	23.2	16	44	350
5	8,780	1,000	236	23.6	16	45	363
5	8,780	1,000	237	23.7	16	46	367
5	8,780	1,000	242	24.2	16	47	369
5	8,780	1,000	268	26.8	16	56	427
5	8,780	1,000	276	27.6	16	57	431
5	8,780	1,000	282	28.2	16	59	445
5	8,780	1,000	311	31.1	16	60	464
5	8,780	1,000	313	31.3	16	62	487
5	8,780	1,000	319	31.9	16	66	581
5	8,780	1,000	337	33.7	16	68	595
5	8,780	1,000	339	33.9	16	69	602
5	8,780	1,000	347	34.7	16	75	631
5	8,780	1,000	356	35.6	16	76	637
5	8,780	1,000	365	36.5	16	84	695
5	8,780	1,000	369	36.9	16	87	719
5	8,780	1,000	380	38.0	16	88	733
5	8,780	1,000	385	38.5	16	91	745
5	8,780	1,000	390	39.0	16	93	771
5	8,780	1,000	409	40.9	16	96	801
5	8,780	1,000	413	41.3	16	103	862
5	43,224	5,000	1,881	37.6	29	118	1,015
5	87,176	10,000	3,650	36.5	32	145	1,197
5	174,128	20,000	7,236	36.2	37	295	2,554
5	262,052	30,000	11,367	37.9	44	325	2,774
5	349,121	40,000	15,842	39.6	44	461	3,988
5	437,046	50,000	20,103	40.2	35	461	3,988

TABLE 10. Surface 3 remains problematic after adding the 'first bad point' at every step

Surface	Height bound	Index bound	# points generated	% points generated	# iter	First bad point	
						Index	Height
3	25,874	1,000	49	4.9	2	13	203
3	25,874	1,000	52	5.2	2	14	248
3	25,874	1,000	54	5.4	2	15	260
3	25,874	1,000	57	5.7	2	16	264
3	25,874	1,000	60	6.0	2	18	335
3	25,874	1,000	62	6.2	2	19	337
3	25,874	1,000	64	6.4	2	20	383
3	25,874	1,000	66	6.6	2	21	413
3	25,874	1,000	67	6.7	2	22	433
3	25,874	1,000	69	6.9	2	23	434
3	25,874	1,000	71	7.1	2	26	526
3	25,874	1,000	73	7.3	2	27	573
3	25,874	1,000	76	7.6	2	28	605
3	25,874	1,000	77	7.7	2	29	630
3	25,874	1,000	78	7.8	2	31	699
3	25,874	1,000	80	8.0	2	32	711
3	25,874	1,000	82	8.2	2	35	754
3	25,874	1,000	85	8.5	2	36	772
3	25,874	1,000	86	8.6	2	37	775
3	25,874	1,000	88	8.8	2	39	808
3	25,874	1,000	90	9.0	2	40	819
3	25,874	1,000	93	9.3	2	41	853
3	25,874	1,000	95	9.5	2	42	868
3	25,874	1,000	98	9.8	2	43	872
3	25,874	1,000	99	9.9	2	44	895
3	25,874	1,000	100	10.0	2	45	895
3	25,874	1,000	106	10.6	3	48	1,021
3	25,874	1,000	108	10.8	3	49	1,032
3	25,874	1,000	109	10.9	3	50	1,042
3	25,874	1,000	110	11.0	3	51	1,061
3	25,874	1,000	111	11.1	3	52	1,062
3	25,874	1,000	112	11.2	3	53	1,079
3	25,874	1,000	113	11.3	3	54	1,097
3	25,874	1,000	116	11.6	3	55	1,120
3	25,874	1,000	117	11.7	3	56	1,131
3	25,874	1,000	118	11.8	3	58	1,226
3	25,874	1,000	120	12.0	3	59	1,270

TABLE 11. Or the first ten bad points

Surface	Height bound	Index bound	# points generated	% points generated	# iter	First bad point	
						Index	Height
3	25,874	1,000	137	13.7	3	69	1,496
3	25,874	1,000	151	15.1	3	82	1,741
3	25,874	1,000	164	16.4	3	95	2,110
3	25,874	1,000	177	17.7	2	107	2,458
3	25,874	1,000	187	18.7	2	118	2,753
3	25,874	1,000	207	20.7	5	134	3,039
3	25,874	1,000	220	22.0	5	146	3,391
3	25,874	1,000	233	23.3	5	160	3,928
3	25,874	1,000	243	24.3	5	174	4,686
3	25,874	1,000	255	25.5	5	184	4,865
3	25,874	1,000	268	26.8	5	197	5,257

TABLE 12. Or even the first twenty bad points

Surface	Height bound	Index bound	# points generated	% points generated	# iter	First bad point	
						Index	Height
3	25,874	1,000	301	30.1	5	226	6,309
3	25,874	1,000	325	32.5	5	248	6,811
3	25,874	1,000	347	34.7	5	269	7,255
3	25,874	1,000	367	36.7	5	290	7,873
3	25,874	1,000	388	38.8	5	314	8,592
3	25,874	1,000	409	40.9	5	338	9,134
3	25,874	1,000	434	43.4	5	359	9,673
3	51,663	2,000	536	26.8	5	359	9,673
3	124,062	5,000	734	14.7	5	359	9,673
3	251,103	10,000	985	9.9	5	359	9,673
3	505,619	20,000	1,298	6.5	7	359	9,673

Higher-Dimensional Varieties

Clay Mathematics Proceedings
Volume 8, 2009

Algebraic varieties with many rational points

Yuri Tschinkel

Abstract. We survey rational points on higher-dimensional algebraic varieties, addressing questions about existence, density, and distribution with respect to heights. Key examples for existence and density problems include hypersurfaces, complete intersections, and K3 surfaces. For varieties closely related to linear algebraic groups, e.g., equivariant compactifications of groups and homogeneous spaces, questions concerning the asymptotic distribution of points of bounded height are closely related to adelic harmonic analysis on the groups. On the other hand, analytic techniques lead naturally to investigations of global geometric invariants of the underlying varieties, studied in the context of the minimal model program.

Contents

Introduction

Let $f \in \mathbb{Z}[t, x_1, \dots, x_n]$ be a polynomial with coefficients in the integers. Consider

$$f(t, x_1, \dots, x_n) = 0,$$

as an equation in the unknowns $t, x_1, \dots, x_n$ or as an algebraic family of equations in $x_1, \dots, x_n$ parametrized by t. We are interested in integer solutions: their existence

2000 *Mathematics Subject Classification.* Primary 14G05, Secondary 11G35, 11G50.
The author was supported by the NSF grant 0602333.

and distribution. Sometimes the emphasis is on individual equations, e.g.,

$$x^n + y^n = z^n,$$

sometimes we want to understand a *typical* equation, i.e., a general equation in some family. To draw inspiration (and techniques) from different branches of algebra it is necessary to consider solutions with values in other rings and fields, most importantly, finite fields $\mathbb{F}_q$, finite extensions of $\mathbb{Q}$, or the function fields $\mathbb{F}_p(t)$ and $\mathbb{C}(t)$. While there is a wealth of *ad hoc* elementary tricks to deal with individual equations, and deep theories focusing on their visible or hidden symmetries, our primary approach here will be via geometry.

Basic geometric objects are the affine space $\mathbb{A}^n$ and the projective space $\mathbb{P}^n = \left(\mathbb{A}^{n+1} \setminus 0\right)/\mathbb{G}_m$, the quotient by the diagonal action of the multiplicative group. Concretely, affine algebraic varieties $X^{\rm aff} \subset \mathbb{A}^n$ are defined by systems of polynomial equations with coefficients in some base ring R; their solutions with values in R, $X^{\rm aff}(R)$, are called R-integral points. Projective varieties are defined by homogeneous equations, and $X^{\rm proj}(R) = X^{\rm proj}(F)$, the F-rational points on $X^{\rm proj}$, where F is the fraction field of R. The geometric advantages of working with "compact" projective varieties translate to important technical advantages in the study of equations, and the theory of rational points is currently much better developed.

The sets of rational points $X(F)$ reflect on the one hand the geometric and algebraic complexity of X (e.g., the dimension of X), and on the other hand the structure of the ground field F (e.g., its topology, analytic structure). It is important to consider the variation of $X(F')$, as F' runs over extensions of F, either algebraic extensions or completions. It is also important to study projective and birational invariants of X, its birational models, automorphisms, fibration structures, deformations. Each point of view contributes its own set of techniques, and it is the interaction of ideas from a diverse set of mathematical cultures that makes the subject so appealing and vibrant.

The focus in these notes will be on smooth projective varieties X defined over $\mathbb{Q}$, with *many* $\mathbb{Q}$-rational points. Main examples are varieties $\mathbb{Q}$-birational to $\mathbb{P}^n$ and hypersurfaces in $\mathbb{P}^n$ of low degree. We will study the relationship between the global *geometry* of X over $\mathbb{C}$ and the distribution of rational points in the Zariski topology and with respect to *heights*. Here are the problems we face:

- Existence of solutions: local obstructions, the Hasse principle, global obstructions;
- Density in various topologies: Zariski density, weak approximation;
- Distribution with respect to heights: bounds on smallest points, asymptotics.

Here is the road map of the paper. Section 1 contains a summary of basic terms from complex algebraic geometry: main invariants of algebraic varieties, classification schemes, and examples most relevant to arithmetic in dimension ≥ 2. Section 2 is devoted to the existence of rational and integral points, including aspects of decidability, effectivity, local and global obstructions. In Section 3 we discuss Lang's conjecture and its converse, focusing on varieties with nontrivial

endomorphisms and fibration structures. Section 4 introduces heights, counting functions, and height zeta functions. We explain conjectures of Batyrev, Manin, Peyre and their refinements. The remaining sections are devoted to geometric and analytic techniques employed in the proof of these conjectures: universal torsors, harmonic analysis on adelic groups, p-adic integration and "estimates".

Acknowledgments. I am very grateful to V. Batyrev, F. Bogomolov, U. Derenthal, A. Chambert-Loir, J. Franke, J. Harris, B. Hassett, A. Kresch, Yu. Manin, E. Peyre, J. Shalika, M. Strauch and R. Takloo-Bighash for the many hours of listening and sharing their ideas. Partial support was provided by National Science Foundation Grants 0554280 and 0602333.

1. Geometry background

We discuss basic notions and techniques of algebraic geometry that are commonly encountered by number theorists. For most of this section, F is an algebraically closed field of characteristic zero. Geometry over algebraically closed fields of positive characteristic, e.g., algebraic closure of a finite field, differs in several aspects: difficulties arising from inseparable morphisms, "unexpected" maps between algebraic varieties, additional symmetries, lack (at present) of *resolution of singularities*. Geometry over nonclosed fields, especially number fields, introduces new phenomena: varieties may have *forms*, not all constructions *descend* to the ground field, parameter counts do not suffice. In practice, it is "equivariant geometry for finite groups", with Galois symmetries acting on all geometric invariants and special loci. The case of surfaces is addressed in detail in [**Has**].

1.1. Basic invariants. Let X be a smooth projective algebraic variety over F. Over ground fields of characteristic zero we can pass to a resolution of singularities and replace any algebraic variety by a smooth projective model. We seek to isolate invariants of X that are most relevant for arithmetic investigations.

There are two natural types of invariants: *birational* invariants, i.e., invariants of the function field $F(X)$, and *projective geometry* invariants, i.e., those arising from a concrete representation of X as a subvariety of $\mathbb{P}^n$. Examples are the *dimension* $\dim(X)$, defined as the transcendence degree of $F(X)$ over F, and the *degree* of X in the given projective embedding. For hypersurfaces $X_f \subset \mathbb{P}^n$ the degree is simply the degree of the defining homogeneous polynomial. In general, it is defined via the Hilbert function of the homogeneous ideal, or geometrically, as the number of intersection points with a general hyperplane of codimension $\dim(X)$.

The degree alone is not a sensitive indicator of the complexity of the variety: Veronese embeddings of $\mathbb{P}^1 \hookrightarrow \mathbb{P}^n$ exhibit it as a curve of degree n. In general, we may want to consider all possible projective embeddings of a variety X. Two such embeddings can be "composed" via the Segre embedding $\mathbb{P}^n \times \mathbb{P}^m \to \mathbb{P}^N$, where $N = nm+n+m$. For example, we have the standard embedding $\mathbb{P}^1 \times \mathbb{P}^1 \hookrightarrow \mathbb{P}^3$, with image a smooth quadric. In this way, projective embeddings of X form a "monoid"; the corresponding abelian group is the *Picard group* $\mathrm{Pic}(X)$. Alternatively, it is the group of isomorphism classes of *line bundles* on X. Cohomologically, in the Zariski (or étale) topology,

$$\mathrm{Pic}(X) = \mathrm{H}^1(X, \mathbb{G}_m),$$

where $\mathbb{G}_m$ is the sheaf of invertible functions. Yet another description is

$$\mathrm{Pic}(X) = \mathrm{Div}(X)/\left(\mathbb{C}(X)^*/\mathbb{C}^*\right),$$

where $\mathrm{Div}(X)$ is the free abelian group generated by codimension one subvarieties of X, and $\mathbb{C}(X)^*$ is the multiplicative group of rational functions of X, each $f \in \mathbb{C}(X)^*$ giving rise to a *principal* divisor $\mathrm{div}(f)$ (divisor of zeroes and poles of f). Sometimes it is convenient to identify divisors with their classes in $\mathrm{Pic}(X)$. Note that Pic is a *contravariant functor*: a morphism $\tilde{X} \to X$ induces a homomorphism of abelian groups $\mathrm{Pic}(X) \to \mathrm{Pic}(\tilde{X})$. There is an exact sequence

$$1 \to \mathrm{Pic}^0(X) \to \mathrm{Pic}(X) \to \mathrm{NS}(X) \to 1,$$

where the subgroup $\mathrm{Pic}^0(X)$ can be endowed with the structure of a connected projective algebraic variety. The finitely generated group $\mathrm{NS}(X)$ is called the *Néron-Severi* group of X. In most applications in this paper, $\mathrm{Pic}^0(X)$ is trivial.

Given a projective variety $X \subset \mathbb{P}^n$, via an explicit system of homogeneous equations, we can easily write down at least one divisor on X, a hyperplane section L in this embedding. Another divisor, the divisor of zeroes of a differential form of top degree on X, can also be computed from the equations. Its class $K_X \in \mathrm{Pic}(X)$, i.e., the class of the line bundle $\Omega_X^{\dim(X)}$, is called the *canonical class*. In general, it is not known how to write down effectively divisors whose classes are not proportional to linear combinations of K_X and L. This can be done for some varieties over $\mathbb{Q}$, e.g., smooth cubic surfaces in $X_3 \subset \mathbb{P}^3$ (see Section 1.9), but already for smooth quartics $X_4 \subset \mathbb{P}^3$ it is unknown how to compute even the rank of $\mathrm{NS}(X)$ (for some partial results in this direction, see Section 1.10).

Elements in $\mathrm{Pic}(X)$ corresponding to projective embeddings generate the *ample* cone $\Lambda_{\mathrm{ample}}(X) \subset \mathrm{Pic}(X)_{\mathbb{R}} := \mathrm{Pic}(X) \otimes \mathbb{R}$; ample divisors arise as hyperplane sections of X in a projective embedding. The closure $\Lambda_{\mathrm{nef}}(X)$ of $\Lambda_{\mathrm{ample}}(X)$ in $\mathrm{Pic}(X)_{\mathbb{R}}$ is called the *nef* cone. An *effective divisor* is a sum with nonnegative coefficients of irreducible subvarieties of codimension one. Their classes span the *effective cone* $\Lambda_{\mathrm{eff}}(X)$; divisors arising as hyperplane sections of projective embeddings of some Zariski open subset of X form the interior of $\Lambda_{\mathrm{eff}}(X)$. These cones and their combinatorial structure encode important geometric information. For example, for all divisors $D \in \Lambda_{\mathrm{nef}}(X)$ and all curves $C \subset X$, the intersection number $D \cdot C \geq 0$ [**Kle66**]. Divisors on the boundary of $\Lambda_{\mathrm{ample}}(X)$ give rise to fibration structures on X; we will discuss this in more detail in Section 1.4.

Example 1.1.1. Let X be a smooth projective variety, $Y \subset X$ a smooth subvariety and $\pi : \tilde{X} = \mathrm{Bl}_Y(X) \to X$ the *blowup* of X in Y, i.e., the complement in $\tilde{X}$ of the *exceptional divisor* $E := \pi^{-1}(Y)$ is isomorphic to $X \setminus Y$, and E itself can be identified with the projectivized normal cone to X at Y. Then

$$\mathrm{Pic}(\tilde{X}) \simeq \mathrm{Pic}(X) \oplus \mathbb{Z}E$$

and

$$K_{\tilde{X}} = \pi^*(K_X) + \mathcal{O}((\mathrm{codim}(Y) - 1)E)$$

(see [**Har77**, Exercise 8.5]). Note that

$$\pi^*(\Lambda_{\mathrm{eff}}(X)) \subset \Lambda_{\mathrm{eff}}(\tilde{X}),$$

but that pullbacks of ample divisors are not necessarily ample.

Example 1.1.2. Let $X \subset \mathbb{P}^n$ be a smooth hypersurface of dimension ≥ 3 and degree d. Then $\mathrm{Pic}(X) = \mathrm{NS}(X) = \mathbb{Z}L$, generated by the class of the hyperplane section, and

$$\Lambda_{\mathrm{ample}}(X) = \Lambda_{\mathrm{eff}}(X) = \mathbb{R}_{\geq 0}L.$$

The canonical class is

$$K_X = -(n+1-d)L.$$

Example 1.1.3. If X is a smooth cubic surface over an algebraically closed field, then $\mathrm{Pic}(X) = \mathbb{Z}^7$. The anticanonical class is proportional to the sum of 27 exceptional curves (lines):

$$-K_X = \frac{1}{9}(D_1 + \cdots + D_{27}).$$

The effective cone $\Lambda_{\mathrm{eff}}(X) \subset \mathrm{Pic}(X)_{\mathbb{R}}$ is spanned by the classes of the lines.

On the other hand, the effective cone of a minimal resolution of the singular cubic surface

$$x_0x_3^2 + x_1^2x_3 + x_2^3 = 0$$

is a simplicial cone (in $\mathbb{R}^7$) [**HT04**].

Example 1.1.4. Let G be a connected solvable linear algebraic group, e.g., the additive group $G = \mathbb{G}_a$, an algebraic torus $G = \mathbb{G}_m^d$ or the group of upper triangular matrices. Let X be an equivariant compactification of G, i.e., there is a morphism $G \times X \to X$ extending the action $G \times G \to G$ of G on itself. Using equivariant resolution of singularities, if necessary, we may assume that X is smooth projective and that the boundary

$$X \setminus G = D = \bigcup_{i \in \mathcal{I}} D_i, \quad \text{with} \quad D_i \text{ irreducible},$$

is a divisor with normal crossings. Every divisor D on X is equivalent to a divisor with support in the boundary since it can be "moved" there by the action of G (see e.g. [**CLT02**, Proposition 1.1]. Thus $\mathrm{Pic}(X)$ is generated by the components D_i, and the relations are given by functions with zeroes and poles supported in D, i.e., by the characters $\mathfrak{X}^*(G)$. We have an exact sequence

$$0 \to \mathfrak{X}^*(G) \to \bigoplus_{i \in \mathcal{I}} \mathbb{Z}D_i \xrightarrow{\pi} \mathrm{Pic}(X) \to 0 \tag{1.1}$$

The cone of effective divisors $\Lambda_{\mathrm{eff}}(X) \subset \mathrm{Pic}(X)_{\mathbb{R}}$ is the image of the simplicial cone $\bigoplus_{i \in \mathcal{I}} \mathbb{R}_{\geq 0}D_i$ under the projection π. The anticanonical class is

$$-K_X = \bigoplus_{i \in \mathcal{I}} \kappa_i D_i, \quad \text{with} \quad \kappa_i \geq 1, \quad \text{for all } i.$$

For unipotent G one has $\kappa_i \geq 2$, for all i [**HT99**].

For higher-dimensional varieties without extra symmetries, the computation of the ample and effective cones, and of the position of K_X with respect to these cones, is a difficult problem. A sample of recent papers on this subject is: [**CS06**], [**Far06**], [**FG03**], [**Cas07**], [**HT03**], [**HT02a**], [**GKM02**]. However, we have the following fundamental result (see also Section 1.4):

THEOREM 1.1.5. *Let X be a smooth projective variety with $-K_X \in \Lambda_{\mathrm{ample}}(X)$. Then $\Lambda_{\mathrm{nef}}(X)$ is a finitely generated rational cone. If $-K_X$ is big and nef then $\Lambda_{\mathrm{eff}}(X)$ is finitely generated.*

Finite generation of the nef cone goes back to Mori (see [**CKM88**] for an introduction). The result concerning $\Lambda_{\mathrm{eff}}(X)$ has been proved in [**Bat92**] in dimension ≤ 3, and in higher dimensions in [**BCHM06**] (see also [**Ara05**], [**Leh08**]).

1.2. Classification schemes. In some arithmetic investigations (e.g., Zariski density or rational points) we rely mostly on birational properties of X; in others (e.g., asymptotics of points of bounded height), we need to work in a fixed projective embedding.

Among *birational* invariants, the most important are those arising from a comparison of X with a projective space:

(1) *rationality*: there exists a birational isomorphism $X \sim \mathbb{P}^n$, i.e., an isomorphism of function fields $F(X) = F(\mathbb{P}^n)$, for some $n \in \mathbb{N}$;
(2) *unirationality*: there exists a dominant map $\mathbb{P}^n \dashrightarrow X$;
(3) *rational connectedness*: for general $x_1, x_2 \in X(F)$ there exists a morphism $f : \mathbb{P}^1 \to X$ such that $x_1, x_2 \in f(\mathbb{P}^1)$.

It is easy to see that

$$(1) \Rightarrow (2) \Rightarrow (3).$$

Over algebraically closed ground fields, these properties are equivalent in dimension two, but (may) diverge in higher dimensions: there are examples with (1) $\neq$ (2) but so far no examples with (2) $\neq$ (3). Finer classifications result when the ground field F is not assumed to be algebraically closed, e.g., there exist unirational but not rational cubic surfaces over nonclosed fields. The first unirational but not rational threefolds over $\mathbb{C}$ were constructed in [**IM71**] and [**CG72**]. The approach of [**IM71**] was to study the group $\mathrm{Bir}(X)$ of birational automorphisms of X; finiteness of $\mathrm{Bir}(X)$, i.e., *birational rigidity*, implies nonrationality.

Interesting unirational varieties arise as quotients V/G, where $V = \mathbb{A}^n$ is a representation space for a faithful action of a linear algebraic group G. For example, the moduli space $\mathcal{M}_{0,n}$ of n points on $\mathbb{P}^1$ is birational to $(\mathbb{P}^1)^n/\mathrm{PGL}_2$. Moduli spaces of degree d hypersurfaces $X \subset \mathbb{P}^n$ are naturally isomorphic to $\mathbb{P}(\mathrm{Sym}^d(\mathbb{A}^{n+1}))/\mathrm{PGL}_{n+1}$. Rationality of V/G is known as *Noether's problem*. It has a positive solution for G being the symmetric group $\mathfrak{S}_n$, the group PGL_2 [**Kat82**], [**BK85**], and in many other cases [**SB89**], [**SB88**]. Counterexamples for some *finite* G were constructed in [**Sal84**], [**Bog87**]; nonrationality is detected by the *unramified Brauer group*, $\mathrm{Br}_{\mathrm{un}}(V/G)$, closely related to the Brauer group of the function field $\mathrm{Br}(F(V/G)) = \mathrm{H}^2_{\text{ét}}(F(V/G), \mathbb{G}_m)$.

Now we turn to invariants arising from projective geometry, i.e., ample line bundles on X. For smooth curves C, an important invariant is the *genus* $\mathrm{g}(C) := \dim(\mathrm{H}^0(C, K_C))$. In higher dimensions, one considers the *Kodaira dimension*

$$\kappa(X) := \limsup \frac{\log(\dim(\mathrm{H}^0(X, nK_X)))}{\log(n)}, \tag{1.2}$$

and the related graded *canonical* section ring

$$R(X, K_X) = \bigoplus_{n \geq 0} \mathrm{H}^0(X, nK_X). \tag{1.3}$$

A fundamental theorem is that this ring is finitely generated [**BCHM06**].

A very rough classification of smooth algebraic varieties is based on the position of the anticanonical class with respect to the cone of ample divisors. Numerically, this is reflected in the value of the Kodaira dimension. There are three main cases:

- *Fano:* $-K_X$ ample, with $\kappa(X) = -\infty$;
- *general type:* K_X ample, $\kappa(X) = \dim(X)$;
- *intermediate type*, e.g., $\kappa(X) = 0$.

The qualitative behavior of rational points on X mirrors this classification (see Section 3). In our arithmetic applications we will mostly encounter Fano varieties and varieties of intermediate type.

For curves, this classification can be read off from the genus: curves of genus 0 are of Fano type, of genus 1 of intermediate type, and of genus ≥ 2 of general type. Other examples of varieties in each group are:

- Fano: $\mathbb{P}^n$, smooth degree d hypersurfaces $X_d \subset \mathbb{P}^n$, with $d \leq n$;
- general type: hypersurfaces $X_d \subset \mathbb{P}^n$, with $d \geq n+2$, moduli spaces of curves of high genus and abelian varieties of high dimension;
- intermediate type: abelian varieties, Calabi-Yau varieties.

There are only finitely many families of smooth Fano varieties in each dimension [**KMM92**]. On the other hand, the universe of varieties of general type is boundless and there are many open classification questions already in dimension 2.

In finer classification schemes such as the *Minimal Model Program* (MMP) it is important to take into account fibration structures and mild singularities (see [**KMM87**] and [**Cam04**]). Analogously, in many arithmetic questions, the passage to fibrations is inevitable (see Section 4.14). These often arise from the section rings

$$R(X, L) = \bigoplus_{n \geq 0} \mathrm{H}^0(X, nL). \tag{1.4}$$

Consequently, one considers the *Iitaka dimension*

$$\kappa(X, L) := \limsup \frac{\log(\dim(\mathrm{H}^0(X, nL)))}{\log(n)}. \tag{1.5}$$

Finally, a pair (X, D), where X is smooth projective and D is an effective divisor in X, gives rise to another set of invariants: the *log Kodaira dimension* $\kappa(X, K_X + D)$ and the *log canonical ring* $R(X, K_X + D)$, whose finite generation is also known in many cases [**BCHM06**]. Again, one distinguishes

- *log Fano:* $\kappa(X, -(K_X + D)) = \dim(X)$;
- *log general type:* $\kappa(X, K_X + D) = \dim(X)$;
- *log intermediate type:* none of the above.

This classification has consequences for the study of *integral* points on the open variety $X \setminus D$.

1.3. Singularities. Assume that X is normal and $\mathbb{Q}$-Cartier, i.e., there exists an integer m such that mK_X is a Cartier divisor. Let $\tilde{X}$ be a normal variety and $f : \tilde{X} \to X$ a proper birational morphism. Denote by E an irreducible f-exceptional divisor and by e its generic point. Let $g = 0$ be a local equation of E. Locally, we can write

$$f^*(\text{generator of } \mathcal{O}(mK_X)) = g^{md(E)}(dy_1 \wedge \cdots \wedge dy_n)^m$$

for some $d(E) \in \mathbb{Q}$ such that $md(E) \in \mathbb{Z}$. If, in addition, $K_{\tilde{X}}$ is a line bundle (e.g., $\tilde{X}$ is smooth), then $mK_{\tilde{X}}$ is linearly equivalent to

$$f^*(mK_X) + \sum_i m \cdot d(E_i)E_i; \quad E_i \text{ exceptional},$$

and numerically

$$K_{\tilde{X}} \sim f^*(K_X) + \sum_i d(E_i)E_i.$$

The number $d(E)$ is called the *discrepancy* of X at the exceptional divisor E. The discrepancy $\mathrm{discr}(X)$ of X is

$$\mathrm{discr}(X) := \inf\{d(E) \,|\, \text{all } f, E\}$$

If X is smooth then $\mathrm{discr}(X) = 1$. In general (see e.g., [**Kol92**, Proposition 1.9]),

$$\mathrm{discr}(X) \in \{-\infty\} \cup [-1, 1].$$

Definition 1.3.1. The singularities of X are called

- *terminal* if $\mathrm{discr}(X) > 0$ and
- *canonical* if $\mathrm{discr}(X) \geq 0$.

It is essential to remember that *terminal* $\Rightarrow$ smooth in codimension 2 and that for surfaces, *canonical* means *Du Val* singularities.

Canonical isolated singularities on surfaces are classified via Dynkin diagrams: Let $f : \tilde{X} \to X$ be the *minimal* desingularization. Then the submodule in $\mathrm{Pic}(\tilde{X})$ spanned by the classes of exceptional curves (with the restriction of the intersection form) is isomorphic to the root lattice of the corresponding Dynkin diagram (exceptional curves give simple roots).

On surfaces, canonical singularities don't influence the expected asymptotics for rational points on the complement to all exceptional curves: for (singular) Del Pezzo surfaces X we still expect an asymptotic of points of bounded anticanonical height of the shape $\mathsf{B}\log(\mathsf{B})^{9-d}$, where d is the degree of X, just like in the smooth case (see Section 4.10). This fails when the singularities are worse than canonical.

Example 1.3.2. Let $w = (w_0, \ldots, w_n) \in \mathbb{N}^n$, with $\gcd(w_0, \ldots, w_n) = 1$ and let

$$X = X(w) = \mathbb{P}(w_0, \ldots, w_n)$$

be a *weighted projective space*, i.e., we have a quotient map

$$(\mathbb{A}^{n+1} \setminus 0) \xrightarrow{\mathbb{G}_m} X,$$

where the torus $\mathbb{G}_m$ acts by

$$\lambda \cdot (x_0, \ldots, x_{n+1}) \mapsto (\lambda^{w_0} x_0, \ldots, \lambda^{w_n} x_n).$$

For $w = (1, \dots, 1)$ it is the usual projective space, e.g., $\mathbb{P}^2 = \mathbb{P}(1,1,1)$. The weighted projective plane $\mathbb{P}(1,1,2)$ has a canonical singularity and the singularity of $\mathbb{P}(1,1,m)$, with $m \geq 3$, is worse than canonical.

For a discussion of singularities on general weighted projective spaces and so called *fake* weighted projective spaces see, e.g., [**Kas08**].

1.4. Minimal Model Program. Here we recall basic notions from the Minimal Model Program (MMP) (see [**CKM88**], [**KM98**], [**KMM87**], [**Mat02**] for more details). The starting point is the following fundamental theorem due to Mori [**Mor82**]:

THEOREM 1.4.1. *Let X be a smooth Fano variety of dimension n. Then through every geometric point of X there passes a rational curve of $-K_X$-degree $\leq n+1$.*

These rational curves move in families. Their specializations are rational curves, which may move again, and again, until one arrives at "rigid" rational curves.

THEOREM 1.4.2 (Cone theorem). *Let X be a smooth Fano variety. Then the closure of the cone of (equivalence classes of) effective curves in $\mathrm{H}_2(X,\mathbb{R})$ is finitely generated by classes of rational curves.*

The generating rational curves are called *extremal rays*; they correspond to codimension-1 faces of the dual cone of nef divisors. Mori's Minimal Model Program links the convex geometry of the nef cone $\Lambda_{\mathrm{nef}}(X)$ with birational transformations of X. Pick a divisor D on the face dual to an extremal ray $[C]$. It is not ample anymore, but it still defines a map

$$X \to \mathrm{Proj}(R(X,D)),$$

which contracts the curve C to a point. The map is one of the following:

- a fibration over a base of smaller dimension, and the restriction of D to a general fiber proportional to the anticanonical class of the fiber, which is a (possibly singular) Fano variety,
- a birational map contracting a divisor,
- a contraction of a subvariety in codimension ≥ 2 (a *small* contraction).

The image could be singular, as in Example 1.3.2, and one of the most difficult issues of MMP was to develop a framework which allows one to maneuver between birational models with singularities in a restricted class, while keeping track of the modifications of the Mori cone of curves. In arithmetic applications, for example proofs of the existence of rational points as in, e.g., [**CTSSD87a**], [**CTSSD87b**], [**CTS89**], one relies on the *fibration method and descent*, applied to some auxiliary varieties. Finding the "right" fibration is an art. Mori's theory gives a systematic approach to these questions.

A variant of Mori's theory, the *Fujita program*, analyzes fibrations arising from divisors on the boundary of the *effective* cone $\Lambda_{\mathrm{eff}}(X)$. This theory turns up in the analysis of height zeta functions in Section 6 (see also Section 4.13).

Let X be smooth projective with $\mathrm{Pic}(X) = \mathrm{NS}(X)$ and a finitely generated effective cone $\Lambda_{\mathrm{eff}}(X)$. For a line bundle L on X define

$$a(L) := \min(a \,|\, aL + K_X \in \Lambda_{\mathrm{eff}}(X)). \tag{1.6}$$

We will also need the notion of the *geometric hypersurface of linear growth*:

$$\Sigma_X^{\mathrm{geom}} := \{L \in \mathrm{NS}(X)_{\mathbb{R}} \,|\, a(L) = 1\} \tag{1.7}$$

Let $b(L)$ be the maximal codimension of a face of $\Lambda_{\mathrm{eff}}(X)$ containing $a(L)L + K_X$. In particular,

$$a(-K_X) = 1 \quad \text{and} \quad b(-K_X) = \mathrm{rk}\,\mathrm{Pic}(X).$$

These invariants arise in Manin's conjecture in Section 4.10 and the analysis of analytic properties of height zeta functions in Section 6.1.

1.5. Campana's program. Recently, Campana developed a new approach to classification of algebraic varieties with the goal of formulating necessary and sufficient conditions for *potential density* of rational points, i.e., Zariski density after a finite extension of the ground field. The key notions are: the *core* of an algebraic variety and *special* varieties. Special varieties include Fano varieties and Calabi–Yau varieties. They are conjectured to have a potentially dense set of rational points. This program is explained in [**Abr**].

1.6. Cox rings. Again, we assume that X is a smooth projective variety with $\mathrm{Pic}(X) = \mathrm{NS}(X)$. Examples are Fano varieties, equivariant compactifications of algebraic groups and holomorphic symplectic varieties. Fix line bundles $L_1, \dots, L_r$ whose classes generate $\mathrm{Pic}(X)$. For $\nu = (\nu_1, \dots, \nu_r) \in \mathbb{Z}^r$ we put

$$L^\nu := L_1^{\nu_1} \otimes \cdots \otimes L_r^{\nu_r}.$$

The *Cox* ring is the multigraded section ring

$$\mathrm{Cox}(X) := \bigoplus_{\nu \in \mathbb{Z}^r} \mathrm{H}^0(X, L^\nu).$$

The nonzero graded pieces of $\mathrm{Cox}(X)$ are in bijection with effective divisors of X. The key issue is finite generation of this ring. This has been proved under quite general assumptions in [**BCHM06**, Corollary 1.1.9]. Assume that $\mathrm{Cox}(X)$ is finitely generated. Then both $\Lambda_{\mathrm{eff}}(X)$ and $\Lambda_{\mathrm{nef}}(X)$ are finitely generated polyhedral (see [**HK00**, Proposition 2.9]). Other important facts are:

- X is a toric variety if and only if $\mathrm{Cox}(X)$ is a polynomial ring [**Cox95**], [**HK00**, Corollary 2.10]; Cox rings of some equivariant compactifications of other semi-simple groups are computed in [**Bri07**];
- $\mathrm{Cox}(X)$ is multigraded for $\mathrm{NS}(X)$, in particular, it carries a natural action of the dual torus T_{NS} (see Section 6.6 for details on the duality between lattices and algebraic tori).

1.7. Universal torsors. We continue to work over an algebraically closed field. Let G be a linear algebraic group and X an algebraic variety. A G-torsor over X is a principal G-bundle $\pi : \mathcal{T}_X \to X$. Basic examples are GL_n-torsors; they arise from vector bundles over X. For instance, each line bundle L gives rise to a $\mathrm{GL}_1 = \mathbb{G}_m$-torsor over X. Up to isomorphism, G-torsors are classified by $\mathrm{H}^1_{\text{ét}}(X, G)$; line bundles are classified by $\mathrm{H}^1_{\text{ét}}(X, \mathbb{G}_m) = \mathrm{Pic}(X)$. When G is commutative, $\mathrm{H}^1_{\text{ét}}(X, G)$ is a group.

Let $G = \mathbb{G}_m^r$ be an algebraic torus and $\mathfrak{X}^*(G) = \mathbb{Z}^r$ its character lattice. A G-torsor over an algebraic variety X is determined, up to isomorphism, by a homomorphism

$$\chi : \mathfrak{X}^*(G) \to \mathrm{Pic}(X). \tag{1.8}$$

Assume that $\mathrm{Pic}(X) = \mathrm{NS}(X) = \mathbb{Z}^r$ and that χ is in fact an isomorphism. The arising G-torsors are called *universal*. The introduction of universal torsors is motivated by the fact that over *nonclosed* fields they "untwist" the action of the Galois group on the Picard group of X (see Sections 1.13 and 2.5). The "extra dimensions" and "extra symmetries" provided by the torsor add crucial freedom in the analysis of the geometry and arithmetic of the underlying variety. Examples of applications to rational points will be presented in Sections 2.5 and 5. This explains the surge of interest in explicit equations for universal torsors, the study of their geometry: singularities and fibration structures.

Assume that $\mathrm{Cox}(X)$ is finitely generated. Then $\mathrm{Spec}\,(\mathrm{Cox}(X))$ contains a universal torsor $\mathcal{T}_X$ of X as an open subset. More precisely, let

$$\overline{\mathcal{T}}_X := \mathrm{Spec}\,(\mathrm{Cox}(X)).$$

Fix an ample class $L^\nu \in \mathrm{Pic}(X)$ and let $\chi_\nu \in \mathfrak{X}^*(T_{\mathrm{NS}})$ be the corresponding character. Then

$$X = \mathrm{Proj}(\bigoplus_{n\geq 0} \mathrm{H}^0(X, \mathcal{O}(nL^\nu))) = \overline{\mathcal{T}}_X /\!\!/ T_{\mathrm{NS}},$$

the geometric invariant theory quotient linearized by χ_ν. The *unstable locus* is

$$Z_\nu := \{t \in \overline{\mathcal{T}}_X \mid f(t) = 0 \ \ \forall f \in \mathrm{Cox}(X)_{n\nu}, \ \ n > 0\}$$

Let W_ν be the set of $t \in \overline{\mathcal{T}_X}$ such that the orbit of t is not closed in $\overline{\mathcal{T}_X} \setminus Z_\nu$, or such that t has a positive-dimensional stabilizer. Geometric invariant theory implies that

$$\overline{\mathcal{T}}_X \setminus W_\nu =: \mathcal{T}_X \to X$$

is a geometric quotient, i.e., $\mathcal{T}_X$ is a T_{NS}-torsor over X.

1.8. Hypersurfaces. We now turn from the general theory to specific varieties. Let $X = X_f \subset \mathbb{P}^n$ be a smooth hypersurface of degree d. We have already described some of its invariants in Example 1.1.2, at least when $\dim(X) \geq 3$. In particular, in this case $\mathrm{Pic}(X) \simeq \mathbb{Z}$ and $T_{\mathrm{NS}} = \mathbb{G}_m$. The universal torsor is the hypersurface in $\mathbb{A}^{n+1} \setminus 0$ given by the vanishing of the defining polynomial f.

In dimension two, there are more possibilities. The most interesting cases are $d = 2, 3$, and 4. A quadric X_2 is isomorphic to $\mathbb{P}^1 \times \mathbb{P}^1$ and has Picard group $\mathrm{Pic}(X_2) \simeq \mathbb{Z} \oplus \mathbb{Z}$. A cubic has Picard group of rank 7. These are examples of *Del Pezzo surfaces* discussed in Section 1.9 and extensively in [**Has**]. They are birational to $\mathbb{P}^2$. A smooth quartic $X_4 \subset \mathbb{P}^3$ is an example of a *K3 surface* (see Section 1.10). We have $\mathrm{Pic}(X_4) = \mathbb{Z}^r$, with r between 1 and 20. They are not rational and, in general, do not admit nontrivial fibrations.

Cubic and quartic surfaces have a rich geometric structure, with large "hidden" symmetries. This translates into many intricate arithmetic issues.

1.9. Del Pezzo surfaces. A smooth projective surface X with ample anticanonical class is called a *Del Pezzo surface*. Standard examples are $\mathbb{P}^2$ and $\mathbb{P}^1\times\mathbb{P}^1$. Over algebraically closed ground fields, all other Del Pezzo surfaces X_r are obtained as blowups of $\mathbb{P}^2$ in $r\le 8$ points in general position (e.g., no three on a line, no 6 on a conic). The number $d=9-r$ is the anticanonical *degree* of X_r. Del Pezzo surfaces of low degree admit the following realizations:

- $d=4$: intersection of two quadrics in $\mathbb{P}^4$;
- $d=3$: hypersurface of degree 3 in $\mathbb{P}^3$;
- $d=2$: hypersurface of degree 4 in the weighted projective space $\mathbb{P}(1,1,1,2)$ given by
$$w^2=f_4(x,y,z),\quad \text{with } f \text{ irreducible, } \deg(f_4)=4.$$
- $d=1$: hypersurface of degree 6 in $\mathbb{P}(1,1,2,3)$ given by
$$w^2=t^3+f_4(x,y)t+f_6(x,y),\quad \text{with } \deg(f_i)=i.$$

Visually and mathematically most appealing are, perhaps, the *cubic surfaces* with $d=3$. Note that for $d=1$, the anticanonical linear series has one *base point*, in particular, $X_8(F)\neq\emptyset$, over F, even when F is not algebraically closed.

Let us compute the geometric invariants of a Del Pezzo surface of degree d, expanding Example 1.1.3. Since $\mathrm{Pic}(\mathbb{P}^2)=\mathbb{Z}L$, the hyperplane class, we have
$$\mathrm{Pic}(X_r)=\mathbb{Z}L\oplus\mathbb{Z}E_1\oplus\cdots\oplus\mathbb{Z}E_r,$$
where E_i are the full preimages of the blown-up points. The canonical class is computed as in Example 1.1.1,
$$K_{X_r}=-3L+(E_1+\cdots+E_r).$$
The intersection pairing defines a quadratic form on $\mathrm{Pic}(X_r)$, with $L^2=1$, $L\cdot E_i=0$, $E_i\cdot E_j=0$, for $i\neq j$, and $E_j^2=-1$. Let W_r be the subgroup of $\mathrm{GL}_{r+1}(\mathbb{Z})$ of elements preserving K_{X_r} and the intersection pairing. For $r\ge 2$ there are other classes with square -1, e.g., preimages of lines passing through two points, conics through five points:
$$L-(E_i+E_j),\quad 2L-(E_1+\cdots+E_5),\quad \text{etc.}$$
The classes whose intersection with K_{X_r} is also -1 are called (classes of) *exceptional curves*; these curves are *lines* in the anticanonical embedding. Their number $n(r)$ can be found in the table below. We have
$$-K_{X_r}=c_r\sum_{j=1}^{n(r)}E_j,$$
the sum over all exceptional curves, where $c_r\in\mathbb{Q}$ can be easily computed, e.g., $c_6=1/9$. The effective cone is spanned by the $n(r)$ classes of exceptional curves, and the nef cone is the cone dual to $\Lambda_{\mathrm{eff}}(X_r)$ with respect to the intersection pairing on $\mathrm{Pic}(X_r)$. Put
$$\alpha(X_r):=\mathrm{vol}\left(\Lambda_{\mathrm{nef}}(X_r)\cap\{C\,|\,(-K_{X_r},C)=1\}\right). \tag{1.9}$$
This "volume" of the nef cone has been computed in [**Der07a**]:

r	1	2	3	4	5	6	7	8
$n(r)$	1	3	6	10	16	27	56	240
$\alpha(X_r)$	1/6	1/24	1/72	1/144	1/180	1/120	1/30	1

Given a Del Pezzo surface over a number field, the equations of the lines can be written down explicitly. This is easy for the diagonal cubic surface

$$x_0^3 + x_1^3 + x_2^3 + x_3^3 = 0.$$

Writing

$$x_i^3 + x_j^3 = \prod_{r=1}^{3}(x_i + \zeta_3^r x_j) = x_\ell^3 + x_k^3 = \prod_{r=1}^{3}(x_\ell + \zeta_3^r x_k) = 0,$$

with $i, j, k, l \in [0, \dots, 3]$, and permuting indices we get all 27 lines. In general, equations for the lines can be obtained by solving the corresponding equations on the Grassmannian of lines.

Degenerations of Del Pezzo surfaces are also interesting and important. Typically, they arise as special fibers of fibrations, and their analysis is unavoidable in the theory of *models* over rings such as $\mathbb{Z}$, or $\mathbb{C}[t]$. A classification of singular Del Pezzo surfaces can be found in [**BW79**], [**DP80**]. Models of Del Pezzo surfaces over curves are discussed in [**Cor96**]. Volumes of nef cones of singular Del Pezzo surfaces are computed in [**DJT08**].

We turn to Cox rings of Del Pezzo surfaces. Smooth Del Pezzo surfaces of degree $d \geq 6$ are toric and their Cox rings are polynomial rings on $12 - d$ generators. The generators and relations of the Cox rings of Del Pezzo surfaces have been computed [**BP04**], [**Der06**], [**STV07**], [**TVAV08**], [**SX08**]. For $r \in \{4, 5, 6, 7\}$ the generators are the nonzero sections from exceptional curves and the relations are induced by fibration structures on X_r (rulings). In degree 1 two extra generators are needed, the independent sections of $\mathrm{H}^0(X_8, -K_{X_8})$.

It was known for a long time that the (affine cone over the) Grassmannian $\mathrm{Gr}(2, 5)$ is a universal torsor for the unique (smooth) degree 5 Del Pezzo surface (this was used in [**SD72**] and [**Sko93**] to prove that every Del Pezzo surface of degree 5 has a rational point). Batyrev conjectured that universal torsors of other Del Pezzo surfaces should embed into (affine cones over) other Grassmannians, and this is why:

One of the most remarkable facts of the theory of Del Pezzo surfaces is the "hidden" symmetry of the collection of exceptional curves in the Picard lattice. Indeed, for $r = 3, 4, 5, \dots, 8$, the group W_r is the *Weyl group* of a *root system*:

$$\mathsf{R}_r \in \{\mathsf{A}_1 \times \mathsf{A}_2, \mathsf{A}_4, \mathsf{D}_5, \mathsf{E}_6, \mathsf{E}_7, \mathsf{E}_8\}, \tag{1.10}$$

and the root lattice itself is the orthogonal to K_{X_r} in $\mathrm{Pic}(X_r)$, the *primitive* Picard group. Let G_r be the simply-connected Lie group with the corresponding root system. The embedding $\mathrm{Pic}(X_{r-1}) \hookrightarrow \mathrm{Pic}(X_r)$ induces an embedding of root lattices $\mathsf{R}_{r-1} \hookrightarrow \mathsf{R}_r$, and identifies a unique simple root α_r in the set of simple roots of R_r, as the complement of the simple roots from R_{r-1}. This defines a parabolic subgroup $P_r \subset G_r$ (see Section 6.4). Batyrev's conjecture was that the flag variety $P_r \backslash G_r$ contains a universal torsor of X_r.

Recent work on Cox rings of Del Pezzo surfaces established this *geometric* connection between smooth Del Pezzo surfaces and Lie groups with root systems of the corresponding type: $r = 5$ was treated in [**Pop01**] and $r = 6, 7$ in [**Der07b**], via explicit manipulations with defining equations. The papers [**SS07**] and [**SS08**] give conceptual, representation-theoretic proofs of these results. It would be important to extend this to singular Del Pezzo surfaces.

Example 1.9.1 (Degree four). Here are some examples of singular degree four Del Pezzo surfaces $X = \{Q_0 = 0\} \cap \{Q = 0\} \subset \mathbb{P}^4$, where $Q_0 = x_0x_1 + x_2^2$ and Q is given in the table below. Let $\tilde{X}$ be the minimal desingularization of X. In all cases below the Cox ring is given by

$$\mathrm{Cox}(\tilde{X}) = F[\eta_1, \dots, \eta_9]/(f)$$

with *one* relation f [**Der07b**]. Note that the Cox ring of a smooth degree 4 Del Pezzo surface has 16 generators and 20 relations (see Example 5.3.2).

Singularities	Q	f
$3\mathsf{A}_1$	$x_2(x_1 + x_2) + x_3x_4$	$\eta_4\eta_5 + \eta_1\eta_6\eta_7 + \eta_8\eta_9$
$\mathsf{A}_1 + \mathsf{A}_3$	$x_3^2 + x_4x_2 + x_0^2$	$\eta_6\eta_9 + \eta_7\eta_8 + \eta_1\eta_3\eta_4^2\eta_5^3$
A_3	$x_3^2 + x_4x_2 + (x_0 + x_1)^2$	$\eta_5\eta_9 + \eta_1\eta_4^2\eta_7 + \eta_3\eta_6^2\eta_8$
D_4	$x_3^2 + x_4x_1 + x_0^2$	$\eta_3\eta_5^2\eta_8 + \eta_4\eta_6^2\eta_9 + \eta_2\eta_7^2$
D_5	$x_1x_2 + x_0x_4 + x_3^2$	$\eta_3\eta_7^2 + \eta_2\eta_6^2\eta_9 + \eta_4\eta_5^2\eta_8^2$.

Example 1.9.2 (Cubics). Here are some singular cubic surfaces $X \subset \mathbb{P}^3$, given by the vanishing of the corresponding cubic form:

$4\mathsf{A}_1$	$x_0x_1x_2 + x_1x_2x_3 + x_2x_3x_0 + x_3x_0x_1$
$2\mathsf{A}_1 + \mathsf{A}_2$	$x_0x_1x_2 = x_3^2(x_1 + x_2 + x_3)$
$2\mathsf{A}_1 + \mathsf{A}_3$	$x_0x_1x_2 = x_3^2(x_1 + x_2)$
$\mathsf{A}_1 + 2\mathsf{A}_2$	$x_0x_1x_2 = x_1x_3^2 + x_3^3$
$\mathsf{A}_1 + \mathsf{A}_3$	$x_0x_1x_2 = (x_1 + x_2)(x_3^2 - x_1^2)$
$\mathsf{A}_1 + \mathsf{A}_4$	$x_0x_1x_2 = x_3^2x_2 + x_3x_1^2$
$\mathsf{A}_1 + \mathsf{A}_5$	$x_0x_1x_2 = x_1^3 - x_3^2x_2$
$3\mathsf{A}_2$	$x_0x_1x_2 = x_3^3$
A_4	$x_0x_1x_2 = x_2^3 - x_3x_1^2 + x_3^2x_2$
A_5	$x_3^3 = x_1^3 + x_0x_3^2 - x_2^2x_3$
D_4	$x_1x_2x_3 = x_0(x_1 + x_2 + x_3)^2$
D_5	$x_0x_1^2 + x_1x_3^2 + x_2^2x_3$
E_6	$x_3^3 = x_1(x_1x_0 + x_2^2)$.

Further examples of Cox rings of singular Del Pezzo surfaces can be found in [**Der06**] and [**DT07**]. In practice, most geometric questions are easier for smooth surfaces, while most arithmetic questions turn out to be easier in the singular case. For a survey of arithmetic problems on rational surfaces, see Sections 2.4 and 3.4, as well as [**MT86**].

Example 1.9.3. In some applications, torsors for subtori of T_{NS} are also used. Let X be the diagonal cubic surface

$$x_0^3 + x_1^3 + x_2^3 + x_3^3 = 0.$$

The following equations were derived in [**CTKS87**]:

$$\mathcal{T}_X := \left\{ \begin{array}{rcl} x_{11}x_{12}x_{13} + x_{21}x_{22}x_{23} + x_{31}x_{32}x_{33} & = & 0 \\ x_{11}x_{21}x_{31} + x_{12}x_{22}x_{32} + x_{13}x_{23}x_{33} & = & 0 \end{array} \right\} \subset \mathbb{P}^8.$$

This is a torsor for $G = \mathbb{G}_m^4$.

1.10. K3 surfaces. Let X be a smooth projective surface with trivial canonical class. There are two possibilities: X could be an abelian surface or a K3 surface. In the latter case, X is simply-connected and $\mathrm{h}^1(X, \mathcal{O}_X) = 0$. The Picard group $\mathrm{Pic}(X)$ of a K3 surface X is a torsion-free $\mathbb{Z}$-module of rank ≤ 20 and the intersection form on $\mathrm{Pic}(X)$ is even, i.e., the square of every class is an even integer. K3 surfaces of with polarizations of small degree can be realized as complete intersections in projective space. The most common examples are K3 surfaces of degree 2, given explicitly as double covers $X \to \mathbb{P}^2$ ramified in a curve of degree 6; or quartic surfaces $X \subset \mathbb{P}^3$.

Example 1.10.1. The Fermat quartic

$$x^4 + y^4 + z^4 + w^4 = 0$$

has Picard rank 20 over $\mathbb{Q}(\sqrt{-1})$. The surface X given by

$$xy^3 + yz^3 + zx^3 + w^4 = 0$$

has $\mathrm{Pic}(X_{\mathbb{Q}}) = \mathbb{Z}^{20}$ (see [**Ino78**] for more explicit examples). All such K3 surfaces are classified in [**Sch08**].

The surface

$$w(x^3 + y^3 + z^3 + x^2z + xw^2) = 3x^2y^2 + 4x^2yz + x^2z^2 + xy^2z + xyz^2 + y^2z^2$$

has geometric Picard rank 1, i.e., $\mathrm{Pic}(X_{\bar{\mathbb{Q}}}) = \mathbb{Z}$ [**vL07**].

Other interesting examples arise from abelian surfaces as follows: Let

$$\begin{array}{rccc} \iota : & A & \to & A \\ & a & \mapsto & -a \end{array}$$

be the standard involution. Its fixed points are the 2-torsion points of A. The quotient A/ι has 16 singularities (the images of the fixed points). The minimal resolution of these singularities is a K3 surface, called a *Kummer surface*. There are several other finite group actions on abelian surfaces such that a similar construction results in a K3 surface, a *generalized* Kummer surface (see [**Kat87**]).

The nef cone of a polarized K3 surface (X, g) admits the following characterization: h is ample if and only if $(h, C) > 0$ for each class C with $(g, C) > 0$ and $(C, C) \geq -2$. The *Torelli theorem* implies an intrinsic description of automorphisms: every automorphism of $\mathrm{H}^2(X, \mathbb{Z})$ preserving the intersection pairing and the nef cone arises from an automorphism of X. There is an extensive literature devoted to the classification of possible automorphism groups [**Nik81**], [**Dol08**]. These automorphisms give examples of interesting algebraic dynamical systems

[**McM02**], [**Can01**]; they can be used to propagate rational points and curves [**BT00**], and to define canonical heights [**Sil91**], [**Kaw08**].

1.11. Threefolds. The classification of smooth Fano threefolds was a major achievement completed in the works of Iskovskikh [**Isk79**], [**IP99a**], and Mori–Mukai [**MM86**]. There are more than 100 families. Among them, for example, cubics $X_3 \subset \mathbb{P}^4$, quartics $X_4 \subset \mathbb{P}^4$ or double covers of $W_2 \to \mathbb{P}^3$, ramified in a surface of degree 6. Many of these varieties, including the above examples, are not rational. Unirationality of cubics can be seen directly by projecting from a line on X_3. The nonrationality of cubics was proved in [**CG72**] using *intermediate Jacobians*. Nonrationality of quartics was proved by establishing *birational rigidity*, i.e., showing triviality of the group of birational automorphisms, via an analysis of *maximal* singularities of such maps [**IM71**]. This technique has been substantially generalized in recent years (see [**Isk01**], [**Puk98**], [**Puk07**], [**Che05**]). Some quartic threefolds are also unirational, e.g., the diagonal, Fermat type, quartic

$$\sum_{i=0}^{4} x_i^4 = 0.$$

It is expected that the *general* quartic is *not* unirational. However, it admits an elliptic fibration: fix a line $\ell \in X_4 \subset \mathbb{P}^4$ and consider a plane in $\mathbb{P}^4$ containing this line, the residual plane curve has degree three and genus 1. A general double cover W_2 does not admit an elliptic or abelian fibration, even birationally [**CP07**].

1.12. Holomorphic symplectic varieties. Let X be a smooth projective simply-connected variety. It is called *holomorphic symplectic* if it carries a unique, modulo constants, nondegenerate holomorphic two-form. Typical examples are K3 surfaces X and their Hilbert schemes $X^{[n]}$ of zero-dimensional length-n subschemes. Another example is the variety of lines of a smooth cubic fourfold; it is deformation equivalent to $X^{[2]}$ of a K3 surface [**BD85**].

These varieties are interesting for the following reasons:

- The symplectic forms allows one to define a *quadratic* form on $\mathrm{Pic}(X)$, the Beauville–Bogomolov form. The symmetries of the lattice carry rich geometric information.
- There is a Torelli theorem, connecting the symmetries of the cohomology lattice with symmetries of the variety.
- there is a *conjectural* characterization of the ample cone and of abelian fibration structures, at least in dimension 4 [**HT01**].

Using this structure as a compass, one can find a plethora of examples with (Lagrangian) abelian fibrations over $\mathbb{P}^n$ or with infinite *endomorphisms*, resp. *birational* automorphisms, which are interesting for arithmetic and algebraic dynamics.

1.13. Nonclosed fields. There is a lot to say: F-rationality, F-unirationality, Galois actions on $\mathrm{Pic}(X_{\bar{F}})$, $\mathrm{Br}(X_{\bar{F}})$, algebraic points, special loci, *descent* of Galois-invariant structures to the ground field etc. Here we touch on just one aspect: the effective computation of the Picard group as a Galois module, for Del Pezzo and K3 surfaces.

Let $X = X_r$ be a Del Pezzo surface over F. A *splitting field* is a normal extension of the ground field over which each exceptional curve is defined. The action of the Galois group Γ factors through a subgroup of the group of symmetries of the exceptional curves, i.e., W_r. In our arithmetic applications we need to know

- $\mathrm{Pic}(X)$ as a Galois module, more specifically, the Galois cohomology
$$\mathrm{H}^1(\Gamma, \mathrm{Pic}(X_{\bar{F}})) = \mathrm{Br}(X)/\mathrm{Br}(F);$$
this group is an obstruction to F-rationality, and also a source of obstructions to the Hasse principle and weak approximation (see Section 2.4);
- the effective cone $\Lambda_{\mathrm{eff}}(X_F)$.

For Del Pezzo surfaces, the possible values of $\mathrm{H}^1(\Gamma, \mathrm{Pic}(X_{\bar{F}}))$ have been computed [**SD93**], [**KST89**], [**Ura96**], [**Cor07**]. This information alone does not suffice. The effective Chebotarev theorem [**LO77**] implies that, given equations defining a Del Pezzo surface, the Galois action on the exceptional curves, i.e., the image of the Galois group in the Weyl group W_r, can be computed in principle. The cone $\Lambda_{\mathrm{eff}}(X_F)$ is spanned by the Galois orbits on these curves.

It would be useful to have a `Magma` implementation of an algorithm computing the Galois representation on $\mathrm{Pic}(X)$, for X a Del Pezzo surface over $\mathbb{Q}$.

Example 1.13.1. The Picard group may be smaller over nonclosed fields: for $X/\mathbb{Q}$ given by
$$x_0^3 + x_1^3 + x_2^3 + x_3^3 = 0$$
$\mathrm{Pic}(X_{\mathbb{Q}}) = \mathbb{Z}^4$. It has a basis e_1, e_2, e_3, e_4 such that $\Lambda_{\mathrm{eff}}(X)$ is spanned by
$$e_2, \quad e_3, \quad 3e_1 - 2e_3 - e_4, \quad 2e_1 - e_2 - e_3 - e_4, \quad e_1 - e_4,$$
$$4e_1 - 2e_2 - 2e_3 - e_4, \quad e_1 - e_2, \quad 2e_1 - 2e_2 - e_4, \quad 2e_1 - e_3$$
(see [**PT01**]).

Example 1.13.2 (Maximal Galois action). Let $X/\mathbb{Q}$ be the cubic surface
$$x^3 + 2xy^2 + 11y^3 + 3xz^2 + 5y^2w + 7zw^2 = 0.$$
Then the Galois group acting on the 27 lines is $\mathsf{W}(\mathsf{E}_6)$ [**EJ08a**] (see [**Eke90**], [**Ern94**], [**VAZ08**], and [**Zar08**], for more examples).

No algorithms for computing even the rank, or the geometric rank, of a K3 surface over a number field are known at present. There are infinitely many possibilities for the Galois action on the Picard lattice.

Example 1.13.3. Let X be a K3 surface over a number field $\mathbb{Q}$. Fix a model $\mathcal{X}$ over $\mathbb{Z}$. For primes p of good reduction we have an injection
$$\mathrm{Pic}(X_{\bar{\mathbb{Q}}}) \hookrightarrow \mathrm{Pic}(X_{\bar{\mathbb{F}}_p}).$$
The rank of $\mathrm{Pic}(X_{\bar{\mathbb{F}}_p})$ is always even. In some examples, it can be computed by counting points over $\mathbb{F}_{p^r}$, for several r, and by using the Weil conjectures.

This local information can sometimes be used to determine the rank of $\mathrm{Pic}(X_{\bar{\mathbb{Q}}})$. Let p, q be distinct primes of good reduction such that the corresponding local ranks are ≤ 2 and the discriminants of the lattices $\mathrm{Pic}(X_{\bar{\mathbb{F}}_p})$, $\mathrm{Pic}(X_{\bar{\mathbb{F}}_q})$ do not differ by a square of a rational number. Then the rank of $\mathrm{Pic}(X_{\bar{\mathbb{Q}}})$ equals 1. This idea has been used in [**vL07**].

2. Existence of points

2.1. Projective spaces and their forms. Let F be a field and $\bar{F}$ an algebraic closure of F. A projective space over F has many rational points: they are dense in the Zariski topology and in the adelic topology. Varieties F-birational to a projective space inherit these properties.

Over nonclosed fields F, projective spaces have *forms*, the so-called *Brauer–Severi* varieties. These are isomorphic to $\mathbb{P}^n$ over $\bar{F}$ but not necessarily over F. They can be classified via the nonabelian cohomology group $\mathrm{H}^1(F, \mathrm{Aut}(\mathbb{P}^n))$, where $\mathrm{Aut}(\mathbb{P}^n) = \mathrm{PGL}_{n+1}$ is the group of algebraic automorphisms of $\mathbb{P}^n$. The basic example is a conic $C \subset \mathbb{P}^2$, e.g.,

$$ax^2 + by^2 + cz^2 = 0, \tag{2.1}$$

with a, b, c square-free and pairwise coprime. It is easy to verify solvability of this equation in $\mathbb{R}$ and modulo p, for all primes p. Legendre proved that (2.1) has primitive solutions in $\mathbb{Z}$ if and only if it has nontrivial solutions in $\mathbb{R}$ and modulo powers of p, for all primes p. This is an instance of a local-to-global principle that will be discussed in Section 2.4.

Checking solvability modulo p is a finite problem which gives a finite procedure to verify solvability in $\mathbb{Z}$. Actually, the proof of Legendre's theorem provides effective bounds for the size of the smallest solution, e.g.,

$$\max(|x|, |y|, |z|) \le 2abc$$

(see [**Kne59**] for a sharper bound), which gives another approach to checking solvability—try all $x, y, z \in \mathbb{N}$ subject to the inequality. If $C(\mathbb{Q}) \neq \emptyset$, then the conic is $\mathbb{Q}$-isomorphic to $\mathbb{P}^1$: draw lines through a $\mathbb{Q}$-point in C.

One could also ask about the number $\mathsf{N}(\mathsf{B})$ of triples of nonzero coprime square-free integers

$$(a, b, c) \in \mathbb{Z}^3, \quad \max(|a|, |b|, |c|) \le \mathsf{B}$$

such that Equation (2.1) has a nontrivial solution. It is [**Guo95**]:

$$\mathsf{N}(\mathsf{B}) = \frac{9}{7\,\Gamma(\frac{3}{2})^3} \prod_p \left(1 - \frac{1}{p}\right)^{3/2} \left(1 + \frac{3}{2p}\right) \frac{\mathsf{B}}{\log(\mathsf{B})^{3/2}} (1 + o(1)), \quad \mathsf{B} \to \infty.$$

I am not aware of a conceptual algebro-geometric interpretation of this density.

In general, forms of $\mathbb{P}^n$ over number fields satisfy the local-to-global principle. Moreover, Brauer–Severi varieties with at least one F-rational point are *split* over F, i.e., isomorphic to $\mathbb{P}^n$ over F. It would be useful to have a routine (in `Magma`) that would check efficiently whether or not a Brauer–Severi variety of small dimension over $\mathbb{Q}$, presented by explicit equations, is split, and to find the smallest solution. The frequency of split fibers in families of Brauer–Severi varieties is studied in [**Ser90**].

2.2. Hypersurfaces. Algebraically, the simplest examples of varieties are hypersurfaces, defined by a single homogeneous equation $f(\mathbf{x}) = 0$. Many classical Diophantine problems reduce to the study of rational points on hypersurfaces. Below we give two proofs and one heuristic argument to motivate the idea that hypersurfaces of low degree should have *many* rational points.

THEOREM 2.2.1. [**Che35**], [**War35**] *Let $X = X_f \subset \mathbb{P}^n$ be a hypersurface over a finite field F given by the equation $f(\mathbf{x}) = 0$. If $\deg(f) \leq n$ then $X(F) \neq \emptyset$.*

PROOF. We reproduce a textbook argument [**BS66**], for $F = \mathbb{F}_p$.
Step 1. Consider the δ-function

$$\sum_{x=1}^{p-1} x^d = \begin{cases} -1 & \mod p \quad \text{if } p-1 \mid d \\ 0 & \mod p \quad \text{if } p-1 \nmid d \end{cases}$$

Step 2. Apply it to a (not necessarily homogeneous) polynomial $\phi \in \mathbb{F}_p[x_0, \dots, x_n]$, with $\deg(\phi) \leq n(p-1)$. Then

$$\sum_{x_0,\dots,x_n} \phi(x_0, \dots, x_n) \equiv 0 \mod p.$$

Indeed, for monomials, we have

$$\sum_{x_0,\dots,x_n} x_1^{d_1} \cdots x_n^{d_n} = \prod (\sum x_j^{d_j}), \quad \text{with} \quad d_0 + \cdots + d_n \leq n(p-1).$$

For some j, we have $0 \leq d_j < p-1$.

Step 3. For $\phi(x) = 1 - f(x)^{p-1}$ we have $\deg(\phi) \leq \deg(f) \cdot (p-1)$. Then

$$\mathsf{N}(f) := \#\{x \mid f(x) = 0\} = \sum_{x_0,\dots,x_n} \phi(x) \equiv 0 \mod p,$$

since $\deg(f) \leq n$.

Step 4. The equation $f(\mathbf{x}) = 0$ has a trivial solution. It follows that

$$\mathsf{N}(f) > 1 \quad \text{and} \quad X_f(\mathbb{F}_p) \neq \emptyset.$$

□

A far-reaching generalization is the following theorem.

THEOREM 2.2.2. [**Esn03**] *If X is a smooth Fano variety over a finite field $\mathbb{F}_q$ then*

$$X(\mathbb{F}_q) \neq \emptyset.$$

Now we pass to the case in which $F = \mathbb{Q}$. Given a form $f \in \mathbb{Z}[x_0, \dots, x_n]$, homogeneous of degree d, we ask how many solutions $\mathbf{x} = (x_0, \dots, x_n) \in \mathbb{Z}^{n+1}$ to the equation $f(\mathbf{x}) = 0$ should we expect? Primitive solutions with $\gcd(x_0, \dots, x_n) = 1$, up to diagonal multiplication with ± 1, are in bijection with rational points on the hypersurface $X_f \subset \mathbb{P}^n$. We have $|f(x)| = O(\mathsf{B}^d)$, for $\|x\| := \max_j(|x_j|) \leq \mathsf{B}$. We may argue that f takes values $0, 1, 2, \dots$ with equal probability, so that the probability of $f(\mathbf{x}) = 0$ is B^{-d}. There are B^{n+1} "events" with $\|x\| \leq \mathsf{B}$. In conclusion, we expect B^{n+1-d} solutions with $\|x\| \leq \mathsf{B}$. There are three cases:

- $n + 1 < d$: as $\mathsf{B} \to \infty$ we should have *fewer and fewer* solutions, and, eventually, none!
- $n + 1 - d = 0$: this is the *stable* regime, we get no information in the limit $\mathsf{B} \to \infty$;
- $n + 1 - d > 0$: the expected number of solutions grows.

We will see many instances when this heuristic reasoning fails. However, it is reasonable, as a first approximation, at least when

$$n+1-d \gg 0.$$

Diagonal hypersurfaces have attracted the attention of computational number theorists (see `http://euler.free.fr`). A sample is given below:

Example 2.2.3.

- There are no rational points (with non-zero coordinates) on the Fano 5-fold $x_0^6 = \sum_{j=1}^6 x_j^6$ with height $\leq 2.6 \cdot 10^6$.
- There are 12 (up to signs and permutations) rational points on $x_0^6 = \sum_{j=1}^7 x_j^6$ of height $\leq 10^5$ (with non-zero coordinates).
- The number of rational points (up to signs, permutations and with non-zero coordinates) on the Fano 5-fold $x_0^6 + x_1^6 = \sum_{j=2}^6 x_j^6$ of height $\leq 10^4$ (resp. $2 \cdot 10^4, 3 \cdot 10^4$) is 12 (resp. 33, 57).

Clearly, it is difficult to generate solutions when $n-d$ is small. On the other hand, we have the following theorem:

THEOREM 2.2.4. [**Bir62**] *If* $n \geq (\deg(f)-1) \cdot 2^{\deg(f)}$, *and* f *is smooth then the number* $\mathsf{N}(f, \mathsf{B})$ *of solutions* $\mathbf{x} = (x_i)$ *with* $\max(|x_i|) \leq \mathsf{B}$ *is*

$$\mathsf{N}(f, \mathsf{B}) = \prod_p \tau_p \cdot \tau_\infty \mathsf{B}^{n+1-d}(1 + 0(1)), \quad \textit{as} \ \ \mathsf{B} \to \infty,$$

where τ_p, τ_∞ *are the p-adic, resp. real, densities. The Euler product converges provided local solutions exist for all* p *and in* $\mathbb{R}$.

We sketch the method of a proof of this result in Section 4.6.

Now we assume that $X = X_f$ is a hypersurface over a function field in one variable $F = \mathbb{C}(t)$. We have

THEOREM 2.2.5. *If* $\deg(f) \leq n$ *then* $X_f(\mathbb{C}(t)) \neq \emptyset$.

PROOF. It suffices to count parameters: Insert $x_j = x_j(t) \in \mathbb{C}[t]$, of degree e, into

$$f = \sum_J f_J x^J = 0,$$

with $|J| = \deg(f)$. This gives a system of $e \cdot \deg(f) + \text{const}$ equations in $e(n+1)$ variables. This system is solvable for $e \gg 0$, provided $\deg(f) \leq n$. □

2.3. Decidability. Hilbert's 10th problem has a negative solution: there is no algorithm to decide whether a or not a Diophantine equation is solvable in integers (see [**Mat00**], [**Mat06**]). In fact, there exist Diophantine equations

$$f_t(x_1, \ldots, x_n) = f(t, x_1, \ldots, x_n)$$

such that the set of $t \in \mathbb{Z}$ with the property that f_t has infinitely many primitive solutions $(x_1, \ldots, x_n)$ is algorithmically random [**Cha94**][1].

[1]The author's abstract: "One normally thinks that everything that is true is true for a reason. I've found mathematical truths that are true for no reason at all. These mathematical truths are beyond the power

There are many results concerning undecidability of general Diophantine equations over other rings and fields (for a recent survey, see [**Poo08b**]). The case of rational points, over a number field, is open; even for a cubic surface we cannot decide, at present, whether or not there are rational points.

2.4. Obstructions. As we have just said, there is no hope of finding an algorithm which would determine the solvability of a Diophantine equation in integers, i.e., there is no algorithm to test for the existence of *integral* points on quasi-projective varieties. The corresponding question for *homogeneous* equations, i.e., for *rational* points, is still open. It is reasonable to expect that at least for certain classes of algebraic varieties, for example, for Del Pezzo surfaces, the existence question can be answered. In this section we survey some recent results in this direction.

Let X_B be a scheme over a base scheme B. We are looking for obstructions to the existence of points $X(B)$, i.e., sections of the structure morphism $X \to B$. Each morphism $B' \to B$ gives rise to a base-change diagram, and each section $x : B \to X$ provides a section $x' : B' \to X_{B'}$.

$$\begin{array}{ccc} X & \longleftarrow & X_{B'} \\ \downarrow & & \downarrow \\ B & \longleftarrow & B' \end{array} \qquad \begin{array}{ccc} X & \longleftarrow & X_{B'} \\ \uparrow{\scriptstyle x} & & \uparrow{\scriptstyle x'} \\ B & \longleftarrow & B' \end{array}$$

This gives rise to a *local* obstruction, since it is sometimes easier to check that $X_{B'}(B') = \emptyset$. In practice, B could be a curve and B' a cover, or an analytic neighborhood of a point on B. In the number-theoretic context, $B = \mathrm{Spec}\,(F)$ and $B' = \mathrm{Spec}\,(F_v)$, where v is a valuation of the number field F and F_v the v-adic completion of F. One says that the *local-global principle*, or the *Hasse principle*, holds, if the existence of F-rational points is implied by the existence of v-adic points in all completions.

Example 2.4.1. The Hasse principle holds for:

(1) smooth quadrics $X_2 \subset \mathbb{P}^n$;
(2) Brauer–Severi varieties;
(3) Del Pezzo surfaces of degree ≥ 5;
(4) Châtelet surfaces $y^2 - az^2 = f(x_0, x_1)$, where f is an irreducible polynomial of degree ≤ 4 [**CTSSD87b**];
(5) hypersurfaces $X_d \subset \mathbb{P}^n$, for $n \gg d$ (see Theorem 2.2.4).

The Hasse principle may fail for cubic curves, e.g.,

$$3x^3 + 4y^3 + 5z^3 = 0.$$

of mathematical reasoning because they are accidental and random. Using software written in Mathematica that runs on an IBM RS/6000 workstation, I constructed a perverse 200-page algebraic equation with a parameter t and 17,000 unknowns. For each whole-number value of the parameter t, we ask whether this equation has a finite or an infinite number of whole number solutions. The answers escape the power of mathematical reason because they are completely random and accidental."

In topology, there is a classical obstruction theory to the existence of sections. An adaptation to algebraic geometry is formulated as follows: Let $\mathfrak{C}$ be a contravariant functor from the category of schemes over a base scheme B to the category of abelian groups. Applying the functor $\mathfrak{C}$ to the diagrams above, we have

$$\begin{array}{ccc} \mathfrak{C}(X) & \longrightarrow & \mathfrak{C}(X_{B'}) \\ \uparrow & & \uparrow \\ \mathfrak{C}(B) & \longrightarrow & \mathfrak{C}(B') \end{array} \qquad \begin{array}{ccc} \mathfrak{C}(X) & \longrightarrow & \mathfrak{C}(X_{B'}) \\ \downarrow{\scriptstyle x} & & \downarrow{\scriptstyle x'} \\ \mathfrak{C}(B) & \longrightarrow & \mathfrak{C}(B') \end{array}$$

If for all sections x', the image of x' in $\mathfrak{C}(B')$ is nontrivial in the cokernel of the map $\mathfrak{C}(B) \to \mathfrak{C}(B')$, then we have a problem, i.e., an obstruction to the existence of B-points on X. So far, this is still a version of a local obstruction. However, a *global obstruction* may arise, when we vary B'.

We are interested in the case when $B = \mathrm{Spec}\,(F)$, for a number field F, with B' ranging over all completions F_v. A global obstruction is possible whenever the map

$$\mathfrak{C}(\mathrm{Spec}\,(F)) \to \prod_v \mathfrak{C}(\mathrm{Spec}\,(F_v))$$

has a nontrivial cokernel. What are sensible choices for $\mathfrak{C}$? Basic contravariant functors on schemes are $\mathfrak{C}(-) := \mathrm{H}^i_{\text{ét}}(-, \mathbb{G}_m)$. For $i = 1$, we get the Picard functor, introduced in Section 1.1. However, by Hilbert's theorem 90,

$$\mathrm{H}^1_{\text{ét}}(F, \mathbb{G}_m) := \mathrm{H}^1_{\text{ét}}(\mathrm{Spec}\,(F), \mathbb{G}_m) = 0,$$

for all fields F, and this won't generate an obstruction. For $i = 2$, we get the (cohomological) Brauer group $\mathrm{Br}(X) = \mathrm{H}^2_{\text{ét}}(X, \mathbb{G}_m)$, classifying sheaves of central simple algebras over X, modulo equivalence (see [**Mil80**, Chapter 4]). By class class field theory, we have an exact sequence

$$0 \to \mathrm{Br}(F) \to \bigoplus_v \mathrm{Br}(F_v) \xrightarrow{\sum_v \mathrm{inv}_v} \mathbb{Q}/\mathbb{Z} \to 0, \tag{2.2}$$

where $\mathrm{inv}_v : \mathrm{Br}(F_v) \to \mathbb{Q}/\mathbb{Z}$ is the *local invariant*. We apply it to the diagram and obtain

$$\begin{array}{ccccccccc} & & \mathrm{Br}(X_F) & \longrightarrow & \bigoplus_v \mathrm{Br}(X_{F_v}) & & & & \\ & & \downarrow{\scriptstyle x} & & \downarrow{\scriptstyle (x_v)_v} & & & & \\ 0 & \longrightarrow & \mathrm{Br}(F) & \longrightarrow & \bigoplus_v \mathrm{Br}(F_v) & \xrightarrow{\sum_v \mathrm{inv}_v} & \mathbb{Q}/\mathbb{Z} & \longrightarrow & 0, \end{array}$$

Define

$$X(\mathbb{A}_F)^{\mathrm{Br}} := \bigcap_{A \in \mathrm{Br}(X)} \{(x_v)_v \in X(\mathbb{A}_F) \mid \sum_v \mathrm{inv}(A(x_v)) = 0\}. \tag{2.3}$$

Let $\overline{X(F)}$ be the closure of $X(F)$ in $X(\mathbb{A}_F)$, in the adelic topology. One says that X satisfies *weak approximation* over F if $\overline{X(F)} = X(\mathbb{A}_F)$. We have

$$X(F) \subset \overline{X(F)} \subseteq X(\mathbb{A}_F)^{\mathrm{Br}} \subseteq X(\mathbb{A}_F).$$

From this we derive the *Brauer–Manin* obstruction to the Hasse principle and weak approximation:

- if $X(\mathbb{A}_F) \neq \emptyset$ but $X(\mathbb{A}_F)^{\mathrm{Br}} = \emptyset$ then $X(F) = \emptyset$, i.e., the Hasse principle fails;
- if $X(\mathbb{A}_F) \neq X(\mathbb{A}_F)^{\mathrm{Br}}$ then weak approximation fails.

Del Pezzo surfaces of degree ≥ 5 satisfy the Hasse principle and weak approximation. Arithmetically most interesting are Del Pezzo surfaces of degree $4, 3$, and 2: these may fail the Hasse principle:

- deg $= 4$: $z^2 + w^2 = (x^2 - 2y^2)(3y^2 - x^2)$ [**Isk71**];
- deg $= 3$: $5x^3 + 12y^3 + 9z^3 + 10w^3 = 0$ [**CG66**];
- deg $= 2$: $w^2 = 2x^4 - 3y^4 - 6z^4$ [**KT04a**].

One says that the Brauer–Manin obstruction to the existence of rational points is the only one if $X(\mathbb{A}_F)^{\mathrm{Br}} \neq \emptyset$ implies that $X(F) \neq \emptyset$. This holds for:

(1) certain curves of genus ≥ 2 (see, e.g., [**Sto07**]);
(2) principal homogeneous spaces for connected linear algebraic groups over F [**San81**];
(3) Del Pezzo surfaces of degree ≥ 3 admitting a conic bundle structure defined over the ground field F;
(4) conjecturally(!), for all geometrically rational surfaces.

However, the Brauer–Manin obstruction is not the only one, in general. Here is a heuristic argument: a smooth hypersurface in $\mathbb{P}^4$ has trivial $\mathrm{Br}(X)/\mathrm{Br}(F)$. It is easy to satisfy local local conditions, so that for a positive proportion of hypersurfaces one has $X(\mathbb{A}_F) \neq \emptyset$ (see [**PV04**]). Consider X of *very large* degree. Lang's philosophy (see Conjecture 3.1.1) predicts that there are very few rational points over *any* finite extension of the ground field. Why should there be points over F? This was made precise in [**SW95**]. The first unconditional result in this direction was [**Sko99**]: there exist surfaces X with empty Brauer–Manin obstructions and étale covers $\tilde{X}$ which acquire new Brauer group elements producing nontrivial obstructions on $\tilde{X}$ and *a posteriori* on X. These type of "multiple-descent", nonabelian, obstructions were systematically studied in [**HS05**], [**HS02**], [**Sko01**] (see also [**Har**], and [**Pey05**], [**Har04**]).

Insufficiency of these nonabelian obstructions for threefolds was established in [**Poo08a**]. The counterexample is a fibration $\phi : X \to C$, defined over $\mathbb{Q}$, such that

- C is a curve of genus ≥ 2 with $C(\mathbb{Q}) \neq \emptyset$ (e.g., a Fermat curve);
- every fiber X_c, for $c \in C(\mathbb{Q})$, is the counterexample
$$z^2 + w^2 = (x^2 - 2y^2)(3y^2 - x^2)$$
from [**Isk71**], i.e., $X_c(\mathbb{A}_{\mathbb{Q}}) \neq \emptyset$, and $X_c(\mathbb{Q}) = \emptyset$;
- $\mathrm{Br}(X) \simeq \mathrm{Br}(C)$, and the same holds for any base change under an étale map $\tilde{C} \to C$.

Then $\tilde{X}(\mathbb{A}_{\mathbb{Q}})^{\mathrm{Br}} \neq \emptyset$, for every étale cover $\tilde{X} \to X$, while $X(\mathbb{Q}) = \emptyset$.

2.5. Descent. Let T be an algebraic torus, considered as a group scheme, and X a smooth projective variety over a number field F. We assume that $\mathrm{Pic}(X_{\bar{F}}) =$

$\mathrm{NS}(X_{\bar F})$. The F-isomorphisms classes of T-torsors

$$\pi : \mathcal{T} \to X$$

are parametrized by $\mathrm{H}^1_{\text{ét}}(X,T)$. A rational point $x \in X(F)$ gives rise to the specialization homomorphism

$$\sigma_x : \mathrm{H}^1_{\text{ét}}(X,T) \to \mathrm{H}^1_{\text{ét}}(F,T),$$

with image a finite set. Thus the partition:

$$X(F) = \bigcup_{\tau \in \mathrm{H}^1_{\text{ét}}(F,T)} \pi_\tau(\mathcal{T}_\tau(F)), \tag{2.4}$$

exhibiting $\mathcal{T}_\tau$ as *descent* varieties.

We now consider the $\Gamma = \mathrm{Gal}(\bar F/F)$-module $\mathrm{NS}(X_{\bar F})$ and the dual torus T_{NS}. The classifying map in Equation 1.8 is now

$$\chi : \mathrm{H}^1(X, T_{\mathrm{NS}}) \to \mathrm{Hom}_\Gamma(\mathrm{NS}(X_{\bar F}), \mathrm{Pic}(X_{\bar F})),$$

a T_{NS}-torsor $\mathcal{T}$ is called *universal* if $\chi([\mathcal{T}]) = \mathrm{Id}$ (it may not exist over the ground field F). The set of forms of a universal torsor $\mathcal{T}$ can be viewed is a principal homogeneous space under $\mathrm{H}^1_{\text{ét}}(F,T)$. The main reasons for working with universal torsors, rather than other torsors are:

- the Brauer–Manin obstruction on X translates to local obstructions on universal torsors, i.e.,
$$X(\mathbb{A}_F)^{\mathrm{Br}} = \bigcup_{\tau \in \mathrm{H}^1_{\text{ét}}(F,T)} \pi_\tau(\mathcal{T}_\tau(\mathbb{A}_F));$$
- the Brauer–Manin obstruction on universal torsors vanishes.

The foundations of the theory are in [**CTS87**] and in the book [**Sko01**].

2.6. Effectivity. In light of the discussion in Section 2.3 it is important to know whether or not the Brauer–Manin obstruction can be computed, effectively in terms of the coefficients of the defining equations. There is an extensive literature on such computations for curves (see the recent papers [**Fly04**], [**BBFL07**] and references therein) and for surfaces (e.g., [**CTKS87**], [**BSD04**], [**Cor07**], [**KT04b**]).

Effective computability of the Brauer-Manin obstruction for all Del Pezzo surfaces over number fields has been proved in [**KT08**]. The main steps are as follows:

(1) Computation of the equations of the exceptional curves and of the action of the Galois group Γ of a splitting field on these curves as in Section 1.13. One obtains the exact sequence of Γ-modules
$$0 \to \text{Relations} \to \oplus \mathbb{Z} E_j \to \mathrm{Pic}(\overline{X}) \to 0.$$
(2) We have
$$\mathrm{Br}(X)/\mathrm{Br}(F) = \mathrm{H}^1(\Gamma, \mathrm{Pic}(\overline{X})).$$
Using the equations for exceptional curves and functions realizing relations between the curves classes in the Picard group one can compute explicitly Azumaya algebras $\{\mathcal{A}_i\}$ representing the classes of $\mathrm{Br}(X)/\mathrm{Br}(F)$.
(3) The local points $X(F_v)$ can be effectively decomposed into a *finite* union of subsets such that each $\mathcal{A}_i$ is constant on each of these subsets. This step uses an effective version of the arithmetic Hilbert Nullstellensatz.

(4) The last step, the computation of local invariants is also effective.

3. Density of points

3.1. Lang's conjecture. One of the main principles underlying arithmetic geometry is the expectation that the trichotomy in the classification of algebraic varieties via the Kodaira dimension in Section 1.2 has an arithmetic manifestation. The broadly accepted form of this is

CONJECTURE 3.1.1 (Lang's conjecture). Let X be a variety of general type, i.e., a smooth projective variety with ample canonical class, defined over a number field F. Then $X(F)$ is not Zariski dense.

What about a converse? The obvious necessary condition for Zariski density of rational points, granted Conjecture 3.1.1, is that X does not dominate a variety of general type. This condition is not enough, as was shown in [**CTSSD97**]: there exist surfaces that do not dominate curves of general type but which have étale covers dominating curves of general type. By the Chevalley–Weil theorem (see, e.g., [**Abr**] in this volume), these covers would have a dense set of rational points, over some finite extension of the ground field, contradicting Conjecture 3.1.1.

As a first approximation, one expects that rational points are potentially dense on Fano varieties, on rationally connected varieties, and on Calabi–Yau varieties. Campana formulated precise conjectures characterizing varieties with potentially dense rational points via the notion of *special* varieties (see Section 1.5). In the following sections we survey techniques for proving density of rational points and provide representative examples illustrating these. For a detailed discussion of geometric aspects related to potential density see [**Abr**], and [**Has03**].

3.2. Zariski density over fixed fields. Here we address Zariski density of rational points in the "unstable" situation, when the density of points is governed by subtle number-theoretical properties, rather than geometric considerations. We have the following fundamental result:

THEOREM 3.2.1. *Let C be a smooth curve of genus $\mathsf{g} = \mathsf{g}(C)$ over a number field F. Then*

- *if $\mathsf{g} = 0$ and $C(F) \neq \emptyset$ then $C(F)$ is Zariski dense;*
- *if $\mathsf{g} = 1$ and $C(F) \neq \emptyset$ then $C(F)$ is an abelian group (the Mordell-Weil group) and there is a constant c_F (independent of C) bounding the order of the torsion subgroup $C(F)_{\mathrm{tors}}$ of $C(F)$* [**Maz77**], [**Mer96**]*; in particular, if there is an F-rational point of infinite order then $C(F)$ is Zariski dense;*
- *if $\mathsf{g} \geq 2$ then $C(F)$ is finite* [**Fal83**], [**Fal91**].

In higher dimensions we have:

THEOREM 3.2.2. *Let X be an algebraic variety over a number field F. Assume that $X(F) \neq \emptyset$ and that X is one of the following*

- *X is a Del Pezzo surface of degree 2 and has a point on the complement to the exceptional curves;*
- *X is a Del Pezzo surface of degree ≥ 3;*
- *X is a Brauer-Severi variety.*

Then $X(F)$ is Zariski dense.

The proof of the first two claims can be found in [**Man86**].

REMARK 3.2.3. Let X/F be a Del Pezzo surface of degree 1 (it always contains an F-rational point, the base point of the anticanonical linear series) or a conic bundle $X \to \mathbb{P}^1$, with $X(F) \neq \emptyset$. It is unknown whether or not $X(F)$ is Zariski dense.

THEOREM 3.2.4. [**Elk88**] *Let $X \subset \mathbb{P}^3$ be the quartic K3 surface given by*

$$x_0^4 + x_1^4 + x_2^4 = x_3^4. \tag{3.1}$$

Then $X(\mathbb{Q})$ is Zariski dense.

The trivial solutions $(1 : 0 : 0 : 1)$ etc are easily seen. The smallest nontrivial solution is

$$(95\,800, 217\,519, 414\,560, 422\,481).$$

Geometrically, over $\bar{\mathbb{Q}}$, the surface given by (3.1) is a Kummer surface, with many elliptic fibrations.

Example 3.2.5. [**EJ06**] Let $X \subset \mathbb{P}^3$ be the quartic given by

$$x^4 + 2y^4 = z^4 + 4t^4.$$

The obvious $\mathbb{Q}$-rational points are given by $y = t = 0$ and $x = \pm z$. The next smallest solution is

$$1484801^4 + 2 \cdot 1203120^4 = 1169407^4 + 4 \cdot 1157520^4.$$

It is unknown whether or not $X(\mathbb{Q})$ is Zariski dense.

3.3. Potential density: techniques. Here is a (short) list of possible strategies to propagate points:

- use the group of automorphisms $\mathrm{Aut}(X)$, if it is infinite;
- try to find a dominant map $\tilde{X} \to X$ where $\tilde{X}$ satisfies potential density (for example, try to prove *unirationality*);
- try to find a fibration structure $X \to B$ where the fibers satisfy potential density in some *uniform way* (that is, the field extensions needed to insure potential density of the fibers V_b can be uniformly controlled).

In particular, it is important for us keep track of minimal conditions which would insure Zariski density of points on varieties.

Example 3.3.1. Let $\pi : X \to \mathbb{P}^1$ be a conic bundle, defined over a field F. Then rational points on X are potentially dense. Indeed, by Tsen's theorem, π has a section $s : \mathbb{P}^1 \to X$ (which is defined over some finite extension F'/F), each fiber has an F'-rational point and it suffices to apply Theorem 3.2.1. Potential density for conic bundles over higher-dimensional bases is an open problem.

If X is an abelian variety then there exists a finite extension F'/F and a point $P \in X(F')$ such that the cyclic subgroup of $X(F')$ generated by P is Zariski dense (see, e.g., [**HT00b**, Proposition 3.1]).

Example 3.3.2. If $\pi : X \to \mathbb{P}^1$ is a Jacobian nonisotrivial elliptic fibration (π admits a section and the j-invariant is nonconstant), then potential density follows from a strong form of the Birch/Swinnerton-Dyer conjecture [**GM97**], [**Man95**]. The key problem is to control the variation of the *root number* (the sign of the functional equation of the L-functions of the elliptic curve) (see [**GM97**]).

On the other hand, rational points on certain elliptic fibrations with *multiple* fibers are not potentially dense [**CTSSD97**].

Example 3.3.3. One geometric approach to Zariski density of rational points on (certain) elliptic fibration can be summarized as follows:

Case 1. Let $\pi : X \to B$ be a Jacobian elliptic fibration and $e : B \to X$ its zero-section. Suppose that we have another section s which is nontorsion in the Mordell-Weil group of $X(F(B))$. Then a specialization argument implies that the restriction of the section to infinitely many fibers of π gives a nontorsion point in the Mordell-Weil group of the corresponding fiber (see [**Ser89**], 11.1). In particular, $X(F)$ is Zariski dense, provided $B(F)$ is Zariski dense in B.

Case 2. Suppose that $\pi : X \to B$ is an elliptic fibration with a *multisection* M (an irreducible curve surjecting onto the base B). After a base change $X \times_B M \to M$ the elliptic fibration acquires the identity section Id (the image of the diagonal under $M \times_B M \to V \times_B M$) *and* a (rational) section

$$\tau_M := d\,\mathrm{Id} - \mathrm{Tr}(M \times_B M),$$

where d is the degree of $\pi : M \to B$ and $\mathrm{Tr}(M \times_B M)$ is obtained (over the generic point) by summing all the points of $M \times_B M$. We will say that M is *nontorsion* if τ_M is nontorsion.

If M is nontorsion and if $M(F)$ is Zariski dense then the same holds for $X(F)$ (see [**BT99**]).

Remark 3.3.4. Similar arguments work for abelian fibrations [**HT00c**]. The difficulty here is to formulate some simple geometric conditions ensuring that a (multi)section leads to points which are not only of infinite order in the Mordell-Weil groups of the corresponding fibers, but in fact generate Zariski dense subgroups.

3.4. Potential density for surfaces. By Theorem 3.2.1, potential density holds for curves of genus $\mathsf{g} \leq 1$. It holds for surfaces which become rational after a finite extension of the ground field, e.g., for all Del Pezzo surfaces. The classification theory in dimension 2 gives us a list of surfaces of Kodaira dimension 0:

- abelian surfaces;
- bielliptic surfaces;
- Enriques surfaces;
- K3 surfaces.

Potential density for the first two classes follows from Theorem 3.2.2. The classification of Enriques surfaces X implies that either $\mathrm{Aut}(X)$ is infinite or X is dominated by a K3 surface $\tilde{X}$ with $\mathrm{Aut}(\tilde{X})$ infinite [**Kon86**]. Thus we are reduced to the study of K3 surfaces.

THEOREM 3.4.1. [**BT00**] *Let X be a K3 surface over* any *field of characteristic zero. If X is elliptic or admits an infinite group of automorphisms then rational points on X are potentially dense.*

SKETCH OF THE PROOF. One needs to find sufficiently nondegenerate rational or elliptic multisections of the elliptic fibration $X \to \mathbb{P}^1$. These are produced using deformation theory. One starts with special K3 surfaces which have rational curves $C_t \subset X_t$ in the desired homology class (for example, Kummer surfaces) and then deforms the pair. This deformation technique has to be applied to twists of the original elliptic surface. □

Example 3.4.2. A smooth hypersurface $X \subset \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1$ of bi-degree $(2,2,2)$ is a K3 surface with $\mathrm{Aut}(X)$ infinite.

Example 3.4.3. Every smooth quartic surface $S_4 \subset \mathbb{P}^3$ which contains a line is an elliptic K3 surface. Indeed, let M be this line and assume that both S_4 and M are defined over a number field F. Consider the 1-parameter family of planes $\mathbb{P}^2_t \subset \mathbb{P}^3$ containing M. The residual curve in the intersection $\mathbb{P}^2_t \cap S_4$ is a plane cubic intersecting M in 3 points. This gives a fibration $\pi : S_4 \to \mathbb{P}^1$ with a *rational* tri-section M.

To apply the strategy of Section 6.1 we need to ensure that M is nontorsion. A sufficient condition, satisfied for generic quartics S_4, is that the restriction of π to M ramifies in a smooth fiber of $\pi : X \to \mathbb{P}^1$. Under this condition $X(F)$ is Zariski dense.

THEOREM 3.4.4. [**HT00a**] *Let $X \subset \mathbb{P}^3$ be a quartic K3 surface containing a line defined over a field F. If X is general, then $X(F)$ is Zariski dense. In all cases, there exists a finite extension F'/F such that $X(F')$ is Zariski dense.*

THEOREM 3.4.5. [**BT00**] *Let X be an elliptic K3 surface over a field F. Then rational points are potentially dense.*

Are there K3 surfaces X over $\mathbb{Q}$ with geometric Picard number 1, $X(\mathbb{Q}) \neq \emptyset$ and $X(\mathbb{Q})$ not Zariski dense?

3.5. Potential density in dimension ≥ 3. Potential density holds for unirational varieties. Classification of (smooth) Fano threefolds and the detailed study of occurring families implies unirationality for all but three cases:

- X_4: quartics in $\mathbb{P}^4$;
- V_1: double covers of a cone over the Veronese surface in $\mathbb{P}^5$ ramified in a surface of degree 6;
- W_2: double covers of $\mathbb{P}^3$ ramified in a surface of degree 6.

We now sketch the proof of potential density for quartics from [**HT00a**]; the case of V_1 is treated by similar techniques in [**BT99**].

The threefold X_4 contains a 1-parameter family of lines. Choose a line M (defined over some extension of the ground field, if necessary) and consider the 1-parameter family of hyperplanes $\mathbb{P}^3_t \subset \mathbb{P}^4$ containing M. The generic hyperplane

section $S_t := \mathbb{P}^3_t \cap X_4$ is a quartic surface with a line. Now we would like to argue as in Example 3.4.3. We need to make sure that M is nontorsion in S_t for a dense set of $t \in \mathbb{P}^1$. This will be the case for general X_4 and M. The analysis of all exceptional cases requires care.

REMARK 3.5.1. It would be interesting to have further (nontrivial) examples of birationally rigid Fano varieties with Zariski dense sets of rational points. Examples of Calabi–Yau varieties over function fields of curves, with geometric Picard number one and dense sets of rational points have been constructed in [**HT08b**]; little is known over number fields.

THEOREM 3.5.2. [**HT00c**] *Let X be a K3 surface over a field F, of degree $2(n-1)$. Then rational points on $X^{[n]}$ are potentially dense.*

The proof relies on the existence of an abelian fibration

$$Y := X^{[n]} \to \mathbb{P}^n,$$

with a nontorsion multisection which has a potentially dense set of rational points. Numerically, such fibrations are predicted by square-zero classes in the Picard group $\mathrm{Pic}(Y)$, with respect to the Beauville–Bogomolov form (see Section 1.12). Geometrically, the fibration is the degree n Jacobian fibration associated to hyperplane sections of X.

THEOREM 3.5.3. [**AV08**] *Let Y be the Fano variety of lines on a general cubic fourfold $X_3 \subset \mathbb{P}^5$ over a field of characteristic zero. Then rational points on Y are potentially dense.*

SKETCH OF PROOF. The key tool is a rational endomorphism $\phi\colon Y \to Y$ analyzed in [**Voi04**]: let $\mathfrak{l}$ on $X_3 \subset \mathbb{P}^5$ be a general line and $\mathbb{P}^2_{\mathfrak{l}} \subset \mathbb{P}^5$ the unique plane everywhere tangent to $\mathfrak{l}$,

$$\mathbb{P}^2_{\mathfrak{l}} \cap X = 2\mathfrak{l} + \mathfrak{l}'.$$

Let $[\mathfrak{l}] \in Y$ be the corresponding point and put $\phi([\mathfrak{l}]) := [\mathfrak{l}']$, where $\mathfrak{l}'$ is the residual line in X_3.

Generically, one can expect that the orbit $\{\phi^n([\mathfrak{l}])\}_{n\in\mathbb{N}}$ is Zariski dense in Y. This was proved by Amerik and Campana in [**AC08**], over *uncountable* ground fields. Over countable fields, one faces the difficulty that the countably many exceptional loci could cover all algebraic points of Y. Amerik and Voisin were able to overcome this obstacle over number fields. Rather than proving density of $\{\phi^n([\mathfrak{l}])\}_{n\in\mathbb{N}}$ they find surfaces $\Sigma \subset Y$, birational to abelian surfaces, whose orbits are dense in Y. The main effort goes into showing that one can choose sufficiently general Σ defined over $\bar{\mathbb{Q}}$, provided that Y is sufficiently general and still defined over a number field. In particular, Y has geometric Picard number one. A case-by-case geometric analysis excludes the possibility that the Zariski closure of $\{\phi^n(\Sigma)\}_{n\in\mathbb{N}}$ is a proper subvariety of F. □

THEOREM 3.5.4. [**HT08a**] *Let Y be the variety of lines of a cubic fourfold $X_3 \subset \mathbb{P}^5$ which contains a cubic scroll T. Assume that the hyperplane section of X_3 containing T has exactly 6 double points in linear general position and that X_3*

does not contain a plane. If X_3 and T are defined over a field F then $Y(F)$ is Zariski dense.

REMARK 3.5.5. In higher dimensions, (smooth) hypersurfaces $X_d \subset \mathbb{P}^n$ of degree d represent a major challenge. The circle method works well when

$$n \gg 2^d$$

while the geometric methods for proving unirationality require at least a super-exponential growth of n (see [**HMP98**] for a construction of a unirational parametrization).

3.6. Approximation. Let X be smooth and projective. Assume that $X(F)$ is dense in each $X(F_v)$. A natural question is whether or not $X(F)$ is dense in the adeles $X(\mathbb{A}_F)$. This *weak approximation* may be obstructed globally, by the Brauer–Manin obstruction, as explained in Section 2.4. There are examples of such obstructions for Del Pezzo surfaces in degree ≤ 4, for conic bundles over $\mathbb{P}^2$ [**Har96**], and for K3 surfaces as in the following example.

Example 3.6.1. [**Wit04**] Let $E \to \mathbb{P}^1$ be the elliptic fibration given by

$$y^2 = x(x-g)(x-h) \quad \text{where} \quad g(t) = 3(t-1)^3(t+3) \quad \text{and} \quad h = g(-t).$$

Its minimal proper regular model X is an elliptic K3 surface that fails weak approximation. The obstruction comes from transcendental classes in the Brauer group of X.

The theory is parallel to the theory of the Brauer–Manin obstruction to the Hasse principle, up to a certain point. The principal new feature is:

THEOREM 3.6.2. [**Min89**] *Let X be a smooth projective variety over a number field with a nontrivial geometric fundamental group. Then weak approximation fails for X.*

This applies to Enriques surfaces [**HS05**].

Of particular interest are varieties which are unirational over the ground field F, e.g., cubic surfaces with an F-rational point. Other natural examples are quotients V/G, where G is a group and V a G-representation, discussed in Section 1.2.

4. Counting problems

Here we consider projective algebraic varieties $X \subset \mathbb{P}^n$ defined over a number field F. We assume that $X(F)$ is Zariski dense. We seek to understand the distribution of rational points with respect to heights.

4.1. Heights. First we assume that $F = \mathbb{Q}$. Then we can define a *height* of integral (respectively rational) points on the affine (respectively projective) space as follows

$$\begin{array}{rcl} \mathsf{H}_{\text{affine}} : \mathbb{A}^n(\mathbb{Z}) = \mathbb{Z}^n & \to & \mathbb{R}_{\geq 0} \\ x = (x_1, \dots, x_n) & \mapsto & \|x\| = \max_j(|x_j|) \\ \mathsf{H} : \mathbb{P}^n(\mathbb{Q}) = (\mathbb{Z}^{n+1}_{\text{prim}} \setminus 0)/\pm & \to & \mathbb{R}_{>0} \\ x = (x_0, \dots, x_n) & \mapsto & \|x\| = \max_j(|x_j|). \end{array}$$

Here $\mathbb{Z}^{n+1}_{\text{prim}}$ are the primitive vectors. This induces heights on points of subvarieties of affine or projective spaces. In some problems it is useful to work with alternative norms, e.g., $\sqrt{\sum x_j^2}$ instead of $\max_j(|x_j|)$. Such choices are referred to as a *change of metrization*. A more conceptual definition of heights and adelic metrizations is given in Section 4.8

4.2. Counting functions. For a subvariety $X \subset \mathbb{P}^n$ put

$$\mathsf{N}(X, \mathsf{B}) := \#\{x \in X(\mathbb{Q}) \,|\, \mathsf{H}(x) \le \mathsf{B}\}.$$

What can be said about

$$\mathsf{N}(X, \mathsf{B}), \quad \text{for } \mathsf{B} \to \infty \ ?$$

Main questions here concern:

- (uniform) upper bounds,
- asymptotic formulas,
- geometric interpretation of the asymptotics.

By the very definition, $\mathsf{N}(X, \mathsf{B})$ depends on the projective embedding of X. For $X = \mathbb{P}^n$ over $\mathbb{Q}$, with the standard embedding via the line bundle $\mathcal{O}(1)$, we get

$$\mathsf{N}(\mathbb{P}^n, \mathsf{B}) = \frac{1}{\zeta(n+1)} \cdot \tau_\infty \cdot \mathsf{B}^{n+1}(1 + o(1)), \ \ \mathsf{B} \to \infty,$$

where τ_∞ is the volume of the unit ball with respect to the metrization of $\mathcal{O}(1)$. But we may also consider the Veronese re-embedding

$$\begin{array}{ccc} \mathbb{P}^n & \to & \mathbb{P}^N \\ x & \mapsto & x^I, \ |I| = d, \end{array}$$

e.g.,

$$\begin{array}{ccc} \mathbb{P}^1 & \to & \mathbb{P}^2, \\ (x_0 : x_1) & \mapsto & (x_0^2 : x_0x_1 : x_1^2). \end{array}$$

The image $y_0y_2 = y_1^2$ has $\sim \mathsf{B}$ points of height $\le \mathsf{B}$. Similarly, the number of rational points on height $\le \mathsf{B}$ in the $\mathcal{O}(d)$-embedding of $\mathbb{P}^n$ will be about $\mathsf{B}^{(n+1)/d}$.

More generally, if $F/\mathbb{Q}$ is a finite extension, put

$$\begin{array}{ccc} \mathbb{P}^n(F) & \to & \mathbb{R}_{>0} \\ x & \mapsto & \prod_v \max(|x_j|_v). \end{array}$$

THEOREM 4.2.1. [**Sch79**]
(4.1)

$$\mathsf{N}(\mathbb{P}^n(F), \mathsf{B}) = \frac{h_F R_F (n+1)^{r_1+r_2-1}}{w_F \zeta_F(n+1)} \left(\frac{2^{r_1}(2\pi)^{r_2}}{\sqrt{\operatorname{disc}(F)}} \right)^{n+1} \mathsf{B}^{n+1}(1 + o(1)), \quad \mathsf{B} \to \infty$$

where

- h_F *is the class number of* F*;*
- R_F *the regulator;*
- r_1 *(resp.* r_2*) the number of real (resp. pairs of complex) embeddings of* F*;*
- $\operatorname{disc}(F)$ *the discriminant;*
- w_F *the number of roots of 1 in* F*;*
- ζ_F *the zeta function of* F*.*

With this starting point, one may try to prove asymptotic formulas of similar precision for arbitrary projective algebraic varieties X, at least under some natural geometric conditions. This program was initiated in [**FMT89**] and it has rapidly grown in recent years.

4.3. Upper bounds. A first step in understanding growth rates of rational points of bounded height is to obtain uniform upper and lower bounds, with effective control of error terms. Results of this type are quite valuable in arguments using fibration structures. Here is a sample:

- [**BP89**], [**Pil96**]: Let $X \subset \mathbb{A}^2$ be a geometrically irreducible affine curve. Then
$$\#\{x \in X(\mathbb{Z}) \,|\, \mathsf{H}_{\mathrm{affine}}(x) \le \mathsf{B}\} \ll_{\deg(X)} \mathsf{B}^{\frac{1}{\deg(X)}} \log(\mathsf{B})^{2\deg(X)+3}.$$
- [**EV05**]: Let $X \subset \mathbb{P}^2$ be a geometrically irreducible curve of genus ≥ 1. Then there is a $\delta > 0$ such that
$$\mathsf{N}(X(\mathbb{Q}), \mathsf{B}) \ll_{\deg(X),\delta} \mathsf{B}^{\frac{2}{\deg(X)}-\delta}.$$

Fibering and using estimates for lower dimensional varieties, one has:

THEOREM 4.3.1. [**Pil95**] *Let $X \subset \mathbb{P}^n$ be a geometrically irreducible variety, and $\epsilon > 0$. Then*
$$\mathsf{N}(X(\mathbb{Q}), \mathsf{B}) \ll_{\deg(X),\dim(X),\epsilon} \mathsf{B}^{\dim(X)+\frac{1}{\deg(X)}+\epsilon}$$

The next breakthrough was accomplished in [**HB02**]; further refinements combined with algebro-geometric tools lead to

THEOREM 4.3.2 ([**BHBS06**], [**Sal07**]). *Let $X \subset \mathbb{P}^n$ be a geometrically irreducible variety, and $\epsilon > 0$. Then*
$$\mathsf{N}(X(\mathbb{Q}), \mathsf{B}) \ll_{\deg(X),\dim(X),\epsilon} \begin{cases} \mathsf{B}^{\dim(X)-\frac{3}{4}+\frac{5}{3\sqrt{3}}+\epsilon} & \deg(X) = 3 \\ \mathsf{B}^{\dim(X)-\frac{2}{3}+\frac{3}{2\sqrt{\deg(X)}}+\epsilon} & \deg(X) = 4, 5 \\ \mathsf{B}^{\dim(X)+\epsilon} & \deg(X) \ge 6 \end{cases}$$

A survey of results on upper bounds, with detailed proofs, is in [**HB06**].

4.4. Lower bounds. Let X be a projective variety over a number field F and let L be a very ample line bundle on X. This gives an embedding $X \hookrightarrow \mathbb{P}^n$. We fix a height H on $\mathbb{P}^n(F)$ and consider the counting function
$$\mathsf{N}(X(F), L, \mathsf{B}) := \#\{x \in X(F) \,|\, \mathsf{H}_L(x) \le \mathsf{B}\},$$
with respect to the induced height H_L (see Section 4.8 for more explanations on heights).

LEMMA 4.4.1. *Let X be a smooth Fano variety over a number field F and $Y := \mathrm{Bl}_Z(X)$ a blowup in a smooth subvariety $Z = Z_F$ of codimension ≥ 2. If $\mathsf{N}(X^\circ(F), -K_X, \mathsf{B}) \gg \mathsf{B}^1$, for all dense Zariski open $X^\circ \subset X$ then the same holds for Y:*
$$\mathsf{N}(Y^\circ(F), -K_Y, \mathsf{B}) \gg \mathsf{B}^1.$$

PROOF. Let $\pi : Y \to X$ be the blowup. We have

$$-K_Y = \pi^*(-K_X) - D$$

with $\mathrm{supp}(D) \subset E$, the exceptional divisor. It remains to use the fact that $\mathsf{H}_D(x)$ is uniformly bounded from below on $(X \setminus D)(F)$ (see, e.g., [**BG06**, Proposition 2.3.9]), so that

$$\mathsf{N}(\pi^{-1}(X^\circ)(F), -K_Y, \mathsf{B}) \geq \mathsf{c} \cdot \mathsf{N}(X^\circ(F), -K_X, \mathsf{B}),$$

for some constant $\mathsf{c} > 0$ and an appropriate Zariski open $X^\circ \subset X$. □

In particular, *split* Del Pezzo surfaces X_r satisfy the lower bound of Conjecture 4.10.1

$$\mathsf{N}(X_r(F), -K_{X_r}, \mathsf{B}) \gg \mathsf{B}^1.$$

Finer lower bounds, in some nonsplit cases have been proved in [**SSD98**]:

$$\mathsf{N}(X_6^\circ(F), -K_{X_6}, \mathsf{B}) \gg \mathsf{B}^1 \log(\mathsf{B})^{r-1},$$

provided the cubic surface X_6 has at least two skew lines defined over F. This gives support to Conjecture 4.10.2. The following theorem gives evidence for Conjecture 4.10.1 in dimension 3.

THEOREM 4.4.2. [**Man93**] *Let X be a Fano threefold over a number field F_0. For every Zariski open subset $X^\circ \subset X$ there exists a finite extension F/F_0 such that*

$$\mathsf{N}(X(F), -K_X, \mathsf{B}) \gg \mathsf{B}^1.$$

This relies on the classification of Fano threefolds (cf. [**IP99b**], [**MM82**], [**MM86**]). One case was missing from the classification when [**Man93**] was published; the Fano threefold obtained as a blowup of $\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1$ in a curve of tri-degree $(1, 1, 3)$ [**MM03**]. Lemma 4.4.1 proves the expected lower bound in this case as well. An open question is whether or not one can choose the extension F independently of X°.

4.5. Finer issues. At the next level of precision we need to take into account more refined arithmetic and geometric data. Specifically, we need to analyze the possible sources of failure of the heuristic $\mathsf{N}(\mathsf{B}) \sim \mathsf{B}^{n+1-d}$ in Section 2.2:

- *Local or global obstructions*: as in $x_0^2 + x_1^2 + x_2^2 = 0$ or $x_0^3 + 4x_1^3 + 10x_2^3 + 25x_3^3 = 0$;
- *Singularities*: the surface $x_1^2x_2^2 + x_2^2x_3^2 + x_3^2x_1^2 = x_0x_1x_2x_3$ has $\sim \mathsf{B}^{3/2}$ points of height $\leq \mathsf{B}$, on every Zariski open subset, too many!
- *Accumulating subvarieties*: On $x_0^3 + x_1^3 + x_2^3 + x_3^3 = 0$ there are $\sim \mathsf{B}^2$ points on $\mathbb{Q}$-lines and and provably $O(\mathsf{B}^{4/3+\epsilon})$ points in the complement [**HB97**]. The expectation is $\mathsf{B}\log(\mathsf{B})^3$, over $\mathbb{Q}$. Similar effects persist in higher dimensions. A quartic $X_4 \subset \mathbb{P}^4$ contains a 1-parameter family of lines, each contributing $\sim \mathsf{B}^2$ to the asymptotic, while the expectation is $\sim \mathsf{B}$. Lines on a cubic $X_3 \subset \mathbb{P}^4$ are parametrized by a surface, which is of general type. We expect $\sim \mathsf{B}^2$ points of height $\leq \mathsf{B}$ on the cubic threefold, and on each line. In [**BG06**, Theorem 11.10.11] it is shown that

$$\mathsf{N}_{\mathrm{lines}}(\mathsf{B}) = c\,\mathsf{B}^2(1 + o(1)), \quad \text{as} \quad \mathsf{B} \to \infty,$$

where the count is over F-rational points on lines defined over F, and the constant c is a *convergent* sum of leading terms of contributions from each line of the type (4.1). In particular, each line contributes a positive density to the main term. On the other hand, one expects the same asymptotic $\sim \mathsf{B}^2$ on the complement of the lines, with the leading term a product of local densities. How to reconcile this? The forced compromise is to discard such *accumulating* subvarieties and to hope that for some Zariski open subset $X^\circ \subset X$, the asymptotic of points of bounded height does reflect the geometry of X, rather than the geometry of its subvarieties.

These finer issues are particularly striking in the case of K3 surfaces. They may have local and global obstructions to the existence of rational points, they may fail the heuristic asymptotic, and they may have accumulating subvarieties, even infinitely many:

Conjecture 4.5.1. (see [**BM90**]) Let X be a K3 surface over a number field F. Let L be a polarization, $\epsilon > 0$ and $Y = Y(\epsilon, L)$ be the union of all F-rational curves $C \subset X$ (i.e., curves that are isomorphic to $\mathbb{P}^1$ over F) that have L-degree $\leq 2/\epsilon$. Then

$$\mathsf{N}(X, L, \mathsf{B}) = \mathsf{N}(Y, L, \mathsf{B}) + O(\mathsf{B}^\epsilon), \quad \text{as} \quad B \to \infty.$$

Theorem 4.5.2. [**McK00**] *Let $X \to \mathbb{P}^1 \times \mathbb{P}^1$ be a double cover ramified over a curve of bidegree $(4,4)$. Then there exists an open cone $\Lambda \subset \Lambda_{\text{ample}}(X)$ such that for every $L \in \Lambda$ there exists a $\delta > 0$ such that*

$$\mathsf{N}(X, L, \mathsf{B}) = \mathsf{N}(Y, L, \mathsf{B}) + O(\mathsf{B}^{2/d-\delta}), \quad \textit{as} \quad \mathsf{B} \to \infty,$$

where d is the minimal L-degree of a rational curve on X and Y is the union of all F-rational curves of degree d.

This theorem exhibits the first layer of an arithmetic stratification predicted in Conjecture 4.5.1.

4.6. The circle method. Let $f \in \mathbb{Z}[x_0, \dots, x_n]$ be a homogeneous polynomial of degree d such that the hypersurface $X_f \subset \mathbb{P}^n$ is nonsingular. Let

$$\mathsf{N}_f(\mathsf{B}) := \#\{\mathbf{x} \in \mathbb{Z}^n \mid f(\mathbf{x}) = 0, \ \|\mathbf{x}\| \leq \mathsf{B}\}$$

be the counting function. In this section we sketch a proof of the following

Theorem 4.6.1. [**Bir62**] *Assume that $n \geq 2^d(d+1)$. Then*

$$\mathsf{N}_f(\mathsf{B}) = \Theta \cdot \mathsf{B}^{n+1-d}(1 + o(1)) \ \mathsf{B} \to \infty, \tag{4.2}$$

where

$$\Theta = \prod_p \tau_p \cdot \tau_\infty > 0,$$

provided $f(\mathbf{x}) = 0$ is solvable in $\mathbb{Z}_p$, for all p, and in $\mathbb{R}$.

The constants τ_p and τ_∞ admit an interpretation as *local densities*; these are explained in a more conceptual framework in Section 4.12.

Substantial efforts have been put into reducing the number of variables, especially for low degrees. Another direction is the extension of the method to systems of equations [**Sch85**] or to more general number fields [**Ski97**].

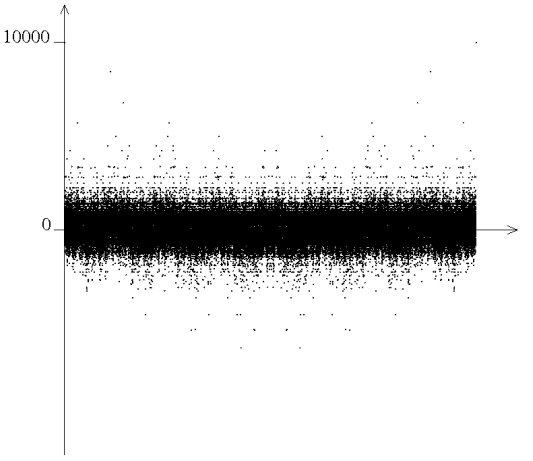


FIGURE 1. Oscillations of $\mathsf{S}(\alpha)$

We now outline the main steps of the proof of the asymptotic formula 4.2. The first is the introduction of a "delta"-function: for $x \in \mathbb{Z}$ we have

$$\int_0^1 e^{2\pi i\alpha x} d\alpha = \begin{cases} 0 & \text{if } x \neq 0, \\ 1 & \text{otherwise.} \end{cases}$$

Now we can write

$$\mathsf{N}_f(\mathsf{B}) = \int_0^1 \mathsf{S}(\alpha)\, d\alpha, \tag{4.3}$$

where

$$\mathsf{S}(\alpha) := \sum_{\mathbf{x}\in\mathbb{Z}^{n+1},\, \|\mathbf{x}\|\leq \mathsf{B}} e^{2\pi i\alpha f(\mathbf{x})}.$$

The function $\mathsf{S}(\alpha)$ is wildly oscillating (see Figure 1), with peaks at $\alpha = a/q$, for small q. Indeed, the probability that $f(\mathbf{x})$ is divisible by q is higher for small q, and each such term contributes 1 to $\mathsf{S}(\alpha)$. The idea of the circle method is to establish the asymptotic of the integral in equation 4.3, for $\mathsf{B} \to \infty$, by extracting the contributions of α close to rational numbers a/q with small q, and finding appropriate bounds for integrals over the remaining intervals.

More precisely, one introduces the *major arcs*

$$\mathfrak{M} := \bigcup_{(a,q)=1,\, q\le \mathsf{B}^{\Delta}} \mathfrak{M}_{a,q},$$

where $\Delta > 0$ is a parameter to be specified, and

$$\mathfrak{M}_{a,q} := \left\{ \alpha \mid \left| \alpha - \frac{a}{q} \right| \le \mathsf{B}^{-d+\delta} \right\}.$$

The *minor arcs* are the complement:

$$\mathfrak{m} := [0,1] \setminus \mathfrak{M}.$$

The goal is to prove the bound

$$\int_{\mathfrak{m}} \mathsf{S}(\alpha)\, d\alpha = O(\mathsf{B}^{n+1-d-\epsilon}), \quad \text{for some } \epsilon > 0, \tag{4.4}$$

and the asymptotic

$$\int_{\mathfrak{M}} \mathsf{S}(\alpha)\, d\alpha = \prod_p \tau_p \cdot \tau_\infty \cdot \mathsf{B}^{n+1-d}(1+o(1)) \quad \text{for } \mathsf{B} \to \infty. \tag{4.5}$$

Remark 4.6.2. Modern refinements employ "smoothed out" intervals, i.e., the delta function of an interval in the major arcs is replaced by a smooth bell curve with support in this interval. In Fourier analysis, "rough edges" translate into bad bounds on the dual side, and should be avoided. An implementation of this idea, leading to savings in the number of variables, can be found in [**HB83**].

There are various approaches to proving upper bounds in equation 4.4; most are a variant or refinement of Weyl's bounds (1916) [**Wey16**]. Weyl considered the following exponential sums:

$$\mathsf{s}(\alpha) := \sum_{0\le x\le \mathsf{B}} e^{2\pi i \alpha x^d}.$$

The main observation is that $|\mathsf{s}(\alpha)|$ is "small", when $|\alpha - a/q|$ is "large". This is easy to see when $d = 1$; summing the geometric series we get

$$|\mathsf{s}(\alpha)| = \left| \frac{1 - e^{2\pi i \alpha(\mathsf{B}+1)}}{1 - e^{2\pi i \alpha}} \right| \ll \frac{1}{\langle\langle \alpha \rangle\rangle},$$

where $\langle\langle \alpha \rangle\rangle$ is the distance to the nearest integer. In general, Weyl's differencing technique is applied to reduce the degree, to eventually arrive at a geometric series.

We turn to major arcs. Let

$$\alpha = \frac{a}{q} + \beta$$

with β *very small*, and getting smaller as a function of B. Here we will assume that $|\beta| \le \mathsf{B}^{-d+\delta'}$, for some small $\delta' > 0$. We put $\mathbf{x} = q\mathbf{y} + \mathbf{z}$, with $\mathbf{z}$ the corresponding

residue class modulo q, and obtain

$$\begin{aligned}\mathsf{S}(\alpha) &= \sum_{\mathbf{x}\in\mathbb{Z}^{n+1},\,\|\mathbf{x}\|\le \mathsf{B}} e^{2\pi i\frac{a}{q}f(\mathbf{x})}e^{2\pi i\beta f(\mathbf{x})}\\ &= \sum_{\|\mathbf{x}\|\le \mathsf{B}} e^{2\pi i\frac{a}{q}f(q\mathbf{y}+\mathbf{z})}e^{2\pi i\beta f(\mathbf{x})}\\ &= \sum_{\mathbf{z}} e^{2\pi i\frac{a}{q}f(\mathbf{z})}\Big(\sum_{\|\mathbf{y}\|\le \mathsf{B}/q} e^{2\pi i\beta f(\mathbf{x})}\Big)\\ &= \sum_{\mathbf{z}} e^{2\pi i\frac{a}{q}f(\mathbf{z})}\int_{\|\mathbf{y}\|\le \mathsf{B}/q} e^{2\pi i\beta f(\mathbf{x})}\,d\mathbf{y}\\ &= \sum_{\mathbf{z}} \frac{e^{2\pi i\frac{a}{q}f(\mathbf{z})}}{q^{n+1}}\int_{\|\mathbf{x}\|\le \mathsf{B}} e^{2\pi i\beta f(\mathbf{x})}\,d\mathbf{x},\end{aligned}$$

where $d\mathbf{y} = q^{n+1}d\mathbf{x}$. The passage $\sum \mapsto \int$ is justified for our choice of small β—the difference will be absorbed in the error term in (4.2). We have obtained

$$\int_0^1 \mathsf{S}(\alpha)\,d\alpha = \sum_{a,q}\sum_{\mathbf{z}} \frac{e^{2\pi i\frac{a}{q}f(\mathbf{z})}}{q^{n+1}}\cdot\int_{|\beta|\le \mathsf{B}^{-d+\delta}}\int_{\|\mathbf{x}\|\le \mathsf{B}} e^{2\pi i\beta f(\mathbf{x})}\,d\mathbf{x}d\beta,$$

modulo a negligible error. We first deal with the integral on the right, called the *singular integral*. Put $\beta' = \beta\mathsf{B}^d$ and $\mathbf{x}' = \mathbf{x}/\mathsf{B}$. The change of variables leads to

$$\int_{|\beta|\le \frac{1}{\mathsf{B}^{d-\delta}}} d\beta\int_{\|\mathbf{x}\|\le \mathsf{B}} e^{2\pi i\beta\mathsf{B}^d f(\frac{\mathbf{x}}{\mathsf{B}})}\mathsf{B}^{n+1}\,d(\frac{\mathbf{x}}{\mathsf{B}}) = \mathsf{B}^{n+1-d}\int_{|\beta'|\le \mathsf{B}^{\delta}}\int_{\|\mathbf{x}'\|\le 1} e^{2\pi i\beta' f(\mathbf{x}')}d(\mathbf{x}').$$

We see the appearance of the main term B^{n+1-d} and the density

$$\tau_\infty := \int_0^1 d\beta'\int_{\|\mathbf{x}\|\le 1} e^{2\pi i\beta' f(\mathbf{x}')}\,d\mathbf{x}'.$$

Now we analyze the *singular series*

$$\tau_Q := \sum_{a,q}\sum_{\mathbf{z}} \frac{e^{2\pi i\frac{a}{q}f(\mathbf{z})}}{q^{n+1}},$$

where the outer sum runs over positive coprime integers a, q, $a < q$ and $q < Q$, and the inner sum over residue classes $\mathbf{z} \in (\mathbb{Z}/q)^{n+1}$. This sum has the following properties:

(1) multiplicativity in q; in particular we have

$$\tau := \prod_p\Big(\sum_{i=0}^{\infty} A(p^i)\Big),$$

with $\tau_Q \to \tau$, for $Q \to \infty$, (with small error term);

(2) and

$$\sum_{i=0}^{k}\frac{A(p^i)}{p^{i(n+1)}} = \frac{\varrho(f,p^k)}{p^{kn}},$$

where

$$\varrho(f,p^k) := \#\{\,\mathbf{z} \mod p^k \mid f(\mathbf{z}) = 0 \mod p^k\,\}.$$

Here, a discrete version of equation (4.3) comes into play:

$$\#\{\text{solutions} \mod p^k\} = \frac{1}{p^k} \sum_{a=0}^{p^k-1} \sum_{\mathbf{z}} e^{2\pi i \frac{af(\mathbf{z})}{p^k}}.$$

However, our sums run over a with $(a,p)=1$. A rearranging of terms leads to

$$\begin{aligned}
\frac{\varrho(f,p^k)}{p^{kn}} &= \sum_{i=0}^{k} \sum_{(a,p^k)=p^i} \sum_{\mathbf{z}} \frac{1}{p^{k(n+1)}} e^{2\pi i \frac{a}{p^k} f(\mathbf{z})} \\
&= \sum_{i=0}^{k} \sum_{(\frac{a}{p^i},p^{k-i})=1} \sum_{\mathbf{z}} \frac{1}{p^{k(n+1)}} e^{2\pi i \frac{a/p^i}{p^{k-i}} f(\mathbf{z})} \cdot p^{(n+1)i} \\
&= \sum_{i=0}^{k} \frac{1}{p^{(n+1)(k-i)}} \sum_{(a,p^{k-i})=1} \sum_{\mathbf{z}} e^{2\pi i \frac{a}{p^{k-i}} f(\mathbf{z})} \\
&= \sum_{i=0}^{k} \frac{1}{p^{(n+1)(k-i)}} \cdot A(p^{k-i}).
\end{aligned}$$

In conclusion,

$$\tau = \prod_p \tau_p, \quad \text{where} \quad \tau_p = \lim_{k\to\infty} \frac{\varrho(f,p^k)}{p^{nk}}. \tag{4.6}$$

As soon as there is at least one (nonsingular) solution $f(\mathbf{z}) = 0 \mod p$, $\tau_p \neq 0$, and in fact, for almost all p,

$$\frac{\varrho(f,p^k)}{p^{kn}} = \frac{\varrho(f,p)}{p^n},$$

by Hensel's lemma. Moreover, if $\tau_p \neq 0$ for all p, the Euler product in equation (4.6) converges.

Let us illustrate this in the example of Fermat type equations

$$f(\mathbf{x}) = a_0 x_0^d + \cdots + a_n x_n^d = 0.$$

Using properties of Jacobi sums one can show that

$$\left| \frac{\varrho(f,p)}{p^n} - 1 \right| \leq \frac{\mathsf{c}}{p^{(n+1)/2}},$$

for some $\mathsf{c} > 0$. The corresponding Euler product

$$\prod_p \frac{\varrho(f,p)}{p^n} \ll \prod_p \left(1 + \frac{\mathsf{c}}{p^{(n+1)/2}}\right)$$

is convergent.

Some historical background: the circle method was firmly established in the series of papers of Hardy and Littlewood *Partitio numerorum*. They comment: *"A method of great power and wide scope, applicable to almost any problem concerning the decomposition of integers into parts of a particular kind, and to many against which it is difficult to suggest any other obvious method of attack."*

4.7. Function fields: heuristics. Here we present Batyrev's heuristic arguments from 1987, which lead to Conjecture 4.10.4. In the function field case, when $F = \mathbb{F}_q(C)$, for some curve C, F-rational points on a projective variety X/F correspond to sections $C \to \mathcal{X}$, where $\mathcal{X}$ is a *model* of X over C. Points of bounded height correspond to sections of bounded degree with respect to an ample line bundle $\mathcal{L}$ over $\mathcal{X}$. Deformation theory allows one to compute the dimension of moduli spaces of sections of fixed degree. The analytic properties of the associated generating series lead to a *heuristic* asymptotic formula.

Let p be a prime and put $q = p^n$. Let Λ be a convex n-dimensional cone in $\mathbb{R}^n$ with vertex at 0. Let

$$f_1, f_2 : \mathbb{R}^n \to \mathbb{R}$$

be two linear functions such that

- $f_i(\mathbb{Z}^n) \subset \mathbb{Z}$;
- $f_2(x) > 0$ for all $x \in \Lambda \setminus \{0\}$;
- there exists an $x \in \Lambda \setminus \{0\}$ such that $f_1(x) > 0$.

For each $\lambda \in \mathbb{Z}^n \cap \Lambda$ let M_λ be a set of cardinality

$$|M_\lambda| := q^{\max(0, f_1(\lambda))},$$

and put $M = \cup_\lambda M_\lambda$. Let

$$\varphi(m) := q^{f_2(\lambda)}, \text{ for } m \in M_\lambda.$$

Then the series

$$\Phi(s) = \sum_{\lambda \in \Lambda \cap \mathbb{Z}^n} \frac{|M_\lambda|}{q^{s f_2(\lambda)}}$$

converges for

$$\Re(s) > a := \max_{x \in \Lambda}(f_1(x)/f_2(x)) > 0.$$

What happens around $s = a$? Choose an $\epsilon > 0$ and decompose the cone

$$\Lambda := \Lambda_\epsilon^+ \cup \Lambda_\epsilon^-,$$

where

$$\begin{array}{rcl} \Lambda_\epsilon^+ & := & \{x \in \Lambda \,|\, f_1(x)/f_2(x) \geq a - \epsilon\} \\ \Lambda_\epsilon^- & := & \{x \in \Lambda \,|\, f_1(x)/f_2(x) < a - \epsilon\} \end{array}$$

Therefore,

$$\Phi(s) = \Phi_\epsilon^+ + \Phi_\epsilon^-,$$

where Φ_ϵ^- converges absolutely for $\Re(s) > a - \epsilon$.

Now we make some assumptions concerning Λ: suppose that for all $\epsilon \in \mathbb{Q}_{>0}$, the cone Λ_ϵ^+ is a rational finitely generated polyhedral cone. Then

$$\Lambda_\epsilon^a := \{x \,|\, f_1(x)/f_2(x) = a\}$$

is a face of Λ_ϵ^+, and thus also finitely generated polyhedral.

LEMMA 4.7.1. *There exists a function $G_\epsilon(s)$, holomorphic for $\Re(s) > a - \epsilon$, such that*

$$\Phi_\epsilon(s) = \frac{G_\epsilon(s)}{(s-a)^b},$$

where b is the dimension of the face Λ_ϵ^a.

PROOF. For $y \in \mathbb{Q}_{>0}$ we put

$$P(y) := \{x \,|\, x \in \Lambda,\ f_2(x) = y\}.$$

Consider the expansion

$$\Phi(s) = \sum_{y\in\mathbb{N}} \sum_{\lambda\in P(y)\cap\mathbb{Z}^n} q^{f_1(\lambda)-sf_2(\lambda)}.$$

Replacing by the integral, we obtain (with $w = yz$)

$$\begin{aligned}
&= \int_0^\infty dy \left(\int_{P(y)} q^{f_1(w)-sf_2(w)} dw \right) \\
&= \int_0^\infty dy \left(\int_{P(1)} y^{n-1} q^{(f_1(z)-sf_2(z))y} dz \right) \\
&= \int_{P(1)} dz \int_0^\infty y^{n-1} q^{(f_1(z)-s)y} dy \qquad (4.7) \\
&= \int_{P(1)} dz \frac{1}{(s-f_1(z))^n} \int_0^\infty u^{n-1} q^{-u} du \\
&= \frac{\Gamma(n)}{(\log(q))^n} \int_{P(1)} \frac{1}{(s-f_1(z))^n} dz.
\end{aligned}$$

It is already clear that we get a singularity at $s = \max(f_1(z))$ on $P(1)$, which is a. In general, let f be a linear function and

$$\Phi(s) := \int_\Delta (s - f(x))^{-n} d\Omega$$

where Δ is a polytope of dimension $n-1$. Then Φ is a rational function in s, with an asymptotic at $s = a$ given by

$$\mathrm{vol}_{f,a} \frac{(b-1)!}{(n-1)!} (s-a)^{-b},$$

where $\Delta_{f,a}$ is the polytope $\Delta \cap \{f(x) = a\}$, $\mathrm{vol}_{f,a}$ is its volume and $b = 1 + \dim(\Delta_{f,a})$. □

Let C be a curve of genus g over the finite field $\mathbb{F}_q$ and F its function field. Let X be a variety over $\mathbb{F}_q$ of dimension n. Then $V := X \times C$ is a variety over F. Every F-rational point x of V gives rise to a section $\tilde{x}$ of the map $V \to C$. We have a pairing

$$\mathrm{A}^1(V) \times \mathrm{A}^n(V) \to \mathbb{Z}$$

between the groups of (numerical) equivalence classes of codimension 1-cycles and codimension n-cycles. We have

$$\mathrm{A}^n(V) = \mathrm{A}^n(X) \otimes \mathrm{A}^1(C) \oplus \mathrm{A}^{n-1}(X) \otimes \mathrm{A}^0(C)$$

and

$$\begin{aligned}
\mathrm{A}^1(V) &= \mathrm{A}^1(X) \oplus \mathbb{Z}, \\
L &= (L_X, \ell), \\
-K_V &= (-K_X, 2-2\mathsf{g}).
\end{aligned}$$

Assume that L is a very ample line bundle on V. Then

$$q^{(L,\tilde{x})}$$

is the height of the point x with respect to L. The height zeta function takes the form (cf. Section 4.9)

$$\begin{aligned}\mathsf{Z}(s) &= \sum_{x\in V(F)} q^{-(L,\tilde{x})s} \\ &= \sum_{y\in \mathrm{A}^n(X)} \tilde{\mathsf{N}}(q) q^{-[(L_X,y)+\ell]s},\end{aligned}$$

where

$$\tilde{\mathsf{N}}(q) := \#\{x \in V(F) \mid \mathrm{cl}(x) = y\}.$$

We proceed to give some *heuristic*(!) bound on $\tilde{\mathsf{N}}(q)$. The cycles in a given class y are parametrized by an algebraic variety M_y and

$$\dim(M_{y(\tilde{x})}) \geq \chi(\mathcal{N}_{V|\tilde{x}})$$

(the Euler characteristic of the normal bundle). More precisely, the local ring on the moduli space is the quotient of a power series ring with $\mathrm{h}^0(\mathcal{N}_{V|\tilde{x}})$ generators by $\mathrm{h}^1(\mathcal{N}_{V|\tilde{x}})$ relations. Our main heuristic assumption is that

$$\tilde{\mathsf{N}}(q) = q^{\dim(M_y)} \sim q^{\chi(\mathcal{N}_{V|\tilde{x}})},$$

modulo smaller order terms. This assumption fails, for example, for points contained in "exceptional" (accumulating) subvarieties.

By the short exact sequence

$$0 \to \mathcal{T}_{\tilde{x}} \to \mathcal{T}_{V|\tilde{x}} \to \mathcal{N}_{V|\tilde{x}} \to 0$$

we have

$$\begin{aligned}\chi(\mathcal{T}_{V|\tilde{x}}) &= (-K_V, \tilde{x}) + (n+1)\chi(\mathcal{O}_{\tilde{x}}), \\ \chi(\mathcal{N}_{V|\tilde{x}}) &= (-K_X, \mathrm{cl}(x)) + n\chi(\mathcal{O}_{\tilde{x}})\end{aligned}$$

From now on we consider a *modified* height zeta function

$$\mathsf{Z}_{\mathrm{mod}}(s) := \sum q^{\chi(\mathcal{N}_{V|\tilde{x}})-(L,\tilde{x})s}.$$

We observe that its analytic properties are determined by the ratio between two linear functions

$$(-K_X, \cdot) \text{ and } (L, \cdot).$$

The relevant cone Λ is the cone spanned by classes of (maximally moving) effective curves. The *finite* generation of this cone for Fano varieties is one of the main results of Mori's theory. Applying the Tauberian theorem 6.1.4 to $\mathsf{Z}_{\mathrm{mod}}(s)$ we obtain the heuristic formula:

$$\mathsf{N}(X^\circ, L, \mathsf{B}) = \mathsf{c}\mathsf{B}^a(\log(\mathsf{B}))^{b-1}(1+o(1)),$$

where

$$a = a(\Lambda, L) = \max_{z\in\Lambda}((-K_X, z)/(L, z))$$

and $b = b(\Lambda, L)$ is the dimension of the face of the cone where this maximum is achieved.

4.8. Metrizations of line bundles. In this section we discuss a refined theory of height functions, based on the notion of an adelically metrized line bundle.

Let F be a number field and $\mathrm{disc}(F)$ the discriminant of F (over $\mathbb{Q}$). The set of places of F will be denoted by $\mathrm{Val}(F)$. We shall write $v|\infty$ if v is archimedean and $v\nmid\infty$ if v is nonarchimedean. For any place v of F we denote by F_v the completion of F at v and by $\mathfrak{o}_v$ the ring of v-adic integers (for $v\nmid\infty$). Let q_v be the cardinality of the residue field $\mathbb{F}_v$ of F_v for nonarchimedean valuations. The local absolute value $|\cdot|_v$ on F_v is the multiplier of the Haar measure, i.e., $d(ax_v)=|a|_v dx_v$ for some Haar measure dx_v on F_v. We denote by $\mathbb{A}=\mathbb{A}_F=\prod_v' F_v$ the adele ring of F. We have the *product formula*

$$\prod_{v\in\mathrm{Val}(F)} |a|_v = 1, \quad \text{for all } \ a\in F^*.$$

Definition 4.8.1. Let X be an algebraic variety over F and L a line bundle on X. A v-adic metric on L is a family $(\|\cdot\|_x)_{x\in X(F_v)}$ of v-adic Banach norms on the fibers L_x such that for all Zariski open subsets $X^\circ\subset X$ and every section $\mathsf{f}\in \mathrm{H}^0(X^\circ, L)$ the map

$$X^\circ(F_v)\to\mathbb{R}, \ \ x\mapsto \|\mathsf{f}\|_x,$$

is continuous in the v-adic topology on $X^\circ(F_v)$.

Example 4.8.2. Assume that L is generated by global sections. Choose a basis $(\mathsf{f}_j)_{j\in[0,\dots,n]}$ of $\mathrm{H}^0(X,L)$ (over F). If f is a section such that $\mathsf{f}(x)\neq 0$ then define

$$\|\mathsf{f}\|_x := \max_{0\le j\le n}\left(|\frac{\mathsf{f}_j}{\mathsf{f}}(x)|_v\right)^{-1},$$

otherwise $\|0\|_x:=0$. This defines a v-adic metric on L. Of course, this metric depends on the choice of $(\mathsf{f}_j)_{j\in[0,\dots,n]}$.

Definition 4.8.3. Assume that L is generated by global sections. An adelic metric on L is a collection of v-adic metrics, for every $v\in\mathrm{Val}(F)$, such that for all but finitely many $v\in\mathrm{Val}(F)$ the v-adic metric on L is defined by means of some *fixed* basis $(\mathsf{f}_j)_{j\in[0,\dots,n]}$ of $\mathrm{H}^0(X,L)$.

We shall write $\|\cdot\|_{\mathbb{A}}:=(\|\cdot\|_v)$ for an adelic metric on L and call a pair $\mathcal{L}=(L,\|\cdot\|_{\mathbb{A}})$ an adelically metrized line bundle. Metrizations extend naturally to tensor products and duals of metrized line bundles, which allows one to define adelic metrizations on arbitrary line bundles L (on projective X): represent L as $L=L_1\otimes L_2^{-1}$ with very ample L_1 and L_2. Assume that L_1, L_2 are adelically metrized. An adelic metrization of L is any metrization which for all but finitely many v is induced from the metrizations on L_1, L_2.

Definition 4.8.4. Let $\mathcal{L}=(L,\|\cdot\|_{\mathbb{A}})$ be an adelically metrized line bundle on X and f an F-rational section of L. Let $X^\circ\subset X$ be the maximal Zariski open subset of X where f is defined and does not vanish. For all $x=(x_v)_v\in X^\circ(\mathbb{A})$ we define the local

$$\mathsf{H}_{\mathcal{L},\mathsf{f},v}(x_v):=\|\mathsf{f}\|_{x_v}^{-1}$$

and the global *height function*

$$\mathsf{H}_{\mathcal{L}}(x):=\prod_{v\in\mathrm{Val}(F)}\mathsf{H}_{\mathcal{L},\mathsf{f},v}(x_v).$$

By the product formula, the restriction of the global height to $X^\circ(F)$ does not depend on the choice of f.

Example 4.8.5. For $X = \mathbb{P}^1 = (x_0 : x_1)$ one has $\mathrm{Pic}(X) = \mathbb{Z}$, spanned by the class $L = [(1:0)]$. For $x = x_0/x_1 \in \mathbb{G}_a(\mathbb{A})$ and $\mathsf{f} = x_1$ we define

$$\mathsf{H}_{\mathcal{L},\mathsf{f},v}(x) = \max(1, |x|_v).$$

The restriction of $\mathsf{H}_{\mathcal{L}} = \prod_v \mathsf{H}_{\mathcal{L},\mathsf{f},v}$ to $\mathbb{G}_a(F) \subset \mathbb{P}^1$ is the usual height on $\mathbb{P}^1$ (with respect to the usual metrization of $\mathcal{L} = \mathcal{O}(1)$).

Example 4.8.6. Let X be an equivariant compactification of a unipotent group G and L a very ample line bundle on X. The space $\mathrm{H}^0(X, L)$, a representation space for G, has a *unique* G-invariant section f, modulo scalars. Indeed, if we had two nonproportional sections, their quotient would be a character of G, which is trivial.

Fix such a section. We have $\mathsf{f}(g_v) \neq 0$, for all $g_v \in G(F_v)$. Put

$$\mathsf{H}_{\mathcal{L},\mathsf{f},v}(g_v) = \|\mathsf{f}(g_v)\|_v^{-1} \quad \text{and } \mathsf{H}_{\mathcal{L},\mathsf{f}} = \prod_v \mathsf{H}_{\mathcal{L},\mathsf{f},v}.$$

By the product formula, the global height is independent of the choice of f.

4.9. Height zeta functions. Let X be an algebraic variety over a global field F, $\mathcal{L} = (L, \|\cdot\|_{\mathbb{A}})$ an adelically metrized ample line bundle on X, $\mathsf{H}_{\mathcal{L}}$ a height function associated to $\mathcal{L}$, X° a subvariety of X, $a_{X^\circ}(\mathcal{L})$ the abscissa of convergence of the height zeta function

$$\mathsf{Z}(X^\circ, \mathcal{L}, s) := \sum_{x \in X^\circ(F)} \mathsf{H}_{\mathcal{L}}(x)^{-s}.$$

Proposition 4.9.1.

(1) *The value of $a_{X^\circ}(\mathcal{L})$ depends only on the class of L in* $\mathrm{NS}(X)$.
(2) *Either $0 \leq a_{X^\circ}(\mathcal{L}) < \infty$, or $a_{X^\circ}(\mathcal{L}) = -\infty$, the latter possibility corresponding to the case of finite $X^\circ(F)$. If $a_{X^\circ}(\mathcal{L}) > 0$ for one ample L then this is so for every ample L.*
(3) *$a_{X^\circ}(\mathcal{L}^m) = \frac{1}{m} a_{X^\circ}(\mathcal{L})$. In general, $a_{X^\circ}(\mathcal{L})$ extends uniquely to a continuous function on $\Lambda_{\mathrm{nef}}(X)^\circ$, which is inverse linear on each half-line unless it identically vanishes.*

Proof. All statements follow directly from the standard properties of heights. In particular,

$$a_{X^\circ}(\mathcal{L}) \leq a(\mathbb{P}^n(F), \mathcal{O}(m)) = \frac{n+1}{m}$$

for some n, m. If $\mathsf{Z}(X^\circ, \mathcal{L}, s)$ converges at some negative s, then it must be a finite sum. Since for two ample heights H, H' we have

$$\mathsf{c}\mathsf{H}^m < \mathsf{H}' < \mathsf{c}'\mathsf{H}^n, \ \ \mathsf{c}, \mathsf{c}', m, n > 0,$$

the value of a can only be simultaneously positive or zero. Finally, if L and L' are close in the (real) topology of $\mathrm{NS}(V)_{\mathbb{R}}$, then $L - L'$ is a linear combination of ample classes with small coefficients, and so $a_{X^\circ}(\mathcal{L})$ is close to $a_{X^\circ}(\mathcal{L}')$. □

By Property (1) of Proposition 4.9.1, we may write $a_{X^\circ}(\mathcal{L}) = a_{X^\circ}(L)$.

Example 4.9.2. For an abelian variety X and ample line bundle L we have

$$\mathsf{H}_{\mathcal{L}}(x) = \exp(\mathsf{q}(x) + \mathsf{l}(x) + O(1)),$$

where q is a positive definite quadratic form on $X(F)\otimes\mathbb{Q}$ and l is a linear form. It follows that $a_X(L) = 0$, although $X(F)$ may well be Zariski dense in X. Also

$$\mathsf{N}(X,\mathcal{L},\mathsf{B}) = \mathsf{c}\log(\mathsf{B})^{r/2}(1+o(1)),$$

where $r = \operatorname{rk} X(F)$. Hence, for $a = 0$, the power of $\log(\mathsf{B})$ in principle cannot be calculated geometrically: it depends on the arithmetic of X and F. The hope is that for $a > 0$ the situation is more stable.

Definition 4.9.3. The *arithmetic hypersurface of linear growth* is

$$\Sigma^{arith}_{X^\circ} := \{L \in \mathrm{NS}(X)_{\mathbb{R}} \,|\, a_{X^\circ}(L) = 1\}.$$

Proposition 4.9.4.

- *If $a_{X^\circ}(L) > 0$ for some L, then $\Sigma^{arith}_{X^\circ}$ is nonempty and intersects each half-line in $\Lambda_{\mathrm{eff}}(X)^\circ$ in exactly one point.*
- *$\Sigma^{<}_{X^\circ} := \{L \,|\, a_{X^\circ}(L) < 1\}$ is convex.*

Proof. The first statement is clear. The second follows from the Hölder inequality: if

$$0 < \sigma, \sigma' \le 1 \quad \text{and} \quad \sigma + \sigma' = 1$$

then

$$\mathsf{H}^{-\sigma}_{\mathcal{L}}(x)\mathsf{H}^{-\sigma'}_{\mathcal{L}'}(x) \le \sigma\mathsf{H}_{\mathcal{L}}(x)^{-1} + \sigma'\mathsf{H}_{\mathcal{L}'}(x)^{-1}$$

so that from $L, L' \in \Sigma^{<}_{X^\circ}$ it follows that $\sigma L + \sigma' L' \in \Sigma^{<}_{X^\circ}$. □

When $\operatorname{rk}\mathrm{NS}(X) = 1$, Σ_{X° is either empty, or consists of one point. Schanuel's theorem 4.2.1 implies that for $\mathbb{P}^n(F)$, this point is the anticanonical class.

Definition 4.9.5. A subvariety $Y \subsetneq X^\circ \subset X$ is called *point accumulating*, or simply *accumulating* (in X° with respect to L), if

$$a_{X^\circ}(L) = a_Y(L) > a_{X^\circ\setminus Y}(L).$$

It is called *weakly accumulating* if

$$a_{X^\circ}(L) = a_Y(L) = a_{X^\circ\setminus Y}(L).$$

Example 4.9.6. If we blow up an F-point of an abelian variety X, the exceptional divisor will be an accumulating subvariety in the resulting variety, although to prove this we must analyze the height with respect to the exceptional divisor, which is not quite obvious.

If $X := \mathbb{P}^{n_1}\times\cdots\times\mathbb{P}^{n_k}$, with $n_j > 0$, then every fiber of a partial projection is weakly accumulating with respect to the anticanonical class.

The role of accumulating subvarieties is different for various classes of varieties, but we will generally try to pinpoint them in a geometric way. For example, on Fano varieties we need to remove the $-K_X$-accumulating subvarieties to ensure stable effects, e.g., the *linear growth conjecture*. Weakly accumulating subvarieties sometimes allow one to obtain lower bounds for the growth rate of $X(F)$ by analyzing subvarieties of smaller dimension (as in Theorem 4.4.2).

4.10. Manin's conjecture. The following picture emerged from the analysis of examples such as $\mathbb{P}^n$, flag varieties, complete intersections of small degree [**FMT89**], [**BM90**].

Let X be a smooth projective variety with ample anticanonical class over a number field F_0. The conjectures below describe the asymptotic of rational points of bounded height in a *stable* situation, i.e., after a sufficiently large finite extension F/F_0 and passing to a sufficiently small Zariski dense subset $X^\circ \subset X$.

Conjecture 4.10.1 (Linear growth conjecture). One has

$$\mathsf{B}^1 \ll \mathsf{N}(X^\circ(F), -K_X, \mathsf{B}) \ll \mathsf{B}^{1+\epsilon}. \tag{4.8}$$

Conjecture 4.10.2 (The power of log).

$$\mathsf{N}(X^\circ(F), -K_X, \mathsf{B}) \asymp \mathsf{B}^1 \log(\mathsf{B})^{r-1}, \tag{4.9}$$

where $r = \operatorname{rk} \operatorname{Pic}(X_F)$.

Conjecture 4.10.3 (General polarizations / linear growth). Every smooth projective X with $-K_X \in \Lambda_{\mathrm{big}}(X)$ has a dense Zariski open subset X° such that

$$\Sigma^{\mathrm{arith}}_{X^\circ} = \Sigma^{\mathrm{geom}}_X,$$

(see Definitions 4.9.3, (1.7)).

The next level of precision requires that $\Lambda_{\mathrm{eff}}(X)$ is a finitely generated polyhedral cone. By Theorem 1.1.5, this holds when X is Fano.

Conjecture 4.10.4 (General polarizations / power of log). For all sufficiently small Zariski open subsets $X^\circ \subset X$ and very ample L one has

$$\mathsf{N}(X^\circ(F), L, \mathsf{B}) \asymp \mathsf{B}^{a(L)} \log(\mathsf{B})^{b(L)-1}, \quad \mathsf{B} \to \infty, \tag{4.10}$$

where $a(L), b(L)$ are the constants defined in Section 1.4.

4.11. Counterexamples. Presently, no counterexamples to Conjecture 4.10.1 are known. However, Conjecture 4.10.2 fails in dimension 3. The geometric reason for this failure comes from Mori fiber spaces, more specifically from "unexpected" jumps in the rank of the Picard group in fibrations.

Let $X \subset \mathbb{P}^n$ be a smooth hypersurface. We know, by Lefschetz, that $\operatorname{Pic}(X) = \operatorname{Pic}(\mathbb{P}^n) = \mathbb{Z}$, for $n \geq 4$. However, this may fail when X has dimension 2. Moreover, the variation of the rank of the Picard group in a family of surfaces X_t over a number field F may be nontrivial, even when *geometrically*, i.e., over the algebraic closure $\bar{F}$ of F, the rank is constant.

The following example appeared in [**BT96b**]: consider a hypersurface $X \subset \mathbb{P}^3_{\mathbf{x}} \times \mathbb{P}^3_{\mathbf{y}}$ given by a form of bidegree (1,3):

$$\sum_{j=0}^{3} x_j y_j^3 = 0.$$

By Lefschetz, the Picard group $\operatorname{Pic}(X) = \mathbb{Z}^2$, with the basis of hyperplane sections of $\mathbb{P}^3_{\mathbf{x}}$, resp. $\mathbb{P}^3_{\mathbf{y}}$, and the anticanonical class is computed as in Example 1.1.2

$$-K_X = (3, 1).$$

Projection onto $\mathbb{P}^3_{\mathbf{y}}$ exhibits X as a $\mathbb{P}^2$-fibration over $\mathbb{P}^3$. The second Mori fiber space structure on X is given by projection to $\mathbb{P}^3_{\mathbf{x}}$, with fibers diagonal cubic surfaces. The restriction of $-K_X$ to each (smooth) fiber $X_{\mathbf{x}}$ is the anticanonical class of the fiber.

The rank $\operatorname{rk}\operatorname{Pic}(X_{\mathbf{x}})$ varies between 1 and 7. For example, if F contains $\sqrt{-3}$, then $\operatorname{rk}\operatorname{Pic}(X_{\mathbf{x}}) = 7$ whenever all x_j are cubes in F. The lower bounds in Section 4.4 show that

$$\mathsf{N}(X^\circ_{\mathbf{x}}, -K_{X_{\mathbf{x}}}, \mathsf{B}) \asymp \mathsf{B}\log(\mathsf{B})^6$$

for all such fibers, all dense Zariski open subsets $X^\circ_{\mathbf{x}}$ and all F. On the other hand, Conjecture 4.10.2 implies that

$$\mathsf{N}(X^\circ, -K_X, \mathsf{B}) \asymp \mathsf{B}\log(\mathsf{B}),$$

for some Zariski open $X^\circ \subset X$, over a sufficiently large number field F. However, every Zariski open subset $X^\circ \subset X$ intersects infinitely many fibers $X_{\mathbf{x}}$ with $\operatorname{rk}\operatorname{Pic}(X_{\mathbf{x}}) = 7$ in a dense Zariski open subset. This is a contradiction.

4.12. Peyre's refinement. The refinement concerns the conjectured asymptotic formula (4.9). Fix a metrization of $-\mathcal{K}_X = (-K_X, \|\cdot\|_{\mathbb{A}})$. The expectation is that

$$\mathsf{N}(X^\circ(F), -\mathcal{K}_X, \mathsf{B}) = c(-\mathcal{K}_X)\cdot \mathsf{B}^1\log(\mathsf{B})^{r-1}(1+o(1)), \quad \text{as} \quad \mathsf{B}\to\infty,$$

with $r = \operatorname{rk}\operatorname{Pic}(X)$. Peyre's achievement was to give a conceptual interpretation of the constant $c(-\mathcal{K}_X)$ [**Pey95**]. Here we explain the key steps of his construction.

Let F be a number field and F_v its v-adic completion. Let X be a smooth algebraic variety over F of dimension d equipped with an adelically metrized line bundle $\mathcal{K} = \mathcal{K}_X = (K_X, \|\cdot\|_{\mathbb{A}})$. Fix a point $x \in X(F_v)$ and let $x_1, \dots, x_d$ be local analytic coordinates in an analytic neighborhood U_x of x giving a homeomorphism

$$\phi : U_x \xrightarrow{\sim} F_v^d.$$

Let $dy_1\wedge\cdots\wedge dy_d$ be the standard differential form on F_v^d and $\mathsf{f} := \phi^*(dy_1\wedge\cdots\wedge dy_d)$ its pullback to U_x. Note that f is a local section of the canonical sheaf K_X and that a v-adic metric $\|\cdot\|_v$ on K_X gives rise to a norm $\|\mathsf{f}(u)\|_v \in \mathbb{R}_{>0}$, for each $u \in U_x$. Let $d\mu_v = dy_1\cdots dy_d$ be the standard Haar measure, normalized by

$$\int_{\mathfrak{o}_v^d} d\mu_v = \frac{1}{\mathfrak{d}_v{}^{d/2}},$$

where $\mathfrak{d}_v$ is the local different (which equals 1 for almost all v).

Define the local v-adic measure $\tilde{\omega}_{\mathcal{K},v}$ on U_x via

$$\int_W \tilde{\omega}_{\mathcal{K},v} = \int_{\phi(W)} \|\mathsf{f}(\phi^{-1}(y))\|_v d\mu_v,$$

for every open $W \subset U_x$. This local measure glues to a measure $\tilde{\omega}_{\mathcal{K},v}$ on $X(F_v)$.

Let $\mathcal{X}$ be a model of X over the integers $\mathfrak{o}_F$ and let v be a place of good reduction. Let $\mathbb{F}_v = \mathfrak{o}_v/\mathfrak{m}_v$ be the corresponding finite field and put $q_v = \#\mathbb{F}_v$. Since X is projective, we have

$$\pi_v : X(F_v) = \mathcal{X}(\mathfrak{o}_v) \to \mathcal{X}(\mathbb{F}_v).$$

We have

$$\begin{aligned}\int_{X(F_v)} \tilde{\omega}_{\mathcal{K},v} &= \sum_{\bar{x}_v \in X(\mathbb{F}_v)} \int_{\pi_v^{-1}(\bar{x}_v)} \tilde{\omega}_{\mathcal{K},v} \\ &= \frac{X(\mathbb{F}_v)}{q_v^d} \\ &= 1 + \frac{\mathrm{Tr}_v(\mathrm{H}^{2d-1}_{\text{ét}}(X_{\bar{\mathbb{F}}_v}))}{\sqrt{q_v}} + \frac{\mathrm{Tr}_v(\mathrm{H}^{2d-2}_{\text{ét}}(X_{\bar{\mathbb{F}}_v}))}{q_v} + \cdots + \frac{1}{q_v^d},\end{aligned}$$

where Tr_v is the trace of the v-Frobenius on the ℓ-adic cohomology of X. Trying to integrate the product measure over $X(\mathbb{A})$ is problematic, since the Euler product

$$\prod_v \frac{X(\mathbb{F}_v)}{q_v^d}$$

diverges. In all examples of interest to us, the cohomology group $\mathrm{H}^{2d-1}_{\text{ét}}(X_{\bar{\mathbb{F}}_v}, \mathbb{Q}_\ell)$ vanishes. For instance, this holds if the anticanonical class is ample. Still the product diverges, since the $1/q_v$ term does not vanish, for projective X. There is a standard regularization procedure: Choose a finite set $S \subset \mathrm{Val}(F)$, including all $v \mid \infty$ and all places of bad reduction. Put

$$\lambda_v = \begin{cases} \mathsf{L}_v(1, \mathrm{Pic}(X_{\bar{\mathbb{Q}}})) & v \notin S \\ 1 & v \in S \end{cases},$$

where $\mathsf{L}_v(s, \mathrm{Pic}(X_{\bar{\mathbb{Q}}}))$ is the local factor of the Artin L-function associated to the Galois representation on the geometric Picard group. Define the regularized *Tamagawa measure*

$$\omega_{\mathcal{K},v} := \lambda_v^{-1} \tilde{\omega}_{\mathcal{K},v}.$$

Write

$$\omega_{\mathcal{K}} := \mathsf{L}_S^*(1, \mathrm{Pic}(X_{\bar{\mathbb{Q}}}))|\mathrm{disc}(F)|^{-d/2} \prod_v \omega_{\mathcal{K},v},$$

where

$$\mathsf{L}_S^*(1, \mathrm{Pic}(X_{\bar{\mathbb{Q}}})) := \lim_{s \to 1} (s-1)^r \mathsf{L}_S^*(s, \mathrm{Pic}(X_{\bar{\mathbb{Q}}}))$$

and r is the rank of $\mathrm{Pic}(X_F)$, and define

$$\tau(-\mathcal{K}_X) := \int_{X(F)} \omega_{\mathcal{K}}. \tag{4.11}$$

Example 4.12.1. Let G be a linear algebraic group over F. It carries an F-rational d-form ω, where $d = \dim(G)$. This form is unique, modulo multiplication by nonzero constants. Fixing ω, we obtain an isomorphism $K_X \simeq \mathcal{O}_G$, the structure sheaf, which carries a natural adelic metrization $(\|\cdot\|_{\mathbb{A}})$.

Let (A, Λ) be a pair consisting of a lattice and a strictly convex (closed) cone in $A_{\mathbb{R}}$: $\Lambda \cap -\Lambda = 0$. Let $(\check{A}, \check{\Lambda})$ be the pair consisting of the dual lattice and the dual cone defined by

$$\check{\Lambda} := \{\check{\lambda} \in \check{A}_{\mathbb{R}} \mid \langle \lambda', \check{\lambda} \rangle \geq 0, \ \ \forall \lambda' \in \Lambda\}.$$

The lattice $\check{A}$ determines the normalization of the Lebesgue measure $d\check{a}$ on $\check{A}_{\mathbb{R}}$ (covolume =1). For $a \in A_{\mathbb{C}}$ define

$$\mathcal{X}_\Lambda(a) := \int_{\check{\Lambda}} e^{-\langle a,\check{a}\rangle} d\check{a}. \tag{4.12}$$

The integral converges absolutely and uniformly for $\Re(a)$ in compacts contained in the interior Λ° of Λ.

Definition 4.12.2. Assume that X is smooth, $\mathrm{NS}(X) = \mathrm{Pic}(X)$ and that $-K_X$ is in the interior of $\Lambda_{\mathrm{eff}}(X)$. We define

$$\alpha(X) := \mathcal{X}_{\Lambda_{\mathrm{eff}}(X)}(-K_X).$$

Remark 4.12.3. This constant measures the volume of the polytope obtained by intersecting the affine hyperplane $(-K_X, \cdot) = 1$ with the *dual* to the cone of effective divisors $\Lambda_{\mathrm{eff}}(X)$ in the dual to the Néron-Severi group. The explicit determination of $\alpha(X)$ can be a serious problem. For Del Pezzo surfaces, these volumes are given in Section 1.9. For example, let X be the moduli space $\bar{\mathcal{M}}_{0,6}$. The dual to the cone $\Lambda_{\mathrm{eff}}(X)$ has 3905 generators (in a 16-dimensional vector space), forming 25 orbits under the action of the symmetric group $\mathbb{S}_6$ [**HT02b**].

Conjecture 4.12.4 (Leading constant). Let X be a Fano variety over F with an adelically metrized anticanonical line bundle $-\mathcal{K}_X = (-K_X, \|\cdot\|_{\mathbb{A}})$. Assume that $X(F)$ is Zariski dense. Then there exists a Zariski open subset $X^\circ \subset X$ such that

$$\mathsf{N}(X^\circ(F), -\mathcal{K}_X, \mathsf{B}) = c(-\mathcal{K}_X)\mathsf{B}^1 \log(\mathsf{B})^{r-1}(1 + o(1)), \tag{4.13}$$

where $r = \mathrm{rk}\,\mathrm{Pic}(X_F)$ and

$$c(-\mathcal{K}_X) = c(X, -\mathcal{K}_X) = \alpha(X)\beta(X)\tau(-\mathcal{K}_X), \tag{4.14}$$

with $\beta(X) := \#\mathrm{Br}(X)/\mathrm{Br}(F)$ (considered in Section 2.4), and $\tau(-\mathcal{K}_X)$ the constant defined in equation 4.11.

4.13. General polarizations. I follow closely the exposition in [**BT98**]. Let E be a finite Galois extension of a number field F such that all of the following constructions are defined over E. Let $(X^\circ, \mathcal{L})$ be a smooth *quasi-projective* d-dimensional variety together with a metrized very ample line bundle $\mathcal{L}$ which embeds X° in some projective space $\mathbb{P}^n$. We denote by $\overline{X}^{\mathcal{L}}$ the normalization of the projective closure of $X \subset \mathbb{P}^n$. In general, $\overline{X}^{\mathcal{L}}$ is singular. We will introduce several notions relying on a resolution of singularities

$$\rho : X \to \overline{X}^{\mathcal{L}}.$$

Naturally, the defined objects will be independent of the choice of the resolution.

For a convex cone $\Lambda \subset \mathrm{NS}(X)_{\mathbb{R}}$ we define

$$a(\Lambda, \mathcal{L}) := a(\Lambda, \rho^*\mathcal{L}).$$

We will always assume that $a(\Lambda_{\mathrm{eff}}(X), \mathcal{L}) > 0$.

Definition 4.13.1. A pair $(X^\circ, \mathcal{L})$ is called *primitive* if there exists a resolution of singularities

$$\rho : X \to \overline{X}^{\mathcal{L}}$$

such that $a(\Lambda_{\text{eff}}(X), \mathcal{L}) \in \mathbb{Q}_{>0}$ and for some $k \in \mathbb{N}$

$$((\rho^*\mathcal{L})^{\otimes a(\Lambda_{\text{eff}}(X),\mathcal{L})} \otimes K_X)^{\otimes k} = \mathcal{O}(D),$$

where D is a *rigid* effective divisor (i.e., $\mathrm{h}^0(X, \mathcal{O}(\nu D)) = 1$ for all $\nu \gg 0$).

Example 4.13.2. of a primitive pair: $(X, -\mathcal{K}_X)$, where X is a smooth projective variety with $-\mathcal{K}_X$ a metrized very ample anticanonical line bundle.

Let $k \in \mathbb{N}$ be such that $a(\Lambda, \mathcal{L})k \in \mathbb{N}$ and consider

$$R(\Lambda, \mathcal{L}) := \bigoplus_{\nu \geq 0} \mathrm{H}^0(X, (((\rho^*\mathcal{L})^{a(\Lambda,\mathcal{L})} \otimes K_X)^{\otimes k})^{\otimes \nu}).$$

In both cases ($\Lambda = \Lambda_{\text{ample}}$ or $\Lambda = \Lambda_{\text{eff}}$) it is expected that $R(\Lambda, \mathcal{L})$ is finitely generated and that we have a fibration

$$\pi = \pi_{\mathcal{L}} : X \to Y^{\mathcal{L}},$$

where $Y^{\mathcal{L}} = \mathrm{Proj}(R(\Lambda, \mathcal{L}))$. For $\Lambda = \Lambda_{\text{eff}}(X)$ the generic fiber of π is expected to be a primitive variety in the sense of Definition 4.13.1. More precisely, there should be a diagram:

$$\rho : \quad \begin{array}{ccc} X & \to & \overline{X}^{\mathcal{L}} \supset X \\ \downarrow & & \\ Y^{\mathcal{L}} & & \end{array}$$

such that:

- $\dim(Y^{\mathcal{L}}) < \dim(X)$;
- there exists a Zariski open $U \subset Y^{\mathcal{L}}$ such that for all $y \in U(\mathbb{C})$ the pair $(X_y, \mathcal{L}_y)$ is primitive (here $X_y = \pi^{-1}(y) \cap X$ and $\mathcal{L}_y$ is the restriction of $\mathcal{L}$ to X_y);
- for all $y \in U(\mathbb{C})$ we have $a(\Lambda_{\text{eff}}(X), \mathcal{L}) = a(\Lambda_{\text{eff}}(X_y), \mathcal{L}_y)$;
- For all $k \in \mathbb{N}$ such that $a(\Lambda_{\text{eff}}(X), \mathcal{L})k \in \mathbb{N}$ the vector bundle

$$\mathcal{L}_k := \mathrm{R}^0\pi_*(((\rho^*\mathcal{L})^{\otimes a(\Lambda_{\text{eff}}(X),\mathcal{L})} \otimes K_X)^{\otimes k})$$

is in fact an ample line bundle on $Y^{\mathcal{L}}$.

Such a fibration will be called an *$\mathcal{L}$-primitive fibration*. A variety may admit several primitive fibrations.

Example 4.13.3. Let $X \subset \mathbb{P}^n_1 \times \mathbb{P}^n_2$ ($n \geq 2$) be a hypersurface given by a bihomogeneous form of bi-degree (d_1, d_2). Both projections $X \to \mathbb{P}^n_1$ and $X \to \mathbb{P}^n_2$ are $\mathcal{L}$-primitive, for appropriate $\mathcal{L}$. In particular, for $n = 3$ and $(d_1, d_2) = (1, 3)$ there are *two* distinct $-\mathcal{K}_X$-primitive fibrations: one onto a point and another onto $\mathbb{P}^3_1$.

4.14. Tamagawa numbers. For smooth projective Fano varieties X with an adelically metrized anticanonical line bundle Peyre defined in [**Pey95**] a Tamagawa number, generalizing the classical construction for linear algebraic groups (see Section 4.12). We need to further generalize this to primitive pairs.

Abbreviate $a(\mathcal{L}) = a(\Lambda_{\text{eff}}(X), \mathcal{L})$ and let $(X, \mathcal{L})$ be a primitive pair such that

$$\mathcal{O}(D) := ((\rho^*\mathcal{L})^{\otimes a(\mathcal{L})} \otimes K_X)^{\otimes k},$$

where k is such that $a(\mathcal{L})k \in \mathbb{N}$ and D is a rigid effective divisor as in Definition 4.13.1. Choose an F-rational section $\mathsf{g} \in \mathrm{H}^0(X, \mathcal{O}(D))$; it is unique up to multiplication by F^*. Choose local analytic coordinates $x_{1,v}, \dots, x_{d,v}$ in a neighborhood U_x of $x \in X(F_v)$. In U_x the section g has a representation

$$\mathsf{g} = \mathsf{f}^{ka(\mathcal{L})}(dx_{1,v} \wedge \cdots \wedge dx_{d,v})^k,$$

where f is a local section of L. This defines a local v-adic measure in U_x by

$$\omega_{\mathcal{L},\mathsf{g},v} := \|\mathsf{f}\|_{x_v}^{a(\mathcal{L})} dx_{1,v} \cdots dx_{d,v},$$

where $dx_{1,v} \cdots dx_{d,v}$ is the Haar measure on F_v^d normalized by $\mathrm{vol}(\mathfrak{o}_v^d) = 1$. A standard argument shows that $\omega_{\mathcal{L},\mathsf{g},v}$ glues to a v-adic measure on $X(F_v)$. The restriction of this measure to $X(F_v)$ does not depend on the choice of the resolution $\rho : X \to \overline{X}^{\mathcal{L}}$. Thus we have a measure on $X(F_v)$.

Denote by $(D_j)_{j\in\mathcal{J}}$ the irreducible components of the support of D and by

$$\mathrm{Pic}(X, \mathcal{L}) := \mathrm{Pic}(X \setminus \bigcup_{j\in\mathcal{J}} D_j).$$

The Galois group Γ acts on $\mathrm{Pic}(X, \mathcal{L})$. Let S be a finite set of places of bad reduction for the data $(\rho, D_j$, etc.), including the archimedean places. Put $\lambda_v = 1$ for $v \in S$, $\lambda_v = \mathsf{L}_v(1, \mathrm{Pic}(X, \mathcal{L}))$ for $v \notin S$ and

$$\omega_{\mathcal{L}} := \mathsf{L}_S^*(1, \mathrm{Pic}(X, \mathcal{L}))|\mathrm{disc}(F)|^{-d/2} \prod_v \lambda_v^{-1} \omega_{\mathcal{L},\mathsf{g},v}.$$

(Here L_v is the local factor of the Artin L-function associated to the Γ-module $\mathrm{Pic}(X, \mathcal{L})$ and $\mathsf{L}_S^*(1, \mathrm{Pic}(X, \mathcal{L}))$ is the residue at $s = 1$ of the partial Artin L-function.) By the product formula, the measure does not depend on the choice of the F-rational section g. Define

$$\tau(X, \mathcal{L}) := \int_{\overline{X(F)}} \omega_{\mathcal{L}},$$

where $\overline{X(F)} \subset X(\mathbb{A})$ is the closure of $X(F)$ in the direct product topology. The convergence of the Euler product follows from

$$\mathrm{h}^1(X, \mathcal{O}_X) = \mathrm{h}^2(X, \mathcal{O}_X) = 0.$$

We have a homomorphism

$$\tilde{\rho} : \mathrm{Pic}(X)_{\mathbb{R}} \to \mathrm{Pic}(X, \mathcal{L})_{\mathbb{R}}$$

and we denote by

$$\Lambda_{\mathrm{eff}}(X, \mathcal{L}) := \tilde{\rho}(\Lambda_{\mathrm{eff}}(X)) \subset \mathrm{Pic}(X, \mathcal{L})_{\mathbb{R}}.$$

Definition 4.14.1. Let $(X, \mathcal{L})$ be a primitive pair as above. Define

$$c(X, \mathcal{L}) := \alpha(X, \mathcal{L})\beta(X, \mathcal{L})\tau(X, \mathcal{L}),$$

where

$$\alpha(X, \mathcal{L}) := \mathcal{X}_{\Lambda_{\mathrm{eff}}(X,\mathcal{L})}(\tilde{\rho}(-K_X)) \quad \text{and} \quad \beta(X, \mathcal{L}) := |\mathrm{H}^1(\Gamma, \mathrm{Pic}(X, \mathcal{L}))|.$$

If $(X, \mathcal{L})$ is not primitive then some Zariski open subset $U \subset X$ admits a primitive fibration: there is a diagram

$$\begin{array}{ccc} X & \to & \overline{X}^{\mathcal{L}} \\ \downarrow & & \\ Y^{\mathcal{L}} & & \end{array}$$

such that for all $y \in Y^{\mathcal{L}}(F)$ the pair $(U_y, \mathcal{L}_y)$ is primitive. Then

$$c(U, \mathcal{L}) := \sum_{y \in Y^0} c(U_y, \mathcal{L}_y), \tag{4.15}$$

where the right side is a possibly infinite, conjecturally(!) convergent sum over the subset $Y^0 \subset Y^{\mathcal{L}}(F)$ of all those fibers U_y where

$$a(\mathcal{L}) = a(\mathcal{L}_y) \text{ and } b(\mathcal{L}) = \operatorname{rk} \operatorname{Pic}(X, \mathcal{L})^{\Gamma} = \operatorname{rk} \operatorname{Pic}(X_y, \mathcal{L}_y)^{\Gamma}.$$

In Section 6 we will see that even if we start with pairs $(X, \mathcal{L})$ where X is a smooth projective variety and $\mathcal{L}$ is a very ample adelically metrized line bundle on X we still need to consider singular varieties arising as fibers of $\mathcal{L}$-primitive fibrations.

It is expected that the invariants of $\mathcal{L}$-primitive fibrations defined above are related to asymptotics of rational points of bounded $\mathcal{L}$-height:

Conjecture 4.14.2 (Leading constant / General polarizations). Let X be a Fano variety over a number field F with an adelically metrized very ample line bundle $\mathcal{L} = (L, \|\cdot\|_{\mathbb{A}})$. Assume that $X(F)$ is Zariski dense. Then there exists a Zariski open subset $X^\circ \subset X$ such that

$$\mathsf{N}(X^\circ(F), \mathcal{L}, \mathsf{B}) = c(X^\circ, \mathcal{L}) \mathsf{B}^{a(\mathcal{L})} \log(\mathsf{B})^{b(\mathcal{L})-1} (1 + o(1)). \tag{4.16}$$

Note that the same variety X may admit several $\mathcal{L}$-primitive fibrations (see Section 4.11). Presumably, there are only finitely many isomorphism types of such fibrations on a given X, at least when X is a Fano variety. Then the recipe would be to consider fibrations with maximal $(a(\mathcal{L}), b(\mathcal{L}))$, ordered lexicographically. In Section 6 we will see many examples of polarized varieties satisfying Conjecture 4.14.2.

4.15. Tamagawa number as a height. Why does the right side of Formula (4.15) converge? The natural idea is to interpret it as a height zeta function, i.e., to think of the Tamagawa numbers of the fibers of an $\mathcal{L}$-primitive fibration as "heights". One problem with this guess is that the "functorial" properties of these notions under field extensions are quite different: Let U_y be a fiber defined over the ground field. The local and global heights of the point on the base $y \in Y^\circ$ don't change under extensions. The local Tamagawa factors of U_y, however, take into account information about $\mathbb{F}_q$-points of U_y, i.e., the density

$$\tau_v = \#U_y(\mathbb{F}_{q_v})/q_v^{\dim(U_y)},$$

for almost all v, which may vary nontrivially.

In the absence of conclusive arguments, let us look at examples. For $\mathbf{x} \in \mathbb{P}^3(\mathbb{Q})$, let $X_{\mathbf{x}} \subset \mathbb{P}^3$ be the diagonal cubic surface fibration

$$x_0y_0^3 + x_1y_1^3 + x_2y_2^3 + x_3y_3^3 = 0, \tag{4.17}$$

considered in Section 4.11. Let $\mathsf{H} : \mathbb{P}^3(\mathbb{Q}) \to \mathbb{R}_{>0}$ be the standard height as in Section 4.1.

THEOREM 4.15.1. [**EJ08b**] *For all $\epsilon > 0$ there exists a $\mathsf{c} = \mathsf{c}(\epsilon)$ such that*

$$\frac{1}{\tau(X_{\mathbf{x}})} \geq \mathsf{c}\,\mathsf{H}\left(\frac{1}{x_0} : \cdots : \frac{1}{x_3}\right)^{1/3-\epsilon}$$

In particular, we have the following fundamental finiteness property: for $\mathsf{B} > 0$ there are finitely many $\mathbf{x} \in \mathbb{P}^3(\mathbb{Q})$ such that $\tau(S_{\mathbf{x}}) > \mathsf{B}$.

A similar result holds for 3 dimensional quartics.

THEOREM 4.15.2. [**EJ07**] *Let $X_{\mathbf{x}}$ be the family of quartic threefolds*

$$x_0y_0^4 + x_1y_1^4 + x_2y_2^4 + x_3y_3^4 + x_4y_4^4 = 0,$$

with $x_0 < 0$ and $x_1, \dots, x_4 > 0$, $x_i \in \mathbb{Z}$. For all $\epsilon > 0$ there exists a $\mathsf{c} = \mathsf{c}(\epsilon)$ such that

$$\frac{1}{\tau(X_{\mathbf{x}})} \geq \mathsf{c}\,\mathsf{H}\left(\frac{1}{x_0} : \cdots : \frac{1}{x_4}\right)^{1/4-\epsilon}.$$

4.16. Smallest points. Let $X \subset \mathbb{P}^n$ be a smooth Fano variety over a number field F. What is the smallest height

$$\mathsf{m} = \mathsf{m}(X(F)) := \min\{\mathsf{H}(x)\}$$

of an F-rational points on X? For a general discussion concerning bounds of solutions of Diophantine equations in terms of the *height* of the equation, see [**Mas02**]. A sample result in this direction is [**Pit71**], [**NP89**]: Let

$$\sum_{i=0}^{n} x_i y_i^d = 0, \tag{4.18}$$

with d odd and let $\mathbf{x} = (x_0, \dots, x_n) \in \mathbb{Z}^{n+1}$ be a vector with nonzero coordinates. For $n \gg d$ (e.g., $n = 2^d + 1$) and any $\epsilon > 0$ there exists a constant c such that (4.18) has a solution $\mathbf{y}$ with

$$\sum_{i=0}^{n} |x_i y_i^d| < \mathsf{c} \prod |x_i|^{d+\epsilon}.$$

For $d \geq 12$, one can work with $n \gg 4d^2 \log(d)$. There have been a several improvements of this result for specific values of d, e.g. [**Cas55**], [**Die03**] for quadrics and [**Bak89**], [**Brü94**] for $d = 3$.

In our setup, the expectation

$$\mathsf{N}(X^\circ(F), -\mathcal{K}_X, \mathsf{B}) = \alpha\beta\tau(-\mathcal{K}_X)\mathsf{B}^1 \log(\mathsf{B})^{r-1}(1 + o(1)),$$

where $r = \operatorname{rk}\operatorname{Pic}(X)$, and the hope that the points are equidistributed with respect to the height leads to the guess that $\mathsf{m}(X)$ is inversely related to $\tau(-\mathcal{K}_X)$, rather than the height of the defining equations. Figure 2 shows the distribution of smallest

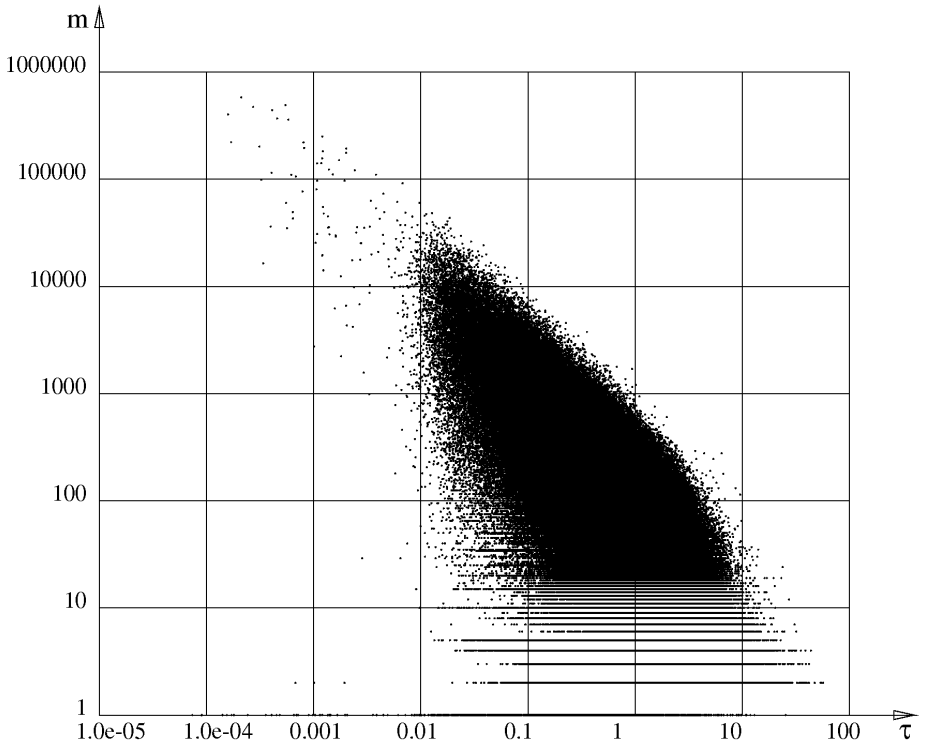


FIGURE 2. Smallest height of a rational point versus the Tamagawa number

points in comparison with the Tamagawa number on a sample of smooth quartic threefolds of the form

$$x_0y^4 = x_1y_1^4 + y_2^4 + y_3^4 + y_4^4, \quad x_0, x_1 = 1, \dots, 1000.$$

On the other hand, there is the following result:

THEOREM 4.16.1. [**EJ07**], [**EJ08b**] *Let $X_x \subset \mathbb{P}^4$ be the quartic threefold given by*

$$xy_0^4 = y_1^4 + y_2^4 + y_3^4 + y_4^4, \quad a \in \mathbb{N}.$$

Then there is no $\mathsf{c} > 0$ such that

$$\mathsf{m}(X_x(\mathbb{Q})) \le \frac{\mathsf{c}}{\tau(-\mathcal{K}_{X_x})}, \quad \forall x \in \mathbb{Z}.$$

Let $X_x \subset \mathbb{P}^3$ be the cubic surface given by

$$xy_0^3 + 4y_1^3 + 2y_2^3 + y_3^3 = 0, \quad x \in \mathbb{N}.$$

Assume the Generalized Riemann Hypothesis. Then there is no $\mathsf{c} > 0$ such that

$$\mathsf{m}(X_x(\mathbb{Q})) \le \frac{\mathsf{c}}{\tau(-\mathcal{K}_{X_x})}, \quad \forall x \in \mathbb{Z}.$$

It may still be the case that

$$\mathsf{m}(X(F)) \leq \frac{\mathsf{c}(\epsilon)}{\tau(-\mathcal{K}_X)^{1+\epsilon}}.$$

5. Counting points via universal torsors

5.1. The formalism. We explain the basic elements of the point counting technique on universal torsors developed in [**Pey04**], [**Sal98**]. The prototype is the projective space:

$$\mathbb{A}^{n+1} \setminus \{0\} \xrightarrow{\mathbb{G}_m} \mathbb{P}^n.$$

The bound $\mathsf{H}(\mathbf{x}) \leq \mathsf{B}$ translates to a bound on $\mathbb{A}^{n+1}(\mathbb{Z})$, it remains to replace the lattice point count on $\mathbb{A}^{n+1}(\mathbb{Z})$ by the volume of the domain. The coprimality on the coordinates leads to the product of local densities formula

$$\mathsf{N}(\mathsf{B}) = \frac{1}{2}\frac{1}{\zeta(n+1)} \cdot \tau_\infty \cdot \mathsf{B}^{n+1}(1+o(1)), \quad \mathsf{B} \to \infty,$$

where τ_∞ is the volume of the unit ball with respect to the norm at infinity.

The lift of points in $\mathbb{P}^n(\mathbb{Q})$ to primitive integral vectors in $\mathbb{Z}^{n+1} \setminus 0$, modulo ± 1 admits a generalization to the context of torsors

$$\mathcal{T}_X \xrightarrow{T_{\mathrm{NS}}} X.$$

Points in $X(\mathbb{Q})$ can be lifted to certain integral points on $\mathcal{T}_X$, uniquely, modulo the action of $T_{\mathrm{NS}}(\mathbb{Z})$ (the analog of the action by ± 1). The height bound on $X(\mathbb{Q})$ lifts to a bound on $\mathcal{T}_X(\mathbb{Z})$. The issue then is to prove, for $\mathsf{B} \to \infty$, that

$$\#\text{ lattice points } \asymp \text{volume of the domain }.$$

The setup for the generalization is as follows. Let X be a smooth projective variety over a number field F. We assume that

- $\mathrm{H}^i(X, \mathcal{O}_X) = 0$, for $i = 1, 2$;
- $\mathrm{Pic}(X_{\bar{F}}) = \mathrm{NS}(X_{\bar{F}})$ is torsion-free;
- $\Lambda_{\mathrm{eff}}(X)$ is a finitely generated rational cone;
- $-K_X$ is in the interior of $\Lambda_{\mathrm{eff}}(X)$;
- $X(F)$ is Zariski dense;
- there is a Zariski open subset without strongly or weakly accumulating subvarieties;
- all universal torsors over X satisfy the Hasse principle and weak approximation.

For simplicity of exposition we will ignore the Galois actions and assume that $\mathrm{NS}(X_F) = \mathrm{NS}(X_{\bar{F}})$. Fix a line bundle L on X and consider the map

$$\begin{array}{ccc} \mathbb{Z} & \to & \mathrm{NS}(X) \\ 1 & \mapsto & [L] \end{array}$$

By duality, we get a homomorphism $\phi_L : T_{\mathrm{NS}} \to \mathbb{G}_m$ and the diagram

$$\begin{array}{ccc} \mathcal{T} & \xrightarrow{\psi_L} & L^* \\ \downarrow & & \downarrow \\ X & = & X \end{array}$$

compatible with the $T_{\rm NS}$-action (where $L^* = L \setminus 0$). Fix a point $t_0 \in \mathcal{T}(F)$ and an adelic metrization $\mathcal{L} = (L, \|\cdot\|_v)$ of L. For each v, we get a map

$$\begin{aligned} \mathsf{H}_{\mathcal{L},\mathcal{T},v} : \mathcal{T}(F_v) &\longrightarrow \mathbb{R}_{>0} \\ t_v &\mapsto \|\psi_L(t)\|_v / \|\psi_L(t_0)\|_v \end{aligned}$$

Fix an adelic height system $\mathsf{H} = \prod_v \mathsf{H}_v$ on X as in Section 4.8, i.e., a basis $L_1, \dots, L_r$ of $\mathrm{Pic}(X)$ and adelic metrizations of these line bundles. This determines compatible adelic metrizations on all $L \in \mathrm{Pic}(X)$. Define

$$\mathcal{T}_{\mathsf{H}_v}(\mathfrak{o}_v) := \{t \in \mathcal{T}(F_v) \,|\, \mathsf{H}_{\mathcal{L},\mathcal{T},v}(t) \le 1 \quad \forall L \in \Lambda_{\rm eff}(X)\}.$$

Let

$$\mathcal{T}_{\mathsf{H}}(\mathbb{A}) := \prod_v \mathcal{T}(F_v)$$

be the *restricted product* with respect to the collection

$$\{\mathcal{T}_{\mathsf{H}_v}(\mathfrak{o}_v)\}_v.$$

This space does not depend on the choice of the points t_0 or on the choice of adelic metrizations.

The next step is the definition of local *Tamagawa measures* on $\mathcal{T}(F_v)$, whose product becomes a global Tamagawa measure on $\mathcal{T}_{\mathsf{H}}(\mathbb{A})$. The main insight is that

- *locally*, in the v-adic topology, $\mathcal{T}(F_v) = X(F_v) \times T_{\rm NS}(F_v)$;
- both factors carry a local Tamagawa measure (defined by the metrizations of the corresponding canonical line bundles);
- the regularizing factor (needed to globalize the measure to the adeles, see Equation 4.11) on $X(F_v)$ is $\lambda_v = \mathsf{L}_v(1, \mathrm{Pic}(X_{\bar{F}})$, for almost all v, and the regularizing factor on $T_{\rm NS}(F_v)$ is λ_v^{-1};
- the regularizing factors cancel and the product measure is integrable over the adelic space $\mathcal{T}_{\mathsf{H}}(\mathbb{A}_F)$.

One chooses a fundamental domain for the action of units $T_{\rm NS}(\mathfrak{o})/W$ (where W is the group of torsion elements), establishes a bijection between the set of rational points $X(F)$ and certain integral points on $\mathcal{T}$ (integral with respect to the unstable locus for the action of $T_{\rm NS}$) in this domain and compares a lattice point count, over these integral points, with the adelic integral, over the space $\mathcal{T}_{\mathsf{H}}(\mathbb{A})$. *If* the difference between these counts goes into the error term, then Conjecture 4.12.4 holds.

The following sections explain concrete realizations of this formalism: examples of universal torsors and counting problems on them.

5.2. Toric Del Pezzo surfaces. A toric surface is an equivariant compactification of the two-dimensional algebraic torus $\mathbb{G}_m^2$. Notation and terminology regarding general toric varieties are explained in Section 6.6. Universal torsors of toric varieties admit a natural embedding into affine space (see Section 1.6).

Example 5.2.1. Let $X = \mathrm{Bl}_Y(\mathbb{P}^2)$ be the blowup of the projective plane in the subscheme

$$Y := (1:0:0) \cup (0:1:0) \cup (0:0:1),$$

a toric Del Pezzo surface of degree 6. We can realize it as a subvariety $X \subset \mathbb{P}^1_{\mathbf{x}} \times \mathbb{P}^1_{\mathbf{y}} \times \mathbb{P}^1_{\mathbf{z}}$ given by $x_0y_0z_0 = x_1y_1z_1$. The anticanonical height is given by

$$\max(|x_0|,|x_1|) \times \max(|y_0|,|y_1|) \times \max(|z_0|,|z_1|).$$

There are six exceptional curves: the preimages of the 3 points and the strict transforms of lines joining two of these points.

Example 5.2.2 (Degree four). There are 3 toric Del Pezzo surfaces of degree 4, given by $X = \{Q_0 = 0\} \cap \{Q = 0\} \subset \mathbb{P}^4$, with $Q_0 = x_0x_1 + x_2^2$ and Q as in the table below.

Singularities	Q
$4\mathsf{A}_1$	$x_3x_4 + x_2^2$
$2\mathsf{A}_1 + \mathsf{A}_2$	$x_1x_2 + x_3x_4$
$2\mathsf{A}_1 + \mathsf{A}_3$	$x_0^2 + x_3x_4$

Example 5.2.3 (Degree three). The unique toric cubic surface X is given by

$$xyz = w^3.$$

The corresponding *fan* is spanned in $\mathbb{Z}^2$ by $(1,1),(1,-2),(-2,1)$. Let $X^\circ = \mathbb{G}_m^2 \subset X$ be the complement to the lines, i.e., the locus with $w \neq 0$. The lines correspond to the three vectors spanning the fan. The universal torsor of the minimal resolution of singularities $\tilde{X}$ of X admits an embedding into $\mathbb{A}^9$, with coordinates corresponding to exceptional curves on $\tilde{X}$. The preimage of X° in $\mathbb{A}^9$ is the complement to the coordinate hyperplanes. The asymptotic

$$\mathsf{N}(X^\circ(F), \mathsf{B}) = c\,\mathsf{B}^1 \log(\mathsf{B})^6(1 + o(1)), \quad \mathsf{B} \to \infty,$$

has been established in [**BT98**] using harmonic analysis (see Section 6.6) and in [**HBM99**], [**Fou98**], [**dlB01**] using the torsor approach.

Example 5.2.4. The toric quartic surface

$$x^2yz = w^4$$

is given by the fan $(2,-1),(0,1),(-2,-1)$. Let X° be the complement to $w = 0$. One has

$$\mathsf{N}(X^\circ(F), \mathsf{B}) = \mathsf{c}\,\mathsf{B}^1 \log(\mathsf{B})^5(1 + o(1)), \quad \mathsf{B} \to \infty,$$

with an explicit constant $\mathsf{c} > 0$ (see Section 6.6). This is more than suggested by the naive heuristic in Section 2.2.

The torsor approach has been successfully implemented for toric varieties over $\mathbb{Q}$ in [**Sal98**] and [**dlB01**].

5.3. Torsors over Del Pezzo surfaces.

Example 5.3.1. A quartic Del Pezzo surface X with two singularities of type A_1 can be realized as a blow-up of the following points

$$\begin{aligned} p_1 &= (0:0:1) \\ p_2 &= (1:0:0) \\ p_3 &= (0:1:0) \\ p_4 &= (1:0:1) \\ p_5 &= (0:1:1) \end{aligned}$$

in $\mathbb{P}^2 = (x_0 : x_1 : x_2)$. The anticanonical line bundle embeds X into $\mathbb{P}^4$:

$$(x_0^2 x_1 : x_0 x_1^2 : x_0 x_1 x_2 : x_0 x_2 (x_0 + x_1 - x_2) : x_1 x_2 (x_0 + x_1 - x_2)).$$

The Picard group is spanned by

$$\mathrm{Pic}(X) = \langle L, E_1, \cdots, E_5 \rangle$$

and $\Lambda_{\mathrm{eff}}(X)$ by

$$\begin{gathered} E_1, \cdots, E_5 \\ L - E_2 - E_3, L - E_3 - E_4, L - E_4 - E_5, L - E_2 - E_5 \\ L - E_1 - E_3 - E_5, L - E_1 - E_2 - E_4. \end{gathered}$$

The universal torsor embeds into the *affine* variety

$$\begin{gathered} (23)(3) - (1)(124)(4) + (25)(5) = 0 \\ (23)(2) - (1)(135)(5) + (34)(4) = 0 \\ (124)(1)(2) - (34)(3) + (45)(5) = 0 \\ (25)(2) - (135)(1)(3) + (45)(4) = 0 \\ (23)(45) + (34)(25) - (1)^2(124)(135) = 0. \end{gathered}$$

(with variables labeled by the corresponding exceptional curves). The complement to the coordinate hyperplanes is a torsor over the complement of the lines on X. Introducing additional variables

$$(24)' := (1)(124), \;\; (35)' := (1)(135)$$

we see that the above equations define a $\mathbb{P}^1$-bundle over a codimension one subvariety of the (affine cone over the) Grassmannian $\mathrm{Gr}(2, 5)$.

We need to estimate the number of 11-tuples of nonzero integers, satisfying the equations above and subject to the inequalities

$$\begin{gathered} |(135)(124)(23)(1)(2)(3)| \le \mathsf{B} \\ |(135)(124)(34)(1)(3)(4)| \le \mathsf{B} \\ \cdots \end{gathered}$$

By symmetry, we can assume that $|(2)| \ge |(4)|$ and write $(2) = (2)'(4) + r_2$. Now we weaken the first inequality to

$$|(135)(124)(23)(1)(4)(2)'(3)| \le \mathsf{B}.$$

There are $O(\mathsf{B}\log(\mathsf{B})^6)$ 7-tuples of integers satisfying this inequality.

Step 1. Use equation $(23)(3) - (1)(124)(4) + (25)(5)$ to reconstruct $(25), (5)$ with ambiguity $O(\log(\mathsf{B}))$.

Step 2. Use $(25)(2) - (135)(1)(3) + (45)(4) = 0$ to reconstruct the residue r_2 modulo (4). Notice that (25) and (4) are "almost" coprime since the corresponding exceptional curves are disjoint.

Step 3. Reconstruct (2) and (45).

Step 4. Use $(23)(2) - (1)(135)(5) + (34)(4)$ to reconstruct (34).

In conclusion, if $X^\circ \subset X$ is the complement to the exceptional curves then

$$\mathsf{N}(X^\circ, -K_X, \mathsf{B}) = O(\mathsf{B}\log(\mathsf{B})^7).$$

We expect that

$$\mathsf{N}(X^\circ, -K_X, \mathsf{B}) = \mathsf{c}\mathsf{B}\log(\mathsf{B})^5(1+o(1))$$

as $B \to \infty$, where c is the constant defined in Section 4.12.

Example 5.3.2. The universal torsor of a smooth quartic Del Pezzo surface, given as a blow-up of the five points

$$\begin{aligned} p_1 &= (1:0:0)\\ p_2 &= (0:1:0)\\ p_3 &= (0:0:1)\\ p_4 &= (1:1:1)\\ p_5 &= (1:a_2:a_3), \end{aligned}$$

assumed to be in general position, is given by the vanishing of polynomials in $\mathbb{A}^{16}$ on the left side of the table below. The right side shows the homogeneous forms defining the D_5-Grassmannian in its Plücker embedding into $\mathbb{P}^{15}$.

$(14)(23)$	$+(12)(34)$	$-(13)(24)$	$(00)(05)-(12)(34)+(13)(24)-(14)(23)$
$(00)(05)$	$+a_3(a_2-1)(12)(34)$	$-a_2(a_3-1)(13)(24)$	
$(23)(03)$	$+(24)(04)$	$-(12)(01)$	$(12)(01)-(23)(03)+(24)(04)-(25)(05)$
$a_2(23)(03)$	$+(25)(05)$	$-(12)(01)$	
$(12)(35)$	$-(13)(25)$	$+(15)(23)$	$(00)(04)-(12)(35)+(13)(25)-(15)(23)$
$(a_2-1)(12)(35)$	$+(00)(04)$	$-(a_3-1)(13)(25)$	
$(12)(45)$	$+(14)(25)$	$-(15)(24)$	$(00)(03)-(12)(45)+(14)(25)-(15)(24)$
$(00)(03)$	$+a_3(14)(25)$	$-(15)(24)$	
$(13)(45)$	$+(14)(35)$	$-(15)(34)$	$(00)(02)-(13)(45)+(14)(35)-(15)(34)$
$(00)(02)$	$+a_2(14)(35)$	$-(15)(34)$	
$(23)(45)$	$+(24)(35)$	$-(25)(34)$	$(00)(01)-(23)(45)+(24)(35)-25)(34)$
$(00)(01)$	$+a_2(24)(35)$	$-a_3(25)(34)$	
$(04)(34)$	$+(02)(23)$	$-(01)(13)$	$(13)(01)-(23)(02)+(34)(04)+35)(05)$
$(05)(35)$	$+a_3(02)(23)$	$-(01)(13)$	
$(a_2-1)(03)(34)$	$+(05)(45)$	$-(a_3-1)(02)(24)$	$(14)(01)-(24)(02)+(34)(03)-(45)(05)$
$(03)(34)$	$+(01)(14)$	$-(02)(24)$	
$(04)(14)$	$+(03)(13)$	$-(02)(12)$	$(12)(02)-(13)(03)+(14)(04)-(15)(05)$
$(05)(15)$	$+a_2(03)(13)$	$-a_3(02)(12)$	
$a_3(02)(25)$	$-a_2(03)(35)$	$-(01)(15)$	$(15)(01)-(25)(02)+(35)(03)-(45)(04)$
$(a_3-1)(02)(25)$	$-(a_2-1)(03)(35)$	$-(04)(45)$	

Connection to the D_5-Grassmannian

Example 5.3.3. The Cayley cubic is the unique cubic hypersurface in $X \subset \mathbb{P}^3$ with 4 double points (A_1-singularities), the maximal number of double points on a cubic surface. It can be given by the equation

$$y_0y_1y_2 + y_0y_1y_3 + y_0y_2y_3 + y_1y_2y_3 = 0.$$

The double points correspond to

$$(1:0:0:0), (0:1:0:0), (0:0:1:0), (0:0:0:1).$$

It can be realized as the blow-up of $\mathbb{P}^2 = (x_1 : x_2 : x_3)$ in the points

$$q_1 = (1:0:0), q_2 = (0:1:0), q_3 = (0:0:1), q_4 = (1:-1:0), q_5 = (1:0:-1), q_6 = (0:1:-1)$$

The points lie on a rigid configuration of 7 lines

$$\begin{array}{cc} x_1 = 0 & (12)(13)(14)(1) \\ x_2 = 0 & (12)(23)(24)(2) \\ x_3 = 0 & (13)(23)(34)(3) \\ x_4 = x_1 + x_2 + x_3 = 0 & (14)(24)(34)(4) \\ x_1 + x_3 = 0 & (13)(24)(13,24) \\ x_2 + x_3 = 0 & (23)(14)(14,23) \\ x_1 + x_2 = 0 & (12)(34)(12,34). \end{array}$$

The proper transform of the line x_j is the (-2)-curve corresponding to (j). The curves corresponding to (ij), (ij,kl) are (-1)-curves. The accumulating subvarieties are exceptional curves. The (anticanonical) embedding $X \hookrightarrow \mathbb{P}^3$ is given by the linear system

$$\begin{aligned} s_1 &= x_1x_2x_3 \\ s_2 &= x_2x_3x_4 \\ s_3 &= x_1x_3x_4 \\ s_4 &= x_1x_2x_4 \end{aligned}$$

The counting problem is: estimate

$$\mathsf{N}(\mathsf{B}) = \#\{(x_1,x_2,x_3) \in \mathbb{Z}^3_{\mathrm{prim}}/\pm, \mid \max_i(|s_i|)/\gcd(s_i) \le \mathsf{B}\},$$

subject to the conditions

$$x_i \neq 0 \ (i = 1,\dots,3), \quad x_j + x_i \neq 0 \ (1 \le i < j \le 3), \quad x_1 + x_2 + x_3 \neq 0.$$

We expect $\sim \mathsf{B}\log(\mathsf{B})^6$ solutions. After dividing the coordinates by their gcd, we obtain

$$\begin{aligned} s_1' &= (1)(2)(3)(12)(13)(23) \\ s_2' &= (2)(3)(4)(23)(24)(34) \\ s_3' &= (1)(3)(4)(13)(14)(34) \\ s_4' &= (1)(2)(4)(12)(14)(24) \end{aligned}$$

These are special sections in the anticanonical series; other decomposable sections are $(1)(2)(12)^2(12,34)$ and $(12,34)(13,24)(14,23)$, for example. Here we use the same notation (i), (ij) etc. for the variables on the universal torsor as for exceptional curves on the minimal resolution $\tilde{X}$ of X. The conic bundles on X produce the following affine equations for the universal torsor:

$$\begin{array}{cccccc} \text{I} & (1)(13)(14) & + & (2)(23)(24) & = & (34)(12,34) \\ \text{II} & (1)(12)(14) & + & (3)(23)(34) & = & (24)(13,24) \\ \text{III} & (2)(12)(24) & + & (3)(13)(34) & = & (14)(14,23) \\ \text{IV} & -(3)(13)(23) & + & (4)(14)(24) & = & (12)(12,34) \\ \text{V} & -(2)(12)(23) & + & (4)(14)(34) & = & (13)(13,24) \\ \text{VI} & -(1)(12)(1) & + & (4)(24)(34) & = & (23)(14,23) \\ \text{VII} & (2)(4)(24)^2 & + & (1)(3)(13)^2 & = & (12,34)(14,23) \\ \text{VIII} & -(1)(2)(12)^2 & + & (3)(4)(34)^2 & = & (13,24)(14,23) \\ \text{IX} & (1)(4)(14)^2 & - & (2)(3)(23)^2 & = & (12,34)(13,24) \end{array}$$

The counting problem is to estimate the number of 13-tuples of *nonzero* integers, satisfying the equations above and subject to the inequality $\max_i\{|s_i'|\} \le \mathsf{B}$. Heath-Brown proved in [**HB03**] that there exist constants $0 < \mathsf{c} < \mathsf{c}'$ such that

$$\mathsf{c}\mathsf{B}\log(\mathsf{B})^6 \le \mathsf{N}(\mathsf{B}) \le \mathsf{c}'\mathsf{B}\log(\mathsf{B})^6.$$

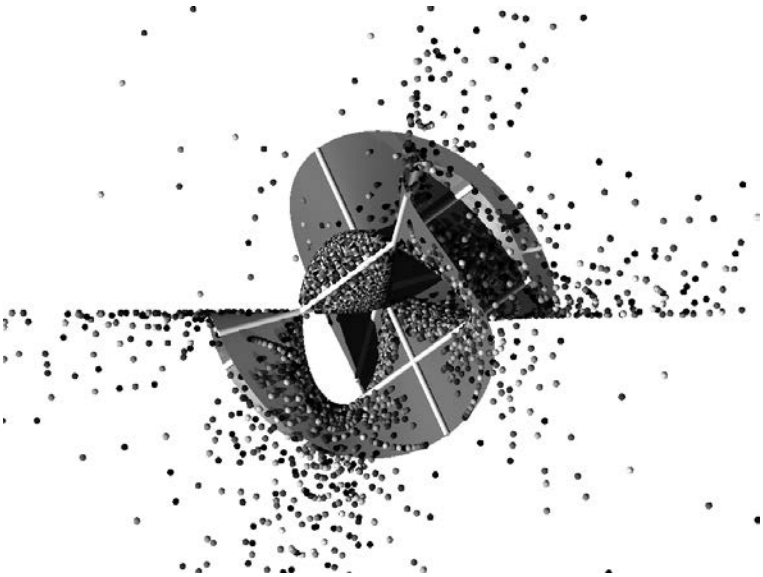


FIGURE 3. The 5332 rational points of height ≤ 100 on
$$x_0x_1x_2 = x_3^2(x_1 + x_2)$$

Example 5.3.4 (The $2\mathsf{A}_2 + \mathsf{A}_3$ cubic surface). The equation
$$x_0x_1x_2 = x_3^2(x_1 + x_2)$$
defines a cubic surface X with singularities of indicated type. It contains 5 lines. The Cox ring has the following presentation [**Der07b**]:
$$\mathrm{Cox}(X) = F[\eta_1, \dots, \eta_{10}]/(\eta_4\eta_6^2\eta_{10} + \eta_1\eta_2\eta_7^2 + \eta_8\eta_9).$$
The figure shows some rational points on this surface. [2] The expected asymptotic
$$\mathsf{N}(X^\circ(\mathbb{Q}), \mathsf{B}) = \mathsf{c}\mathsf{B}^1 \log(\mathsf{B})^6(1 + o(1))$$
on the complement of the 5 lines has not yet been proved.

5.4. Torsors over the Segre cubic threefold. In this section we work over $\mathbb{Q}$. The threefold $X = \bar{\mathcal{M}}_{0,6}$ can be realized as the blow-up of $\mathbb{P}^3$ in the points

	x_0	x_1	x_2	x_3
q_1	1	0	0	0
q_2	0	1	0	0
q_3	0	0	1	0
q_4	0	0	0	1
q_5	1	1	1	1

and in the proper transforms of lines joining two of these points. The Segre cubic is given as the image of X in $\mathbb{P}^4$ under the linear system $2L-(E_1+\cdots+E_5)$ (quadrics

[2] I am grateful to U. Derenthal for allowing me to include it here.

passing through the 5 points):

$$\begin{aligned} s_1 &= (x_2 - x_3)x_1 \\ s_2 &= x_3(x_0 - x_1) \\ s_3 &= x_0(x_1 - x_3) \\ s_4 &= (x_0 - x_1)x_2 \\ s_5 &= (x_1 - x_2)(x_0 - x_3) \end{aligned}$$

It can be realized in $\mathbb{P}^5 = (y_0 : \cdots : y_5)$ as

$$\mathcal{S}_3 := \{\sum_{i=0}^{5} y_i^3 = \sum_{i=0}^{5} y_i = 0\}$$

(exhibiting the $\mathfrak{S}_6$-symmetry.) It contains 15 planes, given by the $\mathfrak{S}_6$-orbit of

$$y_0 + y_3 = y_1 + y_4 = y_3 + y_5 = 0,$$

and 10 singular double points, given by the $\mathfrak{S}_6$-orbit of

$$(1 : 1 : 1 : -1 : -1 : -1).$$

This is the maximal number of nodes on a cubic threefold and $\mathcal{S}_3$ is the unique cubic with this property. The hyperplane sections $\mathcal{S}_3 \cap \{y_i = 0\}$ are *Clebsch* diagonal cubic surfaces (unique cubic surfaces with $\mathfrak{S}_5$ as symmetry group. The hyperplane sections $\mathcal{S}_3 \cap \{y_i - y_j = 0\}$ are *Cayley* cubic surfaces (see Example 5.3.3). The geometry and symmetry of these and similar varieties are described in detail in [**Hun96**]. The counting problem on $\mathcal{S}_3$ is: find the number $\mathsf{N}(\mathsf{B})$ of all 4-tuples of $(x_0, x_1, x_2, x_3) \in \mathbb{Z}^4/\pm$ such that

- $\gcd(x_0, x_1, x_2, x_3) = 1$;
- $\max_{j=1,\dots,5}(|s_j|)/|\gcd(s_1, \dots, s_5) \leq \mathsf{B}$;
- $x_i \neq 0$ and $x_i - x_j \neq 0$ for all $i, j \neq i$.

The last condition is excluding rational points contained in accumulating subvarieties (there are B^3 rational points on planes $\mathbb{P}^2 \subset \mathbb{P}^4$, with respect to the $\mathcal{O}(1)$-height). The second condition is the bound on the *height*.

First we need to determine

$$a(L) = \inf\{a \,|\, aL + K_X \in \Lambda_{\text{eff}}(X)\},$$

where L is the line bundle giving the map to $\mathbb{P}^4$. We claim that $a(L) = 2$. This follows from the fact that

$$\sum_{i,j} (ij)$$

is on the boundary of $\Lambda_{\text{eff}}(X)$ (where (ij) is the class in $\text{Pic}(X)$ of the preimage in X of the line $l_{ij} \subset \mathbb{P}^4$ through q_i, q_j.

Therefore, we expect

$$\mathsf{N}(\mathsf{B}) = O(\mathsf{B}^{2+\epsilon})$$

as $\mathsf{B} \to \infty$. In fact, it was shown in [**BT98**] that $b(L) = 6$. Consequently, one expects

$$\mathsf{N}(\mathsf{B}) = \mathsf{c}\mathsf{B}^2 \log(\mathsf{B})^5(1 + o(1)), \quad \text{as } \mathsf{B} \to \infty.$$

REMARK 5.4.1. The difficult part is to keep track of $\gcd(s_1,\dots,s_5)$. Indeed, if we knew that this gcd $= 1$ we could easily prove the bound $O(\mathsf{B}^{2+\epsilon})$ by observing that there are $O(\mathsf{B}^{1+\epsilon})$ pairs of (positive) integers (x_2-x_3,x_1) (resp. (x_0-x_1,x_2)) satisfying $(x_2-x_3)x_1\le \mathsf{B}$ (resp. $(x_0-x_1)x_2\le \mathsf{B}$). Then we could reconstruct the quadruple

$$(x_2-x_3,x_1,x_0-x_1,x_2)$$

and consequently

$$(x_0,x_1,x_2,x_3)$$

up to $O(\mathsf{B}^{2+\epsilon})$.

Thus it is necessary to introduce gcd between x_j, etc. Again, we use the symbols $(i),(ij),(ijk)$ for variables on the torsor for X corresponding to the classes of the preimages of points, lines, planes resp. Once we fix a point $(x_0,x_1,x_2,x_3)\in \mathbb{Z}^4$ (such that $\gcd(x_0,x_1,x_2,x_3)=1$), the values of these coordinates over the corresponding point on X can be expressed as greatest common divisors. For example, we can write

$$x_3=(123)(12)(13)(23)(1)(2)(3),$$

a product of integers (neglecting the sign of x_3; in the torsor language, we are looking at the orbit of $\mathsf{T}_{\mathrm{NS}}(\mathbb{Z})$). Here is a self-explanatory list:

(123)	x_3	(12)	x_2,x_3
(124)	x_2	(13)	x_1,x_3
(125)	x_2-x_3	(14)	x_1,x_2
(134)	x_1	(15)	x_1-x_3,x_1-x_2
(135)	x_1-x_3	(23)	x_3,x_0-x_3
(145)	x_1-x_2	(24)	x_2,x_0
(234)	x_0	(25)	x_3-x_2,x_0-x_3
(235)	x_0-x_3	(34)	x_1,x_0
(245)	x_0-x_2	(35)	x_1-x_3,x_0-x_1
(345)	x_0-x_1	(45)	x_1-x_2,x_0-x_1.

After dividing s_j by the gcd, we get

$$\begin{aligned}
s_1' &= (125)(134)(12)(15)(25)(13)(14)(34)(1)\\
s_2' &= (123)(245)(12)(13)(23)(24)(25)(45)(2)\\
s_3' &= (234)(135)(23)(24)(34)(13)(15)(35)(3)\\
s_4' &= (345)(124)(34)(35)(45)(12)(14)(24)(4)\\
s_5' &= (145)(235)(14)(15)(45)(23)(35)(25)(5)
\end{aligned}$$

(note the symmetry with respect to the permutation (12345)). We claim that $\gcd(s_1',\dots,s_5')=1$. One can check this directly using the definition of the $(i),(ij)$, and (ijk) as gcd's. For example, let us check that nontrivial divisors $d\neq 1$ of (1) cannot divide any other s_j'. Such a d must divide (123) or (12) or (13) (see s_2'). Assume it divides (12). Then it doesn't divide (13), (14) and (15) (the corresponding divisors are disjoint). Therefore, d divides (135) (by s_3') and (235) (by s_5'). Contradiction (indeed, (135) and (235) correspond to disjoint divisors). Assume that d divides (123). Then it has to divide either (13) or (15) (from s_3') and either (12) or (14) (from s_4'). Contradiction.

The integers $(i), (ij), (ijk)$ satisfy a system of relations (these are equations for the torsor induced from fibrations of $\bar{\mathcal{M}}_{0,6}$ over $\mathbb{P}^1$):

I	x_0	x_1	$x_0 - x_1$
II	x_0	x_2	$x_0 - x_2$
III	x_0	x_3	$x_0 - x_3$
IV	x_1	x_2	$x_1 - x_2$
V	x_1	x_3	$x_1 - x_3$
VI	x_2	x_3	$x_2 - x_3$
VII	$x_0 - x_1$	$x_0 - x_2$	$x_1 - x_2$
VIII	$x_0 - x_1$	$x_0 - x_3$	$x_1 - x_3$
IX	$x_1 - x_2$	$x_1 - x_3$	$x_2 - x_3$
X	$x_2 - x_3$	$x_0 - x_3$	$x_0 - x_2$

which translates to

I	$(234)(23)(24)(2) - (134)(13)(14)(1)$	$=$	$(345)(45)(35)(5)$
II	$(234)(23)(34)(3) - (124)(12)(14)(1)$	$=$	$(245)(25)(45)(5)$
III	$(234)(24)(34)(4) - (123)(12)(13)(1)$	$=$	$(235)(25)(35)(5)$
IV	$(134)(13)(34)(3) - (124)(12)(24)(2)$	$=$	$(145)(15)(45)(5)$
V	$(134)(14)(34)(4) - (123)(12)(23)(2)$	$=$	$(135)(15)(35)(5)$
VI	$(124)(14)(24)(4) - (123)(13)(23)(3)$	$=$	$(125)(15)(25)(5)$
VII	$(345)(34)(35)(3) - (245)(24)(25)(2)$	$=$	$-(145)(14)(15)(1)$
VIII	$(345)(34)(45)(4) - (235)(23)(25)(2)$	$=$	$-(135)(13)(15)(1)$
IX	$(145)(14)(45)(4) - (135)(13)(35)(3)$	$=$	$-(125)(12)(25)(2)$
X	$(125)(12)(15)(1) + (235)(23)(35)(3)$	$=$	$-(245)(24)(45)(4)$

The counting problem now becomes: find all 25-tuples of *nonzero* integers satisfying the equations I – X and the inequality $\max(|s'_j|) \leq \mathsf{B}$.

REMARK 5.4.2. Note the analogy to the case of $\bar{\mathcal{M}}_{0,5}$ (the unique split Del Pezzo surface of degree 5): the variety defined by the above equations is the Grassmannian $\mathrm{Gr}(2,6)$ (in its Plücker embedding into $\mathbb{P}^{24}$).

In [**VW95**] it is shown that there exist constants $\mathsf{c}, \mathsf{c}' > 0$ such that

$$\mathsf{c}\mathsf{B}^2 \log(\mathsf{B})^5 \leq \mathsf{N}(\mathsf{B}) \leq \mathsf{c}'\mathsf{B}^2 \log(\mathsf{B})^5.$$

This uses a different (an intermediate) torsor over X—the determinantal variety given by

$$\det(x_{ij})_{3\times 3} = 0.$$

THEOREM 5.4.3. [**dlB07**]

$$\mathsf{N}(\mathsf{B}) = \frac{1}{24}\tau_\infty \prod_p \tau_p \cdot \mathsf{B}^2 \log(\mathsf{B})^5 \left(1 + O\Big(\frac{(\log\log B)^{1/3}}{(\log B)^{1/3}}\Big)\right),$$

where τ_∞ is the real density of points on X, and

$$\tau_p = \Big(1 - \frac{1}{p}\Big)^6 \Big(1 + \frac{6}{p} + \frac{6}{p^2} + \frac{1}{p^3}\Big)$$

is the p-adic density of points.

The proof of this result uses the Grassmannian $\mathrm{Gr}(2,6)$.

5.5. Flag varieties and torsors. We have seen that for a Del Pezzo surface of degree 5 and for the Segre cubic threefold the universal torsors are flag varieties; and that lifting the count of rational points to these flag varieties yields the expected asymptotic results.

More generally, let G be a semi-simple algebraic group, $P \subset G$ a parabolic subgroup. The flag variety $P\backslash G$ admits an action by any subtorus of the maximal torus in G on the right. Choosing a linearization for this action and passing to the quotient we obtain a plethora of examples of *nonhomogeneous* varieties X whose torsors carry additional symmetries. These may be helpful in the counting rational points on X.

Example 5.5.1. A flag variety for the group G_2 is the quadric hypersurface

$$v_1u_1 + v_2u_2 + v_3u_3 + z^2 = 0,$$

where the torus $\mathbb{G}_m^2 \subset \mathsf{G}_2$ acts as

$$\begin{array}{rcll} v_j & \mapsto & \lambda_j v_j, & j = 1,2 \quad v_3 \mapsto (\lambda_1\lambda_2)^{-1}v_3 \\ u_j & \mapsto & \lambda_j^{-1}u_j, & j = 1,2 \quad u_3 \mapsto \lambda_1\lambda_2 u_3. \end{array}$$

The quotient by $\mathbb{G}_m^2$ is a subvariety in the weighted projective space

$$\mathbb{P}(1,2,2,2,3,3) = (z : x_1 : x_2 : x_3 : y_1 : y_2)$$

with the equations

$$x_0 + x_1 + x_2 + z^2 = 0 \text{ and } x_1x_2x_3 = y_1y_2.$$

6. Analytic approaches to height zeta functions

Consider the variety $X \subset \mathbb{P}^5$ over $\mathbb{Q}$ given by

$$x_0x_1 - x_2x_3 + x_4x_5 = 0.$$

It is visibly a quadric hypersurface and we could apply the circle method as in Section 4.6. It is also the Grassmannian variety $\mathrm{Gr}(2,4)$ and an equivariant compactification of $\mathbb{G}_a^4$. We could count rational points on X taking advantage of any of the underlying structures. In this section we explain counting strategies based on group actions and harmonic analysis.

6.1. Tools from analysis. Here we collect technical results from complex and harmonic analysis which will be used in the treatment of height zeta functions.

For $U \subset \mathbb{R}^n$ let

$$\mathsf{T}_U := \{s \in \mathbb{C}^n \mid \Re(s) \in U\}$$

be the tube domain over U.

Theorem 6.1.1 (Convexity principle). *Let $U \subset \mathbb{R}^n$ be a connected open subset and $\bar{U}$ the convex envelope of U, i.e., the smallest convex open set containing U. Let $\mathsf{Z}(s)$ be a function holomorphic in T_U. Then $\mathsf{Z}(s)$ is holomorphic in $\mathsf{T}_{\bar{U}}$.*

Theorem 6.1.2 (Phragmen-Lindelöf principle). *Let ϕ be a holomorphic function for $\Re(s) \in [\sigma_1, \sigma_2]$. Assume that in this domain ϕ satisfies the following bounds*

- *$|\phi(s)| = O(e^{\epsilon|t|})$, for all $\epsilon > 0$;*
- *$|\phi(\sigma_1 + it)| = O(|t|^{k_1})$ and $|\phi(\sigma_2 + it) = |O(|t|^{k_2})$.*

Then, for all $\sigma \in [\sigma_1, \sigma_2]$ *one has*

$$|\phi(\sigma + it)| = O(|t|^k), \qquad \textit{where } \frac{k - k_1}{\sigma - \sigma_1} = \frac{k_2 - k_1}{\sigma_2 - \sigma_1}.$$

Using the functional equation and known bounds for $\Gamma(s)$ in vertical strips one derives the *convexity bounds*,

$$|\zeta(\frac{1}{2} + it)| = O(|t|^{1/4+\epsilon}), \quad \forall \epsilon > 0, \tag{6.1}$$

and

$$\left|\frac{(s-1)}{s}\zeta(s + it)\right| = O(|t|^{\epsilon}) \text{ for } \Re(s) > 1 - \delta, \tag{6.2}$$

for some sufficiently small $\delta = \delta(\epsilon) > 0$. More generally, we have the following bound for growth rates of Hecke L-functions:

PROPOSITION 6.1.3. *For all* $\epsilon > 0$ *there exists a* $\delta > 0$ *such that*

$$|\mathsf{L}(s, \chi)| \ll (1 + |\Im(\chi)| + |\Im(s)|)^{\epsilon}, \qquad \textit{for } \Re(s) > 1 - \delta, \tag{6.3}$$

for all nontrivial unramified characters χ *of* $\mathbb{G}_m(\mathbb{A}_F)/\mathbb{G}_m(F)$*, i.e.,* χ_v *is trivial on* $\mathbb{G}(\mathfrak{o}_v)$*, for all* $v \nmid \infty$*. Here*

$$\Im(\chi) \in \left(\bigoplus_{v|\infty} \mathbb{G}_m(F_v)/\mathbb{G}_m(\mathfrak{o}_v)\right)^1 \simeq \mathbb{R}^{r_1+r_2-1},$$

with r_1, r_2 *the number of real, resp. pairs of complex embeddings of* F*.*

THEOREM 6.1.4 (Tauberian theorem). *Let* $\{\lambda_n\}$ *be an increasing sequence of positive real numbers, with* $\lim_{n\to\infty} \lambda_n = \infty$*. Let* $\{a_n\}$ *be another sequence of positive real numbers and put*

$$\mathsf{Z}(s) := \sum_{n \geq 1} \frac{a_n}{\lambda_n^s}.$$

Assume that this series converges absolutely and uniformly to a holomorphic function in the tube domain $\mathsf{T}_{>a} \subset \mathbb{C}$*, for some* $a > 0$*, and that it admits a representation*

$$\mathsf{Z}(s) = \frac{h(s)}{(s-a)^b},$$

where h *is holomorphic in* $\mathsf{T}_{>a-\epsilon}$*, for some* $\epsilon > 0$*, with* $h(a) = c > 0$*, and* $b \in \mathbb{N}$*. Then*

$$\mathsf{N}(\mathsf{B}) := \sum_{\lambda_n \leq \mathsf{B}} a_n = \frac{c}{a(b-1)!}\mathsf{B}^a \log(\mathsf{B})^{b-1}(1 + o(1)), \qquad \textit{for } \mathsf{B} \to \infty.$$

A frequently employed result is:

THEOREM 6.1.5 (Poisson formula). *Let* G *be a locally compact abelian group with Haar measure* dg *and* $H \subset G$ *a closed subgroup. Let* $\hat{G}$ *be the Pontryagin dual of* G*, i.e., the group of continuous homomorphisms*

$$\chi : G \to \mathbb{S}^1 \subset \mathbb{C}^*$$

into the unit circle (characters). Let $f : G \to \mathbb{C}$ be a function satisfying some mild assumptions (integrability, continuity) and let

$$\hat{f}(\chi) := \int_G f(g) \cdot \chi(g)\, dg$$

be its Fourier transform. Then there exist Haar measures dh on H and $dh^\perp$ on $H^\perp$, the subgroup of characters trivial on H, such that

$$\int_H f\, dh = \int_{H^\perp} \hat{f}\, dh^\perp. \tag{6.4}$$

A standard application is to $H = \mathbb{Z} \subset \mathbb{R} = G$. In this case $H^\perp = H = \mathbb{Z}$, and the formula reads

$$\sum_{n\in\mathbb{Z}} f(n) = \sum_{n\in\mathbb{Z}} \hat{f}(n).$$

This is a powerful identity which is used, e.g., to prove the functional equation and meromorphic continuation of the Riemann zeta function. We will apply Equation (6.4) in the case when G is the group of adelic points of an algebraic torus or an additive group, and H is the subgroup of rational points. This will allow us to establish a meromorphic continuation of height zeta functions for equivariant compactifications of these groups.

Another application of the Poisson formula arises as follows: Let A be a lattice and Λ a convex cone in $A_{\mathbb{R}}$. Let $d\check{a}$ be the Lebesgue measure on the dual space $\check{A}_{\mathbb{R}}$ normalized by the dual lattice $\check{A}$. Let

$$\mathcal{X}_\Lambda(\mathbf{s}) := \int_{\check{\Lambda}} e^{-\langle \mathbf{s}, \check{a}\rangle} d\check{a}, \quad \Re(\mathbf{s}) \in \Lambda^\circ.$$

be the Laplace transform of the set-theoretic characteristic function of the dual cone $\check{\Lambda}$; it was introduced in Section 4.12. The function $\mathcal{X}_\Lambda$ is holomorphic for $\Re(\mathbf{s})$ contained in the interior Λ° of Λ.

Let $\pi : A \to \tilde{A}$ be a homomorphism of lattices, with finite cokernel A' and kernel $B \subset A$, inducing a surjection $\Lambda \to \tilde{\Lambda}$. Normalize the measure db by $\mathrm{vol}(B_{\mathbb{R}}/B) = 1$. Then

$$\mathcal{X}_{\tilde{\Lambda}}(\pi(\mathbf{s})) = \frac{1}{(2\pi)^{k-\tilde{k}}} \frac{1}{|A'|} \int_{B_{\mathbb{R}}} \mathcal{X}_\Lambda(\mathbf{s} + ib)\, db. \tag{6.5}$$

In particular,

$$\mathcal{X}_\Lambda(\mathbf{s}) = \frac{1}{(2\pi)^d} \int_{M_{\mathbb{R}}} \prod_{j=1}^{n} \frac{1}{(s_j + im_j)}\, dm.$$

6.2. Compactifications of groups and homogeneous spaces. As already mentioned in Section 3, an easy way to generate examples of algebraic varieties with many rational points is to use actions of algebraic groups. Here we discuss the geometric properties of groups and their compactifications.

Let G be a linear algebraic group over a field F, and

$$\varrho : G \to \mathrm{PGL}_{n+1}$$

an algebraic representation over F. Let $x \in \mathbb{P}^n(F)$ be a point. The orbit $\varrho(G)\cdot x \subset \mathbb{P}^n$ inherits rational points from $G(F)$. Let $H \subset G$ be the stabilizer of x. In general, we have an exact sequence

$$1 \to H(F) \to G(F) \to G/H(F) \to \mathrm{H}^1(F, H) \to \cdots$$

We will only consider examples when $(G/H)(F) = G(F)/H(F)$.

By construction, the Zariski closure X of $\varrho(G)\cdot x$ is geometrically isomorphic to an *equivariant* compactification of the homogeneous space G/H. We have a dictionary

$$(\varrho, x \in \mathbb{P}^n) \Leftrightarrow \begin{cases} \text{equivariant compactification } X \supset G/H, \\ G\text{-linearized very ample line bundle } L \text{ on } X. \end{cases}$$

Representations of semi-simple groups do not deform, and can be characterized by combinatorial data: lattices, polytopes, etc. Note, however, that the choice of the initial point $x \in \mathbb{P}^n$ can still give rise to moduli. On the other hand, the classification of representations of unipotent groups is a *wild* problem, already for $G = \mathbb{G}_a^2$. In this case, understanding the moduli of representations of a fixed dimension is equivalent to classifying pairs of commuting matrices, up to conjugacy (see [**GP69**]).

6.3. Basic principles. Here we explain some common features in the study of height zeta functions of compactifications of groups and homogeneous spaces.

In all examples, we have $\mathrm{Pic}(X) = \mathrm{NS}(X)$, a torsion-free abelian group. Choose a basis of $\mathrm{Pic}(X)$ consisting of very ample line bundles $L_1, \ldots, L_r$ and metrizations $\mathcal{L}_j = (L_j, \|\cdot\|_{\mathbb{A}})$. We obtain a *height system*:

$$\mathsf{H}_{\mathcal{L}_j} \,:\, X(F) \to \mathbb{R}_{>0}, \quad \text{for } \; j = 1, \ldots, r,$$

which can be extended to $\mathrm{Pic}(X)_{\mathbb{C}}$, by linearity:

$$\begin{array}{rcl} \mathsf{H} : X(F) \times \mathrm{Pic}(X)_{\mathbb{C}} & \to & \mathbb{R}_{>0}, \\ (x, \mathbf{s}) & \mapsto & \prod_{j=1}^r \mathsf{H}_{\mathcal{L}_j}(x)^{s_j}, \end{array} \tag{6.6}$$

where $\mathbf{s} := \sum_{j=1}^r s_j L_j$. For each j, the 1-parameter zeta function

$$\mathsf{Z}(X, \mathcal{L}_j, s) = \sum_{x \in X(F)} \mathsf{H}_{\mathcal{L}_j}(x)^{-s}$$

converges absolutely to a holomorphic function, for $\Re(s) \gg 0$. It follows that

$$\mathsf{Z}(X, \mathbf{s}) := \sum_{x \in X(F)} \mathsf{H}(x, \mathbf{s})^{-1}$$

converges absolutely to a holomorphic function for $\Re(\mathbf{s})$ contained some cone in $\mathrm{Pic}(X)_{\mathbb{R}}$.

Step 1. One introduces a generalized height pairing

$$\mathsf{H} = \prod_v \mathsf{H}_v \,:\, G(\mathbb{A}) \times \mathrm{Pic}(X)_{\mathbb{C}} \to \mathbb{C}, \tag{6.7}$$

such that the restriction of H to $G(F)\times \mathrm{Pic}(X)$ coincides with the pairing in (6.6). Since X is projective, the height zeta function

$$\mathsf{Z}(g,\mathbf{s})=\mathsf{Z}(X,g,\mathbf{s}):=\sum_{\gamma\in G(F)}\mathsf{H}(\gamma g,\mathbf{s})^{-1} \tag{6.8}$$

converges to a function which is continuous in g and holomorphic in $\mathbf{s}$ for $\Re(\mathbf{s})$ contained in some cone $\Lambda\subset \mathrm{Pic}(X)_{\mathbb{R}}$. The standard height zeta function is obtained by setting $g=e$, the identity in $G(\mathbb{A})$. Our goal is to obtain a meromorphic continuation to the tube domain T over an open neighborhood of $[-K_X]=\kappa\in \mathrm{Pic}(X)_{\mathbb{R}}$ and to identify the poles of Z in this domain.

Step 2. It turns out that

$$\mathsf{Z}(g,\mathbf{s})\in \mathsf{L}^2(G(F)\backslash G(\mathbb{A})),$$

for $\Re(\mathbf{s})\gg 0$. This is immediate in the cocompact case, e.g., for G unipotent or semi-simple anisotropic, and requires an argument in other cases. The L^2-space decomposes into unitary irreducible representations for the natural action of $G(\mathbb{A})$. We get a formal identity

$$\mathsf{Z}(g,\mathbf{s})=\sum_{\varrho}\mathsf{Z}_{\varrho}(g,\mathbf{s}), \tag{6.9}$$

where the summation is over all irreducible unitary representations $(\varrho,\mathcal{H}_\varrho)$ of $G(\mathbb{A})$ occurring in the right regular representation of $G(\mathbb{A})$ in $\mathsf{L}^2(G(F)\backslash G(\mathbb{A}))$.

Step 3. In many cases, the leading pole of $\mathsf{Z}(g,\mathbf{s})$ arises from the trivial representation, i.e., from the integral

$$\int_{G(\mathbb{A}_F)}\mathsf{H}(g,\mathbf{s})^{-1}dg=\prod_v\int_{G(F_v)}\mathsf{H}_v(g_v,\mathbf{s})^{-1}dg_v, \tag{6.10}$$

where dg_v is a Haar measure on $G(F_v)$. To simplify the exposition we assume that

$$X\setminus G=D=\bigcup_{i\in\mathcal{I}}D_i,$$

where D is a divisor with normal crossings whose components D_i are geometrically irreducible.

We choose integral models for X and D_i and observe

$$G(F_v)\subset X(F_v)\xrightarrow{\sim} X(\mathfrak{o}_v)\to X(\mathbb{F}_q)=\bigsqcup_{I\subset\mathcal{I}}D_I^\circ(\mathbb{F}_q),$$

where

$$D_I:=\bigcap_{i\in I}D_i,\ \ D_I^\circ:=D_I\setminus\bigcup_{I'\supsetneq I}D_{I'}.$$

For almost all v, we have

$$\int_{G(F_v)}\mathsf{H}_v(g_v,\mathbf{s})^{-1}dg_v=\tau_v(G)^{-1}\left(\sum_{I\subset\mathcal{I}}\frac{\#D_I^\circ(\mathbb{F}_q)}{q^{\dim(X)}}\prod_{i\in I}\frac{q-1}{q^{s_i-\kappa_i+1}-1}\right), \tag{6.11}$$

where $\tau_v(G)$ is the local Tamagawa number of G and κ_i is the order of the pole of the (unique modulo constants) top-degree differential form on G along D_i. The height

integrals are geometric versions of Igusa's integrals. They are closely related to "motivic" integrals of Batyrev, Kontsevich, Denef and Loeser (see [**DL98**], [**DL99**], and [**DL01**]).

This allows one to regularize explicitly the adelic integral (6.10). For example, for unipotent G we have

$$\int_{G(\mathbb{A}_F)} \mathsf{H}(g,\mathbf{s})^{-1} dg = \prod_i \zeta_F(s_i - \kappa_i + 1) \cdot \Phi(\mathbf{s}), \tag{6.12}$$

with $\Phi(s)$ holomorphic and absolutely bounded for $\Re(s_i) > \kappa_i - \delta$, for all i.

Step 4. Next, one has to identify the leading poles of $\mathsf{Z}_\varrho(g,\mathbf{s})$, and to obtain bounds which are sufficiently uniform in ϱ to yield a meromorphic continuation of the right side of (6.9). This is nontrivial already for abelian groups G (see Section 6.5 for the case when $G = \mathbb{G}_a^n$). Moreover, will need to show *pointwise* convergence of the series, as a function of $g \in G(\mathbb{A})$.

For G abelian, e.g., an algebraic torus, all unitary representation have dimension one, and equation (6.9) is nothing but the usual Fourier expansion of a "periodic" function. The adelic Fourier coefficient is an Euler product, and the local integrals can be evaluated explicitly.

For other groups, it is important to have a manageable parametrization of representations occurring on the right side of the spectral expansion. For example, for unipotent groups such a representation is provided by Kirillov's orbit method (see Section 6.7). For semi-simple groups one has to appeal to Langlands' theory of automorphic representations.

6.4. Generalized flag varieties. The case of generalized flag varieties $X = P\backslash G$ has been treated in [**FMT89**]. Here we will assume that G is a split semi-simple simply connected linear algebraic group over a number field F, and P a parabolic subgroup containing a Borel subgroup P_0 with a Levi decomposition $P_0 = S_0 U_0$. Restriction of characters gives a homomorphism $\mathfrak{X}^*(P) \to \mathfrak{X}^*(P_0)$. Let $\pi : G \to X = P\backslash G$ be the canonical projection. We have an action of P on $G \times \mathbb{A}^1$ via $p \cdot (g,a) \mapsto (pg, \lambda(p)^{-1}a)$. The quotient

$$L_\lambda := P\backslash(G \times \mathbb{A}^1)$$

is a line bundle on X and the assignment $\lambda \mapsto L_\lambda$ gives an isomorphism

$$\mathfrak{X}^*(P) \to \mathrm{Pic}(X).$$

The anticanonical class is given by

$$-K_X = 2\rho_P,$$

the sum of roots of S_0 occurring in the unipotent radical of P. Let Δ_0 be the basis of positive roots of the root system $\Phi(S_0, G)$ determined by P_0. These are labeled by vertices of the Dynkin diagram of G. Let Δ_0^P be the subset of roots orthogonal to $\mathfrak{X}^*(P)$, with respect to the Weyl group invariant intersection form $\langle\, , \rangle$ on $\mathfrak{X}^*(P_0)$, and

$$\Delta_P := \Delta_0 \setminus \Delta_0^P.$$

The cone of effective divisors $\Lambda_{\rm eff}(X) = \Lambda_{\rm nef}(X)$ is the (closure of the) positive Weyl chamber, i.e.,

$$\Lambda_{\rm eff}(X) = \{\lambda \in \mathfrak{X}^*(P)_{\mathbb{R}} \,|\, \langle \lambda, \alpha\rangle \geq 0 \quad \text{for all} \quad \alpha \in \Delta_P\}.$$

Fix a maximal compact subgroup $\mathbf{K}_G = \prod_v \mathbf{K}_{G,v} \subset G(\mathbb{A}_F)$ such that

$$G(\mathbb{A}_F) = P_0(\mathbb{A}_F)\mathbf{K}_G.$$

For $g = pk$, $p = (p_v)_v \in P(\mathbb{A}_F)$ and $k = (k_v)_v \in \mathbf{K}_G$ put

$$H_P(g) := \prod_v H_{P,v}(g_v) \quad \text{with} \quad \langle \lambda, H_{P,v}(g_v)\rangle = \log(|\lambda(p_v)|).$$

This defines an adelically metrized line bundle $\mathcal{L} = \mathcal{L}_\lambda = (L_\lambda, \|\cdot\|_{\mathbb{A}})$ on X by

$$\mathsf{H}_{\mathcal{L}}(x) = e^{-\langle \lambda, H_P(\gamma)\rangle}, \quad \text{with} \quad x = \pi(\gamma). \tag{6.13}$$

The Eisenstein series

$$\mathsf{E}_P^G(s\lambda - \rho_P, g) := \sum_{\gamma \in P(F)\backslash G(F)} e^{\langle s\lambda, H_P(\gamma g)\rangle} \tag{6.14}$$

specializes to the height zeta function

$$\mathsf{Z}(s\lambda) = \mathsf{Z}(X, s\lambda) = \sum_{x\in X(F)} \mathsf{H}_{\mathcal{L}_\lambda}(x)^{-s} = \mathsf{E}_P^G(s\lambda - \rho_P, 1_G).$$

Its analytic properties have been established in [**Lan76**] (see also [**God95**] or [**MW94**, IV, 1.8]); they confirm the conjectures formulated in Section 4.12. The case of function fields is considered in [**LY02**].

6.5. Additive groups. Let X be an equivariant compactification of an additive group $G = \mathbb{G}_a^n$. For example, any blowup $X = {\rm Bl}_Y(\mathbb{P}^n)$, with $Y \subset \mathbb{P}^{n-1} \subset \mathbb{P}^n$, can be equipped with a structure of an equivariant compactification of $\mathbb{G}_a^n$. In particular, the Hilbert schemes of all algebraic subvarieties of $\mathbb{P}^{n-1}$ appear in the moduli of equivariant compactifications X as above. Some features of the geometry of such compactifications have been explored in [**HT99**]. The analysis of height zeta functions has to capture this geometric complexity. In this section we present an approach to height zeta functions developed in [**CLT00a**], [**CLT00b**], and [**CLT02**].

The Poisson formula yields

$$\mathsf{Z}(\mathbf{s}) = \sum_{\gamma\in G(F)} \mathsf{H}(\gamma, \mathbf{s})^{-1} = \int_{G(\mathbb{A}_F)} \mathsf{H}(g, \mathbf{s})^{-1} dg \; + \sum_{\psi\neq\psi_0} \hat{\mathsf{H}}(\psi, \mathbf{s}), \tag{6.15}$$

where the sum runs over all nontrivial characters $\psi \in (G(\mathbb{A}_F)/\mathbb{G}(F))^*$ and

$$\hat{\mathsf{H}}(\psi, \mathbf{s}) = \int_{G(\mathbb{A}_F)} \mathsf{H}(g, \mathbf{s})^{-1}\psi(g) dg \tag{6.16}$$

is the Fourier transform, with an appropriately normalized Haar measure dg.

Example 6.5.1. The simplest case is $G = \mathbb{G}_a = \mathbb{A}^1 \subset \mathbb{P}^1$, over $F = \mathbb{Q}$, with the standard height

$$\mathsf{H}_p(x) = \max(1, |x|_p), \quad \mathsf{H}_\infty(x) = \sqrt{1 + x^2}.$$

We have

$$\mathsf{Z}(s) = \sum_{x\in\mathbb{Q}} \mathsf{H}(x)^{-s} = \int_{\mathbb{A}_{\mathbb{Q}}} \mathsf{H}(x)^{-s}dx + \sum_{\psi\neq\psi_0} \hat{\mathsf{H}}(\psi, s). \tag{6.17}$$

The local Haar measure dx_p is normalized by $\mathrm{vol}(\mathbb{Z}_p) = 1$ so that

$$\mathrm{vol}(|x|_p = p^j) = p^j(1 - \frac{1}{p}).$$

We have

$$\begin{aligned}\int_{\mathbb{Q}_p} \mathsf{H}_p(x)^{-s}dx_p &= \int_{\mathbb{Z}_p} \mathsf{H}_p(x)^{-s}dx_p + \sum_{j\geq 1}\int_{|x|_p=p^j} \mathsf{H}_p(x)^{-s}dx_p \\ &= 1 + \sum_{j\geq 1} p^{-js}\mathrm{vol}(|x|_p = p^j) = \frac{1-p^{-s}}{1-p^{-(s-1)}} \\ \int_{\mathbb{R}}(1+x^2)^{-s/2}dx &= \frac{\Gamma((s-1)/2)}{\Gamma(s/2)}.\end{aligned}$$

Now we analyze the contributions from nontrivial characters. Each such character ψ decomposes as a product of local characters, defined as follows:

$$\begin{aligned}\psi_p = \psi_{p,a_p} &: x_p \mapsto e^{2\pi i a_p\cdot x_p}, \quad a_p \in \mathbb{Q}_p, \\ \psi_\infty = \psi_{\infty,a_\infty} &: x \mapsto e^{2\pi i a_\infty\cdot x}, \quad a_\infty \in \mathbb{R}.\end{aligned}$$

A character is *unramified* at p if it is trivial on $\mathbb{Z}_p$, i.e., $a_p \in \mathbb{Z}_p$. Then $\psi = \psi_a$, with $a \in \mathbb{A}_{\mathbb{Q}}$. A character $\psi = \psi_a$ is unramified for all p iff $a \in \mathbb{Z}$. Pontryagin duality identifies $\hat{\mathbb{Q}}_p = \mathbb{Q}_p$, $\hat{\mathbb{R}} = \mathbb{R}$, and $(\mathbb{A}_{\mathbb{Q}}/\mathbb{Q})^* = \mathbb{Q}$.

Since H_p is invariant under the translation action by $\mathbb{Z}_p$, the local Fourier transform $\hat{\mathsf{H}}_p(\psi_{a_p}, s)$ vanishes unless ψ_p is unramified at p. In particular, only unramified characters are present in the expansion (6.17), i.e., we may assume that $\psi = \psi_a$ with $a \in \mathbb{Z}\setminus 0$. For $p \nmid a$, we compute

$$\hat{\mathsf{H}}_p(s, \psi_a) = 1 + \sum_{j\geq 1} p^{-sj}\int_{|x|_p=p^j} \psi_a(x_p)\, dx_p = 1 - p^{-s}.$$

Putting it all together we obtain

$$\begin{aligned}\mathsf{Z}(s) = &\frac{\zeta(s-1)}{\zeta(s)}\cdot\frac{\Gamma((s-1)/2)}{\Gamma(s/2)} \\ &+ \sum_{a\in\mathbb{Z}}\prod_{p\nmid a}\frac{1}{\zeta_p(s)}\cdot\prod_{p|a}\hat{\mathsf{H}}_p(x_p)^{-s}dx_p\cdot\int_{\mathbb{R}}(1+x^2)^{-s/2}\cdot e^{2\pi i a x}dx\end{aligned}$$

For $\Re(s) > 2-\delta$, we have the upper bounds

$$|\prod_{p|a}\hat{\mathsf{H}}_p(x_p)^{-s}dx_p| \ll |\prod_{p|a}\int_{\mathbb{Q}_p}\mathsf{H}_p(x_p)^{-s}dx_p \ll |a|^{\epsilon} \tag{6.18}$$

$$|\int_{\mathbb{R}}(1+x^2)^{-s/2}\cdot e^{2\pi i a x}dx| \ll_N \frac{1}{(1+|a|)^N}, \quad \text{for any } N\in\mathbb{N}, \tag{6.19}$$

where the second inequality is proved via repeated integration by parts.

Combining these bounds we establish a meromorphic continuation of the right side of Equation (6.17) and thus of $\mathsf{Z}(s)$. It has an isolated pole at $s = 2$ (corresponding to $-K_X = 2L \in \mathbb{Z} = \mathrm{Pic}(\mathbb{P}^1)$). The leading coefficient at this pole is the Tamagawa number defined by Peyre.

Now we turn to the general case. We have seen in Example 1.1.4 that:

- $\mathrm{Pic}(X) = \bigoplus_i \mathbb{Z}D_i$;
- $-K_X = \sum_i \kappa_i D_i$, with $\kappa_i \geq 2$;
- $\Lambda_{\mathrm{eff}}(X) = \bigoplus_i \mathbb{R}_{\geq 0} D_i$.

For each irreducible boundary divisor we let f_i be the unique, modulo scalars, G-invariant section of $\mathrm{H}^0(X, \mathcal{O}(D_i))$. Local and global heights are given as in Definition 4.8.4:

$$\mathsf{H}_{D_i,v}(x) := \|\mathsf{f}_i(x)\|_v^{-1} \quad \text{and} \quad \mathsf{H}_{D_i}(x) = \prod_v \mathsf{H}_{D_i,v}(x).$$

A key fact is that the local heights are invariant under the action of a compact subgroup $\mathbf{K}_{G,v} \subset G(F_v)$, $v \nmid \infty$, with $\mathbf{K}_{G,v} = G(\mathfrak{o}_v)$, for almost all such v. We get a *height pairing*:

$$\begin{array}{rccc} \mathsf{H} & : G(\mathbb{A}_F) \times \mathrm{Pic}(X)_{\mathbb{C}} & \to & \mathbb{C} \\ & (x, \sum_i s_i D_i) & \mapsto & \prod_i \mathsf{H}_{D_i}(x)^{s_i} \end{array}$$

The main term in (6.15) is computed in (6.12); its analytic properties, i.e., location and order of poles, leading constants at these poles, are in accordance with Conjectures in Section 4.10.

We now analyze the "error terms" in equation (6.15), i.e., the contributions from nontrivial characters. A character of $\mathbb{G}_a^n(\mathbb{A}_F)$ is determined by a "linear form" $\langle \mathbf{a}, \cdot \rangle = f_{\mathbf{a}}$, on $\mathbb{G}_a^n$, which gives a rational function $f_{\mathbf{a}} \in F(X)^*$. We have

$$\mathrm{div}(f_{\mathbf{a}}) = E_{\mathbf{a}} - \sum_{i \in \mathcal{I}} d_i(f_{\mathbf{a}}) D_i$$

with $d_i \geq 0$, for all i. Put

$$\mathcal{I}_0(\mathbf{a}) := \{i \,|\, d_i(f_{\mathbf{a}}) = 0\}.$$

Only the trivial character has $\mathcal{I}_0 = \mathcal{I}$. The computation of local integrals in (6.16) is easier at places of good reduction of $f_{\mathbf{a}}$. The contribution to $\hat{\mathsf{H}}(\psi_{\mathbf{a}}, \mathbf{s})$ from nonarchimedean places of bad reduction $S(\mathbf{a}) \subset \mathrm{Val}(F)$ admits an *a priori* bound, replacing the integrand by its absolute value. Iterated integration by parts at archimedean places allows one to establish [**CLT02**, Corollary 10.5]:

$$\hat{\mathsf{H}}(\psi_{\mathbf{a}}, \mathbf{s}) = \prod_{i \in \mathcal{I}_0(\mathbf{a})} \zeta_F(s_i - \kappa_i + 1) \cdot \Phi_{\mathbf{a}}(\mathbf{s}),$$

with $\Phi_{\mathbf{a}}(\mathbf{s})$ holomorphic for $\Re(s_i) > \kappa_i - 1/2 + \epsilon$, $\epsilon > 0$, and bounded by

$$\mathsf{c}(\epsilon, N)(1 + \|\mathbf{s}\|)^{N'}(1 + \|\mathbf{a}\|_\infty)^{-N} \quad N, N' \in \mathbb{N}.$$

We have

$$\hat{\mathsf{H}}(\psi_{\mathbf{a}}, \mathbf{s}) = 0,$$

unless $\psi_{\mathbf{a}}$ is trivial on $\mathbf{K}_{G,f} := \prod_{v \nmid \infty} \mathbf{K}_{G,v}$. Thus only unramified characters $\psi_{\mathbf{a}}$, i.e., with $\mathbf{a}$ in a *lattice*, contribute to the Poisson formula 6.15. One obtains

$$\begin{aligned} \mathsf{Z}(\mathbf{s}) &= \int_{G(\mathbb{A}_F)} \mathsf{H}(g,\mathbf{s})^{-1} dg + \sum_{\mathcal{I}_0 \subsetneq \mathcal{I}} \sum_{\psi_{\mathbf{a}} : \mathcal{I}_0(\mathbf{a}) = \mathcal{I}} \hat{\mathsf{H}}(\psi_{\mathbf{a}}, \mathbf{s}) \\ &= \prod_{i \in \mathcal{I}} \zeta_F(s_i - \kappa_i + 1) \cdot \Phi(\mathbf{s}) + \sum_{I \subsetneq \mathcal{I}} \prod_{i \in I} \zeta_F(s_i - \kappa_i + 1) \cdot \tilde{\Phi}_I(\mathbf{s}), \end{aligned}$$

where $\Phi(\mathbf{s})$ and $\tilde{\Phi}_I(\mathbf{s})$ are holomorphic in $\mathbf{s}$ and admit bounds in vertical strips. Restricting to the line $s(-K_X)$ shows that the pole of highest order is contributed by the first term, i.e., by the trivial character. The leading coefficient at this pole is the adelic height integral; matching of local measures proves Conjecture 4.12.4. Consider the restriction to lines sL, for other L in the interior of $\Lambda_{\mathrm{eff}}(X)$. Let $\mathcal{I}$ correspond to the face of $\Lambda_{\mathrm{eff}}(X)$ which does not contain $a(L)L + K_X$ in its interior. The poles of $\mathsf{Z}(sL)$ are at the predicted value $s = a(L)$, of order $\leq b(L)$. They arise from those characters $\psi_{\mathbf{a}}$ which have $I \subseteq \mathcal{I}_0(\mathbf{a})$. These characters form a subgroup of the group of characters of $G(\mathbb{A}_F)/G(F)$. To show that the sum of the coefficients at these poles does not vanish one applies the Poisson formula (6.4) to this subgroup:

$$\sum_{\psi_{\mathbf{a}}, I \subseteq \mathcal{I}_0(\mathbf{a})} \hat{\mathsf{H}}(\psi_{\mathbf{a}}, sL) = \int_{\mathrm{Ker}} \mathsf{H}(g, sL)^{-1} dg,$$

where Ker is the common kernel of these characters. One can identify

$$\mathrm{Ker} = (G/G_I)(F) \cdot G_I(\mathbb{A}_F),$$

where $G_I \subset G$ is the subgroup defined by the vanishing of the linear forms $\langle \mathbf{a}, \cdot \rangle$. The geometric interpretation of this sum, and of the leading coefficient at the pole, leads to the formalism developed in Section 4.14, specifically to Equation (4.15). In particular, Conjecture 4.14.2 holds.

6.6. Toric varieties. Analytic properties of height zeta functions of toric varieties have been established in [**BT95**], [**BT98**], and [**BT96a**].

An algebraic torus is a linear algebraic group T over a field F such that

$$T_E \simeq \mathbb{G}_{m,E}^d$$

for some finite Galois extension E/F. Such an extension is called a *splitting field* of T. A torus is *split* if $T \simeq \mathbb{G}_{m,F}^d$. The group of algebraic characters

$$M := \mathfrak{X}^*(T) = \mathrm{Hom}(T, E^*)$$

is a torsion-free $\Gamma := \mathrm{Gal}(E/F)$-module. The standard notation for its dual, the cocharacters, is $N := \mathfrak{X}_*(T)$. There is an equivalence of categories:

$$\left\{ \begin{array}{l} d\text{-dimensional integral} \\ \Gamma\text{-representations,} \\ \text{up to equivalence} \end{array} \right\} \Leftrightarrow \left\{ \begin{array}{l} d\text{-dimensional} \\ \text{algebraic tori, split over } E, \\ \text{up to isomorphism} \end{array} \right\}.$$

The local and global theory of tori can be summarized as follows: The local Galois groups $\Gamma_v := \mathrm{Gal}(E_w/F_v) \subset \Gamma$ act on M. Put

$$M_v := M^{\Gamma_v} \quad v \nmid \infty, \ \text{ resp. } \ M_v := M^{\Gamma_v} \otimes \mathbb{R} \quad v \mid \infty,$$

and let N_v be the dual groups. Write $\mathbf{K}_{T,v} \subset T(F_v)$ for the maximal compact subgroup (after choosing an integral model, $\mathbf{K}_{T,v} = T(\mathfrak{o}_v)$, the group of $\mathfrak{o}_v$-valued points of T, for almost all v). Then

$$T(F_v)/\mathbf{K}_{T,v} \hookrightarrow N_v = N^{\Gamma_v}, \tag{6.20}$$

and this map is an isomorphism for v unramified in E/F. Adelically, we have

$$T(\mathbb{A}_F) \supset T^1(\mathbb{A}_F) = \{t \mid \prod_v |m(t_v)|_v = 1 \ \ \forall m \in M^\Gamma\}$$

and

$$T(F) \hookrightarrow T^1(\mathbb{A}_F).$$

THEOREM 6.6.1. *We have*

- $T(\mathbb{A}_F)/T^1(\mathbb{A}_F) = N_{\mathbb{R}}^\Gamma$;
- $T^1(\mathbb{A}_F)/T(F)$ *is compact;*
- $\mathbf{K}_T \cap T(F)$ *is finite;*
- *the homomorphisms* $(T(\mathbb{A}_F)/\mathbf{K}_T \cdot T(F))^* \to \bigoplus_{v|\infty} M_v \otimes \mathbb{R}$ *has finite kernel and image a direct sum of a lattice with* $M_{\mathbb{R}}^\Gamma$.

Over algebraically closed fields, complete *toric varieties*, i.e., equivariant compactifications of algebraic tori, are described and classified by a combinatorial structure (M, N, Σ), where $\Sigma = \{\sigma\}$ is a *fan*, i.e., a collection of strictly convex cones in $N_{\mathbb{R}}$ such that

(1) $0 \in \sigma$ for all $\sigma \in \Sigma$;
(2) $N_{\mathbb{R}} = \bigcup_{\sigma\in\Sigma} \sigma$;
(3) every face $\tau \subset \sigma$ is in Σ;
(4) $\sigma \cap \sigma' \in \Sigma$ and is face of σ, σ'.

A fan Σ is called *regular* if the generators of every $\sigma \in \Sigma$ form part of a basis of N. In this case, the corresponding toric variety X_Σ is smooth. The toric variety is constructed as follows:

$$X_\Sigma := \bigcup_\sigma U_\sigma \ \text{ where } \ U_\sigma := \mathrm{Spec}(F[M \cap \check{\sigma}]),$$

and $\check{\sigma} \subset M_{\mathbb{R}}$ is the cone dual to $\sigma \subset N_{\mathbb{R}}$. The fan Σ encodes all geometric information about X_Σ. For example, 1-dimensional generators $e_1, \dots, e_n$ of Σ correspond to boundary divisors $D_1, \dots, D_n$, i.e., the irreducible components of $X_\Sigma \setminus T$. There is an explicit criterion for projectivity and a description of the cohomology ring, cellular structure, etc.

Over nonclosed ground fields F one has to account for the action of the Galois group of a splitting field E/F. The necessary modifications can be described as follows. The Galois group Γ acts on M, N, Σ. A fan Σ is called Γ-invariant if $\gamma \cdot \sigma \in \Sigma$, for all $\gamma \in \Gamma, \sigma \in \Sigma$. If Σ is a complete regular Γ-invariant fan such that, over the splitting field, the resulting toric variety $X_{\Sigma,E}$ is projective, then it can be descended to a complete algebraic variety $X_{\Sigma,F}$ over the ground field F such that

$$X_{\Sigma,E} \simeq X_{\Sigma,F} \otimes_{\mathrm{Spec}(F)} \mathrm{Spec}(E),$$

as E-varieties with Γ-action. Let $PL(\Sigma)$ be the group of piecewise linear $\mathbb{Z}$-valued functions φ on Σ. An element $\varphi \in PL(\Sigma)$ is determined by a collection of linear

functions $\{m_{\sigma,\varphi}\}_{\sigma\in\Sigma} \subset M$, i.e., by its values $s_j := \varphi(e_j)$, $j = 1, \dots, n$, and we may write $\varphi = \varphi_{\mathbf{s}}$, with $\mathbf{s} = (s_1, \dots, s_n)$. Note that, over a splitting field E,

$$PL(\Sigma) \simeq \mathrm{Pic}^T(X_\Sigma)$$

the group of isomorphism classes of T-linearized line bundles on X_Σ. We have an exact sequence of Γ-modules

$$0 \to M \to PL(\Sigma) \xrightarrow{\pi} \mathrm{Pic}(X_\Sigma) \to 0$$

which leads to

$$0 \to M^\Gamma \to PL(\Sigma)^\Gamma \xrightarrow{\pi} \mathrm{Pic}(X_\Sigma)^\Gamma \to \mathrm{H}^1(\Gamma, M) \to 0$$

This reflects the fact that every divisor is equivalent to a linear combination of boundary divisors $D_1, \dots, D_n$, and φ is determined by its values on $e_1, \dots, e_n$; relations come from characters of T (see Example 1.1.4). The cone of effective divisors is given by:

$$\Lambda_{\mathrm{eff}}(X_\Sigma) = \pi(\mathbb{R}_{\geq 0}D_1 + \cdots + \mathbb{R}_{\geq 0}D_n)$$

and the anticanonical class is

$$-K_\Sigma = \pi(D_1 + \cdots + D_n).$$

Example 6.6.2. Consider the simplest toric variety $\mathbb{P}^1 = \{(x_0 : x_1)\}$, an equivariant compactification of $\mathbb{G}_m$. We have three distinguished Zariski open subsets:

- $\mathbb{P}^1 \supset \mathbb{G}_m = \mathrm{Spec}(F[x, x^{-1}]) = \mathrm{Spec}(F[x^{\mathbb{Z}}])$
- $\mathbb{P}^1 \supset \mathbb{A}^1 = \mathrm{Spec}(F[x]) = \mathrm{Spec}(F[x^{\mathbb{Z}_{\geq 0}}])$,
- $\mathbb{P}^1 \supset \mathbb{A}^1 = \mathrm{Spec}(F[x^{-1}]) = \mathrm{Spec}(F[x^{\mathbb{Z}_{\leq 0}}])$

They correspond to the semigroups:

$$\mathbb{Z} - \text{ dual to } 0, \quad \mathbb{Z}_{\geq 0} - \text{ dual to } \mathbb{Z}_{\geq 0}, \quad \mathbb{Z}_{\leq 0} - \text{ dual to } \mathbb{Z}_{\leq 0}.$$

The local heights can be defined combinatorially, via the introduced explicit charts: on

$$\mathsf{H}_v(x) := \begin{cases} |\frac{x_0}{x_1}|_v & \text{if } |x_0|_v \geq |x_1|_v \\ |\frac{x_1}{x_0}|_v & \text{otherwise.} \end{cases}$$

As usual,

$$\mathsf{H}(x) := \prod_v \mathsf{H}_v(x).$$

In general, for $\varphi \in PL(\Sigma)^\Gamma$, $L = L_\varphi$ and $x = (x_v)_v \in T(F_v)$ define

$$\mathsf{H}_{\mathcal{L},v}(x_v) = \mathsf{H}_{\Sigma,v}(x_v, \varphi) := q_v^{\varphi(\bar{x}_v)}, \qquad \mathsf{H}_\Sigma(x, \varphi) := \prod_v \mathsf{H}_{\Sigma,v}(x, \varphi),$$

where $\bar{x}_v$ is the image of x_v under the homomorphism (6.20), with $q_v = e$, for $v \mid \infty$. One can check that these formulas define an adelic metrization on the T-linearized line bundle $L = L_\varphi$. More generally, for $t = (t_v)_v \in T(\mathbb{A}_F)$ one can define the *t-twisted* adelic metrization $\mathcal{L}(t)$ of $\mathcal{L} = (L, \|\cdot\|)$ via

$$\mathsf{H}_{\mathcal{L}(t),v} = \mathsf{H}_{\Sigma,v}(x_v t_v). \tag{6.21}$$

The product formula implies that

$$\mathsf{H}_{\mathcal{L}(t)} = \mathsf{H}_{\mathcal{L}}, \quad \text{for } t \in T(F).$$

Th height pairing $\mathsf{H}_\Sigma : T(\mathbb{A}_F) \times PL(\Sigma)^\Gamma_{\mathbb{C}} \to \mathbb{C}$ has the following properties:

- its restriction to $T(F) \times PL(\Sigma)^\Gamma_{\mathbb{C}}$ descends to a well-defined pairing
$$T(F) \times \mathrm{Pic}(X_\Sigma)^\Gamma_{\mathbb{C}} \to \mathbb{C};$$
- it is $\mathbf{K}_{T,v}$-invariant, for all v.

The height zeta function

$$\mathsf{Z}_\Sigma(\mathbf{s}) := \sum_{x \in T(F)} \mathsf{H}_\Sigma(x, \varphi_{\mathbf{s}})^{-1}$$

can be analyzed via the Poisson formula (6.4)

$$\mathsf{Z}_\Sigma(\mathbf{s}) := \int_{(T(\mathbb{A}_F)/\mathbf{K}_T \cdot T(F))^*} \hat{\mathsf{H}}_\Sigma(\chi, \mathbf{s})\, d\chi, \tag{6.22}$$

where

$$\hat{\mathsf{H}}_\Sigma(\chi, \mathbf{s}) := \int_{T(\mathbb{A}_F)} \mathsf{H}_\Sigma(x, \varphi_{\mathbf{s}})^{-1} \chi(x)\, dx,$$

and the Haar measure is normalized by $\mathbf{K}_T$. As before, the Fourier transform vanishes for characters χ which are nontrivial on $\mathbf{K}_T$. The integral converges absolutely for $\Re(s_j) > 1$ (for all j), and the goal is to obtain its meromorphic continuation to the left of this domain.

Example 6.6.3. Consider the projective line $\mathbb{P}^1$ over $\mathbb{Q}$. We have

$$0 \to M \to PL(\Sigma) \to \mathrm{Pic}(\mathbb{P}^1) \to 0$$

with $M = \mathbb{Z}$ and $PL(\Sigma) = \mathbb{Z}^2$. The Fourier transforms of local heights can be computed as follows:

$$\hat{\mathsf{H}}_p(\chi_m, \mathbf{s}) = 1 + \sum_{n \geq 1} p^{-s_1 - im} + \sum_{n \geq 1} p^{-s_1 + im} = \frac{\zeta_p(s_1 + im)\zeta_p(s_2 - im)}{\zeta_p(s_1 + s_2)},$$

$$\hat{\mathsf{H}}_\infty(\chi_m, \mathbf{s}) = \int_0^\infty e^{(-s_1 - im)x} dx + \int_0^\infty e^{(-s_2 + im)x} dx = \frac{1}{s_1 + im} + \frac{1}{s_2 - im}.$$

We obtain

$$\mathsf{Z}(\mathbb{P}^1, s_1, s_2) = \int_{\mathbb{R}} \zeta(s_1 + im)\zeta(s_2 - im) \cdot \left(\frac{1}{s_1 + im} + \frac{1}{s_2 - im} \right) dm.$$

The integral converges for $\Re(s_1), \Re(s_2) > 1$, absolutely and uniformly on compact subsets. It remains to establish its meromorphic continuation. This can be achieved by shifting the contour of integration and computing the resulting residues.

It is helpful to compare this approach with the analysis of $\mathbb{P}^1$ as an additive variety in Example 6.5.1.

The Fourier transforms of local height functions $\hat{\mathsf{H}}_{\Sigma,v}(\chi_v, -\mathbf{s})$ in the case of $\mathbb{G}_m^d$ over $\mathbb{Q}$ are given by:

$$\sum_{k=1}^d \sum_{\sigma \in \Sigma(k)} (-1)^k \prod_{e_j \in \sigma} \frac{1}{1 - q_v^{-(s_j + i\langle e_j, m_v \rangle)}} \qquad v \nmid \infty,$$

$$\sum_{\sigma \in \Sigma(d)} \prod_{e_j \in \sigma} \frac{1}{(s_j + i\langle e_j, m_v \rangle)} \qquad v \mid \infty.$$

where $(m_v)_v$ are the local components of the character $\chi = \chi_m$. The general case of nonsplit tori over number fields requires more care. We have an exact sequence of Γ-modules:

$$0 \to M \to PL(\Sigma) \to \mathrm{Pic}(X_\Sigma) \to 0,$$

with $PL(\Sigma)$ a permutation module. Duality gives a sequence of groups:

$$0 \to T_{\mathrm{Pic}}(\mathbb{A}_F) \to T_{PL}(\mathbb{A}_F) \to T(\mathbb{A}_F),$$

with

$$T_{PL}(\mathbb{A}_F) = \prod_{j=1}^{k} R_{F_j/F}\mathbb{G}_m(\mathbb{A}_F) \quad \text{(restriction of scalars)}.$$

We get a map

$$\begin{array}{ccc} (T(\mathbb{A}_F)/\mathbf{K}_T \cdot T(F))^* & \to & \prod_{j=1}^{k} \left(\mathbb{G}_m(\mathbb{A}_{F_j})/\mathbb{G}_m(F_j)\right)^* \\ \chi & \mapsto & (\chi_1, \dots, \chi_k) \end{array} \tag{6.23}$$

This map has finite kernel, denoted by $\mathrm{Ker}(T)$. Assembling local computations, we have

$$\hat{\mathsf{H}}_\Sigma(\chi, \mathbf{s}) = \frac{\prod_{j=1}^{k} \mathsf{L}(s_j, \chi_j)}{Q_\Sigma(\chi, \mathbf{s})} \zeta_{\Sigma,\infty}(\mathbf{s}, \chi), \tag{6.24}$$

where $Q_\Sigma(\chi, \mathbf{s})$ bounded uniformly in χ, in compact subsets in $\Re(s_j) > 1/2 + \delta$, $\delta > 0$, and

$$|\zeta_{\Sigma,\infty}(\mathbf{s}, \chi)| \ll \frac{1}{(1 + \|m\|_\infty)^{d+1}} \cdot \frac{1}{(1 + \|\chi\|_\infty)^{d'+1}}.$$

This implies that

$$\mathsf{Z}_\Sigma(\mathbf{s}) = \int_{M^\Gamma_{\mathbb{R}}} f_\Sigma(\mathbf{s} + im)dm, \tag{6.25}$$

where

$$f_\Sigma(\mathbf{s}) := \sum_{\chi \in (T^1(\mathbb{A}_F)/\mathbf{K}_T \cdot T(F))^*} \hat{\mathsf{H}}_\Sigma(\chi, \mathbf{s})$$

We have

(1) $(s_1 - 1) \dots (s_k - 1) f_\Sigma(\mathbf{s})$ is holomorphic for $\Re(s_j) > 1 - \delta$;
(2) f_Σ satisfies growth conditions in vertical strips (this follows by applying the Phragmen-Lindelöf principle 6.1.2 to bound L-functions appearing in equation (6.24));
(3) $\lim_{s_j \to 1} f_\Sigma(\mathbf{s}) = c(f_\Sigma) \neq 0$.

The integral (6.25) resembles the integral representation (6.5) for $\mathcal{X}_{\Lambda_{\mathrm{eff}}}(\mathbf{s} - 1)$ (defined in Equation 4.12). A technical theorem allows one to compute this integral via iterated residues, in the neighborhood of $\Re(\mathbf{s}) = (1, \dots, 1)$. The Convexity Principle 6.1.1 implies a meromorphic continuation of $\mathsf{Z}_\Sigma(\mathbf{s})$ to a tubular neighborhood of the shifted cone $\Lambda_{\mathrm{eff}}(X_\Sigma)$. The restriction $\mathsf{Z}(s(-K_\Sigma))$ of the height zeta function to the line through the anticanonical class has a pole at $s = 1$ of order $\mathrm{rk}\,\mathrm{Pic}(X_\Sigma)^\Gamma$ with leading coefficient $\alpha(X_\Sigma) \cdot c(f_\Sigma(0))$ (see Definition 4.12.2). The

identification of the factors β and τ in Equation (4.14) requires an application of the Poisson formula to the kernel $\mathrm{Ker}(T)$ from (6.23). One has

$$\bigcap_{\chi\in \mathrm{Ker}(T)} \mathrm{Ker}(\chi) = \overline{T(F)} \subset T(\mathbb{A}_F),$$

the closure of $T(F)$ in the direct product topology. Converting the integral and matching the measures yields

$$c(f_\Sigma(0)) = \sum_{\chi\in \mathrm{Ker}(T)} \hat{\mathsf{H}}(\chi, \mathbf{s}) = \int_{\overline{T(F)}} \mathsf{H}(g,\mathbf{s})^{-1} dg = \beta(X)\cdot \int_{X_\Sigma(F)} \omega_{\mathcal{K}_\Sigma},$$

proving Conjecture 4.12.4. Other line bundles require a version of the technical theorem above, and yet another application of Poisson formula, leading to $\mathcal{L}$-primitive fibrations discussed in Section 4.13.

6.7. Unipotent groups. Let $X \supset G$ be an equivariant compactification of a unipotent group over a number field F and

$$X \setminus G = D = \bigcup_{i\in\mathcal{I}} D_i.$$

Throughout, we will assume that G acts on X on both sides, i.e., that X is a compactification of $G\times G/G$, or a bi-equivariant compactification. We also assume that D is a divisor with normal crossings and its components D_i are geometrically irreducible. The main geometric invariants of X have been computed in Example 1.1.4: The Picard group is freely generated by the classes of D_i, the effective cone is simplicial, and the anticanonical class is the sum of boundary components with nonnegative coefficients.

Local and global heights have been defined in Example 4.8.6:

$$\mathsf{H}_{D_i,v}(x) := \|\mathsf{f}_i(x)\|_v^{-1} \quad \text{and} \quad \mathsf{H}_{D_i}(x) = \prod_v \mathsf{H}_{D_i,v}(x),$$

where f_i is the unique G-invariant section of $\mathrm{H}^0(X, D_i)$. We get a *height pairing*:

$$\mathsf{H} \ : G(\mathbb{A}_F)\times \mathrm{Pic}(X)_{\mathbb{C}} \to \mathbb{C}$$

as in Section 6.3. The bi-equivariance of X implies that H is invariant under the action *on both sides* of a compact open subgroup $\mathbf{K}$ of the finite adeles. Moreover, we can arrange that H_v is smooth in g_v for archimedean v.

The height zeta function

$$\mathsf{Z}(\mathbf{s}, g) := \sum_{\gamma\in G(F)} \mathsf{H}(\mathbf{s}, g)^{-1}$$

is holomorphic in $\mathbf{s}$, for $\Re(\mathbf{s}) \gg 0$. As a function of g it is continuous and in $\mathsf{L}^2(G(F)\backslash G(\mathbb{A}_F))$, for these $\mathbf{s}$. We proceed to analyze its spectral decomposition. We get a formal identity

$$\mathsf{Z}(\mathbf{s}; g) = \sum_{\varrho} \mathsf{Z}_\varrho(\mathbf{s}; g), \tag{6.26}$$

where the sum is over all irreducible unitary representations $(\varrho, \mathcal{H}_\varrho)$ of $G(\mathbb{A}_F)$ occurring in the right regular representation of $G(\mathbb{A}_F)$ in $\mathsf{L}^2(G(F)\backslash G(\mathbb{A}_F))$. They are parametrized by F-rational orbits $\mathcal{O} = \mathcal{O}_\varrho$ under the coadjoint action of G on the

dual of its Lie algebra $\mathfrak{g}^*$. The relevant orbits are *integral*—there exists a lattice in $\mathfrak{g}^*(F)$ such that $\mathsf{Z}_\varrho(\mathbf{s}; g) = 0$ unless the intersection of $\mathcal{O}$ with this lattice is nonempty. The pole of highest order is contributed by the trivial representation and integrality ensures that this representation is "isolated".

Let ϱ be an integral representation as above. It has the following explicit realization: There exists an F-rational subgroup $M \subset G$ such that

$$\varrho = \mathrm{Ind}_M^G(\psi),$$

where ψ is a certain character of $M(\mathbb{A}_F)$. In particular, for the trivial representation, $M = G$ and ψ is the trivial character. Further, there exists a finite set of valuations $S = S_\varrho$ such that $\dim(\varrho_v) = 1$ for $v \notin S$ and consequently

$$\mathsf{Z}_\varrho(\mathbf{s}; g') = \mathsf{Z}^S(\mathbf{s}; g') \cdot \mathsf{Z}_S(\mathbf{s}; g'). \tag{6.27}$$

It turns out that

$$\mathsf{Z}^S(\mathbf{s}; g') := \prod_{v \notin S} \int_{M(F_v)} \mathsf{H}_v(\mathbf{s}; g_v g'_v)^{-1} \psi(g_v)\, dg_v,$$

with an appropriately normalized Haar measure dg_v on $M(F_v)$. The function Z_S is the projection of Z to $\bigotimes_{v \in S} \varrho_v$.

The first key result is the explicit computation of *height integrals*

$$\int_{M(F_v)} \mathsf{H}_v(\mathbf{s}; g_v g'_v)^{-1} \psi(g_v)\, dg_v$$

for almost all v. This has been done in [**CLT02**] for equivariant compactifications of additive groups $\mathbb{G}_a^n$ (see Section 6.5); the same approach works here too. The contribution from the trivial representation can be computed using the formula of Denef-Loeser, as in (6.12):

$$\int_{G(\mathbb{A}_F)} \mathsf{H}(\mathbf{s}; g)^{-1} dg = \prod_i \zeta_F(s_i - \kappa_i + 1) \cdot \Phi(\mathbf{s}),$$

where $\Phi(\mathbf{s})$ is holomorphic in

$$\mathsf{T} := \{\mathbf{s} \mid \Re(s_i) > \kappa_i - \epsilon \quad \forall i\}.$$

(Recall that $-K_X = \sum_i \kappa_i D_i$.) As in the case of additive groups in Section 6.5, this term gives the "correct" pole at $-K_X$. The analysis of 1-dimensional representations, with $M = G$, is similar to the additive case. New difficulties arise from infinite-dimensional ϱ on the right side of the expansion (6.26).

Next we need to estimate $\dim(\varrho_v)$ and the local integrals for nonarchimedean $v \in S_\varrho$. The key result here is that the contribution to the Euler product from these places is a holomorphic function which can be bounded from above by a *polynomial* in the coordinates of ϱ, for $\mathbf{s} \in \mathsf{T}$. The uniform convergence of the spectral expansion comes from estimates at the archimedean places: for every (left or right) G-invariant differential operator ∂ (and $\mathbf{s} \in \mathsf{T}$) there exists a constant $\mathsf{c}(\partial)$ such that

$$\int_{G(F_v)} |\partial \mathsf{H}_v(\mathbf{s}; g_v)^{-1} dg_v|_v \le \mathsf{c}(\partial). \tag{6.28}$$

Let v be real. It is known that ϱ_v can be modeled in $\mathsf{L}^2(\mathbb{R}^r)$, where $2r = \dim(\mathcal{O})$. More precisely, there exists an isometry

$$j \ : \ (\pi_v, \mathsf{L}^2(\mathbb{R}^r)) \to (\varrho_v, \mathcal{H}_v)$$

(an analog of the Θ-distribution). Moreover, the universal enveloping algebra $\mathfrak{U}(\mathfrak{g})$ surjects onto the Weyl algebra of differential operators with polynomial coefficients acting on the smooth vectors $\mathsf{C}^\infty(\mathbb{R}^r) \subset \mathsf{L}^2(\mathbb{R}^r)$. In particular, we can find an operator Δ acting as the (r-dimensional) harmonic oscillator

$$\prod_{j=1}^{r} \left(\frac{\partial^2}{\partial x_j^2} - a_j x_j^2 \right),$$

with $a_j > 0$. We choose an orthonormal basis of $\mathsf{L}^2(\mathbb{R}^r)$ consisting of Δ-eigenfunctions $\{\tilde{\omega}_\lambda\}$ (which are well-known) and analyze

$$\int_{G(F_v)} \mathsf{H}_v(\mathbf{s}; g_v)^{-1} \overline{\omega}_\lambda(g_v) dg_v,$$

where $\omega_\lambda = j^{-1}(\tilde{\omega}_\lambda)$. Using integration by parts we find that for $\mathbf{s} \in \mathsf{T}$ and any $N \in \mathbb{N}$ there is a constant $\mathsf{c}(N, \Delta)$ such that this integral is bounded by

$$(1 + |\lambda|)^{-N} \mathsf{c}(N, \Delta). \tag{6.29}$$

This estimate suffices to conclude that for *each* ϱ the function Z_{S_ϱ} is holomorphic in T.

Now the issue is to prove the convergence of the sum in (6.26). Using any element $\partial \in \mathfrak{U}(\mathfrak{g})$ acting in $\mathcal{H}_\varrho$ by a scalar $\lambda(\partial) \neq 0$ (for example, any element in the center of $\mathfrak{U}(\mathfrak{g})$) we can improve the bound (6.29) to

$$(1 + |\lambda|)^{-N_1} \lambda(\partial)^{-N_2} \mathsf{c}(N_1, N_2, \Delta, \partial)$$

(for any $N_1, N_2 \in \mathbb{N}$). However, we have to ensure the uniformity of such estimates over the set of all ϱ. This relies on a parametrization of coadjoint orbits. There is a finite set $\{\Sigma_{\mathbf{d}}\}$ of "packets" of coadjoint orbits, each parametrized by a locally closed subvariety $Z_{\mathbf{d}} \subset \mathfrak{g}^*$, and for each $\mathbf{d}$ a finite set of F-rational polynomials $\{P_{\mathbf{d},r}\}$ on $\mathfrak{g}^*$ such that the restriction of each $P_{\mathbf{d},r}$ to $Z_{\mathbf{d}}$ is invariant under the coadjoint action. Consequently, the corresponding derivatives

$$\partial_{\mathbf{d},r} \in \mathfrak{U}(\mathfrak{g})$$

act in $\mathcal{H}_\varrho$ by multiplication by the scalar

$$\lambda_{\varrho,r} = P_{\mathbf{d},r}(\mathcal{O}).$$

There is a similar uniform construction of the "harmonic oscillator" $\Delta_{\mathbf{d}}$ for each $\mathbf{d}$. Combining the resulting estimates we obtain the uniform convergence of the right hand side in (6.26).

The last technical point is to prove that both expressions (6.8) and (6.26) for $\mathsf{Z}(\mathbf{s}; g)$ define *continuous* functions on $G(F)\backslash G(\mathbb{A}_F)$. Then (6.9) gives the desired meromorphic continuation of $\mathsf{Z}(\mathbf{s}; e)$. See [**ST04**] for details in the case when G is the Heisenberg group.

Background material on representation theory of unipotent groups can be found in the books [**CG66**], [**Dix96**] and the papers [**Moo65**], [**Kir99**].

6.8. Homogeneous spaces. Recall the setup of Section 6.2: G is an algebraic group acting on PGL_{n+1}, X° the G-orbit through a point $x_0 \in \mathbb{P}^n(F)$ and X the Zariski closure of X° in $\mathbb{P}^n$. Let H be the stabilizer of x_0 so that $X^\circ = H\backslash G$. Thus $X \subset \mathbb{P}^n$ is an equivariant compactification of the homogeneous space X°. Let $L = \mathcal{O}(1)$ be the line bundle on X arising as a hyperplane section in this embedding.

THEOREM 6.8.1. [**GO08**] *Assume that*

- *G is a connected simply connected semisimple F-group;*
- *H is a semisimple maximal connected F-subgroup of G;*
- *for all but finitely many places v, $G(F_v)$ acts transitively on $X^\circ(F_v)$.*

Then there exists a constant $\mathsf{c} > 0$ such that

$$\mathsf{N}(\mathsf{B}) = c \cdot \mathsf{B}^a \log(\mathsf{B})^{b-1}(1 + o(1)), \quad \mathsf{B} \to \infty.$$

Here $a = a(L)$ and $b = b(L)$ are the constants defined in Section 1.4.

The main effort goes into establishing the asymptotic comparison

$$\mathrm{vol}\{x \in X^\circ(\mathbb{A}_F) \mid \mathsf{H}(x) \le \mathsf{B}\} \asymp \#\{x \in X^\circ(F) \mid \mathsf{H}(x) \le \mathsf{B}\},$$

using techniques from ergodic theory. The identification of the constants a, b, c from the adelic volume follows by applying a Tauberian theory to the height integral as in Step 3 of Section 6.3.

The group case, i.e.,

$$X^\circ = G \times G/G,$$

has been treated in [**STBT07**] using spectral methods and in [**GMO06**] using adelic mixing. See also [**Oh08**] for a comprehensive survey of applications of ergodic theory to counting problems.

6.9. Fiber bundles. Let T be an algebraic torus with a left action $X \times T \to X$ on a smooth projective variety X. Let $\mathrm{Pic}^T(X)$ be the group of isomorphism classes of T-linearized line bundles on X. Let $\mathcal{T} \to W$ be a T-torsor over a smooth projective base W. One can form a twisted product

$$Y := X \times^T \mathcal{T},$$

as the quotient of $X \times \mathcal{T}$ by the induced action $(x, \theta)t \mapsto (xt, t^{-1}\theta)$. This is a locally trivial fibration over W with fibers isomorphic to X. Interesting examples of such varieties arise when $\mathcal{T} \to W$ is a universal torsor and X is an equivariant compactification of T. In this case, Y is a compactification of $\mathcal{T}$ and, following the approach in Section 5, we can expect to see connections between arithmetic and geometric properties of Y and W.

Here is a version of this construction combining varieties treated in Sections 6.4 and 6.6: let G be a semi-simple algebraic group, $P \subset G$ a parabolic subgroup and $\eta : P \to T$ a homomorphism to an algebraic torus. Let X be an equivariant compactification of T. Consider the twisted product

$$Y := X \times^P G,$$

i.e., the quotient of $X \times G$ by the P-action

$$(x, g)p \mapsto (x\eta(p), p^{-1}g).$$

This is a locally trivial fiber bundle over $W := P\backslash G$ with fibers X. When P is the Borel subgroup of G, $T = P/U$ the maximal torus, X a smooth equivariant compactification of T, and $\eta : P \to T$ the canonical projection, one obtains equivariant compactifications of G/U, the so-called *horospherical varieties*.

The geometric properties of Y can be read off from the invariants of X, W and η (see [**ST99**], [**CLT01**]):

- there is an exact sequence

$$0 \to \mathfrak{X}^*(T) \xrightarrow{\eta^*} \mathrm{Pic}^T(X) \oplus \mathfrak{X}^*(P) \to \mathrm{Pic}(Y) \to 0;$$

- $\Lambda_{\mathrm{eff}}(Y)$ is the image of $\Lambda^T_{\mathrm{eff}}(X) \oplus \Lambda_{\mathrm{eff}}(W)$ under the natural projection;
- the anticanonical class K_Y is the image of $(-K_X, -K_W)$.

Recall that $\mathfrak{X}^*(P)$ is a finite-index subgroup of $\mathrm{Pic}(W)$. Let $\pi : Y \to W$ denote the projection and L_W, resp. L, a line bundle on W, resp. a T-linearized line bundle on X. For $L \in \mathrm{Pic}^T(X)$ let L^Y be its image in $\mathrm{Pic}(Y)$. There is an exact sequence

$$0 \to \mathrm{Pic}(W) \xrightarrow{\pi^*} \mathrm{Pic}(T) \to \mathrm{Pic}(X) \to 0.$$

Let $(x, g) \in \pi^{-1}(y) \in X \times G$, with $g = pk,\ p \in P(\mathbb{A}_F), k \in \mathbf{K}_G$. One can define an adelic metrization of L^Y (see [**ST99**, Section 4.3]) such that

$$\mathsf{H}_{\mathcal{L}^Y}(y) = \mathsf{H}_{\mathcal{L}(\eta(p))}(x), \tag{6.30}$$

where $\mathcal{L}(\eta(p))$ is the *twisted* adelic metrization of the T-linearized line bundle L on X defined in Section 6.6. Consider the height zeta function

$$\mathsf{Z}(Y, \mathcal{L}^Y \otimes \mathcal{L}_W, s) = \sum_{\gamma \in P(F)\backslash G(F)} \mathsf{H}_{\mathcal{L}_W}(\gamma)^{-s} \sum_{x \in \pi^{-1}(\gamma)} \mathsf{H}_{\mathcal{L}_Y}(x)^{-s}.$$

The key property (6.30) implies that

$$\sum_{x \in \pi^{-1}(\gamma)} \mathsf{H}_{\mathcal{L}_Y}(x)^{-s} = \mathsf{Z}(X, \mathcal{L}(\eta(p_\gamma)), s).$$

Combining this with the Poisson formula 6.22, one obtains, at least formally

$$\mathsf{Z}(Y^\circ, \mathcal{L}^Y \otimes \mathcal{L}_W, s) = \int_{(T(\mathbb{A}_F)/\mathbf{K}_T \cdot T(F))^*} \hat{\mathsf{H}}_\Sigma(\chi, \mathbf{s}) \cdot \mathsf{E}^G_P(\mathbf{s}', \chi \circ \eta)\, d\chi, \tag{6.31}$$

where $(\mathbf{s}, \mathbf{s}') \in \mathrm{Pic}^T(X)_{\mathbb{C}} \oplus \mathfrak{X}^*(P)_{\mathbb{C}}$, $Y^\circ = T \times^P G$, and $\mathsf{E}^G_P(\mathbf{s}', \chi \circ \eta)$ is the Eisenstein series (6.14). Analytic properties of the integral on the right side of (6.31) are established following the approach in Section 6.6. Uniform bounds of the shape

$$|\mathsf{E}^G_P(\mathbf{s}', \chi \circ \eta)| \ll (1 + \|\Im(\mathbf{s}')\| + \|\chi\|)^\epsilon,$$

for $\Re(\mathbf{s}')$ close to 2ρ, follow from Theorem 6.1.2, combined with Proposition 6.1.3.

THEOREM 6.9.1. [**ST99**] *Conjecture 4.14.2 holds for $Y = X \times^P G$.*

Similar constructions can be carried out with parabolic subgroups $P_1, \dots, P_r$ in groups $G_1, \dots, G_r$, and homomorphisms $\eta_r : P_i \to G_{i-1}$, leading to Bott-Samelson varieties

$$Y = P_1 \backslash G_1 \times^{P_2} G_2 \times \dots \times^{P_r} \backslash G_r,$$

which arise as desingularizations of Schubert varieties. Results concerning analytic properties of corresponding height zeta functions can be found in [**Str98**] and [**Str01**].

References

[Abr] D. Abramovich, *Birational geometry for number theorists*, in this volume.

[AC08] E. Amerik and F. Campana, *Fibrations méromorphes sur certaines variétés à fibré canonique trivial*, Pure Appl. Math. Q. **4** (2008), no. 2, part 1, 509–545. MR 2400885

[Ara05] C. Araujo, *The cone of effective divisors of log varieties after Batyrev*, 2005, `arXiv:math/0502174`.

[AV08] E. Amerik and C. Voisin, *Potential density of rational points on the variety of lines of a cubic fourfold*, Duke Math. J. **145** (2008), no. 2, 379–408. MR 2449951

[Bak89] R. C. Baker, *Diagonal cubic equations. II*, Acta Arith. **53** (1989), no. 3, 217–250. MR 1032826 (91b:11100a)

[Bat92] V. V. Batyrev, *The cone of effective divisors of threefolds*, Proceedings of the International Conference on Algebra, Part 3 (Novosibirsk, 1989) (Providence, RI), Contemp. Math., vol. 131, Amer. Math. Soc., 1992, pp. 337–352. MR 94f:14035

[BBFL07] M. J. Bright, N. Bruin, E. V. Flynn, and A. Logan, *The Brauer-Manin obstruction and Sh[2]*, LMS J. Comput. Math. **10** (2007), 354–377. MR 2342713

[BCHM06] C. Birkar, P. Cascini, C. D. Hacon, and J. McKernan, *Existence of minimal models for varieties of log general type*, 2006, `arXiv.org:math/0610203`.

[BD85] A. Beauville and R. Donagi, *La variété des droites d'une hypersurface cubique de dimension* 4, C. R. Acad. Sci. Paris Sér. I Math. **301** (1985), no. 14, 703–706. MR 818549 (87c:14047)

[BG06] E. Bombieri and W. Gubler, *Heights in Diophantine geometry*, New Mathematical Monographs, vol. 4, Cambridge University Press, Cambridge, 2006. MR 2216774 (2007a:11092)

[BHBS06] T. D. Browning, D. R. Heath-Brown, and P. Salberger, *Counting rational points on algebraic varieties*, Duke Math. J. **132** (2006), no. 3, 545–578. MR 2219267 (2007c:14018)

[Bir62] B. J. Birch, *Forms in many variables*, Proc. Roy. Soc. Ser. A **265** (1961/1962), 245–263. MR 0150129 (27 #132)

[BK85] F. A. Bogomolov and P. I. Katsylo, *Rationality of some quotient varieties*, Mat. Sb. (N.S.) **126(168)** (1985), no. 4, 584–589. MR 788089 (86k:14033)

[BM90] V. V. Batyrev and Y. I. Manin, *Sur le nombre des points rationnels de hauteur borné des variétés algébriques*, Math. Ann. **286** (1990), no. 1-3, 27–43. MR 1032922 (91g:11069)

[Bog87] F. A. Bogomolov, *The Brauer group of quotient spaces of linear representations*, Izv. Akad. Nauk SSSR Ser. Mat. **51** (1987), no. 3, 485–516, 688. MR 903621 (88m:16006)

[BP89] E. Bombieri and J. Pila, *The number of integral points on arcs and ovals*, Duke Math. J. **59** (1989), no. 2, 337–357. MR 1016893 (90j:11099)

[BP04] V. V. Batyrev and O. N. Popov, *The Cox ring of a Del Pezzo surface*, Arithmetic of higher-dimensional algebraic varieties (Palo Alto, CA, 2002), Progr. Math., vol. 226, Birkhäuser Boston, Boston, MA, 2004, pp. 85–103. MR 2029863 (2005h:14091)

[Bri07] M. Brion, *The total coordinate ring of a wonderful variety*, J. Algebra **313** (2007), no. 1, 61–99. MR 2326138 (2008d:14067)

[Brü94] J. Brüdern, *Small solutions of additive cubic equations*, J. London Math. Soc. (2) **50** (1994), no. 1, 25–42. MR 1277752 (95e:11111)

[BS66] Z. I. Borevich and I. R. Shafarevich, *Number theory*, Translated from the Russian by Newcomb Greenleaf. Pure and Applied Mathematics, Vol. 20, Academic Press, New York, 1966. MR 0195803 (33 #4001)

[BSD04] M. Bright and P. Swinnerton-Dyer, *Computing the Brauer-Manin obstructions*, Math. Proc. Cambridge Philos. Soc. **137** (2004), no. 1, 1–16. MR 2075039 (2005c:11081)

[BT95] V. V. Batyrev and Y. Tschinkel, *Rational points of bounded height on compactifications of anisotropic tori*, Internat. Math. Res. Notices (1995), no. 12, 591–635. MR 1369408 (97a:14021)

[BT96a] ———, *Height zeta functions of toric varieties*, J. Math. Sci. **82** (1996), no. 1, 3220–3239, Algebraic geometry, 5. MR 1423638 (98b:11067)

[BT96b] ———, *Rational points on some Fano cubic bundles*, C. R. Acad. Sci. Paris Sér. I Math. **323** (1996), no. 1, 41–46. MR 1401626 (97j:14023)

[BT98] ———, *Manin's conjecture for toric varieties*, J. Algebraic Geom. **7** (1998), no. 1, 15–53. MR 1620682 (2000c:11107)

[BT99] F. A. Bogomolov and Y. Tschinkel, *On the density of rational points on elliptic fibrations*, J. Reine Angew. Math. **511** (1999), 87–93. MR 2000e:14025

[BT00] ———, *Density of rational points on elliptic $K3$ surfaces*, Asian J. Math. **4** (2000), no. 2, 351–368. MR 2002b:14025

[BW79] J. W. Bruce and C. T. C. Wall, *On the classification of cubic surfaces*, J. London Math. Soc. (2) **19** (1979), no. 2, 245–256. MR 80f:14021

[Cam04] F. Campana, *Orbifolds, special varieties and classification theory*, Ann. Inst. Fourier (Grenoble) **54** (2004), no. 3, 499–630. MR 2097416 (2006c:14013)

[Can01] S. Cantat, *Dynamique des automorphismes des surfaces $K3$*, Acta Math. **187** (2001), no. 1, 1–57. MR 1864630 (2003h:32026)

[Cas55] J. W. S. Cassels, *Bounds for the least solutions of homogeneous quadratic equations*, Proc. Cambridge Philos. Soc. **51** (1955), 262–264. MR 0069217 (16,1002c)

[Cas07] A.-M. Castravet, *The Cox ring of $\bar{M}_{0,6}$*, 2007, to appear in *Trans. AMS*, `arXiv:07050070`.

[CG66] J. W. S. Cassels and M. J. T. Guy, *On the Hasse principle for cubic surfaces*, Mathematika **13** (1966), 111–120. MR 0211966 (35 #2841)

[CG72] C. H. Clemens and P. A. Griffiths, *The intermediate Jacobian of the cubic threefold*, Ann. of Math. (2) **95** (1972), 281–356. MR 0302652 (46 #1796)

[Cha94] G. J. Chaitin, *Randomness and complexity in pure mathematics*, Internat. J. Bifur. Chaos Appl. Sci. Engrg. **4** (1994), no. 1, 3–15. MR 1276801 (95b:00005)

[Che35] C. Chevalley, *Démonstration d'une hypothèse de M. Artin.*, Abh. Math. Semin. Hamb. Univ. **11** (1935), 73–75.

[Che05] I. A. Cheltsov, *Birationally rigid Fano varieties*, Uspekhi Mat. Nauk **60** (2005), no. 5(365), 71–160. MR 2195677 (2007d:14028)

[CKM88] H. Clemens, J. Kollár, and S. Mori, *Higher-dimensional complex geometry*, Astérisque (1988), no. 166, 144 pp. (1989). MR 90j:14046

[CLT00a] A. Chambert-Loir and Y. Tschinkel, *Points of bounded height on equivariant compactifications of vector groups. I*, Compositio Math. **124** (2000), no. 1, 65–93. MR 1797654 (2001j:11053)

[CLT00b] ———, *Points of bounded height on equivariant compactifications of vector groups. II*, J. Number Theory **85** (2000), no. 2, 172–188. MR 1802710 (2001k:11118)

[CLT01] ———, *Torseurs arithmétiques et espaces fibrés*, Rational points on algebraic varieties, Progr. Math., vol. 199, Birkhäuser, Basel, 2001, pp. 37–70. MR 1875170 (2002m:11054)

[CLT02] ———, *On the distribution of points of bounded height on equivariant compactifications of vector groups*, Invent. Math. **148** (2002), no. 2, 421–452. MR 1906155 (2003d:11094)

[Cor96] A. Corti, *Del Pezzo surfaces over Dedekind schemes*, Ann. of Math. (2) **144** (1996), no. 3, 641–683. MR 1426888 (98e:14037)

[Cor07] P. Corn, *The Brauer-Manin obstruction on del Pezzo surfaces of degree 2*, Proc. Lond. Math. Soc. (3) **95** (2007), no. 3, 735–777. MR 2368282

[Cox95] D. A. Cox, *The homogeneous coordinate ring of a toric variety*, J. Algebraic Geom. **4** (1995), no. 1, 17–50. MR 95i:14046

[CP07] I. A. Cheltsov and J. Park, *Two remarks on sextic double solids*, J. Number Theory **122** (2007), no. 1, 1–12. MR 2287107 (2008a:14031)

[CS06] I. Coskun and J. Starr, *Divisors on the space of maps to Grassmannians*, Int. Math. Res. Not. (2006), Art. ID 35273, 25. MR 2264713 (2007g:14067)

[CTKS87] J.-L. Colliot-Thélène, D. Kanevsky, and J.-J. Sansuc, *Arithmétique des surfaces cubiques diagonales*, Diophantine approximation and transcendence theory (Bonn, 1985), Lecture Notes in Math., vol. 1290, Springer, Berlin, 1987, pp. 1–108. MR 927558 (89g:11051)

[CTS87] J.-L. Colliot-Thélène and J.-J. Sansuc, *La descente sur les variétés rationnelles. II*, Duke Math. J. **54** (1987), no. 2, 375–492. MR 899402 (89f:11082)

[CTS89] J.-L. Colliot-Thélène and P. Salberger, *Arithmetic on some singular cubic hypersurfaces*, Proc. London Math. Soc. (3) **58** (1989), no. 3, 519–549. MR 988101 (90e:11091)

[CTSSD87a] J.-L. Colliot-Thélène, J.-J. Sansuc, and P. Swinnerton-Dyer, *Intersections of two quadrics and Châtelet surfaces. I*, J. Reine Angew. Math. **373** (1987), 37–107. MR 870307 (88m:11045a)

[CTSSD87b] J.-L. Colliot-Thélène, J.-J. Sansuc, and Peter Swinnerton-Dyer, *Intersections of two quadrics and Châtelet surfaces. II*, J. Reine Angew. Math. **374** (1987), 72–168. MR 876222 (88m:11045b)

[CTSSD97] J.-L. Colliot-Thélène, A. N. Skorobogatov, and P. Swinnerton-Dyer, *Double fibres and double covers: paucity of rational points*, Acta Arith. **79** (1997), no. 2, 113–135. MR 98a:11081

[Der06] U. Derenthal, *Geometry of universal torsors*, 2006, Ph.D. thesis, University of Göttingen.

[Der07a] ———, *On a constant arising in Manin's conjecture for del Pezzo surfaces*, Math. Res. Lett. **14** (2007), no. 3, 481–489. MR 2318651

[Der07b] ———, *Universal torsors of del Pezzo surfaces and homogeneous spaces*, Adv. Math. **213** (2007), no. 2, 849–864. MR 2332612

[Die03] R. Dietmann, *Small solutions of quadratic Diophantine equations*, Proc. London Math. Soc. (3) **86** (2003), no. 3, 545–582. MR 1974390 (2003m:11055)

[Dix96] J. Dixmier, *Algèbres enveloppantes*, Les Grands Classiques Gauthier-Villars. [Gauthier-Villars Great Classics], Éditions Jacques Gabay, Paris, 1996, Reprint of the 1974 original. MR 1451138 (98a:17012)

[DJT08] U. Derenthal, M. Joyce, and Z. Teitler, *The nef cone volume of generalized del Pezzo surfaces*, Algebra Number Theory **2** (2008), no. 2, 157–182. MR 2377367

[DL98] J. Denef and F. Loeser, *Motivic Igusa zeta functions*, J. Algebraic Geom. **7** (1998), no. 3, 505–537. MR 1618144 (99j:14021)

[DL99] ———, *Germs of arcs on singular algebraic varieties and motivic integration*, Invent. Math. **135** (1999), no. 1, 201–232. MR 1664700 (99k:14002)

[DL01] ———, *Definable sets, motives and p-adic integrals*, J. Amer. Math. Soc. **14** (2001), no. 2, 429–469 (electronic). MR 1815218 (2002k:14033)

[dlB01] R. de la Bretèche, *Compter des points d'une variété torique*, J. Number Theory **87** (2001), no. 2, 315–331. MR 1824152 (2002a:11067)

[dlB07] ———, *Répartition des points rationnels sur la cubique de Segre*, Proc. Lond. Math. Soc. (3) **95** (2007), no. 1, 69–155. MR 2329549 (2008f:11041)

[Dol08] I. V. Dolgachev, *Reflection groups in algebraic geometry*, Bull. Amer. Math. Soc. (N.S.) **45** (2008), no. 1, 1–60 (electronic). MR 2358376

[DP80] M. Demazure and H. Pinkham (eds.), *Séminaire sur les Singularités des Surfaces*, Lecture Notes in Mathematics, vol. 777, Springer, Berlin, 1980, Held at the Centre de Mathématiques de l'École Polytechnique, Palaiseau, 1976–1977. MR 82d:14021

[DT07] U. Derenthal and Y. Tschinkel, *Universal torsors over del Pezzo surfaces and rational points*, Equidistribution in number theory, an introduction, NATO Sci. Ser. II Math. Phys. Chem., vol. 237, Springer, Dordrecht, 2007, pp. 169–196. MR 2290499 (2007j:14024)

[EJ06] A.-S. Elsenhans and J. Jahnel, *The Diophantine equation $x^4 + 2y^4 = z^4 + 4w^4$*, Math. Comp. **75** (2006), no. 254, 935–940 (electronic). MR 2197001 (2007e:11143)

[EJ07] ———, *On the smallest point on a diagonal quartic threefold*, J. Ramanujan Math. Soc. **22** (2007), no. 2, 189–204. MR 2333743 (2008d:14041)

[EJ08a] ———, *Experiments with general cubic surfaces*, 2008, to appear in *Manin's Festschrift*.

[EJ08b] ———, *On the smallest point on a diagonal cubic surface*, 2008, preprint.

[Eke90] T. Ekedahl, *An effective version of Hilbert's irreducibility theorem*, Séminaire de Théorie des Nombres, Paris 1988–1989, Progr. Math., vol. 91, Birkhäuser Boston, Boston, MA, 1990, pp. 241–249. MR 1104709 (92f:14018)

[Elk88] N. D. Elkies, *On $A^4 + B^4 + C^4 = D^4$*, Math. Comp. **51** (1988), no. 184, 825–835. MR 930224 (89h:11012)

[Ern94] R. Erné, *Construction of a del Pezzo surface with maximal Galois action on its Picard group*, J. Pure Appl. Algebra **97** (1994), no. 1, 15–27. MR 1310745 (96b:14045)

[Esn03] H. Esnault, *Varieties over a finite field with trivial Chow group of 0-cycles have a rational point*, Invent. Math. **151** (2003), no. 1, 187–191. MR 1943746 (2004e:14015)

[EV05] J. Ellenberg and A. Venkatesh, *On uniform bounds for rational points on nonrational curves*, Int. Math. Res. Not. (2005), no. 35, 2163–2181. MR 2181791 (2006g:11133)

[Fal83] G. Faltings, *Endlichkeitssätze für abelsche Varietäten über Zahlkörpern*, Invent. Math. **73** (1983), no. 3, 349–366, English translation in: Arithmetic Geometry, G. Cornell and J.H. Silverman eds. New York: Springer Verlag (1986). MR 718935 (85g:11026a)

[Fal91] G. Faltings, *Diophantine approximation on abelian varieties*, Ann. of Math. **133** (1991), 549–576. MR 1109353 (93d:11066)

[Far06] G. Farkas, *Syzygies of curves and the effective cone of $\overline{\mathcal{M}}_g$*, Duke Math. J. **135** (2006), no. 1, 53–98. MR 2259923 (2008a:14037)

[FG03] G. Farkas and A. Gibney, *The Mori cones of moduli spaces of pointed curves of small genus*, Trans. Amer. Math. Soc. **355** (2003), no. 3, 1183–1199 (electronic). MR 1938752 (2003m:14043)

[Fly04] E. V. Flynn, *The Hasse principle and the Brauer-Manin obstruction for curves*, Manuscripta Math. **115** (2004), no. 4, 437–466. MR 2103661 (2005j:11047)

[FMT89] J. Franke, Y. I. Manin, and Y. Tschinkel, *Rational points of bounded height on Fano varieties*, Invent. Math. **95** (1989), no. 2, 421–435. MR 974910 (89m:11060)

[Fou98] É. Fouvry, *Sur la hauteur des points d'une certaine surface cubique singulière*, Astérisque (1998), no. 251, 31–49, Nombre et répartition de points de hauteur bornée (Paris, 1996). MR 1679838 (2000b:11075)

[GKM02] A. Gibney, S. Keel, and I. Morrison, *Towards the ample cone of $\overline{M}_{g,n}$*, J. Amer. Math. Soc. **15** (2002), no. 2, 273–294 (electronic). MR 1887636 (2003c:14029)

[GM97] G. R. Grant and E. Manduchi, *Root numbers and algebraic points on elliptic surfaces with base $\mathbf{P}^1$*, Duke Math. J. **89** (1997), no. 3, 413–422. MR 99a:14028

[GMO06] A. Gorodnik, F. Maucourant, and H. Oh, *Manin's and Peyre's conjectures on rational points and adelic mixing*, 2006, `arXiv:math/0601127`, to appear in Ann. Sci. Ecole Norm. Sup.

[GO08] A. Gorodnik and H. Oh, *Rational points on homogeneous varieties and equidistribution of adelic points*, 2008, `arXiv:0803.1996`.

[God95] R. Godement, *Introduction à la théorie de Langlands*, Séminaire Bourbaki, Vol. 10, Soc. Math. France, Paris, 1995, pp. Exp. No. 321, 115–144. MR 1610464

[GP69] I. M. Gelfand and V. A. Ponomarev, *Remarks on the classification of a pair of commuting linear transformations in a finite-dimensional space*, Funkcional. Anal. i Priložen. **3** (1969), no. 4, 81–82. MR 0254068 (40 #7279)

[Guo95] C. R. Guo, *On solvability of ternary quadratic forms*, Proc. London Math. Soc. (3) **70** (1995), no. 2, 241–263. MR 1309229 (96d:11040)

[Har] D. Harari, *Non-abelian descent*, in this volume.

[Har77] R. Hartshorne, *Algebraic geometry*, Springer-Verlag, New York, 1977, Graduate Texts in Mathematics, No. 52. MR 0463157 (57 #3116)

[Har96] D. Harari, *Obstructions de Manin transcendantes*, Number theory (Paris, 1993–1994), London Math. Soc. Lecture Note Ser., vol. 235, Cambridge Univ. Press, Cambridge, 1996, pp. 75–87. MR 1628794 (99e:14025)

[Har04] ———, *Weak approximation on algebraic varieties*, Arithmetic of higher-dimensional algebraic varieties (Palo Alto, CA, 2002), Progr. Math., vol. 226, Birkhäuser Boston, Boston, MA, 2004, pp. 43–60. MR 2029861 (2004k:11101)

[Has] B. Hassett, *Rational surfaces over nonclosed fields*, in this volume.

[Has03] ———, *Potential density of rational points on algebraic varieties*, Higher dimensional varieties and rational points (Budapest, 2001), Bolyai Soc. Math. Stud., vol. 12, Springer, Berlin, 2003, pp. 223–282. MR 2011748 (2004j:14021)

[HB83] D. R. Heath-Brown, *Cubic forms in ten variables*, Proc. London Math. Soc. (3) **47** (1983), no. 2, 225–257. MR 703978 (85b:11025)

[HB97] ———, *The density of rational points on cubic surfaces*, Acta Arith. **79** (1997), no. 1, 17–30. MR 1438113 (98h:11083)

[HB02] ———, *The density of rational points on curves and surfaces*, Ann. of Math. (2) **155** (2002), no. 2, 553–595. MR 1906595 (2003d:11091)

[HB03] ———, *The density of rational points on Cayley's cubic surface*, Proceedings of the Session in Analytic Number Theory and Diophantine Equations (Bonn), Bonner Math. Schriften, vol. 360, Univ. Bonn, 2003, p. 33. MR 2075628 (2005d:14033)

[HB06] ———, *Counting rational points on algebraic varieties*, Analytic number theory, Lecture Notes in Math., vol. 1891, Springer, Berlin, 2006, pp. 51–95. MR 2277658 (2007h:14025)

[HBM99] D. R. Heath-Brown and B. Z. Moroz, *The density of rational points on the cubic surface* $X_0^3 = X_1X_2X_3$, Math. Proc. Cambridge Philos. Soc. **125** (1999), no. 3, 385–395. MR 1656797 (2000f:11080)

[HK00] Y. Hu and S. Keel, *Mori dream spaces and GIT*, Michigan Math. J. **48** (2000), 331–348, Dedicated to William Fulton on the occasion of his 60th birthday. MR 1786494 (2001i:14059)

[HMP98] J. Harris, B. Mazur, and R. Pandharipande, *Hypersurfaces of low degree*, Duke Math. J. **95** (1998), no. 1, 125–160. MR 1646558 (99j:14043)

[HS02] D. Harari and A. N. Skorobogatov, *Non-abelian cohomology and rational points*, Compositio Math. **130** (2002), no. 3, 241–273. MR 1887115 (2003b:11056)

[HS05] D. Harari and A. Skorobogatov, *Non-abelian descent and the arithmetic of Enriques surfaces*, Int. Math. Res. Not. (2005), no. 52, 3203–3228. MR 2186792 (2006m:14031)

[HT99] B. Hassett and Y. Tschinkel, *Geometry of equivariant compactifications of* $\mathbf{G}_a^n$, Internat. Math. Res. Notices (1999), no. 22, 1211–1230. MR 1731473 (2000j:14073)

[HT00a] J. Harris and Y. Tschinkel, *Rational points on quartics*, Duke Math. J. **104** (2000), no. 3, 477–500. MR 2002h:14033

[HT00b] B. Hassett and Y. Tschinkel, *Abelian fibrations and rational points on symmetric products*, Internat. J. Math. **11** (2000), no. 9, 1163–1176. MR 1809306 (2002a:14010)

[HT00c] ———, *Abelian fibrations and rational points on symmetric products*, Internat. J. Math. **11** (2000), no. 9, 1163–1176. MR 1809306 (2002a:14010)

[HT01] ———, *Rational curves on holomorphic symplectic fourfolds*, Geom. Funct. Anal. **11** (2001), no. 6, 1201–1228. MR 1878319 (2002m:14033)

[HT02a] ———, *On the effective cone of the moduli space of pointed rational curves*, Topology and geometry: commemorating SISTAG, Contemp. Math., vol. 314, Amer. Math. Soc., Providence, RI, 2002, pp. 83–96. MR 1941624 (2004d:14028)

[HT02b] ———, *On the effective cone of the moduli space of pointed rational curves*, Topology and geometry: commemorating SISTAG, Contemp. Math., vol. 314, Amer. Math. Soc., Providence, RI, 2002, pp. 83–96. MR 1941624 (2004d:14028)

[HT03] ———, *Integral points and effective cones of moduli spaces of stable maps*, Duke Math. J. **120** (2003), no. 3, 577–599. MR 2030096 (2005g:14055a)

[HT04] ———, *Universal torsors and Cox rings*, Arithmetic of higher-dimensional algebraic varieties (Palo Alto, CA, 2002), Progr. Math., vol. 226, Birkhäuser Boston, Boston, MA, 2004, pp. 149–173. MR 2029868 (2005a:14049)

[HT08a] ———, *Flops on holomorphic symplectic fourfolds and determinantal cubic hypersurfaces*, 2008, arXiv:0805.4162.

[HT08b] ———, *Potential density of rational points for K3 surfaces over function fields*, Amer. J. Math. **130** (2008), no. 5, 1263–1278.

[Hun96] B. Hunt, *The geometry of some special arithmetic quotients*, Lecture Notes in Mathematics, vol. 1637, Springer-Verlag, Berlin, 1996. MR 1438547 (98c:14033)

[IM71] V. A. Iskovskih and Y. I. Manin, *Three-dimensional quartics and counterexamples to the Lüroth problem*, Mat. Sb. (N.S.) **86(128)** (1971), 140–166. MR 0291172 (45 #266)

[Ino78] H. Inose, *Defining equations of singular K3 surfaces and a notion of isogeny*, Proceedings of the International Symposium on Algebraic Geometry (Kyoto Univ., Kyoto, 1977) (Tokyo), Kinokuniya Book Store, 1978, pp. 495–502. MR 578868 (81h:14021)

[IP99a] V. A. Iskovskikh and Y. G. Prokhorov, *Fano varieties*, Algebraic geometry, V, Encyclopaedia Math. Sci., vol. 47, Springer, Berlin, 1999, pp. 1–247. MR 2000b:14051b

[IP99b] ———, *Fano varieties*, Algebraic geometry, V, Encyclopaedia Math. Sci., vol. 47, Springer, Berlin, 1999, pp. 1–247. MR 1668579 (2000b:14051b)

[Isk71] V. A. Iskovskih, *A counterexample to the Hasse principle for systems of two quadratic forms in five variables*, Mat. Zametki **10** (1971), 253–257. MR 0286743 (44 #3952)

[Isk79] ———, *Anticanonical models of three-dimensional algebraic varieties*, Current problems in mathematics, Vol. 12 (Russian), VINITI, Moscow, 1979, pp. 59–157, 239. MR 81i:14026b

[Isk01] V. A. Iskovskikh, *Birational rigidity of Fano hypersurfaces in the framework of Mori theory*, Uspekhi Mat. Nauk **56** (2001), no. 2(338), 3–86. MR 1859707 (2002g:14017)

[Kas08] A. Kasprzyk, *Bounds on fake weighted projective space*, 2008, arXiv:0805.1008.

[Kat82] P. I. Katsylo, *Sections of sheets in a reductive algebraic Lie algebra*, Izv. Akad. Nauk SSSR Ser. Mat. **46** (1982), no. 3, 477–486, 670. MR 661143 (84k:17005)

[Kat87] T. Katsura, *Generalized Kummer surfaces and their unirationality in characteristic p*, J. Fac. Sci. Univ. Tokyo Sect. IA Math. **34** (1987), no. 1, 1–41. MR 882121 (88c:14055)

[Kaw08] Sh. Kawaguchi, *Projective surface automorphisms of positive topological entropy from an arithmetic viewpoint*, Amer. J. Math. **130** (2008), no. 1, 159–186. MR 2382145

[Kir99] A. A. Kirillov, *Merits and demerits of the orbit method*, Bull. Amer. Math. Soc. (N.S.) **36** (1999), no. 4, 433–488. MR 1701415 (2000h:22001)

[Kle66] S. L. Kleiman, *Toward a numerical theory of ampleness*, Ann. of Math. (2) **84** (1966), 293–344. MR 0206009 (34 #5834)

[KM98] J. Kollár and S. Mori, *Birational geometry of algebraic varieties*, Cambridge Tracts in Mathematics, vol. 134, Cambridge University Press, Cambridge, 1998, With the collaboration of C. H. Clemens and A. Corti, Translated from the 1998 Japanese original. MR 1658959 (2000b:14018)

[KMM87] Y. Kawamata, K. Matsuda, and K. Matsuki, *Introduction to the minimal model problem*, Algebraic geometry, Sendai, 1985, Adv. Stud. Pure Math., vol. 10, North-Holland, Amsterdam, 1987, pp. 283–360. MR 89e:14015

[KMM92] J. Kollár, Y. Miyaoka, and S. Mori, *Rational connectedness and boundedness of Fano manifolds*, J. Differential Geom. **36** (1992), no. 3, 765–779. MR 94g:14021

[Kne59] M. Kneser, *Kleine Lösungen der diophantischen Gleichung* $ax^2 + by^2 = cz^2$, Abh. Math. Sem. Univ. Hamburg **23** (1959), 163–173. MR 0103880 (21 #2643)

[Kol92] J. Kollár, *Flips and abundance for algebraic threefolds*, Société Mathématique de France, Paris, 1992, Papers from the Second Summer Seminar on Algebraic Geometry held at the University of Utah, Salt Lake City, Utah, August 1991, Astérisque No. 211 (1992). MR 1225842 (94f:14013)

[Kon86] S. Kondō, *Enriques surfaces with finite automorphism groups*, Japan. J. Math. (N.S.) **12** (1986), no. 2, 191–282. MR 914299 (89c:14058)

[KST89] B. È. Kunyavskiĭ, A. N. Skorobogatov, and M. A. Tsfasman, *del Pezzo surfaces of degree four*, Mém. Soc. Math. France (N.S.) (1989), no. 37, 113. MR 1016354 (90k:14035)

[KT04a] A. Kresch and Y. Tschinkel, *On the arithmetic of del Pezzo surfaces of degree 2*, Proc. London Math. Soc. (3) **89** (2004), no. 3, 545–569. MR 2107007 (2005h:14060)

[KT04b] ———, *On the arithmetic of del Pezzo surfaces of degree 2*, Proc. London Math. Soc. (3) **89** (2004), no. 3, 545–569. MR 2107007 (2005h:14060)

[KT08] ———, *Effectivity of Brauer-Manin obstructions*, Adv. Math. **218** (2008), no. 1, 1–27. MR 2409407

[Lan76] R. P. Langlands, *On the functional equations satisfied by Eisenstein series*, Lecture Notes in Mathematics, Vol. 544, Springer-Verlag, Berlin, 1976. MR 0579181 (58 #28319)

[Leh08] B. Lehmann, *A cone theorem for nef curves*, 2008, `arXiv:0807.2294`.

[LO77] J. C. Lagarias and A. M. Odlyzko, *Effective versions of the Chebotarev density theorem*, Algebraic number fields: L-functions and Galois properties (Proc. Sympos., Univ. Durham, Durham, 1975), Academic Press, London, 1977, pp. 409–464. MR 0447191 (56 #5506)

[LY02] K. F. Lai and K. M. Yeung, *Rational points in flag varieties over function fields*, J. Number Theory **95** (2002), no. 2, 142–149. MR 1924094 (2003i:11089)

[Man86] Y. I. Manin, *Cubic forms*, second ed., North-Holland Publishing Co., Amsterdam, 1986. MR 87d:11037

[Man93] ———, *Notes on the arithmetic of Fano threefolds*, Compositio Math. **85** (1993), no. 1, 37–55. MR 1199203 (94c:11053)

[Man95] E. Manduchi, *Root numbers of fibers of elliptic surfaces*, Compositio Math. **99** (1995), no. 1, 33–58. MR 1352567 (96j:11076)

[Mas02] D. W. Masser, *Search bounds for Diophantine equations*, A panorama of number theory or the view from Baker's garden (Zürich, 1999), Cambridge Univ. Press, Cambridge, 2002, pp. 247–259. MR 1975456 (2004f:11027)

[Mat00] Y. V. Matiyasevich, *Hilbert's tenth problem: what was done and what is to be done*, Hilbert's tenth problem: relations with arithmetic and algebraic geometry (Ghent, 1999), Contemp. Math., vol. 270, Amer. Math. Soc., Providence, RI, 2000, pp. 1–47. MR 1802008 (2001m:03084)

[Mat02] K. Matsuki, *Introduction to the Mori program*, Universitext, Springer-Verlag, New York, 2002. MR 1875410 (2002m:14011)

[Mat06] Y. V. Matiyasevich, *Hilbert's tenth problem: Diophantine equations in the twentieth century*, Mathematical events of the twentieth century, Springer, Berlin, 2006, pp. 185–213. MR 2182785 (2006j:01012)

[Maz77] B. Mazur, *Modular curves and the Eisenstein ideal*, Inst. Hautes Études Sci. Publ. Math. (1977), no. 47, 33–186 (1978). MR 80c:14015

[McK00] D. McKinnon, *Counting rational points on $K3$ surfaces*, J. Number Theory **84** (2000), no. 1, 49–62. MR 1782261 (2001j:14031)

[McM02] C. T. McMullen, *Dynamics on $K3$ surfaces: Salem numbers and Siegel disks*, J. Reine Angew. Math. **545** (2002), 201–233. MR 1896103 (2003a:37057)

[Mer96] L. Merel, *Bornes pour la torsion des courbes elliptiques sur les corps de nombres*, Invent. Math. **124** (1996), no. 1-3, 437–449. MR 1369424 (96i:11057)

[Mil80] J. S. Milne, *Étale cohomology*, Princeton Mathematical Series, vol. 33, Princeton University Press, Princeton, N.J., 1980. MR 559531 (81j:14002)

[Min89] Kh. P. Minchev, *Strong approximation for varieties over an algebraic number field*, Dokl. Akad. Nauk BSSR **33** (1989), no. 1, 5–8, 92. MR 984929 (90a:11065)

[MM86] S. Mori and S. Mukai, *Classification of Fano 3-folds with $B_2 \geq 2$. I*, Algebraic and topological theories (Kinosaki, 1984), Kinokuniya, Tokyo, 1986, pp. 496–545. MR 1 102 273

[MM03] ———, *Erratum: "Classification of Fano 3-folds with $B_2 \geq 2$" [Manuscripta Math. **36** (1981/82), no. 2, 147–162]*, Manuscripta Math. **110** (2003), no. 3, 407. MR 1969009

[MM82] ———, *Classification of Fano 3-folds with $B_2 \geq 2$*, Manuscripta Math. **36** (1981/82), no. 2, 147–162. MR 83f:14032

[Moo65] C. C. Moore, *Decomposition of unitary representations defined by discrete subgroups of nilpotent groups*, Ann. of Math. (2) **82** (1965), 146–182. MR 0181701 (31 #5928)

[Mor82] S. Mori, *Threefolds whose canonical bundles are not numerically effective*, Ann. of Math. (2) **116** (1982), no. 1, 133–176. MR 84e:14032

[MT86] Y. I. Manin and M. A. Tsfasman, *Rational varieties: algebra, geometry, arithmetic*, Uspekhi Mat. Nauk **41** (1986), no. 2(248), 43–94. MR 842161 (87k:11065)

[MW94] C. Mœglin and J. Waldspurger, *Décomposition spectrale et séries d'Eisenstein*, Progress in Mathematics, vol. 113, Birkhäuser Verlag, Basel, 1994, Une paraphrase de l'Écriture. [A paraphrase of Scripture]. MR 1261867 (95d:11067)

[Nik81] V. V. Nikulin, *Quotient-groups of groups of automorphisms of hyperbolic forms by subgroups generated by 2-reflections. Algebro-geometric applications*, Current problems in mathematics, Vol. 18, Akad. Nauk SSSR, Vsesoyuz. Inst. Nauchn. i Tekhn. Informatsii, Moscow, 1981, pp. 3–114. MR 633160 (83c:10030)

[NP89] T. Nadesalingam and J. Pitman, *Bounds for solutions of simultaneous diagonal equations of odd degree*, Théorie des nombres (Quebec, PQ, 1987), de Gruyter, Berlin, 1989, pp. 703–734. MR 1024598 (91f:11021)

[Oh08] H. Oh, *Orbital counting via mixing and unipotent flows*, 2008, Lecture notes - Clay Summer school, Pisa 2007.

[Pey95] E. Peyre, *Hauteurs et mesures de Tamagawa sur les variétés de Fano*, Duke Math. J. **79** (1995), no. 1, 101–218. MR 1340296 (96h:11062)

[Pey04] ———, *Counting points on varieties using universal torsors*, Arithmetic of higher-dimensional algebraic varieties (Palo Alto, CA, 2002), Progr. Math., vol. 226, Birkhäuser Boston, Boston, MA, 2004, pp. 61–81. MR 2029862 (2004m:11103)

[Pey05] ———, *Obstructions au principe de Hasse et à l'approximation faible*, Astérisque (2005), no. 299, Exp. No. 931, viii, 165–193, Séminaire Bourbaki. Vol. 2003/2004. MR 2167206 (2007b:14041)

[Pil95] J. Pila, *Density of integral and rational points on varieties*, Astérisque (1995), no. 228, 4, 183–187, Columbia University Number Theory Seminar (New York, 1992). MR 1330933 (96b:11043)

[Pil96] ———, *Density of integer points on plane algebraic curves*, Internat. Math. Res. Notices (1996), no. 18, 903–912. MR 1420555 (97m:11126)

[Pit71] J. Pitman, *Bounds for solutions of diagonal equations*, Acta Arith. **19** (1971), 223–247. (loose errata). MR 0297701 (45 #6753)

[Poo08a] B. Poonen, *Insufficiency of the Brauer–Manin obstruction applied to étale covers*, 2008, `arXiv:0806.1312`.

[Poo08b] ———, *Undecidability in number theory*, Notices Amer. Math. Soc. **55** (2008), no. 3, 344–350. MR 2382821

[Pop01] O. N. Popov, *Del Pezzo surfaces and algebraic groups*, 2001, MA thesis, University of Tübingen.

[PT01] E. Peyre and Y. Tschinkel (eds.), *Rational points on algebraic varieties*, Progress in Mathematics, vol. 199, Birkhäuser Verlag, Basel, 2001. MR 1875168 (2002g:11006)

[Puk98] A. V. Pukhlikov, *Birational automorphisms of higher-dimensional algebraic varieties*, Proceedings of the International Congress of Mathematicians, Vol. II (Berlin, 1998), no. Extra Vol. II, 1998, pp. 97–107. MR 1648060 (99k:14026)

[Puk07] ———, *Birationally rigid varieties. I. Fano varieties*, Uspekhi Mat. Nauk **62** (2007), no. 5(377), 15–106. MR 2373751

[PV04] B. Poonen and J. F. Voloch, *Random Diophantine equations*, Arithmetic of higher-dimensional algebraic varieties (Palo Alto, CA, 2002), Progr. Math., vol. 226, Birkhäuser Boston, Boston, MA, 2004, With appendices by Jean-Louis Colliot-Thélène and Nicholas M. Katz, pp. 175–184. MR 2029869 (2005g:11055)

[Sal84] D. J. Saltman, *Noether's problem over an algebraically closed field*, Invent. Math. **77** (1984), no. 1, 71–84. MR 751131 (85m:13006)

[Sal98] P. Salberger, *Tamagawa measures on universal torsors and points of bounded height on Fano varieties*, Astérisque (1998), no. 251, 91–258, Nombre et répartition de points de hauteur bornée (Paris, 1996). MR 1679841 (2000d:11091)

[Sal07] ———, *On the density of rational and integral points on algebraic varieties*, J. Reine Angew. Math. **606** (2007), 123–147. MR 2337644

[San81] J.-J. Sansuc, *Groupe de Brauer et arithmétique des groupes algébriques linéaires sur un corps de nombres*, J. Reine Angew. Math. **327** (1981), 12–80. MR 631309 (83d:12010)

[SB88] N. I. Shepherd-Barron, *The rationality of some moduli spaces of plane curves*, Compositio Math. **67** (1988), no. 1, 51–88. MR 949271 (89k:14011)

[SB89] ———, *Rationality of moduli spaces via invariant theory*, Topological methods in algebraic transformation groups (New Brunswick, NJ, 1988), Progr. Math., vol. 80, Birkhäuser Boston, Boston, MA, 1989, pp. 153–164. MR 1040862 (91b:14010)

[Sch79] S. Schanuel, *Heights in number fields*, Bull. Soc. Math. France **107** (1979), 433–449.

[Sch85] W. M. Schmidt, *The density of integer points on homogeneous varieties*, Acta Math. **154** (1985), no. 3-4, 243–296. MR 781588 (86h:11027)

[Sch08] M. Schuett, *K3 surfaces with Picard rank 20*, 2008, `arXiv:0804.1558`.

[SD72] P. Swinnerton-Dyer, *Rational points on del Pezzo surfaces of degree* 5, Algebraic geometry, Oslo 1970 (Proc. Fifth Nordic Summer School in Math.), Wolters-Noordhoff, Groningen, 1972, pp. 287–290. MR 0376684 (51 #12859)

[SD93] ———, *The Brauer group of cubic surfaces*, Math. Proc. Cambridge Philos. Soc. **113** (1993), no. 3, 449–460. MR 1207510 (94a:14038)

[Ser89] J.-P. Serre, *Lectures on the Mordell-Weil theorem*, Aspects of Mathematics, E15, Friedr. Vieweg & Sohn, Braunschweig, 1989, Translated from the French and edited by Martin Brown from notes by Michel Waldschmidt. MR 1002324 (90e:11086)

[Ser90] ———, *Spécialisation des éléments de* $\mathrm{Br}_2(\mathbf{Q}(T_1, \cdots, T_n))$, C. R. Acad. Sci. Paris Sér. I Math. **311** (1990), no. 7, 397–402. MR 1075658 (91m:12007)

[Sil91] J. H. Silverman, *Rational points on K3 surfaces: a new canonical height*, Invent. Math. **105** (1991), no. 2, 347–373. MR 1115546 (92k:14025)

[Ski97] C. M. Skinner, *Forms over number fields and weak approximation*, Compositio Math. **106** (1997), no. 1, 11–29. MR 1446148 (98b:14021)

[Sko93] A. N. Skorobogatov, *On a theorem of Enriques-Swinnerton-Dyer*, Ann. Fac. Sci. Toulouse Math. (6) **2** (1993), no. 3, 429–440. MR 1260765 (95b:14018)

[Sko99] ———, *Beyond the Manin obstruction*, Invent. Math. **135** (1999), no. 2, 399–424. MR 1666779 (2000c:14022)

[Sko01] ———, *Torsors and rational points*, Cambridge Tracts in Mathematics, vol. 144, Cambridge University Press, Cambridge, 2001. MR 1845760 (2002d:14032)

[SS07] V. V. Serganova and A. N. Skorobogatov, *Del Pezzo surfaces and representation theory*, Algebra Number Theory **1** (2007), no. 4, 393–419. MR 2368955 (2009b:14070)

[SS08] ———, *On the equations for universal torsors over Del Pezzo surfaces*, 2008, `arXiv:0806.0089` .

[SSD98] J. B. Slater and P. Swinnerton-Dyer, *Counting points on cubic surfaces. I*, Astérisque (1998), no. 251, 1–12, Nombre et répartition de points de hauteur bornée (Paris, 1996). MR 1679836 (2000d:11087)

[ST99] M. Strauch and Y. Tschinkel, *Height zeta functions of toric bundles over flag varieties*, Selecta Math. (N.S.) **5** (1999), no. 3, 325–396. MR 1723811 (2001h:14028)

[ST04] J. Shalika and Y. Tschinkel, *Height zeta functions of equivariant compactifications of the Heisenberg group*, Contributions to automorphic forms, geometry, and number theory, Johns Hopkins Univ. Press, Baltimore, MD, 2004, pp. 743–771. MR 2058627 (2005c:11110)

[STBT07] J. Shalika, R. Takloo-Bighash, and Y. Tschinkel, *Rational points on compactifications of semi-simple groups*, J. Amer. Math. Soc. **20** (2007), no. 4, 1135–1186 (electronic). MR 2328719 (2008g:14034)

[Sto07] M. Stoll, *Finite descent obstructions and rational points on curves*, Algebra Number Theory **1** (2007), no. 4, 349–391. MR 2368954 (2008i:11086)

[Str98] M. Strauch, *Höhentheoretische Zetafunktionen von Faserbündeln über verallgemeinerten Fahnenvarietäten*, Bonner Mathematische Schriften [Bonn Mathematical Publications], 309, Universität Bonn Mathematisches Institut, Bonn, 1998, Dissertation, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, 1997. MR 1937786 (2003k:11110)

[Str01] ———, *Arithmetic stratifications and partial Eisenstein series*, Rational points on algebraic varieties, Progr. Math., vol. 199, Birkhäuser, Basel, 2001, pp. 335–355. MR 1875180 (2002m:11057)

[STV07] M. Stillman, D. Testa, and M. Velasco, *Gröbner bases, monomial group actions, and the Cox rings of del Pezzo surfaces*, J. Algebra **316** (2007), no. 2, 777–801. MR 2358614 (2008i:14054)

[SW95] P. Sarnak and L. Wang, *Some hypersurfaces in* $\mathbf{P}^4$ *and the Hasse-principle*, C. R. Acad. Sci. Paris Sér. I Math. **321** (1995), no. 3, 319–322. MR 1346134 (96j:14014)

[SX08] B. Sturmfels and Zh. Xu, *Sagbi bases of Cox–Nagata rings*, 2008, `arXiv:0803.0892`.

[TVAV08] D. Testa, A. Varilly-Alvarado, and M. Velasco, *Cox rings of degree one Del Pezzo surfaces*, 2008, `arXiv:0803.0353`.

[Ura96] T. Urabe, *Calculation of Manin's invariant for Del Pezzo surfaces*, Math. Comp. **65** (1996), no. 213, 247–258, S15–S23. MR 1322894 (96f:14047)

[VAZ08] A. Varilly-Alvarado and D. Zywina, *Arithmetic* E_8 *lattices with maximal Galois action*, 2008, `arXiv:0803.3063`.

[vL07] R. van Luijk, *K3 surfaces with Picard number one and infinitely many rational points*, Algebra Number Theory **1** (2007), no. 1, 1–15. MR 2322921 (2008d:14058)

[Voi04] C. Voisin, *Intrinsic pseudo-volume forms and* K*-correspondences*, The Fano Conference, Univ. Torino, Turin, 2004, pp. 761–792. MR 2112602 (2006b:14020)

[VW95] R. C. Vaughan and T. D. Wooley, *On a certain nonary cubic form and related equations*, Duke Math. J. **80** (1995), no. 3, 669–735. MR 96j:11038

[War35] E. Warning, *Bemerkung zur vorstehenden Arbeit von Herrn Chevalley*, Abh. Math. Semin. Hamb. Univ. **11** (1935), 76–83.

[Wey16] H. Weyl, *Über die Gleichverteilung von Zahlen mod. Eins*, Math. Ann. **77** (1916), no. 3, 313–352. MR 1511862

[Wit04] O. Wittenberg, *Transcendental Brauer-Manin obstruction on a pencil of elliptic curves*, Arithmetic of higher-dimensional algebraic varieties (Palo Alto, CA, 2002), Progr. Math., vol. 226, Birkhäuser Boston, Boston, MA, 2004, pp. 259–267. MR 2029873 (2005c:11082)

[Zar08] Y. G. Zarhin, *Del Pezzo surfaces of degree 1 and Jacobians*, Math. Ann. **340** (2008), no. 2, 407–435. MR 2368986

Courant Institute of Mathematical Sciences, NYU, 251 Mercer Street, New York, NY 10012, USA

E-mail address: `tschinkel@cims.nyu.edu`

Clay Mathematics Proceedings
Volume 8, 2009

Birational geometry for number theorists

Dan Abramovich

Abstract. We introduce some of the ideas and tools of birational geometry which play a role in conjectures by Bombieri, Lang, Vojta and Campana on the relationship between arithmetic and geometry. After a brief discussion of geometry and arithmetic on curves in Section 0, we discuss Kodaira dimension of a variety and its conjectural relationship with arithmetic properties in Section 1. In Section 2 we outline Campana's approach aiming for a more solid conjectural relationship with arithmetic through the core map. Section 3 outlines the minimal model program and discusses its current status. In Section 4 we review Vojta's conjectures and their relationship to Campana's conjectures and to the *abc* conjecture of Masser-Oesterlé.

Contents

Introduction

When thinking about the course "birational geometry for number theorists" I so naïvely agreed to give at the Göttingen summer school, I could not avoid imagining the spirit of the late Serge Lang, not so quietly beseeching one to do things right, keeping the theorems functorial with respect to ideas, and definitions natural. But most important is the fundamental tenet of Diophantine geometry, for which Lang was one of the strongest and loudest advocates, which was so aptly summarized in the introduction of Hindry-Silverman [**HS00**]:

GEOMETRY DETERMINES ARITHMETIC.

2000 *Mathematics Subject Classification.* Primary 14E30, Secondary 11G35, 13E05, 14E15.
Partial support for research provided by NSF grants DMS-0301695 and DMS-0603284.

To make sense of this, largely conjectural, epithet, it is good to have some loose background in birational geometry, which I will try to provide. For the arithmetic motivation I will explain conjectures of Bombieri, Lang and Vojta, and new and exciting versions of those due to Campana. In fact, I imagine Lang would insist (strongly, as only he could) that Campana's conjectures most urgently need further investigation, and indeed in some sense they form the centerpiece of these notes.

Birational geometry is undergoing revolutionary developments these very days: large portions of the minimal model program were solved soon after the Göttingen lectures [**BCHM06**], and it seems likely that more is to come. Also, a number of people seem to have made new inroads into the long standing resolution of singularities problem. I am not able to report on the latter, but I will give a brief account of the minimal model program as it seems to stand at this point in time.

Our convention: a variety over k is an *absolutely* reduced and irreducible scheme of finite type over k.

Acknowledgements: I thank the CMI and the organizers for inviting me, I thank the colleagues and students at Brown for their patience with my ill prepared preliminary lectures and numerous suggestions, I thank F. Campana for a number of inspiring discussions, H.-H. Tseng and H. Ulfarsson for a number of good comments, and L. Caporaso for the notes of her MSRI lecture [**Cap**], to which my lecture plans grew increasingly close. The treatment of the minimal model program is influenced by lectures of Ch. Hacon and J. McKernan and discussions with them. Many thanks are due to the referee who caught a large number of errors and made numerous suggestions. Of course all remaining errors are entirely my responsibility. Anything new is partially supported by the NSF grants DMS-0301695 and DMS-0603284.

0. Geometry and arithmetic of curves

The arithmetic of algebraic curves is one area where basic relationships between geometry and arithmetic are known, rather than conjectured. Much of the material here is covered in Darmon's lectures of this summer school.

0.1. Closed curves. Consider a smooth projective algebraic curve C defined over a number field k. We are interested in a qualitative relationship between its arithmetic and geometric properties.

We have three basic facts:

0.1.1. A curve of genus 0 becomes rational after at most a quadratic extension k' of k, in which case its set of rational points $C(k')$ is infinite (and therefore dense in the Zariski topology).

0.1.2. A curve of genus 1 has a rational point after a finite extension k' of k (though the degree is not a priori bounded), and has positive Mordell–Weil rank after a further quadratic extension k''/k', in which case again its set of rational points $C(k'')$ is infinite (and therefore dense in the Zariski topology).

We can immediately introduce the following definition:

Definition 0.1.3. Let X be an algebraic variety defined over k. We say that rational points on X are potentially dense if there is a finite extension k'/k such that the set $X(k')$ is dense in $X_{k'}$ in the Zariski topology.

Thus rational points on a curve of genus 0 or 1 are potentially dense.

Finally we have

THEOREM 0.1.4 (Faltings, 1983). *Let C be an algebraic curve of genus > 1 over a number field k. Then $C(k)$ is finite.*

See, e.g. [**Fal83, HS00**].

In other words, rational points on a curve C of genus g are potentially dense if and only if $g \leq 1$.

0.1.5. So far there isn't much birational geometry involved, because we have the old theorem:

THEOREM 0.1.6. *A smooth algebraic curve is uniquely determined by its function field.*

But this is an opportunity to introduce a tool: on the curve C we have a canonical divisor class K_C, such that $\mathcal{O}_C(K_C) = \Omega^1_C$, the sheaf of differentials, also known by the notation ω_C—the dualizing sheaf. We have:

(1) $\deg K_C = 2g - 2 = -\chi^{\mathrm{top}}(C_{\mathbb{C}})$, where $\chi^{\mathrm{top}}(C_{\mathbb{C}})$ is the topological Euler characteristic of the complex Riemann surface $C_{\mathbb{C}}$.
(2) $\dim H^0(C, \mathcal{O}_C(K_C)) = g$.

For future discussion, the first property will be useful. We can now summarize, following [**HS00**]:

0.1.7.

Degree of K_C	rational points
$2g - 2 \leq 0$	potentially dense
$2g - 2 > 0$	finite

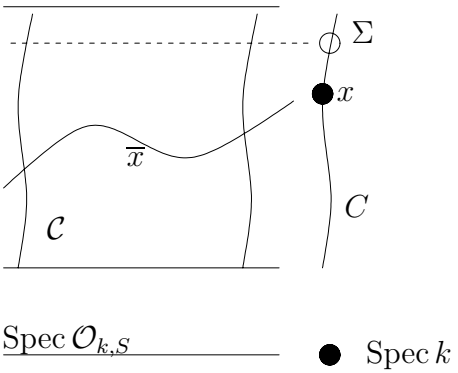
0.2. Open curves.

0.2.1. Consider a smooth quasi-projective algebraic curve C defined over a number field k. It has a unique smooth projective completion $C \subset \overline{C}$, and the complement is a finite set $\Sigma = \overline{C} \smallsetminus C$. Thinking of Σ as a reduced divisor of some degree n, a natural line bundle to consider is $\mathcal{O}_{\overline{C}}(K_{\overline{C}} + \Sigma) \simeq \omega_{\mathcal{O}}(\Sigma)$, the sheaf of differentials with logarithmic poles on Σ, whose degree is again $-\chi^{\mathrm{top}}(C) = 2g - 2 + n$. The sign of $2g - 2 + n$ again serves as the geometric invariant to consider.

0.2.2. Consider for example the affine line. Rational points on the affine line are not much more interesting than those on $\mathbb{P}^1$. But we can also consider the behavior of *integral* points, where interesting results do arise. However, what does one mean by integral points on $\mathbb{A}^1$? The key is that integral points are an invariant of an "integral model" of $\mathbb{A}^1$ over $\mathbb{Z}$.

0.2.3. Consider the ring of integers $\mathcal{O}_k$ and a finite set $S \subset \operatorname{Spec} \mathcal{O}_k$ of finite primes. One can associate to it the ring $\mathcal{O}_{k,S}$ of S-integers, of elements in K which are in $\mathcal{O}_\wp$ for any prime $\wp \notin S$.

Now consider a *model* of C over $\mathcal{O}_{k,S}$, namely a scheme $\mathcal{C}$ of finite type over $\mathcal{O}_{k,S}$ with an isomorphism of the generic fiber $\mathcal{C}_k \simeq C$. It is often useful to start with a model $\overline{\mathcal{C}}$ of $\overline{C}$, and take $\mathcal{C} = \overline{\mathcal{C}} \smallsetminus \overline{\Sigma}$, where $\overline{\Sigma}$ is the closure of Σ in $\overline{\mathcal{C}}$.



Now it is clear how to define integral points: an S-integral point on $\mathcal{C}$ is simply an element of $\mathcal{C}(\mathcal{O}_{k,S})$, in other words, a section of $\mathcal{C} \to \mathrm{Spec}(\mathcal{O}_{k,S})$. This is related to rational points on a proper curve as follows:

0.2.4. If $\Sigma = \emptyset$, and the model chosen is proper, the notions of integral and rational points agree, because of the valuative criterion for properness.

EXERCISE 0.2.5. Prove this!

We have the following facts:

0.2.6. If C is rational and $n \leq 2$, then after possibly enlarging k and S, any integral model of C has an infinite collection of integral points.

EXERCISE 0.2.7. Prove this!

On the other hand, we have:

THEOREM 0.2.8 (Siegel's Theorem). *If $n \geq 3$, or if $g > 0$ and $n > 0$, then for any integral model $\mathcal{C}$ of C, the set of integral points $\mathcal{C}(\mathcal{O}_{k,S})$ is finite.*

A good generalization of Definition 0.1.3 is the following:

DEFINITION 0.2.9. Let X be an algebraic variety defined over k with a model $\mathcal{X}$ over $\mathcal{O}_{k,S}$. We say that integral points on X are potentially dense if there is a finite extension k'/k, and an enlargement S' of the set of places in k' over S, such that the set $\mathcal{X}(\mathcal{O}_{k',S'})$ is dense in $X_{k'}$ in the Zariski topology.

We can apply this definition in the case of a curve C and generalize 0.1.7, as in [**HS00**], as follows:

0.2.10.

degree of $K_{\overline{C}} + \Sigma$	integral points
$2g - 2 + n \leq 0$	potentially dense
$2g - 2 + n > 0$	finite

0.2.11. One lesson we must remember from this discussion is that

For **open** varieties we use **integral** points on **integral models**.

0.3. Faltings implies Siegel. Siegel's theorem was proven years before Faltings's theorem. Yet it is instructive, especially in the later parts of these notes, to give the following argument showing that Faltings's theorem implies Siegel's.

THEOREM 0.3.1 (Hermite-Minkowski, see [**HS00**] page 264). *Let k be a number field, $S \subset \mathrm{Spec}\,\mathcal{O}_{k,S}$ a finite set of finite places, and d a positive integer. Then there are only finitely many extensions k'/k of degree $\leq d$ unramified outside S.*

From which one can deduce

THEOREM 0.3.2 (Chevalley-Weil, see [**HS00**] page 292). *Let $\pi : \mathcal{X} \to \mathcal{Y}$ be a finite étale morphism of schemes over $\mathcal{O}_{k,S}$. Then there is a finite extension k'/k, with S' lying over S, such that $\pi^{-1}\mathcal{Y}(\mathcal{O}_{k,S}) \subset \mathcal{X}(\mathcal{O}_{k',S'})$.*

On the geometric side we have an old topological result

THEOREM 0.3.3. *If C is an open curve with $2g-2+n>0$ and $n>0$, defined over k, there is a finite extension k'/k and a finite unramified covering $D \to C$, such that $g(D)>1$.*

EXERCISE 0.3.4. Combine these theorems to obtain a proof of Siegel's theorem assuming Faltings's theorem.

This is discussed in Darmon's lectures, as well as [**HS00**].

0.3.5. Our lesson this time is that

Rational and integral points can be controlled in finite étale covers.

0.4. Function field case. There is an old and distinguished tradition of comparing results over number fields with results over function fields. To avoid complications I will concentrate on function fields of characteristic 0, and consider closed curves only.

0.4.1. If K is the function field of a complex variety B, then a variety X/K is the generic fiber of a scheme $\mathcal{X}/B$, and a K-rational point $P \in X(K)$ can be thought of as a rational section of $\mathcal{X} \to B$. If B is a smooth curve and $\mathcal{X} \to B$ is proper, then again a K-rational point $P \in X(K)$ is equivalent to a *regular* section $B \to \mathcal{X}$.

EXERCISE 0.4.2. Make sense of this (i.e., prove this)!

0.4.3. The notion of *integral points* is similarly defined using sections. When $\dim B > 1$ there is an intermediate notion of *properly rational points*: a K-rational point p of X is a properly rational point of $\mathcal{X}/B$ if the closure B' of p in $\mathcal{X}$ maps properly to B.

Consider now C/K a curve. Of course it is possible that $\mathcal{C}$ is, or is birationally equivalent to, $C_0 \times B$, in which case we have plenty of constant sections coming from $C_0(\mathbb{C})$, corresponding to constant points in $C(K)$. But that is almost all there is:

THEOREM 0.4.4 (Manin [**Man63**], Grauert [**Gra65**]). *Let k be a field of characteristic 0, let K be a regular extension of k, and C/K a smooth curve. Assume $g(C)>1$. If $C(K)$ is infinite, then there is a curve C_0/k with $(C_0)_K \simeq C$, such that $C(K) \smallsetminus C_0(k)$ is finite.*

EXERCISE 0.4.5. What does this mean for constant curves $C_0 \times B$ in terms of maps from C_0 to B?

Working inductively on transcendence degree, and using Faltings's Theorem, we obtain:

THEOREM 0.4.6. *Let C be a curve of genus >1 over a field k finitely generated over $\mathbb{Q}$. Then the set of k-rational points $C(k)$ is finite.*

EXERCISE 0.4.7. Prove this, using previous results as given!

See [**Sam66**], [**MS02**] for appropriate statements in positive characteristics.

1. Kodaira dimension

1.1. Iitaka dimension. Consider now a smooth, projective variety X of dimension d over a field k of characteristic 0. We seek an analogue of the sign of $2g-2$ in this case. The approach is by counting sections of the canonical line bundle $\mathcal{O}_X(K_X) = \bigwedge^d \Omega^1_X$. Iitaka's book [**Iit82**] is a good reference.

THEOREM 1.1.1. *Let L be a line bundle on X. Assume $h^0(X, L^n)$ does not vanish for all positive integers n. Then there is a unique integer $\kappa = \kappa(X, L)$ with $0 \leq \kappa \leq d$ such that*

$$\limsup_{n\to\infty} \frac{h^0(X, L^n)}{n^\kappa}$$

exists and is nonzero.

DEFINITION 1.1.2.
(1) The integer $\kappa(X, L)$ in the theorem is called *the Iitaka dimension of* (X, L).
(2) In the special case $L = \mathcal{O}_X(K_X)$ we write $\kappa(X) := \kappa(X, L)$ and call $\kappa(X)$ *the Kodaira dimension of* X.
(3) It is customary to set $\kappa(X, L)$ to be either -1 or $-\infty$ if $h^0(X, L^n)$ vanishes for all positive integers n. It is safest to say in this case that the Iitaka dimension is *negative*. I will use $-\infty$.

We will see an algebraic justification for the -1 convention immediately in Proposition 1.1.3, and a geometric justification for the more commonly used $-\infty$ in Paragraph 1.2.7.

An algebraically meaningful presentation of the Iitaka dimension is the following:

PROPOSITION 1.1.3. *Consider the algebra of sections*

$$\mathcal{R}(X, L) := \bigoplus_{n\geq 0} H^0(X, L^n).$$

Then, with the -1 convention,

$$\text{tr. deg}\, \mathcal{R}(X, L) = \kappa(X, L) + 1.$$

DEFINITION 1.1.4. We say that a property holds for a sufficiently high and divisible n if there exists $n_0 > 0$ such that the property holds for every positive multiple of n_0.

A geometric meaning of $\kappa(X, L)$ is given by the following:

PROPOSITION 1.1.5. *Assume $\kappa(X, L) \geq 0$. Then for sufficiently high and divisible n, the dimension of the image of the rational map $\phi_{L^n} : X \dashrightarrow \mathbb{P}H^0(X, L^n)$ is precisely $\kappa(X, L)$.*

Even more precise is:

PROPOSITION 1.1.6. *There is $n_0 > 0$ such that the image $\phi_{L^n}(X)$ is birational to $\phi_{L^{n_0}}(X)$ for all $n > 0$ divisible by n_0.*

DEFINITION 1.1.7.
(1) The birational equivalence class of the variety $\phi_{L^{n_0}}(X)$ is denoted by $I(X, L)$.
(2) The rational map $X \to I(X, L)$ is called the *Iitaka fibration* of (X, L).

(3) In case L is the canonical bundle ω_X, this map is simply called the Iitaka fibration of X, written $X \to I(X)$

The following notion is important:

DEFINITION 1.1.8. The variety X is said to be *of general type* of $\kappa(X) = \dim X$.

REMARK 1.1.9. The name is not as informative as one could wish. It comes from the observation that surfaces not of general type can be nicely classified, whereas there is a whole zoo of surfaces of general type.

EXERCISE 1.1.10. Prove Proposition 1.1.6:

(1) Show that if $n, d > 0$ and $H^0(X, L^n) \neq 0$ then there is a dominant rational map $\phi_{L^{nd}}(X) \dashrightarrow \phi_{L^n}(X)$ such that the following diagram is commutative:

$$\begin{array}{ccc} X & \overset{\phi_{L^{nd}}}{\dashrightarrow} & \phi_{L^{nd}}(X) \\ & \searrow^{\phi_{L^n}} & \downarrow \\ & & \phi_{L^n}(X). \end{array}$$

(2) Conclude that $\dim \phi_{L^n}(X)$ is a constant κ for sufficiently high and divisible n.
(3) Suppose $n > 0$ satisfies $\kappa := \dim \phi_{L^n}(X)$. Show that for any $d > 0$, the function field of $\phi_{L^{nd}}(X)$ is algebraic over the function field $\phi_{L^n}(X)$.
(4) Recall that for any variety X, any subfield L of $K(X)$ containing k is finitely generated. Apply this to the algebraic closure of $\phi_{L^n}(X)$ in $K(X)$ to complete the proof of the proposition.

For details see [**Iit82**].

EXERCISE 1.1.11. Use Proposition 1.1.6 to prove Theorem 1.1.1.

1.2. Properties and examples of the Kodaira dimension.

EXERCISE 1.2.1. Show that $\kappa(\mathbb{P}^n) = -\infty$. Show that $\kappa(A) = 0$ for an abelian variety A.

1.2.2. *Curves:*

EXERCISE. Let C be a smooth projective curve and L a line bundle. Prove that

$$\kappa(C, L) = \begin{cases} 1 & \text{if } \deg_C L > 0, \\ 0 & \text{if } L \text{ is torsion, and} \\ < 0 & \text{otherwise.} \end{cases}$$

In particular,

$$\kappa(C) = \begin{cases} 1 & \text{if } g > 1, \\ 0 & \text{if } g = 1, \text{ and} \\ < 0 & \text{if } g = 0. \end{cases}$$

1.2.3. *Birational invariance:*

EXERCISE. Let $X' \dashrightarrow X$ be a birational map of smooth projective varieties. Show that the spaces $H^0(X, \mathcal{O}_X(mK_X))$ and $H^0(X', \mathcal{O}_{X'}(mK_{X'}))$ are canonically isomorphic. Deduce that $\kappa(X) = \kappa(X')$.

(See [**Har77**], Chapter II, Theorem 8.19).

1.2.4. *Generically finite dominant maps.*

EXERCISE. Let $f : X' \to X$ be a generically finite dominant map of smooth projective varieties.

Show that $\kappa(X') \geq \kappa(X)$.

1.2.5. *Finite étale maps.*

EXERCISE. Let $f : X' \to X$ be a finite étale map of smooth projective varieties.

Show that $\kappa(X') = \kappa(X)$.

1.2.6. *Field extensions:*

EXERCISE. Let k'/k be a field extension, X a variety over k with line bundle L, and $X_{k'}, L_{k'}$ the result of base change.

Show that $\kappa(X, L) = \kappa(X_{k'}, L_{k'})$. In particular $\kappa(X) = \kappa(X_{k'})$.

1.2.7. *Products.*

EXERCISE. Show that, with the $-\infty$ convention,

$$\kappa(X_1 \times X_2, L_1 \boxtimes L_2) = \kappa(X_1, L_1) + \kappa(X_2, L_2).$$

Deduce that $\kappa(X_1 \times X_2) = \kappa(X_1) + \kappa(X_2)$.

This so-called "easy additivity" of the Kodaira dimension is the main reason for the $-\infty$ convention.

1.2.8. *Fibrations.* The following is subtle and difficult:

THEOREM (Siu's theorem on deformation invariance of plurigenera [**Siu98, Siu02**]). *Let $X \to B$ be a smooth projective morphism with connected geometric fibers, and m a positive integer. Then for closed points $b \in B$, the dimension $h^0(X_b, \mathcal{O}(mK_{X_b}))$ is independent of $b \in B$. In particular $\kappa(X_b)$ is independent of the closed point $b \in B$.*

EXERCISE 1.2.9. Let $X \to B$ be a morphism of smooth projective varieties with connected geometric fibers. Let $b \in B$ be such that $X \to B$ is smooth over b, and let $\eta_B \in B$ be the generic point.

Use "cohomology and base change" and Siu's theorem to deduce that

$$\kappa(X_b) = \kappa(X_{\eta_B}).$$

DEFINITION 1.2.10. The Kodaira defect of X is $\delta(X) = \dim(X) - \kappa(X)$.

EXERCISE 1.2.11. Let $X \to B$ be a morphism of smooth projective varieties with connected geometric fibers. Show that the Kodaira defects satisfy $\delta(X) \geq \delta(X_{\eta_B})$. Equivalently $\kappa(X) \leq \dim(B) + \kappa(X_{\eta_B})$.

We remark that before Siu's deformation invariance theorem was proven, a weaker and more technical result, yet still very useful, saying that the Kodaira dimension is constant on "very general fibers" was used.

EXERCISE 1.2.12. Let $Y \to B$ be a morphism of smooth projective varieties with connected geometric fibers, and $Y \to X$ a generically finite map. Show that $\delta(X) \geq \delta(Y_{\eta_B})$. In other words, $\kappa(X) \leq \kappa(Y_{\eta_B}) + \dim B$.

This so-called "easy subadditivity" has many useful consequences.

DEFINITION 1.2.13. We say that X is uniruled if there is a variety B of dimension $\dim X - 1$ and a dominant rational map $B \times \mathbb{P}^1 \dashrightarrow X$.

EXERCISE 1.2.14. If X is uniruled, show that $\kappa(X) = -\infty$.

The converse is an important conjecture, sometimes known as the $(-\infty)$-Conjecture. It is a consequence of the "good minimal model" conjecture:

CONJECTURE 1.2.15. *Assume X is not uniruled. Then $\kappa(X) \geq 0$.*

EXERCISE 1.2.16. If X is covered by a family of elliptic curves, show that $\kappa(X) \leq \dim X - 1$.

1.2.17. *Surfaces.* Surfaces of Kodaira dimension < 2 are "completely classified". Some of these you can place in the following table using what you have learned so far. In the following description we give a representative of the birational class of each type:

κ	description
$-\infty$	ruled surfaces: $\mathbb{P}^2$ or $\mathbb{P}^1 \times C$
0	a. abelian surfaces b. bielliptic surfaces k. K3 surfaces e. Enriques surfaces
1	all other elliptic surfaces

1.2.18. *Iitaka's program.* Here is a central conjecture of birational geometry:

CONJECTURE (Iitaka). *Let $X \to B$ be a surjective morphism of smooth projective varieties. Then*

$$\kappa(X) \geq \kappa(B) + \kappa(X_{\eta_B}).$$

1.2.19. Major progress on this conjecture was made through the years by several geometers, including Fujita [**Fuj78**], Kawamata [**Kaw85**], Viehweg [**Vie82**] and Kollár [**Kol87**]. The key, which makes this conjecture plausible, is the semipositivity properties of the relative dualizing sheaf $\omega_{X/B}$, which originate from work of Arakelov and rely on deep Hodge theoretic arguments.

Two results will be important for these lectures.

THEOREM 1.2.20 (Kawamata). *Iitaka's conjecture follows from the Minimal Model Program: if X_{η_B} has a good minimal model then $\kappa(X) \geq \kappa(B) + \kappa(X_{\eta_B})$.*

THEOREM 1.2.21 (Viehweg). *Iitaka's conjecture holds in case B is of general type, namely:*

Let $X \to B$ be a surjective morphism of smooth projective varieties, and assume $\kappa(B) = \dim B$. Then $\kappa(X) = \dim(B) + \kappa(X_{\eta_B})$.

Note that equality here is forced by the easy subadditivity inequality: $\kappa(X) \leq \dim(B) + \kappa(X_{\eta_B})$ always holds.

EXERCISE 1.2.22. Let X, B_1, B_2 be smooth projective varieties. Suppose $X \to B_1 \times B_2$ is generically finite to its image, and assume both $X \to B_i$ surjective.

(1) Assume B_1, B_2 are of general type. Use Viehweg's theorem and the Kodaira defect inequality to conclude that X is of general type. (Hint for a key step: construct a subvariety of general type $V \subset B_1$, such that $X \times_{B_1} V \to B_2$ is generically finite and surjective.)

(2) Assume $\kappa(B_1), \kappa(B_2) \geq 0$. Show that if Iitaka's conjecture holds true, then $\kappa(X) \geq 0$.

EXERCISE 1.2.23. Let X be a smooth projective variety. Using the previous exercise, show that there is a dominant rational map

$$L_X : X \dashrightarrow L(X)$$

such that

(1) $L(X)$ is of general type, and
(2) the map is universal: if $g : X \dashrightarrow Z$ is a dominant rational map with Z of general type, there is a unique rational map $L(g) : L(X) \dashrightarrow Z$ such that the following diagram commutes:

$$\begin{array}{ccc} X & \overset{L_X}{\dashrightarrow} & L(X) \\ & \searrow^{g} & \downarrow{\scriptstyle L(g)} \\ & & Z. \end{array}$$

I call the map L_X *the Lang map of* X, and $L(X)$ the *Lang variety of* X.

1.3. Uniruled varieties and rationally connected fibrations.

1.3.1. *Uniruled varieties.* For simplicity let us assume here that k is algebraically closed.

As indicated above, a variety X is said to be *uniruled* if there is a $(d-1)$-dimensional variety B and a dominant rational map $B \times \mathbb{P}^1 \dashrightarrow X$. Instead of $B \times \mathbb{P}^1$ one can take any variety $Y \to B$ whose generic fiber is a curve of genus 0. As discussed above, if X is uniruled then $\kappa(X) = -\infty$. The converse is the important $(-\infty)$-Conjecture 1.2.15.

A natural question is, can one "take all these rational curves out of the picture?" The answer is yes, in the best possible sense.

DEFINITION 1.3.2. A smooth projective variety P is said to be *rationally connected* if through any two points $x, y \in P$ there is a morphism from a rational curve $C \to P$ having x and y in its image.

There are various equivalent ways to characterize rationally connected varieties.

THEOREM 1.3.3 (Campana [**Cam92**], Kollár-Miyaoka-Mori [**KMM92**]). *Let P be a smooth projective variety. The following are equivalent:*

(1) P is rationally connected.
(2) Any two points are connected by a chain of rational curves.
(3) For any finite set of points $S \subset P$, there is a morphism from a rational curve $C \to P$ having S in its image.
(4) There is a "very free" rational curve on P—if $\dim P > 2$ this means there is a rational curve $C \subset P$ such that the normal bundle $N_{C \subset P}$ is ample.

Key properties:

THEOREM 1.3.4 ([**Cam92, KMM92**]). *Let X and X' be smooth projective varieties, with X rationally connected.*

(1) If $X \dashrightarrow X'$ is a dominant rational map (in particular when X and X' are birationally equivalent) then X' is rationally connected.

(2) *If X' is deformation-equivalent to X then X' is rationally connected.*
(3) *If $X' = X_{k'}$ where k'/k is an algebraically closed field extension, then X' is rationally connected if and only if X is.*

EXERCISE 1.3.5. A variety is unirational if it is a dominant image of $\mathbb{P}^n$. Show that every unirational variety is rationally connected.

On the other hand, one expects the following:

CONJECTURE 1.3.6 (Kollár). *There is a rationally connected threefold which is not unirational. There should also exist some hypersurface of degree n in $\mathbb{P}^n$, $n \geq 4$ which is not unirational.*

Rational connectedness often arises when there is some negativity of differential forms, as in the following statement. A smooth projective variety X is Fano if its anti-canonical divisor is ample. We have the following:

THEOREM 1.3.7 (Kollár-Miyaoka-Mori, Campana). *A Fano variety is rationally connected.*

CONJECTURE 1.3.8 (Kollár-Miyaoka-Mori, Campana).

(1) *A variety X is rationally connected if and only if*
$$H^0(X, (\Omega^1_X)^{\otimes n}) = 0$$
for every positive integer n.
(2) *A variety X is rationally connected if and only if every positive dimensional dominant image $X \dashrightarrow Z$ has $\kappa(Z) = -\infty$.*

This conjecture follows from the minimal model program; see Conjecture 3.4.3 and 3.4.4.

Now we can break any variety X into a rationally connected fiber over a nonuniruled base:

THEOREM 1.3.9 (Campana, Kollár-Miyaoka-Mori, Graber-Harris-Starr). *Let X be a smooth projective variety. There is a birational morphism $X' \to X$, a variety $Z(X)$, and a dominant morphism $X' \to Z(X)$ with connected fibers, such that*

(1) *The general fiber of $X' \to Z(X)$ is rationally connected, and*
(2) *$Z(X)$ is not uniruled.*

Moreover, $X' \to X$ is an isomorphism in a neighborhood of the general fiber of $X' \to Z(X)$.

The existence of a fibration containing "most" rational curves was proven in the original papers by Campana and Kollár-Miyaoka-Mori. The crucial fact that $Z(X)$ is not uniruled was proven by Graber, Harris and Starr in [**GHS03**].

1.3.10. The rational map $r_X : X \dashrightarrow Z(X)$ is called the *maximally rationally connected fibration of X* (or MRC fibration of X) and $Z(X)$, which is well defined up to birational equivalence, is called the *MRC quotient* of X.

1.3.11. The MRC fibration has the universal property of being "final" for dominant rational maps $X \to B$ with rationally connected fibers.

One can construct similar fibrations with a similar universal property for maps with fibers having $H^0(X_b, (\Omega^1_{X_b})^{\otimes n}) = 0$, or for fibers having no dominant morphism to positive dimensional varieties of nonnegative Kodaira dimension. Conjecturally

these agree with r_X. Also conjecturally, assuming Iitaka's conjecture, there exists $X \dashrightarrow Z'$ which is initial for maps to varieties of non-negative Kodaira dimension. This conjecturally will also agree with r_X. All these conjectures would follow from the "good minimal model" conjecture.

1.3.12. *Arithmetic, finally.* The set of rational points on a rational curve is Zariski-dense. The following is a natural extension:

CONJECTURE 1.3.13 (Campana). *Let P be a rationally connected variety over a number field k. Then rational points on P are potentially dense.*

This conjecture and its sister 1.4.2 below was implicit in works of many, including Bogomolov, Colliot-Thélène, Harris, Hassett, Tschinkel.

1.4. Geometry and arithmetic of the Iitaka fibration. We now want to understand the geometry and arithmetic of varieties such as $Z(X)$, i.e., non-uniruled varieties. In view of Conjecture 1.2.15, we focus on the case $\kappa(X) \geq 0$.

So let X satisfy $\kappa(X) \geq 0$, and consider the Iitaka fibration $X \dashrightarrow I(X)$. The next proposition follows from easy subadditivity and Siu's theorem:

PROPOSITION 1.4.1. *Let F be a general fiber of $X \dashrightarrow I(X)$. Then $\kappa(F) = 0$.*

CONJECTURE 1.4.2 (Campana). *Let F be a variety over a number field k satisfying $\kappa(F) = 0$. Then rational points on F are potentially dense.*

EXERCISE 1.4.3. Recall the Lang map in 1.2.23. Assuming Conjecture 1.2.15, show that $L(X)$ is the result of applying MRC fibrations and Iitaka fibrations, alternating between the former and the latter, until the result stabilizes.

1.5. Lang's conjecture. In this section we let k be a number field, or any field which is finitely generated over $\mathbb{Q}$.

A highly inspiring conjecture in Diophantine geometry is the following:

CONJECTURE (Lang's conjecture, weak form). *Let X be a smooth projective variety of general type over k. Then $X(k)$ is not Zariski-dense in X.*

In fact, motivated by analogy with conjectures on the Kobayashi pseudo-metric of a variety of general type, Lang even proposed the following:

CONJECTURE (Lang's geometric conjecture). *Let X be a smooth complex projective variety of general type. There is a Zariski-closed proper subset $S(X) \subset X$, whose irreducible components are not of general type, and such that every irreducible subset $T \subset X$ not of general type is contained in $S(X)$.*

The notation "$S(X)$" stands for "the special subvariety of X". It is not hard to see that $S(X)$ is defined over any field of definition of X. The two conjectures combine to give:

CONJECTURE (Lang's conjecture, strong form). *Let X be a smooth projective variety of general type over k. Then for any finite extension k'/k, the set $(X \setminus S(X))(k')$ is finite.*

Here is a simple consequence:

PROPOSITION 1.5.1. *Assume Lang's conjecture holds true. Let X be a smooth projective variety over a number field k. Assume there is a dominant rational map $X \to Z$, such that Z is a positive dimensional variety of general type (i.e., $\dim L(X) > 0$). Then $X(k)$ is not Zariski-dense in X.*

1.6. Uniformity of rational points. Lang's conjecture can be investigated whenever one has a variety of general type around. By considering certain subvarieties of the moduli space $\mathcal{M}_{g,n}$ of curves of genus g with n distinct points on them, rather surprising and inspiring implications on the arithmetic of curves arise. This is the subject of the work [**CHM97**] of L. Caporaso, J. Harris and B. Mazur. Here are their key results:

THEOREM 1.6.1. *Assume that the weak Lang's conjecture holds true. Let k be as above, and let $g > 1$ be an integer. Then there exists an integer $N(k, g)$ such that for every algebraic curve C of genus g over k we have*

$$\#C(k) \leq N(k, g).$$

THEOREM 1.6.2. *Assume that the strong Lang's conjecture holds true. Let $g > 1$ be an integer. Then there exists an integer $N(g)$ such that for every finitely generated field k there are, up to isomorphism, only finitely many algebraic curves C of genus g over k with $\#C(k) > N(g)$.*

Further results along these lines, involving higher dimensional varieties and involving stronger results on curves can be found in [**Has96**], [**Abr95**], [**Pac97**], [**AV96**], [**Abr97**]. For instance, P. Pacelli's result in [**Pac97**] says that the number $N(k, g)$ can be replaced for number fields by $N(d, g)$, where $d = [k : \mathbb{Q}]$.

The reader may decide whether this shows the great power of the conjectures or their unlikelihood. I prefer to be agnostic and rely on the conjectures for inspiration.

1.7. The search for an arithmetic dichotomy. As demonstrated in table 0.1.7, potential density of rational points on curves is dictated by geometry. Lang's conjecture carves out a class of higher dimensional varieties for which rational points are, conjecturally, not potentially dense. Can this be extended to a dichotomy as we have for curves?

One can naturally wonder—is the Kodaira dimension itself enough for determining potential density of points? Or else, maybe just the nonexistence of a map to a positive dimensional variety of general type?

1.7.1. *Rational points on surfaces.* The following table, which I copied from a lecture of L. Caporaso [**Cap**], describes what is known about surfaces.

CAPORASO'S TABLE: RATIONAL POINTS ON SURFACES

Kodaira dimension	$X(k)$ potentially dense	$X(k)$ never dense
$\kappa = -\infty$	$\mathbb{P}^2$	$\mathbb{P}^1 \times C$ $(g(C) \geq 2)$
$\kappa = 0$	$E \times E$, many others	**none known**
$\kappa = 1$	many examples	$E \times C$ $(g(C) \geq 2)$
$\kappa = 2$	**none known**	many examples

The bottom row is the subject of Lang's conjecture, and the $\kappa = 0$ row is the subject of Conjecture 1.4.2.

1.7.2. *Failure of the dichotomy using $\kappa(X)$.* The first clear lesson we learn from this is, as Caporaso aptly put it in her lecture, that

Diophantine geometry is not governed by the Kodaira dimension.

On the top row we see that clearly: on a ruled surface over a curve of genus ≥ 2, rational points can never be dense by Faltings's theorem. So it behaves very differently from a rational surface.

Even if one insists on working with varieties of non-negative Kodaira dimension, the $\kappa = 1$ row gives us trouble.

EXERCISE. Take a Lefschetz pencil of cubic curves in $\mathbb{P}^2$, parametrized by t, and assume that it has two sections s_1, s_2 whose difference is not torsion on the generic fiber. We use s_1 as the origin.

(1) Show that the dualizing sheaf of the total space S is $\mathcal{O}_S(-[F])$, where F is a fiber.
(2) Show that the relative dualizing sheaf is $\mathcal{O}_S([F])$. Take the base change $t = s^3$. We still have two sections, still denoted s_1, s_2, such that the difference is not torsion. We view s_1 as origin.

    Show that the relative dualizing sheaf of the new surface X is $\mathcal{O}_X(3[F])$ and its dualizing sheaf is $\mathcal{O}_X([F])$. Conclude that the resulting surface X has Kodaira dimension 1.
(3) For any rational point p on $\mathbb{P}^1$ where the section s_2 of $X \to \mathbb{P}^1$ is not torsion, the fiber has a dense set of rational points.

    In characteristic 0 it can be shown that the set of such points is dense. For instance, by Mazur's theorem the rational torsion points have order at most 12, and therefore they lie on finitely many points of intersection of s_2 with the locus of torsion points of order ≤ 12.
(4) Conclude that X has a dense set of rational points.

1.7.3. *Failure of the dichotomy using the Lang map.* The examples given above still allow for a possible dichotomy based on the existence of a nontrivial map to a variety of general type. But the following example, which fits in the right column on row $\kappa = 1$, shows this doesn't work either. The example is due to Colliot-Thélène, Skorobogatov and Swinnerton-Dyer [**CTSSD97**].

EXAMPLE. Let C be a curve with an involution $\phi : C \to C$, such that the quotient is rational. Consider an elliptic curve E with a 2-torsion point a, and consider the fixed-point free action of $\mathbb{Z}/2\mathbb{Z}$ on $Y = E \times C$ given by

$$(x, y) \mapsto (x + a, \phi(y)).$$

Let the quotient of Y by the involution be X. Then $L(X)$ is trivial, though rational points on X are not potentially dense by Chevalley-Weil and Faltings.

In the next section we address a conjectural approach to a dichotomy—due to F. Campana—which has a chance to work.

1.8. Logarithmic Kodaira dimension and the Lang-Vojta conjectures. We now briefly turn our attention to open varieties, following the lesson in section 0.2.11.

Let $\overline{X}$ be a smooth projective variety, D a reduced normal crossings divisor. We can consider the quasiprojective variety $X = \overline{X} \smallsetminus D$.

The logarithmic Kodaira dimension of X is defined to be the Iitaka dimension $\kappa(X) := \kappa(\overline{X}, K_{\overline{X}} + D)$. We say that X is of *logarithmic general type* if $\kappa(X) = \dim X$.

It can be easily shown that $\kappa(X)$ is independent of the completion $X \subset \overline{X}$, as long as $\overline{X}$ is smooth and D is a normal crossings divisor. More invariance properties can be discussed, but will take us too far afield.

Now to arithmetic: suppose $\mathcal{X}$ is a model of X over $\mathcal{O}_{k,S}$. We can consider integral points $\mathcal{X}(\mathcal{O}_{L,S_L})$ for any finite extension L/k and enlargement S_L of the set of places over S.

The Lang-Vojta conjecture is the following:

CONJECTURE 1.8.1. *If X is of logarithmic general type, then integral points are not potentially dense on X, i.e., $\mathcal{X}(\mathcal{O}_{L,S_L})$ is not Zariski-dense for any L, S_L.*

1.8.2. In case $X = \overline{X}$ is already projective, the Lang-Vojta conjecture reduces to Lang's conjecture: X is simply a variety of general type, integral points on X are the same as rational points, and Lang's conjecture asserts that $X(k)$ is not Zariski-dense in X.

1.8.3. The Lang-Vojta conjecture turns out to be a particular case of a more precise and more refined conjecture of Vojta, which will be discussed in a later section.

2. Campana's program

For this section one important road sign is

THIS SITE IS UNDER CONSTRUCTION
DANGER! HEAVY EQUIPMENT CROSSING

A quick search on the web shows close to the top a number of web sites deriding the idea of "site under construction". Evidently these people have never engaged in research!

2.0.1. Campana's program is a new method of breaking algebraic varieties into "pieces" which builds upon Iitaka's program, but, by using a particular structure on varieties which I will call "Campana constellations" enables one to get closer to a classification which is compatible with arithmetic properties. There is in fact an underlying more refined structure which I call "firmament" for the Campana constellation, which might be the more fundamental structure to study. I believe it truly does say something about rational points.

2.0.2. The term "constellation" is inspired by Aluffi's celestial [**Alu07**], which is in turn inspired by Hironaka.

Campana used the term "orbifold", in analogy to orbifolds used in geometry, but the analogy breaks down very early on. A suggested replacement "orbifold pair" still does not make me too happy. Also, "Campana pair" is a term which Campana himself is not comfortable using, nor could he shorten it to just "pair", which is insufficient. I was told by Campana that he would be happy to use "constellations" if the term catches on.

2.1. One-dimensional Campana constellations.

2.1.1. *The two key examples: elliptic surfaces.* Let us inspect again Caporaso's table of surfaces, and concentrate on $\kappa = 1$. We have in 1.7.2 and 1.7.3 two examples, say $S_1 \to \mathbb{P}^1$ and $S_2 \to \mathbb{P}^1$ of elliptic surfaces of Kodaira dimension 1 fibered over $\mathbb{P}^1$. But their arithmetic behavior is very different.

Campana asked the question: is there an underlying structure on the base $\mathbb{P}^1$ from which we can deduce this difference of behavior?

The key point is that the example in 1.7.3 has $2g + 2$ double fibers lying over a divisor $D \subset \mathbb{P}^1$. This means that the elliptic surface $S_2 \to \mathbb{P}^1$ can be lifted to $S_2 \to \mathcal{P}$, where $\mathcal{P}$ is the orbifold structure $\mathbb{P}^1(\sqrt{D})$ on $\mathbb{P}^1$ obtained by taking

the square root of D. Following the ideas of Darmon and Granville in [**DG95**], one should consider the canonical divisor class $K_{\mathcal{P}}$ of $\mathcal{P}$, viewed as a divisor with rational coefficients on $\mathbb{P}^1$, namely $K_{\mathbb{P}^1} + (1 - 1/2)D$. In general, when one has an m-fold fiber over a divisor D, one wants to take D with coefficient $(1 - 1/m)$.

Darmon and Granville prove, using Chevalley-Weil and Faltings, that such an orbifold $\mathcal{P}$ has potentially dense set of integral points if and only if the Kodaira dimension $\kappa(\mathcal{P}) = \kappa(\mathcal{P}, K_{\mathcal{P}}) < 1$. And the image of a rational point on S_2 is an integral point on $\mathcal{P}$. This fully explains our example: since integral points on $\mathcal{P} = \mathbb{P}^1(\sqrt{D})$ are not Zariski-dense, and since rational points on S_2 map to integral points on $\mathcal{P}$, rational points on S_2 are not dense.

2.1.2. *The multiplicity divisor.* What should we declare the structure to be when we have a fiber that looks like $x^2y^3 = 0$, i.e. has two components of multiplicities 2 and 3? Here Campana departs from the classical orbifold picture: the highest classical orbifold to which the fibration lifts has no new structure lying under such a fiber, because $\gcd(2, 3) = 1$. Campana makes a key observation that a rich and interesting classification theory arises if one instead considers $\min(2, 3) = 2$ as the basis of the structure.

Definition 2.1.3 (Campana). Consider a dominant morphism $f : X \to Y$ with X, Y smooth and $\dim Y = 1$. Define a divisor with rational coefficients $\Delta_f = \sum \delta_p p$ on Y as follows: assume the divisor f^*p on X decomposes as $f^*p = \sum m_i C_i$, where C_i are the distinct irreducible components of the fiber taken with reduced structure. Then set

$$\delta_p \;=\; 1 - \frac{1}{m_p}, \qquad \text{where} \qquad m_p \;=\; \min_i m_i.$$

Definition 2.1.4 (Campana).

(1) A Campana constellation curve (Y/Δ) is a pair consisting of a curve Y along with a divisor $\Delta = \sum \delta_p p$ with rational coefficients, where each δ_p is of the form $\delta_p = 1 - 1/m_p$ for some integer m_p.
(2) The Campana constellation base of $f : X \to Y$ is the structure pair consisting of Y with the divisor Δ_f defined above, denoted (Y/Δ_f).

The word used by Campana is *orbifold*, but as I have argued, the analogy with orbifolds is shattered in this very definition.

The suggested terminology "constellation" will become better justified and much more laden with meaning when we consider Y of higher dimension.

Campana's definition deliberately does not distinguish between the structure coming from a fiber of type $x^2 = 0$ and one of type $x^2y^3 = 0$. We will see later a way to resurrect the difference to some extent using the notion of *firmament*, from which a Campana constellation hangs.

Definition 2.1.5 (Campana). The Kodaira dimension of a Campana constellation curve (Y/Δ) is defined as the following Iitaka dimension:

$$\kappa\,(\,(Y/\Delta)\,) = \kappa(Y, K_Y + \Delta).$$

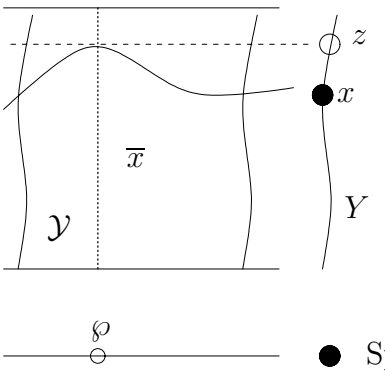
We say that (Y/Δ) is of general type if it has Kodaira dimension 1. We say that it is *special* if it is not of general type.

Exercise 2.1.6. Classify *special* Campana constellation curves over $\mathbb{C}$. See [**Cam05**] for a detailed discussion.

2.1.7. *Models and integral points.* Now to arithmetic. As we learned in Lesson 0.2.11, when dealing with a variety with a structure given by a divisor, we need to speak about *integral points* on an *integral model* of the structure. Thus let $\mathcal{Y}$ be an integral model of Y, proper over $\mathcal{O}_{k,S}$, and denote by $\widetilde{\Delta}$ the closure of Δ. As above we assume that Δ is the union of integral points of Y, denoted z_i, and for simplicity let us assume that they are disjoint (we can always achieve this by enlarging S). It turns out that there is more than one natural notion to consider in our theory - soft and firm. The firm notion will be introduced when firmaments are considered.

Definition. A k-rational point x on Y, considered as an integral point of $\mathcal{Y}$, is said to be a *soft S-integral point on* $(\mathcal{Y}/\widetilde{\Delta})$ if for any integral point z in $\widetilde{\Delta}$, and any nonzero prime $\wp \subset \mathcal{O}_{k,S}$ such that the reductions coincide, $x_\wp = z_\wp$, we have

$$\mathrm{mult}_\wp(x \cap z) \geq m_p.$$



A key property of this definition is:

Proposition 2.1.8. *Assume $f : X \to Y$ extends to a good model $\tilde{f} : \mathcal{X} \to \mathcal{Y}$. Then the image of a rational point on X is a soft S-integral point on $(\mathcal{Y}/\widetilde{\Delta}_f)$.*

So rational points on X can be investigated using integral points on a model of Y. This makes the following very much relevant:

Conjecture 2.1.9 (Campana). *Suppose the Campana constellation curve (Y/Δ) is of general type. Then the set of soft S-integral point on any model $\mathcal{Y}$ is not Zariski-dense.*

This conjecture is not likely to follow readily from Faltings's theorem, as the following example suggests.

Example 2.1.10. Let $n \geq 4$ be an integer. Let $Y \simeq \mathbb{P}^1$ and Δ the divisor supported at $0, 1$ and ∞ with all multiplicities equal to $(n-1)/n$. Then (Y/Δ) is of general type.

Using the same as a model over $\operatorname{Spec}\mathbb{Z}$, we see that a point y on Y is a soft integral point on (Y/Δ) if at every prime where y reduces to $0, 1$ or ∞, the multiplicity of this reduction is at least n.

Considering a triple a, b, c of relatively prime integers with $a^N + b^N = c^N$, the point $(a^N : c^N)$ on Y is a soft integral point as soon as $N \geq n$.

It follows that Campana's conjecture 2.1.9 implies asymptotic Fermat over any number field.

It also seems that the conjecture does not follow readily from any of the methods surrounding Wiles's proof of Fermat. As we'll see in the last section, the conjecture does follow from the *abc* conjecture (which implies asymptotic Fermat). In particular we have the following theorem in the function field case.

THEOREM 2.1.11 (Campana). *Let B be a complex algebraic curve, and K its function field, and let $S \subset B$ be a finite set of closed points. Let (Y/Δ) be a Campana constellation curve of general type defined over the function field K. Then the set of non-constant soft S-integral points on any model $\mathcal{Y} \to B$ is not Zariski-dense.*

2.1.12. *Some examples.*

(1) Consider $f : \mathbb{A}^2 \to \mathbb{A}^1$ given by $t = x^2$. The constellation base has $\Delta = 1/2\,(0)$, where (0) is the origin on $\mathbb{A}^1$. Sections of $\mathcal{O}_Y(K + \Delta)$ are generated by dt, sections of $\mathcal{O}_Y\,(2(K + \Delta))$ by $(dt)^2/t$, and sections of $\mathcal{O}_Y\,(3(K + \Delta))$ by $(dt)^3/t$.
(2) The same structure occurs for $f : \mathbb{A}^2 \to \mathbb{A}^1$ given by $t = x^2y^3$ and $f : \mathbb{A}^2 \to \mathbb{A}^1$ given by $t = x^2y^2$.
(3) For $f : \mathbb{A}^2 \to \mathbb{A}^1$ given by $t = x^2y$ the constellation base is trivial.
(4) for $f : \mathbb{A}^2 \to \mathbb{A}^1$ given by $t = x^3y^4$, we get $\Delta = 2/3\,(0)$. Again sections of $\mathcal{O}_Y\,(2(K + \Delta))$ are generated by $(dt)^2/t$, but sections of $\mathcal{O}_Y\,(3(K + \Delta))$ are generated by $(dt)^3/t^2$.

2.2. Higher dimensional Campana constellations. We turn now to the analogous situation of $f : X \to Y$ with higher dimensional Y.

One seeks to define objects, say Campana constellations $(Y/\boldsymbol{\Delta})$, in analogy to the case of curves, which in some sense should help us understand the geometry and arithmetic of plain varieties mapping to them.

Ideally, these objects should form a category extending the category of varieties, at least with some interesting class of morphisms. Ideally these objects should have a good notion of differential forms which fits into the standard theory of birational geometry, for instance having well-behaved Kodaira dimension. Ideally there should be a notion of integral points on $(Y/\boldsymbol{\Delta})$ which says something about rational points of a "plain" variety X whenever X maps to $(Y/\boldsymbol{\Delta})$.

The theory we describe in this section, which is due to Campana in all but some details, relies on divisorial data. We will describe a category of objects, called Campana constellations, which at the moment only allows dominant morphisms. This means that we do not have a satisfactory description of integral points, since integral points are sections, and sections are not dominant morphisms. The theory of firmaments aims at resolving this problem.

Unfortunately, points on Y are no longer divisors. And divisors on Y are not quite sufficient to describe codimension > 1 behavior. Campana resolves this by considering all birational models of Y separately. This brings him to define various invariants, such as Kodaira dimension, depending on a morphism $X \to Y$ rather than of the structure $(Y/\boldsymbol{\Delta})$ itself. I prefer to put all this data together using the notion of a b-divisor, introduced by Shokurov [**Sho03**], based on ideas by Zariski [**Zar44**]. I was also inspired by Aluffi's [**Alu07**]. The main advantage is that all invariants will be defined directly on the level of $(Y/\boldsymbol{\Delta})$. This structure has the disadvantage that it is not obviously computable in finite or combinatorial terms. It turns out that it is—this again will be addressed using firmaments.

DEFINITION 2.2.1. Let k be a field and Y a variety over k. A *rank 1 discrete valuation* on the function field $\mathcal{K} = \mathcal{K}(Y)$ over k is a surjective group homomorphism $\nu : \mathcal{K}^\times \to \mathbb{Z}$, sending $k^\times$ to 0, satisfying

$$\nu(x+y) \geq \min(\nu(x), \nu(y))$$

with equality unless $\nu(x) = \nu(y)$. We define $\nu(0) = +\infty$.

The *valuation ring of* ν is defined as

$$R_\nu \;=\; \left\{ x \in \mathcal{K} \;\middle|\; \nu(x) \geq 0 \right\}.$$

Denote $Y_\nu = \operatorname{Spec} R_\nu$, and its unique closed point by s_ν.

A rank 1 discrete valuation ν is *divisorial* if there is a birational model Y' of Y and an irreducible divisor $D' \subset Y'$ such that for all nonzero $x \in \mathcal{K}(Y) = \mathcal{K}(Y')$ we have

$$\nu(x) \;=\; \operatorname{mult}_{D'} x.$$

In this case we say ν has *divisorial center* D' in Y'.

DEFINITION 2.2.2. A *b-divisor* $\mathbf{\Delta}$ on Y is an expression of the form

$$\mathbf{\Delta} \;=\; \sum_\nu c_\nu \cdot \nu,$$

a possibly infinite sum over divisorial valuations of $\mathcal{K}(Y)$ with rational coefficients, which satisfies the following finiteness condition:

- for each birational model Y' there are only finitely many ν with divisorial center on Y' having $c_\nu \neq 0$.

A b-divisor is of *orbifold type* if for each ν there is a positive integer m_ν such that $c_\nu = 1 - 1/m_\nu$.

Before we continue, here is an analogue of the strict transform of a divisor:

DEFINITION 2.2.3. Let Y be a variety, X a reduced scheme, and let $f : X \to Y$ be a morphism. Consider an integral scheme Y' with generic point η, a dominant morphism $Y' \to Y$, and the pullback $X \times_Y Y' \to Y'$. The *main part* $\widetilde{X \times_Y Y'}$ of $X \times_Y Y'$ is the closure of the generic fiber $X \times_Y \eta$ inside $X \times_Y Y'$.

Here is a higher dimensional analogue of the divisor underlying the constellation curve of a morphism $X \to Y$:

DEFINITION 2.2.4. Let Y be a variety, X a reduced scheme, and let $f : X \to Y$ be a morphism, surjective on each irreducible component of X. For each divisorial valuation ν on $\mathcal{K}(Y)$ consider $f' : X'_\nu \to Y_\nu$, where X'_ν is a desingularization of the main part of the pullback $X \times_Y Y_\nu$. Write $f'^* s_\nu = \sum m_i C_i$. Define

$$\delta_\nu \;=\; 1 - \frac{1}{m_\nu} \quad \text{with} \quad m_\nu = \min_i m_i.$$

The *Campana b-divisor* on Y associated to a dominant map $f : X \to Y$ is defined to be the b-divisor

$$\mathbf{\Delta}_f = \sum \delta_\nu \nu.$$

EXERCISE 2.2.5. The definition is independent of the choice of desingularization X'_ν.

This makes the b-divisor $\mathbf{\Delta}_f$ invariant under proper birational transformations on X and Y. In particular the notion makes sense for a dominant rational map f.

DEFINITION 2.2.6. (1) A *Campana constellation* $(Y/\mathbf{\Delta})$ consists of a variety Y with a b-divisor $\mathbf{\Delta}$ such that, locally in the étale topology on Y, there is $f : X \to Y$ with $\mathbf{\Delta} = \mathbf{\Delta}_f$.

(2) The Campana constellation base of a morphism $X \to Y$ as above is $(Y/\mathbf{\Delta}_f)$.

(3) The trivial constellation on Y is given by the zero b-divisor.

(4) For each birational model Y', define the Y'-divisorial part of $\mathbf{\Delta}$:

$$\Delta_{Y'} = \sum_{\nu \text{ with divisorial support on } Y'} \delta_\nu \nu.$$

The definition of a constellation feels a bit unsatisfactory because it requires, at least locally, the existence of a morphism f. But using the notion of firmament, especially toroidal firmament, we will make this structure more combinatorial, in such a way that the existence of f is automatic.

2.2.7. Here's why I like the word "constellation": think of a divisorial valuation ν as a sort of "generalized point" on Y. Putting $\delta_\nu > 0$ suggests viewing a "star" at that point. Replacing Y by higher and higher models Y' is analogous to using stronger and stronger telescopes to view farther stars deeper into space. The picture I have in my mind is somewhat reminiscent of the astrological meaning of "constellation", not as just one group of stars, but rather as the arrangement of the entire heavens at the time the "baby" $X \to Y$ is born. But hopefully it is better grounded in reality.

We now consider morphisms. For constellations we work only with dominant morphisms.

DEFINITION 2.2.8. (1) Let $(X/\mathbf{\Delta}_X)$ be a Campana constellation, and $f : X \to Y$ a *proper* dominant morphism. The constellation base $(Y, \mathbf{\Delta}_{f,\mathbf{\Delta}_X})$ is defined as follows: for each divisorial valuation ν of Y and each divisorial valuation μ of X with center D dominating the center E of ν, let

$$m_{\mu/\nu} = m_\mu \cdot \mathrm{mult}_D(f^*E).$$

Define

$$m_\nu = \min_{\mu/\nu} m_{\mu/\nu} \quad \text{and} \quad \delta_\nu = 1 - \frac{1}{m_\nu}.$$

Then set as before

$$\mathbf{\Delta}_{f,\mathbf{\Delta}_X} = \sum_\nu \delta_\nu \nu.$$

(2) Let $(X/\mathbf{\Delta}_X)$ and $(Y/\mathbf{\Delta}_Y)$ be Campana constellations and $f : X \to Y$ a dominant morphism. Then f is said to be a *constellation morphism* if for every divisorial valuation ν on Y and any μ/ν we have $m_\nu \le m_{\mu/\nu}$, where as above $m_{\mu/\nu} = m_\mu \cdot \mathrm{mult}_D(f^*E)$. When f is proper this just means $\mathbf{\Delta}_Y \le \mathbf{\Delta}_{f,\mathbf{\Delta}_X}$.

Now to differential forms:

DEFINITION 2.2.9. A rational m-canonical differential ω on Y is said to be *regular* on $(Y/\mathbf{\Delta})$ if for every divisorial valuation ν on $\mathcal{K}(Y)$, the polar multiplicity of ω at ν satisfies

$$(\omega)_{\infty,\nu} \le m\delta_\nu.$$

In other words, ω is a section of $\mathcal{O}_{Y'}(m(K_{Y'} + \Delta_{Y'}))$ on every birational model Y'.

The Kodaira dimension $\kappa(\,(Y/\mathbf{\Delta})\,)$ is defined using the ring of regular m-canonical differentials on $(Y/\mathbf{\Delta})$.

EXERCISE 2.2.10. This is a birational invariant: if Y and Y' are proper and have the same function field, then $\kappa(\,(Y/\mathbf{\Delta})\,) = \kappa(\,(Y'/\mathbf{\Delta})\,)$.

THEOREM 2.2.11 (Campana [**Cam04**] Section 1.3). *There is a birational model Y' with $\Delta_{Y'}$ a normal crossings divisor such that*

$$\kappa(\,(Y/\mathbf{\Delta})\,) = \kappa(Y', K_{Y'} + \Delta_{Y'}),$$

and moreover the algebra of regular pluricanonical differentials on $(Y/\mathbf{\Delta})$ agrees with the algebra of sections $\bigoplus_{m\geq 0} H^0(Y', \mathcal{O}_{Y'}(m(K_{Y'} + \Delta_{Y'})))$.

Campana calls such a model *admissible*. This is proven using Bogomolov sheaves, an important notion which is a bit far afield for the present discussion. The formalism of firmaments, especially toroidal firmaments, allows one to give a combinatorial proof of this result.

We remark that this theorem means that the new and ground-breaking finite generation theorem of [**BCHM06**] applies, so the algebra of regular pluricanonical differentials on $(Y/\mathbf{\Delta})$ is finitely generated.

It is not difficult to see that any birational model lying over an admissible model is also admissible.

DEFINITION 2.2.12. A Campana constellation $(Y/\mathbf{\Delta})$ is said to be *of general type* if $\kappa(\,(Y/\mathbf{\Delta})\,) = \dim Y$.

A Campana constellation $(X/\mathbf{\Delta})$ is said to be *special* if there is no dominant morphism $(X/\mathbf{\Delta}) \to (Y/\mathbf{\Delta}')$ where $(Y/\mathbf{\Delta}')$ is of general type.

DEFINITION 2.2.13. Let $f : X \to Y$ be a dominant morphism of varieties and $(X/\mathbf{\Delta})$ a Campana constellation, with $\mathbf{\Delta} = \sum \delta_\nu \nu$. The *generic fiber* of $f : (X/\mathbf{\Delta}) \to Y$ is the Campana constellation $(X_\eta, \mathbf{\Delta}_\eta)$, where X_η is the generic fiber of $f : X \to Y$, and

$$\mathbf{\Delta}_\eta = \sum_{\nu|_{f^*\mathcal{K}(Y)^\times}=0} \delta_\nu \nu,$$

namely the part of the b-divisor $\mathbf{\Delta}$ supported on the generic fiber.

DEFINITION 2.2.14.

(1) Given a Campana constellation $(X/\mathbf{\Delta}_X)$, a dominant morphism $f : X \to Y$ is *special* if its generic fiber is special.

(1') In particular, considering X with trivial constellation, a dominant morphism $f : X \to Y$ is *special* if its generic fiber is special as a variety with trivial constellation.

(2) Given a Campana constellation $(X/\mathbf{\Delta}_X)$, a proper dominant morphism $f : X \to Y$ is said to have *general type base* if $(Y/\mathbf{\Delta}_{f,\mathbf{\Delta}_X})$ is of general type.

(2') In particular, considering X with trivial constellation, a proper dominant morphism $f : X \to Y$ is said to have *general type base* if $(Y/\mathbf{\Delta}_f)$ is of general type.

Here is the main classification theorem of Campana:

THEOREM 2.2.15 (Campana). *Let $(X/\mathbf{\Delta}_X)$ be a Campana constellation on a projective variety X. There exists a dominant rational map $c : X \dashrightarrow C(X)$, unique up to birational equivalence, such that*

(1) the map c has special generic fiber, and
(2) the Campana constellation base $(C(X)/\mathbf{\Delta}_{c,\mathbf{\Delta}_X})$ is of general type.

This map is final for (1) and initial for (2).

This is the Campana core map of $(X/\mathbf{\Delta}_X)$, the constellation $(C(X)/\mathbf{\Delta}_{c,\mathbf{\Delta}_X})$ being the core of $(X/\mathbf{\Delta}_X)$. The key case is when X has the trivial constellation, and then $c : X \dashrightarrow (C(X)/\mathbf{\Delta}_c)$ is the Campana core map of X and $(C(X)/\mathbf{\Delta}_c)$ the core of X.

2.2.16. *More examples of constellation bases.* The following is a collection of examples which I find useful to keep in mind. The stated rules for the constellation bases are explained below in 2.2.17.

(1) Consider $f : \mathbb{A}^2 \to \mathbb{A}^2$ given by $s = x^2; t = y$. We want to describe the constellation base. Clearly on $Y = \mathbb{A}^2$, the divisor $\Delta_Y = 1/2\,(s = 0)$. But what should the multiplicity be for a divisor on some blowup of Y?

    The point is that $X \to Y$ is *toric*, and $\mathbf{\Delta}$ can be described using toric geometry. Indeed, the multiplicity at a divisorial valuation ν is precisely dictated by the value of $\nu(s)$, with a simple rule: if $\nu(s)$ is even, we have $m_\nu = 1$ so $\delta_\nu = 0$, otherwise $m_\nu = 2$ and $\delta_\nu = 1/2$. Regular pluricanonical differentials are generated by $(ds \wedge dt)^2/s$.

(2) Consider now $f : \mathbb{A}^2 \to \mathbb{A}^2$ given by $s = x^2; t = y^2$. The rule this time: $m_\nu = 1$ and $\delta_\nu = 0$ if and only if both $\nu(s)$ and $\nu(t)$ are even, otherwise $m_\nu = 2$ and $\delta_\nu = 1/2$. Regular pluricanonical differentials are generated by $(ds \wedge dt)^2/st$.

(3) $f : \mathbb{A}^2 \sqcup \mathbb{A}^2 \to \mathbb{A}^2$ given by $s = x_1^2; t = y_1$ and $s = x_2; t = y_2^2$ The rule this time: $m_\nu = 1$ and $\delta_\nu = 0$ if and only if either $\nu(s)$ or $\nu(t)$ is even, otherwise $m_\nu = 2$ and $\delta_\nu = 1/2$. Regular pluricanonical differentials are generated by $ds \wedge dt$.

(4) $f : X \to \mathbb{A}^2$ given by the singular cover $\operatorname{Spec} \mathbb{C}[s, t, \sqrt{st}]$. The rule: $m_\nu = 1$ and $\delta_\nu = 0$ if and only if either $\nu(s) + \nu(t)$ is even, otherwise $m_\nu = 2$ and $\delta_\nu = 1/2$. Regular pluricanonical differentials are generated by $(ds \wedge dt)^2/st$.

(5) $f : \mathbb{A}^3 \to \mathbb{A}^2$ given by $s = x^2y^3; t = z$. The rule: $m_\nu = 1$ and $\delta_\nu = 0$ if and only if either $\nu(s) = 0$ or $\nu(s) \geq 2$, otherwise $m_\nu = 2$ and $\delta_\nu = 1/2$. Regular pluricanonical differentials are generated by $(ds \wedge dt)^2/s$.

2.2.17. Where does the rule come from? When we have a toric map of affine toric varieties, we have a map of cones $f_\sigma : \sigma_X \to \sigma_Y$. Inside these cones we have lattices N_X and N_Y - I am considering only the part of the lattice lying in the closed cone, so it is only a monoid, not a group. The map f_σ maps N_X into a sub-monoid $\Gamma \subset N_Y$. Each rank-1 discrete valuation ν of Y has a corresponding point $n_\nu \in N_Y$, calculated by the value of ν on the monomials of Y: in the case of $\mathbb{A}^2$ this point is simply $(\nu(s), \nu(t))$. The rule is: m_ν is the minimal positive integer such that

$$m_\nu \cdot n_\nu \in \Gamma.$$

These toric examples form the basis for defining firmaments later on.

2.2.18. *Rational points and the question of integral points.* Campana made the following bold conjecture:

CONJECTURE 2.2.19 (Campana). *Let X/k be a variety over a number field. Then rational points are potentially dense on X if and only if X is special, i.e., if and only if the core of X is a point.*

It is natural to seek a good definition of integral points on a Campana constellation and translate the non-special case of the conjecture above to a conjecture on integral points on Campana constellations of general type.

The following definition covers part of the ground. It seems natural, yet it is not satisfactory as it is quite restrictive. It is also not clear how these points behave in morphisms. We'll be able to go a bit further with firmaments.

DEFINITION 2.2.20. Let $(Y/\mathbf{\Delta})$ be a Campana constellation over a number field k, and assume it is admissible as in Theorem 2.2.11 and the discussion therein. Write as usual $\Delta_Y = \sum(1-1/m_i)\Delta_i$ for the part of $\mathbf{\Delta}$ with divisorial support on Y. Assume given a model $(\mathcal{Y}, \widetilde{\Delta}_Y)$ of (Y, Δ_Y) over $\mathcal{O}_{k,S}$, such that $\mathcal{Y}$ is smooth and $\widetilde{\Delta}_Y$ a horizontal normal crossings divisor. Write $Y_0 = Y \smallsetminus \Delta_Y$.

Consider $y \in Y_0(K)$. We say that y is a *soft S-integral point on* $(Y/\mathbf{\Delta})$ if for any prime $\wp$ where the Zariski closure $\bar{y}$ of y reduces to $\widetilde{\Delta}$ we have

$$\sum \frac{1}{m_i} \operatorname{mult}_\wp \widetilde{\Delta}_i \cdot \bar{y} \quad \geq \quad 1.$$

2.3. Bogomolov vs. Campana: some remarks about their philosophies. Let us take a step back and reconsider what we are doing. After all, we are trying to learn something about the geometry of a variety X from the data of dominant morphisms $X \to Y$ it admits to other varieties. And somehow the effect of such a map is encoded not only in the geometry of Y but in some extra structure.

Campana's approach involves introducing a new category of objects, which I call Campana constellations. For any dominant $f : X \to Y$, this maps leaves an indelible mark, namely a constellation, on the *target* Y, and you learn about X by studying the constellations onto which it maps.

There is an approach which is technically closely related but philosophically diametrically opposed, due to Bogomolov. Bogomolov suggests that since our object of study is X, we need to look for the indelible mark $f : X \to Y$ leaves *on X itself.* Bogomolov proposes to use what has come to be called a *Bogomolov sheaf*: let $d = \dim Y$ and consider the saturated image $\mathcal{F}_m$ of $f^*\omega_Y^m$ in $\operatorname{Sym}^m \Omega_X^d$. These form a sheaf of algebras $\bigoplus_{m\geq 0} \mathcal{F}_m$, and it is said to be of general type if the algebra of sections has dimension $d+1$. Bogomolov suggests that such sheaves should have an important role in the arithmetic and geometric properties of X.

Even if one prefers Bogomolov's approach, I think the achievement of Campana's Theorem 2.2.15 is remarkable and cannot be ignored. For example, it seems that the preprint [**Lu02**] attempted to develop a theory based entirely on Bogomolov sheaves, but the author could not resist veering towards statements such as Theorem 2.2.15.

So let us take a closer look at what we have been doing with Campana's approach.

In essence, what we are trying to capture is a structure on Y that measures a sort of equivalence class of dominant maps $X \to Y$. In some sense, the structure should measure to what extent the map $X \to Y$ has a section, perhaps locally and up to proper birational maps, or perhaps on a suitable choice of discrete valuation rings. There are some reasonable properties this should satisfy:

- It should be local on Y.
- It should be invariant under modifications of X.
- It should behave well under birational modifications of Y.
- There should be a good notion of morphism of such structures, at least on the level of dominant maps.

So far, our notion of constellation satisfies all of the above. We defined constellations in terms of divisorial valuations, which live on the function field of Y, and automatically behave well under birational maps. In fact I modified Campana's original definition, which relied on the divisor Δ_Y, by introducing $\boldsymbol{\Delta}$ precisely for this purpose. One seems to lose in the category of computability, though not so much if one can characterise and find admissible models. The definition was made precisely to guarantee that if S is the spectrum of a complete discrete valuation ring with algebraically closed residue field, and $S \to Y$ is *dominant*, then the map lifts to $S \to (Y/\boldsymbol{\Delta})$ if, and only if, it lifts to $S \to X$.

But consider the following desirable properties, which are not yet achieved:

- The structure should be invariant under smooth maps on X.
- In some sense it should be recovered from an open covering of X.
- It should be computable.
- There should be a notion of morphisms, good enough to work with non-dominant maps and integral points.

It seems that Campana constellations are wonderfully suited for purposes of birational classification. Still they seem to lack some subtle information necessary to have these last properties, such as good definitions of non-dominant morphisms and integral points—at least I have not been successful in doing this directly on constellations in a satisfactory manner. For these purposes I propose the notion of firmaments. At this point I can achieve these desired properties under extenuating circumstances, which at least enables one to state meaningful questions. It is very much possible that at the end a simpler formalism will be discovered, and the whole notion of firmaments will be redundant.

2.4. Firmaments supporting constellations and integral points. The material in this section is very much incomplete as many details are missing and many questions are yet unanswered.

2.4.1. *Firmaments: valuative definition.* Let me first define the notion of firmaments in a way that seems to make things a bit more complicated than constellations, and where it is not clear that any additional desired properties are achieved.

The underlying structure is still a datum attached to every divisorial valuation ν of Y. The datum is a subset $\Gamma_\nu \subset \mathbb{N}$, and the sole requirement on each individual Γ_ν is that

- Γ_ν is the union of finitely many non-zero additive submonoids of $\mathbb{N}$.

and the structure is considered trivial if $\Gamma_\nu = \mathbb{N}$.

There is an additional requirement, namely that this should come locally from a map $X \to Y$, in the way described below.

Definition 2.4.2. Let Y be a variety, X a reduced scheme, and let $f : X \to Y$ be a morphism, surjective on each irreducible component of X. For each divisorial valuation ν on $\mathcal{K}(Y)$ consider $f' : X'_\nu \to Y_\nu$, where X'_ν is a normal-crossings desingularization of the main part of the pullback $X \times_Y Y_\nu$. Write F_ν for the fiber of X'_ν

over s_ν. For each point $x \in F_\nu$, assume that the components of F_ν passing through x have multiplicities $m_1, \dots, m_k$, generating a submonoid

$$\Gamma_\nu^x \;:=\; \langle m_1, \dots, m_k \rangle \;\subset \mathbb{N}.$$

Define

$$\Gamma_\nu \;=\; \bigcup_{x \in F_\nu} \Gamma_\nu^x.$$

Definition 2.4.3. A *firmament* $\mathbf{\Gamma}$ on Y is an assignment

$$\nu \;\mapsto\; \Gamma_\nu \subset \mathbb{N}$$

which, locally in the étale topology of Y, comes from a morphism $X \to Y$ as above.

This condition requiring a local description, which seems harmless, is actually crucial for the properties of firmaments.

A firmament supports a unique constellation:

Definition 2.4.4. Let $\mathbf{\Gamma}$ be a firmament. The *multiplicity* of the divisorial valuation ν is defined as $m_\nu = \min(\Gamma_\nu \smallsetminus \{0\})$. The *constellation hanging by* $\mathbf{\Gamma}$ is

$$\mathbf{\Delta_\Gamma} = \sum \left(1 - \frac{1}{m_\nu}\right) \nu.$$

2.4.5. Note that, according to the definition above, every firmament supports a unique constellation, though a constellation can be supported by more than one firmament. Depending on one's background, this might agree or disagree with the primitive cosmology of one's culture. Think of it this way: as we said before, the word "constellation" refers to the entire "heavens", visible through stronger and stronger telescopes Y'. The word "firmament" refers to an overarching solid structure supporting the heavens, but solid as it may be, it is entirely imaginary and certainly not unique.

2.4.6. *Toroidal formalism.* I wish to convince the reader that this extra structure I piled on top of constellations actually makes things better. For this purpose I need to discuss a toroidal point of view.

In fact the right foundation to use seems to be that of logarithmic structures, rather than toroidal geometry. For the longest time I stuck with toroidal geometry because the book [**Bou15**] had not been written. As [**Ogu**] and [**GR09**] are becoming available my excuses are running out, but I'll leave the translation work for the future.

Definition 2.4.7 ([**KKMSD73**], [**Kat94**], [**AK00**]). (1) A *toroidal embedding* $U \subset X$ is the data of a variety X and a dense open set U with complement a Weil divisor $D = X \smallsetminus U$, such that locally in the étale, or analytic, topology, or formally, near every point, $U \subset X$ admits an isomorphism with (a neighborhood of a point in) $T \subset V$, with T a torus and V a toric variety. (It is sometimes convenient to refer to the toroidal structure using the divisor: (X, D).)

(2) Let $U_X \subset X$ and $U_Y \subset Y$ be toroidal embeddings, then a dominant morphism $f : X \to Y$ is said to be *toroidal* if étale locally near every point of X there is a toric chart for X near x and a toric chart for Y near $f(x)$, such that on these charts f becomes a torus-equivariant morphism of toric varieties.

2.4.8. *The cone complex.* Recall that, to a toroidal embedding $U \subset X$ we can attach an integral polyhedral cone complex Σ_X, consisting of strictly convex cones, attached to each other along faces, and in each cone σ a finitely generated, unit free integral saturated monoid $N_\sigma \subset \sigma$ generating σ as a real cone.

Note that I am departing from usual terminology, by taking N_σ to be the part of the lattice lying in the cone, rather than the associated group. Note also that in [**KKMSD73**], [**Kat94**] the monoid M_σ dual to N_σ is used. While the use of M_σ is natural from the point of view of logarithmic structures, all the action with firmaments happens on its dual N_σ, so I use it instead.

2.4.9. *Valuation rings and the cone complex.* The complex Σ_X can be pieced together using the toric charts, where the picture is well known: for a toric variety V, cones correspond to toric affine opens V_σ, and the lattice N_σ is the monoid of one-parameter subgroups of the corresponding torus having a limit point in V_σ; it is dual to the lattice of effective toric Cartier divisors M_σ, which is the quotient of the lattice of regular monomials $\tilde{M}_\sigma$ by the unit monomials.

For our purposes it is convenient to recall the characterization of toric cones using valuations given in [**KKMSD73**]: let R be a discrete valuation ring with valuation ν, special point s_R and generic point η_R; let $\phi : \operatorname{Spec} R \to X$ be a morphism such that $\phi(\eta_R) \subset U$ and $\phi(s_R)$ lying in a stratum having chart $V = \operatorname{Spec} k[\tilde{M}_\sigma]$. One associates to ϕ the point n_ϕ in N_σ given by the rule

$$n(m) = \nu(\phi^* m) \quad \forall m \in M.$$

In case $R = R_\nu$ is a valuation ring of Y, I'll call this point n_ν. One can indeed give a coherent picture including the case $\phi(\eta_R) \not\subset U$, but I won't discuss this here and delay it for future treatment if one is called for. (It is however important for giving a complete picture of the category and a complete picture of the arithmetic structure.).

2.4.10. *Functoriality.* Given toroidal embeddings $U_X \subset X$ and $U_Y \subset Y$ and a morphism $f : X \to Y$ carrying U_X into U_Y (but not necessarily toroidal) the description above functorially associates a polyhedral morphism $f_\Sigma : \Sigma_X \to \Sigma_Y$ which is integral, that is, $f_\Sigma(N_\sigma) \subset N_\tau$ whenever $f_\Sigma(\sigma) \subset \tau$.

2.4.11. *Toroidalizing a morphism.* While most morphisms are not toroidal, we have the following:

THEOREM (Abramovich-Karu). *Let $f : X \to Y$ be a dominant morphism of varieties. Then there exist modifications $X' \to X$ and $Y' \to Y$ and toroidal structures $U_{X'} \subset X'$, $U_{Y'} \subset Y'$ such that the resulting rational map $f' : X' \to Y'$ is a toroidal morphism:*

$$\begin{array}{ccccc} U_{X'} & \hookrightarrow & X' & \longrightarrow & X \\ \downarrow & & \downarrow{\scriptstyle f'} & & \downarrow{\scriptstyle f} \\ U_{Y'} & \hookrightarrow & Y' & \longrightarrow & Y \end{array}$$

Furthermore, f' can be chosen flat.

We now define toroidal firmaments, and give an alternative definition of firmaments in general:

DEFINITION 2.4.12. A *toroidal firmament* on a toroidal embedding $U \subset X$ with complex Σ is a finite collection $\mathbf{\Gamma} = \{\Gamma^i_\sigma \subset N_\sigma\}$, where

- each $\Gamma^i_\sigma \subset N_\sigma$ is a finitely generate submonoid, not necessarily saturated,

- each Γ^i_σ generates the corresponding σ as a cone,
- the collection is closed under restrictions to faces, i.e., for each Γ^i_σ and each $\tau \prec \sigma$ there is j with $\Gamma^i_\sigma \cap \tau = \Gamma^j_\tau$, and
- it is irredundant, in the sense that $\Gamma^i_\sigma \not\subset \Gamma^j_\sigma$ for different i, j.

A morphism from a toroidal firmament $\mathbf{\Gamma}_X$ on a toroidal embedding $U_X \subset X$ to $\mathbf{\Gamma}_Y$ on $U_Y \subset Y$ is a morphism $f : X \to Y$ with $f(U_X) \subset U_Y$ such that for each σ in Σ_X and each i, and if $f_\Sigma(\sigma) \subset \tau$, we have $f_\Sigma(\Gamma^i_\sigma) \subset \Gamma^j_\tau$ for some j.

We say that the toroidal firmament $\mathbf{\Gamma}_X$ is *induced* by $f : X \to Y$ from $\mathbf{\Gamma}_Y$ if for each $\sigma \in \Sigma_X$ and $\tau \in \Sigma_Y$ such that $f_\Sigma(\sigma) \subset \tau$, we have $\Gamma^i_\sigma = f_\Sigma^{-1}\Gamma^i_\tau \cap N_\sigma$.

Given a proper birational equivalence $\phi : X_1 \dashrightarrow X_2$, then two toroidal firmaments $\mathbf{\Gamma}_{X_1}$ and $\mathbf{\Gamma}_{X_2}$ are said to be *equivalent* if there is a toroidal embedding $U_3 \subset X_3$, and a commutative diagram

$$\begin{array}{ccc} & X_3 & \\ {\scriptstyle f_1}\swarrow & & \searrow{\scriptstyle f_2} \\ X_1 & \overset{\phi}{- - - \to} & X_2, \end{array}$$

where f_i are modifications sending U_3 to U_i, such that the two toroidal firmaments on X_3 induced by f_i from $\mathbf{\Gamma}_{X_i}$ are identical.

A firmament on an arbitrary X is the same as an equivalence class represented by a modification $X' \to X$ with a toroidal embedding $U' \subset X'$ and a toroidal firmament $\mathbf{\Gamma}$ on $\Sigma_{X'}$. A morphism of firmaments is a morphism of varieties which becomes a morphism of toroidal firmaments on some toroidal model.

The trivial firmament is defined by $\Gamma_\sigma = N_\sigma$ for all σ in Σ.

For the discussion below one can in fact replace $\mathbf{\Gamma}$ by the union of the Γ^i_σ, but I am not convinced that makes things better.

Definition 2.4.13. (1) Let $f : X \to Y$ be a flat toroidal morphism of toroidal embeddings. The *base firmament* $\mathbf{\Gamma}_f$ associated to $X \to Y$ is defined by the images $\Gamma^\tau_\sigma = f_\Sigma(N_\tau)$ for each cone $\tau \in \Sigma_X$ over $\sigma \in \Sigma_Y$. We make this collection irredundant by taking the sub-collection of maximal elements.

(2) Let $f : X \to Y$ be a dominant morphism of varieties. The base firmament of f is represented by any $\mathbf{\Gamma}_{f'}$, where $f' : X' \to Y'$ is a flat toroidal birational model of f.

(3) If X is reducible, decomposed as $X = \cup X_i$, but $f : X_i \to Y$ is dominant for all i, we define the base firmament by the (maximal elements of) the union of all the firmaments associated to $X_i \to Y$.

2.4.14. *Equivalence of definitions.* Given a firmament in the new definition, given a toroidal model and given a divisorial valuation ν, we have a corresponding point $n_\nu \in N_\sigma$. We define

$$\Gamma_\nu = \{k \in \mathbb{N} | kn_\nu \in \Gamma_i \text{ for some } i\}.$$

This gives a firmament in the valuative definition.

Conversely, a firmament in the valuative definition has finitely many étale charts $Y_i \to Y$ where the firmament comes from $X_i \to Y_i$. One can toroidalize each $X_i \to Y_i$ simultaneously over some toroidal structure $U \subset Y$, and take the base

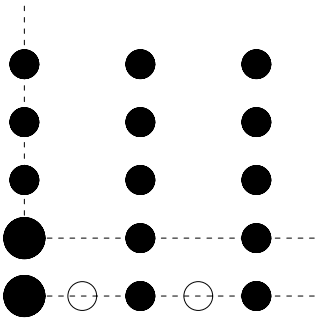
toroidal firmament, associated $\sqcup X_i \to Y$. This gives a firmament in the "new" sense on Y.

One can show that the two procedures are inverse to each other. Again I'll leave this for a later treatment.

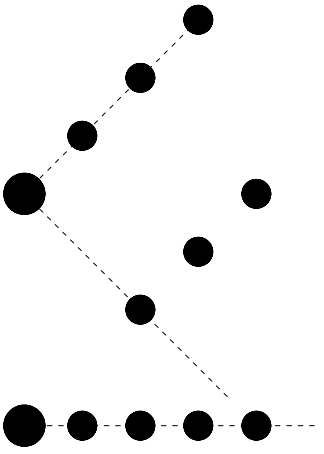
This shows in particular that any firmament supports a unique constellation, thus allowing us access to the differential invariants of constellations.

2.4.15. *Examples revisited.* We can now revisit our examples of base constellations in the one dimensional and higher dimensional cases, and recast them in terms of firmaments. It then becomes evident that the rules we used to calculate the constellations are simply the combinatorial data of firmaments!

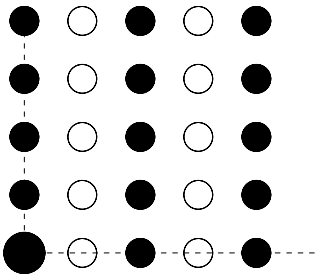
(1) $f : \mathbb{A}^2 \to \mathbb{A}^1$ given by $t = x^2$: $\tau = \mathbb{R}_{\geq 0}; N_\tau = \mathbb{N}; \mathbf{\Gamma} = \{2\mathbb{N}\}$.



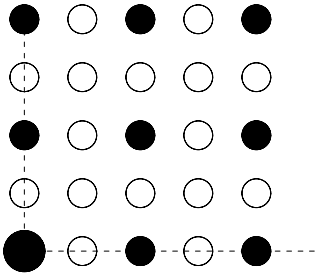
(2) $f : \mathbb{A}^2 \to \mathbb{A}^1$ given by $t = x^2y$: $\mathbf{\Gamma} = \{\mathbb{N}\}$, the trivial structure.



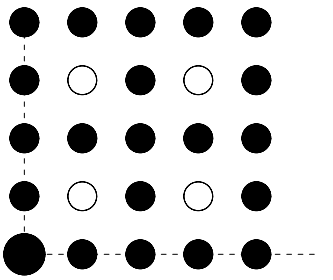
(3) $f : \mathbb{A}^2 \to \mathbb{A}^1$ given by $t = x^2y^2$: $\mathbf{\Gamma} = \{2\mathbb{N}\}$. Supported constellation: $\Delta = D_0/2$
(4) $f : \mathbb{A}^2 \to \mathbb{A}^1$ given by $t = x^2y^3$: $\mathbf{\Gamma} = \{2\mathbb{N} + 3\mathbb{N}\}$. Supported constellation: $\Delta = D_0/2$. Note: this is the same constellation as before, but hanging by different firmaments.
(5) $f : \mathbb{A}^2 \to \mathbb{A}^1$ given by $t = x^3y^4$: $\mathbf{\Gamma} = \{3\mathbb{N} + 4\mathbb{N}\}$. Note: even $\Gamma \smallsetminus \{0\}$ is not saturated in its associated group.
(6) $f : \mathbb{A}^2 \to \mathbb{A}^2$ given by $s = x^2; t = y$: $\mathbf{\Gamma} = \{2\mathbb{N} \times \mathbb{N}\}$.



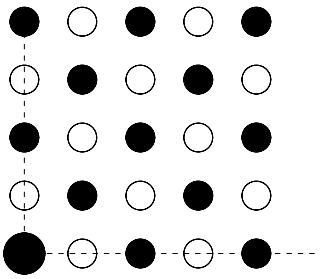
(7) $f : \mathbb{A}^2 \to \mathbb{A}^2$ given by $s = x^2; t = y^2$: $\mathbf{\Gamma} = \{2\mathbb{N} \times 2\mathbb{N}\}$.



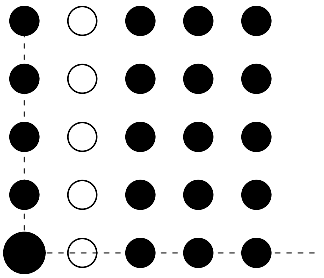
(8) $f : \mathbb{A}^2 \sqcup \mathbb{A}^2 \to \mathbb{A}^2$ given by $s = x_1^2; t = y_1$ and $s = x_2; t = y_2^2$: $\mathbf{\Gamma} = \{2\mathbb{N} \times \mathbb{N}, \mathbb{N} \times 2\mathbb{N}\}$. Note: more than one semigroup. $\Delta_Y = 0$, but on blowup the exceptional gets $1/2$.



(9) $f : X \to \mathbb{A}^2$ given by $\operatorname{Spec} \mathbb{C}[s, t, \sqrt{st}]$: $\mathbf{\Gamma} = \{\langle (2,0), (1,1), (0,2) \rangle\}$.



(10) $f : \mathbb{A}^3 \to \mathbb{A}^2$ given by $s = x^2y^3; t = z$: $\mathbf{\Gamma} = \{(2\mathbb{N} + 3\mathbb{N}) \times \mathbb{N}\}$.



2.4.16. *Arithmetic.* We have learned our lesson—for arithmetic we need to talk about integral points on integral models. I'll restrict to the toroidal case, leaving the general situation to future work.

Definition. An S-integral model of a toroidal firmament $\mathbf{\Gamma}$ on Y consists of an integral toroidal model $\mathcal{Y}'$ of Y'.

Definition 2.4.17. Consider a toroidal firmament $\mathbf{\Gamma}$ on Y/k, and a rational point y such that the firmament is trivial in a neighborhood of y. Let $\mathcal{Y}$ be a toroidal S-integral model.

Then y is a *firm integral point of* $\mathcal{Y}$ with respect to $\mathbf{\Gamma}$ if the section $\operatorname{Spec}\mathcal{O}_{k,S} \to \mathcal{Y}$ is a morphism of firmaments, when $\operatorname{Spec}\mathcal{O}_{k,S}$ is endowed with the trivial firmament.

Explicitly, at each prime $\wp \in \operatorname{Spec}\mathcal{O}_{k,S}$ where y reduces to a stratum with cone σ, consider the associated point $n_{y_\wp} \in N_\sigma$. Then y is firmly S-integral if for every $\wp$ we have $n_{y_\wp} \in \Gamma^i_\sigma$ for some i.

Theorem 2.4.18. *Let $f : X \to Y$ be a proper dominant morphism of varieties over k. There exists a toroidal birational model $X' \to Y'$ and an integral model $\mathcal{Y}'$ such that image of a rational point on X' is a firm S-integral point on $\mathcal{Y}'$ with respect to $\mathbf{\Gamma}_f$.*

In fact, at least after throwing a few small primes into the trash-bin S, a point is S-integral on $\mathcal{Y}'$ with respect to $\mathbf{\Gamma}_f$ if and only if locally in the étale topology on $\mathcal{Y}'$ it lifts to a rational point on X. This is the motivation for the definition.

The following statements are due at least in spirit to Campana.

Conjecture 2.4.19. *Let $(Y/\mathbf{\Delta})$ be a smooth projective Campana constellation supported by firmament $\mathbf{\Gamma}$. Then points on Y integral with respect to Γ are potentially dense if and only if $(Y/\mathbf{\Delta})$ is special.*

Note that this conjecture implies Conjecture 2.2.19: assume this conjecture holds true. Let X be a smooth projective variety. Then rational points are potentially dense if and only if X is special.

3. The minimal model program

For the "quick and easy" introduction to the minimal model program see [**Deb01**]. For a more detailed treatment starting from surfaces see [**Mat02**]. For a full treatment up to 1999 see [**KM98**].

The minimal model program has a beautiful beginning, a rather technical main body of work from the 80s and 90s, and quite an exciting present. In the present account I will skip the technical main body of work.

3.1. Cone of curves.

3.1.1. *Groups of divisors and curves modulo numerical equivalence.* Let X be a smooth complex projective variety.

We denote by $N^1(X)$ the image of $\mathbf{Pic}(X) \to H^2(X,\mathbb{Z})/\text{torsion} \subset H^2(X,\mathbb{Q})$. This is the group of Cartier divisors modulo numerical equivalence.

We denote by $N_1(X)$ the subgroup of $H_2(X,\mathbb{Q})$ generated by the fundamental classes of curves. This is the group of algebraic 1-cycles modulo numerical equivalence.

The intersection pairing restricts to $N^1(X) \times N_1(X) \to \mathbb{Z}$, which over $\mathbb{Q}$ is a perfect pairing.

3.1.2. *Cones of divisors and of curves.* Denote by $\mathrm{Amp}(X) \subset N^1(X)_{\mathbb{Q}}$ the cone generated by classes of ample divisors. We denote by $\mathrm{NEF}(X)$ the closure of $\mathrm{Amp}(X) \subset N^1(X)_{\mathbb{R}}$, called the nef cone of X.

Denote by $NE(X) \subset N_1(X)_{\mathbb{Q}}$ the cone generated by classes of curves. We denote its closure by $\overline{NE}(X)$. The class of a curve C in $NE(X)$ is denoted $[C]$.

THEOREM 3.1.3 (Kleiman). *The class $[D]$ of a Cartier divisor is in the closed cone $\mathrm{NEF}(X)$ if and only if $[D]\cdot[C] \geq 0$ for every algebraic curve $C \subset X$.*

In other words, the cones $\overline{NE}(X)$ and $\mathrm{NEF}(X)$ are dual to each other.

3.2. Bend and break. For any divisor D on X which is not numerically equivalent to 0, the subset

$$(D \leq 0) := \{v \in NE(X) | v \cdot D \leq 0\}$$

is a half-space. The minimal model program starts with the observation that this set is especially important when $D = K_X$. In fact, in the case of surfaces, $(K_X \leq 0) \cap NE(X)$ is a subcone generated by (-1)-curves, which suggests that it must say something in higher dimensions. Indeed, as it turns out, it is in general a nice cone generated by so-called "extremal rays", represented by rational curves $[C]$ which can be contracted in something like a (-1) contraction.

Suppose again X is a smooth, projective variety with K_X not nef. Our first goal is to show that there is *some* rational curve C with $K_X \cdot C < 0$.

The idea is to take an arbitrary curve on X, and to show, using deformation theory, that it has to "move around a lot"—it has so many deformations that eventually it has to break, unless it is already the rational curve we were looking for.

3.2.1. *Breaking curves.* The key to showing that a curve breaks is the following:

LEMMA 3.2.2. *Suppose C is a projective curve of genus > 0 with a point $p \in C$, suppose B is a one dimensional affine curve, $f : C \times B \to X$ a nonconstant morphism such that $\{p\} \times B \to X$ is constant. Then, in the closure of $f(C \times B) \subset X$, there is a rational curve passing through $f(p)$.*

In genus 0 a little more will be needed:

LEMMA 3.2.3. *Suppose C is a projective curve of genus 0 with points $p_1, p_2 \in C$, suppose B is a one dimensional affine curve, $f : C \times B \to X$ a morphism such that $\{p_i\} \times B \to X$ is constant, $i = 1, 2$, and the image is two-dimensional. Then the class $[f(C)] \in NE(X)$ is "reducible": there are effective curves C_1, C_2 passing through p_1, p_2 respectively, such that $[C_1] + [C_2] = [C]$.*

3.2.4. *Some deformation theory.* We need to understand deformations of a map $f : C \to X$ fixing a point or two. The key is that the tangent space of the moduli space of such maps—the deformation space—can be computed cohomologically, and the number of equations of the deformation space is also bounded cohomologically.

LEMMA 3.2.5. *The tangent space of the deformation space of $f : C \to X$ fixing points $p_1, \ldots, p_n$ is*

$$H^0\left(C, f^*T_X(-\sum p_i)\right).$$

The obstructions lie in the next cohomology group:

$$H^1\left(C, f^*T_X(-\sum p_i)\right).$$

The dimension of the deformation space is bounded below:

$$\begin{aligned}\dim Def(f : C \to X, p_1, \ldots, p_n) &\geq \chi\left(C, f^*T_X(-\sum p_i)\right)\\ &= -(K_X \cdot C) + (1 - g(C) - n)\dim X\end{aligned}$$

3.2.6. *Rational curves.* Let us consider the case where C is rational. Suppose we have such a rational curve inside X with $-(K_X \cdot C) \geq \dim X + 2$, and we consider deformations fixing $n = 2$ of its points. Then $-(K_X \cdot C) + (1 - g(C) - 2)\dim X = -(K_X \cdot C) - \dim X \geq 2$. Since C is inside X, the only ways $f : C \to X$ can deform is either by the 1-parameter group of automorphisms, or, beyond 1-parameter, go outside the image of C, and we get an image of dimension at least 2. So the rational curve must break, and one of the resulting components C_1 is a curve with $-(K_X \cdot C_1) < -(K_X \cdot C)$.

Suppose for a moment $-K_X$ is ample, so its intersection number with an effective curve is positive. In this case the process can only stop once we have a curve C_∞ with

$$0 < -(K_X \cdot C_\infty) \leq \dim X + 1.$$

Note that this is optimal—the canonical line bundle on $\mathbb{P}^r$ has degree $r + 1$ on any line.

3.2.7. *Higher genus.* If X is any projective variety with K_X not nef, then there is some curve C with $K_X \cdot C < 0$. To be able to break C we need

$$-(K_X \cdot C) - g(C)\dim X \geq 1.$$

There is apparently a problem: the genus term may offset the positivity of $-(K_X \cdot C)$. One might think of replacing C by a curve covering C, but there is again a problem: the genus increases in coverings roughly by a factor of the degree of the cover, and this offsets the increase in $-(K_X \cdot C)$. There is one case when this does not happen, that is in characteristic p we can take the iterated Frobenius morphism $C^{[m]} \to C$, and the genus of $C^{[m]}$ is $g(C)$. We can apply our bound and deduce that there is a rational curve C' on X. If $-K_X$ is ample we also have $0 < -(K_X \cdot C') \leq \dim X + 1$.

But our variety X was a complex projective variety. What do we do now? We can find a smooth model $\mathcal{X}$ of X over some ring R finitely generated over $\mathbb{Z}$, and for each maximal ideal $\wp \subset R$ the fiber $\mathcal{X}_\wp$ has a rational curve on it.

How do we deduce that there is a rational curve on the original X? If $-K_X$ is ample, the same is true for $-K_{\mathcal{X}}$, and we deduce that there is a rational curve $C_\wp$ on each $\mathcal{X}_\wp$ such that $0 < -(K_{\mathcal{X}_\wp} \cdot C_\wp) \leq \dim X + 1$. These are parametrized by a

Hilbert scheme of finite type over R, and therefore this Hilbert scheme has a point over $\mathbb{C}$, namely there is a rational curve C on X with $0 < -(K_X \cdot C) \leq \dim X + 1$.

In case $-K_X$ is not ample, a more delicate argument is necessary. One fixes an ample line bundle H on $\mathcal{X}$, and given a curve C on X with $-(K_X \cdot C) < 0$ one shows that there is a rational curve C' on each $\mathcal{X}_\wp$ with

$$(H \cdot C') \leq 2 \dim X \frac{H \cdot C}{-(K_X \cdot C)}.$$

Then one continues with a similar Hilbert scheme argument.

3.3. Cone theorem. Using some additional delicate arguments one proves:

THEOREM 3.3.1 (Cone theorem). *Let X be a smooth projective variety. There is a countable collection C_i of rational curves on X with*

$$0 < -(K_X \cdot C_i) \leq \dim X + 1,$$

whose classes $[C_i]$ are discrete in the half space $N_1(X)_{K_X<0}$, such that

$$\overline{NE}(X) \quad = \overline{NE}(X)_{K_X \geq 0} + \sum_i \mathbb{R}_{\geq 0} \cdot [C_i].$$

The rays $\mathbb{R}_{\geq 0} \cdot [C_i]$ are called extremal rays (or, more precisely, extremal K_X-negative rays) of X.

These extremal rays have a crucial property:

THEOREM 3.3.2 (Contraction theorem). *Let X be a smooth complex projective variety and let $R = \mathbb{R}_{\geq 0} \cdot [C]$ be an extremal K_X-negative ray. Then there is a normal projective variety Z and a surjective morphism $c_R : X \to Z$ with connected fibers, unique up to unique isomorphism, such that for an irreducible curve $D \subset X$ we have $c_R(D)$ is a point if and only if $[D] \in R$.*

This map c_R is defined using a base-point-free linear system on X made out of a combination of an ample sheaf H and K_X.

3.4. The minimal model program. If X has an extremal ray which gives a contraction to a lower dimensional variety Z, then the fibers of c_R are rationally connected and we did learn something important about the structure of X: it is uniruled.

Otherwise $c_R : X \to Z$ is birational, but at least we have gotten rid of one extremal ray - one piece of obstruction for K_X to be nef.

One is tempted to apply the contraction theorem repeatedly, replacing X by Z, until we get to a variety with K_X nef. There is a problem: the variety Z is often singular, and the theorems apply to smooth varieties. All we can say about Z is that it has somewhat mild singularities: in general it has rational singularities; if the exceptional locus has codimension 1—the case of a so-called *divisorial* contraction—the variety Z has so called *terminal* singularities. For surfaces, terminal singularities are in fact smooth, and in fact contractions of extremal rays are just (-1)-contraction, and we eventually are led to a minimal model. But in higher dimensions singularities do occur.

The good news is that the theorems can be extended, in roughly the same form, to varieties with terminal singularities. (The methods are very different from what we have seen and I would rather not go into them.) So as long as we only need to deal with divisorial contractions, we can continue as in the surface case.

For non-divisorial contractions—so-called small contractions—we have the following recent major result of Birkar, Cascini, Hacon and McKernan:

THEOREM 3.4.1 (Flip Theorem [**BCHM06**]). *Suppose $c_R : X \to Z$ is a small extremal contraction on a variety X with terminal singularities. Then there exists another small contraction $c_R^+ : X^+ \to Z$ such that X^+ has terminal singularities and $K_{X^+} \cdot C > 0$ for any curve C contracted by c_R^+.*

The transformation $X \dashrightarrow X^+$ is known as a *flip*.

The proof of this theorem goes by way of a spectacular inductive argument, where proofs of existence of minimal models for varieties of general type, finite generation of canonical rings, and finiteness of certain minimal models are intertwined.

CONJECTURE 3.4.2 (Termination Conjecture). *Any sequence of flips is finite.*

This implies the following:

CONJECTURE 3.4.3 (Minimal model conjecture). *Let X be a smooth projective variety. Then either X is uniruled, or there is a birational modification $X \dashrightarrow X'$ such that X' has only terminal singularities and $K_{X'}$ is nef*

Often one combines this with the following:

CONJECTURE 3.4.4 (Abundance). *Let X be a projective variety with terminal singularities and K_X nef. Then for some integer $m > 0$, we have $H^0(X, \mathcal{O}_X(mK_X))$ is base-point-free.*

The two together are sometimes named "the good minimal model conjecture".

The result is known for varieties of general type: it follows from the recent theorem of [**BCHM06**] on finite generation of canonical rings.

As we have seen in previous sections, this conjecture has a number of far reaching corollaries, including Iitaka's additivity conjecture and the $(-\infty)$-conjecture.

4. Vojta, Campana and *abc*

In [**Voj87**], Paul Vojta started a speculative investigation in Diophantine geometry motivated by analogy with value distribution theory. His conjectures go in the same direction as Lang's—they are concerned with bounding the set of points on a variety rather than constructing "many" rational points. Many of the actual proofs in the subject, such as an alternative proof of Faltings's theorem, use razor-sharp tools such as Arakelov geometry. But to describe the relevant conjectures it will suffice to discuss heights from the classical "naïve" point of view. The reader is encouraged to consult Hindry–Silverman [**HS00**] for a user–friendly, Arakelov–free treatment of the theory of heights (including a proof of Faltings's theorem, following Bombieri).

A crucial feature of Vojta's conjectures is that they are not concerned just with rational points, but with algebraic points of bounded degree. To account for varying fields of definition, Vojta's conjecture always has the discriminant of the field of definition of a point P accounted for.

Vojta's conjectures are thus much farther-reaching than Lang's. You might say, much more outrageous. On the other hand, working with all extensions of a bounded degree allows for enormous flexibility in using geometric constructions in the investigation of algebraic points. So, even if one is worried about the validity of the conjectures, they serve as a wonderful testing ground for our arithmetic intuition.

4.1. Heights and related invariants. Consider a point in projective space $P = (x_0 : \ldots : x_r) \in \mathbb{P}^r$, defined over some number field k, with set of places $\mathbb{M}_k$. Define the *naïve height* of P to be

$$H(P) = \prod_{v \in \mathbb{M}_k} \max(\|x_0\|_v, \ldots, \|x_r\|_v).$$

Here $\|x\|_v = |x|$ for a real v, $\|x\|_v = |x|^2$ for a complex v, and $\|x\|_v$ is normalized so that $\|p\| = p^{-[k_v:\mathbb{Q}_p]}$ otherwise. (If the coordinates can be chosen relatively prime algebraic integers, then the product is of course a finite product over the Archimedean places, where everything is as easy as can be expected.)

This height is independent of the homogeneous coordinates chosen, by the product formula.

To keep things independent of a chosen field of definition, and to replace products by sums, one defines the normalized logarithmic height

$$h(P) = \frac{1}{[k : \mathbb{Q}]} \log H(P).$$

Now if X is a variety over k with a very ample line bundle L, one can consider the embedding of X in a suitable $\mathbb{P}^r$ via the complete linear system of $H^0(X, L)$. We define the height $h_L(P)$ to be the height of the image point in $\mathbb{P}^r$.

This definition of $h_L(P)$ is not valid for embeddings by incomplete linear systems, and is not additive in L. But it does satisfy these desired properties "almost": $h_L(P) = h(P) + O(1)$ if we embed by an incomplete linear system, and $h_{L \otimes L'}(P) = h_L(P) + h_{L'}(P)$ for very ample L, L'. This allows us to define

$$h_L(P) = h_A(P) - h_B(P)$$

where A and B are very ample and $L \otimes B = A$. The function $h_L(P)$ is now only well defined as a function on $X(\bar{k})$ up to $O(1)$.

Consider a finite set of places S containing all Archimedean places.

Let now $\mathcal{X}$ be a scheme proper over $\mathcal{O}_{k,S}$, and D a Cartier divisor.

The counting function of $\mathcal{X}, D$ relative to k, S is a function on points of $X(\bar{k})$ not lying on D. Suppose $P \in X(E)$, which we view again as an S-integral point of $\mathcal{X}$. Consider a place w of E not lying over S, with residue field $\kappa(w)$. Then the restriction of D to $P \simeq \operatorname{Spec} \mathcal{O}_{E,S}$ is a fractional ideal with some multiplicity n_w at w. We define the *counting function* as follows:

$$N_{k,S}(D, P) = \frac{1}{[E : k]} \sum_{\substack{w \in \mathbb{M}_E \\ w \nmid S}} n_w \log |\kappa(w)|.$$

A variant of this is the *truncated counting function*

$$N^{(1)}_{k,S}(D, P) = \frac{1}{[E : k]} \sum_{\substack{w \in \mathbb{M}_E \\ w \nmid S}} \min(1, n_w) \log |\kappa(w)|.$$

Counting functions and truncated counting functions depend on the choice of S and a model $\mathcal{X}$, but only up to $O(1)$. We'll thus suppress the subscript S.

One defines the *relative logarithmic discriminant* of E/k as follows: suppose the discriminant of a number field k is denoted D_k. Then define

$$d_k(E) = \frac{1}{[E:k]} \log |D_E| \quad - \quad \log |D_k|.$$

4.2. Vojta's conjectures.

CONJECTURE 4.2.1. *Let X be a smooth proper variety over a number field k, D a normal crossings divisor on X, and A an ample line bundle on X. Let r be a positive integer and $\epsilon > 0$. Then there is a proper Zariski-closed subset $Z \subset X$ containing D such that*

$$N_k(D,P) + d_k(k(P)) \geq h_{K_X(D)}(P) - \epsilon h_A(P) - O(1)$$

for all $P \in X(\bar{k}) \smallsetminus Z$ with $[k(P):k] \leq r$.

In the original conjecture in [**Voj87**], the discriminant term came with a factor $\dim X$. By the time of [**Voj98**] Vojta came to the conclusion that the factor was not well justified. A seemingly stronger version is

CONJECTURE 4.2.2. *Let X be a smooth proper variety over a number field k, D a normal crossings divisor on X, and A an ample line bundle on X. Let r be a positive integer and $\epsilon > 0$. Then there is a proper Zariski-closed subset $Z \subset X$ containing D such that*

$$N_k^{(1)}(D,P) + d_k(k(P)) \geq h_{K_X(D)}(P) - \epsilon h_A(P) - O(1).$$

but in [**Voj98**], Vojta shows that the two conjectures are equivalent.

4.3. Vojta and *abc*. The following discussion is taken from [**Voj98**], section 2.

The Masser-Oesterlé *abc* conjecture is the following:

CONJECTURE 4.3.1. *For any $\epsilon > 0$ there is $C > 0$ such that for all $a, b, c \in \mathbb{Z}$, with $a + b + c = 0$ and $\gcd(a,b,c) = 1$ we have*

$$\max(|a|,|b|,|c|) \quad \leq \quad C \cdot \prod_{p|abc} p^{1+\epsilon}.$$

Consider the point $P = (a : b : c) \in \mathbb{P}^2$. Its height is $\log \max(|a|,|b|,|c|)$. Of course the point lies on the line X defined by $x + y + z = 0$. If we denote by D the divisor of $xyz = 0$, that is the intersection of X with the coordinate axes, and if we set $S = \{\infty\}$, then

$$N^{(1)}_{\mathbb{Q},S}(D,P) = \sum_{p|abc} \log p.$$

So the *abc* conjecture says

$$h(P) \leq (1+\epsilon) N^{(1)}_{\mathbb{Q},S}(D,P) + O(1),$$

which, writing $1 - \epsilon' = (1+\epsilon)^{-1}$, is the same as

$$(1-\epsilon')h(P) \leq N^{(1)}_{\mathbb{Q},S}(D,P) + O(1).$$

This is applied only to rational points on X, so $d_{\mathbb{Q}}(\mathbb{Q}) = 0$. We have $K_X(D) = \mathcal{O}_X(1)$, and setting $A = \mathcal{O}_X(1)$ as well we get that *abc* is equivalent to

$$N^{(1)}_{\mathbb{Q},S}(D,P) \geq h_{K_X(D)}(P) - \epsilon' h_A(P) - O(1),$$

which is exactly what Vojta's conjecture predicts in this case.

Note that the same argument gives the *abc* conjecture over any fixed number field.

4.4. *abc* **and Campana.** Material in this section follows Campana's [**Cam05**].

Let us go back to Campana's constellation curves. Recall Conjecture 2.1.9, in particular a Campana constellation curve of general type over a number field is conjectured to have a finite number of soft S-integral points.

Simple inequalities, along with Faltings's theorem, allow Campana to reduce to a finite number of cases, all on $\mathbb{P}^1$. The multiplicities m_i that occur in these "minimal" divisors Δ on $\mathbb{P}^1$ are

$$(2,3,7),\ (2,4,5),\ (3,3,4),\ (2,2,2,3) \quad \text{and} \quad (2,2,2,2,2).$$

Now one claims that Campana's conjecture in these cases follows from the *abc* conjecture for the number field k. This follows from a simple application of Elkies's [**Elk91**]. It is easiest to verify in case $k = \mathbb{Q}$ when Δ is supported precisely at 3 points, with more points one needs to use Belyi maps (in the function field case one uses a proven generalization of *abc* instead).

We may assume Δ is supported at $0, 1$ and ∞. An integral point on $(\mathbb{P}^1/\Delta)$ in this case is a rational point a/c such that a, c are integers, satisfying the following:

- whenever $p|a$, in fact $p^{n_0}|a$;
- whenever $p|b$, in fact $p^{n_1}|b$; and
- whenever $p|c$, in fact $p^{n_\infty}|c$,

where $b = c - a$.

Now if $M = \max(|a|, |b|, |c|)$ then

$$M^{1/n_0+1/n_1+1/n_\infty} \geq |a|^{1/n_0}|b|^{1/n_1}|c|^{1/n_\infty},$$

and by assumption $a^{1/n_0} \geq \prod_{p|a} p$, and similarly for b, c. In other words

$$M^{1/n_0+1/n_1+1/n_\infty} \geq \prod_{p|abc} p.$$

Since, by assumption, $1/n_0 + 1/n_1 + 1/n_\infty < 1$ we can take any $0 < \epsilon < 1 - 1/n_0 + 1/n_1 + 1/n_\infty$, for which the *abc* conjecture gives $M^{1-\epsilon} < C \prod_{p|abc} p$, for some C. So $M^{1-1/n_0+1/n_1+1/n_\infty-\epsilon} < C$ and M is bounded, so there are only finitely many such points.

4.5. Vojta and Campana. I speculate: Vojta's higher dimensional conjecture implies the non-special part of Campana's conjecture 2.2.19, i.e., if X is non-special its set of rational points is not dense.

The problem is precisely in understanding what happens when a point reduces to the singular locus of D.

References

[Abr95] D. Abramovich, *Uniformité des points rationnels des courbes algébriques sur les extensions quadratiques et cubiques*, C. R. Acad. Sci. Paris Sér. I Math. **321** (1995), no. 6, 755–758. MR 1354720 (96g:14017)

[Abr97] ———, *A high fibered power of a family of varieties of general type dominates a variety of general type*, Invent. Math. **128** (1997), no. 3, 481–494. MR 1452430 (98e:14034)

[AK00] D. Abramovich and K. Karu, *Weak semistable reduction in characteristic 0*, Invent. Math. **139** (2000), no. 2, 241–273. MR 1738451 (2001f:14021)

[Alu07] P. Aluffi, *Celestial integration, stringy invariants, and Chern-Schwartz-MacPherson classes*, Real and complex singularities, Trends Math., Birkhäuser, Basel, 2007, pp. 1–13. MR 2280127 (2008c:14007)

[AV96] D. Abramovich and J. F. Voloch, *Lang's conjectures, fibered powers, and uniformity*, New York J. Math. **2** (1996), 20–34, electronic. MR 1376745 (97e:14031)

[BCHM06] C. Birkar, P. Cascini, C. Hacon, and J. McKernan, *Existence of minimal models for varieties of log general type*, 2006, math.AG/0610203.

[Bou15] N. Bourbaki, *Logarithmic Structures*, Vanish and Perish Press, Furnace Heat, Purgatory, 2015.

[Cam92] F. Campana, *Connexité rationnelle des variétés de Fano*, Ann. Sci. École Norm. Sup. (4) **25** (1992), no. 5, 539–545. MR 1191735 (93k:14050)

[Cam04] ———, *Orbifolds, special varieties and classification theory*, Ann. Inst. Fourier (Grenoble) **54** (2004), no. 3, 499–630. MR 2097416 (2006c:14013)

[Cam05] ———, *Fibres multiples sur les surfaces: aspects geométriques, hyperboliques et arithmétiques*, Manuscripta Math. **117** (2005), no. 4, 429–461. MR 2163487 (2006e:14013)

[Cap] L. Caporaso, *Lecture at MSRI, January 2006*, `http://www.math.brown.edu/~abrmovic/GOTTINGEN/Caporaso-MSRI.pdf`.

[CHM97] L. Caporaso, J. Harris, and B. Mazur, *Uniformity of rational points*, J. Amer. Math. Soc. **10** (1997), no. 1, 1–35. MR 1325796 (97d:14033)

[CTSSD97] J.-L. Colliot-Thélène, A. N. Skorobogatov, and P. Swinnerton-Dyer, *Double fibres and double covers: paucity of rational points*, Acta Arith. **79** (1997), no. 2, 113–135. MR 1438597 (98a:11081)

[Deb01] O. Debarre, *Higher-dimensional algebraic geometry*, Universitext, Springer-Verlag, New York, 2001. MR 1841091 (2002g:14001)

[DG95] H. Darmon and A. Granville, *On the equations* $z^m = F(x, y)$ *and* $Ax^p + By^q = Cz^r$, Bull. London Math. Soc. **27** (1995), no. 6, 513–543. MR 1348707 (96e:11042)

[Elk91] N. D. Elkies, *ABC implies Mordell*, Internat. Math. Res. Notices (1991), no. 7, 99–109. MR 1141316 (93d:11064)

[Fal83] G. Faltings, *Endlichkeitssätze für abelsche Varietäten über Zahlkörpern*, Invent. Math. **73** (1983), no. 3, 349–366. MR 718935 (85g:11026a)

[Fuj78] T. Fujita, *On Kähler fiber spaces over curves*, J. Math. Soc. Japan **30** (1978), no. 4, 779–794. MR 513085 (82h:32024)

[GHS03] T. Graber, J. Harris, and J. Starr, *Families of rationally connected varieties*, J. Amer. Math. Soc. **16** (2003), no. 1, 57–67 (electronic). MR 1937199 (2003m:14081)

[GR09] Ofer Gabber and Lorenzo Ramero, *Foundations of p-adic Hodge theory*, 2009, arXiv:math/0409584v4.

[Gra65] H. Grauert, *Mordells Vermutung über rationale Punkte auf algebraischen Kurven und Funktionenkörper*, Inst. Hautes Études Sci. Publ. Math. (1965), no. 25, 131–149. MR 0222087 (36 #5139)

[Har77] R. Hartshorne, *Algebraic geometry*, Springer-Verlag, New York, 1977, Graduate Texts in Mathematics, No. 52. MR 0463157 (57 #3116)

[Has96] B. Hassett, *Correlation for surfaces of general type*, Duke Math. J. **85** (1996), no. 1, 95–107. MR 1412439 (97i:14025)

[HS00] M. Hindry and J. H. Silverman, *Diophantine geometry*, Graduate Texts in Mathematics, vol. 201, Springer-Verlag, New York, 2000, An introduction. MR 1745599 (2001e:11058)

[Iit82] S. Iitaka, *Algebraic geometry*, Graduate Texts in Mathematics, vol. 76, Springer-Verlag, New York, 1982, An introduction to birational geometry of algebraic varieties, North-Holland Mathematical Library, 24. MR 637060 (84j:14001)

[Kat94] K. Kato, *Toric singularities*, Amer. J. Math. **116** (1994), no. 5, 1073–1099. MR 1296725 (95g:14056)

[Kaw85] Y. Kawamata, *Minimal models and the Kodaira dimension of algebraic fiber spaces*, J. Reine Angew. Math. **363** (1985), 1–46. MR 814013 (87a:14013)

[KKMSD73] G. Kempf, F. F. Knudsen, D. Mumford, and B. Saint-Donat, *Toroidal embeddings. I*, Lecture Notes in Mathematics, Vol. 339, Springer-Verlag, Berlin, 1973. MR 0335518 (49 #299)

[KM98] J. Kollár and S. Mori, *Birational geometry of algebraic varieties*, Cambridge Tracts in Mathematics, vol. 134, Cambridge University Press, Cambridge, 1998, With the collaboration of C. H. Clemens and A. Corti, Translated from the 1998 Japanese original. MR 1658959 (2000b:14018)

[KMM92] J. Kollár, Y. Miyaoka, and S. Mori, *Rationally connected varieties*, J. Algebraic Geom. **1** (1992), no. 3, 429–448. MR 1158625 (93i:14014)

[Kol87] J. Kollár, *Subadditivity of the Kodaira dimension: fibers of general type*, Algebraic geometry, Sendai, 1985, Adv. Stud. Pure Math., vol. 10, North-Holland, Amsterdam, 1987, pp. 361–398. MR 946244 (89i:14029)

[Lu02] S. S. Y. Lu, *A refined Kodaira dimension and its canonical fibration*, 2002, math.AG/0211029.

[Man63] Yu. I. Manin, *Rational points on algebraic curves over function fields*, Izv. Akad. Nauk SSSR Ser. Mat. **27** (1963), 1395–1440, Corrected in [**Man89**]. MR 0157971 (28 #1199)

[Man89] ———, *Letter to the editors: "Rational points on algebraic curves over function fields"*, Izv. Akad. Nauk SSSR Ser. Mat. **53** (1989), no. 2, 447–448. MR 998307 (90f:11039)

[Mat02] K. Matsuki, *Introduction to the Mori program*, Universitext, Springer-Verlag, New York, 2002. MR 1875410 (2002m:14011)

[MS02] R. Moosa and T. Scanlon, *The Mordell-Lang conjecture in positive characteristic revisited*, Model theory and applications, Quad. Mat., vol. 11, Aracne, Rome, 2002, pp. 273–296. MR 2159720 (2007a:11085)

[Ogu] A. Ogus, *Lectures on logarithmic algebraic geometry*, manuscript in progress `http://math.berkeley.edu/~ogus/preprints/log_book/logbook.pdf`.

[Pac97] P. Pacelli, *Uniform boundedness for rational points*, Duke Math. J. **88** (1997), no. 1, 77–102. MR 1448017 (98b:14020)

[Sam66] P. Samuel, *Compléments à un article de Hans Grauert sur la conjecture de Mordell*, Inst. Hautes Études Sci. Publ. Math. (1966), no. 29, 55–62. MR 0204430 (34 #4272)

[Sho03] V. V. Shokurov, *Prelimiting flips*, Tr. Mat. Inst. Steklova **240** (2003), no. Biratsion. Geom. Linein. Sist. Konechno Porozhdennye Algebry, 82–219, English translation in: Proc. Steklov Inst. Math. 2003, no. 1 (240), 75–213. MR 1993750 (2004k:14024)

[Siu98] Y.-T. Siu, *Invariance of plurigenera*, Invent. Math. **134** (1998), no. 3, 661–673. MR 1660941 (99i:32035)

[Siu02] ———, *Extension of twisted pluricanonical sections with plurisubharmonic weight and invariance of semipositively twisted plurigenera for manifolds not necessarily of general type*, Complex geometry (Göttingen, 2000), Springer, Berlin, 2002, pp. 223–277. MR 1922108 (2003j:32027a)

[Vie82] E. Viehweg, *Die Additivität der Kodaira Dimension für projektive Faserräume über Varietäten des allgemeinen Typs*, J. Reine Angew. Math. **330** (1982), 132–142. MR MR641815 (83f:14007)

[Voj87] P. Vojta, *Diophantine approximations and value distribution theory*, Lecture Notes in Mathematics, vol. 1239, Springer-Verlag, Berlin, 1987. MR 883451 (91k:11049)

[Voj98] ———, *A more general abc conjecture*, Internat. Math. Res. Notices (1998), no. 21, 1103–1116. MR 1663215 (99k:11096)

[Zar44] O. Zariski, *The compactness of the Riemann manifold of an abstract field of algebraic functions*, Bull. Amer. Math. Soc. **50** (1944), 683–691. MR MR0011573 (6,186b)

Department of Mathematics, Box 1917, Brown University, Providence, RI, 02912, U.S.A

E-mail address: `abrmovic@math.brown.edu`

Clay Mathematics Proceedings
Volume 8, 2009

Arithmetic over function fields

Jason Michael Starr

Abstract. These notes accompany lectures presented at the Clay Mathematics Institute 2006 Summer School on Arithmetic Geometry. The lectures summarize some recent progress on existence of rational points of projective varieties defined over a function field over an algebraically closed field.

1. Introduction

These notes accompany lectures presented at the Clay Mathematics Institute 2006 Summer School on Arithmetic Geometry. They are more complete than the lectures themselves. Exercises assigned during the lectures are proved as lemmas or propositions in these notes. Hopefully this makes the notes useful to a wider audience than the original participants of the summer school.

This report describes some recent progress on questions in the interface between arithmetic geometry and algebraic geometry. In fact the questions come from arithmetic geometry: what is known about existence and "abundance" of points on algebraic varieties defined over a non-algebraically closed field K. But the answers are in algebraic geometry, i.e., they apply only when the field K is the function field of an algebraic variety over an algebraically closed field. For workers in number theory, such answer are of limited interest. But hopefully the techniques will be of interest, perhaps as simple analogues for more advanced techniques in arithmetic. With regards to this hope, the reader is encouraged to look at two articles on the arithmetic side, [**GHMS04a**] and [**GHMS04b**]. Also, of course, the answers have interesting consequences within algebraic geometry itself.

There are three sections corresponding to the three lectures I delivered in the summer school. The first lecture proves the classical theorems of Chevalley-Warning and Tsen-Lang: complete intersections in projective space of sufficiently low degree defined over finite fields or over function fields always have rational points. These theorems imply corollaries about the Brauer group and Galois cohomology of these fields, which are also described.

The second section introduces rationally connected varieties and presents the proof of Tom Graber, Joe Harris and myself of a conjecture of Kollár, Miyaoka and Mori: every rationally connected fibration over a curve over an algebraically closed field of

2000 *Mathematics Subject Classification.* Primary 14G05, Secondary 11G35, 14F22, 14D15.

characteristic 0 has a section. The proof presented here incorporates simplifications due to A. J. de Jong. Some effort is made to indicate the changes necessary to prove A. J. de Jong's generalization to separably rationally connected fibrations over curves over fields of arbitrary characteristic. In the course of the proof, we give a thorough introduction to the "smoothing combs" technique of Kollár, Miyaoka and Mori and its application to weak approximation for "generic jets" in smooth fibers of rationally connected fibrations. This has been significantly generalized to weak approximation for *all* jets in smooth fibers by Hassett and Tschinkel, cf. [**HT06**]. Some corollaries of the Kollár-Miyaoka-Mori conjecture to Mumford's conjecture, fixed point theorems, and fundamental groups are also described (these were known to follow before the conjecture was proved).

Finally, the last section hints at the beginnings of a generalization of the Kollár-Miyaoka-Mori conjecture to higher-dimensional function fields (not just function fields of curves). A rigorous result in this area is a second proof of A. J. de Jong's *Period-Index Theorem*: for a division algebra D whose center is the function field K of a surface, the index of D equals the order of $[D]$ in the Brauer group of K. This also ties together the first and second sections. Historically the primary motivation for the theorems of Chevalley, Tsen and Lang had to do with Brauer groups and Galois cohomology. The subject has grown beyond these first steps. But the newer results do have consequences for Brauer groups and Galois cohomology in much the same vein as the original results in this subject.

Acknowledgments. I am grateful to the Clay Mathematics Institute for sponsoring such an enjoyable summer school. I am grateful to Brendan Hassett, Yuri Tschinkel and A. J. de Jong for useful conversations on the content and exposition of these notes. And I am especially grateful to the referees whose comments, both positive and negative, improved this article.

2. The Tsen-Lang theorem

A motivating problem in both arithmetic and geometry is the following.

PROBLEM 2.1. Given a field K and a K-variety X find sufficient, resp. necessary, conditions for existence of a K-point of X.

The problem depends dramatically on the type of K: number field, finite field, p-adic field, function field over a finite field, or function field over an algebraically closed field. In arithmetic the number field case is most exciting. However the geometric case, i.e., the case of a function field over an algebraically closed field, is typically easier and may suggest approaches and conjectures in the arithmetic case.

Two results, the Chevalley-Warning theorem and Tsen's theorem, deduce a sufficient condition for existence of K-points by "counting". More generally, counting leads to a relative result: the Tsen-Lang theorem that a strong property about existence of k-points for a field k propagates to a weaker property about K-points for certain field extensions K/k. The prototype result, both historically and logically, is a theorem of Chevalley and its generalization by Warning. The counting result at the heart of the proof is Lagrange's theorem together with the observation that a nonzero single-variable polynomial of degree $\leq q-1$ cannot have q distinct zeroes.

LEMMA 2.2. *For a finite field K with q elements, the polynomial $1 - x^{q-1}$ vanishes on K^* and $x^q - x$ vanishes on all of K. For every integer $n \geq 0$, for the K-algebra homomorphism*

$$\mathrm{ev}_n : K[X_0, \ldots, X_n] \to \mathrm{Hom}_{\mathrm{Sets}}(K^{n+1}, K),$$

$$\mathrm{ev}_n(p(X_0, \ldots, X_n)) = ((a_0, \ldots, a_n) \mapsto p(a_0, \ldots, a_n)),$$

the kernel equals the ideal

$$I_n = \langle X_0^q - X_0, \ldots, X_n^q - X_n \rangle.$$

Finally, the collection $(X_i^q - X_i)_{i=0,\ldots,n}$ is a Gröbner basis with respect to every monomial order refining the grading of monomials by total order. In particular, for every p in I_n some term of p of highest degree is in the ideal $\langle X_0^q, \ldots, X_n^q \rangle$.

PROOF. Because K^* is a group of order $q - 1$, Lagrange's theorem implies $a^{q-1} = 1$ for every element a of K^*, i.e., $1 - x^{q-1}$ vanishes on K^*. Multiplying by x shows that $x^q - x$ vanishes on K. Thus the ideal I_n is at least contained in the kernel of ev_n.

Modulo $X_n^q - X_n$, every element of $K[X_0, \ldots, X_n]$ is congruent to one of the form

$$p(X_0, \ldots, X_n) = p_{q-1} X_n^{q-1} + \cdots + p_0 X_n^0, \quad p_0, \ldots, p_{q-1} \in K[X_0, \ldots, X_{n-1}].$$

(Of course K^n is defined to be $\{0\}$ and $K[X_0, \ldots, X_{n-1}]$ is defined to be K if n equals 0.) Since K has q elements and since a nonzero polynomial of degree $\leq q-1$ can have at most $q-1$ distinct zeroes, for every $(a_0, \ldots, a_{n-1}) \in K^n$ the polynomial $p(a_0, \ldots, a_{n-1}, X_n)$ is zero on K if and only if

$$p_0(a_0, \ldots, a_{n-1}) = \cdots = p_{q-1}(a_0, \ldots, a_{n-1}).$$

Thus $\mathrm{ev}_n(p)$ equals 0 if and only if each $\mathrm{ev}_{n-1}(p_i)$ equals 0. In that case, by the induction hypothesis, each p_i is in I_{n-1} (in case $n = 0$, each p_i equals 0). Then, since $I_{n-1}K[X_0, \ldots, X_n]$ is in I_n, p is in I_n. Therefore, by induction on n, the kernel of ev_n is precisely I_n.

Finally, Buchberger's algorithm applied to the set $(X_0^q - X_0, \ldots, X_n^q - X_n)$ produces S-polynomials

$$S_{i,j} = X_j^q(X_i^q - X_i) - X_i^q(X_j^q - X_j) = X_j(X_i^q - X_i) - X_i(X_j^q - X_j)$$

which have remainder 0. Therefore this set is a Gröbner basis by Buchberger's criterion. □

THEOREM 2.3. [**Che35**],[**War35**] *Let K be a finite field. Let n and r be positive integers and let $F_1, \ldots, F_r$ be nonconstant, homogeneous polynomials in $K[X_0, \ldots, X_n]$. If*

$$\deg(F_1) + \cdots + \deg(F_r) \leq n$$

then there exists $(a_0, \ldots, a_n) \in K^{n+1} - \{0\}$ such that for every $i = 1, \ldots, r$, $F_i(a_0, \ldots, a_n)$ equals 0. Stated differently, the projective scheme $\mathbb{V}(F_1, \ldots, F_r) \subset \mathbb{P}^n_K$ has a K-point.

PROOF. Denote by q the number of elements in K. The polynomial

$$G(X_0, \ldots, X_n) = 1 - \prod_{i=0}^{n} (1 - X_i^{q-1})$$

equals 0 on $\{0\}$ and equals 1 on $K^{n+1} - \{0\}$. For the same reason, the polynomial

$$H(X_0, \ldots, X_n) = 1 - \prod_{j=1}^{r} (1 - F_j(X_0, \ldots, X_n)^{q-1})$$

equals 0 on

$$\{(a_0, \ldots, a_n) \in K^{n+1} | F_1(a_0, \ldots, a_n) = \cdots = F_r(a_0, \ldots, a_n) = 0\}$$

and equals 1 on the complement of this set in K^{n+1}. Since each F_i is homogeneous, 0 is a common zero of $F_1, \ldots, F_r$. Thus the difference $G - H$ equals 1 on

$$\{(a_0, \ldots, a_n) \in K^{n+1} - \{0\} | F_1(a_0, \ldots, a_n) = \cdots = F_r(a_0, \ldots, a_n) = 0\}$$

and equals 0 on the complement of this set in K^{n+1}. Thus, to prove that $F_1, \ldots, F_r$ have a nontrivial common zero, it suffices to prove the polynomial $G - H$ does not lie in the ideal I_n.

Since

$$\deg(F_1) + \cdots + \deg(F_r) \leq n,$$

H has strictly smaller degree than G. Thus the leading term of $G - H$ equals the leading term of G. There is only one term of G of degree $\deg(G)$. Thus, for every monomial ordering refining the grading by total degree, the leading term of G equals

$$(-1)^{n+1} X_0^{q-1} X_1^{q-1} \cdots X_n^{q-1}.$$

This is clearly divisible by none of X_i^q for $i = 0, \ldots, n$, i.e., the leading term of $G - H$ is not in the ideal $\langle X_0^q, \ldots, X_n^q \rangle$. Because $(X_0^q - X_0, \ldots, X_n^q - X_n)$ is a Gröbner basis for I_n with respect to the monomial order, $G - H$ is not in I_n. □

On the geometric side, an analogue of Chevalley's theorem was proved by Tsen, cf. [**Tse33**]. This was later generalized independently by Tsen and Lang, cf. [**Tse36**], [**Lan52**]. Lang introduced a definition which simplifies the argument.

Definition 2.4. [**Lan52**] Let m be a nonnegative integer. A field K is called C_m, or said to have *property* C_m, if it satisfies the following. For every positive integer n and every sequence of positive integers $(d_1, \ldots, d_r)$ satisfying

$$d_1^m + \cdots + d_r^m \leq n,$$

every sequence $(F_1, \ldots, F_r)$ of homogeneous polynomials $F_i \in K[X_0, \ldots, X_n]$ with $\deg(F_i) = d_i$ has a common zero in $K^{n+1} - \{0\}$.

Remark 2.5. In fact the definition in [**Lan52**] is a little bit different than this. For fields having normic forms, Lang proves the definition above is equivalent to his definition. And the definition above works best with the following results.

With this definition, the statement of the Chevalley-Warning theorem is quite simple: every finite field has property C_1. The next result proves that property C_m is preserved by algebraic extension.

Lemma 2.6. *For every nonnegative integer m, every algebraic extension of a field with property C_m has property C_m.*

PROOF. Let K be a field with property C_m and let L'/K be an algebraic extension. For every sequence of polynomials $(F_1, \dots, F_r)$ as in the definition, the coefficients generate a finitely generated subextension L/K of L'/K. Thus clearly it suffices to prove the lemma for finitely generated, algebraic extensions L/K.

Denote by e the finite dimension $\dim_K(L)$. Because multiplication on L is K-bilinear, each homogeneous, degree d_i, polynomial map of L-vector spaces,

$$F_i : L^{\oplus(n+1)} \to L,$$

is also a homogeneous, degree d_i, polynomial map of K-vector spaces. Choosing a K-basis for L and decomposing F_i accordingly, F_i is equivalent to e distinct homogeneous, degree d_i, polynomial maps of K-vector spaces,

$$F_{i,j} : L^{\oplus(n+1)} \to K, \quad j = 1, \dots, e.$$

The set of common zeroes of the collection of homogeneous polynomial maps $(F_i | i = 1, \dots, r)$ equals the set of common zeroes of the collection of homogeneous polynomial functions $(F_{i,j} | i = 1, \dots, r, j = 1, \dots, e)$. Thus it suffices to prove there is a nontrivial common zero of all the functions $F_{i,j}$.

By hypothesis,

$$\sum_{i=1}^{r} \deg(F_i)^m \text{ is no greater than } n.$$

Thus, also

$$\sum_{i=1}^{r}\sum_{j=1}^{e} \deg(F_{i,j})^m = e\sum_{i=1}^{r} \deg(F_i)^m \text{ is no greater than } en.$$

Since K has property C_m and since

$$\dim_K(L^{\oplus(n+1)}), \text{ i.e., } (n+1)\dim_K(L) = e(n+1),$$

is larger than en, the collection of homogeneous polynomials $F_{i,j}$ has a common zero in $L^{\oplus(n+1)} - \{0\}$. □

The heart of the Tsen-Lang theorem is the following proposition.

PROPOSITION 2.7. *Let K/k be a function field of a curve, i.e., a finitely generated, separable field extension of transcendence degree* 1. *If k has property C_m then K has property C_{m+1}.*

This is proved in a series of steps. Let n, r and $d_1, \dots, d_r$ be positive integers such that

$$d_1^{m+1} + \cdots + d_r^{m+1} \leq n.$$

For every collection of homogeneous polynomials

$$F_1, \dots, F_r \in K[X_0, \dots, X_n], \quad \deg(F_i) = d_i,$$

the goal is to prove that the collection of homogeneous, degree d_i, polynomial maps of K-vector spaces

$$F_1, \dots, F_r : K^{\oplus(n+1)} \to K$$

has a common zero. Of course, as in the proof of Lemma 2.6, this is also a collection of homogeneous polynomial maps of k-vector spaces. Unfortunately both of these k-vector spaces are infinite dimensional. However, using geometry, these polynomial maps can be realized as the colimits of polynomial maps of finite dimensional

k-vector spaces. For these maps there is an analogue of the Chevalley-Warning argument replacing the counting argument by a *parameter counting argument* which ultimately follows from the Riemann-Roch theorem for curves. The first step is to give a *projective model* of K/k.

LEMMA 2.8. *For every separable, finitely generated field extension K/k of transcendence degree 1, there exists a smooth, projective, connected curve C over k and an isomorphism of k-extensions $K \cong k(C)$. Moreover the pair $(C, K \cong k(C))$ is unique up to unique isomorphism.*

PROOF. This is essentially the *Zariski-Riemann surface* of the extension K/k. For a proof in the case that k is algebraically closed, see [**Har77**, Theorem I.6.9]. The proof in the general case is similar. □

The isomorphism $K \cong k(C)$ is useful because the infinite dimensional k-vector space $k(C)$ has a plethora of naturally-defined finite dimensional subspaces. For every Cartier divisor D on C, denote by V_D the subspace

$$V_D := H^0(C, \mathcal{O}_C(D)) = \{f \in k(C) | \mathrm{div}(f) + D \geq 0\}.$$

The collection of all Cartier divisors D on C is a partially ordered set, where

$$D' \geq D \text{ if and only if } D' - D \text{ is effective.}$$

The system of subspaces V_D of $k(C)$ is compatible for this partial order, i.e., if $D' \geq D$ then $V_{D'} \supset V_D$. And K is the union of all the subspaces V_D, i.e., it is the colimit of this compatible system of finite dimensional k-vector spaces. Thus for all k-multilinear algebra operations which commute with colimits, the operation on $k(C)$ can be understood in terms of its restrictions to the finite dimensional subspaces $k(C)$. The next lemma makes this more concrete for the polynomial map F.

LEMMA 2.9. *Let C be a smooth, projective, connected curve over a field k and let*

$$F_i \in k(C)[X_0, \ldots, X_n]_{d_i}, \quad i = 1, \ldots, r$$

be a collection of polynomials in the spaces $k(C)[X_0, \ldots, X_n]_{d_i}$ of homogeneous, degree d_i polynomials. There exists an effective, Cartier divisor P on C and for every $i = 1, \ldots, r$ there exists a global section $F_{C,i}$ of the coherent sheaf

$$\mathcal{O}_C(P)[X_0, \ldots, X_n]_{d_i}$$

such that for every $i = 1, \ldots, r$ the germ of $F_{C,i}$ at the generic point of C equals F_i.

REMARK 2.10. In particular, for every Cartier divisor D on C and for every $i = 1, \ldots, r$ there is a homogeneous, degree d, polynomial map of k-vector spaces

$$F_{C,D,i} : V_D^{\oplus(n+1)} \to W_{d_i,P,D}, \quad W_{d_i,P,D} := V_{d_i D+P},$$

such that for every $i = 1, \ldots, r$ the restriction of F_i to $V_D^{\oplus(n+1)}$ equals $F_{C,D,i}$ considered as a map with target K (rather than the subspace V_{dD+P}).

PROOF. The coefficients of each F_i are rational functions on C. Each such function has a polar divisor. Since there are only finitely many coefficients of the finitely many polynomials $F_1, \ldots, F_r$, there exists a single effective, Cartier divisor P on C such that every coefficient is a global section of $\mathcal{O}_C(P)$. □

Because of Lemma 2.9, the original polynomial maps $F_1, \ldots, F_r$ can be understood in terms of their restrictions to the subspaces V_D. The dimensions of these subspaces are determined by the Riemann-Roch theorem.

THEOREM 2.11 (Riemann-Roch for smooth, projective curves). *Let k be a field. Let C be a smooth, projective, connected curve over k. Denote by $\omega_{C/k}$ the sheaf of relative differentials of C over k and denote by $g(C) = \text{genus}(C)$ the unique integer such that $\deg(\omega_{C/k}) = 2g(C) - 2$. For every invertible sheaf $\mathcal{L}$ on C,*

$$h^0(C, \mathcal{L}) - h^0(C, \omega_C \otimes_{\mathcal{O}_C} \mathcal{L}^\vee) = \deg(\mathcal{L}) + 1 - g(C).$$

REMARK 2.12. In particular, if $\deg(\mathcal{L}) > \deg(\omega_C) = 2g(C) - 2$ so that $\omega_C \otimes_{\mathcal{O}_C} \mathcal{L}^\vee$ has negative degree, then $h^0(C, \omega_C \otimes_{\mathcal{O}_C} \mathcal{L}^\vee)$ equals zero. And then

$$h^0(C, \mathcal{L}) = \deg(\mathcal{L}) + 1 - g(C).$$

For a Cartier divisor D satisfying

$$\deg(D) > 2g(C) - 2 \text{ and for each } i = 1, \ldots, r, d_i \deg(D) + \deg(P) > 2g(C) - 2,$$

the Riemann-Roch theorem gives that $V_D^{\oplus(n+1)}$ and $W_{d_i,P,D}$ are finite dimensional k-vector spaces of respective dimensions,

$$\dim_k(V_D^{\oplus(n+1)}) = (n+1)h^0(C, \mathcal{O}_C(D)) = (n+1)(\deg(D) + 1 - g)$$

and

$$\dim_k(W_{d_i,P,D}) = \dim(V_{d_iD+P}) = d_i \deg(D) + \deg(P) + 1 - g.$$

In this case, choosing a basis for $W_{d_i,P,D}$ and decomposing

$$F_{C,D,i} : V_D^{\oplus(n+1)} \to W_{d_i,P,D}$$

into its associated components, there exist $\dim_k(W_{d_i,P,D})$ homogeneous, degree d, polynomial functions

$$(F_{C,D,i})_j : V_D^{\oplus(n+1)} \to k, \quad j = 1, \ldots, \dim_k(W_{d_i,P,D})$$

such that a zero of $F_{C,D,i}$ is precisely the same as a common zero of all the functions $(F_{C,D,i})_j$.

PROOF OF PROPOSITION 2.7. By hypothesis, each d_i and $n + 1 - \sum_{i=1}^r d_i^{m+1}$ are nonzero so that the fractions

$$\frac{2g(C) - 2 - \deg(P)}{d_i} \text{ for each } i = 1, \ldots, r,$$

and

$$\frac{(n + 1 - \sum_{i=1}^r d_i^m)(g-1) + \sum_{i=1}^r d_i^m \deg(P)}{n + 1 - \sum_{i=1}^r d_i^{m+1}}$$

are all defined. Let D be an effective, Cartier divisor on C such that

$$\deg(D) > 2g(C) - 2, \ \deg(D) > \frac{2g(C) - 2 - \deg(P)}{d_i}, i = 1, \ldots, r, \text{ and}$$

$$\deg(D) > \frac{(n + 1 - \sum_{i=1}^r d_i^m)(g-1) + \sum_{i=1}^r d_i^m \deg(P)}{n + 1 - \sum_{i=1}^r d_i^{m+1}}.$$

Because $\deg(D) > 2g(C) - 2$, the Riemann-Roch theorem states that

$$\dim_k(V_D^{\oplus(n+1)}) = (n+1)\dim_k(V_D) = (n+1)(\deg(D) + 1 - g).$$

For every $i = 1, \dots, r$, because d_i is positive and because $\deg(D) > (2g(C) - 2 - \deg(P))/d_i$, also

$$\deg(d_i D + P) = d_i \deg(D) + \deg(P) \text{ is greater than } 2g(C) - 2.$$

Thus the Riemann-Roch theorem states that

$$\dim_k(W_{d_i,P,D}) = \dim_k(V_{d_i D+P}) = d_i \deg(D) + \deg(P) + 1 - g(C).$$

Thus for the collection of polynomial functions $(F_{C,D,i})_j$,

$$\dim_k(V_D^{\oplus(n+1)}) - \sum_{i=1}^{r} \sum_j \deg((F_{C,D,i})_j)^m$$

equals

$$(n+1)(\deg(D) + 1 - g) - \sum_{i=1}^{r} (d_i \deg(D) + \deg(P) + 1 - g(C)) d_i^m =$$

$$(n + 1 - \sum_{i=1}^{r} d_i^{m+1}) \deg(D) - [(n + 1 - \sum_{i=1}^{r} d_i^m)(g-1) + \sum_{i=1}^{r} d_i^m \deg(P)].$$

Because

$$\deg(D) > \frac{(n + 1 - \sum_{i=1}^{r} d_i^m)(g-1) + \sum_{i=1}^{r} d_i^m \deg(P)}{n + 1 - \sum_{i=1}^{r} d_i^{m+1}}$$

and because $n + 1 - \sum_{i=1}^{r} d_i^{m+1}$ is positive, also

$$(n + 1 - \sum_{i=1}^{r} d_i^{m+1}) \deg(D) > [(n + 1 - \sum_{i=1}^{r} d_i^m)(g-1) + \sum_{i=1}^{r} d_i^m \deg(P)].$$

Therefore

$$\dim_k(V_D^{\oplus(n+1)}) \text{ is greater than } \sum_{i=1}^{r} \sum_j \deg((F_{i,C,D})_j)^m.$$

Because of the inequality above, and because k has property C_m, there is a nontrivial common zero of the collection of homogeneous polynomial functions $(F_{C,D,i})_j$, $i = 1, \dots, r$, $j = 1, \dots, \dim_k(W_{d_i,P_i,D})$. Therefore there is a nontrivial common zero of the collection of homogeneous polynomial maps $F_{C,D,i}$, $i = 1, \dots, r$. By Lemma 2.9, the image of this nonzero element in $K^{\oplus(n+1)}$ is a nonzero element which is a common zero of the polynomials $F_1, \dots, F_r$. □

Proposition 2.7 is the main step in the proof of the Tsen-Lang theorem.

THEOREM 2.13 (The Tsen-Lang Theorem). [**Lan52**] *Let K/k be a field extension with finite transcendence degree,* $\text{tr.deg.}(K/k) = t$. *If k has property C_m then K has property C_{m+t}.*

PROOF. The proof of the theorem is by induction on t. When $t = 0$, i.e., when K/k is algebraic, the result follows from Lemma 2.6. Thus assume $t > 0$ and the result is known for $t - 1$. Let $(b_1, \dots, b_t)$ be a transcendence basis for K/k. Let E_t, resp. E_{t-1}, denote the subfield of K generated by k and $b_1, \dots, b_t$, resp. generated by k and $b_1, \dots, b_{t-1}$. Since E_{t-1}/k has transcendence degree $t - 1$, by the induction hypothesis E_{t-1} has property C_{m+t-1}. Now E_t/E_{t-1} is a purely transcendental extension of transcendence degree 1. In particular, it is finitely

generated and separable. Since E_{t-1} has property C_{m+t-1}, by Proposition 2.7 E_t has property C_{m+t}. Finally by Lemma 2.6 again, since K/E_t is algebraic and E_t has property C_{m+t}, also K has property C_{m+t}. □

The homogeneous version of the Nullstellensatz implies a field k has property C_0 if and only if k is algebraically closed. Thus one corollary of Theorem 2.13 is the following.

COROLLARY 2.14. *Let k be an algebraically closed field and let K/k be a field extension of finite transcendence degree t. The field K has property C_t.*

In particular, the case $t = 1$ is historically the first result in this direction.

COROLLARY 2.15 (Tsen's theorem). [**Tse36**] *The function field of a curve over an algebraically closed field has property C_1.*

Chevalley and Tsen recognized that property C_1, which they called *quasi-algebraic closure*, has an important consequence for division algebras. Lang recognized that property C_2 also has an important consequence for division algebras, cf. [**Lan52**, Theorem 13]. Let K be a field. A *division algebra with center K* is a K-algebra D with center K such that every nonzero element of D has a (left-right) inverse. Although this is not always the case, we will also demand that $\dim_K(D)$ is finite.

Denote by $\overline{K}$ the separable closure of K. Every division algebra with center K is an example of a *central simple algebra over K*, i.e., a K-algebra A with center K and $\dim_K(A)$ finite such that $A \otimes_K \overline{K}$ is isomorphic as a $\overline{K}$-algebra to the algebra $\mathrm{Mat}_{n\times n}(\overline{K})$ of $n \times n$ matrices with entries in $\overline{K}$ for some positive integer n. In particular, $\dim_K(A) = n^2$ for a unique positive integer n. For a division algebra D with center K, the unique positive integer n is called the *index* of D.

Let $\phi : A \otimes_K \overline{K} \to \mathrm{Mat}_{n\times n}(\overline{K})$ be an isomorphism of $\overline{K}$-algebras. There is an induced homogeneous, degree n, polynomial map of $\overline{K}$-vector spaces

$$\det \circ \phi : A \to \mathrm{Mat}_{n\times n}(\overline{K}) \to \overline{K}.$$

By the Skolem-Noether theorem, every other isomorphism

$$\phi' : A \otimes_K \overline{K} \to \mathrm{Mat}_{n\times n}(\overline{K})$$

is of the form $\mathrm{conj}_a \circ \phi$ where $a \in \mathrm{Mat}_{n\times n}(\overline{K})$ is an invertible element and

$$\mathrm{conj}_a : \mathrm{Mat}_{n\times n}(\overline{K}) \to \mathrm{Mat}_{n\times n}(\overline{K}), \quad \mathrm{conj}_a(b) = aba^{-1}$$

is conjugation by a. But $\det \circ \mathrm{conj}_a$ equals $\det$. Thus the map $\det \circ \phi$ is independent of the particular choice of ϕ. Since the Galois group of $\overline{K}/K$ acts on the polynomial map through its action on ϕ, the polynomial map is also Galois invariant. Therefore there exists a unique homogeneous, degree n, polynomial map of K-vector spaces

$$\mathrm{Nrm}_{A/K} : A \to K$$

such that for every isomorphism of $\overline{K}$-algebras ϕ, $\det \circ \phi$ equals $\mathrm{Nrm}_{A/K} \otimes 1$.

The homogeneous, polynomial map of K-vectors spaces $\mathrm{Nrm}_{A/K}$ is the *reduced norm* of A. It is multiplicative, i.e.,

$$\forall a, b \in A, \quad \mathrm{Nrm}_{A/K}(ab) = \mathrm{Nrm}_{A/K}(a)\mathrm{Nrm}_{A/K}(b).$$

And the restriction to the center K is the polynomial map $\lambda \mapsto \lambda^n$. These properties characterize the reduced norm. By the same type of Galois invariance argument as above, and using Cramer's rule, an element a of A has a (left and right) inverse if and only if $\mathrm{Nrm}_{A/K}(a)$ is nonzero. In particular, if D is a division algebra the only zero of $\mathrm{Nrm}_{A/K}$ is $a = 0$.

PROPOSITION 2.16. *Let K be a field*

(i) *If K has property C_1, then the only division algebra with center K is K itself.*
(ii) *If K has property C_2 then for every division algebra D with center K the reduced norm map*

$$\mathrm{Nrm}_{D/K} : D \to K$$

is surjective.

PROOF. Let D be a division algebra with center K. Denote by n the index of D. Because $\mathrm{Mat}_{n\times n}(\overline{K})$ has dimension n^2 as a $\overline{K}$-vector space, also D has dimension n^2 as a K-vector space. If K has property C_1, then since the homogeneous polynomial map $\mathrm{Nrm}_{D/K}$ has only the trivial zero,

$$n = \deg(\mathrm{Nrm}_{D/K}) \geq \dim_K(D) = n^2,$$

i.e., $n = 1$. Thus for a field K with property C_1, the only finite dimensional, division algebra with center K has dimension 1, i.e., D equals K.

Next suppose that K has property C_2. Clearly $\mathrm{Nrm}_{D/k}(0)$ equals 0. Thus to prove that

$$\mathrm{Nrm}_{D/K} : D \to K$$

is surjective, it suffices to prove that for every nonzero $c \in K$ there exists b in D with $\mathrm{Nrm}_{D/K}(b) = c$. Consider the homogeneous, degree n, polynomial map

$$F_c : D \oplus K \to K, \quad (a, \lambda) \mapsto \mathrm{Nrm}_{D/K}(a) - c\lambda^n.$$

Since

$$\dim_K(D \oplus K) = n^2 + 1 > \deg(F_c)^2,$$

by property C_2 the map F_c has a zero $(a, \lambda) \neq (0, 0)$, i.e., $\mathrm{Nrm}_{D/k}(a) = c\lambda^n$. In particular, λ must be nonzero since otherwise a is a nonzero element of D with $\mathrm{Nrm}_{D/K}(a) = 0$. But then $b = (1/\lambda)a$ is an element of D with $\mathrm{Nrm}_{D/k}(b) = c$. □

It was later recognized, particularly through the work of Merkurjev and Suslin, that these properties of division algebras are equivalent to properties of Galois cohomology. The *cohomological dimension* of a field K is the smallest integer $\mathrm{cd}(K)$ such that for every Abelian, discrete, torsion Galois module A and for every integer $m > \mathrm{cd}(K)$,

$$H^m(\overline{K}/K, A) = \{0\}.$$

THEOREM 2.17. [**Ser02**, Proposition 5, §I.3.1], [**Sus84**, Corollary 24.9] *Let K be a field.*

(i) *The cohomological dimension of K is ≤ 1 if and only if for every finite extension L/K, the only division algebra with center L is L itself.*
(ii) *If K is perfect, the cohomological dimension of K is ≤ 2 if and only if for every finite extension L/K, for every division algebra D with center L, the reduced norm map $\mathrm{Nrm}_{D/L}$ is surjective.*

3. Rationally connected varieties

The theorems of Chevalley-Warning and Tsen-Lang are positive answers to Problem 2.1 for a certain class of fields. It is natural to ask whether these theorems can be generalized for such fields.

PROBLEM 3.1. Let r be a nonnegative integer. Give sufficient geometric conditions on a variety such that for every C_r field K (or perhaps every C_r field satisfying some additional hypotheses) and for every K-variety satisfying the conditions, X has a K-point.

As with Problem 2.1, this problem is quite vague. Nonetheless there are important partial answers. One such answer, whose proof was sketched in the lectures of Hassett in this same Clay Summer School, is the following.

THEOREM 3.2. [**Man86**] [**CT87**] *Let K be a C_1 field and let X be a projective K-variety. If $X \otimes_K \overline{K}$ is birational to $\mathbb{P}^2_{\overline{K}}$ then X has a K-point.*

This begs the question: What (if anything) is the common feature of rational surfaces and of the varieties occurring in the Chevalley-Warning and Tsen-Lang theorems, i.e., complete intersections in $\mathbb{P}^n$ of hypersurfaces of degrees $d_1, \dots, d_r$ with $d_1 + \cdots + d_r \leq n$? One answer is *rational connectedness*. This is a property that was studied by Kollár-Miyaoka-Mori and Campana, cf. [**Kol96**].

DEFINITION 3.3. Let k be an algebraically closed field. An integral (thus nonempty), separated, finite type, k-scheme X is *rationally connected*, resp. *separably rationally connected*, if there exists an integral, finite type k-scheme M and a morphism of k-schemes

$$u : M \times_k \mathbb{P}^1_k \to X, \quad (m, t) \mapsto u(m, t)$$

such that the induced morphism of k-schemes

$$u^{(2)} : M \times_k \mathbb{P}^1_k \times_k \mathbb{P}^1_k \to X \times_k X, \quad (m, t_1, t_1) \mapsto (u(m, t_1), u(m, t_2))$$

is surjective, resp. surjective and generically smooth.

In a similar way, X is *rationally chain connected*, resp. *separably rationally chain connected*, if for some integer $m \geq 1$, the analogous property holds after replacing $\mathbb{P}^1_k$ by the proper, connected, nodal, reducible curve C_m which is a chain of m smooth rational curves.

Figure 1 shows a rationally connected variety, where every pair of points is contained in an image $u(\mathbb{P}^1)$ of the projective line.

The definition of rational connectedness, resp. rational chain connectedness, mentions a particular parameter space M. However, using the general theory of Hilbert schemes, it suffices to check that every pair (x_1, x_2) of K-points of $X \otimes_k K$ is contained in some rational K-curve, resp. a chain of rational K-curves, (not necessarily from a fixed parameter space) for one *sufficiently large*, algebraically closed, field extension K/k, i.e., for an algebraically closed extension K/k such that for every countable collection of proper closed subvarieties $Y_i \subsetneq X$, there exists a K-point of X contained in none of the sets Y_i. For instance, K/k is sufficiently large if K is uncountable or if K/k contains the fraction field $k(X)/k$ as a subextension.

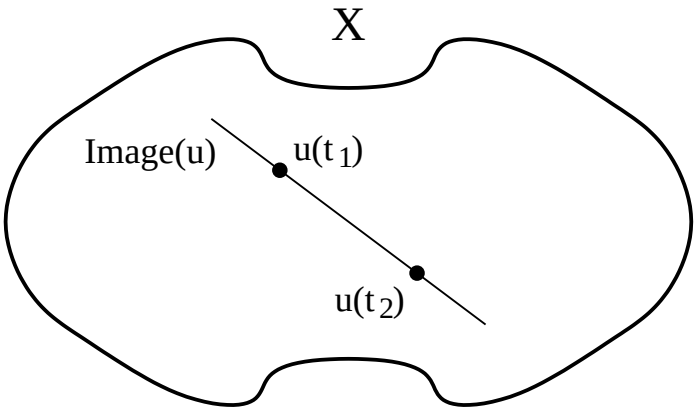


FIGURE 1. Every pair of points in a rationally connected variety lies in an image of the projective line.

A very closely related property is the existence of a very free rational curve. For a d-dimensional variety X, a *very free rational curve* is a morphism

$$f : \mathbb{P}^1_k \to X_{\text{smooth}}$$

into the smooth locus of X such that f^*T_X is ample, i.e.,

$$f^*T_X \cong \mathcal{O}_{\mathbb{P}^1_k}(a_1) \oplus \cdots \oplus \mathcal{O}_{\mathbb{P}^1_k}(a_d),\ a_1, \ldots, a_d > 0.$$

DEFINITION 3.4. Let k be a field and let X be a quasi-projective k-scheme. Denote by X_{smooth} the smooth locus of X. The *very free locus* $X_{\text{v.f.}}$ of X is the union of the images in X_{smooth} of all very free rational curves to $X_{\text{smooth}} \otimes_k K$ as K/k varies over all algebraically closed extensions. More generally, for a flat, quasi-projective morphism of schemes,

$$\pi : X \to B,$$

denoting by $X_{\pi,\text{smooth}}$ the smooth locus of the morphism π, the *very free locus* $X_{\pi,\text{v.f.}}$ is the union in $X_{\pi,\text{smooth}}$ of the images of every very free rational curve in every geometric fiber of $X_{\pi,\text{smooth}}$ over B.

The next theorem explains the relation of these different properties.

THEOREM 3.5. [**Kol96**, §IV.3], [**HT06**], *Unless stated otherwise, all varieties below are d-dimensional, reduced, irreducible, quasi-projective schemes over an algebraically closed field k.*

(0) *In characteristic* 0*, every rationally connected variety is separably rationally connected.*

(i) *For every flat, proper morphism $\pi : X \to B$ (not necessarily of quasi-projective varieties over a field), the subset of B parameterizing points whose geometric fiber is rationally chain connected is stable under specialization. (If one bounds the degree of the chains with respect to a relatively ample invertible $\mathcal{O}_X$-module, then it is a closed subset.)*

(ii) *The very free locus of a quasi-projective variety is open. More generally, for every flat, quasi-projective morphism, $\pi : X \to B$, the subset $X_{\pi,\text{v.f.}}$ of $X_{\pi,\text{smooth}}$ is an open subset.*

(iii) *The very free locus $X_{\text{v.f.}}$ of a quasi-projective variety is (separably) rationally connected in the following strong sense. For every positive integer N, for every positive integer m, and for every positive integer a, for every collection of distinct closed points $t_1, \dots, t_N \in \mathbb{P}^1_k$, for every collection of closed points $x_1, \dots, x_N \in X_{\text{v.f.}}$, and for every specification of an m-jet of a smooth curve in X at each point x_i, there exists a morphism*

$$f : \mathbb{P}^1_k \to X_{\text{v.f.}}$$

such that for every $i = 1, \dots, N$, f is unramified at t_i, $f(t_i)$ equals x_i and the m-jet of t_i in $\mathbb{P}^1_k$ maps isomorphically to the specified m-jet at x_i, and

$$f^*T_X \cong \mathcal{O}_{\mathbb{P}^1_k}(a_1) \oplus \cdots \oplus \mathcal{O}_{\mathbb{P}^1_k}(a_d),\ a_1, \dots, a_d \geq a.$$

(iv) *Every rational curve in X_{smooth} intersecting $X_{\text{v.f.}}$ is contained in $X_{\text{v.f.}}$. Thus for every smooth, rationally chain connected variety, if X contains a very free rational curve then $X_{\text{v.f.}}$ equals all of X.*

(v) *A proper variety X is rationally chain connected if it is* generically *rationally chain connected, i.e., if there exists a morphism u as in the definition such that $u^{(2)}$ is dominant (but not necessarily surjective).*

(vi) *For the morphism $u : M \times_k \mathbb{P}^1_k \to X$, let l be a closed point of M such that $u_l : \mathbb{P}^1 \to X$ has image in X_{smooth} and such that $u^{(2)}$ is smooth at (l, t_1, t_2) for some $t_1, t_2 \in \mathbb{P}^1_k$. Then the morphism u_l is very free. Thus an irreducible, quasi-projective variety X contains a very free curve if and only if there is a separably rationally connected open subset of X_{smooth}. Also, a smooth, quasi-projective variety X in characteristic 0 which is generically rationally connected contains a very free morphism.*

(vii) *For a surjective morphism $f : X \to Y$ of varieties over an algebraically closed field, if X is rationally connected, resp. rationally chain connected, then also Y is rationally connected, resp. rationally chain connected.*

(viii) *For a birational morphism $f : X \to Y$ of proper varieties over an algebraically closed field, if Y is rationally connected then X is rationally chain connected. If the characteristic is zero, then X is rationally connected.*

REMARK 3.6. Item (ii) is proved in Proposition 3.18. The *generic case* of Item (iii), which is all we will need, is proved in Proposition 3.19. The complete result was proved by Hassett and Tschinkel, [**HT06**]. Item (iv) follows from Corollary 3.20. The remaining items are not proved, nor are they used in the proof of the main theorem. For the most part they are proved by similar arguments; complete proofs are in [**Kol96**, §IV.3].

Rational connectedness is analogous to path connectedness in topology, and satisfies the analogues of many properties of path connectedness. One property of path connectedness is this: for a fibration of CW complexes, if the base space and the fibers are path connected, then also the total space is path connected. This led to two conjectures by Kollár, Miyaoka and Mori.

CONJECTURE 3.7. [**Kol96**, Conjecture IV.5.6] Let $\pi : X \to B$ be a surjective morphism of smooth, projective schemes over an algebraically closed field of characteristic 0. If both B and a general fiber of π are rationally connected, then X is also rationally connected.

Conjecture 3.7 is implied by the following conjecture about rationally connected fibrations over curves.

CONJECTURE 3.8. [**Kol96**, Conjecture IV.6.1.1] Let $\pi : X \to B$ be a surjective morphism of projective schemes over an algebraically closed field of characteristic 0. If B is a smooth curve and if a general fiber of f is rationally connected, then there exists a morphism $s : B \to X$ such that $\pi \circ s$ equals Id_B, i.e., s is a section of π.

Our next goal is to prove the following result.

THEOREM 3.9. [**GHS03**] *Conjecture 3.8 of Kollár-Miyaoka-Mori is true. Precisely, let k be an algebraically closed field of characteristic 0 and let $\pi : X \to B$ be a surjective morphism from a normal, projective k-scheme X to a smooth, projective, connected k-curve B. If the geometric generic fiber $X_{\overline{\eta}_B}$ is a normal, integral scheme whose smooth locus contains a very free curve, then there exists a morphism $s : B \to X$ such that $\pi \circ s$ equals Id_B.*

This was generalized by A. J. de Jong to the case that k is algebraically closed of arbitrary characteristic, [**dJS03**]. The key difference has to do with extensions of valuation rings in characteristic 0 and in positive characteristic. Given a flat morphism of smooth schemes in characteristic 0, $\pi : U \to B$, and given codimension 1 points η_D of U and η_Δ of B with $\pi(\eta_D) = \eta_\Delta$, the induced local homomorphism of stalks $\pi_U^* : \widehat{\mathcal{O}}_{B,\eta_\Delta} \to \widehat{\mathcal{O}}_{U,\eta_D}$, is equivalent to

$$k(\Delta)\,[\![t]\!] \to k(D)\,[\![r]\!]\,, \quad t \mapsto ur^m$$

for a unit u and a positive integer m, cf. the proof of Lemma 3.24 below. In particular, it is *rigid* in the sense that $t \mapsto ur^m + vr^{m+1} + \dots$ is equivalent to $t \mapsto ur^m$. However, extensions of positive characteristic valuation rings are not rigid, e.g., $t \mapsto r^p + v_1 r^{p+1}$ is equivalent to $t \mapsto r^p + v_2 r^{p+1}$ only if $v_1 = v_2$. But there is a weak rigidity of local homomorphisms, Krasner's lemma in the theory of non-Archimedean valuations. This is a key step in the generalization to positive characteristic.

Of course when k has characteristic 0, then since X is normal the fiber $X_{\overline{\eta}_B}$ is automatically normal. If X is also smooth (which can be achieved thanks to resolution of singularities in characteristic 0), then also $X_{\overline{\eta}_B}$ is smooth. Then the hypothesis on $X_{\overline{\eta}_B}$ is equivalent to rational connectedness.

3.1. Outline of the proof. The proof that follows is based on a proof by T. Graber, J. Harris and myself (not quite the version we chose to publish) together with several major simplifications due to A. J. de Jong. The basic idea is to choose a smooth curve $C \subset X$ such that $\pi|_C : C \to B$ is finite, and then try to deform C as a curve in X until it specializes to a reducible curve in X, one component of which is the image of a section s of π. Here are some definitions that make this precise.

DEFINITION 3.10. Let $\pi_C : C \to B$ be a finite morphism of smooth, projective k-curves. A *linked curve with handle C* is a reduced, connected, projective curve C_{link} with irreducible components

$$C_{\mathrm{link}} = C \cup L_1 \cup \cdots \cup L_m$$

together with a morphism

$$\pi_{C,\text{link}} : C_{\text{link}} \to B$$

such that

(i) $\pi_{C,\text{link}}$ restricts to π_C on the component C,
(ii) the restriction of $\pi_{C,\text{link}}$ to each *link* component L_i is a constant morphism with image b_i, where $b_1, \dots, b_m$ are distinct closed points of B,
(iii) and each link L_i is a smooth, rational curve intersecting C in a finite number of nodes of C_{link}.

If every link L_i intersects C in a single node of C_{link}, then $(C_{\text{link}}, \pi_{C,\text{link}})$ is called a *comb* and the links L_i are called *teeth*. For combs we will use the notation C_{comb} rather than C_{link}.

A *one-parameter deformation* of a linked curve $(C_{\text{link}}, \pi_{C,\text{link}})$ is a datum of a smooth, connected, pointed curve $(\Pi, 0)$ and a projective morphism

$$(\rho, \pi_{\mathcal{C}}) : \mathcal{C} \to \Pi \times_k B$$

such that ρ is flat and such that $\mathcal{C}_0 := \rho^{-1}(0)$ together with the restriction of $\pi_{\mathcal{C}}$ equals the linked curve $(C_{\text{link}}, \pi_{C,\text{link}})$.

A one-parameter deformation *specializes to a section curve* if there exists a closed point $\infty \in \Pi$ and an irreducible component B_i of $\mathcal{C}_\infty := \rho^{-1}(\infty)$ such that

(i) $\mathcal{C}_\infty$ is reduced at the generic point of B_i
(ii) and the restriction of $\pi_{\mathcal{C}}$ to B_i is an isomorphism

$$\pi_{\mathcal{C}}|_{B_i} : B_i \xrightarrow{\cong} B.$$

Given a linked curve, a one-parameter deformation of the linked curve and a B-morphism $j : C_{\text{link}} \to X$, an *extension* of j is an open neighborhood of 0, $0 \in N \subset B$ and a B-morphism

$$j_N : \mathcal{C}_N \to X, \quad \mathcal{C}_N := \rho^{-1}(N)$$

restricting to j on $\mathcal{C}_0 = C_{\text{link}}$.

Figure 2 shows a linked curve with some links intersecting the handle in more than 1 point. And Figure 3 shows a comb, where every tooth intersects the handle precisely once.

For the purposes of producing a section, the particular parameter space $(\Pi, 0)$ of the one-parameter deformation is irrelevant. Thus, it is allowed to replace the one-parameter deformation by the new one-parameter deformation obtained from a finite base change $(\Pi', 0') \to (\Pi, 0)$. The following lemma is straightforward.

LEMMA 3.11. *Let* $(\Pi, 0, \infty)$ *together with* $(\rho, \pi_{\mathcal{C}}) : \mathcal{C} \to \Pi \times_k B$ *be a one-parameter deformation of* $(C_{\text{link}}, \pi_{C,\text{link}})$ *specializing to a section curve* B_i. *For every morphism of* 2*-pointed, smooth, connected curves*

$$(\Pi', 0', \infty') \to (\Pi, 0, \infty),$$

the base change morphism

$$\Pi' \times_\Pi \mathcal{C} \to \Pi' \times_k B$$

is also a one-parameter deformation of $(C_{\text{link}}, \pi_{C,\text{link}})$ *specializing to the section curve* B_i.

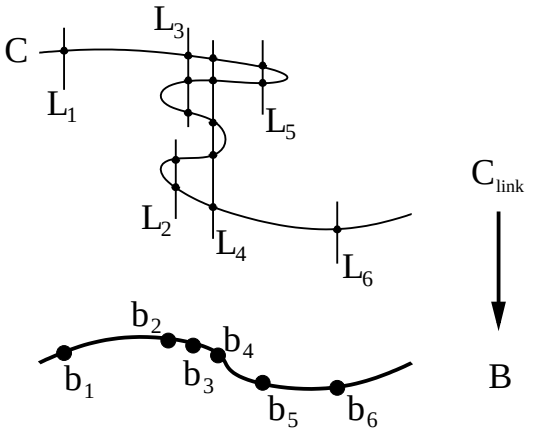


FIGURE 2. A linked curve with handle C and some links intersecting C in more than 1 point.

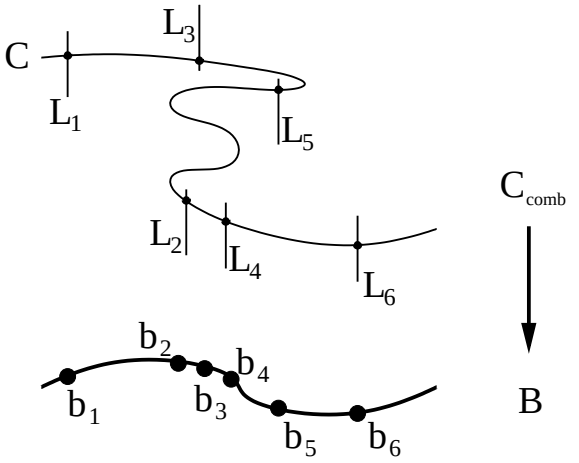


FIGURE 3. A comb is a linked curve where every link, or tooth, intersects the handle precisely once.

The usefulness of these definitions is the following simple consequence of the valuative criterion of properness.

LEMMA 3.12. *Let $(C_{\text{link}}, \pi_{\text{link}})$ be a linked curve together with a B-morphism $j : C_{\text{link}} \to X$. If there exists a one-parameter deformation of the linked curve specializing to a section curve and if there exists an extension of j, then there exists a section $s : B \to X$ of π.*

PROOF. Let R denote the stalk $\mathcal{O}_{\mathcal{C},\eta_{B_i}}$ of $\mathcal{O}_{\mathcal{C}}$ at the generic point η_{B_i} of B_i. By the hypotheses on $\mathcal{C}$ and B_i, R is a discrete valuation ring with residue field $\kappa = k(B_i)$ and fraction field $K = k(\mathcal{C})$. The restriction of j_N to the generic point of $\mathcal{C}$ is a B-morphism

$$j_K : \operatorname{Spec} K \to X.$$

Because $\pi : X \to B$ is proper, by the valuative criterion of properness the B-morphism j_K extends to a B-morphism

$$j_R : \operatorname{Spec} R \to X,$$

which in turn gives a B-morphism from the residue field $\operatorname{Spec} \kappa$ to X, i.e., a rational B-map

$$j_{B_i} : B_i \supset U \to X, \quad U \subset B_i \text{ a dense, Zariski open.}$$

Finally, because B_i is a smooth curve, the valuative criterion applies once more and this rational transformation extends to a B-morphism

$$j_{B_i} : B_i \to X.$$

Because $\pi_{\mathcal{C}}|_{B_i} : B_i \to B$ is an isomorphism, there exists a unique B-morphism

$$s : B \to X$$

such that $j_{B_i} = s \circ \pi_{\mathcal{C}}|_{B_i}$. The morphism s is a section of π. □

Thus the proof of the theorem breaks into three parts:

(i) find a "good" linked curve $j : C_{\text{link}} \to X$,
(ii) find a one-parameter deformation of the linked curve specializing to a section curve,
(iii) and find an extension of j to the one-parameter deformation.

The first step in finding $j : C_{\text{link}} \to X$ is to form a curve C_{init} which is an intersection of X with $\dim(X)-1$ general hyperplanes in projective space. By Bertini's theorem, if the hyperplanes are sufficiently general, then C_{init} will satisfy any reasonable transversality property. Moreover, there is a technique due to Kollár-Miyaoka-Mori—the smoothing combs technique—for improving C_{init} to another curve $C \subset X$ still satisfying the transversality property and also satisfying a positivity property with respect to the vertical tangent bundle of $\pi : X \to B$.

Unfortunately, even after such an improvement, there may be no one-parameter deformation of $\pi|_C : C \to B$ specializing to a section curve. However, after attaching sufficiently many link components over general closed points of B, there does exist a one-parameter deformation of C_{link} specializing to a section curve. This is one aspect of the well-known theorem that for a fixed base curve B and for a fixed degree d, if the number β of branch points is sufficiently large the Hurwitz scheme of degree d covers of B with β branch points is irreducible. (This was proved by Hurwitz when $g(B) = 0$, [**Hur91**], proved by Richard Hamilton for arbitrary genus in his thesis, and periodically re-proved ever since, cf. [**GHS02**].) Because the general fibers of $\pi : X \to B$ are rationally connected, the inclusion $C \subset X$ extends to a B-morphism $j : C_{\text{link}} \to X$.

Finally the positivity property mentioned above implies j extends to the one-parameter deformation, at least after base change by a morphism $\Pi' \to \Pi$.

3.2. Hilbert schemes and smoothing combs. The smoothing combs technique of Kollár-Miyaoka-Mori depends on a result from the deformation theory of Hilbert schemes. Here is the setup for this result. Let $Y \to S$ be a flat, quasi-projective morphism and let

$$(\rho_{\text{Hilb}} : \operatorname{Hilb}(Y/S) \to S, \operatorname{Univ}(Y/S) \subset \operatorname{Hilb}(Y/S) \times_S T)$$

be universal among pairs $(\rho : T \to S, Z \subset T \times_S Y)$ of an S-scheme T and a closed subscheme $Z \subset T \times_S Y$ such that $Z \to T$ is proper, flat and finitely presented. In other words, $\mathrm{Hilb}(Y/S)$ is the *relative Hilbert scheme* of Y over S.

In particular, for every field K the K-valued points of $\mathrm{Hilb}(Y/S)$ are naturally in bijection with pairs (s, Z) of a K-valued point s of S and a closed subscheme Z of $Y_s := \{s\} \times_S Y$. The closed immersion $Z \to Y_s$ is a *regular embedding* if at every point of Z the stalk of the ideal sheaf $\mathcal{I}_{Z/Y_s}$ is generated by a regular sequence of elements in the stalk of $\mathcal{O}_{Y_s}$. In this case the *conormal sheaf* $\mathcal{I}_{Z/Y_s}/\mathcal{I}^2_{Z/Y_s}$ is a locally free $\mathcal{O}_Z$-module, and hence also the *normal sheaf*

$$\mathcal{N}_{Z/Y_s} := \mathit{Hom}_{\mathcal{O}_Z}(\mathcal{I}_{Z/Y_s}/\mathcal{I}^2_{Z/Y_s}, \mathcal{O}_S)$$

is a locally free $\mathcal{O}_Z$-module. The regular embeddings which arise in the proof of Theorem 3.9 are precisely closed immersions of at-worst-nodal curves in a smooth variety.

PROPOSITION 3.13. [**Kol96**, Thm. I.2.10, Lemma I.2.12.1, Prop. I.2.14.2] *If $Z \subset Y_s$ is a regular embedding and if $h^1(Z, \mathcal{N}_{Z/Y_s})$ equals 0, then $\mathrm{Hilb}(Y/S)$ is smooth over S at (s, Z).*

There is a variation of this proposition which is also useful. There is a *flag Hilbert scheme* of Y over S, i.e., a universal pair

$$(\rho_{\mathrm{fHilb}} : \mathrm{fHilb}(Y/S) \to S, \mathrm{Univ}_1(Y/S) \subset \mathrm{Univ}_2(Y/S) \subset \mathrm{Hilb}(Y/S) \times_S T)$$

among all pairs $(\rho : T \to S, Z_1 \subset Z_2 \subset T \times_S Y)$ of an S-scheme T and a nested pair of closed subschemes $Z_1 \subset Z_2 \subset T \times_S Y$ such that for $i = 1, 2$, the projection $Z_i \to T$ is proper, flat and finitely presented. There are obvious forgetful morphisms

$$F_i : \mathrm{fHilb}(Y/S) \to \mathrm{Hilb}(Y/S), \quad F_i(s, Z_1, Z_2) = (s, Z_i).$$

PROPOSITION 3.14. *Let K be a field and (s, Z_1, Z_2) a K-point of $\mathrm{fHilb}(Y/S)$. If each closed immersion $Z_1 \subset Z_2$ and $Z_2 \subset Y_s$ is a regular embedding,*

$$h^1(Z_2, \mathcal{N}_{Z_2/Y_s}) = 0, h^1(Z_1, \mathcal{N}_{Z_1/Z_2}) = 0,$$

and

$$h^i(Z_2, \mathcal{I}_{Z_1/Z_2} \cdot \mathcal{N}_{Z_2/Y_2}) = 0 \text{ for } i = 1, 2,$$

then $\mathrm{fHilb}(Y/S)$ is smooth over S at (s, Z_1, Z_2), for each $i = 1, 2$, $\mathrm{Hilb}(Y/S)$ is smooth over S at (s, Z_i), and each forgetful morphism $F_i : \mathrm{fHilb}(Y/S) \to \mathrm{Hilb}(Y/S)$ is smooth at (s, Z_1, Z_2).

PROOF. Since $h^1(Z_2, \mathcal{N}_{Z_2/Y_s})$ equals 0, $\mathrm{Hilb}(Y/S)$ is smooth at (s, Z_2) by Proposition 3.13. It is easy to see that the forgetful morphism F_2 is equivalent to the relative Hilbert scheme $\mathrm{Hilb}(\mathrm{Univ}(Y/S)/\mathrm{Hilb}(Y/S))$ over $\mathrm{Hilb}(Y/S)$. Thus, applying Proposition 3.13 to this Hilbert scheme, the vanishing of $h^1(Z_1, \mathcal{N}_{Z_1/Z_2})$ implies F_2 is smooth at (s, Z_1, Z_2). Since a composition of smooth morphisms is smooth, also $\mathrm{fHilb}(Y/S)$ is smooth over S at (s, Z_1, Z_2). The long exact sequence of cohomology associated to the short exact sequence

$$0 \longrightarrow \mathcal{I}_{Z_1/Z_2} \cdot \mathcal{N}_{Z_2/Y_s} \longrightarrow \mathcal{N}_{Z_2/Y_s} \longrightarrow \mathcal{N}_{Z_2/Y_s}|_{Z_1} \longrightarrow 0$$

implies that $h^1(Z_1, \mathcal{N}_{Z_2/Y_s}|_{Z_1})$ equals $h^1(Z_2, \mathcal{N}_{Z_2/Y_s})$, which is 0. Thus, the long exact sequence of cohomology associated to

$$0 \longrightarrow \mathcal{N}_{Z_1/Z_2} \longrightarrow \mathcal{N}_{Z_1/Y_s} \longrightarrow \mathcal{N}_{Z_2/Y_s}|_{Z_1} \longrightarrow 0$$

implies that $h^1(Z_2, \mathcal{N}_{Z_1/Y_2})$ equals 0. So again by Proposition 3.13, $\mathrm{Hilb}(Y/S)$ is smooth over S at (s, Z_1). Finally, F_1 is a morphism of smooth S-schemes at (s, Z_1, Z_2). Thus, to prove F_1 is smooth, it suffices to prove it is surjective on Zariski tangent vector spaces. This follows from the fact that

$$h^1(Z_2, \mathcal{I}_{Z_1/Z_2} \cdot \mathcal{N}_{Z_2/Y_s}) = 0.$$

□

Another ingredient in the smoothing combs technique is a simple result about *elementary transforms* of locally free sheaves on a curve: the higher cohomology of the sheaf becomes zero after applying elementary transforms at sufficiently many points.

LEMMA 3.15. *Let C be a projective, at-worst-nodal, connected curve over a field k and let $\mathcal{E}$ be a locally free $\mathcal{O}_C$-module.*

(i) *There exists a short exact sequence of coherent sheaves,*

$$0 \to \mathcal{F}^\vee \to \mathcal{E}^\vee \to \mathcal{T} \to 0$$

such that $\mathcal{T}$ is a torsion sheaf with support in C_{smooth} and such that $h^1(C, \mathcal{F})$ equals 0.

(ii) *Inside the parameter space of torsion quotients $q : \mathcal{E}^\vee \twoheadrightarrow \mathcal{T}$ with support in C_{smooth}, denoting*

$$\mathcal{F}^\vee := \mathrm{Ker}(\mathcal{E}^\vee \to \mathcal{T}) \text{ and } \mathcal{F} := \mathit{Hom}_{\mathcal{O}_C}(\mathcal{F}^\vee, \mathcal{O}_C),$$

the subset parameterizing quotients for which $h^1(C, \mathcal{F}) = 0$ is an open subset.

(iii) *If $h^1(C, \mathcal{F})$ equals 0, then for every short exact sequence of coherent sheaves*

$$0 \to \mathcal{G}^\vee \to \mathcal{E}^\vee \xrightarrow{q'} \mathcal{S} \to 0$$

admitting a morphism $r : \mathcal{S} \to \mathcal{T}$ of torsion sheaves with support in C_{smooth} for which $q = r \circ q'$, $h^1(C, \mathcal{G})$ equals 0.

PROOF. **(i)** By Serre's vanishing theorem, there exists an effective, ample divisor D in the smooth locus of C such that $h^1(C, \mathcal{E}(D))$ equals 0. Define $\mathcal{F} = \mathcal{E}(D)$, define $\mathcal{E} \to \mathcal{F}$ to be the obvious morphism $\mathcal{E} \to \mathcal{E}(D)$, and define $\mathcal{T}$ to be the cokernel of $\mathcal{F}^\vee \to \mathcal{E}^\vee$.

(ii) This follows immediately from the semicontinuity theorem, cf. [**Har77**, §III.12].

(iii) There exists an injective morphism of coherent sheaves $\mathcal{F} \to \mathcal{G}$ with torsion cokernel. Because $h^1(C, \mathcal{F})$ equals 0 and because h^1 of every torsion sheaf is zero, the long exact sequence of cohomology implies that also $h^1(C, \mathcal{G})$ equals 0. □

It is worth noting one interpretation of the sheaf $\mathcal{F}$ associated to a torsion quotient $\mathcal{T}$. Assume that $\mathcal{T}$ is isomorphic to a direct sum of skyscraper sheaves at n distinct points $c_1, \ldots, c_n$ of C_{smooth}. (Inside the parameter space of torsion quotients, those with this property form a dense, open subset.) For each point c_i, the linear functional $\mathcal{E}^\vee|_{c_i} \twoheadrightarrow \mathcal{T}|_{c_i}$ gives a one-dimensional subspace $\mathrm{Hom}_k(\mathcal{T}|_{c_i}, k) \hookrightarrow \mathcal{E}|_{c_i}$. The sheaf $\mathcal{F}$ is precisely the sheaf of rational sections of $\mathcal{E}$ having at worst a simple pole at each point c_i in the direction of this one-dimensional subspace of $\mathcal{E}|_{c_i}$. This is

often called an *elementary transform up* of $\mathcal{E}$ at the point c_i in the specified direction. So Lemma 3.15 says that h^1 becomes zero after sufficiently many elementary transforms up at general points in general directions.

This interpretation is useful because the normal sheaf of a reducible curve can be understood in terms of elementary transforms up. To be precise, let Y be a k-scheme, let C be a proper, nodal curve, let C_0 be a closed subcurve (i.e., a union of irreducible components of C), and let $j : C \to Y$ be a regular embedding such that Y is smooth at every node $p_1, \dots, p_n$ of C which is contained in C_0 and which is not a node of C_0. Then $j_0 : C_0 \to Y$ is also a regular embedding and both $\mathcal{N}_{C/Y}|_{C_0}$ and $\mathcal{N}_{C_0/Y}$ are locally free sheaves on C_0. For each i, there is a branch C_i of C at p_i other than C_0. Denote by T_{C_i,p_i} the tangent direction of this branch in T_{Y,p_i}.

LEMMA 3.16. [**GHS03**, Lemma 2.6] *The restriction $\mathcal{N}_{C/Y}|_{C_0}$ equals the sheaf of rational sections of $\mathcal{N}_{C_0/Y}$ having at most a simple pole at each point p_i in the normal direction determined by T_{C_i,p_i}.*

PROOF. The restrictions of the sheaves $\mathcal{N}_{C/Y}|_{C_0}$ and $\mathcal{N}_{C_0/Y}$ to the complement of $\{p_1, \dots, p_n\}$ are canonically isomorphic. The lemma states that this canonical isomorphism is the restriction of an injection $\mathcal{N}_{C_0/Y} \hookrightarrow \mathcal{N}_{C/Y}|_{C_0}$ which identifies $\mathcal{N}_{C/Y}|_{C_0}$ with the sheaf of rational sections, etc. This local assertion can be verified in a formal neighborhood of each node p_i.

Locally near p_i, $C \to Y$ is formally isomorphic to the union of the two axes inside a 2-plane inside an n-plane, i.e., the subscheme of $\mathbb{A}^n_k$ with ideal $I_{C/Y} = \langle x_1x_2, x_3, \dots, x_n\rangle$. The branch C_0 corresponds to just one of the axes, e.g., the subscheme of $\mathbb{A}^n_k$ with ideal $I_{C_0/Y} = \langle x_2, x_3, \dots, x_n\rangle$. The tangent direction of the other branch C_i is spanned by $(0, 1, 0, \dots, 0)$. Thus it is clear that $I_{C/Y}/I_{C_0/Y}\cdot I_{C/Y}$ is the submodule of $I_{C_0/Y}/I^2_{C_0/Y}$ of elements whose fiber at 0 is contained in the annihilator of T_{C_i,p_i}. Dualizing gives the lemma. □

The final bit of deformation theory needed has to do with deforming nodes. Let C be a proper, nodal curve and let $j : C \to Y$ be a regular embedding. Let p be a node of C and assume that Y is smooth at p. There are two branches C_1 and C_2 of C at p (possibly contained in the same irreducible component of C). The sheaf

$$\mathcal{T} := \mathit{Ext}^1_{\mathcal{O}_C}(\Omega_C, \mathcal{O}_C)$$

is a skyscraper sheaf supported at p and with fiber canonically identified to

$$\mathcal{T}|_p = T_{C_1,p} \otimes_k T_{C_2,p}.$$

The following lemma is as much definition as lemma.

LEMMA 3.17. *There exists a quotient of coherent sheaves*

$$\mathcal{N}_{C/Y} \twoheadrightarrow \mathcal{T}$$

such that for both $i = 1, 2$ the quotient $\mathcal{N}_{C/Y}|_{C_i}/\mathcal{N}_{C_i/Y}$ equals $\mathcal{T}$. A first-order deformation of $C \subset Y$, i.e., a global section of $\mathcal{N}_{C/Y}$ is said to smooth the node p to first-order *if the image of the section in $T_{C_1,p} \otimes_k T_{C_2,p}$ is nonzero. For a deformation*

$$\mathcal{C} \subset \Pi \times_k Y$$

of $C \subset Y$ over a smooth pointed curve $(\Pi, 0)$ (i.e., $\mathcal{C}_0 = C$), if the associated first-order deformation of $C \subset Y$ smooths the node p to first-order, then p is not contained in the closure of the singular locus of the projection,

$$(\Pi - \{0\}) \times_\Pi \mathcal{C} \to (\Pi - \{0\})$$

i.e., a general fiber $\mathcal{C}_t$ of the deformation smooths the node.

This is a well-known result. A good reference for this result, and many other results about deformations of singularities, is [**Art76**], particularly §I.6. Here is a brief remark on the proof. Because $C \subset Y$ is a regular embedding, the conormal sequence is exact on the left, i.e.,

$$0 \longrightarrow \mathcal{I}_{C/Y}/\mathcal{I}_{C/Y}^2 \longrightarrow \Omega_Y|_C \longrightarrow \Omega_C \longrightarrow 0$$

is a short exact sequence. Applying global Ext, there is a connecting map

$$\delta : H^0(C, \mathcal{N}_{C/Y}) \to \mathrm{Ext}^1_{\mathcal{O}_C}(\Omega_C, \mathcal{O}_C).$$

There is also a local-to-global sequence for global Ext inducing a map

$$\mathrm{Ext}^1_{\mathcal{O}_C}(\Omega_C, \mathcal{O}_C) \to H^0(C, \mathit{Ext}^1_{\mathcal{O}_C}(\Omega_C, \mathcal{O}_C)) = H^0(C, \mathcal{T}) = T_{C_1,p} \otimes_k T_{C_2,p}.$$

The composition of these two maps is precisely the map on global sections associated to $\mathcal{N}_{C/Y} \to \mathcal{T}$. The global Ext group is identified with the first-order deformations of C as an abstract scheme, and the Ext term is identified with the first-order deformations of the node. It is worth noting that even if the first-order deformation does not smooth the node, the full deformation $\mathcal{C} \subset \Pi \times_k Y$ may smooth the node if the total space $\mathcal{C}$ is singular at $(0, p)$.

The first result using the smoothing combs technique is the following.

PROPOSITION 3.18. *Let Y be a quasi-projective scheme over an algebraically closed field k. The very free locus $Y_{\mathrm{v.f.}}$ is an open subset of Y. More generally, for a flat, quasi-projective morphism $\pi : Y \to B$, the relative very free locus $Y_{\pi,\mathrm{v.f.}}$ is an open subset of Y.*

Let Y be an irreducible, quasi-projective scheme over an algebraically closed field k. Denote by t_1, resp. t_2, the closed point of $\mathbb{P}^1_k$, $t_1 = 0$, resp. $t_2 = \infty$. Let y_1 and y_2 be closed points of $Y_{\mathrm{v.f.}}$, let a and k be nonnegative integers, and let there be given curvilinear k-jets in Y at each of y_1 and y_2. If the given k-jets are general among all curvilinear k-jets at y_1 and y_2, then there exists a morphism

$$f : (\mathbb{P}^1_k, t_1, t_2) \to (Y_{\mathrm{v.f.}}, y_1, y_2)$$

mapping the k-jet of $\mathbb{P}^1$ at t_i isomorphically to the given k-jet at y_i for $i = 1, 2$ and such that

$$f^*T_Y \cong \mathcal{O}_{\mathbb{P}^1}(a_1) \oplus \cdots \oplus \mathcal{O}_{\mathbb{P}^1}(a_n), \ a_1, \ldots, a_n \geq a.$$

PROOF. In the absolute case, resp. relative case, the very free locus $Y_{\mathrm{v.f.}}$, resp. $Y_{\pi,\mathrm{v.f.}}$, is defined to be the same as the very free locus of the smooth locus Y_{smooth}, resp. $Y_{\pi,\mathrm{smooth}}$. Since the smooth locus is open in Y, and since an open subset of an open subset is an open subset, it suffices to prove the very free locus is open under the additional hypothesis that Y is smooth, resp. that π is smooth.

By the definition of $Y_{\mathrm{v.f.}}$, for each $i = 1, 2$ there exists a very free morphism

$$f_i : (\mathbb{P}^1, 0) \to (Y_{\mathrm{v.f.}}, y_i), \quad f_i^*T_Y \cong \mathcal{O}_{\mathbb{P}^1}(a_1) \oplus \cdots \oplus \mathcal{O}_{\mathbb{P}^1}(a_n), a_1, \ldots, a_n \geq 1.$$

In particular, for each $i = 1, 2$, $h^1(\mathbb{P}^1, f_i^* T_Y(-\underline{0} - \underline{\infty}))$ equals 0, where $\underline{0}$, resp. $\underline{\infty}$, is the Cartier divisor of the point 0, resp. ∞, in $\mathbb{P}^1$. Since the normal sheaf of f_i is a quotient of $f_i^* T_Y$, also $h^1(\mathbb{P}^1, \mathcal{N}_{f_i}(-\underline{0} - \underline{\infty}))$ equals 0. Thus, applying Proposition 3.14 where $Z_1 = \{0, \infty\}$ and $Z_2 = \mathbb{P}^1$, there exist deformations of the morphism f_i such that $f_i(0)$ equals y_i and $f_i(\infty)$ is any point in a nonempty Zariski open subset of Y. The same argument holds in the relative case.

Next assume that Y is irreducible and quasi-projective. Then the smooth locus Y_{smooth} is also irreducible (or empty). Thus, by the same argument as above, a proof of the second result for smooth varieties implies the second result in general. Thus assume Y is also smooth.

Since Y is irreducible, the open for $i = 1$ intersects the open for $i = 2$. Thus there exist very free morphisms f_1 and f_2 such that $f_1(\infty) = f_2(\infty)$. Let C be the nodal curve with two irreducible components C_1 and C_2 each isomorphic to $\mathbb{P}^1$ and with a single node which, when considered as a point in either C_1 or C_2, corresponds to ∞ in $\mathbb{P}^1$. Let $f : C \to Y$ be the unique morphism whose restriction to each component C_i equals f_i. Denote by

$$0 \longrightarrow \mathcal{N}'_{C/Y} \longrightarrow \mathcal{N}_{C/Y} \longrightarrow \mathcal{T} \longrightarrow 0$$

the short exact sequence coming from Lemma 3.17. Using Lemma 3.16, there is an exact sequence

$$0 \longrightarrow \mathcal{N}_{C/Y}|_{C_1}(-\underline{0} - \underline{\infty}) \longrightarrow \mathcal{N}'_{C/Y}(-y_1 - y_2) \longrightarrow \mathcal{N}_{C_2/Y}(-\underline{0}) \longrightarrow 0$$

and an exact sequence

$$0 \longrightarrow \mathcal{N}_{C_1/Y}(-\underline{0} - \underline{\infty}) \longrightarrow \mathcal{N}_{C/Y}|_{C_1}(-\underline{0} - \underline{\infty}) \longrightarrow \kappa_\infty \longrightarrow 0$$

where κ_∞ is the skyscraper sheaf on C_1 supported at ∞. Applying the long exact sequence of cohomology, using that $h^1(C_i, \mathcal{N}_{C_i/Y}(-\underline{0} - \underline{\infty}))$ equals 0 for $i = 1, 2$, and chasing diagrams, this finally gives that $h^1(C, \mathcal{N}'_{C/Y}(-y_1 - y_2))$ also equals 0.

This has two consequences. First, this implies $h^1(C, \mathcal{N}_{C/Y}(-y_1 - y_2))$ equals 0, and thus the space of deformations of C containing y_1 and y_2 is smooth by Proposition 3.14. And second, the map

$$H^0(C, \mathcal{N}_{C/Y}(-y_1 - y_2)) \to T_{C_1,\infty} \otimes T_{C_2,\infty}$$

is surjective. Thus there exist first-order deformations of C containing y_1 and y_2 and smoothing the node at ∞. Since the space of deformations containing y_1 and y_2 is smooth, this first-order deformation is the one associated to a one-parameter deformation

$$\mathcal{C} \subset \Pi \times_k Y$$

of $[C]$ over a smooth, pointed curve $(\Pi, 0)$ (e.g., choose Π to be a general complete intersection curve in the smooth deformation space containing the given Zariski tangent vector). By Lemma 3.17, for a general point t of Π, $\mathcal{C}_t$ is a smooth, connected curve containing y_1 and y_2. Since the arithmetic genus of C is 0, the arithmetic genus of $\mathcal{C}_t$ is also 0, i.e., $\mathcal{C}_t \cong \mathbb{P}^1_k$. Let

$$f_1 : \mathbb{P}^1_k \to \mathcal{C}_t$$

be an isomorphism with $f_1(t_i) = y_i$ for $i = 1, 2$. Because $h^1(C, \mathcal{N}_{C/Y}(-y_1 - y_2))$ equals 0, by the semicontinuity theorem also $h^1(\mathcal{C}_t, \mathcal{N}_{\mathcal{C}_t/Y}(-y_1 - y_2))$ equals 0. This

implies that

$$f_1^*T_Y \cong \mathcal{O}_{\mathbb{P}^1}(a_1) \oplus \cdots \oplus \mathcal{O}_{\mathbb{P}^1}(a_n), \text{ for integers } a_1, \ldots, a_n \geq 1.$$

Next, for every integer a, let $g_a : \mathbb{P}^1 \to \mathbb{P}^1$ be the morphism $z \mapsto z^a$. Then the composition $f_a = f_1 \circ g_1$ is a morphism

$$f_a : (\mathbb{P}^1_k, t_1, t_2) \to (Y_{\text{v.f.}}, y_1, y_2)$$

with

$$f_a^*T_Y = g_a^*(f_1^*T_Y) \cong \mathcal{O}_{\mathbb{P}^1}(a_1) \oplus \cdots \oplus \mathcal{O}_{\mathbb{P}^1}(a_n) \text{ for integers } a_1, \ldots, a_n \geq a,$$

namely the new integer $a_i(f_a)$ equals $a \cdot a_i(f_1)$. Next, choosing $a \geq 2k+1$, this implies that

$$h^1(\mathbb{P}^1, f_a^*T_Y(-(k+1)(t_1+t_2))) = 0.$$

Applying Proposition 3.14 with $\mathbb{P}^1 \times_k Y$ in the place of Y, with the graph of f_a in the place of Z_2 and with $Z_1 = (k+1)(t_1+t_2)$ in the place of Z, deformations of f_a map the k-jet of $\mathbb{P}^1$ at t_1, resp. at t_2, isomorphically to a general k-jet at y_1, resp. at y_2. □

The following proposition is the strongest generalization of Proposition 3.18 we will need. It is stated as a theorem about finding new sections of a rationally connected fibration under the hypothesis that one such section exists. In this sense it may seem premature (and dangerously close to circular logic), since Theorem 3.9 is not yet proved. In fact the proposition is used in the proof of Theorem 3.9 not for the original fibration, but only for a *constant* fibration

$$\text{pr}_{\mathbb{P}^1} : \mathbb{P}^1_k \times_k Y \to \mathbb{P}^1_k$$

which obviously admits sections (constant sections). So there is nothing circular in the application of the proposition to the proof of Theorem 3.9.

PROPOSITION 3.19 (Generic weak approximation). [**KMM92**], [**HT06**] *Let B be a smooth, connected, projective curve over an algebraically closed field k. Let $\pi : U \to B$ be a smooth, quasi-projective morphism having irreducible geometric fibers. Assume there exists a section $s : B \to U$ mapping the generic point of B into the very free locus of the generic fiber of π. Let $(b_1, \ldots, b_M, b'_1, \ldots, b'_{M'})$ be distinct closed points of B such that $s(b_i)$ is in the very free locus $U_{b_i,\text{v.f.}}$ of the fiber U_{b_i} for each $i = 1, \ldots, M$. Let k and a be nonnegative integers. For each i, let x_i be a closed point of $U_{b_i,\text{v.f.}}$ and let there be given a curvilinear k-jet in U at x_i. Assuming each of these k-jets is a general k-jet at x_i, there exists a section $\sigma : B \to U$ such that*

(i) *for each $i = 1, \ldots, M$, $\sigma(b_i)$ equals x_i,*
(ii) *for each $i = 1, \ldots, M'$, $\sigma(b'_i)$ equals $s(b'_i)$,*
(iii) *for every invertible $\mathcal{O}_B$-module $\mathcal{L}$ of degree $\leq a$, $h^1(B, \mathcal{N}_{\sigma(B)/U} \otimes_{\mathcal{O}_B} \mathcal{L}^\vee)$ equals 0,*
(iv) *and σ maps the k-jet of b_i in B isomorphically to the given k-jet at x_i for each i.*

In fact Hassett and Tschinkel proved much more: the result holds for *arbitrary* k-jets transverse to the fibers of π (i.e., for k-jets whose associated Zariski tangent vector is not contained in a fiber of π). In what follows we only need the "generic" result, which is all we prove.

PROOF. Denote by Ω_π the locally free sheaf of relative differentials of π, and denote by T_π the dual locally free sheaf. Choose a large integer N and enlarge the set of pairs $((b_i, x_i))|_{i=1,\dots,M}$ to a set $((b_i, x_i))_{i=1,\dots,N}$ having the same properties above and such that the collection $(b_i)_{i=M+1,\dots,N}$ is a general collection of $N-M$ points in B (this is possible because for all but finitely many closed points of B, $s(b)$ is contained in $U_{b,\text{v.f.}}$). By Proposition 3.18, applied with $k=1$, i.e., in the case that k-jets are simply tangent directions, for every $i = 1, \dots, N$ there exists a morphism

$$f_i : (\mathbb{P}^1, 0, \infty) \to (U_{b_i,\text{v.f.}}, s(b_i), x_i)$$

such that

$$f_i^* T_\pi \cong \mathcal{O}_{\mathbb{P}^1}(a_1) \oplus \cdots \oplus \mathcal{O}_{\mathbb{P}^1}(a_n), \quad a_1, \dots, a_n \geq 1$$

and the tangent direction of $f_i(\mathbb{P}^1)$ at $s(b_i)$ is a general tangent direction in $T_{U_{b_i}, s(b_i)}$. But of course the tangent space $T_{U_{b_i}, s(b_i)}$ equals the normal space $N_{s(B)/U}|_{s(b_i)}$. Thus the tangent direction of $f_i(\mathbb{P}^1)$ at $s(b_i)$ gives a general normal direction to $s(B)$ in U at $s(b_i)$.

Form the comb $j : C_{\text{comb}} \to U$ with handle $s(B)$ and with each morphism f_i being a tooth L_i attached at $s(b_i)$. By Lemma 3.16, $\mathcal{N}_{C_{\text{comb}}/U}|_{s(B)}$ equals the sheaf of rational sections of $\mathcal{N}_{s(B)/U}$ having at most a simple pole at each point $s(b_i)$ in a general normal direction at $s(b_i)$. Assuming the integer N is sufficiently large, Lemma 3.15 then implies that $h^1(B, s^*\mathcal{N}_{C_{\text{comb}}/U})$ equals 0. Moreover, fixing an auxiliary invertible sheaf $\mathcal{M}$ on B of degree $g(B)+1$ and applying Lemma 3.15 to $s^*\mathcal{N}_{s(B)/U}(-(b'_1 + \cdots + b'_{M'})) \otimes_{\mathcal{O}_B} \mathcal{M}^\vee$, for N sufficiently large also $h^1(B, s^*\mathcal{N}_{C_{\text{comb}}/U}(-(b'_1 + \cdots + b'_{M'})) \otimes_{\mathcal{O}_B} \mathcal{M}^\vee)$ equals 0.

For every i, there is a short exact sequence

$$0 \longrightarrow f_i^*\mathcal{N}_{L_i/U_{b_i}} \longrightarrow f_i^*\mathcal{N}_{L_i/U} \longrightarrow f_i^*\mathcal{N}_{U_{b_i}/U} \longrightarrow 0.$$

Of course the normal sheaf $\mathcal{N}_{U_{b_i}/U}$ is just $\mathcal{O}_{U_{b_i}}$ since U_{b_i} is a smooth fiber of a morphism to a curve. Also the tangent direction of $s(B)$ at $s(b_i)$ surjects onto the fiber of $\mathcal{N}_{U_{b_i}/U}$ at $s(b_i)$. Thus the elementary transform up of $\mathcal{N}_{L_i/U}$ at $s(b_i)$ in this direction surjects onto the elementary transform up of $\mathcal{O}^1_{\mathbb{P}}$ at ∞, i.e., it surjects onto $\mathcal{O}_{\mathbb{P}^1}(1)$. Thus, by Lemma 3.16, there is a short exact sequence

$$0 \longrightarrow f_i^*\mathcal{N}_{L_i/U_{b_i}} \longrightarrow f_i^*\mathcal{N}_{C_{\text{comb}}/U} \longrightarrow \mathcal{O}_{\mathbb{P}^1}(1) \longrightarrow 0.$$

Twisting by $\mathcal{O}_{\mathbb{P}^1}(-2)$ and applying the long exact sequence of cohomology associated to the short exact sequence, $h^1(\mathbb{P}^1, f_i^*\mathcal{N}_{C_{\text{comb}}/U}(-\underline{0} - \underline{\infty}))$ equals 0. Combined with the result of the previous paragraph and joining the two types of normal sheaf via the short exact sequence

$$0 \to \bigoplus_{i=1}^{N} \mathcal{N}_{C_{\text{comb}}/U}|_{L_i}(-x_i - s(b_i)) \to \mathcal{N}_{C_{\text{comb}}/U}(-(x_1 + \cdots + x_N) - (b'_1 + \cdots + b'_{M'}))$$

$$\to \mathcal{N}_{C_{\text{comb}}/U}|_{s(B)}(-(b'_1 + \cdots + b'_{M'})) \to 0,$$

the long exact sequence of cohomology implies both that

$$h^1(C_{\text{comb}}, \mathcal{N}_{C_{\text{comb}}/U}(-(x_1 + \cdots + x_N) - (b'_1 + \cdots + b'_{M'}))) = 0,$$

and that the map

$$H^0(C_{\text{comb}}, \mathcal{N}_{C_{\text{comb}}/U}(-(x_1 + \cdots + x_N) - (b'_1 + \cdots + b'_{M'}))) \to$$

$$H^0(B, s^*\mathcal{N}_{C_{\text{comb}}/U}(-(b'_1 + \cdots + b'_{M'})))$$

is surjective.

Thus, by Proposition 3.14, the space of deformations of C_{comb} containing $x_1, \ldots, x_N$ and $b'_1, \ldots, b'_{M'}$ is smooth. And, by Lemma 3.17, to prove there exists a deformation smoothing every node of C_{comb}, it suffices to prove for every i there exists a section of $s^*\mathcal{N}_{C_{\text{comb}}/U}(-(b'_1 + \cdots + b'_{M'}))$ whose image in $T_{s(B),s(b_i)} \otimes_k T_{L_i,s(b_i)}$ is nonzero. Of course this skyscraper sheaf $\mathcal{T}_{s(b_i)}$ is a quotient of the fiber of

$$s^*\mathcal{N}_{C_{\text{comb}}/U}(-(b'_1 + \cdots + b'_{M'}))$$

at b_i. Thus it suffices to prove for every i that

$$h^1(B, s^*\mathcal{N}_{C_{\text{comb}}/U}(-b_i - (b'_1 + \cdots + b'_{M'}))) = 0.$$

Recall the auxiliary invertible sheaf $\mathcal{M}$ of degree $g(B) + 1$. Because the invertible sheaf $\mathcal{M}(-b_i)$ has degree $g(B)$, it is effective, say $\mathcal{O}_B(\Delta_i)$. Thus there exists an injective $\mathcal{O}_B$-module homomorphism

$$s^*\mathcal{N}_{C_{\text{comb}}/U}(-(b'_1 + \cdots + b'_{M'})) \otimes_{\mathcal{O}_B} \mathcal{M}^\vee \hookrightarrow$$

$$s^*\mathcal{N}_{C_{\text{comb}}/U}(-(b'_1 + \ldots b'_{M'})) \otimes_{\mathcal{O}_B} \mathcal{M}^\vee(\Delta_i) = s^*\mathcal{N}_{C_{\text{comb}}/U}(-b_i - (b'_1 + \cdots + b'_{M'}))$$

with torsion cokernel. Since

$$h^1(B, s^*\mathcal{N}_{C_{\text{comb}}/U}(-(b'_1 + \cdots + b'_{M'})) \otimes_{\mathcal{O}_B} \mathcal{M}^\vee) = 0,$$

and since every torsion sheaf has h^1 equal to 0, also

$$h^1(B, s^*\mathcal{N}_{C_{\text{comb}}/U}(-b_i - (b'_1 + \cdots + b'_{M'}))) = 0$$

for every i. Therefore there exist a one-parameter family of deformations $(\mathcal{C}_t)_{t\in\Pi}$ of C_{comb} containing each of $x_1, \ldots, x_M$, containing each of $s(b'_1), \ldots, s(b'_{M'})$ and smoothing every node of C_{comb}, i.e., for t general, $\mathcal{C}_t$ is smooth.

Because π_U maps $s(B)$ to B with degree 1, also π_U maps $\mathcal{C}_t$ to B with degree 1. Because $\mathcal{C}_t$ is smooth, this means the projection $\mathcal{C}_t \to B$ is an isomorphism. Therefore there exists a section $\sigma_t : B \to U$ of π_U with image $\mathcal{C}_t$. In particular, $\sigma_t(b_i) = x_i$ for every $i = 1, \ldots, M$ and $\sigma_t(b'_i) = s(b'_i)$ for every $i = 1, \ldots, M'$. Because

$$h^1(C_{\text{comb}}, \mathcal{N}_{C_{\text{comb}}/U}(-(x_1 + \cdots + x_N))) = 0,$$

by semicontinuity also

$$h^1(B, \sigma_t^*\mathcal{N}_{\sigma_t(B)/U}(-(x_1 + \cdots + x_N))) = 0$$

for t general. In particular, if $N \geq a + g(B)$, then for every invertible sheaf $\mathcal{L}$ of degree $\leq a$, $\mathcal{L}^\vee(x_1 + \cdots + x_N)$ has degree $\geq g(B)$ and thus is effective, say $\mathcal{O}_B(\Delta)$. Therefore there exists an injective sheaf homomorphism

$$\begin{aligned}\sigma_t^*\mathcal{N}_{\sigma_t(B)/U}(-(x_1 + \cdots + x_N)) \hookrightarrow \sigma_t^*\mathcal{N}_{\sigma_t(B)/U}(-(x_1 + \cdots + x_N) + \Delta) \\ = \sigma_t^*\mathcal{N}_{\sigma_t(B)/U} \otimes_{\mathcal{O}_B} \mathcal{L}^\vee\end{aligned}$$

with torsion cokernel. So, by the same type of argument as above,

$$h^1(B, \sigma_t^*\mathcal{N}_{\sigma_t(B)/U} \otimes_{\mathcal{O}_B} \mathcal{L}^\vee) = 0$$

for every invertible sheaf $\mathcal{L}$ of degree $\leq a$.

Finally, applying the last result when $a = (k + 1)(M + M')$ and

$$\mathcal{L} = \mathcal{O}_B((k + 1)(b_1 + \cdots + b_M + b'_1 + \cdots + b'_{M'})),$$

there exists a section $\sigma : B \to U$ of π_U as above and satisfying

$$h^1(B, \sigma^*\mathcal{N}_{\sigma(B)/U}(-(k+1)(b_1 + \cdots + b_M + b'_1 + \cdots + b'_{M'}))) = 0.$$

Therefore, by Proposition 3.14 once more, for a general deformation of $\sigma(B)$ containing $x_1, \ldots, x_M$ and $s(b'_1), \ldots, s(b'_{M'})$, the k-jet of the curve at each point x_i and $s(b'_i)$ is a general curvilinear k-jet in U at that point. □

The main application is to the case when U equals $\mathbb{P}^1 \times_k Y$ where Y is a smooth, irreducible, quasi-projective k-scheme whose very free locus $Y_{\text{v.f.}}$ is nonempty.

COROLLARY 3.20. *Every rational curve in Y intersecting $Y_{\text{v.f.}}$ is contained in $Y_{\text{v.f.}}$. For every integer k, for every integer a, for every collection of distinct, closed points $b_1, \ldots, b_M$ of $\mathbb{P}^1$, for every collection of closed points $y_1, \ldots, y_M$ of $Y_{\text{v.f.}}$ (not necessarily distinct), and for every choice of a curvilinear k-jet in Y at each point y_i, if each k-jet is general among curvilinear k-jets at y_i, then there exists a morphism*

$$f : (\mathbb{P}^1, b_1, \ldots, b_M) \to (Y, y_1, \ldots, y_M)$$

mapping the k-jet of $\mathbb{P}^1$ at b_i isomorphically onto the given k-jet at y_i and such that

$$f^*T_Y \cong \mathcal{O}_{\mathbb{P}^1}(a_1) \oplus \cdots \oplus \mathcal{O}_{\mathbb{P}^1}(a_n), \ a_1, \ldots, a_n \geq a.$$

PROOF. Let $B = \mathbb{P}^1$, let $U = B \times_k Y$ and let π_B be the obvious projection. The sections of π_B are precisely the graphs of morphisms $f : \mathbb{P}^1 \to Y$. In particular, if f is a morphism whose image intersects $Y_{\text{v.f.}}$, then the section $s = (\text{Id}_B, f)$ satisfies the hypotheses of Proposition 3.19. Thus, for every point $b' = b'_1$ of $\mathbb{P}^1$, there exists a section $\sigma = (\text{Id}_{\mathbb{P}^1}, \phi)$ with $\sigma(b') = s(b')$ and with $h^1(B, \sigma^*\mathcal{N}_{\sigma(B)/U}(-2))$ equal to 0. In other words, $\phi : \mathbb{P}^1_k \to Y$ is a morphism with $\phi(b') = f(b')$ and with $h^1(\mathbb{P}^1, \phi^*T_Y(-2))$ equal to 0. Thus ϕ is a very free morphism whose image contains $f(b')$. Therefore every point in the image of f is contained in the very free locus, i.e., every rational curve in Y intersecting $Y_{\text{v.f.}}$ is contained in $Y_{\text{v.f.}}$.

The rest of the corollary is just a straightforward translation of Proposition 3.19 to this context. □

There is one more result in this direction which is useful. The proof is similar to the arguments above.

LEMMA 3.21. [**Kol96**, Lemma II.7.10.1] *Let C_{comb} be a comb with handle C and teeth $L_1, \ldots, L_n$. Let $\rho : \mathcal{C} \to \Pi$ be a one-parameter deformation of C_{comb} over a pointed curve $(\Pi, 0)$ whose general fiber $\mathcal{C}_t$ is smooth. Let $\mathcal{E}$ be a locally free sheaf on $\mathcal{C}$. If $\mathcal{E}|_{L_i}$ is ample for every i and if $h^1(C, (\mathcal{E}|_C) \otimes_{\mathcal{O}_C} \mathcal{M})$ equals 0 for every invertible $\mathcal{O}_C$-module $\mathcal{M}$ of degree $\geq n$, then $h^1(\mathcal{C}_t, \mathcal{E}|_{\mathcal{C}_t})$ equals 0 for general t in Π.*

3.3. Ramification issues. The argument sketched in Subsection 3.1 and the powerful smoothing combs technique from Subsection 3.2 form the core of the proof of Theorem 3.9. However there is a technical issue complicating matters. There may be codimension 1 points of X at which the morphism $\pi : X \to B$ is not smooth. In other words, finitely many scheme-theoretic fibers of π may have irreducible components occurring with multiplicity ≥ 1. This is a well-known issue when working with fibrations. Although there are sophisticated ways to deal with

this (using log structures or Deligne-Mumford stacks), for the purposes of this proof it suffices to deal with this in a more naive manner.

In fact there may be codimension 0 points of X at which π is not smooth, at least if k has positive characteristic. The hypotheses in Theorem 3.9 prevent this, but something slightly weaker suffices. Let B be a smooth k-curve, let X be a reduced, finite type k-scheme and let $\pi : X \to B$ be a flat morphism. From here on, we assume the following hypothesis.

HYPOTHESIS 3.22. The geometric generic fiber of π is reduced. Equivalently, π is smooth at every generic point of X, cf. [**Gro67**, Proposition 4.6.1]. This hypothesis is automatic if $\mathrm{char}(k)$ equals 0.

DEFINITION 3.23. The *good locus* of π is the maximal open subscheme U of X such that U is smooth and such that for every point b of B the reduced scheme of the fiber $\pi^{-1}(b) \cap U$ is smooth. Denote the restriction of π to U by π_U. The morphism π is *good* if the good locus equals all of X. The *log divisor* of π is the Cartier divisor $D_{\pi,\log}$ of U given by

$$D_{\pi,\log} := \sum_{b \in B(k)} \pi_U^*(b) - \pi_U^*(b)_{\mathrm{red}},$$

where $\pi_U^*(b)_{\mathrm{red}}$ is the reduced Cartier divisor.

Since the geometric generic fiber of π is reduced, so is the geometric generic fiber of π_U (or else it is empty if U is empty). Thus the sum in the definition of the log divisor reduces to a sum over those finitely many closed points b of B for which $\pi_U^*(b)$ is nonreduced.

LEMMA 3.24. *The complement of U in X has codimension ≥ 2. If* $\mathrm{char}(k)$ *equals* 0*, then the pullback map on relative differentials*

$$\pi_U^* : \pi_U^*\Omega_{B/k} \to \Omega_{U/k}$$

factors uniquely through the inclusion

$$\pi_U^*\Omega_{B/k} \hookrightarrow \pi_U^*\Omega_{B/k}(D_{\pi,\log})$$

and the cokernel

$$\Omega_{\pi,\log} := \mathrm{Coker}(\pi_U^*\Omega_{B/k}(D_{\pi,\log}) \to \Omega_{U/k})$$

is locally free.

PROOF. To construct U, first remove the closure of the singular locus of the geometric generic fiber of π and next remove the singular locus from the reduced scheme of the finitely many singular fibers. Both of these sets have codimension 2 in X (the first by Hypothesis 3.22).

The proof of the second part uses that $\mathrm{char}(k) = 0$. It can be checked formally locally near every closed point x of U. Denote by b the image $\pi(x)$ in B and denote by D the reduced structure on the irreducible component of $\pi^{-1}(b)$ containing x. Since x is in U, D is a smooth Cartier divisor in U. Let r be a defining equation for D in U and let t be a defining equation for b in B. Near x, $\pi^*(b) = mD +$ other terms. Thus, in $\widehat{\mathcal{O}}_{U,x}$,

$$\pi^*t = a_m r^m + a_{m+1} r^{m+1} + \cdots$$

where a_m is a unit. Because $\mathrm{char}(k) = 0$, the power series

$$u = \sqrt[m]{a_m + a_{m+1}r + \cdots}$$

is a well-defined unit in $\widehat{\mathcal{O}}_{U,x}$. Thus, after replacing r by ur, there exists a regular system of parameters $r, r_2, \ldots, r_n$ for $\widehat{\mathcal{O}}_{U,x}$ with respect to which the pullback homomorphism π^* is the unique local homomorphism with

$$\pi^* t = r^m \text{ and } \pi^*(dt) = mr^{m-1}dr.$$

In particular, the pullback homomorphism π^* on relative differentials locally factors through $\pi^*\Omega_B((e-1)D) = \pi^*\Omega_B(D_{\pi,\log})$. Moreover the stalk of the cokernel of the induced homomorphism is the free module generated by $dr_2, \ldots, dr_n$. □

The locally free quotient $\Omega_{\pi,\log}$ of Ω_π is called the sheaf of *log relative differentials*. Of course it equals the torsion-free quotient of Ω_π. But its true importance comes from the following lemma: given a base change $V \to B$ for which the normalized fiber product $\widetilde{U \times_B V}$ is smooth over V, the sheaf $\Omega_{\widetilde{U \times_B V}/V}$ of relative differentials of $\widetilde{U \times_B V}$ over V equals the pullback of $\Omega_{\pi,\log}$. Thus the relative deformation theory of $\widetilde{U \times_B V}$ over V is already captured by the sheaf $\Omega_{\pi,\log}$ on U. Before stating the lemma precisely, there is some setup.

Let

$$\pi : U \to B, \quad \varpi : V \to B$$

be two good morphisms with respective log divisors $D_{\pi,\log}$ and $E_{\varpi,\log}$. Let b be a closed point of B. Let D be a prime divisor of U in $\mathrm{supp}(D_{\pi,\log}) \cap \pi^{-1}(b)$, and let E be a prime divisor of V in $\mathrm{supp}(E_{\varpi,\log}) \cap \varpi^{-1}(b)$. Denote by $m_D - 1$, resp. $m_E - 1$, the coefficient of D in $D_{\pi,\log}$, resp. the coefficient of E in $E_{\varpi,\log}$. The *normalized fiber product* of U and V along D and E is the normalization $\widetilde{U \times_B V}$ of $U \times_B V$ along $D \times_{\{b\}} E$. Denote by

$$\mathrm{pr}_U : U \times_B V \to U, \quad \mathrm{pr}_V : U \times_B V \to V$$

the two projections, and denote by

$$\widetilde{\mathrm{pr}}_U : \widetilde{U \times_B V} \to U, \quad \widetilde{\mathrm{pr}}_V : \widetilde{U \times_B V} \to V$$

the compositions with the normalization morphism. Denote by Exc the exceptional locus of the morphism, i.e.,

$$\mathrm{Exc} := (\widetilde{\mathrm{pr}}_U^{-1}(D) \cap \widetilde{\mathrm{pr}}_V^{-1}(E))_{\mathrm{reduced}}.$$

From this point forward we explicitly assume that $\mathrm{char}(k)$ equals 0.

Hypothesis 3.25. The algebraically closed ground field k has characteristic 0. In particular, this implies Hypothesis 3.22.

The sheaves Ω_π and $\Omega_{\pi,\log}$ agree over a dense open subset of U, namely $U - \mathrm{supp}(D_{\pi,\log})$. Because $\widetilde{\mathrm{pr}}_V$ and pr_V are isomorphic over a dense open subset of V (namely $V - E$) also $\Omega_{\widetilde{\mathrm{pr}}_V}$ agrees with $\widetilde{\mathrm{pr}}_V^*\Omega_\pi$ on a dense open subset of $\widetilde{U \times_B V}$. Therefore also Ω_ϖ agrees with $\widetilde{\mathrm{pr}}_V^*\Omega_{\pi,\log}$ on a dense open subset of $\widetilde{U \times_B V}$.

LEMMA 3.26. *The morphism*

$$\widetilde{\mathrm{pr}}_V : \widetilde{U \times_B V} \to V$$

is smooth at every point of Exc *if and only if* m_D *divides* m_E. *In this case the reduced normalization equals the blowing up of* $U \times_B V$ *along the closed subscheme* $\mathrm{pr}_U^{-1}(D) \times \mathrm{pr}_V^{-1}((m_E/m_D)E)$ *and* Exc *is contained in the maximal open neighborhood of* $\widetilde{U \times_B V}$ *on which* $\Omega_{\widetilde{\pi}}$ *agrees with* $(\widetilde{\mathrm{pr}_V})^*\Omega_{\pi,\log}$.

PROOF. This is proved in much the same way as the second part of Lemma 3.24. For every closed point x of U and y of V with common image point $b = \pi(x) = \varpi(y)$, there exist a regular system of parameters $(r, r_2, \dots, r_n)$ for $\widehat{\mathcal{O}}_{U,x}$, resp. $(s, s_2, \dots, s_p)$ for $\widehat{\mathcal{O}}_{V,y}$, and a regular parameter t for $\widehat{\mathcal{O}}_{B,b}$ such that

$$\pi^* t = r^{m_D} \text{ and } \varpi^* t = s^{m_E},$$

and thus,

$$\widehat{\mathcal{O}}_{U\times_B V,(x,y)} = k[\![r, r_2, \dots, r_n, s, s_2, \dots, s_p]\!] / \langle r^{m_D} - s^{m_E} \rangle.$$

Denoting by m the greatest common factor of m_D and m_E, the stalk of the normalization equals

$$k[\![u, r, r_2, \dots, r_n, s, s_2, \dots, s_p]\!] / \langle r - u^{m_E/m}, s - u^{m_D/m} \rangle.$$

Thus it is formally smooth as a $k[\![s, s_2, \dots, s_p]\!]$-algebra if and only if m_D/m equals 1, i.e., if and only if m_D divides m_E. In this case it is easy to see that the normalization is the blowing up at the ideal $\langle s, t^{m_E/m_D} \rangle$ and it is easy to see that the module of relative differentials is the free module generated by $dr_2, \dots, dr_n$, i.e., it is the pullback of $\Omega_{\pi,\log}$. □

Having introduced the ideas need to deal with the ramification issues, we now resume the proof of Theorem 3.9. So from this point on we assume the following.

HYPOTHESIS 3.27. The following hypotheses of Theorem 3.9 hold.

(i) The algebraically closed ground field k has characteristic 0, i.e., Hypothesis 3.25 holds.
(ii) The smooth k-curve B is projective and connected.
(iii) The reduced, finite type k-scheme X is normal and projective.
(iv) And the geometric generic fiber of the flat morphism $\pi : X \to B$ is a normal, integral scheme whose smooth locus contains a very free curve.

DEFINITION 3.28. A *log preflexible curve* is a connected, smooth, proper curve $C \subset U$ such that

(i) the generic fiber of C over B is contained in the very free locus of the generic fiber of U over B,
(ii) $\pi_U(C)$ equals B,
(iii) and every intersection point of C with $\mathrm{supp}(D_{\pi,\log})$ is transverse, i.e., the tangent direction of C at the intersection point is not contained in the tangent space of $\mathrm{supp}(D_{\pi,\log})$.

A *linked log preflexible curve* is a B-morphism from a linked curve $j : C_{\mathrm{link}} \to U$ such that the handle C is log preflexible and for every link L_i the image in B of L_i is disjoint from the image in B of $D_{\pi,\log}$.

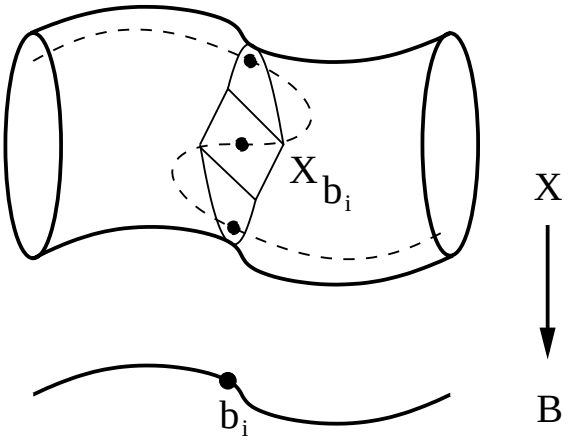


FIGURE 4. Where a log preflexible curve intersects a reduced, possibly singular, fiber the map from the curve to B is unramified.

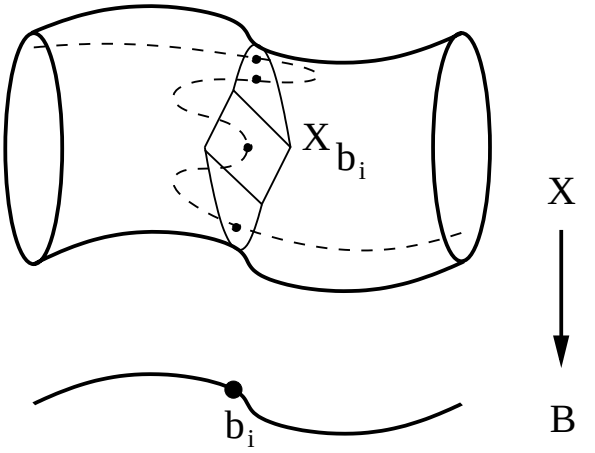


FIGURE 5. Where a log preflexible curve intersects a nonreduced fiber—e.g., the middle component has multiplicity 2—the map from the curve to B is necessarily ramified.

A log preflexible curve C is a *log flexible curve* if

$$h^1(C, T_{\pi,\log}|_C) \text{ equals } 0, \text{ where } T_{\pi,\log} := Hom_{\mathcal{O}_U}(\Omega_{\pi,\log}, \mathcal{O}_U).$$

A linked log preflexible curve is a *linked log flexible curve* if

$$h^1(C_{\text{link}}, j^* T_{\pi,\log}) \text{ equals } 0.$$

Figure 4 shows a log preflexible curve intersecting a singular, but reduced fiber. Because the curve is transverse to the fiber, the morphism to B is unramified. On the other hand, Figure 5 shows a log preflexible curve intersecting a nonreduced fiber—the middle component has multiplicity 2. Necessarily the map from the curve to B is ramified.

LEMMA 3.29. *There exists a log preflexible curve C. In fact, every intersection of X with $\dim(X)-1$ general hyperplanes is a log preflexible curve.*

PROOF. Because $X-U$ has codimension 2 in X, a general complete intersection curve in X is disjoint from $X-U$, i.e., it is contained in U. By hypothesis, $U_{\pi,\text{v.f.}}$ is a dense open subset of U and thus a general complete intersection curve intersects this open. Finally, by Bertini's theorem a general complete intersection curve in U is smooth and intersects $\text{supp}(D_{\pi,\log})$ transversally. □

An important consequence of the smoothing combs technique is the following result.

PROPOSITION 3.30. *There exists a log flexible curve in X. In fact, for every comb in X with log preflexible handle C and with sufficiently many very free teeth in fibers of π_U attached at general points of C and with general tangent directions, there exists a one-parameter deformation of the comb whose general member is a log flexible curve.*

PROOF. By hypothesis, C intersects the very free locus $U_{\pi,\text{v.f.}}$ of the morphism π_U. By the same argument as in the proof of Proposition 3.18, $U_{\pi,\text{v.f.}}$ is open. Therefore all but finitely many points of C are contained in $U_{\pi,\text{v.f.}}$. By Proposition 3.18 applied to 1-jets, i.e., to tangent directions, for each such point c there exists a very free rational curve in $U_{\pi_U(c)}$ containing c and whose tangent direction at c is a general tangent direction in $U_{\pi_U(c)}$.

Let C_{comb} be a comb obtained by attaching to C a number of teeth $L_1,\dots,L_N$ as in the previous paragraph at general points of C (in particular, points where $C\to B$ is unramified) and with general tangent directions in $U_{\pi_U(c)}$. These tangent directions are the same as normal directions to C in U. By the same argument as in the proof of Proposition 3.19, if N is sufficiently large there is a one-parameter deformation

$$\mathcal{C}\subset \Pi\times_k U$$

of C_{comb} such that $\mathcal{C}_t$ is smooth for general t in Π. The properties (i), (ii) and (iii) of Definition 3.28 are all open properties and hold for $\mathcal{C}_0=C_{\text{comb}}$, thus also hold for $\mathcal{C}_t$ so long as t is general.

For each tooth L_i in a fiber U_{b_i}, $T_{\pi,\log}|_{L_i}$ equals $T_{U_{b_i}}|_{L_i}$. Since L_i is very free, this is an ample locally free sheaf. Thus, by Lemma 3.21 with the pullback of $T_{\pi,\log}$ in the place of $\mathcal{E}$, we have that $h^1(\mathcal{C}_t,T_{\pi,\log}|_{\mathcal{C}_t})$ equals 0 for t a general point of Π. Therefore, for t a general point of Π, $\mathcal{C}_t$ is a log flexible curve. □

Because the fibers of π are rationally connected, every log preflexible curve, resp. log flexible curve, extends to a linked log preflexible curve, resp. linked log flexible curve.

LEMMA 3.31. *For every linked curve C_{link} such that each point $b_i=\pi_{C,\text{link}}(L_i)$ is disjoint from $\pi_U(D_{\pi,\log})$, and for every B-morphism $j_0: C\to X$ mapping C isomorphically to a log preflexible curve, resp. log flexible curve, and mapping each fiber C_{b_i} into the very free locus $U_{\pi,\text{v.f.}}$ of π_U, there exists a B-morphism $j: C_{\text{link}}\to X$ which is linked log preflexible, resp. linked log flexible, and restricting to j_0 on C.*

PROOF. Let L_i be a link of C_{link}. Let L_i intersect C in m points $t_1, \dots, t_m$ contained in the fiber over a general point b_i of B. Let $x_1, \dots, x_m$ be the images $j(t_1), \dots, j(t_m)$ in $U_{b_i,\text{v.f.}}$. By Corollary 3.20, there exists a morphism

$$j_i : (L_i, t_1, \dots, t_m) \to ((U_{b_i,\text{v.f.}}, x_1, \dots, x_m)$$

such that

$$j_i^* T_{U_{b_i}} \cong \mathcal{O}_{\mathbb{P}^1}(a_1) \oplus \cdots \oplus \mathcal{O}_{\mathbb{P}^1}(a_n) \text{ for integers } a_1, \dots, a_n \geq m - 1.$$

Because of this,

$$h^1(L_i, j_i^* T_{U_{b_i}}(-(t_1 + \cdots + t_m))) \text{ equals } 0.$$

Define $j : C_{\text{link}} \to U$ to be the unique morphism restricting to j_0 on C and restricting to j_i on each link L_i. Because $j_i(t_k) = j_0(t_k)$ for every link L_i and for every node t_k contained in L_i, this morphism is defined. It is clearly log preflexible.

Next assume that j_0 is log flexible. The claim is that j is also log flexible. To see this, consider the short exact sequence

$$0 \longrightarrow \bigoplus_i j_i^* T_{U_{b_i}}(-C_{b_i}) \longrightarrow j^* T_{\pi,\log} \longrightarrow j_0^* T_{\pi,\log} \longrightarrow 0.$$

By the hypothesis that j_0 is log flexible, the third term has vanishing h^1. And by the construction of j_i, $j_i^* T_{U_{b_i}}(-C_{b_i})$, i.e., $j_i^* T_{U_{b_i}}(-(t_1 + \cdots + t_m))$, has vanishing h^1. Thus, by the long exact sequence of cohomology, also $h^1(C_{\text{link}}, j^* T_{\pi,\log})$ equals 0. Therefore $j : C_{\text{link}} \to U$ is a linked log flexible curve. □

3.4. Existence of log deformations. There is a definition of one-parameter deformation that takes the divisor $D_{\pi,\log}$ into account. Unfortunately, not every curve over B admits a log deformation specializing to a section curve, e.g., étale covers of B are rigid. However, after attaching a sufficient number of links, the linked curve does admit a log deformation specializing to a section curve.

DEFINITION 3.32. Let $(C_{\text{link}}, \pi_{C,\text{link}})$ be a linked curve with handle C. Let $D_C \subset C$ be an effective, reduced, Cartier divisor contained in the smooth locus of C_{link}. A *one-parameter log deformation* of $(C_{\text{link}}, \pi_{C,\text{link}}, D_C)$ is a one-parameter deformation of $(C_{\text{link}}, \pi_{C,\text{link}})$,

$$(\rho, \pi_{\mathcal{C}}) : \mathcal{C} \to \Pi \times_k B$$

together with an effective Cartier divisor $D_{\mathcal{C}} \subset \mathcal{C}$ such that

(i) the pullback of $D_{\mathcal{C}}$ to $\mathcal{C}_0 = C_{\text{link}}$ equals D_C
(ii) and $\pi_{\mathcal{C}}(D_{\mathcal{C}})$ equals $\pi_C(D_C)$, i.e., $D_{\mathcal{C}}$ is vertical over B.

LEMMA 3.33. *For every finite morphism of smooth, projective curves $\pi_C : C \to B$ and for every effective, reduced, Cartier divisor D_C of C, after attaching sufficiently many links to C over general points of B, there exists a one-parameter log deformation specializing to a section curve.*

PROOF. For all sufficiently positive integers e, for a general morphism $g : C \to \mathbb{P}^1$ of degree e, the induced morphism $(\pi_C, g) : C \to B \times_k \mathbb{P}^1$ is unramified and is injective except for finitely many double points, none of which intersects the image of D_C. Denote by $\Sigma \to B \times_k \mathbb{P}^1$ the blowing up along the finitely many double points of $(\pi_C, g)(C)$. Then there is a B-morphism $h : C \to \Sigma$ which is an embedding.

For each point p of D_C, denote by m_p the multiplicity of p in the Cartier divisor $\pi_C^*(\pi_C(p))$. Denote by $\nu_p : \Sigma'_p \to \Sigma$ the m_p-fold iterated blowup of Σ first at p, then at the image of p in the strict transform of $h(C)$, etc. Denote by E_p the final exceptional divisor of this sequence of blowups. The point of this construction is that the strict transform of $h(C)$ intersects E_p at p, and E_p occurs with multiplicity m_p in the Cartier divisor $\Sigma'_p \times_B \{\pi_C(p)\}$. Denote by $\nu : \Sigma' \to \Sigma$ the fiber product over all points p in D_C of $\nu_p : \Sigma'_p \to \Sigma$. Denote by E the Cartier divisor in Σ' being the sum over all p of the pullback of E_p from Σ'_p. Denote by $\pi_{\Sigma'} : \Sigma' \to B$ the composition of $\Sigma' \to \Sigma \to B \times_k \mathbb{P}^1$ with pr_B. Denote by $h' : C \to \Sigma'$ the strict transform of $h(C)$. The point of this construction is that E is a Cartier divisor in Σ' which is vertical over B and such that h^*E equals D_C.

Denote by d the degree of π_C and let $t_1, \ldots, t_d$ be closed points of $\mathbb{P}^1$ such that the Cartier divisor $B \times_k \{t_1, \ldots, t_d\}$ of $B \times_k \mathbb{P}^1$ is disjoint from all double points of $(\pi_C, g)(C)$ and disjoint from $(\pi_C, g)(D)$. Denote by T the strict transform of $B \times_k \{t_1, \ldots, t_d\}$ in Σ'. Form the invertible sheaf $\mathcal{O}_{\Sigma'}(h'(C) - T)$ and the pushforward $\mathcal{E} := \pi_{\Sigma',*}\mathcal{O}_{\Sigma'}(h'(C)-T)$ on B. Because $\pi_{\Sigma'}$ is flat and because $\mathcal{O}_{\Sigma'}(h'(C)-T)$ is locally free, $\mathcal{E}$ is torsion-free. For every point b in $B - \pi_C(D_C)$, Σ'_b is isomorphic to $\mathbb{P}^1$ (via the projection $\Sigma' \to B \times_k \mathbb{P}^1 \to \mathbb{P}^1$). And $\Sigma'_b \cap h'(C)$ and $\Sigma'_b \cap T$ are divisors of the same degree d. Thus $\mathcal{O}_{\Sigma'}(h'(C)-T)|_{\Sigma'_b}$ is isomorphic to $\mathcal{O}_{\Sigma'_b} \cong \mathcal{O}_{\mathbb{P}^1}$. Therefore $\mathcal{E}|_b$ is isomorphic to $H^0(\Sigma'_b, \mathcal{O}_{\Sigma'_b})$, which is one-dimensional. Therefore $\mathcal{E}$ is an invertible sheaf.

By Riemann-Roch and Serre duality, for every sufficiently large degree, for a general effective divisor Δ on B of that degree, $\mathcal{E} \otimes_{\mathcal{O}_B} \mathcal{O}_B(\Delta)$ is globally generated. Choose Δ to be disjoint from $\pi_C(D_C)$ and from the image in B of the finitely many intersection points of $h'(C)$ and T. Since $\mathcal{E} \otimes_{\mathcal{O}_B} \mathcal{O}_B(\Delta)$ is globally generated, there exists a section which is nonzero at every point of Δ. Of course a nonzero section of this sheaf (up to scaling) is precisely the same thing as a divisor V on Σ' such that

$$h'(C) + \pi_{\Sigma'}^*\Delta \sim T + V.$$

For b in $B - \pi_C(D_C)$, if the section is nonzero at b then V does not intersect Σ'_b. The same does not necessarily hold for points b of $\pi_C(D_C)$ since b may lie in the support of $R^1\pi_{\Sigma',*}\mathcal{O}_{\Sigma'}(h'(C)-T)$. Therefore V is a sum of finitely many irreducible components of fibers of $\pi_{\Sigma'}$ (possibly with multiplicity) lying over points not in Δ.

The linked curve $(C_{\mathrm{link}}, \pi_{C,\mathrm{link}})$ is $h'(C) + \pi_{\Sigma'}^*\Delta$ together with the restriction of $\pi_{\Sigma'}$. Denote by Π the pencil of divisors in Σ' spanned by the divisors $h'(C) + \pi_{\Sigma'}^*\Delta$ and $T + V$, with these two divisors marked as 0 and ∞ respectively. Denote by $\mathcal{C} \subset \Pi \times_k \Sigma'$ the corresponding family of divisors. By Bertini's theorem, the general member $\mathcal{C}_t$ is smooth away from the base locus. Now the only singular points of $h'(C) + \pi_{\Sigma'}^*\Delta$ are the points $h'(\pi_C^{-1}(\Delta))$. Since V does not intersect $\pi_{\Sigma'}^*\Delta$, these singular points are not in the base locus. Since $\mathcal{C}_0$ is nonsingular at every basepoint, the same is true for $\mathcal{C}_t$ for t general. Thus a general member $\mathcal{C}_t$ is smooth everywhere.

Define $D_{\mathcal{C}}$ to be the pullback to $\mathcal{C}$ of the Cartier divisor E in Σ'. Because E is vertical over B and because h^*E equals D_C, the deformation $\mathcal{C}$ together with the effective Cartier divisor $D_{\mathcal{C}}$ is a one-parameter log deformation of $(C_{\mathrm{link}}, \pi_{C,\mathrm{link}}, D_C)$. And it specializes at $t = \infty$ to a union of section curves and vertical curves. □

3.5. Completion of the proof. We are finally prepared for the proof of Theorem 3.9.

Proof of Theorem 3.9. By Proposition 3.30, there exists a log flexible curve $j_0 : C \to U$. Denote by D_C the reduced scheme of the intersection $C \cap D_{\pi,\log}$. By Lemma 3.33, after attaching finitely many links to C over the points of a general divisor Δ of B, the linked curve C_{link} together with D_C admits a one-parameter log deformation

$$(\rho, \pi_{\mathcal{C}}) : \mathcal{C} \to \Pi \times_k B, \quad D_{\mathcal{C}} \subset \mathcal{C}$$

of (C_{link}, D_C) specializing to a section curve (in fact $\mathcal{C}_\infty$ is a union of section curves and vertical curves).

By Proposition 3.18, the relative very free locus $U_{\pi,\text{v.f.}}$ is open in U. Thus $C \cap U_{\pi,\text{v.f.}}$ is open in C. So its complement is finitely many points in C. Thus a general divisor Δ is disjoint from the finite set $\pi_U(D_{\pi,\log})$ and from the finite set $\pi_C(C - C \cap U_{\pi,\text{v.f.}})$. Then, by Lemma 3.31, there exists an extension of j_0 to a linked log flexible curve

$$j : C_{\text{link}} \to U.$$

Form the fiber product

$$U_{\mathcal{C}} := \mathcal{C} \times_{\pi_{\mathcal{C}}, B, \pi_U} U.$$

Since π_U is flat, also the projection

$$\text{pr}_{\mathcal{C}} : U_{\mathcal{C}} \to \mathcal{C}$$

is flat. Since $\pi_{\mathcal{C}}$ is surjective, the geometric generic fiber of $\text{pr}_{\mathcal{C}}$ equals the geometric generic fiber of π_U, which is integral. Since $\text{pr}_{\mathcal{C}}$ is flat with integral geometric generic fiber, $U_{\mathcal{C}}$ is integral. Define

$$\nu : \widetilde{U}_{\mathcal{C}} \to U_{\mathcal{C}}$$

to be the blowing up of $U_{\mathcal{C}}$ along the closed subscheme $D_{\mathcal{C}} \times_B D_{\pi,\log}$. Since $U_{\mathcal{C}}$ is integral, also $\widetilde{U}_{\mathcal{C}}$ is integral, and the composition

$$\widetilde{U}_{\mathcal{C}} \to U_{\mathcal{C}} \to \mathcal{C} \to \Pi$$

is surjective. Since Π is a smooth curve, the morphism

$$\widetilde{\rho} : \widetilde{U}_{\mathcal{C}} \to \Pi$$

is flat.

Consider the graph

$$\Gamma_j : C_{\text{link}} = \mathcal{C}_0 \to \mathcal{C}_0 \times_B U = U_{\mathcal{C},0}.$$

Because the links of C_{link} do not intersect $D_{\pi,\log}$, the image of Γ_j is smooth at every point of intersection with $D_C \times_B D_{\pi,\log}$. Since ν is birational, Γ_j gives a rational transformation from C_{link} to $\widetilde{U}_{\mathcal{C},0}$. Since ν is proper, and since C_{link} is smooth at every point of intersection with $D_C \times_B D_{\pi,\log}$, the valuative criterion of properness implies this rational transformation is actually a regular morphism

$$\widetilde{\Gamma}_j : \mathcal{C}_0 \to \widetilde{U}_{\mathcal{C},0}.$$

Clearly this is a section of the projection morphism

$$\text{pr}_{\mathcal{C}_0} : \widetilde{U}_{\mathcal{C},0} \to \mathcal{C}_0.$$

For every point t in $C_{\text{link}} - D_C$, the morphism $\pi_U : U \to B$ is smooth at $j(t)$. Therefore also $U_{\mathcal{C},0} \to \mathcal{C}_0$ is smooth at $\Gamma_j(t)$. And since ν is an isomorphism over

$\Gamma_j(t)$, also $\mathrm{pr}_{\mathcal{C}_0} : \widetilde{U}_{\mathcal{C},0} \to \mathcal{C}_0$ is smooth at $\widetilde{\Gamma}_j(t)$. Also the vertical tangent bundle equals the pullback of the vertical tangent bundle of $\pi_U : U \to B$, which also equals $T_{\pi,\log}$ (since $j(t)$ is not in $D_{\pi,\log}$).

Let t be a point of D_C and let D_t be the unique irreducible component of $D_{\pi,\log}$ containing $j(t)$. Give D_t the reduced structure. Because $j_0(C)$ is transverse to D_t at $j_0(t)$, the ramification index $m_C - 1$ of $\pi_C : C \to B$ at t equals the ramification index $m_D - 1$ of π_U along D_t. Therefore, by Lemma 3.26, the projection

$$\mathrm{pr}_{\mathcal{C}_0} : \widetilde{U}_{\mathcal{C},0} \to \mathcal{C}_0$$

is smooth over the preimage of $\{t\} \times D_t$ for every t and the vertical tangent bundle equals the pullback of $T_{\pi,\log}$. Since $\Gamma_j(t)$ is in $\{t\} \times D_t$, this implies that $\mathrm{pr}_{\mathcal{C}_0}$ is smooth at every point of the image of $\widetilde{\Gamma}_j$ and the vertical tangent bundle of $\mathrm{pr}_{\mathcal{C}_0}$ equals the pullback of $T_{\pi,\log}$.

Since $\widetilde{\Gamma}_j$ is a section with image in the smooth locus of $\mathrm{pr}_{\mathcal{C}_0}$, the normal sheaf $\mathcal{N}$ equals the restriction of the vertical tangent bundle. Therefore $\widetilde{\Gamma}_j^*\mathcal{N}$ equals $j^*T_{\pi,\log}$. Since $j : C_{\mathrm{link}} \to U$ is log flexible, $h^1(C_{\mathrm{link}}, j^*T_{\pi,\log})$ equals 0. Therefore, by Proposition 3.13, the relative Hilbert scheme $\mathrm{Hilb}(\widetilde{U}_{\mathcal{C}}/\Pi)$ is smooth over Π at the point $0' := [\mathrm{Image}(\widetilde{\Gamma}_j)]$. Thus for a general complete intersection curve Π' containing $0'$, the morphism $\Pi' \to \Pi$ is smooth at $0'$.

Replace Π' by the unique irreducible component containing $0'$, and then replace this by its normalization. The result is that Π' is a smooth, projective, connected curve together with a morphism $\Pi' \to \mathrm{Hilb}(\widetilde{U}_{\mathcal{C}}/\Pi)$ so that the induced morphism $\Pi' \to \Pi$ is smooth at $0'$. In particular it is flat, so surjective. Let ∞' denote a closed point of Π' mapping to ∞. Then $(\Pi', 0', \infty') \to (\Pi, 0, \infty)$ is a flat morphism of 2-pointed smooth curves. Thus, by Lemma 3.11, the base change $\Pi' \times_\Pi \mathcal{C}$ is a one-parameter deformation of C_{link} over $(\Pi', 0', \infty')$ specializing to a section curve.

Denote by

$$Z \subset \Pi' \times_\Pi \widetilde{U}_{\mathcal{C}}$$

the pullback of the universal closed subscheme $\mathrm{Univ}(\widetilde{U}_{\mathcal{C}}/\Pi)$ by the morphism $\Pi' \to \mathrm{Hilb}(\widetilde{U}_{\mathcal{C}}/\Pi)$. The composition with $\mathrm{pr}_{\mathcal{C}}$ is a projective morphism

$$Z \subset \Pi' \times_\Pi \widetilde{U}_{\mathcal{C}} \to \Pi' \times_\Pi \mathcal{C}$$

of flat Π'-schemes. Moreover, the fiber over $0' \in \Pi'$ is an isomorphism since the projection $\widetilde{\Gamma}_j(C_{\mathrm{link}}) \to C_{\mathrm{link}}$ is an isomorphism. Therefore the morphism is an isomorphism over $N \times_\Pi \mathcal{C}$ for some open neighborhood N of $0'$ in Π'. (This is well-known; a complete proof is given in [**dJS03**, Lemma 4.7].) Invert this isomorphism and compose it with the morphism

$$\Pi' \times_\Pi \widetilde{U}_{\mathcal{C}} \to \widetilde{U}_{\mathcal{C}} \to U_{\mathcal{C}} \to U.$$

The result is precisely an extension

$$j_N : N \times_\Pi \mathcal{C} \to X$$

of j for the one-parameter deformation $\Pi' \times_\Pi \mathcal{C}$. Therefore, by Lemma 3.12, there exists a section $s : B \to X$ of π. □

3.6. Corollaries. There are a number of consequences of Theorem 3.9 and its generalization to positive characteristic in [**dJS03**]. Many of these consequences were recognized before Conjecture 3.8 was proved.

COROLLARY 3.34. [**Kol96**, Conjecture IV.5.6] *Conjecture 3.7 is true. Moreover, for every smooth, projective, irreducible variety X over an algebraically closed field of characteristic 0, there exists a dense open $X^0 \subset X$ and a projective, smooth morphism $q_0 : X^0 \to Q^0$ such that every fiber of q_0 is rationally connected, and every projective closure of Q^0 is nonuniruled.*

COROLLARY 3.35. [**GHS03**, Corollary 1.7] *The uniruledness conjecture implies Mumford's conjecture. To be precise, assume that for every smooth, projective, irreducible variety X over an algebraically closed field k of characteristic 0, if X is nonuniruled then $h^0(X, \omega_X^{\otimes n})$ is nonzero for some $n > 0$. Then for every smooth, projective, irreducible variety X over k, if X is not rationally connected then $h^0(X, \Omega_X^{\otimes n})$ is nonzero for some $n > 0$.*

The next corollary is a fixed point theorem. In characteristic 0 it can be proved using the Atiyah-Bott fixed point theorem. But in positive characteristic it is a new result. There are examples due to Shioda proving one cannot replace "separably rationally connected" by "rationally connected", cf. [**Shi74**].

COROLLARY 3.36. [**Kol03**] *Let Y be a smooth, projective, separably rationally connected variety over a field k and let $f : Y \to Y$ be a k-automorphism. If* $\mathrm{char}(k)$ *is positive, say p, assume in addition that f has finite order n not divisible by p^2. Then the fixed locus of f is nonempty.*

PROOF. Of course it suffices to prove the case when k is algebraically closed, since the fixed locus of the base change equals the base change of the fixed locus. First assume f has finite order n. If n is prime to $\mathrm{char}(k)$, let B' denote $\mathbb{P}^1$ and let $\mathbb{Z}/n\mathbb{Z}$ act on $\mathbb{P}^1$ by multiplication by a primitive n^{th} root of unity. Note that this action fixes ∞ and has trivial generic stabilizer. If $\mathrm{char}(k) = p$ is positive and if $n = pm$ where m is prime to p, let B' be the normal, projective completion of the affine curve

$$\mathbb{V}(y^m - (x^p - x)) \subset \mathbb{A}^2_k.$$

Let ζ be a primitive m^{th} root of unity, and let a generator of $\mathbb{Z}/m\mathbb{Z}$ act by $(x, y) \mapsto (x, \zeta y)$. Similarly, let a generator of $\mathbb{Z}/p\mathbb{Z}$ act by $(x, y) \mapsto (x + 1, y)$. Clearly these actions commute, and thus define an action of $\mathbb{Z}/n\mathbb{Z}$ on B'. Note this action fixes the unique point ∞ not in the affine chart above, and the action has trivial generic stabilizer.

Let $\mathbb{Z}/n\mathbb{Z}$ act diagonally on $Y \times_k B'$, and let X be the quotient. Also let B be the quotient of the $\mathbb{Z}/n\mathbb{Z}$-action on B'. The projection $\pi : X \to B$ satisfies the hypotheses of Theorem 3.9 (or its generalization in [**dJS03**]). Therefore there exists a section. This is the same as a $\mathbb{Z}/n\mathbb{Z}$-equivariant k-morphism $f : B' \to Y$. In particular, since ∞ is a fixed point in B', $f(\infty)$ is a fixed point in Y.

Next assume k has characteristic 0. By general limit arguments there exists an integral, finitely generated $\mathbb{Z}$-algebra R, a ring homomorphism $R \hookrightarrow k$, a smooth, projective morphism $Y_R \to \operatorname{Spec} R$ whose relative very free locus is all of Y_R, and an R-automorphism $f_R : Y_R \to Y_R$ such that the base change $Y_R \otimes_R k$ equals Y and the base change of f_R equals f. The intersection $(Y_R)^{f_R}$ of the graph of f_R and

the diagonal of $Y_R \times_R Y_R$ is the fixed subscheme of f_R (actually its image under the diagonal morphism). Since $(Y_R)^{f_R}$ is a proper scheme over $\operatorname{Spec} R$, the image in $\operatorname{Spec} R$ is a closed subscheme of $\operatorname{Spec} R$. To prove this closed subscheme equals all of $\operatorname{Spec} R$, and thus contains the image of $\operatorname{Spec} k$, it suffices to prove it contains a Zariski dense set of closed points.

Choose an f-invariant very ample sheaf, choose a basis for the space of global sections, and let A be the $N \times N$ matrix with entries in R giving the action of f on global sections with respect to this basis. The set of maximal ideals in $\operatorname{Spec} R$ with residue field of characteristic $p > N$ is Zariski dense in $\operatorname{Spec} R$. Every invertible matrix over a characteristic p field with order divisible by p^2 has a Jordan block with eigenvalue 1 and size divisible by p. Thus, since $p > N$, the finite order of f_R modulo the prime is not divisible by p^2. Therefore, by the previous case, the reduction of f_R modulo the prime has nonempty fixed locus. Therefore the original automorphism f has nonempty fixed locus. □

This fixed point theorem implies that separably rationally connected varieties are simply connected. When the field k is $\mathbb{C}$, this was first proved by Campana using analytic methods, cf. the excellent reference by Debarre, [**Deb01**, Corollary 4.18].

COROLLARY 3.37 (Campana, Kollár). [**Cam91**], [**Deb03**, 3.6] *Let X be a smooth, projective, and separably rationally connected variety over an algebraically closed field k. The algebraic fundamental group of X is trivial. If $k = \mathbb{C}$, then the topological fundamental group of X is also trivial.*

Kollár has generalized this considerably to prove a result for open subschemes of rationally connected varieties, cf. [**Kol03**].

PROOF. The full proof is included in the beautiful survey by Debarre, [**Deb03**, 3.6]. Here is a brief sketch. First of all, for every quasi-projective, (not necessarily separably) rationally chain connected variety, Campana proved that the algebraic fundamental group is finite and also the topological fundamental group is finite when k equals $\mathbb{C}$ (so that the topological fundamental group is defined). Thus the universal cover $\widetilde{X} \to X$ is finite. Since X is smooth, projective and separably rationally connected, also $\widetilde{X}$ is smooth, projective and separably rationally connected. If the fundamental group of X is nonzero, then it contains a cyclic subgroup $\mathbb{Z}/n\mathbb{Z}$ such that p^2 does not divide n. Of course the action of this group on $\widetilde{X}$ is fixed-point-free. But Corollary 3.36 implies there exists a fixed point. Thus X is simply connected. □

Theorem 3.9 also plays an important role in the proof of a "converse" to that same theorem.

THEOREM 3.38. [**GHMS05**] *Let $\pi : X \to B$ be a surjective morphism of normal, projective, irreducible varieties over an algebraically closed field k of characteristic zero. Assume that for some sufficiently large, algebraically closed field extension K/k, for every k-morphism $C \to B$ from a smooth, projective, K-curve to B, the pullback $\pi_C : C \times_B X \to C$ has a section. Then there exists a closed subvariety $Y \subset X$ such that the geometric generic fiber of $\pi|_Y : Y \to B$ is nonempty, irreducible and rationally connected.*

One corollary of this theorem, in fact the motivation for proving it, was to answer a question first asked by Serre and left unresolved by Theorem 3.9: could it be that a smooth, projective variety X over the function field of a curve has a rational point if it is $\mathcal{O}$-*acyclic*, i.e., if $h^i(X, \mathcal{O}_X)$ equals 0 for all $i > 0$? One reason to ask this is that the corresponding question has a positive answer if "function field" is replaced by "finite field" thanks to N. Katz's positive characteristic analogue of the Atiyah-Bott fixed point theorem, [**DK73**, Exposé XXII, Corollaire 3.2], recently generalized by Esnault, [**Esn03**]. Nonetheless, the answer is negative over function fields.

COROLLARY 3.39. [**GHMS05**] *There exists a surjective morphism $\pi : X \to B$ of smooth, projective varieties over $\mathbb{C}$ such that B is a curve and the geometric generic fiber of π is an Enriques surface, but π has no section. Thus, to guarantee that a fibration over a curve has a section, it is not sufficient to assume the geometric generic fiber is $\mathcal{O}$-acyclic.*

In fact G. Lafon found an *explicit* morphism π as in Corollary 3.39 where B is $\mathbb{P}^1_{\mathbb{C}}$, or in fact $\mathbb{P}^1_k$ for any field k with $\mathrm{char}(k) \neq 2$, and there does not even exist a power series section near $0 \in \mathbb{P}^1_k$, cf. [**Laf04**].

4. The Period-Index theorem

Theorem 3.9 is a generalization of Tsen's theorem, Corollary 2.15, because a sufficiently general complete intersection $\mathbb{V}(F_1, \dots, F_r) \subset \mathbb{P}^n$ with $d_1 + \cdots + d_r \leq n$ is smooth, projective and separably rationally connected (the proof of this is nontrivial, as is the specialization argument reducing Tsen's theorem to the case of complete intersections which are sufficiently general). Is there a similar generalization of the Tsen-Lang theorem?

Joint work with Harris, [**HS05**], proves that the spaces of rational curves on general low degree hypersurfaces are rationally connected. This was later generalized in joint work with A. J. de Jong: complete intersections $X = \mathbb{V}(F_1, \dots, F_r) \subset \mathbb{P}^n$ with $d_1^2 + \cdots + d_r^2 \leq n + 1$ are rationally simply connected in the sense that the space of "good" rational curves in X containing two fixed, general points is itself a rationally connected variety. This is analogous to simple connectedness in topology: a path connected topological space is simply connected if the space of paths connecting two fixed points is itself path connected.

Moreover, de Jong gave a heuristic argument suggesting that for a rationally simply connected fibration over a surface, the only obstruction to existence of a rational section is the elementary obstruction. Given a geometrically integral scheme X defined over a field K, the *elementary obstruction* to existence of a K-point is the existence of a $\mathrm{Gal}(\overline{K}/K)$-invariant splitting of the homomorphism of Abelian Galois modules,

$$\overline{K}^* \hookrightarrow \mathrm{Frac}(X \otimes_K \overline{K})^*,$$

where $\overline{K}$ is the separable closure of K and Frac is the function field. If there exists a K-point of X, evaluation at this point gives a Galois-invariant splitting. The elementary obstruction was introduced by Colliot-Thélène and Sansuc, [**CTS87**].

Its vanishing implies the vanishing of other known obstructions. In particular, it implies the vanishing of a Brauer obstruction

$$\delta : \mathrm{Pic}(X \otimes_K \overline{K})^{\mathrm{Gal}(\overline{K}/K)} \to \mathrm{Br}(K)$$

measuring whether or not a Galois-invariant invertible sheaf $\overline{\mathcal{L}}$ on $X \otimes_K \overline{K}$ is the pullback of an invertible sheaf $\mathcal{L}$ on X.

At the moment, in order to give a rigorous proof, de Jong's heuristic argument requires several additional hypotheses on the rationally simply connected fibration. One case where the hypotheses hold is when all fibers of the fibration are Grassmannian varieties. Although this is very special, it is also quite interesting since it gives a second proof of de Jong's *period-index theorem*.

THEOREM 4.1. [**dJS05**] *Let K be the function field of a surface over an algebraically closed field k. Let $(X, \mathcal{L})$ be a pair of a K-scheme and an invertible $\mathcal{O}_X$-module $\mathcal{L}$. If $(X \otimes_K \overline{K}, \mathcal{L} \otimes_K \overline{K})$ is isomorphic to $(\mathrm{Grass}(r, \overline{K}^n), \mathcal{O}(1))$, then X has a K-point.*

COROLLARY 4.2 (de Jong's Period-Index theorem). [**dJ04**] *Let A be a central simple K-algebra with $A \otimes_K \overline{K} \cong \mathrm{Mat}_{n\times n}(\overline{K})$. Let $r < n$ be an integer such that $r[A]$ equals 0 in $\mathrm{Br}(K)$. Then there exists a left ideal $I \subset A$ such that $\dim_K(I) = rn$. In particular, if $A = D$ is a division algebra then $[D]$ has order n in $\mathrm{Br}(K)$, i.e., the period of D equals the index of D.*

Corollary 4.2 follows from Theorem 4.1 by setting X to be the *generalized Brauer-Severi variety* parameterizing left ideals in A of rank rn. Since $A \otimes_K \overline{K}$ equals $\mathrm{Mat}_{n\times n}(\overline{K})$, $X \otimes_K \overline{K}$ equals $\mathrm{Grass}(r, \overline{K}^n)$. The Brauer obstruction to the existence of an invertible sheaf $\mathcal{L}$ with $\mathcal{L} \otimes_K \overline{K} \cong \mathcal{O}(1)$ is precisely the element $r[A]$ in $\mathrm{Br}(K)$.

The first reduction is "discriminant avoidance", i.e., reduction to the case that the variety X is the generic fiber of a smooth, projective morphism over a smooth, projective surface. Let T be a quasi-compact, integral scheme and let $\mathcal{G}$ be a smooth, affine group scheme over T whose geometric fibers are reductive.

LEMMA 4.3. *For every integer c there exists a datum $(U, \overline{U}, \mathcal{T}_U)$ of a projective, flat T-scheme $\overline{U}$ with integral geometric fibers, an open subset U of $\overline{U}$ and a $\mathcal{G}$-torsor $\mathcal{T}_U$ over U such that*

(i) *U is smooth over T,*
(ii) *the complement $\overline{U} - U$ has codimension $\geq c$ in $\overline{U}$,*
(iii) *and for every $\mathcal{G}$-torsor $\mathcal{T}_K$ over an infinite field K over T, there exists a T-morphism $i : \mathrm{Spec}\, K \to U$ and an isomorphism of $\mathcal{G}$-torsors over K, $i^*\mathcal{T}_U \cong \mathcal{T}_K$.*

The idea is to form the GIT quotient $\overline{U}$ of a linear action of $\mathcal{G}$ on $(\mathbb{P}^N_T, \mathcal{O}(1))$. If the linear representation is "sufficiently large", then $\mathcal{G}$ acts properly and freely on an open subset V of $\mathbb{P}^N_T$ of codimension $\geq c$. Take U to be the quotient of V. Then U is smooth over T and $\overline{U} - U$ has codimension $\geq c$. Finally, for every field K over $\mathcal{O}_T$ and for every $\mathcal{G}$-torsor $\mathcal{T}_K$ over K, the twist $\mathbb{P}^N_T \times_T \mathcal{T}_K/\mathcal{G}$ is isomorphic to $\mathbb{P}^N_K$. Thus there exists a K-point. If K is infinite, then the set of K-points is Zariski dense so that there exists a point in the image of V. This point is only well-defined up to the action of $\mathcal{G}$, but the associated morphism $i : \mathrm{Spec}\, K \to U$ is well-defined. Chasing definitions, $i^*\mathcal{T}_U$ is isomorphic to $\mathcal{T}_K$.

PROPOSITION 4.4. *Let k be an algebraically closed field of characteristic 0, let S be a smooth, projective surface over k, let $\pi : \mathcal{X} \to S$ be a smooth, projective morphism, and let $\mathcal{L}$ be an invertible $\mathcal{O}_{\mathcal{X}}$-module. Let K/k be the fraction field of S and let X be the generic fiber of π.*

(i) *If $(X \otimes_K \overline{K}, \mathcal{L} \otimes_K \overline{K})$ is isomorphic to $(\mathrm{Grass}(r, \overline{K}^n), \mathcal{O}(1))$, then X has a K-point.*
(ii) *Item (i) implies Theorem 4.1*

The point of this proposition is Item (ii), i.e., to prove Theorem 4.1 it suffices to assume that X is the generic fiber of a proper and *everywhere smooth* morphism. The point is that for every algebraically closed field k, there exists the spectrum of a DVR, T, whose residue field is k and whose fraction field has characteristic 0. Now let $\mathcal{G}$ be the automorphism group scheme of $(\mathrm{Grass}(r, \mathcal{O}_T^n), \mathcal{O}(1))$. This satisfies the hypotheses of Lemma 4.3. Taking $c = 3$, there exists a datum $(U, \overline{U}, \mathcal{T}_U)$ as in Lemma 4.3 such that $\overline{U} - U$ has codimension ≥ 3. For the original field K/k, there exists a morphism $i : \mathrm{Spec}\, K \to U$ inducing the pair $(X, \mathcal{L})$. Because $\mathrm{tr.deg.}(K/k) = 2$, the closure of $\mathrm{Image}(i)$ in $\overline{U}$ has dimension ≤ 2. Because $\overline{U} - U$ has codimension ≥ 3, there exists a locally closed subscheme $\mathcal{S}$ of $\overline{U}$ such that

(i) $\mathcal{S} \to T$ is flat,
(ii) the closed fiber S_0 of $\mathcal{S}$ is irreducible with generic point $i(\mathrm{Spec}\, K)$,
(iii) and the generic fiber $\mathcal{S}_\eta$ of $\mathcal{S}$ is a closed subscheme of $\overline{U}_\eta$ completely contained in U_η.

Using specialization arguments, to prove Theorem 4.1 for the restriction of $\mathcal{T}_U$ to the generic point of the closed fiber S_0, it suffices to prove Theorem 4.1 for the restriction of $\mathcal{T}_U$ to the generic point of the geometric generic fiber $S_{\overline{\eta}}$. By construction, this satisfies the additional hypotheses in Proposition 4.4.

Thus, assume the additional hypotheses of Proposition 4.4 are satisfied. After replacing S by the blowing up at the base locus of a Lefschetz pencil of divisors, and replacing $\mathcal{X}$ by its base change, assume there exists a flat, proper morphism $\rho : S \to B$ with smooth, connected generic fiber. Denote by B^0 the maximal open subscheme of B over which ρ is smooth, and denote $S^0 = B^0 \times_B S$ and $\mathcal{X}^0 = B^0 \times_B \mathcal{X}$. There is a pair

$$(\rho_{\mathrm{Sect}} : \mathrm{Section}(\mathcal{X}^0/S^0/B^0) \to B^0, \sigma : \mathrm{Section}(\mathcal{X}^0/S^0/B^0) \times_{B^0} S^0 \to \mathcal{X}^0)$$

which is universal among all pairs (T, σ_T) of a B^0-scheme T and an S_0-morphism $\sigma_T : T \times_{B^0} S^0 \to \mathcal{X}^0$. The universal pair can be constructed in terms of the relative Hilbert scheme. In Grothendieck's terminology, it is $\Pi_{S^0/B^0}\mathcal{X}_0$, cf. [**Gro62**, p. 195-13].

There is a relative Picard scheme $\mathrm{Pic}(S^0/B^0)$ of S^0 over B^0. Associated to the invertible sheaf $\mathcal{L}$ on $\mathcal{X}^0$, there is an invertible sheaf $\sigma^*\mathcal{L}$ on $\mathrm{Section}(\mathcal{X}^0/S^0/B^0) \times_{B^0} S^0$. This induces an *Abel morphism*

$$\alpha : \mathrm{Section}(\mathcal{X}^0/S^0/B^0) \to \mathrm{Pic}(S^0/B^0).$$

Of course $\mathrm{Pic}(S^0/B^0)$ breaks up according the relative degree of the line bundle,

$$\mathrm{Pic}(S^0/B^0) = \coprod_{d \in \mathbb{Z}} \mathrm{Pic}^d(S^0/B^0).$$

Pulling this back via the Abel morphism gives a decomposition

$$\mathrm{Section}(\mathcal{X}^0/S^0/B^0) = \coprod_{d\in\mathbb{Z}} \mathrm{Section}^d(\mathcal{X}^0/S^0/B^0)$$

together with Abel morphisms

$$\alpha_d : \mathrm{Section}^d(\mathcal{X}^0/S^0/B^0) \to \mathrm{Pic}^d(S^0/B^0).$$

For all $d \geq 0$, there are sections of the projection $\mathrm{Pic}^d(S^0/B^0) \to B^0$. The image of this section is a curve B_0' isomorphic to the smooth curve B^0. If the generic fiber of α_d is a dense open subset of a rationally connected variety, then Theorem 3.9 together with the generic version of weak approximation, Proposition 3.19, implies there exists a rational section of the restriction of α_d over B_0' (this uses a slight specialization argument, because the restriction of α_d may not be a rationally connected fibration). Thus it suffices to prove that for $d \gg 0$,

(i) the fiber of α_d over the geometric generic point of $\mathrm{Pic}^d(S^0/B^0)$ is not empty,
(ii) the fiber is also irreducible,
(iii) the fiber is also isomorphic to an open subset of a rationally connected variety.

Moreover, and this will be important, it suffices to prove there exists a canonically defined open subscheme $W \subset \mathrm{Section}^d(\mathcal{X}^0/S^0/B^0)$ such that (i)–(iii) hold for $\alpha_d|_W$.

Note that (i)–(iii) are really statements about the morphism $\mathcal{X}_{\overline{\eta_B}} \to S_{\overline{\eta_B}}$ of fibers over the geometric generic point of B. Thus, it is again a question about a fibration over a projective curve, namely the curve $C = S_{\overline{\eta_B}}$ over the algebraically closed field $\kappa = \overline{k(B)}$. So Proposition 4.4 (i) follows from the following result.

PROPOSITION 4.5. *Let κ be an algebraically closed field of characteristic 0. Let C be a smooth, projective, connected curve over κ. Let $\pi : \mathcal{X}_C \to C$ be a smooth, projective morphism and let $\mathcal{L}$ be an invertible sheaf on $\mathcal{X}_C$. Assume the geometric generic fiber of $(\mathcal{X}_C, \mathcal{L})$ over C is isomorphic to $(\mathrm{Grass}(r, \overline{\kappa(C)}^n), \mathcal{O}(1))$. Then for $d \gg 0$ there exists a canonically defined open subset*

$$W_d \subset \mathrm{Section}^d(\mathcal{X}_C/C/\mathrm{Spec}\,\kappa)$$

such that the geometric generic fiber of

$$\alpha_d : W_d \hookrightarrow \mathrm{Section}^d(\mathcal{X}_C/C/\mathrm{Spec}\,\kappa) \to \mathrm{Pic}^d(C/\mathrm{Spec}\,\kappa)$$

satisfies (i), (ii) and (iii) above.

In order to prove Corollary 4.2, it suffices to prove this in the special case that $\mathcal{X}_C \to C$ is the parameter scheme for rank rn left ideals in an Azumaya algebra $\mathcal{A}$ over C with $\mathcal{A} \otimes_{\mathcal{O}_C} \overline{\kappa(C)} \cong \mathrm{Mat}_{n\times n}(\overline{\kappa(C)})$. An *Azumaya algebra* over a scheme T is a coherent $\mathcal{O}_T$-algebra which is étale locally isomorphic to $\mathrm{Mat}_{n\times n}(\mathcal{O}_T)$ for some integer n.

Because of Tsen's theorem, Corollary 2.15, and Proposition 2.16(i), there exists a locally free $\mathcal{O}_C$-module $\mathcal{E}$ of rank n such that $\mathcal{A} \cong \mathit{End}(\mathcal{E})$. The locally free sheaf $\mathcal{E}$ is only well-defined up to the operation $\mathcal{E} \mapsto \mathcal{E} \otimes_{\mathcal{O}_C} \mathcal{N}$, for any invertible sheaf $\mathcal{N}$. The choice of a locally free sheaf $\mathcal{E}$ and an isomorphism of algebras gives an isomorphism of $\mathrm{Sect}(\mathcal{X}_C/C/\mathrm{Spec}\,\kappa)$ with the parameter scheme of locally free

quotients $\mathcal{E} \twoheadrightarrow \mathcal{Q}$ of rank r. To see this, associate to each quotient the left ideal of endomorphisms that factor as

$$\mathcal{E} \twoheadrightarrow \mathcal{Q} \xrightarrow{\phi} \mathcal{E}$$

as ϕ varies over all $\mathcal{O}_C$-module homomorphisms. Replacing $\mathcal{E}$ by $\mathcal{E} \otimes_{\mathcal{O}_C} \mathcal{N}$ gives a new isomorphism sending the original quotient to the twist $\mathcal{E} \otimes_{\mathcal{O}_C} \mathcal{N} \twoheadrightarrow \mathcal{Q} \otimes_{\mathcal{O}_C} \mathcal{N}$. The real effect of this change has to do with the Abel map. Up to a constant translation by a point of $\mathrm{Pic}(C/\mathrm{Spec}\,\kappa)$, which does not change (i)–(iii), the Abel map is identified with the map sending a quotient to $\det(\mathcal{Q})$ in $\mathrm{Pic}(C/\mathrm{Spec}\,\kappa)$. After replacing $\mathcal{E}$ by $\mathcal{E} \otimes_{\mathcal{O}_C} \mathcal{N}$, the constant changes by adding $[\mathcal{N}^{\otimes r}]$. Thus, it is clear that the Abel map really should only be considered as well-defined up to an additive constant.

If d is sufficiently large, there exist quotients such that $\mathcal{Q}$ is stable. Note that $\mathcal{Q}$ is stable if and only if $\mathcal{Q} \otimes_{\mathcal{O}_C} \mathcal{N}$ is stable. Therefore the open subset W_d of $\mathrm{Section}(\mathcal{X}_C/C/\mathrm{Spec}\,\kappa)$ parameterizing stable quotients is well-defined and canonical. Fix an integer d_0. For every integer e, the moduli space of stable, rank r locally free sheaves on C of degree d_0 is isomorphic to the moduli space for degree $d_0 + re$ via the map sending $\mathcal{Q}$ to $\mathcal{Q}(D)$, where D is any fixed Cartier divisor of degree e. For each fixed locally free sheaf $\mathcal{Q}$ of rank $r < n$, if e is sufficiently large, there exists a surjection $\mathcal{E} \twoheadrightarrow \mathcal{Q}(D)$ for all Cartier divisors of degree e. Because the moduli space of stable bundles is quasi-compact, it follows that there exists a single integer e_0 such that for every $e \geq e_0$ and every stable locally free sheaf $\mathcal{Q}$ of rank r and degree d_0, for every Cartier divisor D of degree e there exists a surjection $\mathcal{E} \twoheadrightarrow \mathcal{Q}(D)$. By the same sort of argument, if e is sufficiently large then $h^1(C, \mathit{Hom}_{\mathcal{O}_C}(\mathcal{E}, \mathcal{Q}(D)))$ equals 0.

Repeating this argument with d_0 replaced by each of $d_0, d_0+1, d_0+2, \ldots, d_0+r-1$, there exists an integer d_1 such that for every $d \geq d_1$ (i.e., $d = d_0 + re$, etc.), for every stable, locally free sheaf $\mathcal{Q}$ of rank r and degree d, there exists a surjection

$$\mathcal{E} \twoheadrightarrow \mathcal{Q}$$

and also $h^1(C, \mathit{Hom}_{\mathcal{O}_C}(\mathcal{E}, \mathcal{Q}))$ equals 0. In other words, the forgetful morphism from the space of quotients $\mathcal{E} \twoheadrightarrow \mathcal{Q}$ to the space of stable sheaves $\mathcal{Q}$ is smooth and surjective, and the geometric fibers are each isomorphic to an open subset of an affine space $\mathrm{Hom}_{\mathcal{O}_C}(\mathcal{E}, \mathcal{Q})$. Moreover, the fiber of the Abel map α_d is the inverse image of the space of stable sheaves with fixed determinant. As is well-known, the moduli space of stable sheaves over C of fixed rank r and fixed determinant is a unirational variety of dimension $(r^2-1)(g(C)-1)$. Thus the fiber of the Abel map fibers over a rationally connected variety and the fibers are rationally connected. By Corollary 3.34, it follows that a general fiber of

$$\alpha_d|_W : W_d \hookrightarrow \mathrm{Section}^d(\mathcal{X}_C/C/\mathrm{Spec}\,\kappa) \to \mathrm{Pic}^d(C/\mathrm{Spec}\,\kappa)$$

is isomorphic to an open subset of a rationally connected variety, i.e., (i), (ii) and (iii) hold. Therefore Corollary 4.2 is true.

References

[Art76] M. Artin, *Lectures on deformations of singularities*, Lectures on Mathematics and Physics, vol. 54, Tata Institue of Fundamental Research, Bombay, 1976.

[Cam91] F. Campana, *On twistor spaces of the class $\mathcal{C}$*, J. Differential Geom. **33** (1991), no. 2, 541–549. MR 1094468 (92g:32059)

[Che35] C. Chevalley, *Demonstration d'une hypothèse de E. Artin*, Abh. Math. Sem. Hansischen Univ. **11** (1935), 73.

[CT87] J.-L. Colliot-Thélène, *Arithmétique des variétés rationnelles et problèmes birationnels*, Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Berkeley, Calif., 1986) (Providence, RI), Amer. Math. Soc., 1987, pp. 641–653. MR 934267 (89d:11051)

[CTS87] J.-L. Colliot-Thélène and J.-J. Sansuc, *La descente sur les variétés rationnelles. II*, Duke Math. J. **54** (1987), no. 2, 375–492. MR 899402 (89f:11082)

[Deb01] O. Debarre, *Higher-dimensional algebraic geometry*, Universitext, Springer-Verlag, New York, 2001. MR 1841091 (2002g:14001)

[Deb03] ———, *Variétés rationnellement connexes (d'après T. Graber, J. Harris, J. Starr et A. J. de Jong)*, Astérisque (2003), no. 290, Exp. No. 905, ix, 243–266, Séminaire Bourbaki. Vol. 2001/2002. MR 2074059

[dJ04] A. J. de Jong, *The period-index problem for the Brauer group of an algebraic surface*, Duke Math. J. **123** (2004), no. 1, 71–94. MR 2060023 (2005e:14025)

[dJS03] A. J. de Jong and J. Starr, *Every rationally connected variety over the function field of a curve has a rational point*, Amer. J. Math. **125** (2003), no. 3, 567–580. MR 1981034 (2004h:14018)

[dJS05] A. J. de Jong and J. Starr, *Almost proper GIT-stacks and discriminant avoidance*, preprint, available `http://www.math.columbia.edu/~dejong/`, 2005.

[DK73] P. Deligne and N. Katz, *Groupes de monodromie en géométrie algébrique. II*, Lecture Notes in Mathematics, vol. 340, Springer-Verlag, Berlin, 1973, Séminaire de Géométrie Algébrique du Bois-Marie 1967–1969 (SGA 7 II), Dirigé par P. Deligne et N. Katz. MR 0354657 (50 #7135)

[Esn03] H. Esnault, *Varieties over a finite field with trivial Chow group of 0-cycles have a rational point*, Invent. Math. **151** (2003), no. 1, 187–191. MR 1943746 (2004e:14015)

[GHMS04a] T. Graber, J. Harris, B. Mazur, and J. Starr, *Arithmetic questions related to rationally connected varieties*, The legacy of Niels Henrik Abel, Springer, Berlin, 2004, pp. 531–542. MR 2077583 (2005g:14097)

[GHMS04b] ———, *Jumps in Mordell-Weil rank and arithmetic surjectivity*, Arithmetic of higher-dimensional algebraic varieties (Palo Alto, CA, 2002), Progr. Math., vol. 226, Birkhäuser Boston, Boston, MA, 2004, pp. 141–147. MR 2029867 (2005d:14035)

[GHMS05] ———, *Rational connectivity and sections of families over curves*, Ann. Sci. École Norm. Sup. (4) **38** (2005), no. 5, 671–692. MR 2195256 (2006j:14044)

[GHS02] T. Graber, J. Harris, and J. Starr, *A note on Hurwitz schemes of covers of a positive genus curve*, preprint, 2002.

[GHS03] ———, *Families of rationally connected varieties*, J. Amer. Math. Soc. **16** (2003), no. 1, 57–67 (electronic). MR 1937199 (2003m:14081)

[Gro62] A. Grothendieck, *Fondements de la géométrie algébrique. [Extraits du Séminaire Bourbaki, 1957–1962.]*, Secrétariat mathématique, Paris, 1962. MR 0146040 (26 #3566)

[Gro67] ———, *Éléments de géométrie algébrique. IV. Étude locale des schémas et des morphismes de schémas.*, Inst. Hautes Études Sci. Publ. Math. 20 (1964), 101-355; ibid. 24 (1965), 5-231; ibid. 28 (1966), 5-255; ibid. (1967), no. 32, 5–361, `http://www.numdam.org/item?id=PMIHES_1965__24__5_0`. MR 0173675 (30 #3885), 0199181 (33 #7330), 0217086 (36 #178), 0238860 (39 #220)

[Har77] R. Hartshorne, *Algebraic geometry*, Springer-Verlag, New York, 1977, Graduate Texts in Mathematics, No. 52. MR 0463157 (57 #3116)

[HS05] J. Harris and J. Starr, *Rational curves on hypersurfaces of low degree. II*, Compos. Math. **141** (2005), no. 1, 35–92. MR 2099769 (2006c:14039)

[HT06] B. Hassett and Y. Tschinkel, *Weak approximation over function fields*, Invent. Math. **163** (2006), no. 1, 171–190. MR 2208420 (2007b:14109)

[Hur91] A. Hurwitz, *Ueber Riemann'sche Flächen mit gegebenen Verzweigungspunkten*, Mathematische Annalen **39** (1891), 1–61.

[KMM92] J. Kollár, Y. Miyaoka, and S. Mori, *Rationally connected varieties*, J. Algebraic Geom. **1** (1992), no. 3, 429–448. MR 1158625 (93i:14014)

[Kol96] J. Kollár, *Rational curves on algebraic varieties*, Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. A Series of Modern Surveys in Mathematics [Results in

Mathematics and Related Areas. 3rd Series. A Series of Modern Surveys in Mathematics], vol. 32, Springer-Verlag, Berlin, 1996. MR 1440180 (98c:14001)

[Kol03] ———, *Rationally connected varieties and fundamental groups*, Higher dimensional varieties and rational points (Budapest, 2001), Bolyai Soc. Math. Stud., vol. 12, Springer, Berlin, 2003, pp. 69–92. MR 2011744 (2005g:14042)

[Laf04] G. Lafon, *Une surface d'Enriques sans point sur* $\mathbb{C}((t))$, C. R. Math. Acad. Sci. Paris **338** (2004), no. 1, 51–54. MR 2038084 (2004k:14035)

[Lan52] S. Lang, *On quasi algebraic closure*, Ann. of Math. (2) **55** (1952), 373–390. MR 0046388 (13,726d)

[Man86] Yu. I. Manin, *Cubic forms*, second ed., North-Holland Mathematical Library, vol. 4, North-Holland Publishing Co., Amsterdam, 1986, Algebra, geometry, arithmetic, Translated from the Russian by M. Hazewinkel. MR 833513 (87d:11037)

[Ser02] J.-P. Serre, *Galois cohomology*, English ed., Springer Monographs in Mathematics, Springer-Verlag, Berlin, 2002, Translated from the French by Patrick Ion and revised by the author. MR 1867431 (2002i:12004)

[Shi74] T. Shioda, *An example of unirational surfaces in characteristic p*, Math. Ann. **211** (1974), 233–236. MR 0374149 (51 #10349)

[Sus84] A. A. Suslin, *Algebraic K-theory and the norm residue homomorphism*, Current problems in mathematics, Vol. 25, Itogi Nauki i Tekhniki, Akad. Nauk SSSR Vsesoyuz. Inst. Nauchn. i Tekhn. Inform., Moscow, 1984, pp. 115–207. MR 770942 (86j:11121)

[Tse33] C. Tsen, *Divisionalgebren über Funktionenkörper*, Nachr. Ges. Wiss. Göttingen (1933), 335.

[Tse36] ———, *Quasi-algebraische-abgeschlossene Funktionenkörper*, J. Chinese Math. **1** (1936), 81–92.

[War35] E. Warning, *Bemerkung zur vorstehenden Arbeit von Herrn Chevalley.*, Abhandl. Hamburg **11** (1935), 76–83 (German).

Department of Mathematics, Stony Brook University, Stony Brook, NY 11794
E-mail address: jstarr@math.sunysb.edu

Clay Mathematics Proceedings
Volume 8, 2009

Galois + Équidistribution = Manin-Mumford

Nicolas Ratazzi et Emmanuel Ullmo

Résumé. Ce texte est une version rédigée du premier exposé donné par le second auteur durant l'école d'été "Arithmetic Geometry" à l'université de Göttingen à l'été 2006. Les exposés avaient pour but de donner une idée de la preuve récente de la conjecture d'André–Oort sous l'hypothèse de Riemann généralisée due à Klingler, Yafaev et le second auteur.

1. Introduction

Le texte qui suit est une version rédigée du premier exposé donné par le second auteur durant l'école d'été "Arithmetic Geometry" à l'université de Göttingen à l'été 2006. C'est un plaisir d'avoir l'occasion de remercier les organisateurs pour leur invitation. Les exposés avaient pour but de donner une idée de la preuve récente de la conjecture d'André–Oort sous l'hypothèse de Riemann généralisée due à Klingler, Yafaev et le second auteur [**UY06**], [**KY06**].

La conjecture d'André–Oort est un analogue pour les variétés de Shimura de la conjecture de Manin-Mumford démontré par Raynaud [**Ray83**]. Il était donc naturel d'essayer d'adapter la stratégie de preuve de la conjecture d'André-Oort dans le cas des variétés abéliennes. Le texte qui suit propose cette traduction et donne donc une démonstration de la conjecture de Manin-Mumford.

Rappelons tout d'abord l'énoncé de la conjecture de Manin-Mumford.

Théorème 1.1. *Soient K un corps de nombres, A/K une variété abélienne sur K et V/K une sous-variété géométriquement irréductible de A. Si $V(\overline{K})$ contient un ensemble de points de torsion dense dans V pour la topologie de Zariski alors V est un translaté par un point de torsion d'une sous-variété abélienne.*

De nombreuses preuves de cette conjecture ont été obtenues. La première preuve donnée par Raynaud [**Ray83**] utilise des méthodes p-adiques. Hindry [**Hin88**] donne une preuve utilisant la théorie de Galois et l'approximation diophantienne. Hrushovski montre la conjecture en utilisant des idées provenant de la logique (théorie des modèles des corps). Pink et Roessler [**PR02, PR04**] donnent une preuve par des techniques de géométrie algébrique qui s'inspire de la preuve de Hrushovski. Enfin une preuve utilisant la théorie d'Arakelov via l'équidistribution des orbites sous Galois des points de petite hauteur d'une conjecture plus forte due

2000 *Mathematics Subject Classification.* Primary 11G10, 11G18, Secondary 14G35, 14K15.

à Bogomolov est obtenue par Zhang [**Zha98**] et le deuxième auteur de cette note [**Ull98**].

La preuve présentée ici s'inspire des récentes stratégies de preuve de la conjecture d'André-Oort, qui est un analogue dans le cadre des variétés de Shimura de la conjecture de Manin-Mumford. Dans ce cadre les méthodes galoisiennes dues à Edixhoven et Yafaev [**EY03**] se combinent aux méthodes issues de la théorie ergodique développées par Clozel et le second auteur [**CU05**]. Cette stratégie est expliquée de manière générale dans un travail récent de Yafaev et le second auteur [**UY06**] et présentée dans un cas particulier simple de la conjecture d'André-Oort sous l'hypothèse de Riemann dans ce volume [**UY**].

Dans le cadre des variétés abéliennes nous combinons des résultats galoisiens dus à Serre à des techniques élémentaires d'équidistribution de sous-variétés abéliennes. Les résultats galoisiens utilisés ici sont au centre de la méthode de Hindry et la preuve donnée ici ne peut qu'être considérée comme une variante de la preuve de Hindry. Il est notable que la traduction dans le cadre abélien des idées d'Edixhoven et Yafaev [**EY03**] donne assez naturellement la preuve de Hindry de la conjecture de Manin-Mumford et qu'il n'est pas utile d'utiliser les techniques ergodiques dans ce cadre. Nous ne savons pas si il est possible de s'en dispenser aussi dans une preuve de la conjecture d'André-Oort.

Nous espérons que la preuve de la conjecture de Manin-Mumford présentée dans cette note pourra aider le lecteur intéressé par la conjecture d'André-Oort à comprendre la stratégie mise en oeuvre pour les variétés de Shimura. Pour rendre la présentation plus agréable et naturelle nous avons utilisé un résultat non publié de Daniel Bertrand d'effectivité dans le lemme de Poincaré pour les variétés abéliennes. Nous le remercions pour les notes qu'il a eu la gentillesse de nous transmettre ainsi que pour la permission de présenter ici ses résultats que nous avons insérés en appendice.

1.1. Notations et conventions. On dira que V/k est une *variété (définie) sur un corps* k si V est un k-schéma de type fini, géométriquement réduit. Si V/k est une variété définie sur un corps k et si L est une k-algèbre, on notera V_L la variété produit fibré de V et $\mathrm{Spec}\,(L)$ au dessus de $\mathrm{Spec}\,(k)$.

Dans toute la suite, on fixe $A/\overline{\mathbb{Q}}$ une variété abélienne. On se donne K un corps de nombres sur lequel A est définie ainsi que toutes ses sous-variétés abéliennes (un tel corps existe et peut-être choisi de degré $3^{(2\dim A)^4}$ sur un corps de définition de A, cf. par exemple [**MW93**] lemma 2.2.). On se donne également un fibré en droites $\mathcal{L}$ très ample sur A/K de sorte à avoir une notion de degré projectif deg relativement à $\mathcal{L}$.

Si E est un ensemble de points de $A(\overline{\mathbb{Q}})$, on notera $\overline{E}$ son adhérence de Zariski dans A/K. Enfin on notera A_{tors} l'ensemble des points de torsion de $A(\overline{\mathbb{Q}})$ et, si V est une sous-variété de A, on notera V_{tors} l'ensemble $V(\overline{\mathbb{Q}}) \cap A_{\mathrm{tors}}$ des points de torsion de A situés sur V.

Définition 1.1. Soit $V/\overline{\mathbb{Q}}$ une sous-variété irréductible de A. On dit que V est une *variété de torsion* s'il existe une sous-variété abélienne B de A et un point de torsion $\xi \in A_{\mathrm{tors}}$ tels que $V = B + \xi$.

Une variété V/K définie sur un corps de nombres K est dite *de torsion* si $V_{\overline{\mathbb{Q}}}$ est une réunion de sous-variétés de torsion.

On utilisera les symboles $\ll$ et $\gg$ pour dire inférieur à (respectivement supérieur à), à une constante ne dépendant que de $(A, K, \mathcal{L}, V)$ près.

2. Effectivité dans le lemme de Poincaré

Supposant connu le lemme de réductibilité de Poincaré, qui affirme que toute variété abélienne est isogène à un produit de variétés abéliennes simples, nous donnons ici une version effective de ce résultat due à Bertrand. Cet énoncé (plus fort que ce dont nous aurions réellement besoin) permet de présenter les choses de façon naturelle dans la preuve du résultat principal (cf. la remarque 3.1. du paragraphe 3). Nous en donnons en appendice la preuve, telle qu'on la trouve dans l'appendice de [**Ber**].

PROPOSITION 2.1. **(Bertrand)** *Pour toute sous-variété abélienne B de A, il existe une sous-variété abélienne B' de A telle que*

$$A = B + B' \text{ et telle que } \operatorname{card}(B \cap B') \ll 1.$$

3. Preuve du théorème 1.1

Notons

$$\Sigma_V = \left\{ X/\overline{\mathbb{Q}} \text{ sous-variété de torsion de } A \ \mid X \subset V_{\overline{\mathbb{Q}}} \right\}.$$

Définition 3.1. Une suite $(\Sigma_n)_{n\in\mathbb{N}}$ de Σ_V est *générique pour V* si pour toute sous-variété W de $V_{\overline{\mathbb{Q}}}$, distincte de $V_{\overline{\mathbb{Q}}}$, l'ensemble $\{n \in \mathbb{N} \mid \Sigma_n \subset W\}$ est fini.

Soit $(\Sigma_n)_{\in\mathbb{N}}$ une suite générique pour V (une telle suite existe d'après l'hypothèse faite sur V_{tors}. En fait il existe même une suite générique constituée de points de torsion.). Pour tout entier $n \in \mathbb{N}$ choisissons arbitrairement une représentation

$$\Sigma_n = A_n + \xi_n$$

où A_n est une sous-variété abélienne de A et ξ_n un point de torsion de A.

En notant, pour tout $n \in \mathbb{N}$, A'_n la variété associée à A_n par la proposition 2.1, on voit qu'il existe $(a_n, a'_n) \in A_n \times A'_n$ de torsion tels que $\xi_n = a_n + a'_n$. Quitte à remplacer ξ_n par a'_n on peut supposer (et nous le ferons dans la suite) que

$$\forall n \in \mathbb{N}, \quad \Sigma_n = A_n + \xi_n \quad \text{ avec } \xi_n \text{ de torsion dans } A'_n.$$

Pour tout entier $n \in \mathbb{N}$, on note d_n l'ordre du point ξ_n ainsi obtenu. Deux situations peuvent alors apparaître :

1. La suite $(d_n)_{n\in\mathbb{N}}$ est bornée. Dans ce cas un argument ergodique va nous permettre de conclure.
2. La suite $(d_n)_{n\in\mathbb{N}}$ est non-bornée. Dans ce cas la combinaison d'un argument galoisien et d'un argument diophantien vont nous permettre de conclure par un procédé itératif.

Remarque 3.1. Notons que le fait que la suite $(d_n)_{n\in\mathbb{N}}$ est ou n'est pas bornée ne dépend pas du choix du point ξ_n pris dans A'_n (ceci précisément grâce au résultat de Bertrand). En effet soient ξ_n et ξ'_n deux points de torsion de A'_n, d'ordre respectifs d_n et d'_n, tels que $\Sigma_n = A_n + \xi_n = A_n + \xi'_n$. On a

$$\xi_n - \xi'_n \in A_n \cap A'_n.$$

En notant C une borne sur le cardinal de $A_n \cap A'_n$ quand n varie, on constate donc que

$$\frac{1}{C} d_n \leq d'_n \leq C d_n.$$

Autrement dit avec ce choix de variétés abéliennes (A'_n), les suites (d_n) et (d'_n) obtenues sont (ou non) bornées en même temps.

3.1. Cas borné. L'ensemble des points de torsion de A d'ordre borné (par une constante donnée M) étant fini, on peut, en passant à une sous-suite, supposer qu'il existe $\xi \in A_{\mathrm{tors}}$ tel que

$$\forall n \in \mathbb{N}, \quad \Sigma_n = \xi + A_n.$$

Soient C une variété abélienne complexe et μ_C la mesure de Haar sur $C(\mathbb{C})$.

Proposition 3.1. *Quitte à extraire une sous-suite, la suite $(\mu_{A_n})_{n\in\mathbb{N}}$ converge vaguement vers la mesure μ_B où B est une sous-variété abélienne de A, contenant A_n pour tout $n \in \mathbb{N}$.*

Démonstration : La preuve de cette proposition est donnée à la section 4. Le lecteur peut aussi consulter [**Ull05**] proposition 4.1. □

De cette proposition on déduit que, quitte à extraire une sous-suite, on a

$$\forall n \in \mathbb{N}, \qquad \Sigma_n \subset \xi + B \subset V_{\overline{\mathbb{Q}}}.$$

Par généricité, $\xi + B$ ne peut être strictement incluse dans $V_{\overline{\mathbb{Q}}}$, donc V est de torsion, ce qui conclut. □

3.2. Cas non-borné. La preuve de ce cas utilise essentiellement les outils développés par Hindry [**Hin88**]. Ceci étant la stratégie est légèrement différente. Étant donné un corps de nombres K nous noterons G_K le groupe de Galois de $\overline{K}$ sur K.

Théorème 3.1. **(Serre)** *Soit A/K une variété abélienne définie sur un corps de nombres K.*

1. *Il existe une constante $c(A,K) > 0$ telle que pour tout point $x \in A_{\mathrm{tors}}$ d'ordre n et pour tout entier m premier à n, il existe $\sigma \in G_K$ tel que*
$$\left[m^{c(A,K)}\right] x = \sigma(x).$$
2. *Pour tout $\varepsilon > 0$ il existe une constante $C_1(A,K,\varepsilon) > 0$ telle que pour tout $x \in A_{\mathrm{tors}}$ d'ordre n on a*
$$|G_K \cdot x| \geq C_1(A,K,\varepsilon) n^{1-\varepsilon}.$$

Démonstration : Le point 1. est un théorème difficile de Serre (cf. [**Ser00**] Théorème 2' p. 34) dont on peut trouver une preuve dans [**Win02**] (Théorème 3 paragraphe 2.3) et dans [**Ser**]. Le second point est un corollaire du premier. En effet, soit x un point de torsion de A d'ordre n. Par le point 1., il existe une constante $c > 0$ telle que pour tout entier m ne divisant pas n, on a $[m^c]x \in G_K \cdot x$. En particulier, en notant $\varphi(n)$ l'indicatrice d'Euler de n, on a

$$|G_K \cdot x| \geq \left|\left\{x^c \mid x \in (\mathbb{Z}/n\mathbb{Z})^\times\right\}\right| = \frac{\varphi(n)}{|\{x \in (\mathbb{Z}/n\mathbb{Z})^\times \mid x^c = 1\}|}.$$

On sait que $\varphi(n) \gg n^{1-\varepsilon}$, il suffit donc de savoir minorer le cardinal de l'ensemble de $x \in (\mathbb{Z}/n\mathbb{Z})^\times$ tels que $x^c = 1$. Écrivons la décomposition de n en facteurs premiers, $n = \prod_{i=1}^r p_i^{k_i}$, et posons pour tout $i \leq r$, $n_i = p_i^{k_i-1}(p_i - 1)$. En écrivant la décomposition de $(\mathbb{Z}/n\mathbb{Z})^\times$ en produit de groupes cycliques selon les p_i, on voit que

$$\left|\left\{x \in (\mathbb{Z}/n\mathbb{Z})^\times \mid x^c = 1\right\}\right| \leq 2\prod_{i=1}^r \operatorname{pgcd}(c, n_i) \leq 2c^r$$

le facteur 2 étant là pour ne pas avoir d'ennui si 2 intervient dans la décomposition en facteurs premiers de n. Par ailleurs le nombre $r = \omega(n)$ de premiers divisant n est tel que $\omega(n) \ll \log n / \log\log n$. Ainsi $c^r \ll n^\varepsilon$ ce qui permet de conclure. □

LEMME 3.1. **(Hindry)** *Soient n un entier et $X/\overline{\mathbb{Q}}$ une sous-variété irréductible de A. On a*

$$\deg [n]X = \frac{n^{2\dim X}}{|\mathrm{Stab}(X) \cap \ker[n]|} \deg X.$$

Démonstration : cf. [**Hin88**] lemme 6.(ii) ou [**DH00**] proposition 2.3. □

LEMME 3.2. *Soient $X/\overline{\mathbb{Q}}$ une sous-variété irréductible de A et $d \geq 2$ un entier. Si $[d]X \subset X$ alors X est de torsion.*

Démonstration : C'est une conséquence du calcul du degré précédent. En effet soient $s \in \mathbb{N}$ et $G_X = \mathrm{Stab}(X)$. La variété G_X^0 est une variété abélienne d'indice fini $|G_X : G_X^0|$ dans G_X et on a

$$|\mathrm{Stab}(X) \cap \ker[d^s]| \leq \left|G_X : G_X^0\right| \left|G_X^0 \cap \ker[d^s]\right| = d^{2s\dim G_X^0} \left|G_X : G_X^0\right|.$$

Or l'hypothèse nous assure donc que

$$\deg X = \deg [d^s]X \geq \frac{d^{2s\dim X}}{|G_X : G_X^0|\, d^{2s\dim G_X^0}} \deg X.$$

Prenant s assez grand on en déduit que $\dim X = \dim G_X^0 = \dim G_X$. Or $G_X = \bigcap_{x\in X} X - x$, donc X est de la forme $G_X^0 + x$ où x est un point de A. L'hypothèse $[d]X \subset X$ entraine que $[d]X = X$ et donc que $[d-1]x \in G_X^0$. Notamment en notant B la variété abélienne G_X^0, il existe un point ξ de $(d-1)$-torsion dans A tel que $X = \xi + B$. Donc X est de torsion. □

Soit $n \in \mathbb{N}$. On choisit p premier à d_n et on pose $d = p^{c(A,K)}$ comme dans le théorème 3.1 précédent. Considérons l'intersection $V^1 := V \cap [d]V$. Si $\dim V^1 = \dim V$ on a $V^1 = [d]V$ par irréductibilité de V et donc $[d]V \subset V$. Mais alors le lemme 3.2 prouve que V est de torsion. Sinon V^1 est de dimension strictement inférieure à V. Par ailleurs, on a :

LEMME 3.3. *La variété Σ_n est incluse dans $V^1_{\overline{\mathbb{Q}}}$.*

Démonstration : On a $[d]\Sigma_n = [d]A_n + [d]\xi_n = A_n + \sigma(\xi_n)$ par le point 1. du théorème 3.1. Par ailleurs, A_n et V sont définie sur K donc

$$\Sigma_n = \sigma^{-1}\left([d]\Sigma_n\right) \subset \sigma^{-1}\left([d]V\right) = [d]V.$$

Ainsi Σ_n est bien incluse dans $[d]V$ donc dans $V^1_{\overline{\mathbb{Q}}}$. □

On note désormais V_1 une composante K-irréductible de V^1 telle que $(V_1)_{\overline{\mathbb{Q}}}$ contient Σ_n.

LEMME 3.4. **(Bézout)** *Soient X et Y deux sous-variétés d'un espace projectif $\mathbb{P}_n$. En notant $Z_1, \dots Z_r$ des composantes irréductibles de $X \cap Y$, on a*

$$\sum_{i=1}^{r} \deg Z_i \leq \deg X \cdot \deg Y.$$

Démonstration : Il s'agit d'un résultat type Bézout que l'on trouve par exemple dans Fulton [**Ful98**] Example 8.4.6. □

On va donc calculer un majorant du degré de V_1. Pour cela il sufit de savoir estimer (grossièrement) le degré de $[d]V$. Ceci est une conséquence immédiate du lemme 3.1 précédent. On a :

$$\deg [d]X \leq d^{2g} \deg X. \tag{1}$$

Utilisant le lemme 3.4 et l'inégalité (1) on obtient la majoration suivante pour V_1 :

$$\deg V_1 \ll d^{2g}.$$

Partant de V_1 en lieu et place de V et itérant ceci au plus $m := \dim V - \dim \Sigma_n$ fois on aboutit à l'alternative suivante :

1. *ou bien* on a construit une variété de torsion contenant strictement Σ_n,
2. *ou bien* on a fabriqué une variété V_m telle que Σ_n est une composante irréductible de V_m sur $\overline{\mathbb{Q}}$ et vérifiant de plus

$$\deg V_m \ll d^c \tag{2}$$

pour une certaine constante c ne dépendant que de V.

Si on est dans le cas 2. de l'alternative, alors la variété V_m étant définie sur K elle contient comme composante la variété $\bigcup_{\sigma \in \mathrm{Gal}(\overline{K}/K)} \sigma(\Sigma_n)$. Or on a le lemme suivant :

LEMME 3.5. *Avec les notations précédentes on a*

$$\mathcal{O}(\Sigma_n) := \mathrm{card}\left\{\sigma(\Sigma_n) \ | \sigma \in \mathrm{Gal}(\overline{K}/K)\right\} \gg \frac{d_n^{\frac{1}{2}}}{\deg A_n}.$$

Démonstration : Par le point 2. du théorème 3.1 appliqué avec $\varepsilon = \frac{1}{2}$ on a

$$\mathrm{Card}\left(\mathrm{Gal}(\overline{K}/K) \cdot \xi_n\right) \gg d_n^{\frac{1}{2}}.$$

Par ailleurs, le point ξ_n étant choisi dans A'_n, et les variétés A_n et A'_n étant des variétés abéliennes sur K, on a

$$\sigma(\Sigma_n) = \Sigma_n \iff \xi_n - \sigma(\xi_n) \in A_n \cap A'_n.$$

La proposition 2.1 bornant le cardinal de $A_n \cap A'_n$ permet donc de conclure. □

En combinant l'équation 2 et le lemme 3.5, on obtient finalement

$$d_n^{\frac{1}{2}} \ll \mathcal{O}(\Sigma_n) \deg(A_n) = \deg\left(\bigcup_{\sigma \in \mathrm{Gal}(\overline{K}/K)} \sigma(\Sigma_n)\right) \leq \deg V_m \ll d^c.$$

Par le théorème des nombres premiers, on peut prendre d de l'ordre de $(\log d_n)^{c(A,K)}$. Avec un tel choix de d, on voit que pour n assez grand ceci est impossible. C'est donc que l'on est dans le cas 1. de l'alternative indiquée précédemment.

3.3. Conclusion. Partant d'une suite Σ_n, soit l'on est dans le cas borné auquel cas, la preuve s'arrête immédiatement, soit l'on est dans le cas non-borné. Dans ce cas, on a vu que l'on peut alors construire une nouvelle suite générique $(\Sigma'_n)_{n\in\mathbb{N}}$ de sous-variétés de torsion incluses dans V, avec pour tout $n \gg 0$

$$\dim \Sigma'_n \geq \dim \Sigma_n + 1.$$

On réeffectue toute la preuve avec cette nouvelle suite, et au bout d'au plus $\dim V$ étapes on aboutit à la conclusion : V est de torsion. □

4. Preuve de la proposition 3.1

Le but de cette section est de donner la preuve de la proposition 3.1. Cette preuve élémentaire repose essentiellement sur la théorie des séries de Fourier.

4.1. Le cas plat. Dans cette partie on note $G = \mathbb{Q}^n$, $\Lambda = \mathbb{Z}^n$ et $X = \mathbb{Z}^n\backslash\mathbb{R}^n$. Soit $\pi : \mathbb{R} \to X$ la surjection canonique. On dit qu'une sous-variété S de X est spéciale si elle est de la forme

$$S = \pi(H \otimes_{\mathbb{Q}} \mathbb{R})$$

pour un sous-$\mathbb{Q}$-vectoriel H de G. On dispose alors sur X d'une mesure de probabilité μ_S, $(H \otimes \mathbb{R})$-invariante, de support S. On notera alors $\mu = \mu_X$ la mesure de Lebesgue normalisée sur X

On notera par abus de langage de la même manière une fonction sur X et la fonction $\mathbb{Z}^n$-invariante sur $\mathbb{R}^n$ correspondante.

On dit qu'une suite de sous-variétés Y_n de X est stricte si pour toute sous-variété spéciale S_1,

$$\{n \in \mathbb{N},\ Y_n \subset S_1\}$$

est un ensemble fini.

Proposition 4.1. *Soit T_n une suite stricte de sous-variétés spéciales de X. Soit μ_n la mesure invariante normalisée de support T_n. Pour toute fonction continue sur X, on a*

$$\int_{T_n} f\ d\mu_n \longrightarrow \int_X f\ d\mu. \tag{3}$$

Pour $x \in \mathbb{R}$, on note $\overline{x}$ sa classe dans $\mathbb{Z}\backslash\mathbb{R}$. Pour $(k_1, \ldots, k_n)$, on note $\chi_{k_1,\ldots,k_n}$ le caractère de X défini par

$$\chi_{k_1,\ldots,k_n}(\overline{x}_1, \ldots, \overline{x}_n) = \exp(2i\pi \sum_{j=1}^{n} k_j x_j).$$

On obtient ainsi tous les caractères de X. Si $\chi = \chi_{k_1,\ldots,k_n}$ pour $(k_1, \ldots, k_n)$ non tous nuls, on note

$$H_\chi = H_{k_1,\ldots,k_n}$$

le $\mathbb{Q}$-hyperplan de G d'équation

$$\sum_{j=1}^{n} k_j x_j = 0.$$

Remarquons que $H_\chi = H_{\chi'}$ avec $\chi = \chi_{k_1,\ldots,k_n}$ et $\chi' = \chi_{k'_1,\ldots,k'_n}$ si et seulement si il existe $\alpha \in \mathbb{Q}$ tel que $k'_i = \alpha k_i$ pour tout i.

On note alors

$$S_\chi = S_{k_1,\dots,k_n} = \pi(H_{k_1,\dots,k_n} \otimes \mathbb{R})$$

la sous-variété spéciale maximale associée. De même on note

$$\widetilde{S_\chi} = H_{k_1,\dots,k_n} \otimes \mathbb{R}.$$

On obtient ainsi toute les sous-variétés spéciales maximales de X. Par ailleurs la théorie élémentaire des séries de Fourier nous donne

LEMME 4.1. *Une suite de mesure ν_n sur X converge faiblement vers une mesure ν si et seulement si pour tout caractère de X, on a*

$$\nu_n(\chi) \longrightarrow \nu(\chi). \tag{4}$$

LEMME 4.2. *Soit S une variété spéciale et χ un caractère non trivial de X. La restriction χ_S de χ à S est un caractère de S et $\chi_S = 1$ si et seulement si $S \subset S_\chi$.*

Preuve. La restriction de χ à S_χ est triviale donc si $S \subset S_\chi$ alors la restriction de χ à S est triviale.

Réciproquement, on peut écrire $S = \pi(\widetilde{S})$ pour un sous-$\mathbb{R}$-espace vectoriel $\widetilde{S}$ de $\mathbb{R}^n$. Soit $x = (x_1, \dots, x_n) \in \widetilde{S}$ et χ un caractère trivial sur S. Pour tout $t \in \mathbb{R}$ $\chi(\pi(tx)) = 1$ car $tx \in \widetilde{S}$. Pour tout $t \in \mathbb{R}$ on a $t\sum_{I=1}^n k_i x_i \in \mathbb{Z}$. On en déduit que $\sum_{I=1}^n k_i x_i = 0$ donc que $(x_1, \dots, x_n) \in \widetilde{S_\chi}$. D'ou $\pi(x) \in S_\chi$ et donc $S \subset S_\chi$.

Preuve de la proposition 4.1. Soit S une variété spéciale et μ_S sa mesure invariante normalisée. Pour tout caractère χ de X, $\int_S \chi \, d\mu_S = 1$ si la restriction de χ à S est le caractère trivial 1 et vaut 0 sinon. D'après le lemme 4.2 $\int_S \chi \, d\mu_S = 1$ si $S \subset S_\chi$ et 0 sinon.

Soient donc T_n une suite stricte de sous-variétés spéciales de X. et μ_n la mesure invariante normalisée de T_n. Soit χ un caractère non trivial de X. Comme T_n est une suite stricte, pour tout n assez grand (dépendant de χ), T_n n'est pas contenu dans S_χ. D'après ce qui précède, on voit que pour tout n assez grand on a

$$\int_{T_n} \chi \, d\mu_n = 0.$$

on a donc

$$\lim_{n\to\infty} \int_{T_n} \chi \, d\mu_n = 0.$$

On termine la preuve de la proposition 4.1 en utilisant le lemme 4.1.

Une conséquence de la proposition 4.1 est l'énoncé:

COROLLAIRE 4.1. *Soit S_n une suite de sous-variétés spéciales de $\mathbb{Z}^n\backslash\mathbb{R}^n$. En passant au besoin à une sous-suite, il existe une sous-variété spéciale S contenant les S_n telle que la suite de mesures canoniques μ_n de S_n converge faiblement vers la mesure canonique de S.*

Soit $\mathcal{E}$ l'ensemble des sous-variétés spéciales contenant une infinité de termes de la suite. Soit S un élément minimal de $\mathcal{E}$. En passant à une sous-suite on peut supposer que les S_i sont contenus dans S pour tout i. La variété S est de la forme $\mathbb{Z}^k\backslash\mathbb{R}^k$ et par définition les S_n forment une suite stricte de sous-variétés spéciales de S. Le résultat se déduit alors de la proposition 4.1.

4.2. Application aux variétés abéliennes. Soit $A = \Gamma\backslash\mathbb{C}^n$ une variété abélienne. En identifiant $\mathbb{C}^n$ à $\mathbb{R}^{2n}$, on peut appliquer la théorie de la partie précédente. Les sous-variétés abéliennes sont des sous-variétés spéciales. On obtient alors la proposition 3.1.

PROPOSITION 4.2. *Soit B_n une suite de sous-variétés abéliennes de A. En passant au besoin à une sous-suite, il existe une sous-variété abélienne B de A, telle que telle que pour tout n $B_n \subset B$ et telle que la suite de mesures μ_n canoniquement associées converge vers la mesure μ_B canoniquement associée à B.*

Preuve. On sait déjà qu'il existe une sous-variété spéciale S de A ayant la propriété du théorème, il faut voir que S est une variété abélienne. Les B_i sont de la forme $\Gamma \cap V_i\backslash V_i$ pour des sous-$\mathbb{C}$ espaces vectoriels V_i de $\mathbb{C}^n$. Par ailleurs $S = \Gamma \cap V\backslash V$ pour un $\mathbb{R}$ sous-espace vectoriel de $\mathbb{C}^n = \mathbb{R}^{2n}$. Comme les B_n forment une suite stricte de S, S est engendré comme groupe par un nombre fini des B_i et V est somme d'un nombre fini des V_i. On en déduit que V est muni d'une structure de $\mathbb{C}$-espace vectoriel donc que B est un tore complexe puis que B est une sous-variété abélienne de A car d'après ([**LB92**] p 73) un sous-tore complexe d'une variété abélienne est une variété abélienne.

5. Appendice : effectivité dans le lemme de Poincaré, selon Bertrand [**Ber**]

Nous donnons ici la preuve, reprise de l'appendice de [**Ber**], de la proposition 2.1. Pour l'énoncé lui-même, voir aussi [**LB92**], Chap. 5, Exercice 5.

5.1. Rappels. On fixe une $\mathbb{Q}$-algèbre $\mathbb{D}$ de dimension finie que l'on suppose être une algèbre à division, c'est-à-dire munie d'un élément unité 1 et telle que tous ses éléments non nuls sont inversibles pour la multiplication. Dans notre situation, avec une variété abélienne simple A, nous utiliserons ceci pour $\mathbb{D} := \mathrm{End}(A) \otimes \mathbb{Q}$. Le centre de $\mathbb{D}$ ne jouera pas de rôle.

Définition 5.1. Un anneau $\mathcal{O}$ est un *ordre dans* $\mathbb{D}$ si

1. il est de type fini sur $\mathbb{Z}$,
2. il est contenu dans $\mathbb{D}$,
3. il contient 1,
4. $\mathcal{O} \cdot \mathbb{Q} = \mathbb{D}$ (où $\mathcal{O} \cdot \mathbb{Q}$ désigne l'image de l'application naturelle de $\mathcal{O} \otimes_{\mathbb{Z}} \mathbb{Q}$ dans $\mathbb{D}$).

Définition 5.2. Soit $\mathcal{O}$ un ordre dans $\mathbb{D}$. On dit que Λ est un *$\mathcal{O}$-réseau* si c'est un $\mathcal{O}$-module à droite de type fini qui est sans $\mathbb{Z}$-torsion.

Donnons maintenant un cas particulier du théorème de Jordan-Zassenhaus (cf. par exemple [**CR81**] p.534).

THÉORÈME 5.1. (Jordan-Zassenhaus) *Soient $\mathbb{D}$ une algèbre à division de dimension finie, $\mathcal{O}$ un ordre dans $\mathbb{D}$, et V un $\mathbb{D}$-module à droite de rang fini. Il n'existe qu'un nombre fini de classes d'isomorphismes de $\mathcal{O}$-réseaux L contenus dans V tels que $V = L \cdot \mathbb{Q}$.*

5.2. Le résultat de Bertrand.

LEMME 5.1. *Soient $\mathbb{D}$ une $\mathbb{Q}$-algèbre à division de dimension finie, $n \geq 1$ un entier et $\mathcal{O}$ un ordre maximal dans $\mathbb{D}$. Il existe un entier $c_1 = c_1(\mathcal{O}, n) > 0$ telle que pour tout sous-$\mathcal{O}$-module à droite L de $\mathcal{O}^n$, on a :*

1. *Il existe un sous-$\mathcal{O}$-module libre F de L tel que $|L/F|$ divise c_1.*
2. *Si $\mathcal{O}^n/L$ est sans $\mathbb{Z}$-torsion, alors il existe un sous-$\mathcal{O}$-module libre M de $\mathcal{O}^n$ tel que $L \cap M = \{0\}$ et tel que $L + M$ est d'indice divisant c_1 dans $\mathcal{O}^n$.*

Démonstration : Le module L est de rang fini, m, comme $\mathcal{O}$-module. De plus, c'est un sous-module de $\mathcal{O}^n$ qui est sans $\mathbb{Z}$-torsion. C'est donc un $\mathcal{O}$-réseau. Par le théorème de Jordan-Zassenhaus 5.1, il n'y a qu'un nombre fini de classes d'isomorphismes de sous-$\mathcal{O}$-réseaux M de $\mathbb{D}^m$ tels que $M \cdot \mathbb{Q} = \mathbb{D}^m \simeq L \cdot \mathbb{Q}$. Ainsi, L est isomorphe à un élément M_0 appartenant à cette famille, finie de cardinal ne dépendant que de $\mathcal{O}$ et de $m \leq n$, de $\mathcal{O}$-réseaux. Pour chaque élément de cette famille, on choisit un sous-$\mathcal{O}$-module libre d'indice fini et on prend l'image dans L du sous-module correspondant à M_0. On obtient ainsi un sous-$\mathcal{O}$-module libre de L, d'indice fini borné indépendamment de L. Ceci prouve le 1.

Pour le point 2. : le module $\mathcal{O}^n/L$ est sans $\mathbb{Z}$-torsion donc c'est un $\mathcal{O}$-réseau. De plus par maximalité de l'ordre $\mathcal{O}$, le $\mathcal{O}$-réseau $\mathcal{O}^n/L$ est un $\mathcal{O}$-module projectif (cf. [**CR81**] Theorem 26.12 (ii) p.565). Ainsi L admet un $\mathcal{O}$-réseau supplémentaire M_1 dans $\mathcal{O}^n$. Par le point 1. , et quitte à remplacer c_1 par son carré, on en déduit un sous-$\mathcal{O}$-module M de M_1 comme annoncé. □

PROPOSITION 5.1. (= Proposition 2.1) *Il existe un entier $c_2 = c_2(A) > 0$ ne dépendant que de A tel que pour toute sous-variété abélienne B de A, il existe une sous-variété B' de A telle que*

$$A = B + B' \text{ et } \operatorname{card}(B \cap B') \leq c_2.$$

Démonstration : Notons tout d'abord que si $\varphi : A \to \widetilde{A}$ est une isogénie de degré N, l'énoncé sur $\widetilde{A}$ l'entraîne sur A, avec $c_2(A) = c_2(\widetilde{A})N$. D'après le théorème de réductibilité de Poincaré, on peut donc supposer sans perte de généralité que A est un produit de puissances de variétés abéliennes deux à deux non isogènes.

En deuxième lieu, notons que si A_1 et A_2 sont deux variétés abéliennes telles que $\operatorname{Hom}(A_1, A_2) = 0$, alors, toute sous-variété abélienne B de $A = A_1 \times A_2$ est de la forme $B_1 \times B_2$, où B_1 et B_2 sont les projections de B sur A_1 et A_2. Par conséquent, $c_2(A_1 \times A_2) := c_2(A_1) \times c_2(A_2)$ convient, et on peut sans perte de généralité supposer que notre variété abélienne A est une puissance A_0^n d'une variété abélienne simple. Alors, $\operatorname{End}(A_0)$ est un ordre d'une algèbre à division $\mathbb{D}$, et il existe une variété abélienne $\widetilde{A}_0$, isogène à A_0, telle $\operatorname{End}(\widetilde{A}_0)$ est un ordre maximal $\mathcal{O}$ de $\mathbb{D}$.

On peut finalement supposer sans perte de généralité que la variété abélienne ambiante est de la forme A^n avec A simple et telle que $\mathcal{O} := \operatorname{End}(A)$ est un ordre maximal dans l'algèbre à division $\mathbb{D} = \operatorname{End}(A) \otimes \mathbb{Q}$.

Soit alors B une sous-variété abélienne de A^n. Elle est isogène à A^m avec $m \leq n$. On a $\mathcal{O}^n \simeq \operatorname{Hom}(A, A^n)$. Notons

$$L_B = \{f \in \mathcal{O}^n \mid f(A) \subset B\} \simeq \operatorname{Hom}(A, B).$$

C'est un sous-$\mathcal{O}$-module à droite de $\mathcal{O}^n$. On peut donc appliquer le point 1. du lemme 5.1 précédent qui fournit un sous-$\mathcal{O}$-module libre L de L_B d'indice divisant

l'entier $c_1(A)$. Par ailleurs, A étant un groupe divisible on voit également sur la définition de L_B que $\mathcal{O}^n/L_B$ est sans $\mathbb{Z}$-torsion.

Le point 2. du même lemme 5.1 fournit alors un sous-$\mathcal{O}$-module libre de type fini M de $\mathcal{O}^n$ tel que $M \cap L_B = \{0\}$ et tel que $L_B + M$ est d'indice divisant $c_1(A)$ dans $\mathcal{O}^n$. Ainsi L et M sont deux sous-$\mathcal{O}$-modules libres de $\mathcal{O}^n$ tels que $L \cap M = 0$ et $L + M$ est d'indice $\leq c_1(A)^2$ dans $\mathcal{O}^n$.

Notons $\{\lambda_1, \ldots, \lambda_m\}$ une base de L sur $\mathcal{O}$ et $\{\mu_1, \ldots, \mu_r\}$ une base de M sur $\mathcal{O}$ (avec $m + r = n$). Considérons de plus l'homomorphisme de A^m dans A^n

$$\underline{\lambda} : A^m \to A^n \text{ défini par } \underline{\lambda}(x_1, \ldots, x_m) = \sum_{i=1}^{m} \lambda_i(x_i),$$

et notons de même $\underline{\mu} : A^r \to A^n$. On pose $B' := \underline{\mu}(A^r)$. Quant à l'image de $\underline{\lambda}$, c'est B elle-même, puisque L étant contenu dans L_B, elle est par définition contenue dans B, tandis que les λ_i étant linéairement indépendants sur $\mathcal{O}$, sa dimension est égale à $m \dim(A) = \dim(B)$.

Considérons maintenant l'endomorphisme de A^n

$$(\underline{\lambda}, \underline{\mu}) : A^n \to A^n, \quad (x_1, \ldots, x_n) \mapsto \sum_{i=1} \lambda_i(x_i) + \sum_{i=1}^{r} \mu_i(x_{m+i}),$$

dont l'image est par définition $B + B'$. Comme les λ_i, μ_j sont linéairement indépendants sur $\mathcal{O}$, son image a pour dimension $(m + r) \dim(A) = \dim(A^n)$. Il est donc surjectif et $B + B' = A^n$. D'autre part, $L + M$ est d'indice $\nu \leq c_1(A)^2$ dans $\mathcal{O}^n$, donc il existe une matrice carrée γ d'ordre n à coefficients dans $\mathcal{O}$ telle que $(\underline{\lambda}, \underline{\mu}) \circ \gamma = \nu I_n$. D'après [**Dra83**] p.131, on a alors aussi $\gamma \circ (\underline{\lambda}, \underline{\mu}) = \nu I_n$, de sorte que le noyau de $(\underline{\lambda}, \underline{\mu})$ est composé de points de torsion d'ordre divisant ν. Finalement, $(\underline{\lambda}, \underline{\mu})$ est une isogénie de degré borné par $\nu^{2n^2 dim(A)2} := c_2(A)$. Enfin, $(\underline{\lambda}, \underline{\mu})$ est la composée des isogénies $\underline{\lambda} \times \underline{\mu} : A^m \times A^r \to B \times B'$ et $+ : B \times B' \to A^n$, donc

$$|B \cap B'| = |\ker(+)| \leq c_2(A).$$

Ainsi, la sous-variété abélienne B' de la variété ambiante A^n répond à la question. □

Références

[Ber] D. Bertrand, *Minimal heights and polarizations on abelian varieties*, MSRI, Preprint 06220-87, Berkeley, June 1987.

[CR81] C. W. Curtis et I. Reiner, *Methods of representation theory. Vol. I*, John Wiley & Sons Inc., New York, 1981, With applications to finite groups and orders, Pure and Applied Mathematics, A Wiley-Interscience Publication. MR 632548 (82i:20001)

[CU05] L. Clozel et E. Ullmo, *Équidistribution de sous-variétés spéciales*, Ann. of Math. (2) **161** (2005), no. 3, 1571–1588. MR 2180407 (2006j:11083)

[DH00] S. David et M. Hindry, *Minoration de la hauteur de Néron-Tate sur les variétés abéliennes de type C. M*, J. Reine Angew. Math. **529** (2000), 1–74. MR 1799933 (2001j:11054)

[Dra83] P. K. Draxl, *Skew fields*, London Mathematical Society Lecture Note Series, vol. 81, Cambridge University Press, Cambridge, 1983. MR 696937 (85a:16022)

[EY03] B. Edixhoven et A. Yafaev, *Subvarieties of Shimura varieties*, Ann. of Math. (2) **157** (2003), no. 2, 621–645. MR 1973057 (2004c:11103)

[Ful98] W. Fulton, *Intersection theory*, second ed., Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge, vol. 2, Springer-Verlag, Berlin, 1998. MR 1644323 (99d:14003)

[Hin88] M. Hindry, *Autour d'une conjecture de Serge Lang*, Invent. Math. **94** (1988), no. 3, 575–603. MR 969244 (89k:11046)

[KY06] B. Klingler et A. Yafaev, *On the André-Oort conjecture*, 2006, preprint.

[LB92] H. Lange et C. Birkenhake, *Complex abelian varieties*, Grundlehren der Mathematischen Wissenschaften, vol. 302, Springer-Verlag, Berlin, 1992. MR 1217487 (94j:14001)

[MW93] D. Masser et G. Wüstholz, *Periods and minimal abelian subvarieties*, Ann. of Math. (2) **137** (1993), no. 2, 407–458. MR 1207211 (94g:11040)

[PR02] R. Pink et 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, pp. 539–546. MR 1989204 (2004f:14062)

[PR04] ———, *On ψ-invariant subvarieties of semiabelian varieties and the Manin-Mumford conjecture*, J. Algebraic Geom. **13** (2004), no. 4, 771–798. MR 2073195 (2005d:14061)

[Ray83] M. Raynaud, *Sous-variétés d'une variété abélienne et points de torsion*, Arithmetic and geometry, Vol. I, Progr. Math., vol. 35, Birkhäuser Boston, Boston, MA, 1983, pp. 327–352. MR 717600 (85k:14022)

[Ser] J.-P. Serre, *Groupes linéaires modulo p et points d'ordre fini des variétés abéliennes*, notes de E. Bayer-Fluckiger sur le cours donné par Jean-Pierre Serre au Collège de France de janvier à mars 1986, accessible en ligne à l'adresse http://alg-geo.epfl.ch/~bayer/html/notes.html.

[Ser00] ———, *Œuvres. Collected papers. IV*, Springer-Verlag, Berlin, 2000, 1985–1998. MR 1730973 (2001e:01037)

[Ull98] E. Ullmo, *Positivité et discrétion des points algébriques des courbes*, Ann. of Math. (2) **147** (1998), no. 1, 167–179. MR 1609514 (99e:14031)

[Ull05] ———, *Manin-Mumford, André-Oort, the equidistribution point of view*, 2005, notes de cours à l'école d' été Equidistribution en théorie des nombres, Montréal, disponible à l'adresse http://www.math.u-psud.fr/~ullmo/Prebublications/coursMontrealfinal.pdf.

[UY] E. Ullmo et A. Yafaev, *The André-Oort conjecture for products of modular curves*, dans ce volume.

[UY06] ———, *Galois orbits and equidistribution of special subvarieties of Shimura varieties: towards the André-Oort conjecture*, 2006, preprint.

[Win02] J.-P. Wintenberger, *Démonstration d'une conjecture de Lang dans des cas particuliers*, J. Reine Angew. Math. **553** (2002), 1–16. MR 1944805 (2003i:11075)

[Zha98] S.-W. Zhang, *Equidistribution of small points on abelian varieties*, Ann. of Math. (2) **147** (1998), no. 1, 159–165. MR 1609518 (99e:14032)

Université Paris-Sud, Bâtiment 425, 91405 Orsay Cedex, France
E-mail address: nicolas.ratazzi@math.u-psud.fr

Université Paris-Sud, Bâtiment 425, 91405 Orsay Cedex, France
E-mail address: emmanuel.ullmo@math.u-psud.fr

Clay Mathematics Proceedings
Volume 8, 2009

The André-Oort conjecture for products of modular curves

Emmanuel Ullmo and Andrei Yafaev

Abstract. In this paper we prove (assuming the Generalised Riemann Hypothesis) the André-Oort conjecture for products of modular curves using a combination of Galois-theoretic and ergodic-theoretic methods.

1. Introduction

The André-Oort conjecture stated below has been recently proved by Klingler, Ullmo and Yafaev (see [**UY06**] and [**KY06**]) in full generality assuming the Generalised Riemann Hypothesis.

Conjecture 1.1 (André-Oort). *Let S be a Shimura variety and let Σ be a set of special points in S. Every irreducible component of the Zariski closure of Σ is a special (or Hodge type) subvariety of S.*

For generalities on this conjecture, in particular for the notions of special points and subvarieties we refer, for example, to [**Yaf07**]. The purpose of this note is to present a proof of this conjecture in the special case where S is a product of an arbitrary number of modular curves. It is our hope that this will help in understanding the strategy used in [**UY06**] and [**KY06**], as many of the technical problems occurring in the general case do not present themselves in the case considered in this paper but all of the main ideas of the proof are conserved. The main result of this paper is the following.

Theorem 1.2. *Assume the GRH for imaginary quadratic fields. Let $n \geq 1$ be an integer and let S be a product of n modular curves. Let Σ be a set of special points in S. The irreducible components of the Zariski closure of Σ are special subvarieties.*

Note that this case of the conjecture has already been dealt with by Edixhoven [**Edi05**] but his strategy does not seem to be easily generalisable as it relies on the very particular geometric properties of the Shimura variety under consideration. We also point out that our strategy yields a proof of the Manin-Mumford conjecture as well (the "abelian counterpart" of the André-Oort conjecture). We refer to [**RU**] for details on this.

2000 *Mathematics Subject Classification.* Primary 11G18, Secondary 14G35, 14L05.

The strategy of the proof is based on the following alternative in the geometry of Shimura varieties. Let S be a Shimura variety and let Z_n be a sequence of irreducible special subvarieties of S. Let F be some number field over which S admits a canonical model. After possibly replacing Z_n by a subsequence and assuming the GRH for CM-fields, at least one of the following cases occurs.

(1) The cardinality of the sets $\{\sigma(Z_n), \sigma \in \mathrm{Gal}(\overline{\mathbb{Q}}/F)\}$ is unbounded as $n \to \infty$ (and therefore Galois-theoretic techniques can be used).
(2) The sequence of probability measures μ_n canonically associated to Z_n weakly converges to some μ_Z, the probability measure canonically associated to a special subvariety Z of S. Moreover, for every n large enough, Z_n is contained in Z.

Which of the two cases occurs depends on the geometric nature of the subvarieties Z_n.

Let us explain this in more detail in the case considered in this paper. So let S be a product of n modular curves. We assume that S is $(\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H})^n = \mathbb{C}^n$. Special subvarieties are products of factors which are of one of the following forms:

(1) A special point (equivalently CM point) of some $\mathbb{C}^m$, $m \leq n$.
(2) A modular curve $\Gamma\backslash\mathbb{H}$ (for some congruence subgroup of Γ of $\mathrm{SL}_2(\mathbb{Z})$) embedded in a product of copies of $\mathbb{C}$.

A special subvariety is called *strongly* special if it does not have any CM factors. Sequences of strongly special subvarieties are precisely those for which the second case of the alternative occurs (this is a consequence of a theorem of Clozel-Ullmo that we will recall later). The sequences of special subvarieties that do have special factors are those for which the first case of the alternative occurs.

The strategy of the proof is as follows. For a special subvariety Z, we let $c(\Omega_Z)$ be the number of CM factors, therefore $c(\Omega_Z) = 0$ means precisely that Z is strongly special. Let X be a subvariety of S containing a Zariski dense set Σ of special subvarieties. We can assume (after possibly replacing Σ by a Zariski dense subset) that $c(\Omega_Z)$ is constant as Z ranges through Σ; let's call $c(\Sigma)$ this number. If $c(\Sigma) = 0$, then X is special by the theorem of Clozel and Ullmo, otherwise the size of the Galois orbit of Z is unbounded as Z ranges through Σ. Using the explicit description of the Galois action on special points and a characterisation of special subvarieties in terms of Hecke correspondences, we show that every Z with sufficiently large Galois orbit is contained in a special subvariety Z' with $c(\Omega_{Z'}) < c(\Omega_Z)$. Thus we construct a Zariski-dense set Σ' of special subvarieties with $c(\Sigma') < c(\Sigma)$. We then reiterate the process with Σ' instead of Σ. Eventually we obtain a Zariski dense set of strongly special subvarieties.

Acknowledgements.

The first author is grateful to the organisers of the Summer School in Arithmetic geometry in Göttingen in August 2006 for inviting him and giving him an opportunity to give a series of lectures on the André-Oort conjecture. The second author is grateful to the Université de Paris-Sud for hospitality and financial support. Both authors are grateful to the Scuola Normale Superiore Di Pisa and to the Université de Montreal for inviting them in April and July 2005 respectively. We thank the referee for pointing out a serious mistake in the first version of this paper.

2. Preliminaries.

Before we state and prove our main result, we recall some definitions and prove some preliminary results which will be used in the course of the proof. Let us first recall the following definition from Edixhoven ([**Edi05**], definition 1.1).

DEFINITION 2.1. *Let I be a finite set of cardinality r. For every i in I, let Γ_i be a congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$ and S be the product of the $X_{\Gamma_i} := \Gamma_i\backslash\mathbb{H}$ for $i \in I$. A closed irreducible subvariety Z of S is called special (of type $\Omega = \Omega_Z$) if I has a partition $\Omega = (I_1, \ldots, I_t)$ such that Z is a product of subvarieties Z_i of $S_i = \prod_{j\in I_i} \Gamma_j\backslash\mathbb{H}$, each of one of the forms:*

1. *I_i is a one-element set and Z_i is a CM point.*
2. *Z_i is the image of $\mathbb{H}$ in S_i under the map sending τ in $\mathbb{H}$ to the image of $(g_s\tau)_{s\in I_i}$ in S_i, where the g_s are some elements of $\mathrm{GL}_2(\mathbb{Q})$ with positive determinant.*

Given a special subvariety Z of type Ω, we define $c(\Omega)$ to be the number of CM factors. A special subvariety Z is called strongly special if $c(\Omega) = 0$

We now prove a few lemmas that will be used in the course of the proof.

LEMMA 2.2. *Let Z be a strongly special subvariety of $\mathbb{C}^n$. Then Z is defined (as an absolutely irreducible subscheme) over an abelian extension L of $\mathbb{Q}$ such that $\mathrm{Gal}(L/\mathbb{Q})$ is killed by the multiplication by 2, i.e. for every σ in $\mathrm{Gal}(L/\mathbb{Q})$, $\sigma^2 = 1$.*

Proof. This is a consequence of the explicit description of the Galois action on irreducible components of strongly special subvarieties via a reciprocity law. We refer to section 2 of [**UY06**] for details on this.

The inclusion $Z \hookrightarrow \mathbb{C}^n$ corresponds to the inclusion of Shimura data

$$(\mathrm{PGL}_2^m, \mathbb{H}^{\pm m}) \hookrightarrow (\mathrm{PGL}_2^n, \mathbb{H}^{\pm n})$$

for some $m \leq n$. This is a consequence of the explicit description of special subvarieties of $\mathbb{C}^n$ given above. Let $\rho\colon \mathrm{SL}_2^m \longrightarrow \mathrm{PGL}_2^m$ be the simply connected covering. Its kernel is killed by 2. Then the reciprocity morphism defining the Galois action on connected components is a morphism

$$r\colon \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \longrightarrow \mathrm{PGL}_2^m(\mathbb{A}_f)/\mathrm{PGL}_2^m(\mathbb{Q})\rho(\mathrm{SL}_2^m(\mathbb{A}_f))$$

It is now clear that its image (which is isomorphic to $\mathrm{Gal}(L/\mathbb{Q})$) is killed by 2. □

Using the above lemma we now prove the following.

LEMMA 2.3. *Let $Z = \{x_1, \ldots, x_s\} \times Z'$ be a special subvariety of $\mathbb{C}^n$ (Z' is a strongly special subvariety of $\mathbb{C}^{n-s}$ and $(x_1, \ldots, x_s)$ is a special point of $\mathbb{C}^s$). Let O_{x_i} be the ring of complex multiplication of the point x_i. Let l be a prime splitting every O_{x_i}. Let $T_{\mathbf{l^2}}$ be the Hecke correspondence defined by the element of the product of r copies of $\mathrm{GL}_2(\mathbb{Q})^+$ which is $\left(\begin{smallmatrix} l^2 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ on the first s components and 1 elsewhere.*

There exists an element σ of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ such that

$$\sigma(Z) \subset T_{\mathbf{l^2}} Z$$

Proof. Let K be the composite of the fields K_{x_i} of complex multiplication of the points x_i and let R be the ring $O_{x_1} \otimes \cdots \otimes O_{x_s}$. The ring R is an order in K and the prime l splits in R. Let τ be the Frobenius element in $\mathrm{Gal}(\overline{\mathbb{Q}}/K)$ for a prime ideal lying over l. The theory of complex multiplication of elliptic curves shows that $\tau^2(x_i) \subset T_{l^2}(x_i)$ where T_{l^2} is the usual Hecke correspondence given by the element

$\left(\begin{smallmatrix} l^2 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ of $\mathrm{GL}_2(\mathbb{Q})^+$. By the lemma above $\tau^2(Z') = Z'$. It follows that we can take $\sigma = \tau^2$. $\square$

As in [**Edi05**], we will make use of lower bounds for Galois orbits of CM points. Let Γ be a congruence subgroup of $\mathrm{SL}_2(\mathbb{Z})$ and x a special point of $\Gamma\backslash\mathbb{H}$. Let O_x be the ring of complex multiplication of x (an order in an imaginary quadratic field) and d_x be the absolute value of the discriminant of O_x. Let Z be a special subvariety of S of type Ω with $s = c(\Omega) > 0$. Let, as in the above lemma, $\{x_1, \ldots, x_s\}$ be the set of CM points occurring as a CM factor of Z and let $d_Z = \max_{1 \leq i \leq s} d_{x_i}$. In the following statement and in the rest of this paper, the symbol $\gg$ stands for "up to a uniform constant". We hope that this notation will cause no confusion.

PROPOSITION 2.4. *Let $0 < \epsilon < 1/2$ be a real number. Let Z be a special subvariety of type Ω with $c(\Omega) > 0$. The following inequality holds:*

$$|\{\sigma(Z), \sigma \in \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})\}| \gg d_Z^{\frac{1}{2}-\epsilon}$$

Proof. The cardinality of this set is at least that of the Galois orbit of the special point $(x_1, \ldots, x_s)$ of a product of s modular curves. The lower bound for this Galois orbit is a consequence of the Brauer-Siegel theorem. We refer to section 5 of [**Edi05**] for details. $\square$

To finish this section, we state the following consequence of the main theorem of [**CU05**] that will be used in a crucial way in the course of our proof.

THEOREM 2.5 (Clozel-Ullmo). *Let S be a product of modular curves and let Z_n be a sequence of strongly special subvarieties. The sequence Z_n is equidistributed in the following sense. There exists a subsequence Z_{n_k} and a strongly special subvariety Z such that the sequence of probability measures μ_{n_k} canonically associated to Z_{n_k} weakly converges to μ_Z, the probability measure canonically associated to Z. Furthermore, Z contains Z_{n_k} for all k large enough.*

3. A characterization.

In this section we consider a subvariety X containing a special subvariety Z with $c(\Omega_Z) > 0$. We prove that if Z is contained in its image by some suitable Hecke correspondence, then X contains a special subvariety Z' containing Z properly. This is a key ingredient of our proof.

We will make use of the notion of degree of a subvariety V of $\mathbb{P}^{1^r}$ that we now recall. The Chow ring of $\mathbb{P}^{1^r}$ is $\mathbb{Z}[\epsilon_1, \ldots, \epsilon_r]$ with $\epsilon_i^2 = 0$. Suppose that V is irreducible of codimension i. The class $[V]$ of V in the Chow group $CH^i(\mathbb{P}^{1^r})$ is

$$[V] = \sum_{|I|=i} a_I \epsilon_I$$

where ϵ_I is the product of the ϵ_i for i in I and a_I is the degree of the projection of V onto the product of copies of $\mathbb{P}^{1^r}$ indexed by the complement $I^\vee$ of I in $\{1, \ldots, r\}$. We define the degree of V to be $\deg(V) = \sum_I a_I$. The variety S admits a quasi-finite morphism to $\mathbb{C}^r$. Let V' be a subvariety of S and let V be the closure of the image of V' in $\mathbb{P}^{1^r}$. The morphism from V' to V is quasi-finite. We define the degree of V' to be $\deg(V') = \deg(V)[\mathbb{C}(V') : \mathbb{C}(V)]$, where $\mathbb{C}(V')$ and $\mathbb{C}(V)$ denote the function fields of V' and V, respectively. The degree $[\mathbb{C}(V') : \mathbb{C}(V)]$ is at most the index of the product of the Γ_i in the product of the $\mathrm{SL}_2(\mathbb{Z})$.

PROPOSITION 3.1. *Let $r \geq 3$ be an integer. Let X be an irreducible hypersurface of $\mathbb{C}^r$ such that for every $I \subset \{1, \ldots, r\}$ of cardinality $r-1$, the projections $p_I \colon X \longrightarrow \mathbb{C}^{|I|}$ are dominant (in particular generically finite). Let l be a prime. Let $s \leq r$ be an integer and let $T_{\mathbf{l^2}}$ be the Hecke correspondence defined by the element of the product of r copies of $\mathrm{GL}_2(\mathbb{Q})^+$ which is $\left(\begin{smallmatrix} l^2 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ on the first s components and 1 elsewhere.*

Suppose that l is larger than $\deg(X)$ and 13. Then the variety $T_{\mathbf{l^2}}X$ is irreducible.

Proof. Let $Y(l^2)$ be $\Gamma(l^2)\backslash\mathbb{H}$ where $\Gamma(l^2) := \ker(\mathrm{SL}_2(\mathbb{Z}) \longrightarrow \mathrm{SL}_2(\mathbb{Z}/l^2\mathbb{Z}))$. Let π be the quotient map $Y(l^2)^r \longrightarrow \mathbb{C}^r$. Let l be a prime as in the statement above. The proof of Proposition 4.2 of [**Edi05**] shows that $\pi^{-1}X$ is irreducible. As $T_{\mathbf{l^2}}X$ is the image of $\pi^{-1}X$ by some (other) morphism, $T_{\mathbf{l^2}}X$ is irreducible as well. □

We recall (see [**Edi05**]) that if X is an irreducible subvariety of $\mathbb{C}^r$ then a minimal projection p_I for X is given by a subset $I \subset \{1, \ldots, r\}$ such that $p_I X$ is a hypersurface of $\mathbb{C}^{|I|}$ such that for all subsets I' of I with $|I'| = |I| - 1$, the projection $p_{I'}X$ is dominant. We now prove our fundamental characterisation.

PROPOSITION 3.2. *Let X be a subvariety of $\mathbb{C}^r$ all of whose irreducible components have the same minimal projections.*

Suppose that every irreducible component X_i of X contains a special subvariety Z_i with $s = c(\Omega_{Z_i}) > 0$ independent of i and such that

$$Z_i = \{x_1, \ldots, x_s\} \times Z_i'$$

where, as usual, x_is are CM points and Z_i' is strongly special.

Suppose that the first projection of X (or of one of its irreducible components) is dominant.

Suppose that there exists a prime l greater than 13 and greater than $\deg(X)$ such that

$$X \subset T_{\mathbf{l^2}}X$$

Then X is a direct product $X = \mathcal{X}_1 \times \mathcal{X}_2$ of subvarieties of $\mathbb{C}^s$ and $\mathbb{C}^{r-s}$ respectively. Furthermore, the irreducible components of $\mathcal{X}_1$ are special. In particular, each component X_i of X contains a special subvariety Z_i' containing Z_i with $c(\Omega_{Z_i'}) < c(\Omega_{Z_i})$.

Proof. We will use the following lemma.

LEMMA 3.3. *Let I be a subset of $\{1, \ldots, r\}$ which is minimal for every component of X. Then either I is contained in $\{1, \ldots, s\}$ or I is contained in $\{s+1, \ldots, r\}$. Furthermore, if $|I| \geq 3$, then I is contained in $\{s+1, \ldots, r\}$.*

Proof. Suppose that $|I| \geq 3$. We will show that in this case the intersection of I with $\{1, \ldots, s\}$ is empty. Suppose it is not and write $I = I_1 \coprod I_2$ where I_1 is the intersection of I with $\{1, \ldots, s\}$. Let x_2 be a point of $p_{I_2}X$. As I is minimal, the irreducible components of the subvarieties $p_I X$ and $p_I(T_{\mathbf{l}}X)$ of $\mathbb{C}^{|I|}$ are hypersurfaces. The degree of $p_I X$ is at most the degree of X. Proposition 3.1 above shows that for every component Y of $p_I X$, $p_I(T_{\mathbf{l}})Y$ is irreducible, hence $p_I X$ and $p_I T_{\mathbf{l}} X$ have the same number of irreducible components. The inclusion $X \subset T_{\mathbf{l}}X$ implies that $p_I X$ contains a $p_I T_{\mathbf{l}}$-orbit. Lemma 4.4 of [**Edi05**] shows that the orbits of the usual Hecke correspondence T_l on $\mathbb{C}$ are dense. Hence $p_I X$ contains $\mathbb{C}^{I_1} \times \{x_2\}$. It follows that components of $p_I X$ are of the form $\mathbb{C}^{I_1} \times p_{I_2}X_i$

(X_i being the components of X) which contradicts the minimality of I (note that for any i in I_1, the projection on $(I_1 - \{i\}) \cup I_2$ is not dominant).

Suppose now that $|I| = 2$. We will see that either I is contained in $\{1, \dots, s\}$ or in $\{s+1, \dots, r\}$. Indeed, suppose that I is not contained in one of the two sets. Let Z be a special subvariety contained in X as in the statement. It follows that $p_I Z$ is of the form $\{x\} \times \mathbb{C}$ where x is some special point. By minimality of I, this implies that components of $p_I X$ are of the form $\{x\} \times \mathbb{C}$. But this again contradicts the minimality of the set I.

Finally, it can occur that I is a one-element set but then $p_I(X)$ is a special point among $\{x_2, \dots, x_s\}$ and hence I is again contained in $\{1, \dots, s\}$. $\square$

By Lemma 3.5 of [**Edi05**], X is a union of components of the intersection of the $p_I^{-1} p_I X$ with I ranging through minimal sets for X. The intersection of the $p_I^{-1} p_I X$ for I contained in $\{1, \dots, s\}$ is of the form $\mathcal{X}'_1 \times \mathbb{C}^{r-s}$. The intersection for the I contained in $\{s+1, \dots, r\}$ is of the form $\mathbb{C}^s \times \mathcal{X}'_2$. It follows that X is a union of components of the product $\mathcal{X}'_1 \times \mathcal{X}'_2$, say $X = \mathcal{X}_1 \times \mathcal{X}_2$.

It remains to see that the components of $\mathcal{X}_1$ are special. Fix a component say $X_1 = Y_1 \times Y_2$ of X.

Let I be a minimal set for Y_1. Then I is contained in $\{1, \dots, s\}$ and the lemma above shows that either $|I| = 2$ or $|I| = 1$. By Proposition 3.6 of [**Edi05**], it suffices to show that the closure of the image of every minimal projection $p_I(Y_1)$ is special.

If $|I| = 1$, then the projection $p_I Y_1$ of Y_1 is a special point.

Suppose now that $|I| = 2$. Then (the closure of) $p_I Y_1$ is an irreducible curve in $\mathbb{C}^2$. Let $\mathcal{Y}$ be this curve. Furthermore, the degrees of the two projections of $\mathcal{Y}$ are finite (because I is minimal) and bounded above by $\deg(X)$. The proof of Proposition 4.1 of [**Edi05**] shows that $\mathcal{Y}$ is special.

As every minimal set I for Y_1 satisfies $|I| \leq 2$ and $p_I Y_1$ is special, Proposition 3.6 of [**Edi05**] shows that Y_1 is special.

Write $Z_1 = \{x_1, \dots, x_s\} \times \mathcal{Z}_1$ where $\mathcal{Z}_1$ is a strongly special subvariety of $\mathbb{C}^{r-s}$. We now take $Z'_1 = Y_1 \times \mathcal{Z}_1$. $\square$

4. Proof of the main theorem.

This section is devoted to the proof of the following theorem, which is the main result of this note.

THEOREM 4.1. *Assume the GRH for imaginary quadratic fields. Let X be a subvariety of a product S of r modular curves containing a Zariski dense set Σ of special subvarieties. Then the irreducible components of X are special.*

We can assume that all the subvarieties in Σ are of the same type Ω. As X contains a Zariski dense set of special points, X is defined over $\overline{\mathbb{Q}}$. We replace X by a (finite) union of its $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$-conjugates and hence assume that X is irreducible over $\mathbb{Q}$. If $c(\Omega) = 0$, then X is special by the theorem 2.5. Using Proposition 2.1 of [**EY03**], we see that we can (and do) assume that S is the product of r copies of $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ (this is the fact that the level structure does not matter for the André-Oort conjecture). We can also assume that the projections of X to every factor are dominant (simply remove the factors to which X projects as just one, necessarily special, point).

Let s denote $c(\Omega)$. Renumbering the I_i and possibly replacing Σ by a Zariski dense subset allows us to assume that the cardinality of I_i is one for $i = 1, \dots, s$

and the corresponding projection is a CM point. In other words a subvariety Z in Σ can be written as

$$Z = \{x_1, \ldots, x_s\} \times Z'$$

with Z' strongly special (in the product of the $\Gamma_i \backslash \mathbb{H}$ for $i > s$). We can also assume that $d_Z = d_{x_1}$ (renumbering the x_i). Our main theorem is a consequence of the following proposition, the proof of which will occupy the rest of this section.

PROPOSITION 4.2. *Suppose that $c(\Omega) > 0$, then X contains a Zariski dense set of special subvarieties of type Ω' with $c(\Omega') < c(\Omega)$.*

LEMMA 4.3. *Assume the GRH for imaginary quadratic fields. Fix an integer $A \geq 3$. As soon as d_Z is larger than some absolute constant, there exists a prime l which is split in every O_{x_i} and satisfies*

$$(\log d_Z)^A < l < (\log d_Z)^{A+1}$$

Proof. Let us quickly recall a consequence of the effective Chebotarev theorem (that assumes the GRH) in the form presented in the section 6 of [**Edi05**]. We also use [**Edi01**] appendix N. Edixhoven shows that for a given real number $x > (\log d_Z)^3$ (and bigger than some absolute constant), the number $\pi(x)$ of primes $l \leq x$ split in every O_{x_i} satisfies

$$|\pi(x) - \frac{\mathrm{Li}(x)}{n_K}| \leq \frac{\sqrt{x}}{3n_K}(\log(d_Z) + n_K \log(d_Z)) \tag{1}$$

where n_K is the degree of the composite K of the fields of complex multiplication of the x_i. Here $\mathrm{Li}(x) = \int_2^x dt/\log(t)$. Edixhoven further shows in the appendix N to [**Edi01**] that if x is larger than $(\log d_Z)^3$, then $(\log d_Z)\log(x)/3\sqrt{x} < 1/2$. Using this and the facts that $Li(x)\log(x)/x$ tends to 1 and $(\log x)^2/\sqrt{x}$ tends to 0 when x tends to infinity, we deduce that for $x \geq (\log d_Z)^3$ and larger than some absolute constant

$$\frac{x}{3n_K \log(x)} \leq \pi(x) \leq \frac{3x}{2n_K \log(x)}$$

The number of primes we are interested in is

$$\pi((\log d_Z)^{A+1}) - \pi((\log d_Z)^A)$$

Hence the number of primes l satisfying $(\log d_Z)^A < l < (\log d_Z)^{A+1}$ is at least

$$\frac{(\log d_Z)^A}{3n_K \log\log(d_Z)}(\frac{(\log d_Z)}{A+1} - \frac{9}{2A})$$

which is clearly larger than 1 provided d_Z is larger than some absolute constant. □

As Z ranges through Σ, d_Z is unbounded because the first projection of X is dominant and the projection of Σ is Zariski dense (infinite) in $\mathbb{C}$. We can now finish the proof of Theorem 4.2 and hence of 4.1 by induction.

Using equation (1), we choose a prime $l > \max(3, \deg(X))$ split in every O_{x_i} and satisfying

$$l < (\log d_Z)^3$$

If X is contained in $T_{\mathfrak{l}^2}X$, then Proposition 3.2 shows that X contains a special subvariety Z' containing Z with $c(\Omega_{Z'}) < c(\Omega_Z)$.

Suppose now that a geometrically irreducible component X' of X is not contained in $T_{\mathfrak{l}^2}X$. As both X and $T_{\mathfrak{l}^2}X$ are defined over $\mathbb{Q}$, either the intersection of X with $T_{\mathfrak{l}^2}X$ is proper or X is contained in $T_{\mathfrak{l}^2}X$. We make use of the following lemma.

LEMMA 4.4. *Suppose that X is not contained in $T_{\mathfrak{l}^2}X$. We can choose a hypersurface H in $(\mathbb{P}^1_{\overline{\mathbb{Q}}})^r$ such that*

(1) *X is not contained in H but $T_{\mathfrak{l}^2}(X) \subset H$.*
(2) *$\deg(H) \ll \deg(X) l^{2s}$.*

Proof. Let $\overline{X}$ be the closure of X in $(\mathbb{P}^1_{\overline{\mathbb{Q}}})^r$. Let $[\overline{X}] = \sum_{|I|=r-\dim(X)} a_I \epsilon_I$ be the decomposition of the cycle $[\overline{X}]$ in $CH^{r-\dim(X)}(\mathbb{P}^{1\,r})$. Then

$$[\overline{T_{\mathfrak{l}^2}(X)}] = \sum_{|I|=r-\dim(X)} (l^2+l)^s a_I \epsilon_I$$

We choose H to be a hypersurface such that

$$[H] = \Big(\dim(X)! \cdot \sum_{|I|=r-\dim(X)} (l^2+l)^s a_I \Big) \sum_{i=1,\dots,r} \epsilon_i$$

As usual, write $Z = \{x_1, \dots, x_s\} \times Z'$. As l splits every O_{x_i}, Lemma 2.3 implies that some Galois conjugate of Z is contained in $T_{\mathfrak{l}^2} Z \subset T_{\mathfrak{l}^2} X$. By rationality of the Hecke operator $T_{\mathfrak{l}^2}$ we find that $Z \subset X \cap T_{\mathfrak{l}^2} X$. Hence $Z \subset X \cap H$. □

LEMMA 4.5. *Let Y be a $\mathbb{Q}$-component of $X \cap H$ containing Z. The projection of Y on the first factor is dominant.*

Proof. Suppose that the first projection of Y is not dominant. Then some geometrically irreducible component of Y is of the form $\{x_1\} \times Y'$. Therefore, the image of the first projection of Y contains the Galois orbit of x_1. It follows that $[Y]$ is divisible by $O(x_1)\epsilon_1$, where $O(x_1)$ denotes the cardinality of this Galois orbit. It follows that the degree of Y is at least $O(x_1)$. The fact that

$$O(x_1) \gg d_Z^{\frac{1}{2}-\epsilon}$$

contradicts the fact that the degree of H is bounded by a uniform power of $\log(d_Z)$. □

We replace $X := X_1$ by $X_2 := Y$, given by the previous lemma. The degree of X_2 is $\ll (\log d_Z)^A$ where A is some uniform integer. Using Lemma 4.3 we can find a prime l_2 split in every O_{x_i} such that

$$\deg(X_2) < l_2 \ll (\log d_Z)^{A+1}.$$

We now apply the construction just made recursively. If the inclusion did not occur at any of the previous stages, then we end up in the following situation.

(1) $\dim(X_k) = \dim(Z) + 1$
(2) $\deg X_k \ll (\log d_Z)^C$ where C is some uniform integer.
(3) $|\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \cdot Z| \gg d_Z^{\frac{1}{2}-\epsilon}$.
(4) $Z \subset X_k$

Using effective Chebotarev, provided d_Z is large enough, we can choose l_k splitting the O_{x_i}'s such that for an absolute constant C

(1) $l^{2s} \ll (\log d_Z)^C = o(d_Z^{\frac{1}{2}-\epsilon})$ for a small $\epsilon > 0$.
(2) $l_k > \deg(X_k)$.

Then by the lemma 2.3, $Z \subset X_k \cap T_{\mathfrak{l}_k^2} X_k$. The inequalities above show that the intersection $X_k \cap T_{\mathfrak{l}_k^2} X_k$ is not proper, hence X_k is contained in $T_{\mathfrak{l}_k^2} X_k$ and by 3.2 X_k and therefore X contains a special subvariety Z' containing Z with $c(\Omega_{Z'}) < c(\Omega_{Z'})$. This finishes the proof.

References

[CU05] L. Clozel and E. Ullmo, *Équidistribution de sous-variétés spéciales*, Ann. of Math. (2) **161** (2005), no. 3, 1571–1588. MR 2180407 (2006j:11083)

[Edi01] B. Edixhoven, *On the André-Oort conjecture for Hilbert modular surfaces*, Moduli of abelian varieties (Texel Island, 1999), Progr. Math., vol. 195, Birkhäuser, Basel, 2001, pp. 133–155. MR 1827018 (2002c:14042)

[Edi05] ———, *Special points on products of modular curves*, Duke Math. J. **126** (2005), no. 2, 325–348. MR 2115260 (2006g:11119)

[EY03] B. Edixhoven and A. Yafaev, *Subvarieties of Shimura varieties*, Ann. of Math. (2) **157** (2003), no. 2, 621–645. MR 1973057 (2004c:11103)

[KY06] B. Klingler and A. Yafaev, *On the André-Oort conjecture*, 2006, preprint.

[RU] N. Ratazzi and E. Ullmo, *Galois+Equidistribution=Manin-Mumford*, in this volume.

[UY06] E. Ullmo and A. Yafaev, *Galois orbits and equidistribution of special subvarieties of Shimura varieties: towards the André-Oort conjecture*, 2006, preprint, with an appendix by P. Gille and L. Moret-Bailly.

[Yaf07] A. Yafaev, *The André-Oort conjecture—a survey*, L-functions and Galois representations, London Math. Soc. Lecture Note Ser., vol. 320, Cambridge Univ. Press, Cambridge, 2007, Papers from the symposium held at the University of Durham, Durham, July 19–30, 2004, pp. 381–406. MR 2392360

Université de Paris-Sud, Bat 425 et institut universitaire de France, 91405, Orsay Cedex France

E-mail address: ullmo@math.u-psud.fr

University College London, Department of Mathematics, 25 Gordon street, WC1H OAH, London, United Kingdom

E-mail address: yafaev@math.ucl.ac.uk

Clay Mathematics Proceedings
Volume 8, 2009

Moduli of abelian varieties and p-divisible groups

Ching-Li Chai and Frans Oort

Abstract. This is a set of notes for a course we gave in the second week of August in the 2006 CMI Summer School at Göttingen. Our main topic is geometry and arithmetic of $\mathcal{A}_g \otimes \mathbb{F}_p$, the moduli space of polarized abelian varieties of dimension g in positive characteristic. We illustrate properties of $\mathcal{A}_g \otimes \mathbb{F}_p$, and some of the available techniques by treating two topics: 'Density of ordinary Hecke orbits' and 'A conjecture by Grothendieck on deformations of p-divisible groups'.

Contents

We present proofs of two recent results. The main point is that the *methods* used for these proofs are interesting. The emphasis will be on the various techniques available.

In characteristic zero we have strong tools at our disposal: besides algebraic-geometric theories we can use analytic and topological methods. It would seem that we are at a loss in positive characteristic. However the opposite is true. Phenomena

2000 *Mathematics Subject Classification.* Primary 14L05, Secondary 14K10, 14L15.

The first author was partially supported by a grant DMS04-00482 from the National Science Foundation.

occurring only in positive characteristic provide us with strong tools to study moduli spaces. And, as it turns out again and again, several results in characteristic zero can be derived using reduction modulo p. The discussion of *tools in positive characteristic* will be the focus of our notes.

Here is a list of some of the central topics:

- Serre-Tate theory.
- Abelian varieties over finite fields.
- Monodromy: ℓ-adic and p-adic, geometric and arithmetic.
- Dieudonné modules and Newton polygons.
- Theory of Dieudonné modules, Cartier modules and displays.
- Cayley-Hamilton and deformations of p-divisible groups.
- Hilbert modular varieties.
- Purity of the Newton polygon stratification in families of p-divisible groups.

The strategy is that we have chosen certain central topics, and for those we took ample time for explanation and for proofs. Besides that we need certain results which we label as "Black Box". These are results which we need for our proofs, which are either fundamental theoretical results (but it would take too much time to explain their proofs), or lemmas which are computational, important for the proof, but not very interesting to explain in a course. We hope that we explain well enough what every relevant statement is. We write:

BB A Black Box, please accept that this result is true.
Th One of the central results, we will explain it.
Extra A result which is interesting but was not discussed in the course.

Notation to be used will be explained in Section 10. In order to be somewhat complete we will gather related interesting results, questions and conjectures in Section 11. Part of our general convention is that K denotes a field of characteristic $p > 0$, unless otherwise specified, and k denotes an algebraically closed field.

We assume that the reader is familiar with the basic theory of abelian varieties at the level of Chapter II of [**54**] and [**55**], Chapter 6; we consider abelian varieties over an arbitrary field, and abelian schemes over a base scheme. Alternative references: [**16**], [**81**]. For the main characters of our play: *abelian varieties, moduli spaces, and p-divisible groups*, we give references and definitions in Section 10.

1. Introduction: Hecke orbits, and the Grothendieck conjecture

In this section we discuss the two theorems we are going to consider.

1.1. An abelian variety A of dimension g over a field $K \supset \mathbb{F}_p$ is said to be *ordinary* if

$$\#\left(A[p](k)\right) = p^g.$$

More generally, the number $f = f(A)$ such that $\#\left(A[p](k)\right) = p^f$ is called the *p-rank* of A. It is a fact that the p-rank of A is at most $\dim(A)$; an abelian variety is ordinary if its p-rank is equal to its dimension. See 10.10 for other equivalent definitions.

An elliptic curve E over a field $K \supset \mathbb{F}_p$ is said to be *supersingular* if it is not ordinary; equivalently, E is supersingular if $E[p](k) = 0$ for any overfield $k \supset K$. This

terminology stems from Deuring: an elliptic curve in characteristic zero is said to determine a *singular* j-value if its endomorphism ring over an algebraically closed field (of characteristic 0) is larger than $\mathbb{Z}$ (therefore of rank 2 over $\mathbb{Z}$), while a *supersingular* elliptic curve E over an algebraically closed field $k \supset \mathbb{F}_p$ has $\mathrm{rk}_{\mathbb{Z}}(\mathrm{End}(E)) = 4$. Since an elliptic curve is non-singular, a better terminology would be "an elliptic curve with a singular j-invariant".

We say an abelian variety A of dimension g over a field K is *supersingular* if there exists an isogeny $A \otimes_K k \sim E^g$, where E is a supersingular elliptic curve. An equivalent definition for an abelian variety in characteristic p to be supersingular is that all of its slopes are equal to $1/2$; see 4.38 for the definition of slopes and the Newton polygon. Supersingular abelian varieties have p-rank zero. For $g = 2$ one can show that (supersingular) $\Leftrightarrow$ ($f = 0$), where f is the p-rank. For $g > 2$ there exist abelian varieties of p-rank zero which are not supersingular, see 5.22.

Hecke orbits

Definition 1.2. Let A and B be abelian varieties over a field K. Let $\Gamma \subset \mathbb{Q}$ be a subring. A Γ-isogeny from A to B is an element ψ of $\mathrm{Hom}(A, B) \otimes_{\mathbb{Z}} \Gamma$ which has an inverse in $\mathrm{Hom}(B, A) \otimes_{\mathbb{Z}} \Gamma$, i.e., there exists an element $\psi' \in \mathrm{Hom}(B, A) \otimes_{\mathbb{Z}} \Gamma$ such that $\psi'\,\psi = \mathrm{id}_A \otimes 1$ in $\mathrm{Hom}(A, A) \otimes_{\mathbb{Z}} \Gamma$ and $\psi\,\psi' = \mathrm{id}_B \otimes 1$ in $\mathrm{Hom}(B, B) \otimes_{\mathbb{Z}} \Gamma$.

Remark.

(i) When $\Gamma = \mathbb{Q}$ (resp. $\Gamma = \mathbb{Z}_{(p)}$, resp. $\mathbb{Z}[1/\ell]$), we say that ψ is a *quasi-isogeny* (resp. *prime-to-p quasi-isogeny*, resp. an *ℓ-power quasi-isogeny*). A *prime-to-p isogeny* (resp. *ℓ-power isogeny*) is an isogeny which is also a $\mathbb{Z}_{(p)}$-isogeny (resp. a $\mathbb{Z}[1/\ell]$-isogeny). Here $\mathbb{Z}_{(p)} = \mathbb{Q} \cap \mathbb{Z}_p$ is the localization of $\mathbb{Z}$ at the prime ideal $(p) = p\mathbb{Z}$.

(ii) A $\mathbb{Q}$-isogeny ψ (resp. $\mathbb{Z}_{(p)}$-isogeny, resp. $\mathbb{Z}[1/\ell]$-isogeny) can be realized by a diagram

$$A \xleftarrow{\alpha} C \xrightarrow{\beta} B\,,$$

where α and β are isogenies such that there exists an integer $N \in \Gamma^\times$ (resp. an integer N prime to p, resp. an integer $N \in \ell^{\mathbb{N}}$) such that $N \cdot \mathrm{Ker}(\alpha) = N \cdot \mathrm{Ker}(\beta) = 0$.

Definition 1.3. Let $[(A, \lambda)] = x \in \mathcal{A}_g(K)$ be the moduli point of a polarized abelian variety over a field K.

(i) We say that a point $[(B, \mu)] = y$ of $\mathcal{A}_g$ is in the *Hecke orbit* of x if there exists a field Ω and

$$\text{a } \mathbb{Q}\text{-isogeny } \varphi : A_\Omega \to B_\Omega \text{ such that } \varphi^*(\mu) = \lambda.$$

Notation: $y \in \mathcal{H}(x)$. The set $\mathcal{H}(x)$ is called the *Hecke orbit* of x.

(ii) *Hecke-prime-to-p-orbits.* If in the previous definition moreover φ is a $\mathbb{Z}_{(p)}$-isogeny, we say $[(B, \mu)] = y$ is in the Hecke-prime-to-p-orbit of x.
Notation: $y \in \mathcal{H}^{(p)}(x)$.

(iii) *Hecke-ℓ-orbits.* Fix a prime number ℓ different from p. We say $[(B, \mu)] = y$ is in the Hecke-ℓ-orbit of x if in the previous definition moreover φ is a $\mathbb{Z}[1/\ell]$-isogeny.
Notation: $y \in \mathcal{H}_\ell(x)$.

(iv) **Notation:** Suppose that $x = [(A, \lambda)]$ is a point of $\mathcal{A}_{g,1}$, i.e., λ is principal. Write

$$\mathcal{H}^{(p)}_{\mathrm{Sp}}(x) := \mathcal{H}^{(p)}(x) \cap \mathcal{A}_{g,1}\,, \qquad \mathcal{H}^{\mathrm{Sp}}_{\ell}(x) := \mathcal{H}_{\ell}(x) \cap \mathcal{A}_{g,1} \quad (\ell \neq p)\,.$$

Remark.

(i) Clearly we have $\mathcal{H}_{\ell}(x) \subseteq \mathcal{H}^{(p)}(x) \subseteq \mathcal{H}(x)$. Similarly we have

$$\mathcal{H}^{\mathrm{Sp}}_{\ell}(x) \subseteq \mathcal{H}^{(p)}_{\mathrm{Sp}}(x) \quad \text{for} \quad x \in \mathcal{A}_{g,1}.$$

(ii) Note that $y \in \mathcal{H}(x)$ is equivalent to requiring the existence of a diagram

$$(B, \mu) \xleftarrow{\psi} (C, \zeta) \xrightarrow{\varphi} (A, \lambda).$$

such that $\psi^*\mu = \zeta = \varphi^*\lambda$, where ϕ and ψ are isogenies, $[(B, \mu)] = y$, $[(A, \lambda)] = x$. If we have such a diagram such that both ψ and φ are $\mathbb{Z}_p$-isogenies (resp. $\mathbb{Z}[1/\ell]$-isogenies), then $y \in \mathcal{H}^{(p)}(x)$ (resp. $y \in \mathcal{H}_{\ell}(x)$.)

(iii) We have given the definition of the so-called Sp_{2g}-Hecke-orbit. One can also define the (slightly bigger) $\mathrm{CSp}_{2g}(\mathbb{Z})$-Hecke-orbits by the usual Hecke correspondences, see [**28**], VII.3, also see 1.7 below.

(iv) The diagram which defines $\mathcal{H}(x)$ as above gives representable correspondences between components of the moduli scheme; these correspondences could be denoted by Sp-Isog, whereas the correspondences considered in [**28**], VII.3 could be denoted by CSp-Isog.

1.4. Why are Hecke orbits interesting? Here we work first over $\mathbb{Z}$. A short answer is that they are manifestations of the Hecke symmetry on $\mathcal{A}_g$. The Hecke symmetry is a salient feature of the moduli space $\mathcal{A}_g$; methods developed for studying Hecke orbits have been helpful for understanding the Hecke symmetry.

To explain what the Hecke symmetry is, we will focus on $\mathcal{A}_{g,1}$, the moduli space of principally polarized abelian varieties and the prime-to-p power, the projective system of $\mathcal{A}_{g,1,n}$ over $\mathbb{Z}[1/n]$. The group $\mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$ of finite prime-to-p adelic points of the symplectic group Sp_{2g} operates on this tower, and induces finite correspondences on $\mathcal{A}_{g,1}$; these finite correspondences are known as *Hecke correspondences*. By the term Hecke symmetry we refer to both the action on the tower of moduli spaces and the correspondences on $\mathcal{A}_{g,1}$. The prime-to-p Hecke orbit $\mathcal{H}^{(p)}(x) \cap \mathcal{A}_{g,1}$ is exactly the orbit of x under the Hecke correspondences coming from the group $\mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$. In characteristic 0, the moduli space of g-dimensional principally polarized abelian varieties is uniformized by the Siegel upper half space $\mathbb{H}_g$, consisting of all symmetric $g \times g$ matrices in $\mathrm{M}_g(\mathbb{C})$ whose imaginary part is positive definite: $\mathcal{A}_{g,1}(\mathbb{C}) \cong \mathrm{Sp}_{2g}(\mathbb{Z})\backslash\mathbb{H}_g$. The group $\mathrm{Sp}_{2g}(\mathbb{R})$ operates transitively on $\mathbb{H}_g$, and the action of the rational elements $\mathrm{Sp}_{2g}(\mathbb{Q})$ give a family of algebraic correspondences on $\mathrm{Sp}_{2g}(\mathbb{Z})\backslash\mathbb{H}_g$. These algebraic correspondences are of fundamental importance for harmonic analysis on arithmetic quotients such as $\mathrm{Sp}_{2g}(\mathbb{Z})\backslash\mathbb{H}_g$.

Remark/Exercise 1.5. (Characteristic zero.) The Hecke orbit of a point in the moduli space $\mathcal{A}_g \otimes \mathbb{C}$ in *characteristic zero* is dense in that moduli space for both the metric topology and the Zariski topology.

1.6. Hecke orbits of elliptic curves. Consider the moduli point $[E] = j(E) = x \in \mathcal{A}_{1,1} \cong \mathbb{A}^1$ of an elliptic curve in characteristic p. Here $\mathcal{A}_{1,1}$ stands for $\mathcal{A}_{1,1} \otimes \mathbb{F}_p$. Note that every elliptic curve has a unique principal polarization.

(1) *If E is supersingular $\mathcal{H}(x) \cap \mathcal{A}_{1,1}$ is a finite set; we conclude that $\mathcal{H}(x)$ is nowhere dense in $\mathcal{A}_1$.*

Indeed, the supersingular locus in $\mathcal{A}_{1,1}$ is closed, there do exist ordinary elliptic curves, and hence that locus is finite; Deuring and Igusa computed the exact number of geometric points in this locus.

(2) **Remark/Exercise.** *If E is ordinary, its Hecke-ℓ-orbit is dense in $\mathcal{A}_{1,1}$.*

There are several ways to prove this. Easy and direct considerations show that in this case $\mathcal{H}_\ell(x) \cap \mathcal{A}_{1,1}$ is not finite; note that every component of $\mathcal{A}_1$ has dimension one, and conclude $\mathcal{H}(x)$ is dense in $\mathcal{A}_1$.

Remark. For elliptic curves we have defined (supersingular) $\Leftrightarrow$ (non-ordinary). For $g = 2$ one can show that (supersingular) $\Leftrightarrow$ ($f = 0$), where f is the p-rank. For $g > 2$ there exist abelian varieties of p-rank zero which are not supersingular, see 5.22.

1.7. A bigger Hecke orbit. We work over $\mathbb{Z}$. We define the notion of CSp-Hecke orbits on $\mathcal{A}_{g,1}$. Two K-points $[(A, \lambda)]$, $[(B, \mu)]$ of $\mathcal{A}_{g,1}$ are in the same CSp-Hecke orbit (resp. prime-to-p CSp-Hecke orbit, resp. ℓ-power CSp-Hecke orbit) if there exists an isogeny $\varphi : A \otimes k \to B \otimes k$ and a positive integer n (resp. a positive integer n which is relatively prime to p, resp. a positive integer which is a power of ℓ) such that $\varphi^*(\mu) = n \cdot \lambda$. Such Hecke correspondences are representable by a morphism $\mathrm{Isog}_g \to \mathcal{A}_g \times \mathcal{A}_g$ on $\mathcal{A}_g$, also see [**28**], VII.3.

The set of all such (B, μ) for a fixed $x := [(A, \lambda)]$ is called the CSp-Hecke orbit (resp. $\mathrm{CSp}(\mathbb{A}_f^{(p)})$-Hecke orbit, resp. $\mathrm{CSp}(\mathbb{Q}_\ell)$-Hecke orbit) of x, and denoted $\mathcal{H}^{\mathrm{CSp}}(x)$ (resp. $\mathcal{H}^{(p)}_{\mathrm{CSp}}(x)$, resp. $\mathcal{H}^{\mathrm{CSp}}_\ell(x)$). Note that

$$\mathcal{H}^{\mathrm{CSp}}(x) \supset \mathcal{H}^{(p)}_{\mathrm{CSp}}(x) \supset \mathcal{H}^{\mathrm{CSp}}_\ell(x),$$

$\mathcal{H}^{\mathrm{CSp}}(x) \supset \mathcal{H}(x)$, $\mathcal{H}^{(p)}_{\mathrm{CSp}}(x) \supset \mathcal{H}^{(p)}(x)$ and $\mathcal{H}^{\mathrm{CSp}}_\ell(x) \supset \mathcal{H}_\ell(x)$. This slightly bigger Hecke orbit will play no role in this paper. However, it is nice to see the relation between the Hecke orbits defined previously in 1.3, which could be called the Sp-Hecke orbits and Sp-Hecke correspondences, with the CSp-Hecke orbits and CSp-Hecke correspondences.

Theorem 1.8. **Th** (Density of ordinary Hecke orbits) *Let $[(A, \lambda)] = x \in \mathcal{A}_g \otimes \mathbb{F}_p$ be the moduli point of a polarized* ordinary *abelian variety in characteristic p.*

(i) *If the polarization λ is separable, then $\left(\mathcal{H}^{(p)}(x) \cap \mathcal{A}_{g,1}\right)^{\mathrm{Zar}} = \mathcal{A}_{g,1}$. If $\deg(\lambda) \in \ell^{\mathbb{N}}$ for a prime number $\ell \neq p$, then $\left(\mathcal{H}_\ell(x) \cap \mathcal{A}_{g,1}\right)^{\mathrm{Zar}} = \mathcal{A}_{g,1}$.*

(ii) *From* (i) *we conclude that $\mathcal{H}(x)$ is dense in $\mathcal{A}_g$, with no restriction on the degree of λ.*

See Theorem 9.1. This theorem was proved by Ching-Li Chai in 1995; see [**9**], Theorem 2 on page 477. Although CSp-Hecke orbits were used in [**9**], the same argument works for Sp-Hecke orbits as well. We present a proof of this theorem; we follow [**9**] partly, but also present a new insight which was necessary for solving the general Hecke orbit problem. This final strategy will provide us with a proof which seems easier than the one given previously. More information on the general Hecke orbit problem can be obtained from [**10**] as long as [**13**] is not yet available.

Exercise 1.9. (Any characteristic.) Let k be any algebraically closed field (of any characteristic). Let E be an elliptic curve over k such that $\mathrm{End}(E) = \mathbb{Z}$. Let

ℓ be a prime number different from the characteristic of k. Let E' be an elliptic curve such that there exists an isomorphism $E'/\underline{(\mathbb{Z}/\ell)}_k \cong E$. Let λ be the principal polarization on E, let μ be the pullback of λ to E', hence μ has degree ℓ^2, and let $\mu' = \mu/\ell^2$, hence μ' is a principal polarization on E'. *Note that* $[(E', \mu')] \in \mathcal{H}(x)$. *Show that* $[(E', \mu')] \notin \mathcal{H}^{\mathrm{Sp}}(x)$.

Exercise 1.10. Let E be an elliptic curve in characteristic p which is not supersingular (hence ordinary); let μ be any polarization on E, and $x := [(E, \mu)]$. Show $\mathcal{H}^{\mathrm{Sp}}(x)$ is dense in $\mathcal{A}_1$.

1.11. (1) Let $\mathrm{Isog}_{\mathrm{g,Sp}}$ be the moduli space which classifies diagrams of polarized g-dimensional abelian schemes

$$(B, \mu) \xleftarrow{\psi} (C, \zeta) \xrightarrow{\varphi} (A, \lambda)$$

in characteristic p such that $\psi^*\mu = \zeta = \varphi^*\lambda$. Consider a component I of $\mathrm{Isog}_{g,\mathrm{Sp}}$ defined by diagrams as in 1.7 with $\deg(\psi) = b$ and $\deg(\varphi) = c$. If b is not divisible by p, the first projection $\mathcal{A}_g \leftarrow I$ is étale; if c is not divisible by p, the second projection $I \to \mathcal{A}_g$ is étale.
(2) Consider $\mathrm{Isog}^{\mathrm{ord}}_{g,\mathrm{Sp}} \subset \mathrm{Isog}_{g,\mathrm{Sp}}$, the largest subscheme (it is locally closed) lying over the ordinary locus (either in the first projection, or in the second projection, which is the same).

Exercise. Show that the two projections $(\mathcal{A}_g)^{\mathrm{ord}} \leftarrow \mathrm{Isog}^{\mathrm{ord}}_{g,\mathrm{Sp}} \to (\mathcal{A}_g)^{\mathrm{ord}}$ are both surjective, finite and flat.

(3) **Extra** Let Z be an irreducible component of $\mathrm{Isog}_{g,\mathrm{Sp}}$ over which the polarizations μ, λ are principal, and such that ζ is a multiple of a principal polarization. Then the projections $\mathcal{A}_{g,1} \leftarrow Z \to \mathcal{A}_{g,1}$ are both surjective and proper. This follows from [**28**], VII.4. The previous exercise (2) is easy; the fact (3) here is more difficult; it uses the computation in [**56**].

1.12. We explain the reason to focus our attention on $\mathcal{A}_{g,1} \otimes \mathbb{F}_p$, the moduli space of principally polarized abelian varieties in characteristic p.
(1) **BB** In [**56**] it has been proved that $(\mathcal{A}_g)^{\mathrm{ord}}$ *is dense in* $\mathcal{A}_g = \mathcal{A}_g \otimes \mathbb{F}_p$.
(2) We show that for an ordinary $[(A, \lambda)] = x$ we have

$$(\mathcal{H}_\ell(z) \cap \mathcal{A}_{g,1})^{\mathrm{Zar}} = \mathcal{A}_{g,1} \quad \forall z \in \mathcal{A}_{g,1} \quad \Longrightarrow \quad (\mathcal{H}(x))^{\mathrm{Zar}} = \mathcal{A}_g.$$

Work over k. In fact, consider an irreducible component T of $\mathcal{A}_g$. As proved in [**56**] there is an ordinary point $y = [(B, \mu)] \in T$. By [**54**], Corollary 1 on page 234, we see that there is an isogeny $(B, \mu) \to (A, \lambda)$, where λ is a principal polarization. By 1.11 (2) we see that density of $\mathcal{H}_\ell(x) \cap \mathcal{A}_{g,1}$ in $\mathcal{A}_{g,1}$ implies density of $\mathcal{H}_\ell(x) \cap T$ in T. □

Therefore, from now on we shall be mainly interested in Hecke orbits in the principally polarized case.

Theorem 1.13. **Extra** (Ching-Li Chai and Frans Oort) *For any* $[(A, \mu)] = x \in \mathcal{A}_g \otimes \mathbb{F}_p$ *with* $\xi = \mathcal{N}(A)$, *the Hecke orbit* $\mathcal{H}(x)$ *is dense in the Newton polygon stratum* $\mathcal{W}_\xi(\mathcal{A}_g \otimes \mathbb{F}_p)$.

A proof will be presented in [**13**]. For a definition of Newton polygon strata and the fact that they are closed in the moduli space, see 1.19, 1.20. Note that in

case $f(A) \leq g-2$ the ℓ-Hecke orbit is not dense in $\mathcal{W}_\xi(\mathcal{A}_g \otimes \mathbb{F}_p)$. In [**68**], 6.2 we find a precise conjectural description of the Zariski closure of $\mathcal{H}_\ell(x)$; that conjecture has now been proven and it implies 1.13.

Lemma 1.14. **BB** (Chai) *Let $[(A,\lambda)] = x \in \mathcal{A}_{g,1}$. Suppose that A is supersingular. Then $\mathcal{H}^{(p)}(x) \cap \mathcal{A}_{g,1}$ is finite, therefore $\mathcal{H}_\ell(x) \cap \mathcal{A}_{g,1}$ is finite for every prime number $\ell \neq p$. Conversely if $\mathcal{H}_\ell(x) \cap \mathcal{A}_{g,1}$ is finite for a prime number $\ell \neq p$, then x is supersingular.*

See [**9**], Proposition 1 on page 448 for a proof. Note that $\mathcal{H}(x)$ equals *the whole supersingular Newton polygon stratum: the prime-to-p Hecke orbit is small, but the Hecke orbit including p-power quasi-isogenies is large.* Lemma 1.14 will be used in 3.22.

A conjecture by Grothendieck

Definition 1.15. ***p*-divisible groups.** Let $h \in \mathbb{Z}_{>0}$ be a positive integer, and let S be a base scheme. A p-divisible group of height h over S is an inductive system of finite, locally free commutative group scheme G_i over S indexed by $i \in \mathbb{N}$, satisfying the following conditions.

(1) The group scheme $G_i \to S$ is of rank p^{ih} for every $i \geq 0$. In particular G_0 is the constant trivial group scheme over S.
(2) The subgroup scheme $G_{i+1}[p^i]$ of p^i-torsion points in G_{i+1} is equal to G_i for every $i \geq 0$.
(3) For each $i \geq 0$, the endomorphism $[p]_{G_{i+1}} : G_{i+1} \to G_{i+1}$ of G_{i+1} factors as $\iota_{i+1,i} \circ \psi_{i+1,i}$, where $\psi_{i+1,i} : G_{i+1} \to G_i$ is a faithfully flat homomorphism, and $\iota_{i+1,i} : G_i \hookrightarrow G_{i+1}$ is the inclusion.

Homomorphisms between p-divisible groups are defined by

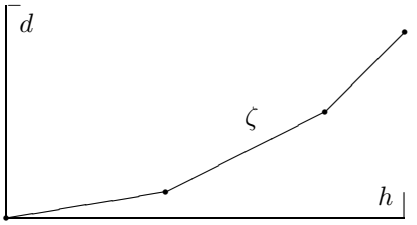
$$\mathrm{Hom}(\{G_i\},\{H_j\}) \quad = \varprojlim_i \varinjlim_j \mathrm{Hom}(G_i, H_j).$$

A p-divisible group is also called a Barsotti-Tate group. It is clear that one can generalize the definition of p-divisible groups so that the height is a locally constant function $h : S \to \mathbb{Z}$ on the base scheme S. For more information see [**38**], Section 1. Also see 10.6, and see Section 10 for further information.

In order to being able to handle the isogeny class of $A[p^\infty]$ we need the notion of Newton polygons.

1.16. Newton polygons. Suppose we are given integers $h, d \in \mathbb{Z}_{\geq 0}$; here h = "height", d = "dimension". In the case of abelian varieties we will choose $h = 2g$, and $d = g$. A Newton polygon γ (related to h and d) is a polygon $\gamma \subset \mathbb{R} \times \mathbb{R}$, such that:

- γ starts at $(0,0)$ and ends at (h,d);
- γ is lower convex;
- every slope β of γ has the property that $0 \leq \beta \leq 1$;
- the breakpoints of γ are in $\mathbb{Z} \times \mathbb{Z}$; hence $\beta \in \mathbb{Q}$.



In the above, γ being lower convex means that the region in $\mathbb{R}^2$ above γ is a convex subset of $\mathbb{R}^2$, or equivalently, γ is the graph of a piecewise linear continuous function $f : [0, h] \to \mathbb{R}$ such that $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)$ for all $x, y \in [0, h]$.

Note that a Newton polygon determines (and is determined by) its slope sequence

$$\beta_1, \ldots, \beta_h \in \mathbb{Q} \text{ with } 0 \leq \beta_1 \leq \cdots \leq \beta_h \leq 1 \quad \leftrightarrow \quad \zeta.$$

Remark.
(i) The last condition above implies that the multiplicity of any slope β in a Newton polygon is a multiple of the denominator of β.
(ii) We imposed the condition that all slopes are between 0 and 1 because we only consider Newton polygons attached to p-divisible groups or abelian varieties. This condition should be eliminated when one considers Newton polygons attached to general (iso)crystals.

Sometimes we will give a Newton polygon by data of the form $\sum_i (m_i, n_i)$, where $m_i, n_i \in \mathbb{Z}_{\geq 0}$, with $\gcd(m_i, n_i) = 1$, and $m_i/(m_i + n_i) \leq m_j/(m_j + n_j)$ for $i \leq j$. The Newton polygon attached to $\sum_i (m_i, n_i)$ can be described as follows. Its height h is given by the formula $h = \sum_i (m_i + n_i)$, its dimension d is given by the formula $d = \sum_i m_i$, and the multiplicity of any rational number β as a slope is $\sum_{m_i = \beta(m_i+n_i)} (m_i + n_i)$. Conversely it is clear that every Newton polygon can be encoded in a unique way in such a form.

Remark. The Newton polygon of a polynomial. Let $g \in \mathbb{Q}_p[T]$ be a monic polynomial of degree h. We are interested in the p-adic values of its zeroes (in an algebraic closure of $\mathbb{Q}_p$). These can be computed by the Newton polygon of this polynomial. Write $g = \sum_j \gamma_j T^{h-j}$. Plot the pairs $(j, v_p(\gamma_j))$ for $0 \leq j \leq h$, where v_p is the valuation on $\mathbb{Q}_p$ with $v_p(p) = 1$. Consider the lower convex hull of $\{(j, v_p(\gamma_j)) \mid j\}$. This is a Newton polygon according to the definition above. The slopes of the sides of this polygon are precisely the p-adic values of the zeroes of g, ordered in non-decreasing order. (Suggestion: prove this as an exercise.)

Later we will see: *a p-divisible group X over a field of characteristic p determines a Newton polygon.* In Section 4 a correct and precise definition will be given. Isogenous p-divisible groups have the same Newton polygon. Moreover a theorem by Dieudonné and Manin says that the isogeny class of a p-divisible group over an algebraically closed field $k \supset \mathbb{F}_p$ is uniquely determined by its Newton polygon; see [**48**], "Classification Theorem" on page 35 and 4.42.

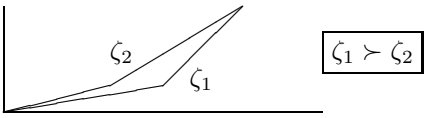
(Incorrect.) Here we indicate what the Newton polygon of a p-divisible group is (in a slightly incorrect way...). Consider "the Frobenius endomorphism" of X. This has a "characteristic polynomial". This polynomial determines a Newton polygon, which we write as $\mathcal{N}(X)$, the Newton polygon of X. For an abelian variety A we write $\mathcal{N}(A)$ instead of $\mathcal{N}(A[p^\infty])$.

Well, this "definition" is correct over $\mathbb{F}_p$ as ground field. However, over any other field, $F : X \to X^{(p)}$ is not an endomorphism, and the above "construction" fails. Over a finite field there is a method which repairs this, see 3.8. However we need the Newton polygon of an abelian variety over an arbitrary field. Please accept for the time being the "explanation" given above: $\mathcal{N}(X)$ is the "Newton polygon of the Frobenius on X", which will be made precise later, see Section 4. There is also a more conceptual way of defining the Newton polygon than the definition in Section 4: the slopes measures divisibility properties of tensor constructions of the crystal attached to a p-divisible group; see [**40**].

Examples.
(1) The Newton polygon of $\mathbb{G}_m[p^\infty]_{\mathbb{F}_p}$ has one slope (counting multiplicity), which is equal to 1. In fact, on $\mathbb{G}_m$ the Frobenius endomorphism is $[p]$.
(2) The Newton polygon of the constant p-divisible group $\underline{\mathbb{Q}_p/\mathbb{Z}_p}_{\mathbb{F}_p}$ has one slope (counting multiplicity), which is equal to 0.
(3) The Newton polygon of an ordinary elliptic curve has two slopes, equal to 0 and to 1, each with multiplicity one.
(4) The Newton polygon of a supersingular elliptic curve has two slopes, both equal to 1/2.

1.17. Newton polygons go up under specialization. Grothendieck observed in 1970 that "Newton polygons go up" under specialization. See 1.20, 4.47 for more information. In order to study this and related questions we introduce the notation of a *partial ordering* between Newton polygons.

We write $\zeta_1 \succ \zeta_2$ if ζ_1 is "below" ζ_2, i.e., if no point of ζ_1 is strictly above ζ_2:



Note that we use this notation only if Newton polygons with the same endpoints are considered. A note on convention: we write $\prec$ instead of $\preceq$, so we have $\zeta \succ \zeta$ for any Newton polygon ζ.

This notation may seem unnatural. However if ζ_1 is strictly below ζ_2 the stratum defined by ζ_1 is larger than the stratum defined by ζ_2; this explains the choice for this notation.

1.18. Later in Section 4 we will show that isogenous p-divisible groups have the same Newton polygon. We will also see in 4.40 that if $\mathcal{N}(X)$ is given by $\{\beta_i \mid 1 \le i \le h\}$ then $\mathcal{N}(X^t)$ is given by $\{1-\beta_h, \dots, 1-\beta_1\}$.

A Newton polygon ξ, given by the slopes $\beta_1 \le \cdots \le \beta_h$ is called *symmetric* if $\beta_i = 1 - \beta_{h+1-i}$ for all i. We see that $X \sim X^t$ implies that $\mathcal{N}(X)$ is symmetric; in particular for an abelian variety A we see that $\mathcal{N}(A)$ is symmetric. This was proved over finite fields by Manin, see [**48**], page 70; for any base field we can use the duality theorem over any base, see [**61**], Th. 19.1, also see 10.11.

1.19. If S is a base scheme over $\mathbb{F}_p$, and $\mathcal{X} \to S$ is a p-divisible group over S and ζ is a Newton polygon we write

$$\mathcal{W}_\zeta(S) := \{s \in S \mid \mathcal{N}(\mathcal{X}_s) \prec \zeta\} \subset S$$

and

$$\mathcal{W}^0_\zeta(S) := \{s \in S \mid \mathcal{N}(\mathcal{X}_s) = \zeta\} \subset S.$$

Theorem 1.20. **BB** (Grothendieck and Katz; see [**40**], 2.3.2)

$$\mathcal{W}_\zeta(S) \subset S \quad \textit{is a closed set.}$$

Working over $S = \mathrm{Spec}(K)$, where K is a perfect field, $\mathcal{W}_\zeta(S)$ and $\mathcal{W}^0_\zeta(S)$ will be given the induced reduced scheme structure.

As the set of Newton polygons of a given height is finite we conclude:

$$\mathcal{W}^0_\zeta(S) \subset S \quad \textit{is a locally closed set.}$$

Notation. Let ξ be a symmetric Newton polygon. We write $W_\xi = \mathcal{W}_\xi(\mathcal{A}_{g,1} \otimes \mathbb{F}_p)$.

1.21. We have seen that "Newton polygons go up under specialization". Does a kind of converse hold? In 1970 Grothendieck conjectured the converse. In the appendix of [**34**] is a letter of Grothendieck to Barsotti, with the following passage on page 150: *"The wishful conjecture I have in mind now is the following: the necessary conditions ... that G' be a specialization of G are also sufficient. In other words, starting with a BT group $G_0 = G'$, taking its formal modular deformation ... we want to know if every sequence of rational numbers satisfying ... these numbers occur as the sequence of slopes of a fiber of G as some point of S."*

Theorem 1.22. **Th** (The Grothendieck Conjecture, Montreal 1970) *Let K be a field of characteristic p, and let X_0 be a p-divisible group over K. We write $\mathcal{N}(\mathcal{X}_0) =: \beta$ for its Newton polygon. Given a Newton polygon γ "below" β, i.e., $\beta \prec \gamma$, there exists a deformation X_η of X_0 such that $\mathcal{N}(\mathcal{X}_\eta) = \gamma$.*

See §8. This was proved by Frans Oort in 2001. For a proof see [**20**], [**65**], [**66**].

We say "X_η is a deformation of X_0" if there exists an integral scheme S over K, with generic point $\eta \in S$ and $0 \in S(K)$, and a p-divisible group $\mathcal{X} \to S$ such that $\mathcal{X}_0 = X_0$ and $\mathcal{X}_\eta = X_\eta$.

A (quasi-) polarized version will be given later.

In this paper we record a proof of this theorem, and we will see that this is an important tool in understanding Newton polygon strata in $\mathcal{A}_g$ in characteristic p.

Why is the proof of this theorem difficult? A direct approach seems obvious: write down deformations of X_0, compute Newton polygons of all fibers, and inspect whether all relevant Newton polygons appear in this way. However, computing the Newton polygon of a p-divisible group in general is difficult (but see Section 5 on how to circumvent this in an important special case). Moreover, abstract deformation theory is easy, but in general Newton polygon strata are "very singular"; in Section 7 we describe how to "move out" of a singular point to a non-singular point of a Newton polygon stratum. Then, at non-singular points the deformation theory can be described more easily, see Section 5. By a combination of these two methods we achieve a proof of the Grothendieck conjecture. Later we will formulate and prove the analogous "polarized case" of the Grothendieck conjecture, see Section 8.

We see: a direct approach did not work, but the detour via "deformation to $a \leq 1$" plus the results via Cayley-Hamilton gave the essential ingredients for a proof. Note the analogy of this method with the approach to liftability of abelian varieties to characteristic zero, as proposed by Mumford, and carried out in [**56**].

2. Serre-Tate theory

In this section we explain the deformation theory of abelian varieties and p-divisible groups. The content can be divided into several parts:

(1) In 2.1 we give the formal definitions of deformation functors for abelian varieties and p-divisible groups.
(2) In contrast to the deformation theory for general algebraic varieties, the deformation theory for abelian varieties and p-divisible groups can be efficiently dealt with by linear algebra, as follows from the crystalline deformation theory of Grothendieck-Messing. It says that, over an extension by an ideal with a divided power structure, deforming abelian varieties or p-divisible groups is the same as lifting the Hodge filtration. See Theorem 2.4 for the precise statement, and Theorem 2.11 for the behavior of the theory under duality. The smoothness of the moduli space $\mathcal{A}_{g,1,n}$ follows quickly from this theory.
(3) The Serre-Tate theorem: deforming an abelian variety is the same as deforming its p-divisible group. See Theorem 2.7 for a precise statement. A consequence is that the deformation space of a polarized abelian variety admits an natural action by a large p-adic group, see 2.14. In general this action is poorly understood.
(4) There is one case when the action on the deformation space mentioned in (3) above is linearized and well-understood. This is the case when the abelian variety is ordinary. The theory of Serre-Tate coordinates says that the deformation space of an ordinary abelian variety has a natural structure as a formal torus. See Theorem 2.19 for the statement. In this case the action on the local moduli space mentioned in (3) above preserves the group structure and gives a linear representation on the character group of the Serre-Tate formal torus. This phenomenon has important consequences later. A local rigidity result, Theorem 2.26, is important for the Hecke orbit problem in that it provides an effective linearization of the Hecke orbit problem. Also, computing the deformation using the Serre-Tate coordinates is often easy; the reader is encouraged to try Exercise 2.25.

Here is a list of recommended references.

p-divisible groups: [**49**], [**38**].
Crystalline deformation theory: [**49**], [**5**].
Serre-Tate Theorem: [**49**], [**41**].
Serre-Tate coordinates: [**42**].

2.1. Deformations of abelian varieties and of p-divisible groups.

Definition. Let K be a perfect field of characteristic p. Denote by $W(K)$ the ring of p-adic Witt vectors with coordinates in K.
(i) Denote by $\mathrm{Art}_{W(K)}$ the category of Artinian local algebras over $W(K)$. An object of $\mathrm{Art}_{W(K)}$ is a pair (R, j), where R is an Artinian local algebra and $j : W(K) \to R$ is a local homomorphism of local rings. A morphism in $\mathrm{Art}_{W(K)}$ from

(R_1, j_1) to (R_2, j_2) is a homomorphism $h : R_1 \to R_2$ between Artinian local rings such that $h \circ j_1 = j_2$.
(ii) Denote by Art_K the category of Artinian local K-algebras. An object in Art_K is a pair (R, j), where R is an Artinian local algebra and $j : K \to R$ is a ring homomorphism. A morphism in Art/K from (R_1, j_1) to (R_2, j_2) is a homomorphism $h : R_1 \to R_2$ between Artinian local rings such that $h \circ j_1 = j_2$. Notice that Art_K is a fully faithful subcategory of $\mathrm{Art}_{W(K)}$.

Definition. Denote by Sets the category whose objects are sets and whose morphisms are maps between sets.

Definition. Let A_0 be an abelian variety over a perfect field $K \supset \mathbb{F}_p$. The deformation functor of A_0 is a functor

$$\mathrm{Def}(A_0/W(K)) : \mathrm{Art}_{W(K)} \to \mathrm{Sets}$$

defined as follows. For every object (R, j) of $\mathrm{Art}_{W(K)}$, $\mathrm{Def}(A_0/W(K))(R, j)$ is the set of isomorphism classes of pairs $(A \to \mathrm{Spec}(R), \epsilon)$, where $A \to \mathrm{Spec}(R)$ is an abelian scheme, and

$$\epsilon : A \times_{\mathrm{Spec}(R)} \mathrm{Spec}(R/\mathfrak{m}_R) \xrightarrow{\sim} A_0 \times_{\mathrm{Spec}(K)} \mathrm{Spec}(R/\mathfrak{m}_R)$$

is an isomorphism of abelian varieties over $R/\mathfrak{m}_R$. Denote by $\mathrm{Def}(A_0/K)$ the restriction of the deformation functor $\mathrm{Def}(A_0/W(K))$ to the faithful subcategory Art_K of $\mathrm{Art}_{W(K)}$.

Definition. Let A_0 be an abelian variety over a perfect field $K \supset \mathbb{F}_p$, and let λ_0 be a polarization on A_0. The deformation functor of (A_0, λ_0) is a functor

$$\mathrm{Def}(A_0/W(K)) : \mathrm{Art}_{W(K)} \to \mathrm{Sets}$$

defined as follows. For every object (R, ϵ) of $\mathrm{Art}_{W(K)}$, $\mathrm{Def}(A_0/W(K))(R, \epsilon)$ is the set of isomorphism classes of pairs $(A, \lambda) \to \mathrm{Spec}(R), \epsilon)$, where $(A, \lambda) \to \mathrm{Spec}(R)$ is a polarized abelian scheme, and

$$\epsilon : (A, \lambda) \times_{\mathrm{Spec}(R)} \mathrm{Spec}(R/\mathfrak{m}_R) \xrightarrow{\sim} (A_0, \lambda_0) \times_{\mathrm{Spec}(K)} \mathrm{Spec}(R/\mathfrak{m}_R)$$

is an isomorphism of polarized abelian varieties over $R/\mathfrak{m}_R$. Let $\mathrm{Def}((A_0, \lambda_0)/K)$ denote the restriction of $\mathrm{Def}(A_0/W(K))$ to the faithful subcategory Art_K of $\mathrm{Art}_{W(K)}$.

Exercise. Let X_0 be a p-divisible group over a perfect field $K \supset \mathbb{F}_p$, and let $\lambda_0 : X_0 \to X_0^t$ be a polarization of X_0. Define the deformation functor $\mathrm{Def}(X_0/W(K))$ for X_0 and the deformation functor $\mathrm{Def}((X_0, \lambda_0)/W(K))$ imitating the above definitions for abelian varieties.

Definition 2.2. Let R be a commutative ring, and let $I \subset R$ be an ideal of I. A *divided power structure* (a DP structure for short) on I is a collection of maps $\gamma_i : I \to R$, $i \in \mathbb{N}$, such that

- $\gamma_0(x) = 1 \quad \forall x \in I$,
- $\gamma_1(x) = x \quad \forall x \in I$,
- $\gamma_i(x) \in I \quad \forall x \in I, \forall i \geq 1$,
- $\gamma_j(x + y) = \sum_{0 \leq i \leq j} \gamma_i(x)\gamma_{j-i}(y) \quad \forall x, y \in I, \forall j \geq 0$,
- $\gamma_i(ax) = a^i \quad \forall a \in R, \forall x \in I, \forall i \geq 1$,
- $\gamma_i(x)\gamma_j(y) = \frac{(i+j)!}{i!j!}\gamma_{i+j}(x) \quad \forall i, j \geq 0, \forall x \in I$,

- $\gamma_i(\gamma_j(x)) = \frac{(ij)!}{i!(j!)^i}\gamma_{ij}(x) \quad \forall i, j \geq 1, \forall x \in I.$

A divided power structure $(R, I, (\gamma_i)_{i\in\mathbb{N}})$ as above is *locally nilpotent* if there exist $n_0 \in \mathbb{N}$ such that $\gamma_i(x) = 0$ for all $i \geq n_0$ and all $x \in I$. A *locally nilpotent* DP *extension* of a commutative ring R_0 is a locally nilpotent DP structure $(R, I, (\gamma_i)_{i\in\mathbb{N}})$ together with an isomorphism $R/I \xrightarrow{\sim} R_0$.

Remark 2.3.

(i) The basic idea for a divided power structure is that $\gamma_i(x)$ should "behave like" $x^i/i!$ and make sense even though dividing by $i!$ is illegitimate. The reader can easily verify that $\gamma_i(x) = x^i/i!$ is the unique divided power structure on (R, I) when $R \supseteq \mathbb{Q}$.

(ii) Given a divided power structure on (R, I) as above, one can define an exponential homomorphism $\exp : I \to 1 + I \subset R^\times$ by

$$\exp(x) = 1 + \sum_{n\geq 1} \gamma_n(x),$$

and a logarithmic homomorphism $\log : (1 + I) \to I$ by

$$\log(1 + x) = \sum_{n\geq 1} (n - 1)!\,(-1)^{n-1}\,\gamma_n(x).$$

These homomorphisms establish an isomorphism $(1 + I) \xrightarrow{\sim} I$.

(iii) Let R be a commutative ring with 1, and let I be an ideal of R such that $I^2 = (0)$. Define a DP structure on I by requiring that $\gamma_i(x) = 0$ for all $i \geq 2$ and all $x \in I$. This DP structure on a square-zero ideal I will be called the *trivial DP structure* on I.

(iv) The notion of a divided power structure was first introduced in the context of cohomology of Eilenberg-Mac Lane spaces. Grothendieck realized that one can use the divided power structure to define the crystalline first Chern class of line bundles, by analogy with the classical definition in characteristic 0, thanks to the isomorphism $(1 + I) \xrightarrow{\sim} I$ provided by a divided power structure. This observation motivated the definition of the crystalline site based on the notion of divided power structure.

(v) Theorem 2.4 below reduces deformation theory for abelian varieties and p-divisible groups to linear algebra, provided the augmentation ideal I has a divided power structure. An extension of a ring R_0 by a square-zero ideal I constitutes a standard "input data" in deformation theory; on (R, I) we have the trivial divided power structure. So we can feed such input data into the crystalline deformation theory summarized in Theorem 2.4 below to translate the deformation of an abelian scheme $A \to \mathrm{Spec}(R_0)$ over a square-zero extension $R \twoheadrightarrow R_0$ into a question about lifting Hodge filtrations, which is a question in linear algebra.

The statement of the black-boxed Theorem 2.4 below is a bit long. Roughly it says that attached to any DP-extension $(R, I, (\gamma_i)$ of the base of an abelian scheme (or a p-divisible group) over R/I, we can attach a (covariant) Dieudonné crystal, which is canonically isomorphic to the first de Rham homology of any lifting over R of the abelian scheme, if such a lifting exists. Moreover, lifting the abelian scheme to R is equivalent to lifting the Hodge filtration to the Dieudonné crystal. Notice that the first de Rham homology of abelian varieties over base schemes in

characteristic 0 enjoys similar properties, through the Gauss-Manin connection and the Kodaira-Spencer map.

Theorem 2.4. **BB** (Grothendieck-Messing) *Let $X_0 \to \mathrm{Spec}(R_0)$ be an p-divisible group over a commutative ring R_0.*

(i) *To every locally nilpotent* DP *extension $(R, I, (\gamma_i)_{i\in\mathbb{N}})$ of R_0 there is a functorially attached locally free R-module $\mathbb{D}(X_0)_R = \mathbb{D}(X0)_{(R,I,(\gamma_i))}$ of rank $\mathrm{ht}(X_0)$. The functor $\mathbb{D}(X_0)$ is called the covariant Dieudonné crystal attached to X_0.*

(ii) *Let $(R, I, (\gamma_i)_{i\in\mathbb{N}})$ be a locally nilpotent* DP *extension of R_0. Suppose that $X \to \mathrm{Spec}(R)$ is a p-divisible group extending $X_0 \to \mathrm{Spec}(R_0)$. Then there is a functorial short exact sequence*

$$0 \to \mathrm{Lie}(X^t/R)^\vee \to \mathbb{D}(X_0)_R \to \mathrm{Lie}(X/R) \to 0\,.$$

Here $\mathrm{Lie}(X/R)$ is the tangent space of the p-divisible group $X \to \mathrm{Spec}(R)$, which is a projective R-module of rank $\dim(X/R)$, and $\mathrm{Lie}(X^t/R)^\vee$ is the R-dual of the tangent space of the Serre dual $X^t \to \mathrm{Spec}(R)$ of $X \to \mathrm{Spec}(R)$.

(iii) *Let $(R, I, (\gamma_i)_{i\in\mathbb{N}})$ be a locally nilpotent* DP *extension of R_0. Suppose that $A \to \mathrm{Spec}(R)$ is an abelian scheme such that there exists an isomorphism*

$$\beta : A[p^\infty] \times_{\mathrm{Spec}(R)} \mathrm{Spec}(R_0) \xrightarrow{\sim} X_0$$

of p-divisible groups over R_0. Then there exists a natural isomorphism

$$\mathbb{D}(X_0)_R \to \mathrm{H}_1^{\mathrm{DR}}(A/R)\,,$$

where $\mathrm{H}_1^{\mathrm{DR}}(A/R)$ is the first de Rham homology of $A \to \mathrm{Spec}(R)$. Moreover the above isomorphism identifies the short exact sequence

$$0 \to \mathrm{Lie}(A[p^\infty]^t/R)^\vee \to \mathbb{D}(X_0)_R \to \mathrm{Lie}(A[p^\infty]/R) \to 0$$

described in (ii) *with the Hodge filtration*

$$0 \to \mathrm{Lie}(A^t/R)^\vee \to \mathrm{H}_1^{\mathrm{DR}}(A/R) \to \mathrm{Lie}(A/R) \to 0$$

on $\mathrm{H}_1^{\mathrm{DR}}(A/R)$.

(iv) *Let $(R, I, (\gamma_i)_{i\in\mathbb{N}})$ be a locally nilpotent* DP *extension of R_0. Denote by $\mathfrak{E}$ the category whose objects are short exact sequences*

$$0 \to F \to \mathbb{D}(X_0)_R \to Q \to 0$$

such that F and Q are projective R-modules, plus an isomorphism from the short exact sequence

$$(0 \to F \to \mathbb{D}(X_0)_R \to Q \to 0) \otimes_R R_0$$

of projective R_0-modules to the short exact sequence

$$0 \to \mathrm{Lie}(X_0^t)^\vee \to \mathbb{D}(X_0)_{R_0} \to \mathrm{Lie}(X) \to 0$$

attached to the p-divisible group $X_0 \to \mathrm{Spec}(R_0)$ as a special case of (ii) *above. The morphisms in $\mathfrak{E}$ are maps between diagrams. Then the functor from the category of p-divisible groups over R lifting X_0 to the category $\mathfrak{E}$ described in* (ii) *is an equivalence of categories.*

Theorem 2.4 is a summary of the main results in Chapter IV of [**49**].

Corollary 2.5. *Let X_0 be a p-divisible group over a perfect field $K \supset \mathbb{F}_p$. Let $d = \dim(X_0)$, $c = \dim(X_0^t)$. The deformation functor $\mathrm{Def}(X_0/W(K))$ of X_0 is representable by a smooth formal scheme over $W(K)$ of dimension cd. In other words, $\mathrm{Def}(X_0/W(K))$ is non-canonically isomorphic to the functor represented by the formal spectrum $\mathrm{Spf}\,(W(K)[[x_1, \ldots, x_{cd}]])$.*

PROOF. Recall that formal smoothness of $\mathrm{Def}(X_0/W(K))$ means that the natural map

$$\mathrm{Def}(X_0/W(K))(R) \to \mathrm{Def}(X_0/W(K))(R')$$

attached to any surjective ring homomorphism $R \to R'$ is surjective, for any Artinian local rings $R, R' \in \mathrm{Art}_{W(K)}$. To prove this, we may and do assume that the kernel I of the surjective homomorphism $R \to R'$ satisfies $I^2 = (0)$. Apply Theorem 2.4 to the trivial DP structure on pairs (R, I) with $I^2 = (0)$, we see that $\mathrm{Def}(X_0/W(K))$ is formally smooth over $W(K)$. Applying Theorem 2.4 again to the pair $K[t]/(t^2), tK[t]/(t^2)$, we see that the dimension of the tangent space of $\mathrm{Def}(X_0/K)$ is equal to cd. □

2.6. We set up notation for the Serre-Tate Theorem 2.7, which says that deforming an abelian variety A_0 over a field of characteristic p is the same as deforming the p-divisible group $A_0[p^\infty]$ attached to A_0. Recall that $A_0[p^\infty]$ is the inductive system formed by the p^n-torsion subgroup schemes $A_0[p^n]$ of A_0, where n runs through positive integers. In view of Theorem 2.4 one can regard the p-divisible group $A_0[p^\infty]$ as a refinement of the first homology group of A_0.

Let p be a prime number. Let S be a scheme such that p is locally nilpotent in $\mathcal{O}_S$. Let $I \subset \mathcal{O}_S$ be a coherent sheaf of ideals such that I is locally nilpotent. Let $S_0 = \mathrm{Spec}(\mathcal{O}_S/I)$. Denote by AV_S the category of abelian schemes over S. Denote by $\mathrm{AVBT}_{S_0,S}$ the category whose objects are triples $(A_0 \to S_0, X \to S, \epsilon)$, where $A_0 \to S_0$ is an abelian scheme over S_0, $X \to S$ is a p-divisible group over S, and $\epsilon : X \times_S S_0 \to A_0[p^\infty]$ is an isomorphism of p-divisible groups. A morphism from $(A_0 \to S_0, X \to S, \epsilon)$ to $(A_0' \to S_0, X' \to S, \epsilon')$ is a pair (h, f), where $h_0 : A_0 \to A_0'$ is a homomorphism of abelian schemes over S_0, and $f : X \to X'$ is a homomorphism of p-divisible groups over S, such that $h[p^\infty] \circ \epsilon = \epsilon' \circ (f \times_S S_0)$. Let

$$\mathbb{G}_{S_0,S} : \mathrm{AV}_S \to \mathrm{AVBT}_{S_0,S}$$

be the functor which sends an abelian scheme $A \to S$ to the triple

$$((A \times_S S_0, A[p^\infty], \mathrm{can}),$$

where can is the canonical isomorphism $A[p^\infty] \times_S S_0 \xrightarrow{\sim} (A \times_S S_0)[p^\infty]$.

Theorem 2.7. **BB** (Serre-Tate) *Notation and assumptions as in the above paragraph. The functor $\mathbb{G}_{S_0,S}$ is an equivalence of categories.*

Remark. See [**46**]. A proof of Theorem 2.7 first appeared in print in [**49**]. See also [**41**].

Corollary 2.8. *Let A_0 be an abelian variety over a perfect field K. Let*

$$\mathbb{G} : \mathrm{Def}(A_0/W(K)) \to \mathrm{Def}(A_0[p^\infty]/W(K))$$

be the functor which sends any object

$$\Big(A \to \mathrm{Spec}(R), \epsilon : A \times_{\mathrm{Spec}(R)} \mathrm{Spec}(R/\mathfrak{m}_R) \xrightarrow{\sim} A_0 \times_{\mathrm{Spec}(k)} \mathrm{Spec}(R/\mathfrak{m}_R)\Big)$$

in $\mathrm{Def}(A_0/W(K))$ *to the object*

$$(A[p^\infty] \to \mathrm{Spec}(R), \epsilon[p^\infty])$$
$$\epsilon[p^\infty] : A[p^\infty] \times_{\mathrm{Spec}(R)} \mathrm{Spec}(R/\mathfrak{m}_R) \xrightarrow{\sim} A_0[p^\infty] \times_{\mathrm{Spec}(K)} \mathrm{Spec}(R/\mathfrak{m}_R)$$

in $\mathrm{Def}(A_0[p^\infty]/W(K))$. *The functor* $\mathbb{G}$ *is an equivalence of categories.*

Remark. In words, Corollary 2.8 says that deforming an abelian variety is the same as deforming its p-divisible group.

Corollary 2.9. *Let* A_0 *be a* g*-dimensional abelian variety over a perfect field* $K \supset \mathbb{F}_p$. *The deformation functor* $\mathrm{Def}(A_0/W(K))$ *of* A_0 *is representable by a smooth formal scheme over* $W(K)$ *of relative dimension* g^2.

Proof. We have $\mathrm{Def}(A_0/W(K)) \cong \mathrm{Def}(A_0[p^\infty]/W(K))$ by Theorem 2.7. Corollary 2.9 follows from Corollary 2.5. □

2.10. Let R_0 be a commutative ring. Let $A_0 \to \mathrm{Spec}(R_0)$ be an abelian scheme. Let $\mathbb{D}(A_0) := \mathbb{D}(A_0[p^\infty])$ be the covariant Dieudonné crystal attached to A_0. Let $\mathbb{D}(A_0^t)$ be the covariant Dieudonné crystal attached to the dual abelian scheme A^t. Let $\mathbb{D}(A_0)^\vee$ be the dual of $\mathbb{D}(A_0)$, i.e.,

$$\mathbb{D}(A_0)^\vee_{(R,I,(\gamma_i))} = \mathrm{Hom}_R(\mathbb{D}(A_0)_R, R)$$

for any locally nilpotent DP extension $(R, I, (\gamma_i)_{i\in\mathbb{N}})$ of $R_0 = R/I$.

Theorem 2.11. **BB** *We have functorial isomorphisms*

$$\varphi_{A_0} : \mathbb{D}(A_0)^\vee \xrightarrow{\sim} \mathbb{D}(A_0^t)$$

for abelian varieties A_0 *over* K *with the following properties.*

(1) *The composition*

$$\mathbb{D}(A_0^t)^\vee \xrightarrow[\sim]{\varphi^\vee_{A_0}} (\mathbb{D}(A_0)^\vee)^\vee = \mathbb{D}(A_0) \xrightarrow[\sim]{j_{A_0}} \mathbb{D}((A_0^t)^t)$$

is equal to

$$-\varphi_{A_0^t} : \mathbb{D}(A_0^t)^\vee \xrightarrow{\sim} \mathbb{D}((A_0^t)^t)\,,$$

where the isomorphism $\mathbb{D}_{A_0} \xrightarrow[\sim]{j_{A_0}} \mathbb{D}((A_0^t)^t)$ *is induced by the canonical isomorphism*

$$A_0 \xrightarrow{\sim} (A_0^t)^t\,.$$

(2) *For any locally nilpotent DP extension* $(R, I, (\gamma_i)_{i\in\mathbb{N}})$ *of* $R_0 = R/I$ *and any lifting* $A \to \mathrm{Spec}(R)$ *of* $A_0 \to \mathrm{Spec}(R_0)$ *to* R, *the following diagram*

$$\begin{array}{ccccccccc}
0 & \longrightarrow & \mathrm{Lie}(A/R)^\vee & \longrightarrow & \mathbb{D}(A_0)^\vee_R & \longrightarrow & (\mathrm{Lie}(A^t/R)^\vee)^\vee & \longrightarrow & 0 \\
 & & \downarrow \cong & & \downarrow \varphi_{A_0} & & \downarrow = & & \\
0 & \longrightarrow & \mathrm{Lie}((A^t)^t/R)^\vee & \longrightarrow & \mathbb{D}(A_0^t)_R & \longrightarrow & \mathrm{Lie}(A^t/R) & \longrightarrow & 0
\end{array}$$

commutes. Here the bottom horizontal exact sequence is as in 2.4, *the top horizontal sequence is the dual of the short exact sequence in* 2.4, *and the left vertical isomorphism is induced by the canonical isomorphism* $A \xrightarrow{\sim} (A^t)^t$.

Theorem 2.11 is proved in [**5**], Chapter 5, §1.

Below are three applications of Theorem 2.4, Theorem 2.7 and Theorem 2.11; their proofs are left as exercises. The first two, Corollary 2.12 and Corollary 2.13, are basic properties of the moduli space of polarized abelian varieties. The group action in Corollary 2.14 is called the action of the local stabilizer subgroup. This "local symmetry" of the local moduli space will play an important role later in the proof of the density of ordinary Hecke orbits.

Corollary 2.12. *Let (A_0, λ_0) be a g-dimensional principally polarized abelian variety over a perfect field $K \supset \mathbb{F}_p$. The deformation functor* $\mathrm{Def}((A_0, \lambda)/W(K))$ *of A_0 is representable by a smooth formal scheme over $W(K)$ of dimensional $g(g+1)/2$.*

Remark. Corollary 2.12 can be reformulated as follows. Let η_0 be a K-rational symplectic level-n structure on A_0, $n \geq 3$, $(n, p) = 1$, and let $x_0 = [(A_0, \lambda_0, \eta_0)] \in \mathcal{A}_{g,1,n}(K)$. The formal completion $\mathcal{A}^{/x_0}_{g,1,n}$ of the moduli space $\mathcal{A}^{/x_0}_{g,1,n} \to \mathrm{Spec}(W(K))$ is non-canonically isomorphic to $\mathrm{Spf}\left(W(K)[[x_1, \ldots, x_{g(g+1)/2}]]\right)$.

Corollary 2.13. *Let (A_0, λ_0) be a polarized abelian variety over a perfect field $K \supset \mathbb{F}$; let* $\deg(\lambda_0) = d^2$.

(i) *The natural map* $\mathrm{Def}((A_0, \lambda_0)/W(K)) \to \mathrm{Def}(A_0/W(K))$ *is represented by a closed embedding of formal schemes.*

(ii) *Let n be a positive integer, $n \geq 3$, $(n, pd) = 1$. Let η_0 be a K-rational symplectic level-n structure on (A_0, λ_0). Let $x_0 = [(A_0, \lambda_0, \eta_0)] \in \mathcal{A}_{g,d,n}(K)$. The formal completion $\mathcal{A}^{/x_0}_{g,d,n}$ of the moduli space $\mathcal{A}_{g,d,n} \to \mathrm{Spec}(W(K))$ at the closed point x_0 is isomorphic to the local deformation space*

$$\mathrm{Def}((A_0, \lambda_0)/W(K)).$$

Corollary 2.14. (i) *Let A_0 be an variety over a perfect field $K \supset \mathbb{F}$. There is a natural action of the profinite group* $\mathrm{Aut}(A_0[p^\infty])$ *on the smooth formal scheme* $\mathrm{Def}(A_0/W(K))$.

(ii) *Let λ_0 be a principal polarization on an abelian variety A_0 over a perfect field K. Denote by* $\mathrm{Aut}((A_0, \lambda_0)[p^\infty])$ *the closed subgroup of* $\mathrm{Aut}(A_0[p^\infty])$ *consisting of all automorphisms of* $\mathrm{Aut}(A[p^\infty])$ *compatible with the quasi-polarization $\lambda_0[p^\infty]$. The natural action in* (i) *above induces a natural action on the closed formal subscheme* $\mathrm{Def}(A_0, \lambda_0)$ *of* $\mathrm{Def}(A_0)$.

Remark. In the situation of (ii) above, the group $\mathrm{Aut}(A_0, \lambda_0)$ of polarization-preserving automorphisms of A_0 is finite, while $\mathrm{Aut}((A_0, \lambda_0)[p^\infty])$ is a compact p-adic Lie group of positive dimension of positive dimension if $\dim(A_0) > 0$. The group $\mathrm{Aut}(A_0, \lambda)$ (resp. $\mathrm{Aut}((A_0, \lambda_0)[p^\infty])$) operates on $\mathrm{Def}(A_0, \lambda_0)$ (resp. $\mathrm{Def}((A_0, \lambda_0)[p^\infty])$) by "changing the marking". By Theorem 2.7, we have a natural isomorphism $\mathrm{Def}(A_0, \lambda_0) \xrightarrow{\sim} \mathrm{Def}((A_0, \lambda_0)[p^\infty])$, which is equivariant for the inclusion homomorphism $\mathrm{Aut}(A_0, \lambda_0) \hookrightarrow \mathrm{Aut}((A_0, \lambda_0)[p^\infty])$. In other words, the action of $\mathrm{Aut}(A_0, \lambda)$ on $\mathrm{Def}(A_0, \lambda_0)$ extends to an action by $\mathrm{Aut}((A_0, \lambda_0)[p^\infty])$.

2.15. Étale and toric p-divisible groups: notation. A p-divisible group X over a base scheme S is said to be étale (resp. toric) if and only if $X[p^n]$ is étale (resp. of multiplicative type) for every $n \geq 1$; see the end of 10.6.

Remark. Let $E \to S$ be an étale p-divisible group, where S is a scheme. The p-adic Tate module of E, defined by

$$\mathrm{T}_p(E) := \varprojlim_n E[p^n]\,,$$

is representable by a smooth $\mathbb{Z}_p$-sheaf on $S_{\text{ét}}$ whose rank is equal to $\mathrm{ht}(E/S)$. Here the rank of E is a locally constant function on the base scheme S. When S is the spectrum of a field K, $\mathrm{T}_p(E)$ "is" a free $\mathbb{Z}_p$-module with an action by $\mathrm{Gal}(K^{\mathrm{sep}}/K)$; see 10.5.

Remark. Attached to any toric p-divisible group $T \to S$ is its character group

$$X^*(T) := \underline{\mathrm{Hom}}(T, \mathbb{G}_m[p^\infty])$$

and cocharacter group

$$X_*(T) := \underline{\mathrm{Hom}}(\mathbb{G}_m[p^\infty], T)\,.$$

The character group of T can be identified with the p-adic Tate module of the Serre-dual T^t of T, and T^t is an étale p-divisible group over S. Both $X^*(T)$ and $X_*(T)$ are smooth $\mathbb{Z}_p$-sheaves of rank $\dim(T/S)$ on $S_{\text{ét}}$, and they are naturally dual to each other.

Definition 2.16. Let S be either a scheme such that p is locally nilpotent in $\mathcal{O}_S$, or an adic formal scheme such that p is locally topologically nilpotent in $\mathcal{O}_S$. A p-divisible group $X \to S$ is *ordinary* if X sits in the middle of a short exact sequence

$$0 \to T \to X \to E \to 0$$

where T (resp. E) is a multiplicative (resp. étale) p-divisible group. Such an exact sequence is unique up to unique isomorphism.

Remark. Suppose that X is an ordinary p-divisible group over $S = \mathrm{Spec}(K)$, where K is a perfect field $K \supset \mathbb{F}_p$. Then there exists a unique splitting of the short exact sequence $0 \to T \to X \to E \to 0$ over K.

Proposition 2.17. **BB** *Suppose that S is a scheme over $W(K)$ and p is locally nilpotent in $\mathcal{O}_S$. Let $S_0 = \underline{\mathrm{Spec}}(\mathcal{O}_S/p\mathcal{O}_S)$, the closed subscheme of S defined by the ideal $p\mathcal{O}_S$ of the structure sheaf $\mathcal{O}_S$. If $X \to S$ is a p-divisible group such that $X \times_S S_0$ is ordinary, then $X \to S$ is ordinary.*

Proposition 2.17 is a consequence of the rigidity of finite étale group schemes and commutative finite group schemes of multiplicative type. See SGA3, Exposé X.

2.18. We set up notation for Theorem 2.19 on the theory of Serre-Tate local coordinates. Let $K \supset \mathbb{F}_p$ be a perfect field and let X_0 be an ordinary p-divisible group over K. This means that there is a natural split short exact sequence

$$0 \to T_0 \to X_0 \to E_0 \to 0$$

where T_0 (resp. E_0) is a multiplicative (resp. étale) p-divisible group over K. Let

$$T_i \to \mathrm{Spec}(W(K)/p^iW(K)) \quad (\text{resp. } E_i \to \mathrm{Spec}(W(K)/p^iW(K))\,)$$

be the multiplicative (resp. étale) p-divisible group over $\mathrm{Spec}(W(K)/p^iW(K))$ which lifts T_0 (resp. E_0) for each $i \geq 1$. Both T_i and E_i are unique up to unique isomorphism. Taking the limit of $T_i[p^n]$ (resp. $E_i[p^i]$) as $i \to \infty$, we get a multiplicative (resp. étale) BT_n-group $T^\sim \to \mathrm{Spec}(W(K))$ (resp. $E^\sim \to \mathrm{Spec}(W(K))$) over $W(K)$.

Denote by $T^\wedge$ the formal torus over $W(K)$ attached to T_0. More explicitly, it is the scheme theoretic inductive limit of $T_i[p^n]$ as i and n both go to ∞; see also 10.9 (4) and 10.21. Another equivalent description is that $T^\wedge = X_*(T_0) \otimes_{\mathbb{Z}_p} \mathbb{G}_m^\wedge$, where $\mathbb{G}_m^\wedge$ is the formal completion of $\mathbb{G}_m \to \mathrm{Spec}(W(K))$ along its unit section, and $X_*(T_0)$ is the étale smooth free $\mathbb{Z}_p$-sheaf of rank $\dim(T_0)$ on the étale site $\mathrm{Spec}(K)_{\text{ét}}$, which is isomorphic to the étale sites $\big(\mathrm{Spec}(W(K)/p^iW(K))\big)_{\text{et}}$ and $(\mathrm{Spec}W(K))_{\text{ét}}$ for all i because the étale topology is insensitive to nilpotent extensions.

Theorem 2.19 below says that the deformation space of an ordinary p-divisible group X_0 as above has a natural structure as a formal torus over $W(K)$, whose dimension is equal to the product of the heights of the étale part E_0 and the multiplicative part T_0.

Theorem 2.19. *Notation and assumption as above.*

(i) *Every deformation $X \to \mathrm{Spec}(R)$ of X_0 over an Artinian local $W(K)$-algebra R is an ordinary p-divisible group over R. Therefore X sits in the middle of a short exact sequence*

$$0 \to T^\sim \times_{\mathrm{Spec}(W(K))} \mathrm{Spec}(R) \to X \to E^\sim \times_{\mathrm{Spec}(W(K))} \mathrm{Spec}(R) \to 0\,.$$

(ii) *The deformation functor $\mathrm{Def}(X_0/W(K)$ has a natural structure, via the Baer sum construction, as a functor from $\mathrm{Art}_{W(K)}$ to the category AbG of abelian groups. In particular the unit element in $\mathrm{Def}(X_0/W(K)(R)$ corresponds to the p-divisible group*

$$\big(T^\sim \times_{\mathrm{Spec}(W(K))} E^\sim\big) \times_{\mathrm{Spec}(W(K))} \mathrm{Spec}(R)$$

over R.

(iii) *There is a natural isomorphism of functors*

$$\begin{aligned}\mathrm{Def}(X_0/W(K)) \xleftarrow{\sim} \underline{\mathrm{Hom}}_{\mathbb{Z}_p}(\mathrm{T}_p(E_0), T^\wedge) &= \mathrm{T}_p(E_0)^\vee \otimes_{\mathbb{Z}_p} X_*(T_0) \otimes_{\mathbb{Z}_p} \mathbb{G}_m^\wedge \\ &= \underline{\mathrm{Hom}}_{\mathbb{Z}_p}\big(\mathrm{T}_p(E_0) \otimes_{\mathbb{Z}_p} X^*(T_0), \mathbb{G}_m^\wedge\big)\,.\end{aligned}$$

In other words, the deformation space $\mathrm{Def}(X_0/W(K))$ of X_0 has a natural structure as a formal torus over $W(K)$ whose cocharacter group is isomorphic to the $\mathrm{Gal}(K^{\mathrm{alg}}/K)$-module $\mathrm{T}_p(E)^\vee \otimes_{\mathbb{Z}_p} X_*(T_0)$.

Proof. The statement (i) is follows from Proposition 2.17, so is (ii). It remains to prove (iii).

By étale descent, we may and do assume that K is algebraically closed. By (i), over any Artinian local $W(K)$-algebra R, we see that $\mathrm{Def}(X_0/W(K))(R)$ is the set of isomorphism classes of extensions of $E^\sim \times_{W(K)} \mathrm{Spec}(R)$ by $T^\sim \times_{W(K)} \mathrm{Spec}(R)$. Write T_0 (resp. E_0) as a product of a finite number of copies of $\mathbb{G}_m[p^\infty]$ (resp. $\mathbb{Q}_p/\mathbb{Z}_p$), we only need to verify the statement (iii) in the case when $T_0 = \mathbb{G}_m[p^\infty]$ and $E_0 = \mathbb{Q}_p/\mathbb{Z}_p$.

Let R be an Artinian local $W(K)$-algebra. We have seen that

$$\mathrm{Def}(\mathbb{Q}_p/\mathbb{Z}_p, \mathbb{G}_m[p^\infty])(R)$$

is naturally isomorphic to the inverse limit $\varprojlim_n \mathrm{Ext}^1_{\mathrm{Spec}(R),\mathbb{Z}/p^n\mathbb{Z}}(p^{-n}\mathbb{Z}/\mathbb{Z}, \mu_{p^n})$, where the Ext group is computed in the category of sheaves of $(\mathbb{Z}/p^n\mathbb{Z})$-modules for the flat topology on $\mathrm{Spec}(R)$. By Kummer theory, we have

$$\mathrm{Ext}^1_{\mathrm{Spec}(R),\mathbb{Z}/p^n\mathbb{Z}}(p^{-n}\mathbb{Z}/\mathbb{Z}, \mu_{p^n}) = R^\times/(R^\times)^{p^n} = (1+\mathfrak{m}_R)/(1+\mathfrak{m}_R)^{p^n}\,;$$

the second equality follows from the hypothesis that K is perfect. One checks that the map

$$\mathrm{Ext}^1_{\mathrm{Spec}(R),\mathbb{Z}/p^{n+1}\mathbb{Z}}(p^{-n-1}\mathbb{Z}/\mathbb{Z},\mu_{p^{n+1}}) \to \mathrm{Ext}^1_{\mathrm{Spec}(R),\mathbb{Z}/p^{n}\mathbb{Z}}(p^{-n}\mathbb{Z}/\mathbb{Z},\mu_{p^{n}})$$

obtained by "restriction to the subgroup of $[p^n]$-torsions" corresponds to the natural surjection

$$(1+\mathfrak{m}_R)/(1+\mathfrak{m}_R)^{p^{n+1}} \twoheadrightarrow (1+\mathfrak{m}_R)/(1+\mathfrak{m}_R)^{p^n}\,.$$

We know that $p \in \mathfrak{m}_R$ and $\mathfrak{m}_R$ is nilpotent. Hence there exists an n_0 such that $(1+\mathfrak{m}_R)^{p^n} = 1$ for all $n \geq n_0$. Taking the inverse limit as $n \to \infty$, we see that the natural map

$$1+\mathfrak{m}_R \to \varprojlim_n \mathrm{Ext}^1_{\mathrm{Spec}(R),\mathbb{Z}/p^{n}\mathbb{Z}}(p^{-n}\mathbb{Z}/\mathbb{Z},\mu_{p^{n}})$$

is an isomorphism. □

Corollary 2.20. *Let $K \supset \mathbb{F}_p$ be a perfect field, and let A_0 be an ordinary abelian variety. Let $\mathrm{T}_p(A_0) := \mathrm{T}_p(A_0[p^\infty]_{\text{ét}})$, $\mathrm{T}_p(A_0^t) := \mathrm{T}_p(A_0^t[p^\infty]_{\text{ét}})$. Then*

$$\mathrm{Def}(A_0/W(K)) \cong \underline{\mathrm{Hom}}_{\mathbb{Z}_p}(\mathrm{T}_p(A_0) \otimes_{\mathbb{Z}_p} \mathrm{T}_p(A_0^t), \mathbb{G}_m^\wedge)\,.$$

Exercise 2.21. Let R be a commutative ring with 1. Compute

$$\mathrm{Ext}^1_{\mathrm{Spec}(R),(\mathbb{Z}/n\mathbb{Z})}(n^{-1}\mathbb{Z}/\mathbb{Z},\mu_n),$$

the group of isomorphism classes of extensions of the constant group scheme $n^{-1}\mathbb{Z}/\mathbb{Z}$ by μ_n over $\mathrm{Spec}(R)$ in the category of finite flat group schemes over $\mathrm{Spec}(R)$ which are killed by n.

Notation. Let R be an Artinian local $W(k)$-algebra, where $k \supset \mathbb{F}_p$ is an algebraically closed field. Let $X \to \mathrm{Spec}(R)$ be an ordinary p-divisible group such that the closed fiber $X_0 := X \times_{\mathrm{Spec}(R)} \mathrm{Spec}(k)$ is an ordinary p-divisible group over k. Denote by $q(X/R;\cdot,\cdot)$ the $\mathbb{Z}_p$-bilinear map

$$q(X/R;\cdot,\cdot) : \mathrm{T}_p(X_{0,\text{ét}}) \times \mathrm{T}_p(X^t_{0,\text{ét}}) \to 1+\mathfrak{m}_R$$

correspond to the deformation $X \to \mathrm{Spec}(R)$ of the p-divisible group X_0 as in Corollary 2.20. Here we have used the natural isomorphism $X^*(X_{0,\mathrm{mult}}) \cong \mathrm{T}_p(X^t_{0,\text{ét}})$, so that the Serre-Tate coordinates for the p-divisible group $X \to \mathrm{Spec}(R)$ is a $\mathbb{Z}_p$-bilinear map $q(X/R;\cdot,\cdot)$ on $\mathrm{T}_p(X_{0,\text{ét}}) \times \mathrm{T}_p(X^t_{0,\text{ét}})$. The abelian group $1+\mathfrak{m}_R \subset R^\times$ is regarded as a $\mathbb{Z}_p$-module, so "$\mathbb{Z}_p$-bilinear" makes sense. Let $\mathrm{can} : X_0 \xrightarrow{\sim} (X_0^t)^t$ be the canonical isomorphism from X_0 to its double Serre dual, and let $\mathrm{can}_* : \mathrm{T}_p(X_{0,\text{ét}}) \xrightarrow{\sim} \mathrm{T}_p((X_0^t)^t_{\text{ét}})$ be the isomorphism induced by can.

The relation between the Serre-Tate coordinate $q(X/R;\cdot,\cdot)$ of a deformation of X_0 and the Serre-Tate coordinates $q(X^t/R;\cdot,\cdot)$ of the Serre dual X^t of X is given by 2.22. The proof is left as an exercise.

Lemma 2.22. *Let $X \to \mathrm{Spec}(R)$ be an ordinary p-divisible group over an Artinian local $W(k)$-algebra R. Then we have*

$$q(X;u,v_t) = q(X^t;v_t,\mathrm{can}_*(u)) \qquad \forall u \in \mathrm{T}_p(X_{0,\text{ét}}),\ \forall v \in \mathrm{T}_p(X^t_{0,\text{ét}})\,.$$

The same statement holds when the ordinary p-divisible group $X \to \mathrm{Spec}(R)$ is replaced by an ordinary abelian scheme $A \to \mathrm{Spec}(R)$.

From the functoriality of the construction in 2.19, it is not difficult to verify the following.

Proposition 2.23. *Let X_0, Y_0 be ordinary p-divisible groups over a perfect field $K \supset \mathbb{F}$. Let R be an Artinian local ring over $W(K)$. Let $X \to \mathrm{Spec}(R)$, $Y \to \mathrm{Spec}(R)$ be abelian schemes whose closed fibers are X_0 and Y_0 respectively. Let $q(X/R;\cdot,\cdot)$, $q(Y/R;\cdot,\cdot)$ be the Serre-Tate coordinates for X and Y respectively. Let $\beta : X_0 \to Y_0$ be a homomorphism of abelian varieties over k. Then β extends to a homomorphism from X to Y over $\mathrm{Spec}(R)$ if and only if*

$$q(X/R; u, \beta^t(v_t)) = q(Y/R; \beta(u), v_t) \quad \forall u \in \mathrm{T}_p(X_0),\ \forall v_t \in \mathrm{T}_p(Y_0^t)\,.$$

Corollary 2.24. *Let A_0 be an ordinary abelian variety over a perfect field $K \supset \mathbb{F}_p$. Let $\lambda_0 : A_0 \to A_0^t$ be a polarization on A_0. Then*

$$\mathrm{Def}((A_0, \lambda_0)/W(K)) \cong \underline{\mathrm{Hom}}_{\mathbb{Z}_p}(S, \mathbb{G}_m{}^{\wedge})\,,$$

where S is defined as

$$\mathrm{T}_p(A_0[p^\infty]_{\text{ét}}) \otimes_{\mathbb{Z}_p} \mathrm{T}_p(A_0^t[p^\infty]_{\text{ét}}) \Big/ \big(u \otimes \mathrm{T}_p(\lambda_0)(v) - v \otimes \mathrm{T}_p(\lambda_0)(u)\big)_{u,v \in \mathrm{T}_p(A[p^\infty]_{\text{ét}})}\,.$$

Exercise 2.25. Notation as in 2.24. Let $p^{e_1}, \dots, p^{e_g}$ be the elementary divisors of the $\mathbb{Z}_p$-linear map $\mathrm{T}_p(\lambda_0) : \mathrm{T}_p(A_0[p^\infty]_{\text{ét}}) \to \mathrm{T}_p(A_0^t[p^\infty]_{\text{ét}})$, $g = \dim(A_0)$, $e_1 \le e_2 \le \cdots \le e_g$. The torsion submodule S_{torsion} of S is isomorphic to $\bigoplus_{1 \le i < j \le g} (\mathbb{Z}_p / p^{e_i}\mathbb{Z}_p)$.

Theorem 2.26. **BB** (local rigidity) *Let $k \supset \mathbb{F}_p$ be an algebraically closed field. Let*

$$T \cong ((\mathbb{G}_m^{\wedge})^n = \mathrm{Spf}\, k[[u_1, \dots, u_n]]$$

be a formal torus, with group law given by

$$u_i \ \mapsto\ u_i \otimes 1 + 1 \otimes u_i + u_i \otimes u_i \qquad i = 1, \dots n\,.$$

Let $X = \mathrm{Hom}_k(\mathbb{G}_m^{\wedge}, T) \cong \mathbb{Z}_p^n$ be the cocharacter group of T; notice that $\mathrm{GL}(X)$ operates naturally on T. Let $G \subset \mathrm{GL}(X \otimes_{\mathbb{Z}_p} \mathbb{Q}_p) \cong \mathrm{GL}_n$ be a reductive linear algebraic subgroup over $\mathbb{Q}_p$. Let Z be an irreducible closed formal subscheme of T which is stable under the action of an open subgroup U of $G(\mathbb{Q}_p) \cap \mathrm{GL}(X)$. Then Z is a formal subtorus of T.

See Theorem 6.6 of [**7**] for a proof of 2.26; see also [**12**].

Corollary 2.27. *Let $x_0 = [(A_0, \lambda_0, \eta_0)] \in \mathcal{A}_{g,1,n}(\mathbb{F})$ be an $\mathbb{F}$-point of $\mathcal{A}_{g,1,n}$, where $\mathbb{F}$ is the algebraic closure of $\mathbb{F}_p$. Assume that the abelian variety A_0 is ordinary. Let $Z(x_0)$ be the Zariski closure of the prime-to-p Hecke orbit $\mathcal{H}^{(p)}_{\mathrm{Sp}_{2g}}(x_0)$ on $\mathcal{A}_{g,1,n}$. The formal completion $Z(x_0)^{/x_0}$ of $Z(x_0)$ at x_0 is a formal subtorus of the Serre-Tate formal torus $\mathcal{A}^{/x_0}_{g,1,n}$.*

Proof. This is immediate from 2.26 and the local stabilizer principle; see 9.5 for the statement of the local stabilizer principle. ∎

Remark. Corollary 2.27 puts a serious restriction on the Zariski closure $Z(x_0)$ of the Hecke orbit of an ordinary point x_0 in $\mathcal{A}_{g,1,n}(\mathbb{F})$. In fact the argument shows that the formal completion of $Z(x_0)$ at any closed point y_0 of the smooth ordinary locus of Z_{x_0} is a formal subtorus of the Serre-Tate torus at y_0. This constitutes the linearization step toward proving that $Z(x_0) = A_{g,1,n}$. See Proposition 6.14 and Step 4 of the proof of Theorem 9.2, where Theorem 2.26 plays a crucial role.

3. The Tate-conjecture: ℓ-adic and p-adic

Most results of this section will not be used directly in our proofs. However, this is such a beautiful part of mathematics that we wish to tell more than we really need.

Basic references: [**79**] and [**37**]; [**78**], [**19**], [**62**].

3.1. Let A be an abelian variety over a field K of arbitrary characteristic. The ring $\mathrm{End}(A)$ is an algebra over $\mathbb{Z}$, which has no torsion, and which is free of finite rank as $\mathbb{Z}$-module. We write $\mathrm{End}^0(A) = \mathrm{End}(A) \otimes_{\mathbb{Z}} \mathbb{Q}$. Let $\mu : A \to A^t$ be a polarization. An endomorphism $x : A \to A$ defines $x^t : A^t \to A^t$. We define an anti-involution

$$^\dagger : \mathrm{End}^0(A) \to \mathrm{End}^0(A), \quad \text{by} \quad x^t \cdot \mu = \mu \cdot x^\dagger,$$

called the *Rosati involution*; see 10.13. In case μ is a principal polarization the Rosati involution maps $\mathrm{End}(A)$ into itself.

The Rosati involution is *positive definite* on $D := \mathrm{End}^0(A)$, meaning that $x \mapsto \mathrm{Tr}(x \cdot x^\dagger)$ is a positive definite quadratic form on $\mathrm{End}^0(A)$; for references see Proposition II in 3.10. Such algebras have been classified by Albert, see 10.14.

Definition 3.2. A field L is said to be a CM-field if L is a finite extension of $\mathbb{Q}$ (hence L is a number field), and there is a subfield $L_0 \subset L$ such that $L_0/\mathbb{Q}$ is totally real (i.e., every $\psi_0 : L_0 \to \mathbb{C}$ gives $\psi_0(L_0) \subset \mathbb{R}$) *and* L/L_0 *is quadratic totally imaginary* (i.e., $[L : L_0] = 2$ and for every $\psi : L \to \mathbb{C}$ we have $\psi(L) \not\subset \mathbb{R}$.

Equivalently, L is a CM-field if there exists an element of order 2 in the center of the Galois group $\mathrm{Gal}(M/\mathbb{Q})$ of the Galois closure M of L over $\mathbb{Q}$, which is equal to the complex conjugation for every archimedean place of M.

Remark. The quadratic extension L/L_0 gives an involution $\iota \in \mathrm{Aut}(L/L_0)$. For every embedding $\psi : L \to \mathbb{C}$ this involution corresponds with the restriction of complex conjugation on $\mathbb{C}$ to $\psi(L)$.

Even more is known about the endomorphism algebra of an abelian variety over a finite field. Tate showed that

Theorem 3.3. (Tate) *An abelian variety over a finite field admits* sufficiently many Complex Multiplications.

This is equivalent with: *Let A be a simple abelian variety over a finite field. Then there is a* CM*-field of degree* $2 \cdot \dim(A)$ *contained in* $\mathrm{End}^0(A)$.

A proof can be found in [**78**], [**79**]; also see 10.17 for a stronger statement. See 10.15 for the definition of "abelian varieties with sufficiently many Complex Multiplications". A consequence of this theorem is the following.

Let A be an abelian variety over $\mathbb{F} = \overline{\mathbb{F}_p}$. Suppose that A is simple, and hence that $\mathrm{End}^0(A)$ is a division algebra; this algebra has finite rank over $\mathbb{Q}$; the possible structures of endomorphism algebras of an abelian variety have been classified by Albert, see 10.14. In this case

- either A is a supersingular elliptic curve, and $D := \mathrm{End}^0(A) = \mathbb{Q}_{p,\infty}$, which is the (unique) quaternion algebra central over $\mathbb{Q}$, which is unramified for every finite prime $\ell \neq p$, i.e., $D \otimes \mathbb{Q}_\ell$ is the 2×2 matrix algebra over $\mathbb{Q}_\ell$, and $D/\mathbb{Q}$ is ramified at p and at ∞; here D is of Albert Type III(1);

- or A is not a supersingular elliptic curve; in this case D is of Albert Type IV(e_0, d) with $e_0 \cdot d = g := \dim(A)$.

In particular (to be used later).

Corollary 3.4. *Let A be an abelian variety over $\mathbb{F} := \overline{\mathbb{F}_p}$. There exists $E = F_1 \times \cdots \times F_r$, a product of totally real fields, and an injective homomorphism $E \hookrightarrow \mathrm{End}^0(A)$ such that $\dim_{\mathbb{Q}}(E) = \dim(A)$.*

Examples.
(1) E is a supersingular elliptic curve over $K = \mathbb{F}_q$. Then either $D := \mathrm{End}^0(E)$ is isomorphic with $\mathbb{Q}_{p,\infty}$, or D is an imaginary quadratic field over $\mathbb{Q}$ in which p is not split.
(2) E is a non-supersingular elliptic curve over $K = \mathbb{F}_q$. Then $D := \mathrm{End}^0(E)$ is an imaginary quadratic field over $\mathbb{Q}$ in which p is split.
(3) If A is simple over $K = \mathbb{F}_q$ such that $D := \mathrm{End}^0(A)$ is commutative, then $D = L = \mathrm{End}^0(A)$ is a CM-field of degree $2 \cdot \dim(A)$ over $\mathbb{Q}$.
(4) In characteristic zero the endomorphism algebra of a simple abelian variety which admits smCM is *commutative*. However in positive characteristic an Albert Type IV(e_0, d) with $e_0 > 1$ can appear. For example, see [**79**], page 67: for any prime number $p > 0$, and for any $g > 2$ there exists a simple abelian variety over $\mathbb{F}$ such that $D = \mathrm{End}^0(A)$ is a division algebra of rank g^2 over its center L, which is a quadratic imaginary field over $\mathbb{Q}$.

3.5. Weil numbers and CM-fields.

Definition. Let p be a prime number, $n \in \mathbb{Z}_{>0}$; write $q = p^n$. A Weil q-number is an algebraic integer π such that for every embedding $\psi : \mathbb{Q}(\pi) \to \mathbb{C}$ we have

$$|\psi(\pi)| \quad = \quad \sqrt{q}.$$

We say that π and π' are *conjugated* if there exists an isomorphism $\mathbb{Q}(\pi) \cong \mathbb{Q}(\pi')$ mapping π to π'.
Notation: $\pi \sim \pi'$. We write $W(q)$ for the set of Weil q-numbers and $W(q)/\sim$ for the set of conjugacy classes of Weil q-numbers.

Proposition 3.6. *Let π be a Weil q-number. Then*
(I) *either for at least one $\psi : \mathbb{Q}(\pi) \to \mathbb{C}$ we have $\pm\sqrt{q} = \psi(\pi) \in \mathbb{R}$; in this case we have:*
(Ie) *n is even, $\sqrt{q} \in \mathbb{Q}$, and $\pi = +p^{n/2}$, or $\pi = -p^{n/2}$; or*
(Io) *n is odd, $\sqrt{q} \in \mathbb{Q}(\sqrt{p})$, and $\psi(\pi) = \pm p^{n/2}$.*
In particular in case (I) we have $\psi(\pi) \in \mathbb{R}$ for every ψ.
(II) *Or for every $\psi : \mathbb{Q}(\pi) \to \mathbb{C}$ we have $\psi(\pi) \notin \mathbb{R}$ (equivalently: for at least one ψ we have $\psi(\pi) \notin \mathbb{R}$). In case (II) the field $\mathbb{Q}(\pi)$ is a CM-field.*

Proof. Exercise.

Remark 3.7. We see a characterization of Weil q-numbers. In case (I) we have $\pi = \pm\sqrt{q}$. If $\pi \notin \mathbb{R}$:

$$\beta := \pi + \frac{q}{\pi} \quad \text{is totally real,}$$

and π is a zero of

$$T^2 - \beta \cdot T + q, \quad \text{with} \quad \beta < 2\sqrt{q}.$$

In this way it is easy to construct Weil q-numbers.

3.8. Let A be an abelian variety over a finite field $K = \mathbb{F}_q$ with $q = p^n$. Let $F : A \to A^{(p)}$ be the relative Frobenius morphism for A. Iterating this Frobenius map n times, observing there is a canonical identification $A^{(p^n)} = A$, we obtain $(\pi : A \to A) \in \mathrm{End}(A)$. If A is simple, the subring $\mathbb{Q}(\pi) \subset \mathrm{End}^0(A)$ is a subfield, and we can view π as an algebraic integer.

Theorem 3.9. **Extra** (Weil) *Let $K = \mathbb{F}_q$ be a finite field, let A be a simple abelian variety over K. Then π is a Weil q-number.*

This is the famous "Weil conjecture" for an abelian variety over a finite field. See [**86**], page 70; [**87**], page 138; [**54**], Theorem 4 on page 206.

Exercise 3.10. Use Propositions I and II below to prove the following statements, thereby proving Theorem 3.9.

(i) *Suppose that A is a simple abelian variety over a field K, and let $L = \mathrm{Centre}(\mathrm{End}^0(A))$. A Rosati involution on $D := \mathrm{End}^0(A)$ induces the complex conjugation on L (for every embedding $L \hookrightarrow \mathbb{C}$).*

(ii) *If moreover K is a finite field, $\pi = \pi_A$ is a Weil q-number.*

Proposition. I. *For a simple abelian variety A over $K = \mathbb{F}_q$ we have*

$$\pi_A \cdot (\pi_A)^\dagger = q.$$

Here $\dagger : D \to D := \mathrm{End}^0(A)$ is the Rosati involution attached to a polarization of A.

One proof can be found in [**54**], formula (i) on page 206; also see [**16**], Corollary 19.2 on page 144.

Another proof of (I) can be given by duality. We have

$$\left(F_{A/S} : A \to A^{(p)}\right)^t \quad = \quad V_{A^t/S} : (A^{(p)})^t \to A^t\,,$$

where $V_{A^t/S}$ is the Verschiebung of the abelian scheme A^t/S dual to A/S; see 10.24. From this formula we see that

$$\pi_{A^t} \cdot (\pi_A)^t = (F_{A^t})^n \cdot (V_{A^t})^n = p^n = q,$$

where we use the shorthand notation F^n for the n times iterated relative Frobenius morphism, and the same for V^n. See [GM], 5.21, 7.34 and Section 15. □

Proposition. II. *For any polarized abelian variety A over a field the Rosati involution $\dagger : D \to D := \mathrm{End}^0(A)$ is a positive definite bilinear form on D, i.e., for any non-zero $x \in D$ we have $\mathrm{Tr}(x \cdot x^\dagger) > 0$.*

See [**54**], Theorem 1 on page 192, see [**16**], Theorem 17.3 on page 138.

Remark 3.11. Given $\pi = \pi_A$ of a simple abelian variety over $\mathbb{F}_q$ one can determine the structure of the division algebra $\mathrm{End}^0(A)$, see [**79**], Theorem 1. See 10.17.

Theorem 3.12. **Extra** (Honda and Tate) *By $A \mapsto \pi_A$ we obtain a bijective map*

$$\{\textit{abelian varieties simple over } \mathbb{F}_q\}/\sim_{\mathbb{F}_q} \quad \xrightarrow{\sim} \quad W(q)/\sim$$

between the set of $\mathbb{F}_q$-isogeny classes of abelian varieties simple over $\mathbb{F}_q$ and the set of conjugacy classes of Weil q-numbers.

See [**79**], Theorem 1 on page 96.

3.13. Let π be a Weil q-number. Let $\mathbb{Q} \subset L \subset D$ be the central algebra determined by π. It is known that

$$[L:\mathbb{Q}] =: e, \quad [D:L] =: d^2, \quad 2g := e{\cdot}d. \qquad \text{See 10.12.}$$

As we have seen in Proposition 3.6 there are three possibilities:

(ℝe) *Either* $\pi = \sqrt{q} \in \mathbb{Q}$, *and* $q = p^n$ *with* n *an* **even** *positive integer.*

Type III(1), $g = 1$

In this case $\pi = +p^{n/2}$, or $\pi = -p^{n/2}$. Hence $L = L_0 = \mathbb{Q}$. We see that $D/\mathbb{Q}$ has rank 4, with ramification exactly at ∞ and at p. We obtain $g = 1$, we have that $A = E$ is a supersingular elliptic curve, $\mathrm{End}^0(A)$ is of Type III(1), a definite quaternion algebra over $\mathbb{Q}$. This algebra was denoted by Deuring as $\mathbb{Q}_{p,\infty}$. Note that "all endomorphisms of E are defined over K", i.e., for any

$$\forall \quad K \subset K' \quad \text{we have} \quad \mathrm{End}(A) = \mathrm{End}(A \otimes K').$$

(ℝo) *Or* $q = p^n$ *with* n *an* **odd** *positive integer,* $\pi = \sqrt{q} \in \mathbb{R} \notin \mathbb{Q}$.

Type III(2), $g = 2$

In this case $L_0 = L = \mathbb{Q}(\sqrt{p})$, a real quadratic field. We see that D ramifies exactly at the two infinite places with invariants equal to $(n/2){\cdot}2/(2n) = 1/2$. Hence D/L_0 is a definite quaternion algebra over L_0; it is of Type III(2). We conclude $g = 2$. If $K \subset K'$ is an extension of odd degree we have $\mathrm{End}(A) = \mathrm{End}(A \otimes K')$. If $K \subset K'$ is an extension of even degree, $A \otimes K'$ is non-simple, it is K'-isogenous to a product of two supersingular elliptic curves, and $\mathrm{End}^0(A \otimes K')$ is a 2×2 matrix algebra over $\mathbb{Q}_{p,\infty}$, and

$$\forall K' \quad \text{with} \quad 2 \mid [K':K] \quad \text{we have} \quad \mathrm{End}(A) \neq \mathrm{End}(A \otimes K').$$

(ℂ) *For at least one embedding* $\psi : \mathbb{Q}(\pi) \to \mathbb{C}$ *we have* $\psi(\pi) \notin \mathbb{R}$.

Type IV(e_0, d), $g := e_0{\cdot}d$

In this case all conjugates of $\psi(\pi)$ are non-real. We can determine $[D:L]$ knowing all $v(\pi)$ by 10.17 (3); here d is the greatest common divisor of all denominators of $[L_v : \mathbb{Q}_p]{\cdot}v(\pi)/v(q)$, for all $v \mid p$. This determines $2g := e{\cdot}d$. The endomorphism algebra is of Type IV(e_0, d). For $K = \mathbb{F}_q \subset K' = \mathbb{F}_{q^m}$ we have

$$\mathrm{End}(A) = \mathrm{End}(A \otimes K') \quad \Longleftrightarrow \quad \mathbb{Q}(\pi) = \mathbb{Q}(\pi^m).$$

Exercise 3.14. Let $m, n \in \mathbb{Z}$ with $m > n > 0$; write $g = m + n$ and $q = p^g$. Consider the polynomial $T^2 + p^n T + p^g$, and let π be a zero of this polynomial.

(a) *Show that* π *is a* p^g*-Weil number; compute the p-adic values of all conjugates of* π.

(b) By the previous theorem we see that π defines the isogeny class of an abelian variety A over $\mathbb{F}_q$. It can be shown that A has dimension g, and that $\mathcal{N}(A) = (m,n) + (n,m)$, see [**79**], page 98. *This gives a proof of a conjecture of Manin,* see 5.21.

3.15. ℓ-adic monodromy. (Any characteristic.) Let K be a base field, of any characteristic. Write $G_K = \mathrm{Gal}(K^{\mathrm{sep}}/K)$. Let ℓ be a prime number, not equal to $\mathrm{char}(K)$. Note that this implies that $T_\ell(A) = \varprojlim_j A[\ell^j](K^{\mathrm{sep}})$ can be considered as a group isomorphic with $(\mathbb{Z}_\ell)^{2g}$ with a continuous G_K-action. See 10.5, 10.8.

Theorem 3.16. **Extra** (Tate, Faltings, and many others) *Suppose K is of finite type over its prime field. (Any characteristic.) The canonical map*

$$\mathrm{End}(A) \otimes_{\mathbb{Z}} \mathbb{Z}_\ell \quad \xrightarrow{\sim} \quad \mathrm{End}(T_\ell(A)) \cong \mathrm{End}_{G_K}((\mathbb{Z}_\ell)^{2g})$$

is an isomorphism.

This was conjectured by Tate. In 1966 Tate proved this in case K is a finite field, see [**78**]. The case of function fields in characteristic p was proved by Zarhin and by Mori, see [**90**], [**91**], [**52**]; also see [**51**], pp. 9/10 and VI.5 (pp. 154-161).

The case K is a number field was open for a long time; it was finally proved by Faltings in 1983, see [**26**]. For the case of a function field in characteristic zero, see [**29**], Theorem 1 on page 204.

Remark 3.17. **Extra** The previous result holds over a number field, but the Tate map need not be an isomorphism for an abelian variety over a local field.

Example. Lubin and Tate, see [**47**], 3.5; see [**63**], 14.9. *There exists a finite extension $L \supset \mathbb{Q}_p$ and an abelian variety over L such that*

$$\mathrm{End}(A) \otimes_{\mathbb{Z}} \mathbb{Z}_\ell \quad \subsetneqq \quad \mathrm{End}(T_\ell(A)).$$

We give details of a proof of this fact (slightly more general than in the paper by Lubin and Tate). Choose a prime number p, and choose a supersingular elliptic curve E_0 over $K = \mathbb{F}_q$ such that the endomorphism ring $R := \mathrm{End}(E_0)$ has rank 4 over $\mathbb{Z}$. In that case R is a maximal order in the endomorphism algebra $D := \mathrm{End}^0(E_0)$, which is a quaternion division algebra central over $\mathbb{Q}$. Let I be the index set of all subfields L_i of D, and let

$$\Lambda := \bigcup_{i \in I} (L_i \otimes \mathbb{Q}_p) \subset D \otimes \mathbb{Q}_p.$$

Claim.

$$\Lambda \subsetneqq D_p := D \otimes \mathbb{Q}_p.$$

Indeed, the set I is countable, and $[L_i : \mathbb{Q}] \leq 2$ for every i. Hence Λ is a countable union of 2-dimensional $\mathbb{Q}_p$-vector spaces inside $D_p \cong (\mathbb{Q}_p)^4$. The claim follows.

Hence we can choose $\psi_0 \in R_p := R \otimes \mathbb{Z}_p$ such that $\psi_0 \notin \Lambda$: first choose ψ_0' in D_p outside Λ, then multiply with a power of p in order to make $\psi_0 = p^n \cdot \psi_0'$ integral.

Consider $X_0 := E_0[p^\infty]$. The pair (X_0, ψ_0) can be lifted to characteristic zero, see [**63**], Lemma 14.7, hence to (X, ψ) defined over an order in a finite extension L of $\mathbb{Q}_p$. We see that $\mathrm{End}^0(X) = \mathbb{Q}_p(\psi)$, which is a quadratic extension of $\mathbb{Q}_p$. By the theorem of Serre and Tate, see 2.7, we derive an elliptic curve E, which is a lifting of E_0, such that $E[p^\infty] = X$. Clearly $\mathrm{End}(E) \otimes \mathbb{Z}_p \subset \mathrm{End}(X)$.

Claim. $\mathrm{End}(E) = \mathbb{Z}$.
In fact, if $\mathrm{End}(E)$ were bigger, we would have $\mathrm{End}(E) \otimes \mathbb{Z}_p = \mathrm{End}(X)$. Hence

$\psi \in \mathrm{End}^0(E) \subset \Lambda$, which is a contradiction. This finishes the proof of the example:

$$\mathrm{End}(E) = \mathbb{Z} \quad \text{and} \quad \dim_{\mathbb{Q}_p} \mathrm{End}^0(X) = 2 \quad \text{and} \quad \mathrm{End}(E) \otimes \mathbb{Z}_p \subsetneqq \mathrm{End}(X).$$

However, surprise, in the "anabelian situation" of a hyperbolic curve over a p-adic field, the analogous situation, gives an isomorphism for fundamental groups, see [**50**]. We see: the Tate conjecture as in 3.16 does not hold over p-adic fields but the Grothendieck "anabelian conjecture" is true for hyperbolic curves over p-adic fields. Grothendieck took care to formulate his conjecture with a number field as base field, see [**75**], page 19; we see that this care is necessary for the original Tate conjecture for abelian varieties, but for hyperbolic curves this condition can be relaxed.

3.18. We would like to have a p-adic analogue of 3.16. For this purpose it is convenient to have p-divisible groups instead of Tate ℓ-groups, and in fact the following theorem now has been proved to be true.

Theorem 3.19. **BB** (Tate and De Jong) *Let R be an integrally closed, Noetherian integral domain with field of fractions K (any characteristic). Let X, Y be p-divisible groups over* $\mathrm{Spec}(R)$. *Let $\beta_K : X_K \to Y_K$ be a homomorphism. There exists (a unique) $\beta : X \to Y$ over* $\mathrm{Spec}(R)$ *extending β_K.*

This was proved by Tate, under the extra assumption that the characteristic of K is zero. For the case $\mathrm{char}(K) = p$, see [**19**], 1.2 and [**18**], Theorem 2 on page 261.

Theorem 3.20. **BB** (Tate and De Jong) *Let K be a field finitely generated over $\mathbb{F}_p$. Let A and B be abelian varieties over K. The natural map*

$$\mathrm{Hom}(A, B) \otimes \mathbb{Z}_p \quad \xrightarrow{\sim} \quad \mathrm{Hom}(A[p^\infty], B[p^\infty])$$

is an isomorphism.

This was proved by Tate in case K is a finite field; a proof was written up in [**85**]. The case of a function field over $\mathbb{F}_p$ was proved by Johan de Jong, see [**19**], Theorem 2.6. This case follows from the result by Tate and from the proceeding result 3.19 on extending homomorphisms.

3.21. Ekedahl-Oort strata. **BB** In [**67**] a new technique is developed, which will be used below. We sketch some of the details of that method. We will only indicate details relevant for the polarized case (and we leave aside the much easier unpolarized case).

A finite group scheme N (say over a perfect field) for which $N[V] = \mathrm{Im}(F_N)$ and $N[F] = \mathrm{Im}(V_N)$ is called a BT_1 group scheme (a p-divisible group scheme truncated at level 1). By a theorem of Kraft, independently observed by Oort, for a given rank over an algebraically closed field k the number of isomorphism classes of BT_1 group schemes is finite, see [**43**]. For any abelian variety A, the group scheme $A[p]$ is a BT_1 group scheme. A principal polarization λ on A induces a form on $A[p]$, and the pair $(A, \lambda)[p]$ is a polarized BT_1 group scheme, see [**67**], Section 9 (there are subtleties in case $p = 2$: the form has to be taken, over a perfect field, on the Dieudonné module of $A[p]$).

3.21.1. *The number of isomorphism classes of polarized BT_1 group schemes $(N, \langle , \rangle)$ over k of a given rank is finite; see the classification in* [**67**], 9.4.

Let φ be the isomorphism type of a polarized BT_1 group scheme. Consider $S_\varphi \subset \mathcal{A}_{g,1}$, the set of all $[(A, \lambda)]$ such that $(A, \lambda)[p]$ geometrically belongs to the isomorphism class φ.

3.21.2. *It can be shown that S_φ is a locally closed set*; it is called an EO-stratum. We obtain $\mathcal{A}_{g,1} = \bigsqcup_\varphi S_\varphi$, a disjoint union of locally closed sets. *This is a stratification*, in the sense that *the boundary of a stratum is a union of lower dimensional strata.*

One of the main theorems of this theory is that

3.21.3. *The set S_φ is quasi-affine* (i.e., open in an affine scheme) *for every φ*, see [**67**], 1.2.

The finite set Φ_g of such isomorphism types has two partial orderings, see [**67**], 14.3. One of these, denoted by $\varphi \subset \varphi'$, is defined by the property that S_φ is contained in the Zariski closure of $S_{\varphi'}$.

3.22. An application. *Let $x \in \mathcal{A}_{g,1}$. Let*

$$\left(\mathcal{H}_\ell^{\mathrm{Sp}}(x)\right)^{\mathrm{Zar}} = (\mathcal{H}_\ell(x) \cap \mathcal{A}_{g,1})^{\mathrm{Zar}} \quad \subset \quad \mathcal{A}_{g,1}$$

be the Zariski closure of the ℓ-power Hecke orbit of x in $\mathcal{A}_{g,1}$. This closed set in $\mathcal{A}_{g,1}$ contains a supersingular point.

Use 3.21 and the second part of 1.14. □

4. Dieudonné modules and Cartier modules

In this section we explain the theory of Cartier modules and Dieudonné modules. These theories provide equivalence of categories of geometric objects such as commutative smooth formal groups or p-divisible groups on the one side, and modules over certain non-commutative rings on the other side. As a result, questions on commutative smooth formal groups or p-divisible groups, which are apparently non-linear in nature, are translated into questions in linear algebra over rings. Such results are essential for any serious computation.

There are many versions and flavors of Dieudonné theory. We explain the Cartier theory for commutative smooth formal groups over general commutative rings, and the covariant Dieudonné modules for p-divisible groups over perfect fields of characteristic $p > 0$. Since the Cartier theory works over general commutative rings, one can "write down" explicit deformations over complete rings such as $k[[x_1, \ldots, x_n]]$ or $W(k)[[x_1, \ldots, x_n]]$, something rarely feasible in algebraic geometry. For our purpose it is those commutative formal groups which are formal completions of p-divisible groups that are really relevant; see 10.9 for the relation between such p-divisible formal groups and the connected p-divisible groups.

Remarks on notation:

(i) In the first part of this section, on Cartier theory, R denotes a commutative ring with 1, or a commutative $\mathbb{Z}_{(p)}$-algebra with 1.

(ii) In this section, we used V and F as elements in the Cartier ring $\mathrm{Cart}_p(R)$ or the smaller Dieudonné ring $R_K \subset \mathrm{Cart}_p(R)$ for a perfect field K. In the rest of this article, the notations $\mathcal{V}$ and $\mathcal{F}$ are used; $\mathcal{V}$ corresponds to

the relative Frobenius morphism and $\mathcal{F}$ corresponds to the Verschiebung morphisms for commutative smooth formal groups or p-divisible groups over K.

A synopsis of Cartier theory. The main theorem of Cartier theory says that there is an equivalence between the category of commutative smooth formal groups over R and the category of left modules over a non-commutative ring $\mathrm{Cart}_p(R)$ satisfying certain conditions. See 4.27 for a precise statement.

The Cartier ring $\mathrm{Cart}_p(R)$ plays a crucial role. This is a topological ring which contains elements V, F and $\{\langle a\rangle \mid a \in R\}$. These elements form a set of topological generators, in the sense that every element of $\mathrm{Cart}_p(R)$ has a unique expression as a convergent sum in the following form

$$\sum_{m,n\geq 0} V^m \langle a_{mn}\rangle F^n\,,$$

with $a_{mn} \in R$ for all $m, n \geq 0$; moreover for each $m \in \mathbb{N}$, there exists a constant $C_m > 0$ such that $a_{mn} = 0$ for all $n \geq C_m$. Every convergent sum as above is an element of $\mathrm{Cart}_p(R)$. These topological generators satisfy the following commutation relations:

- $F\,\langle a\rangle = \langle a^p\rangle\,F$ for all $a \in R$;
- $\langle a\rangle V = V\,\langle a^p\rangle$ for all $a \in R$;
- $\langle a\rangle\,\langle b\rangle = \langle ab\rangle$ for all $a, b \in R$;
- $FV = p$;
- $V^m\langle a\rangle F^m\,V^n\langle b\rangle F^n \;=\; p^r\,V^{m+n-r}\langle a^{p^{n-r}}b^{p^{m-r}}\rangle F^{m+n-r}$ for all $a, b \in R$ and all $m, n \in \mathbb{N}$, where $r = \min\{m, n\}$.

Moreover, the ring of p-adic Witt vectors $W_p(R)$ is embedded in $\mathrm{Cart}_p(R)$ by the formula

$$W_p(R) \;\ni\; \underline{c} = (c_0, c_1, c_2, \ldots) \;\longmapsto\; \sum_{n\geq 0} V^n\langle c_n\rangle F^n \;\in\; \mathrm{Cart}_p(R)\,.$$

The topology of $\mathrm{Cart}_p(R)$ is given by the decreasing filtration

$$\mathrm{Fil}^n\left(\mathrm{Cart}_p(R)\right) := V^n \cdot \mathrm{Cart}_p(R)\,,$$

making $\mathrm{Cart}_p(R)$ a complete and separated topological ring. Under the equivalence of categories mentioned above, a left $\mathrm{Cart}_p(R)$ module corresponds to a finite dimensional smooth commutative formal group G over R if and only if

- $V : M \to M$ is injective,
- $M \xrightarrow{\sim} \varprojlim_n M/V^nM$, and
- M/VM is a projective R-module of finite type.

If so, then $\mathrm{Lie}(G/R) \cong M/VM$, and M is a finitely generated $\mathrm{Cart}_p(R)$-module. See 4.18 for the definition of $\mathrm{Cart}_p(R)$, 4.19 for the commutation relations in $\mathrm{Cart}_p(R)$, and 4.23 for some other properties of R.

References for Cartier theory. We highly recommend [**93**], where the approach in §2 of [**73**] is fully developed. Other references for Cartier theory are [**44**] and [**36**].

Remarks on Dieudonné theories.
(1) As already mentioned, the effect of a Dieudonné theory for p-divisible groups (and/or formal groups) is to translate questions for p-divisible groups (and/or formal groups) into questions in linear algebra for modules over suitable rings. A survey of Dieudonné theories can be found in [**6**]. The book [**24**] is a good introduction to the classical contravariant Dieudonné theory over a perfect field $K \supset \mathbb{F}_p$.

(2) The covariant Dieudonné theory described in this section is the dual version of the classical contravariant theory. For a p-divisible group X over a perfect field $K \supset \mathbb{F}_p$, the covariant Dieudonné module $\mathbb{D}(X)$ described in Theorem 4.33 is functorially isomorphic to $\mathbb{D}_{\text{classical}}(X^t)$, the classical contravariant Dieudonné module of the Serre dual X^t of X as defined in [**48**] and [**24**].

(3) The Cartier theory is a Dieudonné theory for commutative formal groups. As explained in 10.9, a p-divisible group X over an Artinian local ring R with residue characteristic p whose maximal étale quotient is trivial can be recovered from the formal completion $X^\wedge$ of X. Such p-divisible groups are called *p-divisible formal groups*. Given a p-divisible formal group X over an Artinian local ring R with residue characteristic p, the formal completion $X^\wedge$ is a smooth commutative formal group over R, and the Cartier theory provides us with a module V-flat V-reduced left $\text{Cart}_p(R)$-module $M_p(X^\wedge)$.

(4) The Cartier module attached to a p-divisible formal group X over a perfect field $K \supset \mathbb{F}_p$ is canonically isomorphic to the Dieudonné module $\mathbb{D}(X)$; see Theorem 4.33 (2). See also 4.34 for the relation with the Dieudonné crystal attached to X.

(5) Let $(R, \mathfrak{m})$ be a complete Noetherian local ring with residue characteristic p. Suppose that X is a p-divisible group over R such that $X \times_{\text{Spec}(R)} \text{Spec}(R/\mathfrak{m})$ is a p-divisible formal group. Let $X^\wedge$ be the formal completion of X, defined as the scheme-theoretic inductive limit of the finite flat group schemes $X[p^n] \times_{\text{Spec}(R)} \text{Spec}(R/\mathfrak{m}^i)$ over $\text{Spec}(R/m^i)$ as m and i go to ∞. Then $X^\wedge$ is a commutative smooth formal group over R, whose closed fiber is a p-divisible formal group. Conversely, suppose that X' is a commutative smooth formal group over R whose closed fiber X_0 is the formal completion of a p-divisible formal group X_0 over the residue field $R/\mathfrak{m}$. Then for each $i > 0$, the formal group $X' \times_{\text{Spec}R} \text{Spec}(R/\mathfrak{m}^i)$ is the formal completion of a p-divisible formal group $X_i \to \text{Spec}(R/\mathfrak{m}^i)$, uniquely determined up to unique isomorphism. The projective limit of the p-divisible groups $X_i \to \text{Spec}(R/\mathfrak{m}^i)$ "is" a p-divisible group X over R whose closed fiber is the p-divisible formal group X_0 over $R/\mathfrak{m}$; see 10.21 and 10.9. Notice that the fibers of $X \to \text{Spec}(R/pR))$ may not be p-divisible formal groups; the universal p-divisible group over the equi-characteristic deformation space of a supersingular elliptic curve provides an example.

The upshot of the previous paragraph is that we can apply the Cartier theory to construct and study deformation of p-divisible formal groups. Suppose that R is a complete Noetherian local ring whose residue field $R/\mathfrak{m}$ has characteristic p. In order to produce a deformation over R of a p-divisible formal group X_0 over $R/\mathfrak{m}$, it suffices to "write down" a V-flat V-reduced left $\text{Cart}_p(R)$-module M such that the tensor product $\text{Cart}_p(R/\mathfrak{m}) \otimes_{\text{Cart}_p(R)} M$ is isomorphic to the Cartier module

$M_p(X_0)$ attached to X_0. This way of explicitly constructing deformations over rings such as $K[[x_1,\dots,x_N]]$ and $W(K)[[x_1,\dots,x_N]]$ whose analog in deformation theory for general algebraic varieties is often intractable, becomes manageable. This method is essential for Sections 5, 7, 8.

We strongly advise readers with no prior experience with Cartier theory to accept the synopsis above as a "big black box" and use the materials in 4.1–4.27 as a dictionary only when necessary. Instead, we suggest such readers start with 4.28–4.52, get familiar with the ring $\mathrm{Cart}_p(K)$ in the case when $R = K$, a perfect field of characteristic p, play with some examples of finitely generated modules over $\mathrm{Cart}_p(K)$ in conjunction with the theory of covariant Dieudonné modules over perfect fields in characteristic p and do some of the exercises. See the Lemma/Exercise after Def. 4.28 for a concrete definition of the Cartier ring $\mathrm{Cart}_p(K)$ as the V-adic completion of the Dieudonné ring R_K.

Cartier theory

Definition 4.1. Let R be a commutative ring with 1.

(1) Let Nilp_R be the category of all nilpotent R-algebras, consisting of all commutative R-algebras N without unit such that $N^n = (0)$ for some positive integer n.

(2) A *commutative smooth formal group* over R is a covariant functor $G : \mathrm{Nilp}_R \to \mathrm{Ab}$ from Nilp_R to the category of all abelian groups such that the following properties are satisfied:
- G commutes with finite inverse limits;
- G is formally smooth, i.e., every surjection $N_1 \to N_2$ in Nilp_R induces a surjection $G(N_1) \to G(N_2)$;
- G commutes with arbitrary direct limits.

(3) The *Lie algebra* of a commutative smooth formal group G is defined to be $G(N_0)$, where N_0 is the object in Nilp_R whose underlying R-module is R, and $N_0^2 = (0)$.

Remark. Let G be a commutative smooth formal group over R, then G extends uniquely to a functor $G^\sim$ on the category $\mathrm{ProNilp}_R$ of all filtered projective system of nilpotent R-algebras which commutes with filtered projective limits. This functor $G^\sim$ is often denoted G by abuse of notation.

Example. Let A be a commutative smooth group scheme of finite presentation over R. For every nilpotent R-algebra N, denote by $R \oplus N$ the commutative R-algebra with multiplication given by

$$(u_1, n_1) \cdot (u_2, n_2) = (u_1u_2, u_1n_2 + u_2n_1 + n_1n_2) \qquad \forall u_1, u_2 \in R \ \ \forall n_1, n_2 \in N\,.$$

The functor which sends an object N in Nilp_R to the abelian group

$$\mathrm{Ker}\,(A(R \oplus N) \to A(R))$$

is a commutative smooth formal group over R, denoted by $A^\wedge$. Note that the functor $A^\wedge$ commutes with arbitrary inductive limits because A does.

Here are two special cases: we have

$$\mathbb{G}_a^\wedge(N) = N \qquad \text{and} \qquad \mathbb{G}_m^\wedge(N) = 1 + N \subset (R \oplus N)^\times$$

for all $N \in \mathrm{Ob}(\mathrm{Nilp}_R)$.

Definition 4.2. We define a *restricted version* of the smooth formal group attached to the universal Witt vector group over R, denoted by Λ_R, or Λ when the base ring R is understood.

$$\Lambda_R(N) = 1 + t\,R[t] \otimes_R N \subset ((R \oplus N)[t])^\times \qquad \forall\ N \in \mathrm{Ob}(\mathrm{Nilp}_R)\,.$$

In other words, the elements of $\Lambda(N)$ consists of all polynomials of the form $1 + u_1\,t + u_2\,t^2 + \cdots + u_r\,t^r$ for some $r \geq 0$, where $u_i \in N$ for $i = 1, \ldots, r$. The group law of $\Lambda(N)$ comes from multiplication in the polynomial ring $(R \oplus N)[t]$ in one variable t.

Remarks.

(i) The formal group Λ plays the role of a free generator in the category of (smooth) formal groups; see Theorem 4.4.

(ii) When we want to emphasize that the polynomial $1+\sum_{i\geq 1}\ u_i\,t^i$ is regarded as an element of $\Lambda(N)$, we denote it by $\lambda(1 + \sum_{i\geq 1}\ u_i\,t^i)$.

(iii) Note that the functor $N \mapsto 1+N\,t\,R[[t]] \subset (R\oplus N)[[t]]^\times$ is not a commutative smooth formal group because it does not commute with arbitrary direct limits.

Exercise 4.3. Let $R[[X]]^+ = X\,R[[X]]$ be the set of all formal power series over R with constant term 0; it is an object in $\mathrm{ProNilp}_R$. Show that $\Lambda(R[[X]]^+)$ equals

$$\left\{ \prod_{m,n\geq 1} (1 - a_{mn}\,X^m\,t^n) \;\middle|\; a_{m,n} \in R,\ \forall\, m\, \exists\, C_m > 0 \text{ s.t. } a_{mn} = 0\ \forall\ n \geq C_m \right\}.$$

Theorem 4.4. **BB** *Let $H : \mathrm{Nilp}_R \to \mathrm{Ab}$ be a commutative smooth formal group over R. Let $\Lambda = \Lambda_R$ be the functor defined in* 4.2. *Then the map*

$$Y_H : \mathrm{Hom}(\Lambda_R, H) \to H(R[[X]]^+)$$

which sends each homomorphism $\alpha : \Lambda \to H$ of group-valued functors to the element

$$\alpha_{R[[X]]^+}(1 - Xt) \in H(R[[X]]^+)$$

is a bijection.

Remark. The formal group Λ is in some sense a free generator of the additive category of commutative smooth formal groups, a phenomenon reflected in Theorem 4.4.

Definition 4.5. (i) Define $\mathrm{Cart}(R)$ to be $(\mathrm{End}(\Lambda_R))^{\mathrm{op}}$, the opposite ring of the endomorphism ring of the smooth formal group Λ_R. According to Theorem 4.4, for every commutative smooth formal group $H : \mathrm{Nilp}_R \to \mathrm{Ab}$, the abelian group $H(R[[X]]^+) = \mathrm{Hom}(\Lambda_R, H)$ is a *left* module over $\mathrm{Cart}(R)$.
(ii) We define some special elements of the Cartier ring $\mathrm{Cart}(R)$, naturally identified with $\Lambda(R[[X]])$ via the bijection $Y = Y_\Lambda : \mathrm{End}(\Lambda) \xrightarrow{\sim} \Lambda(R[[X]]^+)$ in Theorem 4.4.

- $V_n := Y^{-1}(1 - X^n\,t)\,, n \geq 1$,
- $F_n := Y^{-1}(1 - X\,t^n)\,, n \geq 1$,
- $[c] := Y^{-1}(1 - c\,X\,t)\,, c \in R$.

Corollary. For every commutative ring with 1 we have

$$\mathrm{Cart}(R) = \left\{ \sum_{m,n\geq 1} V_m\,[c_{mn}]\,F_n \;\middle|\; c_{mn}\in R,\ \forall\, m\,\exists\, C_m>0 \text{ s.t. } c_{mn}=0\ \forall\ n\geq C_m \right\}.$$

Proposition 4.6. **BB** *The following identities hold in* $\mathrm{Cart}(R)$.

(1) $V_1 = F_1 = 1$, $F_n\,V_n = n$.
(2) $[a]\,[b] = [ab]$ *for all* $a, b \in R$
(3) $[c]\,V_n = V_n\,[c^n]$, $F_n\,[c] = [c^n]\,F_n$ *for all* $c \in R$, *all* $n \geq 1$.
(4) $V_m\,V_n = V_n\,V_m = V_{mn}$, $F_m\,F_n = F_n\,F_m = F_{mn}$ *for all* $m, n \geq 1$.
(5) $F_n\,V_m = V_m\,F_n$ *if* $(m, n) = 1$.
(6) $(V_n[a]F_n)\cdot(V_m[b]F_m) = r\,V_{\frac{mn}{r}}\left[a^{\frac{m}{r}}b^{\frac{n}{r}}\right]F_{\frac{mn}{r}}$, $r = (m, n)$, *for all* $a, b \in R$, $m, n \geq 1$.

Definition 4.7. The ring $\mathrm{Cart}(R)$ has a natural filtration $\mathrm{Fil}^\bullet\mathrm{Cart}(R)$ by right ideals, where $\mathrm{Fil}^j\mathrm{Cart}(R)$ is defined by

$$\left\{ \sum_{m\geq j}\sum_{n\geq 1} V_m[a_{mn}]F_n \mid a_{mn}\in R,\ \forall\, m\geq j,\ \exists\, C_m>0 \text{ s.t. } a_{mn}=0 \text{ if } n\geq C_m \right\}$$

for every integer $j \geq 1$. The Cartier ring $\mathrm{Cart}(R)$ is complete with respect to the topology given by the above filtration. Moreover each right ideal $\mathrm{Fil}^j\mathrm{Cart}(R)$ is open and closed in $\mathrm{Cart}(R)$.

Remark. The definition of the Cartier ring gives a functor

$$R \longmapsto \mathrm{Cart}(R)$$

from the category of commutative rings with 1 to the category of complete filtered rings with 1.

Definition 4.8. Let R be a commutative ring with 1.

(1) A *V-reduced* left $\mathrm{Cart}(R)$-module is a left $\mathrm{Cart}(R)$-module M together with a separated decreasing filtration of M

$$M = \mathrm{Fil}^1 M \supset \mathrm{Fil}^2 M \supset \cdots \supset \mathrm{Fil}^n M \supset \mathrm{Fil}^{n+1} \supset \cdots$$

such that each $\mathrm{Fil}^n M$ is an abelian subgroup of M and
   (i) $(M, \mathrm{Fil}^\bullet M)$ is complete with respect to the topology given by the filtration $\mathrm{Fil}^\bullet M$. In other words, the natural map

$$\mathrm{Fil}^n M \to \varprojlim_{m\geq n} (\mathrm{Fil}^n M/\mathrm{Fil}^m M)$$

   is a bijection for all $n \geq 1$.
   (ii) $V_m \cdot \mathrm{Fil}^n M \subset \mathrm{Fil}^{mn} M$ for all $m, n \geq 1$.
   (iii) The map V_n induces a bijection $V_n : M/\mathrm{Fil}^2 M \xrightarrow{\sim} \mathrm{Fil}^n M/\mathrm{Fil}^{n+1} M$ for every $n \geq 1$.
   (iv) $[c] \cdot \mathrm{Fil}^n M \subset \mathrm{Fil}^n M$ for all $c \in R$ and all $n \geq 1$.
   (v) For every $m, n \geq 1$, there exists an $r \geq 1$ such that $F_m \cdot \mathrm{Fil}^r M \subset \mathrm{Fil}^n M$.

(2) A V-reduced left $\mathrm{Cart}(R)$-module $(M, \mathrm{Fil}^\bullet M)$ is *V-flat* if $M/\mathrm{Fil}^2 M$ is a flat R-module. The R-module $M/\mathrm{Fil}^2 M$ is called the *tangent space* of $(M, \mathrm{Fil}^\bullet M)$.

Definition 4.9. Let $H : \mathrm{Nilp}_R \to \mathrm{Ab}$ be a commutative smooth formal group over R. The abelian group $\mathrm{M}(H) := H(R[[X]]^+)$ has a natural structure as a left $\mathrm{Cart}(R)$-module according to Theorem 4.4 The $\mathrm{Cart}(R)$-module $\mathrm{M}(H)$ has a natural filtration, with

$$\mathrm{Fil}^n\mathrm{M}(H) := \mathrm{Ker}(H(R[[X]]^+) \to H(R[[X]]^+/X^nR[[X]]))\,.$$

We call the pair $(\mathrm{M}(H), \mathrm{Fil}^\bullet\mathrm{M}(H))$ the *Cartier module attached to* H.

Definition 4.10. Let M be a V-reduced left $\mathrm{Cart}(R)$-module and let Q be a right $\mathrm{Cart}(R)$-module.

(i) For every integer $m \geq 1$, let $Q_m := \mathrm{Ann}_Q(\mathrm{Fil}^m\mathrm{Cart}(R))$ be the subgroup of Q consisting of all elements $x \in Q$ such that $x \cdot \mathrm{Fil}^m\mathrm{Cart}(R) = (0)$. Clearly we have $Q_1 \subseteq Q_2 \subseteq Q_3 \subseteq \cdots$.

(ii) For each $m, r \geq 1$, define $Q_m \odot M^r$ to be the image of $Q_m \otimes \mathrm{Fil}^r M$ in $Q \otimes_{\mathrm{Cart}(R)} M$.

Notice that if $r \geq m$ and $s \geq m$, then $Q_m \odot M^r = Q_m \odot M^s$. Hence $Q_m \odot M^m \subseteq Q_n \odot M^n$ if $m \leq n$.

(iii) Define the *reduced tensor product* $Q\overline{\otimes}_{\mathrm{Cart}(R)}M$ by

$$Q\overline{\otimes}_{\mathrm{Cart}(R)}M = Q \otimes_{\mathrm{Cart}(R)} M \Big/ \big(\bigcup_m (Q_m \odot M^m)\big)\ .$$

Remark. The reduced tensor product is used to construct the arrow in the "reverse direction" in the equivalence of category in 4.11 below.

Theorem 4.11. **BB** *Let R be a commutative ring with* 1. *There is a canonical equivalence of categories between the category of smooth commutative formal groups over R as defined in* 4.1 *and the category of V-flat V-reduced left* $\mathrm{Cart}(R)$*-modules, defined as follows.*

$$\begin{array}{ccc} \{\text{smooth formal groups over } R\} & \xrightarrow{\sim} & \{V\text{-flat } V\text{-reduced left } \mathrm{Cart}(R)\text{-mod}\} \\ G & \longmapsto & \mathrm{M}(G) = \mathrm{Hom}(\Lambda, G) \\ \Lambda\overline{\otimes}_{\mathrm{Cart}(R)}M & \longmapsfrom & M \end{array}$$

Recall that $\mathrm{M}(G) = \mathrm{Hom}(\Lambda, G)$ *is canonically isomorphic to $G(X\,R[[X]])$, the group of all formal curves in the smooth formal group G. The reduced tensor product $\Lambda\overline{\otimes}_{\mathrm{Cart}(R)}M$ is the functor whose value at any nilpotent R-algebra N is $\Lambda(N)\overline{\otimes}_{\mathrm{Cart}(R)}M$.*

The Cartier ring $\mathrm{Cart}(R)$ contains the ring of universal Witt vectors $W^\sim(R)$ as a subring which contains the unit element of $\mathrm{Cart}(R)$.

Definition 4.12.

(1) The *universal Witt vector group* $W^\sim$ is defined as the functor from the category of all commutative algebras with 1 to the category of abelian groups such that

$$W^\sim(R) = 1 + T\,R[[T]] \subset R[[T]]^\times$$

for every commutative ring R with 1.

When we regard a formal power series $1 + \sum_{m\geq 1} u_m\,T^m$ in $R[[T]]$ as an element of $W^\sim(R)$, we use the notation $\omega(1 + \sum_{m\geq 1} u_m\,T^m)$. It is

easy to see that every element of $W^\sim(R)$ has a unique expression as

$$\omega\left(\prod_{m\geq 1}(1-a_m\,T^m)\right).$$

Hence $W^\sim$ is isomorphic to $\operatorname{Spec}\mathbb{Z}[x_1,x_2,x_3,\ldots]$ as a scheme; the R-valued point such that $x_i\mapsto a_i$ is denoted by $\omega(\underline{a})$, where $\underline{a}$ is short for $(a_1,a_2,a_3,\ldots)$, and $\omega(\underline{a})=\omega(\prod_{m\geq 1}(1-a_m\,T^m))$.

(2) The group scheme $W^\sim$ has a natural structure as a ring scheme, such that multiplication on $W^\sim$ is determined by the formula

$$\omega(1-a\,T^m)\cdot\omega(1-b\,T^n)=\omega\left(\left(1-a^{\frac{n}{r}}\,b^{\frac{m}{r}}\,T^{\frac{mn}{r}}\right)^r\right),\quad \text{where } r=\gcd(m,n).$$

(3) There are two families of endomorphisms of the group scheme $W^\sim$: V_n and F_n, $n\in\mathbb{N}_{\geq 1}$. Also, for each commutative ring R with 1 and each element $c\in R$ we have an endomorphism $[c]$ of $W^\sim\times_{\operatorname{Spec}\mathbb{Z}}\operatorname{Spec}R$. These operators make $W^\sim(R)$ a left $\operatorname{Cart}(R)$-module; they are defined as follows

$$\begin{aligned} V_n &: \omega(f(T)) \mapsto \omega(f(T^n)),\\ F_n &: \omega(f(T)) \mapsto \textstyle\sum_{\zeta\in\mu_n}\omega(f(\zeta\,T^{\frac{1}{n}})), \qquad (\text{formally})\\ [c] &: \omega(f(T)) \mapsto \omega(f(cT)). \end{aligned}$$

The formula for $F_n(\omega(f(T)))$ means that $F_n(\omega(f(T)))$ is defined as the unique element such that $V_n(F_n(\omega(f(T))))=\sum_{\zeta\in\mu_n}\omega(f(\zeta\,T))$.

Exercise 4.13. Show that the Cartier module of $\mathbb{G}_m^\wedge$ over R is naturally isomorphic to $W^\sim(R)$ as a module over $\operatorname{Cart}(R)$.

Proposition 4.14. **BB** *Let R be a commutative ring with 1.*

(i) *The subset S of $\operatorname{Cart}(R)$ consisting of all elements of the form*

$$\sum_{n\geq 1}V_n[a_n]F_n\,,\quad a_n\in R\ \forall\, n\geq 1$$

forms a subring of $\operatorname{Cart}(R)$.

(ii) *The injective map*

$$W^\sim(R)\hookrightarrow\operatorname{Cart}(R),\qquad \omega(\underline{a})\mapsto\sum_{n\geq 1}V_n\,[a_n]\,F_n$$

is an injective homomorphism of rings which sends 1 to 1; its image is the subring S defined in (i).

Definition 4.15. It is a fact that every prime number $\ell\neq p$ is invertible in $\operatorname{Cart}(\mathbb{Z}_{(p)})$. Define elements ϵ_p and $\epsilon_{p,n}$ of the Cartier ring $\operatorname{Cart}(\mathbb{Z}_{(p)})$ for $n\in\mathbb{N}$, $(n,p)=1$ by

$$\begin{aligned} \epsilon_p=\epsilon_{p,1} &= \sum_{\substack{(n,p)=1\\ n\geq 1}}\frac{\mu(n)}{n}V_nF_n=\prod_{\substack{\ell\neq p\\ \ell\ \text{prime}}}\left(1-\frac{1}{\ell}V_\ell F_\ell\right)\\ \epsilon_{p,n} &= \frac{1}{n}V_n\epsilon_pF_n \end{aligned}$$

where μ is the Möbius function on $\mathbb{N}_{\geq 1}$, characterized by the following properties: $\mu(mn)=\mu(m)\,\mu(n)$ if $(m,n)=1$, and for every prime number ℓ we have $\mu(\ell)=-1$,

$\mu(\ell^i) = 0$ if $i \geq 2$. For every commutative with 1 over $\mathbb{Z}_{(p)}$, the image of ϵ_p in $\mathrm{Cart}(R)$ under the canonical ring homomorphism $\mathrm{Cart}(\mathbb{Z}_{(p)}) \to \mathrm{Cart}(R)$ is also denoted by ϵ_p.

Exercise 4.16. Let R be a $\mathbb{Z}_{(p)}$-algebra, and let $(a_m)_{m\geq 0}$ be a sequence in R. Prove the equality

$$\epsilon_p\left(\omega\left(\prod_{m\geq 1}(1-a_m\,T^m)\right)\right) = \epsilon_p\left(\omega\left(\prod_{n\geq 0}(1-a_{p^n}\,T^{p^n})\right)\right) = \omega\left(\prod_{n\geq 0}E(a_{p^n}\,T^{p^n})\right),$$

in $W^{\sim}(R)$, where

$$E(X) = \prod_{(n,p)=1}(1-X^n)^{\frac{\mu(n)}{n}} = \exp\left(-\sum_{n\geq 0}\frac{X^{p^n}}{p^n}\right) \in 1 + X\mathbb{Z}_{(p)}[[X]]$$

is the inverse of the classical Artin-Hasse exponential.

Proposition 4.17. **BB** *Let R be a commutative $\mathbb{Z}_{(p)}$-algebra with 1. The following equalities hold in* $\mathrm{Cart}(R)$.

(i) ${\epsilon_p}^2 = \epsilon_p$.
(ii) $\sum_{p\nmid n,\, n\geq 1} \epsilon_{p,n} = 1$.
(iii) $\epsilon_p V_n = 0$, $F_n\epsilon_p = 0$ *for all n with $p \nmid n$.*
(iv) ${\epsilon_{p,n}}^2 = \epsilon_{p,n}$ *for all $n \geq 1$ with $p \nmid n$.*
(v) $\epsilon_{p,n}\,\epsilon_{p,m} = 0$ *for all $m \neq n$ with $p \nmid mn$.*
(vi) $[c]\,\epsilon_p = \epsilon_p\,[c]$ *and* $[c]\,\epsilon_{p,n} = \epsilon_{p,n}\,[c]$ *for all $c \in R$ and all n with $p \nmid n$.*
(vii) $F_p\epsilon_{p,n} = \epsilon_{p,n}F_p$, $V_p\epsilon_{p,n} = \epsilon_{p,n}V_p$ *for all n with $p \nmid n$.*

Definition 4.18. Let R be a commutative ring with 1 over $\mathbb{Z}_{(p)}$.

(i) Denote by $\mathrm{Cart}_p(R)$ the subring $\epsilon_p\mathrm{Cart}(R)\epsilon_p$ of $\mathrm{Cart}(R)$. Note that ϵ_p is the unit element of $\mathrm{Cart}_p(R)$.
(ii) Define elements $F, V \in \mathrm{Cart}_p(R)$ by

$$F = \epsilon_p F_p = F_p\epsilon_p = \epsilon_p F_p \epsilon_p\,, \quad V = \epsilon_p V_p = V_p\epsilon_p = \epsilon_p V_p \epsilon_p\,.$$

(iii) For every element $c \in R$, denote by $\langle c\rangle$ the element

$$\epsilon_p[c]\epsilon_p = \epsilon_p[c] = [c]\epsilon_p \in \mathrm{Cart}_p(R).$$

Exercise 4.19. Prove the following identities in $\mathrm{Cart}_p(R)$.

(1) $F\,\langle a\rangle = \langle a^p\rangle\,F$ for all $a \in R$.
(2) $\langle a\rangle V = V\,\langle a^p\rangle$ for all $a \in R$.
(3) $\langle a\rangle\,\langle b\rangle = \langle ab\rangle$ for all $a, b \in R$.
(4) $FV = p$.
(5) $VF = p$ if and only if $p = 0$ in R.
(6) Every prime number $\ell \neq p$ is invertible in $\mathrm{Cart}_p(R)$. The prime number p is invertible in $\mathrm{Cart}_p(R)$ if and only if p is invertible in R.
(7) $V^m\langle a\rangle F^m\,V^n\langle b\rangle F^n = p^r\,V^{m+n-r}\langle a^{p^{n-r}}b^{p^{m-r}}\rangle F^{m+n-r}$ for all $a, b \in R$ and all $m, n \in \mathbb{N}$, where $r = \min\{m, n\}$.

Definition 4.20. Let R be a commutative $\mathbb{Z}_{(p)}$-algebra with 1. Denote by Λ_p the image of ϵ_p in Λ. In other words, Λ_p is the functor from the category Nilp_R of nilpotent commutative R-algebras to the category Ab of abelian groups such that

$$\Lambda_p(N) = \Lambda(N) \cdot \epsilon_p$$

for any nilpotent R-algebra N.

Definition 4.21.

(1) Denote by W_p the image of ϵ_p, i.e., $W_p(R) := \epsilon_p(W^\sim(R))$ for every $\mathbb{Z}_{(p)}$-algebra R. Equivalently, $W_p(R)$ is the intersection of the kernels $\mathrm{Ker}(F_\ell)$ of the operators F_ℓ on $W^\sim(R)$, where ℓ runs through all prime numbers different from p.

(2) Denote the element

$$\omega(\prod_{n=0}^{\infty} E(c_n\, T^{p^n})) \in W_p(R)$$

by $\omega_p(\underline{c})$.

(3) The endomorphism V_p, F_p of the group scheme $W^\sim$ induces endomorphisms of the group scheme W_p, denoted by V and F respectively.

Remark. The functor W_p has a natural structure as a ring-valued functor induced from that of $W^\sim$; it is represented by the scheme $\mathrm{Spec}\,\mathbb{Z}_{(p)}[y_0, y_1, y_2, \ldots, y_n, \ldots]$ such that the element $\omega_p(\underline{c})$ has coordinates $\underline{c} = (c_0, c_1, c_2, \ldots)$.

Exercise 4.22. Let R be a commutative $\mathbb{Z}_{(p)}$-algebra with 1. Let $E(T) \in \mathbb{Z}_{(p)}[[T]]$ be the inverse of the Artin-Hasse exponential as in Exer. 4.16.

(i) Prove that for any nilpotent R-algebra N, every element of $\Lambda_p(N)$ has a unique expression as a finite product

$$\prod_{i=0}^{m} E(u_i\, t^{p^i})$$

for some $m \in \mathbb{N}$, and $u_i \in N$ for $i = 0, 1, \ldots, m$.

(ii) Prove that Λ_p is a smooth commutative formal group over R.

(iii) Prove that every element of $W_p(R)$ can be uniquely expressed as an infinite product

$$\omega(\prod_{n=0}^{\infty} E(c_n\, T^{p^n})) \in W_p(R) =: \omega_p(\underline{c})\,.$$

(iv) Show that the map from $W_p(R)$ to the product ring $\prod_0^\infty R$ defined by

$$\omega_p(\underline{c}) \;\longmapsto \big(\, w_n(\underline{c}) \,\big)_{n\geq 0} \qquad \text{where} \quad w_n(\underline{c}) := \sum_{i=0}^{n} p^{n-i}\, c_{n-i}^{p^i}\,,$$

is a ring homomorphism.

Proposition 4.23.

(i) *The local Cartier ring* $\mathrm{Cart}_p(R)$ *is complete with respect to the decreasing sequence of right ideals* $V^i\mathrm{Cart}_p(R)$.

(ii) *Every element of* $\mathrm{Cart}_p(R)$ *can be expressed in a unique way as a convergent sum in the form*

$$\sum_{m,n\geq 0} V^m\langle a_{mn}\rangle F^n$$

with all $a_{mn} \in R$*, and for each* m *there exists a constant* C_m *such that* $a_{mn} = 0$ *for all* $n \geq C_m$*.*

(iii) *The set of all elements of* $\mathrm{Cart}_p(R)$ *which can be represented as a convergent sum of the form*

$$\sum_{m\geq 0} V^m\langle a_m\rangle F^m\,, \quad a_m \in R$$

is a subring of $\mathrm{Cart}_p(R)$*. The map*

$$w_p(\underline{a}) \;\mapsto \sum_{m\geq 0} V^m\langle a_m\rangle F^m \quad \underline{a} = (a_0, a_1, a_2, \ldots),\ a_i \in R\ \forall\, i \geq 0$$

establishes an isomorphism from the ring of p*-adic Witt vectors* $W_p(R)$ *to the above subring of* $\mathrm{Cart}_p(R)$*.*

Exercise 4.24. Prove that $\mathrm{Cart}_p(R)$ is naturally isomorphic to $\mathrm{End}(\Lambda_p)^{\mathrm{op}}$, the opposite ring of the endomorphism ring of $\mathrm{End}(\Lambda_p)$.

Definition 4.25. Let R be a commutative $\mathbb{Z}_{(p)}$-algebra.

(i) A *V-reduced* left $\mathrm{Cart}_p(R)$-module M is a left $\mathrm{Cart}_p(R)$-module such that the map $V : M \to M$ is injective and the canonical map $M \to \varprojlim_n (M/V^nM)$ is an isomorphism.

(ii) A V-reduced left $\mathrm{Cart}_p(R)$-module M is *V-flat* if M/VM is a flat R-module.

Theorem 4.26. *Let* R *be a commutative* $\mathbb{Z}_{(p)}$*-algebra with* 1*.*

(i) *There is an equivalence of categories between the category of V-reduced left* $\mathrm{Cart}(R)$*-modules and the category of V-reduced left* $\mathrm{Cart}_p(R)$*-modules, defined as follows.*

$$\begin{array}{ccc} \{\ V\text{-reduced left } \mathrm{Cart}(R)\text{-mod }\} & \xrightarrow{\sim} & \{\ V\text{-reduced left } \mathrm{Cart}_p(R)\text{-mod }\} \\ M & \longmapsto & \epsilon_p M \\ \mathrm{Cart}(R)\epsilon_p \widehat{\otimes}_{\mathrm{Cart}_p(R)} M_p & \longmapsfrom & M_p \end{array}$$

(ii) *Let* M *be a V-reduced left* $\mathrm{Cart}(R)$*-module* M*, and let* M_p *be the V-reduced let* $\mathrm{Cart}_p(R)$*-module* M_p *attached to* M *as in* (i) *above. There is a canonical isomorphism* $M/\mathrm{Fil}^2 M \cong M_p/VM_p$*. In particular* M *is V-flat if and only if* M_p *is V-flat. Similarly* M *is a finitely generated* $\mathrm{Cart}(R)$*-module if and only if* M_p *is a finitely generated* $\mathrm{Cart}_p(R)$*-module.*

Theorem 4.27. *Let* R *be a commutative* $\mathbb{Z}_{(p)}$*-algebra with* 1*. There is a canonical equivalence of categories between the category of smooth commutative formal groups over* R *as defined in* 4.1 *and the category of V-flat V-reduced left* $\mathrm{Cart}_p(R)$*-modules,*

defined as follows.

$$\begin{array}{ccc} \{\text{smooth formal groups over } R\} & \xrightarrow{\sim} & \{V\text{-flat } V\text{-reduced left } \mathrm{Cart}_p(R)\text{-mod}\} \\ G & \longmapsto & \mathrm{M}_p(G) = \epsilon_p \,\mathrm{Hom}(\Lambda, G) \\ \Lambda_p \otimes_{\mathrm{Cart}_p(R)} M & \longmapsfrom & M \end{array}$$

Dieudonné modules.

In the rest of this section, K stands for a perfect field of characteristic $p > 0$. We have $FV = VF = p$ in $\mathrm{Cart}_p(K)$. It is well known that the ring of p-adic Witt vectors $W(K)$ is a complete discrete valuation ring with residue field K, whose maximal ideal is generated by p. Denote by $\sigma : W(K) \to W(K)$ the Teichmüller lift of the automorphism $x \mapsto x^p$ of K. With the Witt coordinates we have $\sigma : (c_0, c_1, c_2, \ldots) \mapsto (c_0^p, c_1^p, c_2^p, \ldots)$. Denote by $L = B(K)$ the field of fractions of $W(K)$.

Definition 4.28. Denote by R_K the (non-commutative) ring generated by $W(K)$, F and V, subject to the following relations

$$F \cdot V = V \cdot F = p, \quad F \cdot x = {}^{\sigma}x \cdot F, \quad x \cdot V = V \cdot {}^{\sigma}x \qquad \forall x \in W(K)\,.$$

Remark. There is a natural embedding $R_K \hookrightarrow \mathrm{Cart}_p(K)$; we use it to identify R_K as a dense subring of the Cartier ring $\mathrm{Cart}_p(K)$. For every continuous left $\mathrm{Cart}_p(K)$-module M, the $\mathrm{Cart}_p(K)$-module structure on M is determined by the induced left R_K-module structure on M.

Lemma/Exercise.
(i) The ring R_K is naturally identified with the ring

$$W(K)[V, F] := \left(\bigoplus_{i<0} p^{-i} V^i W(K)\right) \bigoplus \left(\bigoplus_{i\geq 0} V^i W(K)\right),$$

i.e., elements of $W(K)[V, F]$ are sums of the form $\sum_{i\in\mathbb{Z}} a_i V^i$, where $a_i \in L$ for all $i \in \mathbb{Z}$, $\mathrm{ord}_p(a_i) \geq \max(0, -i)$ $\forall i \in \mathbb{Z}$, and $a_i = 0$ for all but finitely many i's. The commutation relation between $W(K)$ and V^i is

$$x \cdot V^i = V^i \cdot {}^{\sigma^i}x \quad \text{for all } x \in W(K) \text{ and all } i \in \mathbb{Z}\,.$$

(ii) The ring $\mathrm{Cart}_p(K)$ is naturally identified with the set $W(K)[[V, F\rangle\rangle$, consisting of all non-commutative formal power series of the form $\sum_{i\in\mathbb{Z}} a_i V^i$ such that $a_i \in L$ $\forall i \in \mathbb{Z}$, $\mathrm{ord}_p(a_i) \geq \max(0, -i)$ $\forall i \in \mathbb{Z}$, and $\mathrm{ord}_p(a_i) + i \to \infty$ as $|i| \to \infty$.
(iii) Check that the ring structure on $W(K)[V, F]$ extends to $W(K)[[V, F\rangle\rangle$ by continuity. In other words, the inclusion $W(K)[V, F] \hookrightarrow W(K)[[V, F\rangle\rangle$ is a ring homomorphism, and $W(K)[V, F]$ is dense in $W(K)[[V, F\rangle\rangle$ with respect to the V-adic topology on $W(K)[[V, F\rangle\rangle \cong \mathrm{Cart}_p(K)$. The latter topology on $W(K)[[V, F\rangle\rangle$ is equivalent to the topology given by the discrete valuation v on $W(K)[[V, F\rangle\rangle$ defined by

$$v\left(\sum_{i\in\mathbb{Z}} a_i V^i\right) = \mathrm{Min}\,\{\mathrm{ord}_p(a_i) + i \,|\, i \in \mathbb{Z}\}\,.$$

Definition 4.29.

(1) A *Dieudonné module* is a left R_K-module M such that M is a free $W(K)$-module of finite rank.

(2) Let M be a Dieudonné module over K. Define the α-rank of M to be the natural number $a(M) = \dim_K(M/(VM + FM))$.

Compare with $a(G)$ as defined in 5.4.

Definition 4.30.

(i) For any natural number $n \geq 1$ and any scheme S, denote by $\underline{(\mathbb{Z}/n\mathbb{Z})}_S$ the constant group scheme over S attached to the finite group $\mathbb{Z}/n\mathbb{Z}$. The scheme underlying $\underline{(\mathbb{Z}/n\mathbb{Z})}_S$ is the disjoint union of n copies of S, indexed by the finite group $\mathbb{Z}/n\mathbb{Z}$, see 10.22.

(ii) For any natural number $n \geq 1$ and any scheme S, denote by $\mu_{n,S}$ the kernel of $[n] : \mathbb{G}_{m/S} \to \mathbb{G}_{m/S}$. The group scheme $\mu_{n,S}$ is finite and locally free over S of rank n; it is the Cartier dual of $\underline{(\mathbb{Z}/n\mathbb{Z})}_S$.

(iii) For any field $K \supset \mathbb{F}_p$, define a finite group scheme α_p over K to be the kernel of the endomorphism

$$\mathrm{Fr}_p : \mathbb{G}_{a/K} = \mathrm{Spec}(K[X]) \to \mathbb{G}_{a/K} = \mathrm{Spec}(K[X])$$

of $\mathbb{G}_a$ over K defined by the K-homomorphism from the K-algebra $K[X]$ to itself which sends X to X^p. We have $\alpha_p = \mathrm{Spec}(K[X]/(X^p))$ as a scheme. The comultiplication on the coordinate ring of α_p is induced by $X \mapsto X \otimes X$.

Proposition 4.31. **BB** *Let X be a p-divisible group over a perfect field $K \supset \mathbb{F}_p$. Then there exists a canonical splitting*

$$X \cong X_{\mathrm{tor}} \times_{\mathrm{Spec}(K)} X_{\ell\ell} \times_{\mathrm{Spec}(K)} X_{\text{ét}}$$

where $X_{\text{ét}}$ is the maximal étale quotient of X, X_{mult} is the maximal toric p-divisible subgroup of X, and $X_{\ell\ell}$ is a p-divisible group with no non-trivial étale quotient nor non-trivial multiplicative p-divisible subgroup.

Remark.

(i) The analogous statement for finite group schemes over K can be found in [**48**, Chapter 1], from which 4.31 follows. See also [**24**], [**25**].

(ii) See 10.9 for a similar statement for p-divisible groups over an Artinian local ring.

Definition 4.32. Let m, n be non-negative integers such that $\gcd(m, n) = 1$. Let $k \supset \mathbb{F}_p$ be an algebraically closed field. Let $G_{m,n}$ be the p-divisible group whose Dieudonné module is

$$\mathbb{D}(G_{m,n}) = R_K/R_K \cdot (V^n - F^m)\,.$$

Theorem 4.33. **BB**

(1) *There is an equivalence of categories between the category of p-divisible groups over K and the category of Dieudonné modules over R_K. Denote by $\mathbb{D}(X)$ the covariant Dieudonné module attached to a p-divisible group over K. This equivalence is compatible with direct product and exactness, i.e., short exact sequences correspond under the above equivalence of categories.*

(2) *Let X be a p-divisible group over K such that X is a p-divisible formal group in the sense that the maximal étale quotient of X is trivial. Denote by $X^\wedge$ the formal group attached to X, i.e., $X^\wedge$ is the formal completion of X along the zero section of X. Then there is a canonical isomorphism*

$\mathbb{D}(X) \xrightarrow{\sim} \mathrm{M}_p(X^\wedge)$ *between the Dieudonné module of* X *and the Cartier module of* $X^\wedge$ *which is compatible with the actions by* F, V *and elements of* $W(K)$.

(3) *Let* X *be a* p*-divisible group over* K *and* $\mathbb{D}(X)$ *the covariant Dieudonné module of* X. *Then* $\mathrm{ht}(X) = \mathrm{rank}_{W(K)}(\mathbb{D}(X))$, *and we have a functorial isomorphism* $\mathrm{Lie}(X) \cong \mathbb{D}(X)/V \cdot \mathbb{D}(X)$.

(4) *Let* X^t *be the Serre-dual of the* p*-divisible group of* X. *Then the Dieudonné module* $\mathbb{D}(X^t)$ *can be described in terms of* $\mathbb{D}(X)$ *as follows. The underlying* $W(K)$*-module is the linear dual* $\mathbb{D}(X)^\vee := \mathrm{Hom}_{W(K)}(\mathbb{D}(X), W(K))$ *of* $\mathbb{D}(X)$. *The actions of* V *and* F *on* $\mathbb{D}(X)^\vee$ *are defined as follows:*

$$(V \cdot h)(m) \;=\; {}^{\sigma^{-1}}(h(Fm)), \quad (F \cdot h)(m) \;=\; {}^{\sigma}(h(Vm))$$

for all $h \in \mathbb{D}(X)^\vee = \mathrm{Hom}_{W(K)}(\mathbb{D}(X), W(K))$ *and all* $m \in \mathbb{D}(X)$.

(5) *A* p*-divisible group* X *over* K *is étale if and only if* $V : \mathbb{D}(X) \to \mathbb{D}(X)$ *is bijective, or equivalently,* $F : \mathbb{D}(X) \to \mathbb{D}(X)$ *is divisible by* p. *A* p*-divisible group* X *over* K *is multiplicative if and only if* $V : \mathbb{D}(X) \to \mathbb{D}(X)$ *is divisible by* p, *or equivalently,* $F : \mathbb{D}(X) \to \mathbb{D}(X)$ *is bijective. A* p*-divisible group* X *over* K *has no non-trivial étale quotient nor non-trivial multiplicative* p*-divisible subgroup if and only if both* F *and* V *are topologically nilpotent on* $\mathbb{D}(X)$.

Remark 4.34.

(1) See [**57**] for Theorem 4.33.

(2) When $p > 2$, the Dieudonné module $\mathbb{D}(X)$ attached to a p-divisible group over K can also be defined in terms of the covariant Dieudonné crystal attached to X described in 2.4. In short, $\mathbb{D}(X)$ "is" $\mathbb{D}(X/W(K))_{W(K)}$, the limit of the "values" of the Dieudonné crystal at the divided power structures

$$(W(K)/p^m W(K), pW(K)/p^m W(K), \gamma)$$

as $m \to \infty$, where these are the reductions modulo p^m of the natural DP-structure on $(W(K), pW(K))$. Recall that the natural DP-structure on $(W(K), pW(K))$ is given by $\gamma_i(x) = \frac{x^i}{i!} \quad \forall\, x \in pW(K)$; the condition that $p > 2$ implies that the induced DP structure on

$$(W(K)/p^m W(K), pW(K)/p^m W(K)$$

is nilpotent.

Proposition 4.35. **BB** *Let* X *be a* p*-divisible group over* K. *We have a natural isomorphism*

$$\mathrm{Hom}_K(\alpha_p, X[p]) \cong \mathrm{Hom}_{W(K)}\left(\mathbb{D}(X)^\vee/(V\mathbb{D}(X)^\vee + F\mathbb{D}(X)^\vee), B(K)/W(K)\right),$$

where $B(K) = \mathrm{frac}(W(K))$ *is the fraction field of* $W(K)$. *In particular we have*

$$\dim_K(\mathrm{Hom}_K(\alpha_p, X[p])) = a(\mathbb{D}(X)).$$

The natural number $a(\mathbb{D}(X))$ *of a* p*-divisible group* X *over* K *is zero if and only if* X *is an extension of an étale* p*-divisible group by a multiplicative* p*-divisible group.*

For the notation $a(M)$ see 4.29, and for $a(G)$ see 5.4.

Exercise 4.36.

(i) Prove that $\mathrm{ht}(G_{m,n}) = m + n$.

(ii) Prove that $\dim(G_{m,n}) = m$.
(iii) Show that $G_{0,1}$ is isomorphic to the étale p-divisible group $\mathbb{Q}_p/\mathbb{Z}_p$, and $G_{1,0}$ is isomorphic to the multiplicative p-divisible group $\mu_\infty = \mathbb{G}_m[p^\infty]$.
(iv) Show that $\mathrm{End}(G_{m,n}) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$ is a central division algebra over $\mathbb{Q}_p$ of dimension $(m+n)^2$, and compute the Brauer invariant of this central division algebra.
(iv) Relate $G_{m,n}$ to $G_{n,m}$.
(v) Determine all pairs (m,n) such that $\mathrm{End}(G_{m,n})$ is the maximal order of the division algebra $\mathrm{End}(G_{m,n}) \otimes_{\mathbb{Z}_p} \mathbb{Q}_p$.

Theorem 4.37. **BB** *Let $k \supset \mathbb{F}_p$ be an algebraically closed field. Let X be a simple p-divisible group over k, i.e., X has no non-trivial quotient p-divisible groups. Then X is isogenous to $G_{m,n}$ for a uniquely determined pair of natural numbers m,n with $\gcd(m,n) = 1$, i.e., there exists a surjective homomorphism $X \to G_{m,n}$ with finite kernel.*

Definition 4.38.

(i) The slope of $G_{m,n}$ is $m/(m+n)$ with multiplicity $m+n$. The Newton polygon of $G_{m,n}$ is the line segment in the plane from $(0,0)$ to $(m+n,m)$. The slope sequence of $G_{m,n}$ is the sequence $(m/(m+n), \dots, m/(m+n))$ with $m+n$ entries.
(ii) Let X be a p-divisible group over a field $K \supset \mathbb{F}_p$, and let k be an algebraically closed field containing K. Suppose that X is isogenous to

$$G_{m_1,n_1} \times_{\mathrm{Spec}(k)} \cdots \times_{\mathrm{Spec}(k)} G_{m_r,n_r}$$

$\gcd(m_i,n_i) = 1$ for $i = 1,\dots,r$, and $m_i/(m_i+n_i) \le m_{i+1}/(m_{i+1}+n_{i+1})$ for $i = 1,\dots,r-1$. Then the Newton polygon of X is defined by the data $\sum_{i=1}^r (m_i,n_i)$. Its slope sequence is the concatenation of the slope sequence for $G_{m_1,n_1},\dots,G_{m_r,n_r}$.

Example. A p-divisible group X over K is étale (resp. multiplicative) if and only if all of its slopes are equal to 0 (resp. 1).

Exercise 4.39. Suppose that X is a p-divisible group over K such that X is isogenous to $G_{1,n}$ (resp. $G_{m,1}$). Show that X is isomorphic to $G_{1,n}$ (resp. $G_{m,1}$).

Exercise 4.40. Let $\beta_1 \le \cdots \le \beta_h$ be the slope sequence of a p-divisible group over K of height h. Prove that the slope sequence of the Serre dual X^t of X is $1-\beta_h,\dots,1-\beta_1$. (HINT: First show that $G_{m,n}^t \cong G_{n,m}$.)

Conclusion 4.41. Let $K \supset \mathbb{F}_p$ be a field, and let k be an algebraically closed field containing K.

- Any p-divisible group X over K admits an isogeny $X \otimes k \sim \prod_i G_{m_i,n_i}$.
- The Newton polygon $\mathcal{N}(G_{m,n})$ is isoclinic (all slopes are the same) of height $m+n$ and slope $m/(m+n)$.
- In this way the Newton polygon $\mathcal{N}(X)$ is determined. Write $h = h(X)$ for the height of X and $d = \dim(X)$ for the dimension of X. The Newton polygon $\mathcal{N}(X)$ ends at $(h(X), \dim(X))$.
- The isogeny class of a p-divisible group over any algebraically closed field k uniquely determines (and is uniquely determined by) its Newton polygon:

Theorem 4.42. (Dieudonné and Manin, see [**48**], "Classification theorem" on page 35)

$$\{p\text{-divisible groups } X \text{ over } k\}/\sim_k \quad \xrightarrow{\sim} \quad \{\text{Newton polygon}\}$$

In words, p-divisible groups over an algebraically closed field $k \supset \mathbb{F}_p$ are classified up to isogeny by their Newton polygons.

Exercise 4.43. Show that there are infinitely many non-isomorphic p-divisible groups with slope sequence $(1/2, 1/2, 1/2, 1/2)$ (resp. $(1/3, 1/3, 1/3, 2/3, 2/3, 2/3)$ over any infinite perfect field $K \supset \mathbb{F}_p$.

Exercise 4.44. Determine all Newton polygons attached to a p-divisible group of height 6, and the symmetric Newton polygons among them.

Exercise 4.45. Recall that the set of all Newton polygons is partially ordered; $\zeta_1 \prec \zeta_2$ if and only if ζ_1, ζ_2 have the same end points, and ζ_2 lies below ζ_1. Show that this poset is *ranked*, i.e., any two maximal chains between two elements of this poset have the same length.

Exercise 4.46. Prove the equivalence of the statements in 10.10 that characterize ordinary abelian varieties.

Theorem 4.47. **BB** *Let S be a scheme such that p is locally nilpotent in $\mathcal{O}_S$. Let $X \to S$ be a p-divisible group over S. Suppose that a point s is a specialization of a point $s' \in S$. Let $\mathcal{N}(X_s)$ and $\mathcal{N}(X_{s'})$ be the Newton polygon of the fibers X_s and $X_{s'}$ of X, respectively. Then $\mathcal{N}(X_s) \prec \mathcal{N}(X_{s'})$, i.e., the Newton polygon $\mathcal{N}(X_s)$ of the specialization lies above (or is equal to) the Newton polygon $\mathcal{N}(X_{s'})$.*

This result first appeared in a letter from Grothendieck to Barsotti dated May 11, 1970; see the Appendix in [**34**]. See [**40**], 2.3.2 for the proof of a stronger result, that the locus in the base scheme S with Newton polygon $\prec \xi$ is closed for any Newton polygon ξ; see 1.19, 1.20.

Exercise 4.48.

(i) Construct an example of a specialization of ordinary p-divisible group of dimension 3 and height 6 to a p-divisible group with slopes 1/3 and 2/3, using the theory of Cartier modules.

(ii) Construct an example of a non-constant p-divisible group with constant slope.

4.49. Here is an explicit description of the Newton polygon of an abelian variety A over a finite field $\mathbb{F}_q \supset \mathbb{F}_p$. We may and do assume that A is simple over $\mathbb{F}_q$. By Tate's Theorem 10.17, we know that the abelian variety A is determined by its q-Frobenius π_A up to $\mathbb{F}_q$-rational isogeny. Then the slopes of A are determined by the p-adic valuations of π_A as follows. For every rational number $\lambda \in [0, 1]$, the multiplicity of the slope λ in the Newton polygon of A is

$$\sum_{v \in \varphi_{p,\lambda}} [\mathbb{Q}(\pi_A)_v : \mathbb{Q}_p]$$

where $\varphi_{p,\lambda}$ is the finite set consisting of all places v of $\mathbb{Q}(\pi_A)$ above p such that $v(\pi_A) = \lambda \cdot v(q)$.

Exercise 4.50. Let E be an ordinary elliptic curve over $\mathbb{F}$. Let $L = \mathrm{End}(E)^0$. Show that L is an imaginary quadratic field which is split at p and $\mathrm{End}(E) \otimes_{\mathbb{Z}} \mathbb{Z}_p \cong \mathbb{Z}_p \times \mathbb{Z}_p$.

Exercise 4.51. Let L be an imaginary quadratic field which is split at p. Let r, s be positive rational numbers such that $\gcd(r, s) = 1$ and $2r < s$. Use Tate's

Theorem 10.17 to show that there exists a simple s-dimensional abelian variety A over a finite field $\mathbb{F}_q \supset \mathbb{F}_p$ such that $\mathrm{End}(A)^0 = L$ and the slopes of the Newton polygon of A are $\frac{r}{s}$ and $\frac{s-r}{s}$.

Exercise 4.52. Let m, n be positive integers with $\gcd(m,n) = 1$, and let $h = m+n$. Let D be a central division algebra over $\mathbb{Q}_p$ of dimension h^2 with Brauer invariant n/h. This means that there exists a homomorphism $j : \mathrm{frac}(W(\mathbb{F}_{p^h})) \to D$ of $\mathbb{Q}_p$-algebras and an element $u \in D^\times$ with $\frac{\mathrm{ord}(u)}{\mathrm{ord}(p)} \equiv \frac{n}{h} \pmod{\mathbb{Z}}$ such that

$$u \cdot j(x) \cdot u^{-1} = j(\sigma(x)) \quad \forall x \in W(\mathbb{F}_{p^h})\,.$$

Here ord denotes the normalized valuation on D, and σ is the canonical lifting of Frobenius on $W(\mathbb{F}_{p^h})$. Changing u by a suitable power of p, we may and do assume that $u,\ pu^{-1} \in \mathcal{O}_D$, where $\mathcal{O}_D$ is the maximal order of D.

Let M be the left $W(\mathbb{F}_{p^h})$-module underlying $\mathcal{O}_D$, where the left $W(\mathbb{F}_{p^h})$-module structure is given by left multiplication with elements of $j(W(\mathbb{F}_{p^h}))$. Let $F : M \to M$ be the operator $z \mapsto u \cdot z$, and let $V : M \to M$ be the operator $z \mapsto pu^{-1} \cdot z$. This makes M a module over the Dieudonné ring $R_{\mathbb{F}_{p^h}}$.

(1) Show that right multiplication by elements of $\mathcal{O}_D$ induces an isomorphism

$$\mathcal{O}_D^{\mathrm{opp}} \xrightarrow{\sim} \mathrm{End}^0_{R_{\mathbb{F}_{p^h}}}(M)\,.$$

(2) Show that $\mathcal{O}_D^{\mathrm{opp}} \xrightarrow{\sim} \mathrm{End}^0_{R_K}(W(K) \otimes_{W(\mathbb{F}_{p^h})} M)$ for every perfect field $K \supset \mathbb{F}_{p^h}$.

(3) Show that there exists a $W(\mathbb{F}_{p^h})$-basis of M $e_0, e_1, \ldots, e_{h-1}$ such that if we extend $e_0, e_1, \ldots, e_{h-1}$ to a cyclic sequence $(e_i)_{\in\mathbb{Z}}$ by the condition that $e_{i+h} = p \cdot e_i$ for all $i \in \mathbb{Z}$, we have

$$F{\cdot}e_i = e_{i+n}, \quad V{\cdot}e_i = e_{i+m}, \quad p{\cdot}e_i = e_{i+m+n}, \quad M_{m,n} := \oplus_{0\le i<m+n}\, W{\cdot}e_i.$$

(4) Show that the p-divisible group $H_{m,n}$ corresponding to the Dieudonné module M has dimension m and slope m/h.

(5) Suppose that X is a p-divisible group over a perfect field $K \supset \mathbb{F}_p$ such that $\mathrm{End}^0(X)$ is isomorphic to $\mathcal{O}_D^{\mathrm{opp}}$. Show that $K \supset \mathbb{F}_{p^h}$ and X is isomorphic to $H_{m,n} \times_{\mathrm{Spec}(\mathbb{F}_{p^h})} \mathrm{Spec}(K)$.

Proposition 4.53. **BB**

(i) *There is an equivalence of categories between the category of finite group schemes over the perfect base field K and the category of left R_K-modules which are $W(K)$-modules of finite length. Denote by $\mathbb{D}(G)$ the left R_K-module attached to a finite group scheme G over K.*

(ii) *Suppose that $0 \to G \to X \xrightarrow{\beta} Y \to 0$ is a short exact sequence, where G is a finite group scheme over K, and $\beta : X \to Y$ is an isogeny between p-divisible groups over K. Then we have a natural isomorphism*

$$\mathbb{D}(G) \xrightarrow{\sim} \mathrm{Ker}\left(\mathbb{D}(X) \otimes_{W(K)} B(K)/W(K) \xrightarrow{\beta} \mathbb{D}(Y) \otimes_{W(K)} B(K)/W(K)\right)$$

of left R_K-modules.

Remark.

(i) We say that $\mathbb{D}(G)$ is the covariant Dieudonné module of G, abusing the terminology, because $\mathbb{D}(G)$ is not a free $W(K)$-module.

(ii) Proposition 4.53 is a covariant version of the classical contravariant Dieudonné theory in [**24**] and [**48**]. See also [**60**].

4.54. Remarks on the operators F and V. For group schemes in characteristic p we have the Frobenius homomorphism, and for commutative group schemes the Verschiebung; see 10.23, 10.24. Also for Dieudonné modules such homomorphisms are studied. However, some care has to be taken. In the *covariant* Dieudonné module theory the Frobenius homomorphism on commutative group schemes corresponds to the operator V on the related modules, and the Verschiebung homomorphism on commutative flat group schemes gives the operator F on modules; for details see [**67**], 15.3. In case confusion is possible we write F (resp. V) for the Frobenius (resp. Verschiebung) homomorphism on group schemes and $\mathcal{V} = \mathbb{D}(F)$ (resp. $\mathcal{F} = \mathbb{D}(V)$) for the corresponding operator on modules,

5. Cayley-Hamilton: a conjecture by Manin and the weak Grothendieck conjecture

Main reference: [**65**].

5.1. In this section we develop non-commutative generalizations of the Cayley Hamilton theorem: a matrix $\mathcal{F}$ is a zero of its own characteristic polynomial.

Exercise 5.2. Prove the classical Cayley-Hamilton theorem for a matrix over a commutative ring R: let X be an $n \times n$ matrix with entries in R with characteristic polynomial $g(T) = \mathrm{Det}(X - T \cdot \mathbf{1}_n) \in R[T]$; then the matrix $g(X)$ is the zero matrix.

Here are some suggestions for a proof:

(a) For any commutative ring R, and an $n \times n$ matrix X with entries in R there exists a ring homomorphism $h : \mathbb{Z}[t_{1,1}, \cdots, t_{i,j}, \cdots, t_{n,n}] \to R$ such that the matrix $(t) = (t_{i,j} \mid 1 \leq i, j \leq n)$ is mapped to X.

(b) Let $h^{\sim} : R[T] \to \mathbb{Z}[t_{ij}][T]$ be the ring homomorphism induced by h. Let $G(T) = \mathrm{Det}((t) - T \cdot \mathbf{1}_n) \in \mathbb{Z}[t_{ij}][T]$, so that $h^{\sim}(G) = g$. Conclude that it suffices to prove the statement for a commutative ring that contains $\mathbb{Z}[t_{1,1}, \ldots, t_{i,j}, \ldots, t_{n,n}]$.

(c) Construct $\mathbb{Z}[t_{i,j} \mid 1 \leq i, j \leq n] \hookrightarrow \mathbb{C}$, and apply the classical Cayley-Hamilton theorem for $\mathbb{C}$ (which is a consequence of the theorem of canonical forms). Alternatively, show that the matrix (t) considered over $\mathbb{C}$ has mutually different eigenvalues.

Here are suggestions for a different proof:

(1) Show it suffices to prove this for an algebraically closed field of characteristic zero.

(2) Show that the classical Cayley-Hamilton theorem holds for a matrix which is in diagonal form with all diagonal elements mutually different.

(3) Show that the set of all conjugates of matrices as in (2) is Zariski dense in $\mathrm{Mat}(n \times n)$. Finish the proof.

5.3. We will develop a useful analog of this Cayley-Hamilton theorem over the Dieudonné ring. Note that over a non-commutative ring there is no reason that any straightforward analog of Cayley-Hamilton should be true. However, given a specific element in a special situation, we construct an operator $g(\mathcal{F})$ which annihilates that specific element in the Dieudonné module. Warning: In general $g(\mathcal{F})$ does not annihilate all elements of the Dieudonné module.

Notation 5.4. Let G be a group scheme over a field $K \supset \mathbb{F}_p$. Consider $\alpha_p = \mathrm{Ker}(F : \mathbb{G}_a \to \mathbb{G}_a)$. Choose a perfect field L containing K. Note that $\mathrm{Hom}(\alpha_p, G_L)$ is a right module over $\mathrm{End}(\alpha_p \otimes_{\mathbb{F}_p} L) = L$. We define

$$a(G) = \dim_L \left(\mathrm{Hom}(\alpha_p, G_L)\right).$$

Remarks. For any field L we write α_p instead of $\alpha_p \otimes_{\mathbb{F}_p} L$ if no confusion is likely.

The group scheme $\alpha_{p,K}$ over a field K corresponds under 5.7 (in any case) or by Dieudonné theory (in case K is perfect) to the module K^+ with operators $\mathcal{F} = 0$ and $\mathcal{V} = 0$.

If K is not perfect it might happen that

$$\dim_K \left(\mathrm{Hom}(\alpha_p, G)\right) < \dim_L \left(\mathrm{Hom}(\alpha_p, G_L)\right);$$

see the Exercise 5.8 below.

However, if L is perfect and $L \subset L'$ is any field extension then

$$\dim_L \left(\mathrm{Hom}(\alpha_p, G_L)\right) = \dim_{L'} \left(\mathrm{Hom}(\alpha_p, G_{L'})\right).$$

Hence the definition of $a(G)$ is independent of the chosen perfect extension L.

Exercise 5.5.

(i) Let N be a finite group scheme over a perfect field K. Assume that F and V on N are nilpotent on N, and suppose that $a(N) = 1$. *Show that the Dieudonné module $\mathbb{D}(N)$ is generated by one element over the Dieudonné ring.*

(ii) Let A be an abelian variety over a perfect field K. Assume that the p-rank of A is zero, and that $a(A) = 1$. *Show that the Dieudonné module $\mathbb{D}(A[p^\infty])$ is generated by one element over the Dieudonné ring.*

Remark. We will see that if $a(X_0) = 1$, then the Newton polygon stratum $\mathcal{W}_{\mathcal{N}(X_0)}(\mathrm{Def}(X_0))$ in $\mathcal{D}(X_0)$ *is non-singular.* See 1.19 and 5.11. Similarly, let (A, λ) be a principally polarized abelian variety, $\xi = \mathcal{N}(A)$. The Newton polygon stratum $\mathcal{W}_\xi(\mathcal{A}_{g,1,n})$ will be shown to be regular at the point (A, λ) (here we work with a fine moduli scheme: assume $n \geq 3$). In the above $\mathcal{W}_\xi(\mathcal{A}_{g,1,n})$ denotes the locus in $\mathcal{A}_{g,1,n}$ with Newton polygon $\prec \xi$ (i.e., lying above ξ); similarly for $\mathcal{W}_{\mathcal{N}(X_0)}(\mathcal{D}(X_0))$; see 1.19, 5.11.

We see that we can a priori consider a set of points where the Newton polygon stratum is guaranteed to be non-singular. That is the main result of this section. Then, in Section 7 we show that such points are dense in both cases considered, p-divisible groups and *principally* polarized abelian varieties.

We give an example, with K non-perfect, where $\dim_K \left(\mathrm{Hom}(\alpha_p, G)\right) < a(G)$; we see that the condition "L is perfect" is necessary in 5.4 .

5.6. "Dieudonné modules" over non-perfect fields? This is a difficult topic. However, in one special case statements and results are easy.

***p*-Lie algebras.** Basic reference [**25**]. We will need this theory only in the commutative case. For more general statements see [**25**], II.7.

Let $K \supset \mathbb{F}_p$ be a field. A commutative finite group scheme of *height one* over K is a finite commutative group scheme N over K such that $(F : N \to N^{(p)}) = 0$, the zero map. Denote the category of such objects by GF_K.

A commutative finite dimensional p-Lie algebra M over K is a pair (M, g), where M is a finite dimensional vector space over K, and $g : M \to M$ is a homomorphism of additive groups with the property

$$g(b{\cdot}x) \quad = \quad b^p{\cdot}g(x).$$

Denote the category of such objects by Liep_K.

Theorem 5.7. **BB** *There is an equivalence of categories*

$$\mathcal{D}_K : \mathrm{GF}_K \xrightarrow{\sim} \mathrm{Liep}_K.$$

This equivalence commutes with base change. If K is a perfect field this functor coincides with the Dieudonné module functor: $\mathcal{D}_K = \mathbb{D}$, and the operator g on $\mathcal{D}_K(N)$ corresponds to the operator F on $\mathbb{D}(N)$ for every $N \in \mathrm{Ob}(\mathrm{GF})_K$.

See [**25**], II.7.4

Exercise.

(i) Classify all commutative group schemes of rank p over k, an algebraically closed field of characteristic p.
(ii) Classify all commutative group schemes of rank p over a perfect field $K \supset \mathbb{F}_p$.

Remark/Exercise 5.8.

(1) Let K be a non-perfect field, with $b \in K$ and $\sqrt[p]{b} \notin K$. Let (M, g) be the commutative finite dimensional p-Lie algebra defined by:

$$M = K{\cdot}x \oplus K{\cdot}y \oplus K{\cdot}z, \quad g(x) = bz,\ g(y) = z,\ g(z) = 0.$$

Let N be the finite group scheme of height one defined by this p-Lie algebra, i.e., such that $\mathcal{D}_K(N) = (M, g)$, see 5.6. *Show*:

$$\dim_K \left(\mathrm{Hom}(\alpha_p, N)\right) = 1, \quad \dim_k \left(\mathrm{Hom}(\alpha_p, N \times_{\mathrm{Spec}(K)} \mathrm{Spec}(k))\right) = 2,$$

where $k = k^{\mathrm{alg}} \supset K$.

(2) Let $N_2 = W_2[F]$ be the kernel of $F : W_2 \to W_2$ over $\mathbb{F}_p$; here W_2 is the 2-dimensional group scheme of Witt vectors of length 2. In fact one can define N_2 by $\mathbb{D}(N_2) = \mathbb{F}_p{\cdot}r \oplus \mathbb{F}_p{\cdot}s$, $\mathcal{V}(r) = 0 = \mathcal{V}(s) = \mathcal{F}(s)$ and $\mathcal{F}(r) = s$. Let $L = K(\sqrt[p]{b})$. *Show that*

$$N \ncong_K (\alpha_p \oplus W_2[F]) \otimes K \quad \text{and} \quad N \otimes L \cong_L (\alpha_p \oplus W_2[F]) \otimes L.$$

Remark. In [**45**], I.5 Definition (1.5.1) should be given over a perfect field K. We thank Chia-Fu Yu for drawing our attention to this flaw.

5.9. We fix integers $h \geq d \geq 0$, and we write $c := h - d$. We consider Newton polygons ending at (h, d). For such a Newton polygon β we write

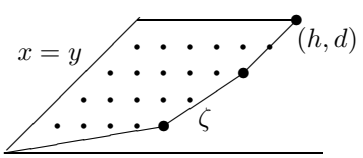
$$\diamond(\beta) := \{(x, y) \in \mathbb{Z} \times \mathbb{Z} \mid y < d, \quad y < x, \quad (x, y) \prec \beta\};$$

here we denote by $(x, y) \prec \beta$ the property "(x, y) is on or above β"; we write

$$\boxed{\dim(\zeta) := \#(\diamond(\zeta)).}$$

Let $\diamond = \{(x, y) \in \mathbb{Z} \times \mathbb{Z} \mid 0 \leq y < d, \quad y < x \leq y + d\}$.

Example.



$$\zeta = 2\times(1,0)+(2,1)+(1,5) =$$
$$= 6\times\tfrac{1}{6}+3\times\tfrac{2}{3}+2\times\tfrac{1}{1};\quad h=11.$$
Here $\dim(\zeta)=\#(\diamond(\zeta))=22$.

Note that for $\rho = d\cdot(1,0)+c\cdot(0,1)$ we have $\dim(\rho)=dc$.

Theorem 5.10. (Newton polygon strata for p-divisible groups) *Suppose* $a(X_0)\leq 1$. *Write* $D=\mathcal{D}(X_0)$. (For notation see 10.21.) *For every* $\beta\succ\gamma=\mathcal{N}(X_0)$, *the Newton polygon stratum:* $\mathcal{W}_\beta(D)$ *is formally smooth and* $\dim(\mathcal{W}_\beta(D))=\dim(\beta)$. *The strata* $\mathcal{W}_\beta(D)$ *are nested as given by the partial ordering on Newton polygons, i.e.,*

$$\mathcal{W}_\beta(D)\subset\mathcal{W}_\delta(D)\quad\Longleftrightarrow\quad\diamond(\beta)\subset\diamond(\delta)\quad\Longleftrightarrow\quad\beta\prec\delta.$$

Generically on $\mathcal{W}_\beta(D)$ *the fibers have Newton polygon equal to* β.

For the notion "generic" for a p-divisible group over a formal scheme, see 10.21.

5.11. In fact, this can be visualized and made more precise as follows. Choose variables $T_{r,s}$, with $1\leq r\leq d=\dim(X_0)$, $1\leq s\leq h=\text{height}(X_0)$ and write these in a diagram

$$\begin{array}{ccccccccc}
 & & & & & & 0 & \cdots & 0 & -1\\
 & & & & & T_{d,h} & \cdot & \cdots & T_{1,h} & \\
 & & & & \vdots & \vdots & \vdots & \cdot & & \\
 & T_{d,d+2} & \cdots & T_{i,d+2} & \cdots & T_{2,d+2} & T_{1,d+2} & & & \\
T_{d,d+1} & \cdots & T_{i,d+1} & \cdots & \cdots & T_{1,d+1} & & & &
\end{array}$$

We show that

$$\mathcal{D}^\wedge=\text{Def}(X_0)=\text{Spf}(k[[Z_{(x,y)}\mid(x,y)\in\diamond]]),\qquad T_{r,s}=Z_{(s-r,s-1-d)}.$$

Moreover, for any $\beta\succ\mathcal{N}(G_0)$ we write

$$R_\beta=\frac{k[[Z_{(x,y)}\mid(x,y)\in\diamond]]}{(Z_{(x,y)}\ \ \forall(x,y)\notin\diamond(\beta))}\quad\cong\quad k[[Z_{(x,y)}\mid(x,y)\in\diamond(\beta)]].$$

Claim.

$$(\text{Spec}(R_\beta)\subset\text{Spec}(R))\quad=\quad(\mathcal{W}_\beta(D)\subset\mathcal{D}).$$

Clearly this claim proves the theorem. We will give a proof of the claim, and hence of this theorem by using the theory of displays and the following tools.

Convention. Let d,c be non-negative integers, and let $h=c+d$. For any $h\times h$ matrix

$$(a)=\begin{pmatrix}A & B\\ C & D\end{pmatrix},$$

its associated F-matrix is

$$(\mathcal{F})=(pa):=\begin{pmatrix}A & pB\\ C & pD\end{pmatrix},$$

where A is a $d\times d$ matrix, B is a $d\times c$ matrix, C is a $c\times d$ matrix and D is a $c\times c$ matrix.

Definition 5.12. We consider matrices which can appear as F-matrices associated with a display. Let $d, c \in \mathbb{Z}_{\geq 0}$, and $h = d + c$. Let W be a ring. We say that a display matrix $(a_{i,j})$ of size $h \times h$ is in *normal form* form over W if the F-matrix is of the following form:

$$\left(\begin{array}{cccccccccc}
0 & 0 & \cdots & 0 & a_{1d} & pa_{1,d+1} & \cdots & \cdots & \cdots & pa_{1,h} \\
1 & 0 & \cdots & 0 & a_{2d} & \cdots & & pa_{i,j} & & \cdots \\
0 & 1 & \cdots & 0 & a_{3d} & & & 1 \leq i \leq d & & \\
\vdots & \vdots & \ddots & \ddots & \vdots & & & d \leq j \leq h & & \\
0 & 0 & \cdots & 1 & a_{dd} & pa_{d,d+1} & \cdots & \cdots & \cdots & pa_{d,h} \\
 & & & & & & & & & \\
0 & \cdots & \cdots & 0 & 1 & 0 & \cdots & \cdots & \cdots & 0 \\
0 & \cdots & & \cdots & 0 & p & 0 & \cdots & \cdots & 0 \\
0 & \cdots & & \cdots & 0 & 0 & p & 0 & \cdots & 0 \\
 & & & & & & & & & \\
0 & \cdots & & \cdots & 0 & 0 & 0 & \ddots & 0 & 0 \\
0 & \cdots & & \cdots & 0 & 0 & \cdots & \cdots & p & 0
\end{array}\right) \qquad (\mathcal{F})$$

with $a_{i,j} \in W$, $a_{1,h} \in W^*$; i.e., it consists of blocks of sizes (d or c) $\times$ (d or c); in the left hand upper corner, which is of size $d \times d$, there are entries in the last column, named $a_{i,d}$, and the entries immediately below the diagonal are equal to 1; the left and lower block has only one element not equal to zero, and it is 1; the right hand upper corner is unspecified and its entries are written $pa_{i,j}$; the right hand lower corner, which is of size $c \times c$, has only entries immediately below the diagonal, and they are all equal to p.

Note that if a Dieudonné module M is defined by a matrix in displayed normal form, then either its p-rank $f(M)$ is maximal, $f = d$, and this happens if and only if $a_{1,d}$ is not divisible by p, or $f(M) < d$, and in that case $a(M) = 1$. The p-rank is zero if and only if $a_{i,d} \equiv 0 \pmod{p}$, $\forall 1 \leq i \leq d$.

Lemma 5.13. **BB** *Let M be the Dieudonné module of a p-divisible group G over k with $f(G) = 0$. Suppose $a(G) = 1$. Then there exists a W-basis for M on which $\mathcal{F}$ has a matrix which is in normal form. In this case the entries $a_{1,d}, \ldots, a_{d,d}$ are divisible by p, they can be chosen to be equal to zero.*

Lemma 5.14. (of Cayley-Hamilton type) *Let L be a field of characteristic p, let $W = W_\infty(L)$ be its ring of infinite Witt vectors. Let X be a p-divisible group, with* $\dim(G) = d$, *and* $\text{height}(G) = h$, *with Dieudonné module $\mathbb{D}(X) = M$. Suppose there is a W-basis of M, such that the display matrix $(a_{i,j})$ on this base gives an $\mathcal{F}$-matrix in normal form as in* 5.12. *We write $e = X_1 = e_1$ for the first base vector. Then for the expression*

$$P := \sum_{i=1}^{d} \sum_{j=d}^{h} p^{j-d} a_{i,j}^{\sigma^{h-j}} \mathcal{F}^{h+i-j-1}$$

we have

$$\mathcal{F}^h \cdot e = P \cdot e \,.$$

Note that we take powers of $\mathcal{F}$ in the σ-linear sense, i.e., if the display matrix is

$(a) = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$ whose associated F-matrix is $(\mathcal{F}) = (pa) = \begin{pmatrix} A & pB \\ C & pD \end{pmatrix}$

then $\mathcal{F}^n$ is given by the matrix

$$(\mathcal{F}^n) = (pa)\cdot(pa^{\sigma})\cdot\cdots\cdot(pa^{\sigma^{n-1}}).$$

The exponent $h+i-j-1$ runs from $0 = h+1-h-1$ to $h-1 = h+d-d-1$.

Note that we do not claim that P and $\mathcal{F}^h$ have the same effect on all elements of M.

PROOF. Note that $\mathcal{F}^{i-1}e_1 = e_i$ for $i \le d$.

CLAIM. *For $d \le s < h$ we have*:

$$\mathcal{F}^s X = \left(\sum_{i=1}^{d} \sum_{j=d}^{s} \mathcal{F}^{s-j} p^{j-d} a_{i,j} \mathcal{F}^{i-1} \right) X + p^{s-d} e_{s+1}.$$

This is correct for $s = d$. The induction step from s to $s+1 < h$ follows from

$$\mathcal{F} e_{s+1} = \left(\sum_{i=1}^{d} p\ a_{i,s+1} F^{i-1} \right) X + p e_{s+2}\,.$$

This proves the claim. Computing $\mathcal{F}(\mathcal{F}^{h-1}X)$ gives the desired formula. □

Proposition 5.15. *Let k be an algebraically closed field of characteristic p, let $W = W_\infty(K)$ be its ring of infinite Witt vectors. Suppose G is a p-divisible group over k such that for its Dieudonné module the map $\mathcal{F}$ is given by a matrix in normal form. Let P be the polynomial given in the previous proposition. The Newton polygon $\mathcal{N}(G)$ of this p-divisible group equals the Newton polygon given by the polynomial P.*

PROOF. Consider the $W[F]$-submodule $M' \subset M$ generated by $X = e_1$. Note that M' contains $X = e_1, e_2, \ldots, e_d$. Also, it contains $\mathcal{F}e_d$, which equals e_{d+1} plus a linear combination of the previous ones; hence $e_{d+1} \in M'$. In the same way we see: $pe_{d+2} \in M'$, and $p^2 e_{d+3} \in M'$ and so on. This shows that $M' \subset M = \bigoplus_{i \le h} W\cdot e_i$ is of finite index. We see that $M' = W[F]/W[F]\cdot(F^h - P)$. From this we see by the classification of p-divisible groups up to isogeny, that the result follows by [**48**], II.1; also see [**24**], pp. 82-84. By [**24**], page 82, Lemma 2 we conclude that the Newton polygon of M' in case of the monic polynomial $\mathcal{F}^h - \sum_0^m b_i \mathcal{F}^{m-i}$ is given by the lower convex hull of the pairs $\{(i, v(b_i)) \mid i\}$. Hence the proposition is proved. □

Corollary 5.16. *We take the notation as above. Suppose that every element $a_{i,j}$, $1 \le i \le c$, $c \le j \le h$, is either equal to zero, or is a unit in $W(k)$. Let S be the set of pairs (i,j) with $0 \le i \le c$ and $c \le j \le h$ for which the corresponding element is non-zero:*

$$(i,j) \in S \quad \Longleftrightarrow \quad a_{i,j} \neq 0.$$

Consider the image T under

$$S \to T \subset \mathbb{Z} \times \mathbb{Z} \quad \text{given by} \quad (i,j) \mapsto (j+1-i, j-c).$$

Then $\mathcal{N}(X)$ is the lower convex hull of the set $T \subset \mathbb{Z} \times \mathbb{Z}$ and the point $(0,0)$; note that $a_{1,h} \in W^*$, hence $(h, h-c=d) \in T$. This can be visualized in the following

diagram (we have pictured the case $d \le h-d$):

$$\begin{array}{cccccccccccc}
 & & & & & & & a_{c,h} & \cdots & & a_{1,h} \\
 & & & & & & . & \cdots & & . & \\
 & & & & & a_{c,2c+2} & . & . & . & & \\
 & & & & a_{c,2c+1} & . & . & . & a_{1,2c+1} & & \\
 & & & . & \vdots & \vdots & \vdots & . & & & \\
 & a_{c,c+1} & \cdots & a_{i,c+1} & \cdots & a_{2,c+1} & a_{1,c+1} & & & & \\
a_{c,c} & \cdots & a_{i,c} & \cdots & \cdots & a_{1,c} & & & & &
\end{array}$$

Here the element $a_{c,c}$ is in the plane with coordinates $(x=1, y=0)$ and $a_{1,h}$ has coordinates $(x=h, y=h-c=d)$. One erases the spots where $a_{i,j}=0$, and one leaves the places where $a_{i,j}$ is as unit. The lower convex hull of these points and $(0,0)$ (and $(h, h-c)$) equals $\mathcal{N}(X)$.

Theorem 5.10 proves the following statement:

Conjecture. (The weak Grothendieck conjecture) *Given Newton polygons $\beta \prec \delta$ there exists a family of p-divisible groups over an integral base having δ as Newton polygon for the generic fiber, and β as Newton polygon for a closed fiber.*

However, we will prove a much stronger result later.

5.17. For principally quasi-polarized p-divisible groups and for principally polarized abelian varieties we have an analogous method.

5.18. We fix an integer g. For every *symmetric* Newton polygon ξ of height $2g$ we define

$$\triangle(\xi) = \{\, (x,y) \in \mathbb{Z} \times \mathbb{Z} \mid \ y < x \le g, \ \ (x,y) \prec \xi \,\},$$

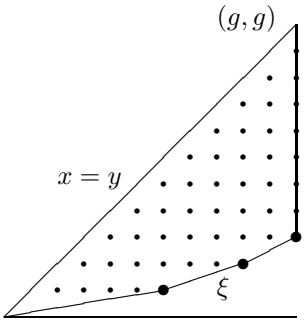
and we write

$$\boxed{\operatorname{sdim}(\xi) := \#(\triangle(\xi)).}$$

Define $\triangle$ by

$$\triangle = \{\, (x,y) \in \mathbb{Z} \times \mathbb{Z} \mid 0 \le y < x \le g \,\}.$$

Example.



$\dim(\mathcal{W}_\xi(\mathcal{A}_{g,1} \otimes \mathbb{F}_p)) = \#(\triangle(\xi))$

$\xi = (5,1) + (2,1) + 2{\cdot}(1,1) + (1,2) + (1,5)$,

g=11; slopes: $\{6 \times \frac{5}{6}, 3 \times \frac{2}{3}, 4 \times \frac{1}{2}, 3 \times \frac{1}{3}, 6 \times \frac{1}{6}\}$.

This case: $\dim(\mathcal{W}_\xi(\mathcal{A}_{g,1} \otimes \mathbb{F}_p)) = \operatorname{sdim}(\xi) = 48$ (see 8.12)

Suppose given a p-divisible group X_0 over k of dimension g with a principal quasi-polarization λ. We write $\mathcal{N}(X_0) = \gamma$; this is a symmetric Newton polygon. We write $D = \mathcal{D}(X_0, \lambda)$ for the universal deformation space; in particular $D = \operatorname{Spec}(R)$, where $\operatorname{Def}(X_0, \lambda) = \operatorname{Spf}(R)$; see 10.21. For every symmetric Newton polygon ξ with $\xi \succ \gamma$ we define $W_\xi(D) \subset D$ as the maximal closed, reduced subscheme of D carrying all fibers with Newton polygon equal to or above ξ; this

space exists and is closed in D by Grothendieck-Katz, see [**40**], Theorem 2.3.1 on page 143. Note that $W_\rho = D$, where $\rho = g\cdot((1,0)+(0,1))$.

Theorem 5.19. (NP-strata for principally quasi-polarized formal groups) *Suppose $a(X_0) \leq 1$. Write $D = \mathcal{D}(X_0,\lambda)$. For every symmetric $\xi \succ \gamma := \mathcal{N}(X_0)$ we have: $\mathcal{W}_\xi(D)$ is formally smooth, with $\dim(\mathcal{W}_\xi(D)) = \mathrm{sdim}(\xi)$. The strata $W'_\xi := \mathcal{W}_\xi(D) \subset \mathcal{D}(X_0,\lambda)$ are nested as given by the partial ordering on symmetric Newton polygons, i.e.,*

$$W'_\xi \quad\subset\quad W'_\delta \quad\Longleftrightarrow\quad \triangle(\xi)\subset\triangle(\delta) \quad\Longleftrightarrow\quad \xi \prec \delta.$$

Generically on W'_ξ the fibers have Newton polygon equal to ξ. We can choose a coordinate system on $\mathcal{D}(X_0,\lambda)$ in which all W'_ξ are given by linear equations.

Corollary 5.20. *Suppose given a principally polarized abelian variety (A_0,λ_0) over k with $a(A_0)\leq 1$. Strata in $\mathcal{D}(X_0,\lambda_0)$ according to Newton polygons are exactly as in* 5.19. *In particular, the fiber above the generic point of W'_ξ is a principally polarized abelian scheme over $\mathrm{Spec}(B_\xi)$ having Newton polygon equal to ξ.*

Proof. (of 5.20, assuming 5.19) We write $(A_0,\lambda_0)[p^\infty] =: (X_0,\lambda_0)$. By Serre-Tate theory, see [**42**], Section 1, the formal deformation spaces of (A_0,λ_0) and of (X_0,λ_0) are canonically isomorphic, say $(A',\lambda) \to \mathrm{Spf}(R)$ and $(\mathcal{X}',\lambda)\to\mathrm{Spf}(R)$ and $(A',\lambda)[p^\infty]\cong(\mathcal{X}',\lambda)$. By Chow-Grothendieck, see [**32**], III^1.5.4 (this is also called a theorem of "GAGA-type"), the formal polarized abelian scheme is algebraizable, and we obtain the universal deformation as a polarized abelian scheme $(A,\lambda)\to\mathrm{Spec}(R)$; see 10.21. We consider the generic point of $W_\xi\subset\mathcal{D}(X_0,\lambda_0)=\mathrm{Spec}(R)$. The Newton polygon of fibers can be read off from the fibers in $(\mathcal{X},\lambda)\to\mathrm{Spec}(R)$. This proves that 5.20 follows from 5.19. □

Proof. (of 5.19) The proof of this theorem is analogous to the proof of 5.10. We use the diagram

$$\begin{array}{ccccccc} & & & & & & -1 \\ & & & & T_{g,g} & \cdots & T_{1,g} & \\ & & & & \vdots & . & & \\ & & & . & & & & \\ 1 & T_{g,1} & \cdots & T_{1,1} & & & \end{array}$$

Here $T_{i,j}$, $1\leq i,j\leq g$, is written on the place with coordinates $(g-i+j,j-1)$. We use the ring

$$B := \frac{k[[T_{i,j};1\leq i,j\leq g]]}{(T_{k\ell}-T_{\ell k})},\quad T_{i,j}=Z_{(g-i+j,j-1)},\quad (g-i+j,j-1)\in\triangle.$$

Note that $B = k[[T_{i,j}\mid 1\leq i\leq j\leq g]] = k[[Z_{x,y}\mid (x,y)\in\triangle]]$. For a symmetric ξ with $\xi\succ\mathcal{N}(X_0)$ we consider

$$B_\xi = \frac{k[[T_{i,j};1\leq i,j\leq g]]}{(T_{k\ell}-T_{\ell k},\quad\text{and}\quad Z_{(x,y)}\ \ \forall (x,y)\notin\triangle(\xi))} \quad\cong\quad k[[Z_{(x,y)}\mid (x,y)\in\triangle(\xi)]].$$

With these notations, applying 5.14 and 5.16 we finish the proof of 5.19 as we did in the proof of 5.10 above. □

If the condition $a(A_0)\leq 1$ in the theorem and corollary above is replaced by $a(A_0)=0$ all fibers above this deformation space are ordinary.

5.21. A conjecture by Manin. Let A be an abelian variety. The Newton polygon $\mathcal{N}(A)$ is symmetric (see 1.18). A conjecture by Manin expects the converse to hold:

Conjecture. (see [**48**], page 76, Conjecture 2) *For any symmetric Newton polygon ξ there exists an abelian variety A such that $\mathcal{N}(A) = \xi$.*

This was proved in the Honda-Tate theory, see 3.12, 3.14. *We sketch a pure characteristic p proof,* see [**65**], Section 5. It is not difficult to show that there exists a principally polarized supersingular abelian variety (A_0, λ_0) with $a(A_0) = 1$, see [**65**], Section 4; this also follows from [**45**], 4.9. By 5.19 it follows that $\mathcal{W}^0_\xi(\mathcal{D}(A_0, \lambda_0))$ is non-empty, *which proves the Manin conjecture.* □

5.22. Let $g \in \mathbb{Z}_{\geq 3}$. There exists an abelian variety in characteristic p which has p-rank equal to zero, and which is not supersingular. In fact choose $\xi = \sum(m_i, n_i)$, a symmetric Newton polygon with $m_i > 0$ and $n_i > 0$ for every i and $(m_i, n_i) \neq (1, 1))$ for at least one i. For example $\xi = (1, g-1) + (g-1, 1)$ or $\xi = (2, 1) + (g-3)(1, 1) + (1, 2)$. By the Manin conjecture there exists an abelian variety A with $\mathcal{N}(A) = \xi$. We see that A is not supersingular, and that the p-rank $f(A)$ equals zero.

6. Hilbert modular varieties

We discuss Hilbert modular varieties over $\mathbb{F}$ in this section. (Recall that $\mathbb{F}$ is the algebraic closure of $\mathbb{F}_p$.) A Hilbert modular variety attached to a totally real number field F classifies "abelian varieties with real multiplication by $\mathcal{O}_F$". An abelian variety A is said to have "real multiplication by $\mathcal{O}_F$" if $\dim(A) = [F : \mathbb{Q}]$ and there is an embedding $\mathcal{O}_F \hookrightarrow \mathrm{End}(A)$; the terminology "fake elliptic curve" was used by some authors. The moduli space of such objects behaves very much like the modular curve, except that its dimension is equal to $[F : \mathbb{Q}]$. Similar to the modular curve, a Hilbert modular variety attached to a totally real number field F has a family of Hecke correspondences coming from the group $\mathrm{SL}_2(F \otimes_{\mathbb{Q}} \mathbb{A}_f^{(p)})$ or $\mathrm{GL}_2(\mathbb{A}_f^{(p)})$ depending on the definition one uses. Hilbert modular varieties are closely related to modular forms for GL_2 over totally real fields and the arithmetic of totally real fields.

Besides their intrinsic interest, Hilbert modular varieties play an essential role in the Hecke orbit problem for Siegel modular varieties. This connection results from a special property of $\mathcal{A}_{g,1,n}$ which is not shared by all modular varieties of PEL type: For every $\mathbb{F}$-point x_0 of $\mathcal{A}_{g,1,n}$, there exists a Hilbert modular variety $\mathcal{M}$ and an isogeny correspondence R on $\mathcal{A}_{g,1,n}$ such that x_0 is contained in the image of $\mathcal{M}$ under the isogeny correspondence R. See 9.10 for a precise formulation, and also the beginning of §8.

References. [**71**], [**22**], [**80**] Chap X, [**23**], [**31**], [**88**].

Let $F_1, \ldots, F_r$ be totally real number fields, and let $E := F_1 \times \cdots \times F_r$. Let $\mathcal{O}_E = \mathcal{O}_{F_1} \times \cdots \times \mathcal{O}_{F_r}$ be the product of the rings of integers of $F_1, \ldots, F_r$. Let $\mathcal{L}_i$ be an invertible $\mathcal{O}_{F_i}$-module, and let $\mathcal{L}$ be the invertible $\mathcal{O}_E$-module $\mathcal{L} = \mathcal{L}_1 \times \cdots \times \mathcal{L}_r$.

Definition 6.1. Notation as above. A *notion of positivity* on an invertible $\mathcal{O}_E$-module $\mathcal{L}$ is a union $\mathcal{L}^+$ of connected components of $\mathcal{L} \otimes_{\mathbb{Q}} \mathbb{R}$ such that $\mathcal{L} \otimes_{\mathbb{Q}} \mathbb{R}$ is the disjoint union of $\mathcal{L}^+$ and $-\mathcal{L}^+$.

Definition 6.2.

(i) An $\mathcal{O}_E$-linear abelian scheme is a pair $(A \to S, \iota)$, where $A \to S$ is an abelian scheme, and $\iota : \mathcal{O}_E \to \mathrm{End}_S(A)$ is an injective ring homomorphism such that $\iota(1) = \mathrm{Id}_A$. Note that every $\mathcal{O}_E$-linear abelian scheme $(A \to S, \iota)$ as above decomposes as a product $(A_1 \to S, \iota_1) \times \cdots \times (A_r \to S, \iota_r)$. Here (A_i, ι_i) is an $\mathcal{O}_{F_i}$-linear abelian scheme for $i = 1, \ldots, r$, and $A = A_1 \times_S \cdots \times_S A_r$.

(ii) An $\mathcal{O}_E$-linear abelian scheme $(A \to S, \iota)$ is said to be of HB-type if $\dim(A/S) = \dim_{\mathbb{Q}}(E)$.

(iii) An $\mathcal{O}_E$-linear polarization of an $\mathcal{O}_E$-linear abelian scheme is a polarization $\lambda : A \to A^t$ such that $\lambda \circ \iota(u) = \iota(u)^t \circ \lambda$ for all $u \in \mathcal{O}_E$.

Exercise 6.3. Suppose that $(A \to S, \iota)$ is an $\mathcal{O}_E$-linear abelian scheme, and

$$(A \to S, \iota) = (A_1 \to S, \iota_1) \times \cdots \times (A_r \to S, \iota_r)$$

as in (i). Show that $(A_1 \to S, \iota_1)$ is an $\mathcal{O}_{F_i}$-linear abelian scheme of HB-type for $i = 1, \ldots, r$.

Exercise 6.4. Show that every $\mathcal{O}_E$-linear abelian variety of HB-type over a field admits an $\mathcal{O}_E$-linear polarization.

Definition 6.5. Let $E_p = \prod_{j=1}^{s} F_{v_j}$ be a product of finite extension fields F_{v_j} of $\mathbb{Q}_p$. Let $\mathcal{O}_{E_p} = \prod_{j=1}^{s} \mathcal{O}_{F_{v_j}}$ be the product of the rings of elements in F_{v_j} which are integral over $\mathbb{Z}_p$.

(i) An $\mathcal{O}_{E_p}$-linear p-divisible group is a pair $(X \to S, \iota)$, where $X \to S$ is a p-divisible group, and $\iota : \mathcal{O}_E \otimes_{\mathbb{Z}} \mathbb{Z}_p \to \mathrm{End}_S(X)$ is an injective ring homomorphism such that $\iota(1) = \mathrm{Id}_X$. Every $(\mathcal{O}_{E_p})$-linear p-divisible group $(X \to S, \iota)$ decomposes canonically into a product $(X \to S, \iota) = \prod_{j=1}^{s} (X_j, \iota_j)$, where (X_j, ι_j) is an $\mathcal{O}_{F_{v_j}}$-linear p-divisible group, defined to be the image of the idempotent in $\mathcal{O}_{E_p}$ corresponding to the factor $\mathcal{O}_{F_{v_j}}$ of $\mathcal{O}_{E_p}$.

(ii) An $\mathcal{O}_{E_p}$-linear p-divisible group $(X \to S, \iota)$ is said to have rank two if in the decomposition $(X \to S, \iota) = \prod_{j=1}^{s} (X_j, \iota_j)$ in (i) above we have $\mathrm{ht}(X_j/S) = 2\,[F_{v_j} : \mathbb{Q}_p]$ for all $j = 1, \ldots, s$.

(iii) An $\mathcal{O}_{E_p}$-linear polarization $(\mathcal{O}_E \otimes_{\mathbb{Z}} \mathbb{Z}_p)$-linear p-divisible group $(X \to S, \iota)$ is a symmetric isogeny $\lambda : X \to X^t$ such that $\lambda \circ \iota(u) = \iota(u)^t \circ \lambda$ for all $u \in \mathcal{O}_{E_p}$.

(iv) A rank-two $\mathcal{O}_{E_p}$-linear p-divisible group $(X \to S, \iota)$ is of HB-type if it admits an $\mathcal{O}_{E_p}$-linear polarization.

Exercise 6.6. Show that for every $\mathcal{O}_E$-linear abelian scheme of HB-type $(A \to S, \iota)$, the associated $(\mathcal{O}_E \otimes_{\mathbb{Z}} \mathbb{Z}_p)$-linear p-divisible group $(A[p^\infty], \iota[p^\infty])$ is of HB-type.

Definition 6.7. Let $E = F_1 \times \cdots \times F_r$, where $F_1, \ldots, F_r$ are totally real number fields. Let $\mathcal{O}_E = \mathcal{O}_{F_1} \times \cdots \times \mathcal{O}_{F_r}$ be the product of the ring of integers of $F_1, \ldots, F_r$. Let $k \supset \mathbb{F}_p$ be an algebraically closed field as before. Let $n \geq 3$ be an integer such that $(n, p) = 1$. Let $(\mathcal{L}, \mathcal{L}^+)$ be an invertible $\mathcal{O}_E$-module with a notion of positivity. The Hilbert modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ over k is a smooth scheme over k of dimension $[E : \mathbb{Q}]$ such that for every k-scheme S the set of S-valued points of $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+n}$ is the set of isomorphism class of 6-tuples $(A \to S, \iota, \mathcal{L}, \mathcal{L}^+, \lambda, \eta)$, where

(i) $(A \to S, \iota)$ is an $\mathcal{O}_E$-linear abelian scheme of HB-type;
(ii) $\lambda : \mathcal{L} \to \mathrm{Hom}^{\mathrm{sym}}_{\mathcal{O}_E}(A, A^t)$ is an $\mathcal{O}_E$-linear homomorphism such that $\lambda(u)$ is an $\mathcal{O}_E$-linear polarization of A for every $u \in \mathcal{L} \cap \mathcal{L}^+$, and the homomorphism $A \otimes_{\mathcal{O}_E} \mathcal{L} \xrightarrow{\sim} A^t$ induced by λ is an isomorphism of abelian schemes.
(iii) η is an $\mathcal{O}_E$-linear level-n structure for $A \to S$, i.e., an $\mathcal{O}_E$-linear isomorphism from the constant group scheme $(\mathcal{O}_E/n\mathcal{O}_E)^2_S$ to $A[n]$.

Remark 6.8. Let $(A \to S, \iota, \lambda, \eta)$ be an $\mathcal{O}_E$-linear abelian scheme with polarization sheaf by $(\mathcal{L}, \mathcal{L}^+)$ and a level-n structure satisfying the condition in (ii) above. Then the $\mathcal{O}_E$-linear polarization λ induces an $\mathcal{O}_E/n\mathcal{O}_E$-linear isomorphism

$$(\mathcal{O}_E/n\mathcal{O}_E) = \bigwedge^2 (\mathcal{O}_E/n\mathcal{O}_E)^2 \xrightarrow{\sim} \mathcal{L}^{-1}\mathcal{D}_E^{-1} \otimes_{\mathbb{Z}} \mu_n$$

over S, where $\mathcal{D}_E$ denotes the invertible $\mathcal{O}_E$-module $\mathcal{D}_{F_1} \times \cdots \times \mathcal{D}_{F_r}$. This isomorphism is a discrete invariant of the quadruple $(A \to S, \iota, \lambda, \eta)$. The above invariant defines a morphism f_n from the Hilbert modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ to the finite étale scheme $\Xi_{E,\mathcal{L},n}$ over k, where the finite étale k-scheme $\Xi_{E,\mathcal{L},n}$ is defined by $\Xi_{E,\mathcal{L},n} := \underline{\mathrm{Isom}}(\mathcal{O}_E/n\mathcal{O}_E, \mathcal{L}^{-1}\mathcal{D}_E^{-1} \otimes_{\mathbb{Z}} \mu_n)$. Notice that $\Xi_{E,\mathcal{L},n}$ is an $(\mathcal{O}_E/n\mathcal{O}_E)^\times$-torsor; it is constant over k because k is algebraically closed. The morphism f_n is faithfully flat.

Although we defined the Hilbert modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ over an algebraically closed field $k \supset \mathbb{F}_p$, we could have defined it over $\mathbb{F}_p$. Then we should use the étale $(\mathcal{O}_E/n\mathcal{O}_E)^\times$-torsor $\Xi_{E,\mathcal{L},n} := \underline{\mathrm{Isom}}(\mathcal{O}_E/n\mathcal{O}_E, \mathcal{L}^{-1}\mathcal{D}_E^{-1} \otimes_{\mathbb{Z}} \mu_n)$ over $\mathbb{F}_p$, and we have a faithfully flat morphism $f_n : \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n} \to \Xi_{E,\mathcal{L},n}$ over $\mathbb{F}_p$.

Remark 6.9.

(i) We have followed [**22**] in the definition of Hilbert modular varieties, except that E is a product of totally real number fields, rather than a totally real number field as in [**22**].
(ii) The product decompositions

$$\mathcal{O}_E = \mathcal{O}_{F_1} \times \cdots \times \mathcal{O}_{F_r} \quad \text{and} \quad (\mathcal{L}, \mathcal{L}^+) = (\mathcal{L}_1, \mathcal{L}_1^+) \times \cdots \times (\mathcal{L}_r, \mathcal{L}_r^+)$$

induce a natural isomorphism

$$\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n} \xrightarrow{\sim} \mathcal{M}_{F_1,\mathcal{L}_1,\mathcal{L}_1^+,n} \times \cdots \times \mathcal{M}_{F_r,\mathcal{L}_r,\mathcal{L}_r^+ n}\,.$$

Remark 6.10. The $\mathcal{O}_E$-linear homomorphism λ in Def. 6.7 should be thought of as specifying a family of $\mathcal{O}_E$-linear polarizations, instead of only one polarization: every element $u \in \mathcal{L} \cap \mathcal{L}^+$ gives a polarization $\lambda(u)$ on $A \to S$. Notice that given a point $x_0 = [(A, \iota, \lambda, \eta)]$ in $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}(k)$, there may not exist an $\mathcal{O}_E$-linear principal polarization on A, because that means that the element of the strict ideal class group represented by $(\mathcal{L}, \mathcal{L}^+)$ is trivial. However, every point $[(A, \iota, \lambda, \eta)]$ of $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ admits an $\mathcal{O}_E$-linear polarization of degree prime to p, because there exists an element $u \in \mathcal{L}^+$ such that $\mathrm{Card}(\mathcal{L}/\mathcal{O}_E \cdot u)$ is not divisible by p. In [**89**] and [**88**] a version of Hilbert modular varieties was defined by specifying a polarization degree d which is prime to p. The resulting Hilbert modular variety is not necessarily irreducible over $\mathbb{F}$; rather it is a disjoint union of modular varieties of the form $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$.

Theorem 6.11. **BB** *Notation as above.*

(i) *The modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ over the algebraically closed field $k \supset \mathbb{F}_p$ is normal and is a local complete intersection. Its dimension is equal to* $\dim_{\mathbb{Q}}(E)$.
(ii) *Every fiber of $f_n : \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n} \to \Xi_{E,\mathcal{L},n}$ is irreducible.*
(iii) *The morphism f_n is smooth outside a closed subscheme of $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ of codimension at least two.*

Remark.

(i) See [**22**] for a proof of Theorem 6.11 which uses the arithmetic toroidal compactification constructed in [**71**].
(ii) The modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ is not smooth over k if any one of the totally real fields F_i is ramified above p.

6.12. Hecke orbits on Hilbert modular varieties. Let E, $\mathcal{L}$ and $\mathcal{L}^+$ be as before. Denote by $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+}{}^{\sim}$ the projective system $\left(\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}\right)_n$ of Hilbert modular varieties over $\mathbb{F}$, where n runs through all positive integers such that $n \geq 3$ and $\gcd(n,p) = 1$. It is clear that the profinite group $\mathrm{SL}_2(\mathcal{O}_E \otimes_{\mathbb{Z}} \mathbb{Z}^{\wedge,(p)})$ operates on the tower $\mathcal{M}^{\sim}_{E,\mathcal{L},\mathcal{L}^+}$, by pre-composing with the $\mathcal{O}_E$-linear level structures. Here $\mathbb{Z}^{\wedge,(p)} = \prod_{\ell \neq p} \mathbb{Z}_\ell$. The transition maps in the projective system are

$$\pi_{mn,n} : \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,mn} \to \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n} \qquad (mn,p) = 1, n \geq 3, m \geq 1\,.$$

The map $\pi_{mn,n}$ is defined by the following construction. Let

$$[m] : (\mathcal{O}_E/n\mathcal{O}_E)^2 \to (\mathcal{O}_E/mn\mathcal{O}_E)^2$$

be the injection induced by "multiplication by m". Given a point (A,ι,λ,η) of $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,mn}$, the composition $\eta \circ [m]$ factors through the inclusion $i_{m,n} : A[m] \hookrightarrow A[mn]$ to give a level-n structure η' such that $\eta \circ [m] = i_{m,n} \circ \eta'$.

Let $\Xi^{\sim}_E$ be the projective system $(\Xi_{E,n})_n$, where n also runs through all positive integers such that $n \geq 3$ and $(n,p) = 1$. The transition maps are defined similarly. The maps $f_n : \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n} \to \Xi_{E,n}$ define a map $f^{\sim} : \mathcal{M}^{\sim}_{E,\mathcal{L},\mathcal{L}^+} \to \Xi^{\sim}_E$ between projective systems.

It is clear that the profinite group $\mathrm{SL}_2(\mathcal{O}_E \otimes_{\mathbb{Z}} \mathbb{Z}^{\wedge,(p)})$ operates on the right of the tower $\mathcal{M}^{\sim}_{E,\mathcal{L},\mathcal{L}^+}$, by pre-composing with the $\mathcal{O}_E$-linear level structures. Moreover this action is compatible with the map $f^{\sim} : \mathcal{M}^{\sim}_{E,\mathcal{L},\mathcal{L}^+} \to \Xi^{\sim}_E$ between projective systems.

The above right action of the compact group $\mathrm{SL}_2(\mathcal{O}_E \otimes_{\mathbb{Z}} \mathbb{Z}^{\wedge,(p)})$ on the projective system $\mathcal{M}^{\sim}_{E,\mathcal{L},\mathcal{L}^+}$ extends to a right action of $\mathrm{SL}_2(E \otimes_{\mathbb{Q}} \mathbb{A}_f^{(p)})$ on $\mathcal{M}^{\sim}_{E,\mathcal{L},\mathcal{L}^+}$. Again this action is compatible with $f^{\sim} : \mathcal{M}^{\sim}_{E,\mathcal{L},\mathcal{L}^+} \to \Xi^{\sim}_E$. This action can be described as follows. A geometric point of $\mathcal{M}^{\sim}_{E,\mathcal{L},\mathcal{L}^+}$ is a quadruple $(A,\iota_A,\lambda_A,\eta^{\sim}_A)$, where the infinite prime-to-p level structure

$$\eta^{\sim}_A : \prod_{\ell \neq p} (\mathcal{O}_E[1/\ell]/\mathcal{O}_E) \xrightarrow{\sim} \prod_{\ell \neq p} A[\ell^\infty]$$

is induced by a compatible system of level-n structures, n running through integers such that $(n,p) = 1$ and $n \geq 3$. Suppose that we have a $\gamma \in \mathrm{SL}_2(E \otimes_{\mathbb{Q}} \mathbb{A}_f^{(p)})$, and $m\,\gamma$ belongs to $\mathrm{M}_2(\mathcal{O}_E \otimes_{\mathbb{Z}} \mathbb{Z}^{\wedge,(p)})$, where m is a non-zero integer which is prime to p. Then the image of the point $(A,\iota_A,\lambda_A,\eta^{\sim}_A)$ under γ is a quadruple $(B,\iota_B,\lambda_B,\eta^{\sim}_B)$ such that there exists an $\mathcal{O}_E$-linear prime-to-p isogeny $m\beta : B \to A$ such that the

diagram

$$\begin{array}{ccc} \coprod_{\ell\neq p}(\mathcal{O}_E[1/\ell]/\mathcal{O}_E)^2 & \xrightarrow{\eta_A^{\sim}} & \coprod_{\ell\neq p} A[\ell^\infty] \\ {\scriptstyle m\gamma}\uparrow & & \uparrow{\scriptstyle m\beta} \\ \coprod_{\ell\neq p}(\mathcal{O}_E[1/\ell]/\mathcal{O}_E)^2 & \xrightarrow[\eta_B^{\sim}]{} & \coprod_{\ell\neq p} B[\ell^\infty] \end{array}$$

commutes. Note that ι_B and λ_B are determined by the requirement that $m\beta$ is an $\mathcal{O}_E$-linear isogeny and $m^{-1}\cdot m\beta$ respects the polarizations λ_A and λ_B. In the above notation, as the point $(A,\iota_A,\lambda_A,\eta_A^{\sim})$ varies, we get a prime-to-p quasi-isogeny $\beta = m^{-1}\cdot(m\beta)$ attached to γ, between the universal abelian schemes.

On a fixed level $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$, the action of $\mathrm{SL}_2(E\otimes_{\mathbb{Q}}\mathbb{A}_f^{(p)})$ on the projective system $\mathcal{M}^{\sim}_{E,\mathcal{L},\mathcal{L}^+}$ induces a family of finite étale correspondences, which will be called $\mathrm{SL}_2(E\otimes_{\mathbb{Q}}\mathbb{A}_f^{(p)})$-Hecke correspondences on $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$, or prime-to-$p$ SL_2-Hecke correspondences for short. Suppose x_0 is a geometric point of $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$, and $x^{\sim}$ is point of $\mathcal{M}^{\sim}_{E,\mathcal{L},\mathcal{L}^+}$ lifting x_0. Then the prime-to-p SL_2-Hecke orbit of x_0, denoted $\mathcal{H}^{(p)}_{\mathrm{SL}_2}(x_0)$, is the image in $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ of the orbit $\mathrm{SL}_2(E\otimes_{\mathbb{Q}}\mathbb{A}_f^{(p)})\cdot x_0^{\sim}$. The set $\mathcal{H}^{(p)}_{\mathrm{SL}_2}(x_0)$ is countable.

Theorem 6.13. *Let $x_0=[(A_0,\iota_0,\lambda_0,\eta_0)]\in\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}(k)$ be a closed point of $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ such that A_0 is an ordinary abelian scheme. Let $\Sigma_{E,p}=\{\wp_1,\dots,\wp_s\}$ be the set of all prime ideals of $\mathcal{O}_E$ containing p. Then we have a natural isomorphism*

$$\mathcal{M}^{/x_0}_{E,\mathcal{L},\mathcal{L}^+,n}\cong\prod_{j=1}^{s}\underline{\mathrm{Hom}}_{\mathbb{Z}_p}\left(\mathrm{T}_p(A_0[\wp_j^\infty]_{\text{ét}})\otimes_{(\mathcal{O}_E\otimes\mathbb{Z}_p)}\mathrm{T}_p(A_0^t[\wp_j^\infty]_{\text{ét}}),\,\mathbb{G}_m^{\wedge}\right).$$

In particular, the formal completion of the Hilbert modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ at the ordinary point x_0 has a natural structure as a $[E:\mathbb{Q}]$-dimensional $(\mathcal{O}_E\otimes_{\mathbb{Z}}\mathbb{Z}_p)$-linear formal torus, non-canonically isomorphic to $(\mathcal{O}_E\otimes_{\mathbb{Z}}\mathbb{Z}_p)\otimes_{\mathbb{Z}_p}\mathbb{G}_m^{\wedge}$.

Proof. By the Serre-Tate theorem, we have

$$\mathcal{M}^{/x_0}_{E,\mathcal{L},\mathcal{L}^+,n}\cong\prod_{j=1}^{s}\underline{\mathrm{Hom}}_{\mathcal{O}_E\otimes\mathbb{Z}_p}\left(\mathrm{T}_p(A_0[\wp_j^\infty]_{\text{ét}}),\,A_0[\wp_j^\infty]^{\wedge}_{\text{mult}}\right),$$

where $A_0[\wp_j^\infty]^{\wedge}_{\text{mult}}$ is the formal torus attached to $A_0[\wp_j^\infty]_{\text{mult}}$, or equivalently the formal completion of A_0. The character group of the last formal torus is naturally isomorphic to the p-adic Tate module $\mathrm{T}_p(A_0^t[\wp_j^\infty]_{\text{ét}})$ attached to the maximal étale quotient of $A_0^t[\wp_j^\infty]_{\text{ét}}$). ∎

Proposition 6.14. *Notation as in* 6.13. *Assume that $k=\mathbb{F}$, so that*

$$x_0=[(A_0,\iota_0,\lambda_0,\eta_0)]\in\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}(\mathbb{F})$$

and A_0 is an ordinary $\mathcal{O}_E$-linear abelian variety of HB-type over $\mathbb{F}$.

(i) *There exist totally imaginary quadratic extensions K_i of F_i, $i=1,\dots,r$ such that*

$$\mathrm{End}^0_{\mathcal{O}_E}(A_0)\cong K_1\times\cdots\times K_r=:K\,.$$

Moreover, for every prime ideal $\wp_j$ of $\mathcal{O}_E$ containing p, we have

$$\mathrm{End}_{\mathcal{O}_E}(A_0)\otimes_{\mathcal{O}_E}\mathcal{O}_{E_{\wp_j}} \xrightarrow{\sim} \mathrm{End}_{\mathcal{O}_{E_{\wp_j}}}(A_0[\wp_j^\infty]_{\mathrm{mult}})\times \mathrm{End}_{\mathcal{O}_{E_{\wp_j}}}(A_0[\wp_j^\infty]_{\text{ét}})$$
$$\cong \mathcal{O}_{E_{\wp_j}}\times\mathcal{O}_{E_{\wp_j}} \xleftarrow{\sim} \mathcal{O}_K\otimes_{\mathcal{O}_E}\mathcal{O}_{E_{\wp_j}}\,.$$

In particular, the quadratic extension K_i/F_i is split above every place of F_i above p, for all $i=1,\dots,r$.

(ii) *Let $\mathrm{H}_{x_0}=\{u\in(\mathcal{O}_E\otimes\mathbb{Z}_p)^\times \mid u\cdot\bar{u}=1\}$, where $u\mapsto\bar{u}$ denotes the product of the complex conjugations on $K_1,\dots,K_r$. Then both projections*

$$\mathrm{pr}_1:\mathrm{H}_{x_0}\to\prod_{\wp\in\Sigma_{E,p}}\Big(\mathrm{End}_{\mathcal{O}_{E_\wp}}(A_0[\wp_j^\infty]_{\mathrm{mult}})\Big)^\times\cong\prod_{\wp\in\Sigma_{E,p}}\mathcal{O}_{E_\wp}^\times$$

and

$$\mathrm{pr}_2:\mathrm{H}_{x_0}\to\prod_{\wp\in\Sigma_{E,p}}\Big(\mathrm{End}_{\mathcal{O}_{E_\wp}}(A_0[\wp^\infty]_{\text{ét}})\Big)^\times\cong\prod_{\wp\in\Sigma_{E,p}}\mathcal{O}_{E_\wp}^\times$$

are isomorphisms. Here $\Sigma_{E,p}$ denotes the set consisting of all prime ideals of $\mathcal{O}_E$ which contain p.

(iii) *The group H_{x_0} operates on the $(\mathcal{O}_E\otimes_{\mathbb{Z}}\mathbb{Z}_p)$-linear formal torus $\mathcal{M}^{/x_0}_{E,\mathcal{L},\mathcal{L}^+,n}$ through the character*

$$\mathrm{H}_{x_0}\ \ni t\longmapsto \mathrm{pr}_1(t)^2\in(\mathcal{O}_E\otimes_{\mathbb{Z}}\mathbb{Z}_p)^\times\,.$$

(iv) *Notation as in* (ii) *above. Let Z be a reduced, irreducible closed formal subscheme of the formal scheme $\mathcal{M}^{/x_0}_{E,\mathcal{L},\mathcal{L}^+,n}$ which is stable under the natural action of an open subgroup U_{x_0} of H_{x_0} on $\mathcal{M}^{/x_0}_{E,\mathcal{L},\mathcal{L}^+,n}$. Then there exists a subset $S\subset\Sigma_{E,p}$ such that*

$$Z=\prod_{\wp\in S}\underline{\mathrm{Hom}}_{\mathbb{Z}_p}\left(\mathrm{T}_p(A_0[\wp^\infty]_{\text{ét}})\otimes_{(\mathcal{O}_E\otimes\mathbb{Z}_p)}\mathrm{T}_p(A_0^t[\wp^\infty]_{\text{ét}}),\,\mathbb{G}_m^\wedge\right)$$

Proof. The statement (i) is a consequence of Tate's theorem on endomorphisms of abelian varieties over a finite field, see [**79**]. The statement (ii) follows from (i). The statement (iii) is immediate from the displayed canonical isomorphism in Theorem 6.13. It remains to prove (iv).

By Theorem 2.26 and Theorem 6.13, we know that Z is a formal subtorus of the formal torus

$$\mathcal{M}^{/x_0}_{E,\mathcal{L},\mathcal{L}^+,n}=\prod_{\wp\in\Sigma_{E,p}}\underline{\mathrm{Hom}}_{\mathbb{Z}_p}\left(\mathrm{T}_p(A_0[\wp^\infty]_{\text{ét}})\otimes_{(\mathcal{O}_E\otimes\mathbb{Z}_p)}\mathrm{T}_p(A_0^t[\wp^\infty]_{\text{ét}}),\,\mathbb{G}_m^\wedge\right)\,.$$

Let $X_*(Z)$ be the group of formal cocharacters of the formal torus Z. We know that $X_*(Z)$ is a $\mathbb{Z}_p$-submodule of the cocharacter group

$$\prod_{\wp\in\Sigma_{E,p}}(\mathrm{T}_p(A_0[\wp^\infty]_{\text{ét}}))^\vee\otimes_{(\mathcal{O}_E\otimes\mathbb{Z}_p)}\left(\mathrm{T}_p(A_0^t[\wp^\infty]_{\text{ét}})\right)^\vee$$

of $\mathcal{M}^{/x_0}_{E,\mathcal{L},\mathcal{L}^+,n}$, which is co-torsion free. Moreover $X_*(Z)$ is stable under the action of H_{x_0}. Denote by $\mathcal{O}$ the closed subring of $\prod_{\wp\in\Sigma_{E,p}}\mathcal{O}_\wp$ generated by the image of the projection pr_1 in (ii). Since the image of H_{x_0} under the projection pr_1 is an open subgroup of $\prod_{\wp\in\Sigma_{E,p}}\mathcal{O}_\wp^\times$, the subring $\mathcal{O}$ of $\prod_{\wp\in\Sigma_{E,p}}\mathcal{O}_\wp$ is an order of

$\prod_{\wp\in\Sigma_{E,p}}\mathcal{O}_\wp$. So $X_*(Z)\otimes\mathbb{Q}$ is stable under the action of $\prod_{\wp\in\Sigma_{E,p}}E_\wp$. It follows that there exists a subset $S\subset\Sigma_{E,p}$ such that $X_*(Z)\otimes_{\mathbb{Z}_p}\mathbb{Q}_p$ is equal to

$$\prod_{\wp\in S}\left(\mathrm{T}_p(A_0[\wp^\infty]_{\text{ét}})\right)^\vee\otimes_{\mathcal{O}_E\otimes\mathbb{Z}_p}\left(\mathrm{T}_p(A_0^t[\wp^\infty]_{\text{ét}})\right)^\vee .$$

Since $X_*(Z)$ is a co-torsion free $\mathbb{Z}_p$-submodule of

$$\prod_{\wp\in\Sigma_{E,p}}\left(\mathrm{T}_p(A_0[\wp^\infty]_{\text{ét}})\right)^\vee\otimes_{(\mathcal{O}_E\otimes\mathbb{Z}_p)}\left(\mathrm{T}_p(A_0^t[\wp^\infty]_{\text{ét}})\right)^\vee ,$$

we see that $X_*(Z)=\left(\prod_{\wp\in S}\left(\mathrm{T}_p(A_0[\wp^\infty]_{\text{ét}})\right)^\vee\otimes_{\mathcal{O}_E\otimes\mathbb{Z}_p}\left(\mathrm{T}_p(A_0^t[\wp^\infty]_{\text{ét}})\right)^\vee\right)$. ∎

Corollary 6.15. *Let $x_0=[(A_0,\iota_0,\lambda_0,\eta_0)]\in\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}(\mathbb{F})$ be an ordinary $\mathbb{F}$-point of the Hilbert modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ as in* 6.14. *Let Z be a reduced closed subscheme of $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ such that $x_0\in Z(\mathbb{F})$. Assume that Z is stable under all $\mathrm{SL}_2(\mathbb{A}_f^{(p)})$-Hecke correspondences on $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}(\mathbb{F})$. Then there exists a subset S_{x_0} of the set $\Sigma_{E,p}$ of prime ideals of $\mathcal{O}_E$ containing p such that*

$$Z^{/x_0}=\prod_{\wp\in S}\underline{\mathrm{Hom}}_{\mathbb{Z}_p}\left(\mathrm{T}_p(A_0[\wp^\infty]_{\text{ét}})\otimes_{(\mathcal{O}_E\otimes\mathbb{Z}_p)}\mathrm{T}_p(A_0^t[\wp^\infty]_{\text{ét}}),\,\mathbb{G}_m^\wedge\right) .$$

Here $Z^{/x_0}$ is the formal completion of Z at the closed point x_0.

Proof. Notation as in 6.14. Recall that $K=\mathrm{End}^0_{\mathcal{O}_E}(A_0)$. Denote by U_K the unitary group attached to K; U_K is a linear algebraic group over $\mathbb{Q}$ such that $\mathrm{U}_K(\mathbb{Q})=\{u\in K^\times\mid u\cdot\bar{u}=1\}$. By 6.14 (i), $\mathrm{U}_K(\mathbb{Q}_p)$ is isomorphic to $(E\otimes\mathbb{Q}_p)^\times$. Denote by $\mathrm{U}_K(\mathbb{Z}_p)$ the compact open subgroup of $\mathrm{U}_K(\mathbb{Q}_p)$ corresponding to the subgroup $(\mathcal{O}_E\otimes\mathbb{Z}_p)^\times\subset(\mathcal{O}_E\otimes\mathbb{Q}_p)^\times$. This group $\mathrm{U}_K(\mathbb{Z}_p)$ is isomorphic to the group H_{x_0} in 6.14 (ii), via the projection to the first factor in the displayed formula in 6.14 (i). We have a natural action of $\mathrm{U}_K(\mathbb{Z}_p)$ on

$$\mathrm{Def}((A_0,\iota_0,\lambda_0)[p^\infty])/\mathbb{F})\cong\mathcal{M}^{/x_0}_{E,\mathcal{L},\mathcal{L}^+,n}$$

by the definition of the deformation functor $\mathrm{Def}((A_0,\iota_0,\lambda_0)[p^\infty])$.

Denote by $\mathrm{U}_K(\mathbb{Z}_{(p)})$ the subgroup $\mathrm{U}_K(\mathbb{Q})\cap\mathrm{U}_K(\mathbb{Z}_p)$ of $\mathrm{U}_K(\mathbb{Q})$; in other words $\mathrm{U}_K(\mathbb{Z}_{(p)})$ consisting of all elements $u\in\mathrm{U}_K(\mathbb{Q})$ such that u induces an automorphism of $A_0[p^\infty]$. Since Z is stable under all $\mathrm{SL}_2(\mathbb{A}_f^{(p)})$-Hecke correspondences, the formal completion $Z^{/x_0}$ at x_0 of the subvariety $Z\subset\mathcal{M}^{/x_0}_{E,\mathcal{L},\mathcal{L}^+,n}$ is stable under the natural action of the subgroup $\mathrm{U}_K(\mathbb{Z}_{(p)})$ of $\mathrm{U}_K(\mathbb{Q})$. By the weak approximation theorem for linear algebraic groups (see [**70**], 7.3, Theorem 7.7 on page 415), $\mathrm{U}_K(\mathbb{Z}_{(p)})$ is p-adically dense in $\mathrm{U}_K(\mathbb{Z}_p)$. So $Z^{/x_0}\subset\mathcal{M}^{/x_0}_{E,\mathcal{L},\mathcal{L}^+,n}$ is stable under the action of $\mathrm{U}_K(\mathbb{Z}_p)$ by continuity. We conclude the proof by invoking 6.14 (iii) and (iv). ∎

Exercise 6.16. Let (A,ι) be an $\mathcal{O}_E$-linear abelian variety of HB-type over a perfect field $K\supset\mathbb{F}_p$. Show that $\mathrm{M}_p((A,\iota)[p^\infty])$ is a free $(\mathcal{O}_E\otimes_{\mathbb{Z}}\mathbb{Z}_p)$-module of rank two.

Exercise 6.17. Let $x=[(A,\iota,\lambda,\eta)]\in\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}(k)$ be a geometric point of a Hilbert modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$, where $k\supset\mathbb{F}_p$ is an algebraically closed field. Assume that $\mathrm{Lie}(A/k)$ is a free $(\mathcal{O}_E\otimes_{\mathbb{Z}}k)$-module of rank one. Show that $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ is smooth at x over k.

Exercise 6.18. Let $k \supset \mathbb{F}_p$ be an algebraically closed field. Assume that p is unramified in E, i.e., $E \otimes_{\mathbb{Z}} \mathbb{Z}_p$ is a product of unramified extension of $\mathbb{Q}_p$. Show that $\mathrm{Lie}(A/k)$ is a free $(\mathcal{O}_E \otimes_{\mathbb{Z}} k)$-module of rank one for every geometric point $[(A, \iota, \lambda, \eta)] \in \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}(k)$.

Exercise 6.19. Give an example of a geometric point

$$x = [(A, \iota, \lambda, \eta)] \in \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}(k)$$

such that $\mathrm{Lie}(A/k)$ is not a free $(\mathcal{O}_E \otimes_{\mathbb{Z}} k)$-module of rank one.

7. Deformations of p-divisible groups to $a \leq 1$

Main references: [**20**], [**66**].

In this section we will prove and use the following rather technical result.

Theorem 7.1. **Th** (Deformation to $a \leq 1$) *Let X_0 be a p-divisible group over a field K. There exists an integral scheme S, a point $0 \in S(K)$ and a p-divisible group $\mathcal{X} \to S$ such that the fiber $\mathcal{X}_0$ is isomorphic to X_0, and for the generic point $\eta \in S$ we have*

$$\mathcal{N}(X_0) = \mathcal{N}(X_\eta) \quad \textit{and} \quad a(X_\eta) \leq 1.$$

See [**20**], 5.12 and [**66**], 2.8.

Note that if X_0 is ordinary (i.e., every slope of $\mathcal{N}(X_0)$ is either 1 or 0), there is not much to prove: $a(X_0) = 0 = a(X_\eta)$; if however X_0 is not ordinary, the theorem says something non-trivial and in that case we end with $a(X_\eta) = 1$.

At the end of this section we discuss the quasi-polarized case.

7.2. *In this section we prove* Theorem 7.1 *in case X_0 is simple.* Surprisingly, this is the most difficult step. We will see, in Section 8, that once we have the theorem in this special case, 7.1 and 7.14 will follow without much trouble.

The proof (and the only one we know) of this special case given here is a combination of general theory, and a computation. We start with one of the tools.

Theorem 7.3. **BB** (Purity of the Newton polygon stratification) *Let S be an integral scheme, and let $X \to S$ be a p-divisible group. Let $\gamma = \mathcal{N}(X_\eta)$ be the Newton polygon of the generic fiber. Let $S \supset D = S_{\neq\gamma} := \{s \mid \mathcal{N}(A_s) \not\preceq \gamma\}$* (Note that D is closed in S by Grothendieck-Katz.) *Then either D is empty or* $\mathrm{codim}(D \subset S) = 1$.

We know two proofs of this theorem, and both proofs are non-trivial. See [**20**], Theorem 4.1. Also see [**82**], th. 6.1; this second proof of purity was analyzed and re-proved [**82**], [**59**], [**83**], [**92**].

When this result was first announced, it was met by disbelief. Why? If you follow the proof by Katz, see [**40**], 2.3.2, you see that $D = S_{\neq\gamma} \subset S$ is given by "many" defining equations. From that point of view "codimension one" seems unlikely. In fact it is not known (to our knowledge) whether there exists a scheme structure on $D = S_{\neq\gamma}$ such that $(D, \mathcal{O}_D) \subset S$ is a Cartier divisor (locally principal) (i.e., locally a complete intersection, or locally a set-theoretic complete intersection).

7.4. Minimal p-divisible groups. We define the p-divisible group $H_{m,n}$ as in [**20**], 5.3; also see [**69**]. See also Exercise 4.52 for another description of $H_{m,n}$ when $K \supset \mathbb{F}_{p^{m+n}}$. Let $K \supset \mathbb{F}_p$ be a perfect field. Let M be a free $W(K)$-module of rank $m+n$, with free generators $e_0, \dots, e_{m+n-1}$. Extend $e_0, \dots, e_{m+n-1}$ to a family $(e_i)_{i\in\mathbb{Z}}$ of elements of M indexed by $\mathbb{Z}$ by the requirement that $e_{i+m+n} = p\cdot e_i$ for all $i \in \mathbb{Z}$. Define a σ-linear operator $\mathcal{F} : M \to M$ and a σ^{-1}-linear operator $\mathcal{V} : M \to M$ by

$$\mathcal{F}\cdot e_i = e_{i+n}\,, \quad \mathcal{V}\cdot e_i = e_{i+m} \qquad \forall i \in \mathbb{Z}\,.$$

This is a Dieudonné module, and the p-divisible group, whose covariant Dieudonné module is M, is denoted $H_{m,n}$.

Remark. We see that $H_{m,n}$ is defined over $\mathbb{F}_p$; for any field L we will write $H_{m,n}$ instead of $H_{m,n} \otimes L$ if no confusion can occur.

Remark. The p-divisible group $H_{m,n}$ is the "minimal p-divisible group" with Newton polygon equal to δ, the isoclinic Newton polygon of height $m+n$ and slope $m/(m+n)$. For properties of minimal p-divisible groups see [**69**]. Such groups are of importance in understanding various stratifications of $\mathcal{A}_g$.

Remark. Suppose that the perfect field K contains $\mathbb{F}_{p^h}$, where $h := m+n$. Then the p-divisible group $H_{m,n}$ defined above coincides with the one defined in Exercise 4.52; this is clear from 4.52 (3). Moreover $\mathrm{End}^0(H_{m,n})$ is an h^2-dimensional central division algebra over $\mathbb{Q}_p$ with Brauer invariant $m/(m+n)$, and $\mathrm{End}(H_{m,n})$ is the maximal order of $\mathrm{End}^0(H_{m,n})$. See the paragraph after the statement of [**20**, 5.4], where the opposite sign convention for Brauer invariants is used.

With the sign convention used both here and also in 4.52, that $\mathrm{End}^0(H_{m,n})$ has Brauer invariant $m/(m+n)$ means that there exists an injective $\mathbb{Q}_p$-linear ring homomorphism

$$j\colon \mathrm{frac}W(\mathbb{F}_{p^h}) \longrightarrow \mathrm{End}^0(H_{m,n})$$

and an element $\Phi \in \mathrm{End}^0(H_{m,n})^\times$ such that

$$\Phi \cdot j(x) \cdot \Phi^{-1} = j(\sigma(x)) \qquad \forall x \in \mathrm{frac}(W(\mathbb{F}_{p^h}))$$

and

$$\frac{\mathrm{ord}(\Phi)}{\mathrm{ord}(p)} = \frac{m}{h} \pmod{\mathbb{Z}}\,.$$

Let's compute the Brauer invariant of $\mathrm{End}^0(H_{m,n})$. Let ϕ be the $W(K)$-linear endomorphism of M such that $\phi(e_i) = e_{i+m}$ for all $i = 0, 1, \dots, h-1$. Choose and fix an integer c such that $c\cdot n \equiv 1 \pmod{h}$. For every $x \in W(\mathbb{F}_h)$, let $\jmath(x)$ be the $W(K)$-linear endomorphism of M such that

$$\jmath(x)\colon e_i \;\mapsto\; \sigma^{ci}(x)\cdot e_i \qquad \forall\, i = 0, 1, \dots, h-1\,.$$

It is easy to see that ϕ (resp. $\jmath(x)$) commutes with $\mathcal{F}$ and $\mathcal{V}$, hence defines an element $\Phi \in \mathrm{End}(H_{m,n})$ (resp. $j(x) \in \mathrm{End}(H_{m,n})$), and the commutation relation

$$\Phi \circ j(x) = j(\sigma(x)) \circ \Phi \qquad \forall x \in W(\mathbb{F}_{p^h})$$

is satisfied. Because $\phi^h = p^m \cdot \mathrm{Id}_M$, we have $h \cdot \mathrm{ord}(\Phi) = m \cdot \mathrm{ord}(p)$. So the Brauer invariant of $\mathrm{End}^0(H_{m,n})$ is indeed $m/(m+n)$. See 4.52 (5) for an alternative proof, where $\mathrm{End}^0(H_{m,n})$ is identified with the opposite algebra D^{opp} of the central division algebra D over $\mathbb{Q}_p$ with Brauer invariant $n/(m+n)$.

7.5. The simple case, notation. We follow [**20**], §5, §6. In order to prove 7.1 in case X_0 is simple we fix notations, to be used for the rest of this section. Let $m \geq n > 0$ be relatively prime integers. We will write $r = (m-1)(n-1)/2$. We write δ for the isoclinic Newton polygon with slope $m/(m+n)$ with multiplicity $m+n$.

We want to understand all p-divisible groups isogenous to $H := H_{m,n}$ (m and n will remain fixed).

Lemma 7.6. **BB** *Work over a perfect field K. For every $X \sim H$ there is an isogeny $\varphi : H \to X$ of degree p^r.*

A proof of this lemma is not difficult and is left as an Exercise.

7.7. Construction. Consider the functor

$$S \quad \mapsto \quad \{(\varphi, X) \mid \varphi : H \times S \to X, \ \ \deg(\varphi) = p^r\}.$$

from the category of schemes over $\mathbb{F}_p$ to the category of sets. This functor is representable; denote the representing object by $(T = T_{m,n}, H_T \to \mathcal{G}) \to \mathrm{Spec}(\mathbb{F}_p)$. Note, using the lemma, that for any $X \sim H$ over a perfect field K there exists a point $x \in T(K)$ such that $X \cong \mathcal{G}_x$.

Discussion. The scheme $T = T_{m,n}$ constructed above is closely related to the Rapoport-Zink spaces $\mathcal{M} = \mathcal{M}(H_{m,n})$ in [**72**], Theorem 2.16, as follows. The formal scheme $\mathcal{M}$ represents a functor on the category Nilp of all $W(\mathbb{F}_p)$-schemes S such that p is locally nilpotent on S; the value $\mathcal{M}(S)$ for an object S in Nilp is the set of isomorphism classes $(X \to S, \rho : H_{m,n} \times_{\mathrm{Spec}(\mathbb{F}_p)} \overline{S} \to X \times_S \overline{S})$, where $X \to S$ is a p-divisible group, $\overline{S} = S \times_{\mathrm{Spec}(W(\mathbb{F}_p))} \mathrm{Spec}(\mathbb{F}_p)$, and ρ is a quasi-isogeny over $\overline{S}$. From the definition of T we get a morphism $f : T \to \mathcal{M}_r \times_{\mathrm{Spec}(W(\mathbb{F}_p))} \mathrm{Spec}(\mathbb{F}_p)$, where $\mathcal{M}_r$ is the open-and-closed formal subscheme whose points (X, ρ) have the property that the degree of the quasi-isogeny ρ is equal to p^r. Let $\overline{\mathcal{M}}_r^{\mathrm{red}}$ be the scheme with the same topological space as $\mathcal{M}_r$ whose structure sheaf is the quotient of $\mathcal{O}_{\mathcal{M}}/(p, \mathcal{I})$ by the nilpotent radical of $\mathcal{O}_{\mathcal{M}}/(p, \mathcal{I})$, where I is a sheaf of definition of the formal scheme $\mathcal{M}$. Let T^{red} be the reduced subscheme underlying T, and let $f^{\mathrm{red}} : T^{\mathrm{red}} \to \overline{\mathcal{M}}_r^{\mathrm{red}}$ be the morphism induced by f. Then Lemma 7.6 and the fact that $\mathrm{End}(H)^0$ is a division algebra imply that $f : T(k) \to \mathcal{M}_r(k)$ is a bijection for any algebraically closed field $k \supset \mathbb{F}_p$, so $f^{\mathrm{red}} : T^{\mathrm{red}} \to \overline{\mathcal{M}}_r^{\mathrm{red}}$ is an isomorphism.

Theorem 7.8. **Th** *The scheme T is geometrically irreducible of dimension r over $\mathbb{F}_p$. The set $T(a = 1) \subset T$ is open and dense in T.*

See [**20**], Theorem 5.11. *Note that* 7.1 *follows from this theorem in case $X_0 \sim H_{m,n}$.* We focus on a proof of 7.8.

Remark. Suppose we have proved the case that $X_0 \sim H_{m,n}$. Then by duality we have $X_0^t \sim H_{m,n}^t = H_{n,m}$, and this case follows also. Hence it suffices to consider only the case $m \geq n > 0$.

Notational Remark. In this section we will not consider abelian varieties. The letters A, B, etc. in this section will not be used for abelian varieties. Semi-modules will only be considered in this section and in later sections these letters again will be used for abelian varieties.

Definition 7.9. We say that $A \subset \mathbb{Z}$ is a *semi-module* or more precisely, an (m, n)-semi-module, if

- A is bounded from below, and if
- for every $x \in A$ we have $a + m, a + n \in A$.

We write $A = \{a_1, a_2, \dots\}$ with $a_j < a_{j+1}\ \forall j$. We say that semi-modules A, B are *equivalent* if there exists $t \in \mathbb{Z}$ such that $B = A + t := \{x + t \mid x \in A\}$.

We say that A is *normalized* if:

(1) $A \subset \mathbb{Z}_{\geq 0}$,
(2) $a_1 < \cdots < a_r \leq 2r$,
(3) $A = \{a_1, \dots, a_r\} \cup [2r, \infty)$;
notation: $[y, \infty) := \mathbb{Z}_{\geq y}$.

Write $A^t = \mathbb{Z}\backslash(2r - 1 - A) = \{y \in \mathbb{Z} \mid 2r - 1 - y \notin A\}$.

Explanation. For a semi-module A the set $\mathbb{Z}\backslash A$ of course is a "$(-m, -n)$-semi-module". Hence $\{y \mid y \notin A\}$ is a semi-module; then normalize.

Example. Write $\langle 0\rangle$ for the semi-module generated by 0, i.e., consisting of all integers of the form $im + jn$ for $i, j \geq 0$.

Exercise.

(4) Note that $\langle 0\rangle$ indeed is normalized. Show that $2r - 1 \notin \langle 0\rangle$.
(5) Show: if A is normalized then A^t is normalized.
(6) $A^{tt} = A$.
(7) For every B there is a unique normalized A such that $A \sim B$.
(8) If A is normalized then: $A = \langle 0\rangle \iff 0 \in A \iff 2r - 1 \notin A$.

7.10. Construction. *Work over a perfect field. For every $X \sim H_{m,n}$ there exists a semi-module.*

An isogeny $X \to H$ gives an inclusion

$$\mathbb{D}(X) \hookrightarrow \mathbb{D}(H) = M = \bigoplus_{0 \leq i < m+n} W{\cdot}e_i.$$

Write $M^{(i)} = \pi^i{\cdot}M$. Define

$$B := \{j \mid \mathbb{D}(X) \cap M^{(j)} \neq \mathbb{D}(X) \cap M^{(j+1)}\},$$

i.e., B is the set of values where the filtration induced on $\mathbb{D}(X)$ jumps. It is clear that B is a semi-module. Let A be the unique normalized semi-module equivalent to B.

Notation. The normalized semi-module constructed in this way will be called the *type* of X, denoted by $\mathrm{Type}(X)$.

Let A be a normalized semi-module. We denote by $U_A \subset T$ the set where the semi-module A is realized:

$$U_A = \{t \in T \mid \mathrm{Type}(\mathcal{G}_t) = A\}.$$

Proposition 7.11.

(1) *$U_A \hookrightarrow T$ is locally closed, $T = \bigsqcup_A U_A$.*
(2) *$A = \langle 0\rangle \iff a(X) = 1$.*
(3) *$U_{\langle 0\rangle}$ is geometrically irreducible and has dimension r.*
(4) *If $A \neq \langle 0\rangle$ then every component of U_A has dimension strictly less than r.*

For a proof see [**20**], the proof on page 233, and 6.5 and 6.15. The argument is not very deep but somewhat involved (combinatorics and studying explicit equations).

7.12. **BB** Let Y_0 be any p-divisible group over a field K, of dimension d and let c be the dimension of Y_0^t. The universal deformation space is isomorphic to $\mathrm{Spf}(K[[t_1, \cdots, t_{cd}]])$ and the generic fiber of that universal deformation is ordinary; in this case its Newton polygon ρ has c slopes equal to 1 and d slopes equal to 0. See 2.5. See [**38**], 4.8, [**20**], 5.15.

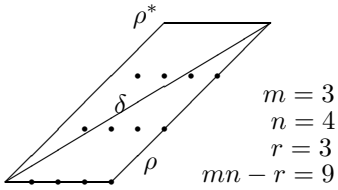
7.13. We prove 7.8, using 7.3 and 7.11. Note that the Zariski closure $(U_{\langle 0\rangle})^{\mathrm{Zar}} \subset T$ is geometrically irreducible, and has dimension r; we want to show equality $(U_{\langle 0\rangle})^{\mathrm{Zar}} = T$. Suppose there would be an irreducible component T' of T not contained in $(U_{\langle 0\rangle})^{\mathrm{Zar}}$. By 7.11 (3) and (4) we see that $\dim(T') < r$. Let $y \in T'$, with corresponding p-divisible group Y_0.

Consider the formal completion $T^{/y}$ of T at y. Write $D = \mathrm{Def}(Y_0)$ for the universal deformation space of Y_0. The moduli map $T^{/y} \to D = \mathrm{Def}(Y_0)$ is an immersion, see [**20**], 5.19. Let $T'' \subset D$ be the image of $(T')^{/y}$ in D; we conclude that no irreducible component of T'' is contained in any irreducible component of the image of $T^{/y} \to D$ in D, i.e., every component of T'' is an component of $\mathcal{W}_\delta(D)$. Clearly $\dim(T') = \dim(T'') < r$.

Obvious, but crucial observation. *Consider the graph of all Newton polygons*

$$\zeta \quad \textit{with} \quad \delta \prec \zeta \prec \rho.$$

The longest path in this graph has length $\leq mn - r$.



Proof. Consider the Newton polygon ρ, in this case given by n slopes equal to 0 and m slopes equal to 1. Note that $\gcd(m, n) = 1$, hence the Newton polygon δ does not contain integral points except its beginning and end points. Consider the interior of the parallelogram given by ρ and by ρ^*, the upper convex polygon given by: first m slopes equal to 1 and then n slopes equal to 0. The number of interior points of this parallelogram equals $(m-1)(n-1)$. Half of these are above δ, and half of these are below δ. Write $\delta \precneqq (i, j)$ for the property "(i, j) is strictly below δ", and $(i, j) \prec \rho$ for "(i, j) is upon or above ρ". We see:

$$\#\left(\{(i,j) \mid \delta \precneqq (i,j) \prec \rho\}\right) = (m-1)(n-1)/2 + (m+n-1) = mn - r.$$

We use the following fact: If $\zeta_1 \precneqq \zeta_2$, then there is an integral point on ζ_2 strictly below ζ_1. One can even show that all maximal chains of Newton polygons in the fact above have the same length, and in fact equal to

$$\#\left(\{(i,j) \mid \delta \precneqq (i,j) \prec \rho\}\right).$$

This finishes the proof of the claim. □

As $\dim(\mathrm{Def}(Y_0)) = mn$ this observation implies by purity, see 7.3, that every irreducible component of $\mathcal{W}_\delta(D)$ had dimension at least r. This is a contradiction to the assumption of the existence of T', i.e., $\dim(T') = \dim(T'') < r$. Hence $(U_{\langle 0\rangle})^{\mathrm{Zar}} = T$. This proves Theorem 7.8. Hence we have proved Theorem 7.1 in the case when X_0 is isogenous to $H_{m,n}$. □

Theorem 7.14. **Th** (Deformation to $a \leq 1$ in the principally quasi-polarized case) *Let X_0 be a p-divisible group over a field K with a principal quasi-polarization $\lambda_0 : X_0 \to X_0^t$. There exists an integral scheme S, a point $0 \in S(K)$ and a principally quasi-polarized p-divisible group $(\mathcal{X}, \lambda) \to S$ such that there is an isomorphism $(X_0, \lambda_0) \cong (\mathcal{X}, \lambda)_0$, and for the generic point $\eta \in S$ we have:*

$$\mathcal{N}(X_0) = \mathcal{N}(X_\eta) \quad \textit{and} \quad a(X_\eta) \leq 1.$$

See [**20**], 5.12 and [**66**], 3.10.

Corollary 7.15. **Th** (Deformation to $a \leq 1$ in the case of principally polarized abelian varieties) *Let (A_0, λ_0) be a principally polarized abelian variety over K. There exists an integral scheme S, a point $0 \in S(K)$ and a principally polarized abelian scheme $(A, \lambda) \to S$ such that there is an isomorphism $(A_0, \lambda_0) \cong (A, \lambda)_0$, and for the generic point $\eta \in S$ we have*

$$\mathcal{N}(A_0) = \mathcal{N}(A_\eta) \quad \textit{and} \quad a(X_\eta) \leq 1.$$

7.16. The non-principally polarized case. Note that the analog of the theorem and of the corollary is not correct in general in the *non-principally polarized* case. Here is an example, see [**39**], 6.10, and also see [**45**], 12.4 and 12.5 where more examples are given. *Consider $g = 3$, let σ be the supersingular Newton polygon; it can be proved that for any $x \in \mathcal{W}_\sigma(\mathcal{A}_{3,p})$ we have $a(A_x) \geq 2$.*

We will show that for $\xi_1 \prec \xi_2$ we have in the principally polarized case:

$$\mathcal{W}^0_{\xi_1}(\mathcal{A}_{g,1}) =: W^0_{\xi_1} \quad \subset \quad (W^0_{\xi_2})^{\mathrm{Zar}} = W_{\xi_2} := \mathcal{W}_{\xi_2}(\mathcal{A}_{g,1}).$$

In the non-principally polarized case this inclusion and the equality $(W^0_{\xi_2})^{\mathrm{Zar}} = W_{\xi_2}$ do not hold in general as is demonstrated by the following example. Let $g = 3$, and $\xi_1 = \sigma$ the supersingular Newton polygon, and $\xi_2 = (2, 1) + (1, 2)$. Clearly $\xi_1 \prec \xi_2$. By [**39**], 6.10, there is a component of $\mathcal{W}_\sigma(\mathcal{A}_{g,p^2})$ of dimension 3; more generally see [**45**], Theorem 10.5 (ii) for the case of $\mathcal{W}_\sigma(\mathcal{A}_{g,p^{[(g-1)^2/2]}})$ and components of dimension equal to $g(g-1)/2$. As the p-rank 0 locus in $\mathcal{A}_g$ has pure dimension equal to $g(g+1)/2 + (f-g) = g(g-1)/2$, see [**56**], Theorem 4.1, this shows the existence of a polarized supersingular abelian variety (of dimension 3, respectively of any dimension at least 3) which cannot be deformed to a non-supersingular abelian variety with p-rank equal to zero.

Many more examples where $(W^0_\xi)^{\mathrm{Zar}} \neq W_\xi$ follow from [**58**], Section 3.

8. Proof of the Grothendieck conjecture

Main reference: [**66**].

Definition 8.1. **Extra** Let X be a p-divisible group over a base S. A filtration

$$0 = X^{(0)} \subset X^{(1)} \subset \cdots \subset X^{(s)} = X$$

of X by p-divisible subgroups $X_i \to S$ is *the slope filtration* of X if there exists rational numbers $\tau_1, \tau_2, \ldots, \tau_s$ with $1 \geq \tau_1 > \tau_2 > \cdots > \tau_s \geq 0$ such that $Y_i := X^{(i)}/X^{(i-1)}$ is an isoclinic p-divisible group over S with slope τ_i for $i = 1, \ldots, s$.

Remark. Clearly, if a slope filtration exists, it is unique.

From the Dieudonné-Manin classification it follows that the slope filtration on X exists if K is perfect.

By Grothendieck and Zink we know that for every p-divisible group over any field K the slope filtration exists; see [**94**], Corollary 13.

In general for a p-divisible group $X \to S$ over a base a slope filtration on X/S does not exists. Even if the Newton polygon is constant in a family, in general the slope filtration does not exist.

Definition 8.2. We say that $0 = X^{(0)} \subset X^{(1)} \subset \cdots \subset X^{(s)} = X$ is a *maximal filtration* of $X \to S$ if every geometric fiber of $Y^{(i)} := X^{(i)}/X^{(i-1)}$ for $1 \leq i \leq s$ is *simple* and isoclinic of slope τ_i with $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_s$.

Lemma. **BB** *For every X over k a maximal filtration exists.*

See [**66**], 2.2.

Lemma 8.3. **BB** *Let $\{X_0^{(i)}\}$ be a p-divisible group X_0 with maximal filtration over k. There exists an integral scheme S and a p-divisible group X/S with a maximal filtration $\{X^{(i)}\} \to S$ and a closed point $0 \in S(k)$ such that $\mathcal{N}(Y^{(i)})$ is constant for $1 \leq i \leq s$, such that $\{X^{(i)}\}_0 = \{X_0^{(i)}\}$ and such that for the generic point $\eta \in S$ we have $a(X_\eta) \leq 1$.*

See [**66**], Section 2. A proof of this lemma uses Theorem 7.8.

In Section 7 we proved 7.8, and obtained as a corollary 7.1 in the case of a simple p-divisible group. From the previous lemma we derive a proof for Theorem 7.1.

Definition 8.4. We say that a p-divisible group X_0 over a field K is a *specialization* of a p-divisible group X_η over a field L if there exists an integral scheme $S \to \mathrm{Spec}(K)$, a k-rational point $0 \in S(K)$, and $\mathcal{X} \to S$ such that $X_0 = \mathcal{X}_0$, and for the generic point $\eta \in S$ we have $L = K(\eta)$ and $X_\eta = \mathcal{X}_\eta$.

This can be used for p-divisible groups, for abelian schemes, etc.

Proposition 8.5. *Let X_0 be a specialization of $X_\eta = Y_0$, and let Y_0 be a specialization of Y_ρ. Then X_0 is a specialization of Y_ρ.*

Using Theorem 5.10 and Theorem 7.1 along with the proposition above, we derive a proof of the Grothendieck Conjecture (Theorem 1.22). □

Corollary 8.6. (of Theorem 1.22) *Let X_0 be a p-divisible group, $\beta = \mathcal{N}(X_0)$. Every component of the locus $\mathcal{W}_\beta(\mathrm{Def}(X_0))$ has dimension $\diamond(\beta)$.*

Definition 8.7. Let (X, λ) be a principally polarized p-divisible group over S. We say that a filtration

$$0 = X^{(0)} \subset X^{(1)} \subset \cdots \subset X^{(s)} = X$$

of X by p-divisible subgroups $X^{(0)}, \ldots, X^{(s)}$ over S is *a maximal symplectic filtration* of (X, λ) if:

- each quotient $Y^{(i)} := X^{(i)}/X^{(i-1)}$ for $i = 1, \ldots, s$ is a p-divisible group over S,
- every geometric fiber of $Y^{(i)}$ for $1 \leq i \leq s$ is simple of slope τ_i,
- $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_s$, and
- $\lambda : X \to X^t$ induces an isomorphism

$$\lambda_i : Y^{(i)} \to (Y^{(s+1-i)})^t \quad \text{for} \quad 0 < i \leq (s+1)/2.$$

Lemma 8.8. *For every principally polarized (X, λ) over k there exists a maximal symplectic filtration.*

See [**66**], 3.5.

8.9. Using this definition and this lemma, we show the principally polarized analog 7.15 of 7.14; see [**66**], Section 3. Hence Corollary 7.15 follows. Using 7.15 and Theorem 5.19 we derive a proof for:

Theorem 8.10. (An analog of the Grothendieck conjecture) *Let $K \supset \mathbb{F}_p$. Let (X_0, λ_0) be a principally quasi-polarized p-divisible group over K. Write $\mathcal{N}(X_0) = \xi$ for its Newton polygon. Given a Newton polygon ζ "below" ξ, i.e., $\xi \prec \zeta$, there exists a deformation (X_η, λ_η) of (X_0, λ_0) such that $\mathcal{N}(\mathcal{X}_\eta) = \zeta$.*

Corollary 8.11. *Let $K \supset \mathbb{F}_p$. Let (A_0, λ_0) be a principally polarized abelian variety over K. Write $\mathcal{N}(A_0) = \xi$ for its Newton polygon. Given a Newton polygon ζ "below" ξ, i.e., $\xi \prec \zeta$, there exists a deformation (A_η, λ_η) of (A_0, λ_0) such that $\mathcal{N}(\mathcal{X}_\eta) = \zeta$.*

Corollary 8.12. *Let ξ be a symmetric Newton polygon. Every component of the stratum $W_\xi = \mathcal{W}_\xi(\mathcal{A}_{g,1})$ has dimension equal to $\triangle(\xi)$.*

9. Proof of the density of ordinary Hecke orbits

In this section we give a proof of Theorem 1.8 on the density of ordinary Hecke orbits, restated as Theorem 9.1 below. To establish Theorem 1.8, we need the analogous statement for a Hilbert modular variety; see 9.2 for the precise statement.

Here is a list of tools we will use; many have been explained in previous sections.

(i) Serre-Tate coordinates, see §2.
(ii) Local stabilizer principle, see 9.5 and 9.6.
(iii) Local rigidity for group actions on formal tori, see 2.26.
(iv) Consequence of EO stratification, see 9.7.
(iv) Hilbert trick, see 9.10.

The logical structure of the proof of Theorem 1.8 is as follows. We first prove the density of ordinary Hecke orbits on Hilbert modular varieties. Then we use the Hilbert trick to show that the Zariski closure of any prime-to-p Hecke orbit on $\mathcal{A}_{g,1,n}$ contains a *hypersymmetric ordinary point*. Finally we use the local stabilizer principle and the local rigidity to conclude the proof of 1.8. Here by a hypersymmetric ordinary point we just mean that the underlying abelian variety is isogenous to $E \times \cdots \times E$, where E is an ordinary elliptic curve over $\mathbb{F}$; see [**15**] for the general notion of hypersymmetric abelian varieties.

The Hilbert trick is based on the following observation. Given an ordinary point $x = [(A_x, \lambda_x, \eta_x)] \in \mathcal{A}_{g,1,n}(\mathbb{F})$, the prime-to-$p$ Hecke orbit of x contains, up to a possibly inseparable isogeny correspondence, the (image of) the prime-to-p Hecke orbit of a point $h = [(A_y, \iota_y, \lambda_y, \eta_y)]$ of a Hilbert modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}$ such that A_y is isogenous to A_x, because $\mathrm{End}^0(A_x)$ contains a product $E = F_1 \times \cdots \times F_r$ of totally real fields with $[E : \mathbb{Q}] = g$. So if we can establish the density of the prime-to-p Hecke orbit of y in $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}$, then we know that the Zariski closure of the prime-to-p Hecke orbit of x contains the image of the Hilbert modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}$ in $\mathcal{A}_{g,1,n}$ under a finite isogeny correspondence, i.e., a scheme T over $\mathbb{F}$ and finite $\mathbb{F}$-morphisms $g : T \to \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}$ and $f : T \to \mathcal{A}_{g,1,n}$ such that the pullback by g of the universal abelian scheme over $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}$ is isogenous to the pullback by f of the universal abelian scheme over $\mathcal{A}_{g,1,n}$. Since $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}$

contains ordinary hypersymmetric points, $\left(\mathcal{H}^{(p)}_{\mathrm{Sp}}(x)\right)^{\mathrm{Zar}}$ also contains an ordinary hypersymmetric point. Then the linearization method afforded by the combination of the *local stabilizer principle* and the *local rigidity* implies that the dimension of $\left(\mathcal{H}^{(p)}_{\mathrm{Sp}}(x)\right)^{\mathrm{Zar}}$ is equal to $g(g+1)/2$, hence $\left(\mathcal{H}^{(p)}_{\mathrm{Sp}}(x)\right)^{\mathrm{Zar}} = \mathcal{A}_{g,1,n}$ because $\mathcal{A}_{g,1,n}$ is geometrically irreducible, see 10.26.

To prove the density of ordinary Hecke orbits on a Hilbert modular variety, the linearization method is again crucial. Since a Hilbert modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}$ is "small", there are only a finite number of possibilities as to what (the formal completion of) the Zariski closure of an ordinary Hecke orbit can be; the possibilities are indexed by the set of all subsets of prime ideals of $\mathcal{O}_E$. To pin down the number of possibilities down to one, one can use either the consequence of EO-stratification that the Zariski closure of any Hecke-invariant subvariety of a Hilbert modular variety contains a supersingular point, or de Jong's theorem on extending homomorphisms between p-divisible groups. We follow the first approach here; see 9.11 and [**7**, §8] for the second approach.

Theorem 9.1. *Let $n \geq 3$ be an integer prime to p. Let $x = [(A_x, \lambda_x, \eta_x)] \in \mathcal{A}_{g,1,n}(\mathbb{F})$ such that A_x is ordinary.*

(i) *The prime-to-p $\mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$-Hecke orbit of x is dense in the moduli space $\mathcal{A}_{g,1,n}$ over $\mathbb{F}$ for any prime number $\ell \neq p$, i.e.,*

$$\left(\mathcal{H}^{(p)}_{\mathrm{Sp}}(x)\right)^{\mathrm{Zar}} = \mathcal{A}_{g,1,n}\,.$$

(ii) *The $\mathrm{Sp}_{2g}(\mathbb{Q}_\ell)$-Hecke orbit of x dense in the moduli space $\mathcal{A}_{g,1,n}$ over $\mathbb{F}$, i.e.,*

$$\left(\mathcal{H}^{\mathrm{Sp}}_{\ell}(x)\right)^{\mathrm{Zar}} = \mathcal{A}_{g,1,n}\,.$$

Theorem 9.2. *Let $n \geq 3$ be an integer prime to p. Let $E = F_1 \times \cdots \times F_r$, where $F_1, \ldots, F_r$ are totally real number fields. Let $\mathcal{L}$ be an invertible $\mathcal{O}_E$-module, and let $\mathcal{L}^+$ be a notion of positivity for $\mathcal{L}$. Let $y = [(A_y, \iota_y, \lambda_y, \eta_y)] \in \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}(\mathbb{F})$ be a point of the Hilbert modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ such that A_y is ordinary. Then the $\mathrm{SL}_2(E \otimes_{\mathbb{Q}} \mathbb{A}_f^{(p)})$-Hecke orbit of y on $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ is Zariski dense in $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ over $\mathbb{F}$.*

Proposition 9.3. *Let $n \geq 3$ be a integer prime to p.*

(i) *Let $x \in \mathcal{A}_{g,1,n}(\mathbb{F})$ be a closed point of $\mathcal{A}_{g,1,n}$. Let $Z(x)$ be the Zariski closure of the prime-to-p Hecke orbit $\mathcal{H}^{(p)}_{\mathrm{Sp}}(x)$ in $\mathcal{A}_{g,1,n}$ over $\mathbb{F}$. Then $Z(x)$ is smooth at x over $\mathbb{F}$.*

(ii) *Let $y \in \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}(\mathbb{F})$ be a closed point of a Hilbert modular variety $\mathcal{M}_{F,\mathcal{L},\mathcal{L}^+,n}$. Let $Z_F(y)$ be the Zariski closure of the prime-to-p Hecke orbit $\mathcal{H}^{(p)}_{\mathrm{SL}_2}(y)$ on $\mathcal{M}_{F,\mathcal{L},\mathcal{L}^+,n}$ over $\mathbb{F}$. Then $Z_F(y)$ is smooth at y over $\mathbb{F}$.*

Proof. We give the proof of (ii) here. The proof of (i) is similar and left to the reader.

Because Z_F is reduced, there exists a dense open subset $U \subset Z_F$ which is smooth over $\mathbb{F}$. This open subset U must contain an element y' of the dense subset $\mathcal{H}^{(p)}_{\mathrm{SL}_2}(y)$ of Z_F, so Z_F is smooth over $\mathbb{F}$ at y'. Since prime-to-p Hecke

correspondences are defined by schemes over $\mathcal{M}_{F,\mathcal{L},\mathcal{L}^+,n} \times_{\mathrm{Spec}(\mathbb{F})} \mathcal{M}_{F,\mathcal{L},\mathcal{L}^+,n}$ such that both projections to $\mathcal{M}_{F,\mathcal{L},\mathcal{L}^+,n}$ are étale, Z_F is smooth at y as well. □

Remark.

(i) Proposition 9.3 is an analog of the following well-known fact. Let X be a reduced scheme over an algebraically closed field k on which an algebraic group operates transitively. Then X is smooth over k.

(ii) The proof of Proposition 9.3 also shows that all irreducible components of $Z(x)$ (resp. $Z_F(y)$) have the same dimension: For any non-empty subset $U_1 \subset Z_F(y)$ and any open subset $W_1 \ni y$, there exist a non-empty subset $U_2 \subset U_1$, an open subset $W_2 \ni y$ and a non-empty étale correspondence between U_2 and W_2.

Theorem 9.4. **BB** *Let Z be a reduced closed subscheme of $\mathcal{A}_{g,1,n}$ over $\mathbb{F}$ such that no maximal point of Z is contained in the supersingular locus of $\mathcal{A}_{g,1,n}$. If Z is stable under all $\mathrm{Sp}_{2g}(\mathbb{Q}_\ell)$-Hecke correspondences on $\mathcal{A}_{g,1,n}$, then Z is stable under all $\mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$-Hecke correspondences.*

Remark. This is proved in [**11**, Proposition 4.6].

Local stabilizer principle

Let $k \supset \mathbb{F}_p$ be an algebraically closed field. Let Z be a reduced closed subscheme of $\mathcal{A}_{g,1,n}$ over k. Let $z = [(A_z, \lambda_z, \eta_z)] \in Z(k) \subset \mathcal{A}_{g,1,n}(k)$ be a closed point of Z. Let $*_z$ be the Rosati involution on $\mathrm{End}^0(A_z)$. Denote by H_z the unitary group attached to the semisimple algebra with involution $(\mathrm{End}^0(A_z), *_z)$, defined by

$$\mathrm{H}_z(R) = \left\{ x \in (\mathrm{End}^0(A_z) \otimes_{\mathbb{Q}} R)^\times \mid x \cdot *_0(x) = *_0(x) \cdot x = \mathrm{Id}_{A_z} \right\}$$

for any $\mathbb{Q}$-algebra R. Denote by $\mathrm{H}_z(\mathbb{Z}_p)$ the subgroup of $\mathrm{H}_z(\mathbb{Q}_p)$ consisting of all elements $x \in \mathrm{H}_z(\mathbb{Q}_p)$ such that x induces an automorphism of $(A_z, \lambda_z)[p^\infty]$. Denote by $\mathrm{H}_z(\mathbb{Z}_{(p)})$ the group $\mathrm{H}_z(\mathbb{Q}) \cap \mathrm{H}_z(\mathbb{Z}_p)$, i.e., its elements consist of all elements $x \in \mathrm{H}_z(\mathbb{Q})$ such that x induces an automorphism of $(A_z, \lambda_z)[p^\infty]$. Note that the action of $\mathrm{H}_z(\mathbb{Z}_p)$ on $A_z[p^\infty]$ makes $\mathrm{H}_z(\mathbb{Z}_p)$ a subgroup of $\mathrm{Aut}((A_z, \lambda_z)[p^\infty])$. Denote by $\mathcal{A}_{g,1,n}^{/z}$ (resp. $Z^{/z}$) the formal completion of $\mathcal{A}_{g,1,n}$ (resp. Z) at z. The compact p-adic group $\mathrm{Aut}((A_z, \lambda_z)[p^\infty])$ operates naturally on the deformation space $\mathrm{Def}\,((A_z, \lambda_z)[p^\infty]/k)$. So we have a natural action of $\mathrm{Aut}((A_z, \lambda_z)[p^\infty])$ on the formal scheme $\mathcal{A}_{g,1,n}^{/z}$ via the canonical isomorphism

$$\mathcal{A}_{g,1,n}^{/z} = \mathrm{Def}\,((A_z, \lambda_z)/k) \xrightarrow[\sim]{\text{Serre-Tate}} \mathrm{Def}\,((A_z, \lambda_z)[p^\infty]/k)\ .$$

Theorem 9.5. (local stabilizer principle) *Notation as above. Suppose that Z is stable under all $\mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$-Hecke correspondences on $\mathcal{A}_{g,1,n}$. Then the closed formal subscheme $Z^{/z}$ in $\mathcal{A}_{g,1,n}^{/z}$ is stable under the action of the subgroup $\mathrm{H}_z(\mathbb{Z}_p)$ of $\mathrm{Aut}((A_z, \lambda_z)[p^\infty])$.*

Proof. Consider the projective system $\mathcal{A}_{g,1}^{\sim} = \varprojlim_m \mathcal{A}_{g,1,m}$ over k, where m runs through all integers $m \geq 1$ which are prime to p. The pro-scheme $\mathcal{A}_{g,1}^{\sim}$ classifies triples $(A \to S, \lambda, \eta)$, where $A \to S$ is an abelian scheme up to prime-to-p isogenies, λ is a principal polarization of $A \to S$, and

$$\eta : \mathrm{H}_1(A_z, \mathbb{A}_f^{(p)}) \xrightarrow{\sim} \underline{\mathrm{H}}_1(A/S, \mathbb{A}_f^{(p)})$$

is a symplectic prime-to-p level structure. Here we have used the first homology groups of A_z attached to the base point z to produce the standard representation of the symplectic group Sp_{2g}. Take $S_z = \mathcal{A}^{/z}_{g,1,n}$, let $(A^{/z}, \lambda^{/z}) \to \mathcal{A}^{/z}_{g,1,n}$ be the restriction of the universal principally polarized abelian scheme to $\mathcal{A}^{/z}_{g,1,n}$, and let $\eta^{/z}$ be the tautological prime-to-p level structure, we get an S_z-point of the tower $\mathcal{A}^{\sim}_{g,1}$ that lifts $S_z \hookrightarrow \mathcal{A}_{g,1,n}$.

Let γ be an element of $\mathrm{H}_z(\mathbb{Z}_{(p)})$. Let γ_p (resp. $\gamma^{(p)}$) be the image of γ in the local stabilizer subgroup $\mathrm{H}_z(\mathbb{Z}_p) \subset \mathrm{Aut}((A_z, \lambda_z)[p^\infty])$ (resp. in $\mathrm{H}_z(\mathbb{A}_f^{(p)})$). From the definition of the action of $\mathrm{Aut}((A_z, \lambda_z)[p^\infty])$ on $\mathcal{A}^{/z}_{g,1,n}$ we have a commutative diagram

$$\begin{array}{ccc} (A^{/z}, \lambda^{/z})[p^\infty] & \xrightarrow{f_\gamma[p^\infty]} & (A^{/z}, \lambda^{/z})[p^\infty] \\ \downarrow & & \downarrow \\ \mathcal{A}^{/z}_{g,1,n} & \xrightarrow{u_\gamma} & \mathcal{A}^{/z}_{g,1,n} \end{array}$$

where u_γ is the action of γ_p on $\mathcal{A}^{/z}_{g,1,n}$ and $f_\gamma[p^\infty]$ is an isomorphism over u_γ whose fiber over z is equal to γ_p. Since γ_p comes from a prime-to-p quasi-isogeny, $f_\gamma[p^\infty]$ extends to a prime-to-p quasi-isogeny f_γ over u_γ, such that the diagram

$$\begin{array}{ccc} A^{/z} & \xrightarrow{f_\gamma} & A^{/z} \\ \downarrow & & \downarrow \\ \mathcal{A}^{/z}_{g,1,n} & \xrightarrow{u_\gamma} & \mathcal{A}^{/z}_{g,1,n} \end{array}$$

commutes and f_γ preserves the polarization $\lambda^{/z}$. Clearly the fiber of f_γ at z is equal to γ as a prime-to-p isogeny from A_z to itself. From the definition of the action of the symplectic group $\mathrm{Sp}(\mathrm{H}_1(A_z, \mathbb{A}_f^{(p)}), \langle \cdot, \cdot \rangle)$ one sees that u_γ coincides with the action of $(\gamma^{(p)})^{-1}$ on $\mathcal{A}^{\sim}_{g,1}$. Since Z is stable under all $\mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$-Hecke correspondences, we conclude that $Z^{/z}$ is stable under the action of u_γ, for every $\gamma \in \mathrm{H}_z(\mathbb{Z}_{(p)})$.

By the weak approximation theorem for linear algebraic groups (see [**70**], 7.3, Theorem 7.7 on page 415), $\mathrm{H}_z(\mathbb{Z}_{(p)})$ is p-adically dense in $\mathrm{H}_z(\mathbb{Z}_p)$. So $Z^{/z}$ is stable under the action of $\mathrm{H}_z(\mathbb{Z}_p)$ by the continuity of the action of $\mathrm{Aut}((A_z, \lambda_z)[p^\infty])$. □

Remark. The group $\mathrm{H}_z(\mathbb{Z}_{(p)})$ can be thought of as the "stabilizer subgroup" at z inside the family of prime-to-p Hecke correspondences: Every element $\gamma \in \mathrm{H}_z(\mathbb{Z}_{(p)})$ gives rise to a prime-to-p Hecke correspondence with z as a fixed point.

We set up notation for the local stabilizer principle for Hilbert modular varieties. Let $E = F_1 \times \cdots \times F_r$, where $F_1, \ldots, F_r$ are totally real number fields. Let $\mathcal{L}$ be an invertible $\mathcal{O}_E$-module, and let $\mathcal{L}^+$ be a notion of positivity for $\mathcal{L}$. Let $m \geq 3$ be a positive integer which is prime to p. Let Y be a reduced closed subscheme of $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}$ over $\mathbb{F}$. Let $y = [(A_y, \iota_y, \lambda_y, \eta_y)] \in \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}(\mathbb{F})$ be a closed point in $Y \subset \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}$. Let $*_y$ be the Rosati involution attached to λ on the semisimple algebra $\mathrm{End}^0_{\mathcal{O}_E}(A_y) = \mathrm{End}_{\mathcal{O}_E}(A_y) \otimes_{\mathcal{O}_E} E$. Denote by H_y the unitary group over $\mathbb{Q}$

attached to $(\mathrm{End}^0_{\mathcal{O}_E}(A_y), *_y)$, so

$$\mathrm{H}_y(R) = \left\{ u \in \left(\mathrm{End}^0_{\mathcal{O}_E}(A_y) \otimes_{\mathbb{Q}} R\right)^\times \mid u \cdot *_y(u) = *_y(u) \cdot u = \mathrm{Id}_{A_y} \right\}$$

for every $\mathbb{Q}$-algebra R. Let $\mathrm{H}_y(\mathbb{Z}_p)$ be the subgroup of $\mathrm{H}_y(\mathbb{Q}_p)$ consisting of all elements of $\mathrm{H}_y(\mathbb{Q}_p)$ which induce an automorphism of $(A_y[p^\infty], \iota_y[p^\infty], \lambda_y[p^\infty])$. Denote by $\mathrm{H}_y(\mathbb{Z}_{(p)})$ the intersection of $\mathrm{H}_y(\mathbb{Q})$ and $\mathrm{H}_y(\mathbb{Z}_p)$ inside $\mathrm{H}_y(\mathbb{Q}_p)$, i.e., it consists of all elements $u \in \mathrm{H}_y(\mathbb{Q})$ such that u induces an automorphism of $(A_y, \iota_y, \lambda_y)[p^\infty]$.

The compact p-adic group $\mathrm{Aut}((A_y, \iota_y, \lambda_y)[p^\infty])$ operates naturally on the local deformation space $\mathrm{Def}\,((A_y, \iota_y, \lambda_y)[p^\infty]/k)$. So we have a natural action of the compact p-adic group $\mathrm{Aut}((A_y, \iota_y, \lambda_y)[p^\infty])$ on the formal scheme $\mathcal{M}^{/y}_{E,\mathcal{L},\mathcal{L}^+,m}$ via the canonical isomorphism

$$\mathcal{M}^{/y}_{E,\mathcal{L},\mathcal{L}^+,m} = \mathrm{Def}\,((A_y, \iota_y, \lambda_y)/k) \xrightarrow[\sim]{\text{Serre-Tate}} \mathrm{Def}\,((A_y, \iota_y, \lambda_y)[p^\infty]/k)\ .$$

Theorem 9.6. *Notation as above. Assume that the closed subscheme*

$$Y \subset \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}$$

over $\mathbb{F}$ is stable under all $\mathrm{SL}_2(E \otimes_{\mathbb{Q}} \mathbb{A}_f^{(p)})$-Hecke correspondences on the Hilbert modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}$. Then the closed formal subscheme $Y^{/y}$ of $\mathcal{M}^{/y}_{E,\mathcal{L},\mathcal{L}^+,m}$ is stable under the action by elements of the subgroup

$$\mathrm{H}_y(\mathbb{Z}_p) \subset \mathrm{Aut}(A_y[p^\infty], \iota_y[p^\infty], \lambda_y[p^\infty]).$$

Proof. The proof of Theorem 9.6 is similar to that of Theorem 9.5, and is already contained in the proof of Corollary 6.15. □

Theorem 9.7. **BB** *Let $n \geq 3$ be an integer relatively prime to p. Let ℓ be a prime number, $\ell \neq p$.*

(i) *Every closed subset of $\mathcal{A}_{g,n}$ over $\mathbb{F}$ which is stable under all Hecke correspondences on $\mathcal{A}_{g,n}$ coming from $\mathrm{Sp}_{2g}(\mathbb{Q}_\ell)$ contains a supersingular point.*

(ii) *Similarly, every closed subset in a Hilbert modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ over $\mathbb{F}$ which is stable under all $\mathrm{SL}_2(E \otimes \mathbb{Q}_\ell)$-Hecke correspondences on $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ contains a supersingular point.*

Remark. Theorem 9.7 follows from the main theorem of [**67**] and Proposition 9.8 below. See also 3.22.

Proposition 9.8. **BB** *Let $k \supset \mathbb{F}_p$ be an algebraically closed field. Let ℓ be a prime number, $\ell \neq p$. Let $n \geq 3$ be an integer prime to p.*

(i) *Let $x = [(A_x, \lambda_x, \eta_x)] \in \mathcal{A}_{g,1,n}(k)$ be a closed point of $\mathcal{A}_{g,1,n}$. If A_x is supersingular, then the prime-to-p Hecke orbit $\mathcal{H}^{(p)}_{\mathrm{Sp}_{2g}}(x)$ is finite. Conversely, if A_x is not supersingular, then the ℓ-adic Hecke orbit $\mathcal{H}^{\mathrm{Sp}_{2g}}_{\ell}(x)$ is infinite for every prime number $\ell \neq p$.*

(ii) *Let $y = [(A_y, \iota_y, \lambda_y, \eta_y)] \in \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}(k)$ be a closed point of a Hilbert modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$. If A_y is supersingular, then the prime-to-p Hecke orbit $\mathcal{H}^{(p)}_{\mathrm{SL}_{2,E}}(y)$ is finite. Conversely, if A_y is not supersingular, then the v-adic Hecke orbit $\mathcal{H}^{\mathrm{SL}_{2,E}}_{v}(y)$ is infinite for every prime ideal $\wp_v$ of $\mathcal{O}_E$ which does not contain p.*

Remark.

(1) The statement (i) is proved in Proposition 1, p. 448 of [**9**], see 1.14. The proof of (ii) is similar. The key to the proof of the second part of (i) is a bijection

$$\mathcal{H}_\ell^{\mathrm{Sp}_{2g}}(x) \overset{\sim}{\leftarrow} \left(\mathrm{H}_x(\mathbb{Q}) \cap \prod_{\ell' \neq \ell} K_\ell \right) \backslash \mathrm{Sp}_{2g}(\mathbb{Q}_\ell)/K_\ell$$

where ℓ' runs through all prime numbers not equal to ℓ or p, H_x is the unitary group attached to $(\mathrm{End}^0(A_x), *_x)$ as in Theorem 9.5. The compact groups $K_{\ell'}$ and K_ℓ are defined as follows: for every prime number $\ell' \neq p$, $K_{\ell'} = \mathrm{Sp}_{2g}(\mathbb{Z}_{\ell'})$ if $(\ell', n) = 1$, and $K_{\ell'}$ consists of all elements $u \in \mathrm{Sp}_{2g}(\mathbb{Z}_{\ell'})$ such that $u \equiv 1 \pmod{n}$ if $\ell' | n$. We have an injection $\mathrm{H}_x(\mathbb{A}_f^{(p)}) \to \mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$ as in Theorem 9.5, so that the intersection $\mathrm{H}_x(\mathbb{Q}) \cap \prod_{\ell' \neq \ell} K_\ell$ makes sense. The second part of (i) follows from the group-theoretic fact that a double coset as above is finite if and only if H_x is a form of Sp_{2g}.

(2) When the abelian variety A_x in (i) (resp. A_y in (ii)) is ordinary, one can also use the canonical lifting to $W(k)$ to show that $\mathcal{H}_\ell^{\mathrm{Sp}_{2g}}(x)$ (resp. $\mathcal{H}_v^{\mathrm{SL}_{2,E}}(y)$) is infinite.

The following irreducibility statement is handy for the proof of Theorem 9.2, because it shortens the argument and simplifies the logical structure of the proof.

Theorem 9.9. **BB** *Let W be a locally closed subscheme of $\mathcal{M}_{F,n}$ over $\mathbb{F}$ which is smooth over $\mathbb{F}$ and stable under all $\mathrm{SL}_2(F \otimes \mathbb{A}_f^{(p)})$-Hecke correspondences. Assume that the $\mathrm{SL}_2(F \otimes \mathbb{A}_f^{(p)})$-Hecke correspondences operate transitively on the set $\Pi_0(W)$ of irreducible components of W, and some (hence all) maximal point of W corresponds to a non-supersingular abelian variety. Then W is irreducible.*

Remark. The argument in [**11**] works in the situation of 9.9. The following observations may be helpful.

(i) The group $\mathrm{SL}_2(F \otimes \mathbb{A}_f^{(p)})$ has no proper subgroup of finite index. This statement can be verified directly without difficulty. It can also be explained in a more general context: The linear algebraic group $\mathrm{Res}_{F/\mathbb{Q}}(\mathrm{SL}_2)$ over $\mathbb{Q}$ is semisimple, connected and simply connected. Therefore every subgroup of finite index in $\mathrm{SL}_2(F \otimes \mathbb{Q}_\ell)$ is equal to $\mathrm{SL}_2(F \otimes \mathbb{Q}_\ell)$, for every prime number ℓ.

(ii) The only part of the argument in [**11**] that needs to be supplemented is the end of (4.1), where the fact that Sp_{2g} is simple over $\mathbb{Q}_\ell$ is used. Let G_ℓ be the image group of the ℓ-adic monodromy ρ_Z attached to Z. By definition, G_ℓ is a closed subgroup of $\mathrm{SL}_2(F \otimes \mathbb{Q}_\ell) = \prod_{v|\ell} \mathrm{SL}_2(F_v)$, where v runs through all places of F above ℓ. In the present situation of a Hilbert modular variety $\mathcal{M}_F$, we need to know the fact that the projection of G_ℓ to the factor $\mathrm{SL}_2(F_v)$ is non-trivial for every place v of F above ℓ and for every $\ell \neq p$.

Theorem 9.10. (Hilbert trick) *Given $x_0 \in \mathcal{A}_{g,1,n}(\mathbb{F})$, then there exist*

(a) *totally real number fields $F_1, \ldots, F_r$ such that $\sum_{i=1}^r [E_i : \mathbb{Q}] = g$,*

(b) *an invertible $\mathcal{O}_E$-module $\mathcal{L}$ with a notion of positivity $\mathcal{L}^+$, i.e., $\mathcal{L}^+$ is a union of connected components of $\mathcal{L}\otimes_{\mathbb{Q}}\mathbb{R}$ such that $\mathcal{L}\otimes\mathbb{R}$ is the disjoint union of $\mathcal{L}^+$ with $-\mathcal{L}^+$,*

(c) *a positive integer a and a positive integer m such that $(m,p)=1$ and $m\equiv 0 \pmod n$,*

(d) *a finite flat morphism $g : \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a} \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$,*

(e) *a finite morphism $f : \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a} \to \mathcal{A}^{\mathrm{ord}}_{g,n}$,*

(f) *a point $y_0 \in \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}(\mathbb{F})$*

such that the following properties are satisfied.

(i) *There exists a projective system $\mathcal{M}^{\mathrm{ord},\sim}_{E,\mathcal{L},\mathcal{L}^+;a}$ of finite étale coverings of $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m;a}$ on which the group $\mathrm{SL}_2(E\otimes\mathbb{A}_f^{(p)})$ operates. This action of $\mathrm{SL}_2(E\otimes\mathbb{A}_f^{(p)})$ induces Hecke correspondences on $\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$*

(ii) *The morphism g is equivariant with respect to Hecke correspondences coming from $\mathrm{SL}_2(E\otimes\mathbb{A}_f^{(p)})$. In other words, there is an $\mathrm{SL}_2(E\otimes\mathbb{A}_f^{(p)})$-equivariant morphism $g^\sim$ from the projective system $\mathcal{M}^{\mathrm{ord},\sim}_{E,\mathcal{L},\mathcal{L}^+;a}$ to the projective system $\left(\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,md}\right)_{d\in\mathbb{N}-p\mathbb{N}}$ which lifts g.*

(iii) *There exists an injective homomorphism $j_E : \mathrm{SL}_2(E\otimes_{\mathbb{Q}}\mathbb{A}_f^{(p)}) \to \mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})$ such that the finite morphism f is Hecke equivariant with respect to j_E.*

(iv) *We have $f(y_0)=x_0$.*

(v) *For every geometric point $z\in\mathcal{M}^{\mathrm{ord}}_{E,m;a}$, the abelian variety underlying the fiber over $g(z)\in\mathcal{M}^{\mathrm{ord}}_{E,m}$ of the universal abelian scheme over $\mathcal{M}^{\mathrm{ord}}_{E,m}$ is isogenous to the abelian variety underlying the fiber over $f(z)\in\mathcal{A}^{\mathrm{ord}}_{g,n}(\mathbb{F})$ of the universal abelian scheme over $\mathcal{A}^{\mathrm{ord}}_{g,n}(\mathbb{F})$.*

Remark. The scheme $\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$ is defined in Step 3 of the proof of Theorem 9.10.

Lemma. *Let A be an ordinary abelian variety over $\mathbb{F}$ which is simple. Then*

(i) *$K := \mathrm{End}^0(A)$ is a totally imaginary quadratic extension of a totally real number field F;*

(ii) *$[F:\mathbb{Q}] = \dim(A)$;*

(iii) *F is fixed by the Rosati involution attached to any polarization of A;*

(iii) *Every place $\wp$ of F above p splits in K.*

Proof. The statements (i)–(iv) are immediate consequences of Tate's theorem for abelian varieties over finite fields; see [**79**]. ∎

Lemma. *Let K be a CM field, let $E := \mathrm{M}_d(K)$, and let $*$ be a positive involution on E which induces the complex conjugation on K. Then there exists a CM field L which contains K and a K-linear ring homomorphism $h : L \to E$ such that $[L:K]=d$ and $h(L)$ is stable under the involution $*$.*

Proof. This is an exercise in algebra. A proof using Hilbert irreducibility can be found on p. 458 of [**9**]. ∎

Proof of Theorem 9.10 (Hilbert trick).
Step 1. Consider the abelian variety A_0 attached to the given point

$$x_0 = [(A_0,\lambda_0,\eta_0)] \in \mathcal{A}^{\mathrm{ord}}_{g,1,n}(\mathbb{F}).$$

By the two lemmas above there exist totally real number fields $F_1, \dots, F_r$ and an embedding $\iota_0 : E := F_1 \times \cdots \times F_r \hookrightarrow \mathrm{End}^0(A_0)$ such that E is fixed under the Rosati involution on $\mathrm{End}^0(A_0)$ attached to the principal polarization λ_0, and $[E : \mathbb{Q}] = g = \dim(A_0)$.

The intersection of E with $\mathrm{End}(A_0)$ is an order $\mathcal{O}_1$ of E, so we can regard A_0 as an abelian variety with action by $\mathcal{O}_1$. We claim that there exists an $\mathcal{O}_E$-linear abelian variety B and an $\mathcal{O}_1$-linear isogeny $\alpha : B \to A_0$. This claim follows from a standard "saturation construction" as follows. Let d be the order of the finite abelian group $\mathcal{O}_E/\mathcal{O}_1$. Since A_0 is ordinary, one sees by Tate's theorem (the case when K is a finite field in Theorem 3.16) that $(d, p) = 1$. For every prime divisor $\ell \neq p$ of d, consider the ℓ-adic Tate module $\mathrm{T}_\ell(A_0)$ as a lattice inside the free rank two E-module $\mathrm{V}_\ell(A_0)$. Then the lattice Λ_ℓ generated by $\mathcal{O}_E \cdot \mathrm{T}_\ell(A_0)$ is stable under the action of $\mathcal{O}_E$ by construction. The finite set of lattices $\{\Lambda_\ell : \ell | d\}$ defines an $\mathcal{O}_E$-linear abelian variety B and an $\mathcal{O}_1$-linear isogeny $\beta_0 : A_0 \to B$ which is killed by a power d^i of d. Let $\alpha : B \to A_0$ be the isogeny such that $\alpha \circ \beta_0 = [d^i]_{A_0}$. The claim is proved.

Step 2. The construction in Step 1 gives us a triple (B, α, ι_{x_0}), where B is an abelian variety B over $\mathbb{F}$, $\alpha : B \to A_x$ is an isogeny over $\mathbb{F}$, and $\iota_B : \mathcal{O}_E \to \mathrm{End}(B)$ is an injective ring homomorphism such that $\alpha^{-1} \circ \iota_x(u) \circ \alpha = \iota_B(u)$ for every $u \in \mathcal{O}_E$. Let $\mathcal{L}_B := \mathrm{Hom}^{\mathrm{sym}}_{\mathcal{O}_E}(B, B^t)$ be set of all $\mathcal{O}_E$-linear symmetric homomorphisms from B to the dual B^t of B. The set $\mathcal{L}_B$ has a natural structure as an $\mathcal{O}_E$-module. By Tate's theorem (the case when K is a finite field in Theorem 3.16, see 10.17) one sees that $\mathcal{L}_B$ is an invertible $\mathcal{O}_E$-module, and the natural map

$$\lambda_B : B \otimes_{\mathcal{O}_E} \mathcal{L}_B \to B^t$$

is an $\mathcal{O}_E$-linear isomorphism. The subset of elements in $\mathcal{L}$ which are polarizations defines a notion of positivity $\mathcal{L}^+$ on $\mathcal{L}$ such that $\mathcal{L}_B \cap \mathcal{L}_B^+$ is the subset of $\mathcal{O}_E$-linear polarizations on (B, ι_B).

Step 3. Recall that the Hilbert modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}_+,n}$ classifies (the isomorphism class of) all quadruples $(A \to S, \iota_A, \lambda_A, \eta_A)$, where $(A \to S, \iota_A)$ is an $\mathcal{O}_E$-linear abelian schemes, $\lambda_A : \mathcal{L} \to \mathrm{Hom}^{\mathrm{sym}}_{\mathcal{O}_E}(A, A^t)$ is an injective $\mathcal{O}_E$-linear map such that the resulting morphism $\mathcal{L} \otimes A \xrightarrow{\sim} A^t$ is an isomorphism of abelian schemes and every element of $\mathcal{L} \cap \mathcal{L}^+$ gives rise to an $\mathcal{O}_E$-linear polarization, and η_A is an $\mathcal{O}_E$-linear level structure on (A, ι_A). In the preceding paragraph, if we choose an $\mathcal{O}_E$-linear level-n structure η_B on (B, ι_B), then $y_1 := [(B, \iota_B, \lambda_B, \eta_B)]$ is an $\mathbb{F}$-point of the Hilbert modular variety $\mathcal{M}_{E,\mathcal{L}_B,\mathcal{L}_B^+,n}$. The element $\alpha^*(\lambda_0)$ is an $\mathcal{O}_E$-linear polarization on B, hence it is equal to $\lambda_B(\mu_0)$ for a unique element $\mu_0 \in \mathcal{L} \cap \mathcal{L}^+$.

Choose a positive integer m_1 with $\gcd(m_1, p) = 1$ and $a \in \mathbb{N}$ such that $\mathrm{Ker}(\alpha)$ is killed by $m_1 p^a$. Let $m = m_1 n$. Let $(A, \iota_A, \lambda_A, \eta_A) \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$ be the universal polarized $\mathcal{O}_E$-linear abelian scheme over the ordinary locus $\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$ of $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}$. Define a scheme $\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$ over $\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$ by

$$\underline{\mathrm{Isom}}^{\mathcal{O}_E}_{\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}} \Big((B, \iota_B, \lambda_B)[p^a] \times_{\mathrm{Spec}(\mathbb{F})} \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}, (A, \iota_A, \lambda_A)[p^a] \Big) .$$

In other words $\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$ is the moduli space of $\mathcal{O}_E$-linear ordinary abelian varieties with level-mp^a structure, where we have used the $\mathcal{O}_E$-linear polarized truncated p-divisible group $(B, \iota_B, \lambda_B)[p^m]$ as the "model" for the mp^a-torsion subgroup

scheme of the universal abelian scheme over $\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$. Let

$$g : \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a} \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$$

be the structural morphism of $\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$, the source of g being an fppf sheaf of sets on the target of g. Notice that the structural morphism $g : \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a} \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$ has a natural structure as a torsor over the constant finite flat group scheme

$$\underline{\mathrm{Aut}}\,((B,\iota_B,\lambda_B)[p^a]) \times_{\mathrm{Spec}(\mathbb{F})} \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}\,.$$

We have constructed the finite flat morphism g as promised in Theorem 9.10 (d). We record some properties of this morphism.

The group $\underline{\mathrm{Aut}}\,(B,\iota_B,\lambda_B)[p^a])$ sits in the middle of a short exact sequence

$$0 \to \underline{\mathrm{Hom}}_{\mathcal{O}_E}\,(B[p^a]_{\text{ét}}, B[p^a]_{\mathrm{mult}}) \to \underline{\mathrm{Aut}}\,((B,\iota_B,\lambda_B)[p^a]) \to \underline{\mathrm{Aut}}_{\mathcal{O}_E}\,(B[p^a]_{\text{ét}}) \to 0\,.$$

The morphism $g : \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a} \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$ factors as

$$\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a} \xrightarrow{g_1} \mathcal{M}^{\mathrm{ord,ét}}_{E,\mathcal{L},\mathcal{L}^+,m;a} \xrightarrow{g_2} \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}\,,$$

where g_1 is defined as the push-forward by the surjection

$$\underline{\mathrm{Aut}}\,((B,\iota_B,\lambda_B)[p^a]) \twoheadrightarrow \underline{\mathrm{Aut}}_{\mathcal{O}_E}\,(B[p^a]_{\text{ét}})$$

of the $\underline{\mathrm{Aut}}\,((B,\iota_B,\lambda_B)[p^a])$-torsor $\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$. Note that the morphism g_1 is finite flat and purely inseparable, and $\mathcal{M}^{\mathrm{ord,ét}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$ is integral. Moreover $\mathcal{M}^{\mathrm{ord,ét}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$ and $\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$ are irreducible by [**74**], [**23**], [**71**] and [**22**].

Step 4. Let $\pi_{n,m} : \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m} \to \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$ be the natural projection. Denote by

$$A[mp^a] \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$$

the kernel of $[mp^a]$ on $A \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$, and let $g^*A[mp^a] \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$ be the pullback of $A[mp^a] \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$ by g. By construction the $\mathcal{O}_E$-linear finite flat group scheme $g^*A[mp^a] \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$ is constant via a tautological trivialization

$$\psi : \underline{\mathrm{Aut}}\,(B,\iota_B,\lambda_B)[p^a]) \times_{\mathrm{Spec}(\mathbb{F})} \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m} \xrightarrow{\sim} \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$$

Choose a point $y_0 \in \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}(\mathbb{F})$ such that $(\pi_{n,m} \circ g)(y_0) = y_1$. The fiber over y_0 of $g^*A[mp^a] \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$ is naturally identified with $B[mp^a]$. Let $K_0 := \mathrm{Ker}(\alpha : B \to A_0)$, and let

$$K := \psi\left(K_0 \times_{\mathrm{Spec}(\mathbb{F})} \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a} \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}\right)\,,$$

the subgroup scheme of $g^*A[mp^a]$ which corresponds to the constant group K_0 under the trivialization ψ. The element $\mu_0 \in \mathcal{L} \cap \mathcal{L}^+$ defines a polarization on the abelian scheme $g^*A \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$, the pullback by g of the universal polarized $\mathcal{O}_E$-linear abelian scheme over $A \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$. The group K is a maximal totally isotropic subgroup scheme of $g^*\mathrm{Ker}(\lambda_A(\mu_0)) \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$, because $g^*\mathrm{Ker}(\lambda_A(\mu_0))$ is constant and K_0 is a maximal totally isotropic subgroup scheme of $\mathrm{Ker}(\lambda_B(\mu_0))$.

Consider the quotient abelian scheme

$$(A' \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}) := (g^*A \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a})/K.$$

Recall that we have defined an element $\mu_0 \in \mathcal{L} \cap \mathcal{L}^+$ in Step 3. The polarization $g^*(\lambda_A(\mu_0))$ on the abelian scheme $g^*A \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$ descends to the quotient abelian scheme $A' \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$, giving it a principal polarization $\lambda_{A'}$. Moreover the n-torsion subgroup scheme $A'[n] \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$ is constant, as can be checked easily. Choose a level-n structure $\eta_{A'}$ for A'. The triple $(A', \lambda_{A'}, \eta_{A'})$ over $\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$ defines a morphism $f : \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a} \to \mathcal{A}^{\mathrm{ord}}_{g,1,n}$ by the modular definition of $\mathcal{A}^{\mathrm{ord}}_{g,1,n}$, since every fiber of $A' \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$ is ordinary by construction. We have constructed the morphism f as required in 9.10 (e), and also the point y_0 as required in 9.10 (f).

Step 5. So far we have constructed the morphisms g and f as required in Theorem 9.10. To construct the homomorphism j_E as required in (iii), one uses the first homology group $V := \mathrm{H}_1(B, \mathbb{A}_f^{(p)})$, and the symplectic pairing $\langle \cdot, \cdot \rangle$ induced by the polarization $\alpha^*(\lambda_0) = \lambda_B(\mu_0)$ constructed in Step 3. Notice that V has a natural structure as a free $E \otimes_{\mathbb{Q}} \mathbb{A}_f^{(p)}$-module of rank two. Also, V is a free $\mathbb{A}_f^{(p)}$-module of rank $2g$. So we get an embedding $j_E : \mathrm{SL}_{E \otimes \mathbb{A}_f^{(p)}}(V) \hookrightarrow \mathrm{Sp}_{\mathbb{A}_f^{(p)}}(V, \langle \cdot, \cdot \rangle)$. We have finished the construction of j_E.

We define $\mathcal{M}^{\mathrm{ord},\sim}_{E,\mathcal{L},\mathcal{L}^+,a}$ to be the projective system $\varprojlim_{md} \mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,md;a}$, where d runs through all positive integers which are prime to p. This finishes the last construction needed for Theorem 9.10.

By construction we have $f(y_0) = x_0$, which is statement (iv). The statement (v) is clear by construction. The statements (i)–(iii) can be verified without difficulty from the construction. □

Proof of Theorem 9.2. (Density of ordinary Hecke orbits in $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$)
Reduction step.

From the product decomposition

$$\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n} = \mathcal{M}_{F_1,\mathcal{L}_1,\mathcal{L}_1^+,n} \times_{\mathrm{Spec}(\mathbb{F})} \cdots \times_{\mathrm{Spec}(\mathbb{F})} \mathcal{M}_{F_r,\mathcal{L}_r,\mathcal{L}_r^+,n}$$

of the Hilbert modular variety $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,n}$, we see that it suffices to prove Theorem 9.2 when $r = 1$, i.e., $E = F_1 =: F$ is a totally real number field. Assume this is the case from now on.

The rest of the proof is divided into four steps.

Step 1. (Serre-Tate coordinates for Hilbert modular varieties)
Claim. The Serre-Tate local coordinates at a closed ordinary point $z \in \mathcal{M}^{\mathrm{ord}}_{F,\mathcal{L},\mathcal{L}^+,n}$ of a Hilbert modular variety $\mathcal{M}_{F,\mathcal{L},\mathcal{L}^+,n}$ admits a canonical decomposition

$$\mathcal{M}^{/z}_{F,\mathcal{L},\mathcal{L}^+,n} \cong \prod_{\wp \in \Sigma_{F,p}} \mathcal{M}^z_\wp\,, \qquad \mathcal{M}^z_\wp = \underline{\mathrm{Hom}}_{\mathcal{O}_{F,\wp}} \left(\mathrm{T}_p(A_z[\wp^\infty]_{\text{ét}}),\ e_\wp \cdot A_z^{/0} \right)\,,$$

where

- the indexing set $\Sigma_{F,p}$ is the finite set consisting of all prime ideals of $\mathcal{O}_F$ above p,
- the $(\mathcal{O}_F \otimes \mathbb{Z}_p)$-linear formal torus $A_z^{/0}$ is the formal completion of the ordinary abelian variety A_z,
- $e_\wp$ is the irreducible idempotent in $\mathcal{O}_F \otimes_{\mathbb{Z}} \mathbb{Z}_p$ so that $e_\wp \cdot (\mathcal{O}_F \otimes_{\mathbb{Z}} \mathbb{Z}_p)$ is equal to the factor $\mathcal{O}_{F_\wp}$ of $\mathcal{O}_F \otimes_{\mathbb{Z}} \mathbb{Z}_p$.

Notice that $e_\wp A_z^{/0}$ is the formal torus attached to the multiplicative p-divisible group $A_z[\wp^\infty]_{\mathrm{mult}}$ over $\mathbb{F}$.

PROOF OF CLAIM. The decomposition $\mathcal{O}_F \otimes_{\mathbb{Z}} \mathbb{Z}_p = \prod_{\wp\in\Sigma_{F,p}} \mathcal{O}_{F_\wp}$ induces a decomposition of the formal scheme $\mathcal{M}^{/z}_{F,\mathcal{L},\mathcal{L}^+,n}$ into a product

$$\mathcal{M}^{/z}_{F,\mathcal{L},\mathcal{L}^+,n} = \prod_{\wp\in\Sigma_{F,p}} \mathcal{M}^z_\wp$$

for every closed point z of $\mathcal{M}_{F,\mathcal{L},\mathcal{L}^+,n}$: Let $(A/R,\iota)$ be an $\mathcal{O}_F$-linear abelian scheme over an Artinian local ring R. Then we have a decomposition

$$A[p^\infty] = \prod_{\wp\in\Sigma_{F,p}} A[\wp^\infty]$$

of the p-divisible group attached to A, where each $A[\wp^\infty]$ is a deformation of $A\times_{\mathrm{Spec}(R)}\mathrm{Spec}(R/\mathfrak{m})$ over $\mathrm{Spec}(R)$.

If z corresponds to an ordinary abelian variety A_z, then $\mathcal{M}^z_\wp$ is canonically isomorphic to the $\mathcal{O}_{F,\wp}$-linear formal torus $\underline{\mathrm{Hom}}_{\mathcal{O}_{F,\wp}}(A_z[\wp^\infty]_{\text{ét}}, e_\wp\cdot A_z^{/0})$, which is the factor "cut out" in the $(\mathcal{O}_F\otimes_{\mathbb{Z}}\mathbb{Z}_p)$-linear formal torus

$$\mathcal{M}^{/z}_{F,\mathcal{L},\mathcal{L}^+,n} = \underline{\mathrm{Hom}}_{\mathcal{O}_F\otimes\mathbb{Z}_p}\left(\mathrm{T}_p(A_z[p^\infty],\ A_z^{/0}\right)$$

by the idempotent $e_\wp$ in $\mathcal{O}_F\otimes\mathbb{Z}_p$. Each factor $\mathcal{M}^z_\wp$ in the above decomposition is a formal torus of dimension $[F_\wp:\mathbb{Q}_p]$, with a natural action by $\mathcal{O}_{F,\wp}^\times$; it is non-canonically isomorphic to the $\mathcal{O}_\wp$-linear formal torus ${A_z}^{/0}$. □

Step 2. (Linearization)

CLAIM. For every closed point $z\in Z_F^{\mathrm{ord}}(\mathbb{F})$ in the ordinary locus of Z_F, there exists a non-empty subset $S_z\subset\Sigma_{F,p}$ such that $Z_F^{/z} = \prod_{\wp\in S_z}\mathcal{M}^z_\wp$, where $\mathcal{M}^z_\wp$ is the factor of the Serre-Tate formal torus $\mathcal{M}^{/z}_{F,\mathcal{L},\mathcal{L}^+,n}$ corresponding to $\wp$.

PROOF OF CLAIM. The $\mathcal{O}_F$-linear abelian variety A_z is an ordinary abelian variety defined over $\mathbb{F}$. Therefore $\mathrm{End}^0_{\mathcal{O}_F}(A_z)$ is a totally imaginary quadratic extension field K of F which is split over every prime ideal $\wp$ of $\mathcal{O}_F$ above p, by Tate's theorem (the case when K is a finite field in Theorem 3.16). By the local stabilizer principle, $Z_F^{/z}$ is stable under the norm-one subgroup U of $(\mathcal{O}_K\otimes_{\mathbb{Z}}\mathbb{Z}_p)^\times$. Since every prime $\wp$ of $\mathcal{O}_F$ above p splits in $\mathcal{O}_K$, U is isomorphic to $\prod_{\wp\in\Sigma_{F,p}}\mathcal{O}_{F,\wp}^\times$ through its action on the $(\mathcal{O}_F\otimes\mathbb{Z}_p)$-linear formal torus $A_z^{/0}$. The factor $\mathcal{O}_{F,\wp}^\times$ of U operates on the $\mathcal{O}_{F,\wp}$-linear formal torus $\mathcal{M}^z_\wp$ through the character $t\mapsto t^2$, i.e., a typical element $t\in U=\prod_{\wp\in\Sigma_{F,p}}\mathcal{O}_{F,\wp}^\times$ operates on the $(\mathcal{O}_F\otimes\mathbb{Z}_p)$-linear formal torus $\mathcal{M}^{/z}_{F,\mathcal{L},\mathcal{L}^+,n}$ through the element $t^2\in U=(\mathcal{O}_F\otimes\mathbb{Z}_p)^\times$. The last assertion can be seen through the formula

$$\mathcal{M}^{/z}_{F,\mathcal{L},\mathcal{L}^+,n} = \underline{\mathrm{Hom}}_{\mathcal{O}_F\otimes\mathbb{Z}_p}\left(\mathrm{T}_p(A_z[p^\infty]_{\text{ét}}),\ A_z^{/0}\right),$$

because any element t of $U\xrightarrow{\sim}\mathcal{O}_F\otimes_{\mathbb{Z}}\mathbb{Z}_p$ operates via t (resp. t^{-1}) on the $(\mathcal{O}_F\otimes\mathbb{Z}_p)$-linear formal torus $e_\wp A_z^{/0}$ (resp. the $(\mathcal{O}_F\otimes\mathbb{Z}_p)$-linear p-divisible group $A_z[p^\infty]_{\text{ét}}$).

The local rigidity theorem 2.26 implies that $Z_F^{/z}$ is a formal subtorus of the Serre-Tate formal torus $\mathcal{M}_F^{/z}$. For every $\wp\in\Sigma_{F,p}$, let $X_{\wp,*}$ be the cocharacter group of the $\mathcal{O}_{F_\wp}$-linear formal torus $\mathcal{M}^z_\wp$, so that $\prod_{\wp\in\Sigma_{F,p}}X_{\wp,*}$ is the cocharacter group of the Serre-Tate formal torus $\mathcal{M}^{/z}_{F,\mathcal{L},\mathcal{L}^+,n}$. Let Y_* be the cocharacter group

of the formal torus $Z_F^{/z}$. We know that Y_* is a co-torsion free $\mathbb{Z}_p$-submodule of the rank-one free $\left(\prod_{\wp\in\Sigma_{F,p}} \mathcal{O}_{F,\wp}\right)$-module $\prod_{\wp\in\Sigma_{F,p}} X_{\wp,*}$, and Y_* is stable under multiplication by elements of the subgroup $\prod_{\wp\in\Sigma_{F,p}}(\mathcal{O}_{F,\wp}^\times)^2$ of $\prod_{\wp\in\Sigma_{F,p}} \mathcal{O}_{F,\wp}^\times$. It is easy to see that the additive subgroup generated by $\prod_{\wp\Sigma_{F,p}}(\mathcal{O}_{F,\wp}^\times)^2$ is equal to $\prod_{\wp\in\Sigma_{F,p}} \mathcal{O}_{F,\wp}$, i.e., Y_* is a $\left(\prod_{\wp\in\Sigma_{F,p}} \mathcal{O}_{F,\wp}\right)$-submodule of $\prod_\wp X_{\wp,*}$. Hence there exists a subset $S_z \subseteq \Sigma_{F,p}$ such that $Y_* = \prod_{\wp\in S_z} X_{\wp,*}$. Since the prime-to-$p$ Hecke orbit $\mathcal{H}^{(p)}_{\mathrm{SL}_{2,F}}(x)$ is infinite by 9.8, we have $0 < \dim(Z_F) = \sum_{\wp\in S_z}[F_\wp : \mathbb{Q}_p]$, hence $S_z \neq \emptyset$ for every ordinary point $z \in Z_F(x)(\mathbb{F})$. We have proved the Claim in Step 2. □

Step 3. (Globalization)
Claim. The finite set S_z is independent of the point z, i.e., there exists a subset $S \subset \Sigma_{F,p}$ such that $S_z = \Sigma_{F,p}$ for all $z \in Z_F^{\mathrm{ord}}(\mathbb{F})$.
Proof of Claim. Consider the diagonal embedding $\Delta_Z : Z_F \to Z_F \times_{\mathrm{Spec}(\mathbb{F})} Z_F$, the diagonal embedding $\Delta_{\mathcal{M}} : \mathcal{M}_{F,n} \to \mathcal{M}_{F,n} \times_{\mathrm{Spec}(\mathbb{F})} \mathcal{M}_{F,n}$, and the map $\Delta_{Z,\mathcal{M}}$ from Δ_Z to $\Delta_{\mathcal{M}}$ induced by the inclusion $Z_F \hookrightarrow \Delta_{\mathcal{M}}$. Let $\mathcal{P}_Z$ be the formal completion of $Z_F \times_{\mathrm{Spec}(\mathbb{F})} Z_F$ along $\Delta_Z(Z_F)$, and let $\mathcal{P}_{\mathcal{M}}$ be the formal completion of $\mathcal{M}_{F,n} \times_{\mathrm{Spec}(\mathbb{F})} \mathcal{M}_{F,n}$ along $\Delta_{\mathcal{M}}(\mathcal{M}_{F,n})$. The map $\Delta_{Z,\mathcal{M}}$ induces a closed embedding $i_{Z,\mathcal{M}} : \mathcal{P}_Z \hookrightarrow \mathcal{P}_{\mathcal{M}}$. We regard $\mathcal{P}_Z$ (resp. $\mathcal{P}_{\mathcal{M}}$) as a formal scheme over Z_F (resp. $\mathcal{M}_{F,n}$) via the first projection.

The decomposition $\mathcal{O}_F \otimes_{\mathbb{Z}} \mathbb{Z}_p = \prod_{\wp\in\Sigma_{F,p}} \mathcal{O}_{F,\wp}$ induces a fiber product decomposition

$$\mathcal{P}_{\mathcal{M}} = \prod_{\wp\in\Sigma_{F,p}} (\mathcal{P}_\wp \to \mathcal{M}_{F,n})$$

over the base scheme $\mathcal{M}_{F,n}$, where $\mathcal{P}_\wp \to \mathcal{M}_{F,n}$ is a smooth formal scheme of relative dimension $[F_\wp : \mathbb{Q}_p]$ with a natural section $\delta_\wp$, for every $\wp \in \Sigma_{F,p}$, and the formal completion of the fiber of $(\mathcal{M}_\wp, \delta_\wp)$ over any closed point z of the base scheme $\mathcal{M}_{F,n}$ is canonically isomorphic to the formal torus $\mathcal{M}_\wp^z$. In fact one can show that $\mathcal{M}_\wp \to \mathcal{M}_{F,n}$ has a natural structure as a formal torus of relative dimension $[F_\wp : \mathbb{Q}_p]$, with $\delta_\wp$ as the zero section; we will not need this fact here. Notice that $\mathcal{P}_Z \to Z_F$ is a closed formal subscheme of $\mathcal{P}_{\mathcal{M}} \times_{\mathcal{M}_{F,n}} Z_F \to Z_F$. The above consideration globalizes the "pointwise" construction of formal completions at closed points.

By Proposition 9.9, Z_F is irreducible. We conclude from the irreducibility of Z_F that there is a non-empty subset $S \subset \Sigma_{F,p}$ such that the restriction of $\mathcal{P}_Z \to Z_F$ to the ordinary locus Z_F^{ord} is equal to the fiber product over Z_F^{ord} of formal schemes $\mathcal{P}_\wp \times_{\mathcal{M}_{F,n}} Z_F^{\mathrm{ord}} \to Z_F^{\mathrm{ord}}$ over Z_F^{ord}, where $\wp$ runs through the subset $S \subseteq \Sigma_{F,p}$. We have proved the Claim in Step 3. □

Remark.

(i) Without using Proposition 9.9, the above argument shows that for each irreducible component Z_1 of Z_F^{ord}, there exists a subset $S \subset \Sigma_{F,p}$ such that $S_z = S$ for every closed point $z \in Z_1(\mathbb{F})$.

(ii) There is an alternative proof of the claim that S_z is independent of z: By Step 2, Z_F^{ord} is smooth over $\mathbb{F}$. Consider the relative cotangent sheaf $\Omega^1_{Z_F^{\mathrm{ord}}/\mathbb{F}}$, which is a locally free $\mathcal{O}_{Z_F^{\mathrm{ord}}}$-module. We recall that $\Omega^1_{\mathcal{M}^{\mathrm{ord}}_{F,\mathcal{L},\mathcal{L}^+,n}/\mathbb{F}}$

has a natural structure as a $\mathcal{O}_F \otimes_{\mathbb{Z}} \mathbb{F}_p$-module, from the Serre-Tate coordinates explained in Step 1. By Step 2, we have

$$\Omega^1_{Z_F^{\mathrm{ord}}/\mathbb{F}} \otimes_{\mathcal{O}_{Z_F,z}} \mathcal{O}^{\wedge}_{Z_F,z} = \sum_{\wp \in S_z} e_\wp \cdot \Omega^1_{\mathcal{M}^{\mathrm{ord}}_{F,\mathcal{L},\mathcal{L}^+,n}/\mathbb{F}} \otimes_{\mathcal{O}_{\mathcal{M}^{\mathrm{ord}}_{F,\mathcal{L},\mathcal{L}^+,n}}} \mathcal{O}^{\wedge}_{Z_F,z}$$

for every $z \in Z_F^{\mathrm{ord}}(\mathbb{F})$, where $\mathcal{O}^{\wedge}_{Z_F,z}$ is the formal completion of the local ring of Z_F at z. Therefore for each irreducible component Z_1 of Z_F^{ord} there exists a subset $S \subset \Sigma_{F,p}$ such that

$$\Omega^1_{Z_1/\mathbb{F}} = \sum_{\wp \in S} e_\wp \cdot \Omega^1_{\mathcal{M}^{\mathrm{ord}}_{F,\mathcal{L},\mathcal{L}^+,n}/\mathbb{F}} \otimes_{\mathcal{O}_{\mathcal{M}^{\mathrm{ord}}_{F,\mathcal{L},\mathcal{L}^+,n}}} \mathcal{O}_{Z_1/\mathbb{F}}\,.$$

Hence $S_z = S$ for every $z \in Z_1(\mathbb{F})$. This argument was used in [**9**]; see Proposition 5 on p. 473 in *loc. cit.*

(iii) One advantage of the globalization argument in Step 3 is that it makes the final Step 4 of the proof of Theorem 9.2 easier, as compared with the two-page proof of Proposition 7 on p. 474 of [**9**].

Step 4. We have $S = \Sigma_{F,p}$. Therefore $Z_F = \mathcal{M}_{F,n}$.

Proof of Step 4.

Notation as in Step 3 above. For every closed point s of Z_F, the formal completion $Z_F^{/s}$ contains the product $\prod_{\wp \in S} \mathcal{M}^s_\wp$. By Theorem 9.7, Z_F contains a supersingular point s_0. Consider the formal completion $Z^{\wedge} := Z_F^{/s_0}$, which contains $W^{\wedge} := \prod_{\wp \in S} \mathcal{M}^{s_0}_\wp$, and the generic point $\eta_{W^{\wedge}}$ of $\mathrm{Spec}\left(\mathrm{H}^0(W^{\wedge}, \mathcal{O}_{W^{\wedge}})\right)$ is a maximal point of $\mathrm{Spec}\left(\mathrm{H}^0(Z^{\wedge}, \mathcal{O}_{Z^{\wedge}})\right)$. The restriction of the universal abelian scheme to $\eta_{W^{\wedge}}$ is an ordinary abelian variety. Hence $S = \Sigma_{F,p}$, otherwise $A_{\eta_{W^{\wedge}}}$ has slope $1/2$ with multiplicity at least $2\sum_{\wp \notin S}[F_\wp : \mathbb{Q}_p]$. Theorem 9.2 is proved. □

Remark. The proof of Theorem 9.2 can be finished without using Proposition 9.9 as follows. We saw in the Remark after Step 3 that S_z depends only on the irreducible component of Z_F^{ord} which contains z. The argument in Step 4 shows that at least the subset $S \subset \Sigma_{F,p}$ attached to one irreducible component Z_1 of Z_F^{ord} is equal to $\Sigma_{F,p}$. So $\dim(Z_1) = \dim(\mathcal{M}_{F,\mathcal{L},\mathcal{L}^+,n}) = [F : \mathbb{Q}]$. Since $\mathcal{M}_{F,\mathcal{L},\mathcal{L}^+,n}$ is irreducible, we conclude that $Z_F = \mathcal{M}_{F,\mathcal{L},\mathcal{L}^+,n}$.

Proof of Theorem 9.1. (Density of ordinary Hecke orbits in $\mathcal{A}_{g,1,n}$.)

Reduction step. By Theorem 9.4, the weaker statement 9.1 (i) implies 9.1 (ii). So it suffices to prove 9.1 (i).

Remark. Our argument can be used to prove (ii) directly without appealing to Theorem 9.4. But some statements, including the local stabilizer principal for Hilbert modular varieties, need to be modified.

Step 1. (Hilbert trick)

Given $x \in \mathcal{A}_{g,n}(\mathbb{F})$, Apply Theorem 9.10 to produce a finite flat morphism

$$g : \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a} \to \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}\,,$$

where E is a product of totally real number fields, a finite morphism

$$f : \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a} \to \mathcal{A}_{g,n}\,,$$

and a point $y_0 \in \mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}(\mathbb{F})$ such that the following properties are satisfied.

(i) There exists a projective system $\mathcal{M}^{\mathrm{ord},\sim}_{E,\mathcal{L},\mathcal{L}^+;a}$ of finite étale coverings of $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m;a}$ on which the group $\mathrm{SL}_2(E\otimes\mathbb{A}_f^{(p)})$ operates. This action of $\mathrm{SL}_2(E\otimes\mathbb{A}_f^{(p)})$ induces Hecke correspondences on $\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$

(ii) The morphism g is equivariant with respect to Hecke correspondences coming from $\mathrm{SL}_2(E\otimes\mathbb{A}_f^{(p)})$. In other words, there is a $\mathrm{SL}_2(E\otimes\mathbb{A}_f^{(p)})$-equivariant morphism $g^\sim$ from the projective system $\mathcal{M}^{\mathrm{ord},\sim}_{E,\mathcal{L},\mathcal{L}^+,a}$ to the projective system $\left(\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,mn}\right)_{n\in\mathbb{N}-p\mathbb{N}}$ which lifts g.

(iii) The finite morphism f is Hecke equivariant with respect to an injective homomorphism

$$j_E : \mathrm{SL}_2(E\otimes_{\mathbb{Q}}\mathbb{A}_f^{(p)}) \to \mathrm{Sp}_{2g}(\mathbb{A}_f^{(p)})\,.$$

(iv) For every geometric point $z\in\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$, the abelian variety underlying the fiber over $g(z)\in\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$ of the universal abelian scheme over $\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$ is isogenous to the abelian variety underlying the fiber over $f(z)\in\mathcal{A}^{\mathrm{ord}}_{g,n}(\mathbb{F})$ of the universal abelian scheme over $\mathcal{A}^{\mathrm{ord}}_{g,n}(\mathbb{F})$.

(v) We have $f(y_0)=x_0$.

Let $y:=g(y_0)\in\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$.

Step 2. Let Z_{y_0} be the Zariski closure of the $\mathrm{SL}_2(E\otimes\mathbb{A}_f^{(p)})$-Hecke orbit of y_0 on $\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m;a}$, and let Z_y be the Zariski closure of the $\mathrm{SL}_2(E\otimes\mathbb{A}_f^{(p)})$-Hecke orbit on $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}$. By Theorem 9.2 we know that $Z_y=\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}$. Since g is finite flat, we conclude that $g(Z_{y_0})=Z_y\cap\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}=\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$. We know that $f(Z_{y_0})\subset Z_x$ because f is Hecke-equivariant.

Step 3. Let E_1 be an ordinary elliptic curve over $\mathbb{F}$. Let y_1 be an $\mathbb{F}$-point of $\mathcal{M}_{E,\mathcal{L},\mathcal{L}^+,m}$ such that A_{y_1} is isogenous to $E_1\otimes_{\mathbb{Z}}\mathcal{O}_E$ and $\mathcal{L}_{y_1}$ contains an $\mathcal{O}_E$-submodule of finite index in $\lambda_{E_1}\otimes\mathcal{O}_E$, where λ_{E_1} denotes the canonical principal polarization on E_1. In the above the tensor product $E_1\otimes_{\mathbb{Z}}\mathcal{O}_E$ is taken in the category of fppf sheaves over $\mathbb{F}$; the tensor product is represented by an abelian variety isomorphic to the product of g copies of E_1, with an action by $\mathcal{O}_E$. It is not difficult to check that such a point y_1 exists.

Let z_1 be a point of Z_{y_0} such that $g(z_1)=y_1$. Such a point y_1 exists because $g(Z_{y_0})=\mathcal{M}^{\mathrm{ord}}_{E,\mathcal{L},\mathcal{L}^+,m}$ The point $x_1=f(z_1)$ is contained in the Zariski closure $Z(x)$ of the prime-to-p Hecke orbit of x on $\mathcal{A}_{g,n}$. Moreover A_{x_1} is isogenous to the product of g copies of E_1 by property (iv) in Step 1. So $\mathrm{End}^0(A_{x_1})\cong\mathrm{M}_g(K)$, where $K=\mathrm{End}^0(E)$ is an imaginary quadratic extension field of $\mathbb{Q}$ which is split above p. The local stabilizer principle says that $Z(x)^{/x_1}$ is stable under the natural action of an open subgroup of $\mathrm{SU}(\mathrm{End}^0(A_{x_1}),\lambda_{x_1})(\mathbb{Q}_p)\cong\mathrm{GL}_g(\mathbb{Q}_p)$.

Step 4. We know that $Z(x)$ is smooth at the ordinary point x over k, so $Z(x)^{/x}$ is reduced and irreducible. By the local stabilizer principle 9.5, $Z(x)^{/x}$ is stable under the natural action of the open subgroup H_x of $\mathrm{SU}(\mathrm{End}^0(A_{x_1}),*_{x_1})$ consisting of all elements $\gamma\in\mathrm{SU}(\mathrm{End}^0(A_{x_1}),*_{x_1})(\mathbb{Q}_p)$ such that $\gamma(A_{x_1}[p^\infty])=A_{x_1}[p^\infty]$. By Theorem 2.26, $Z(x)^{/x_1}$ is a formal subtorus of the formal torus $\mathcal{A}'^{x_1}_{g,n}$, which is stable under the action of an open subgroup of $\mathrm{SU}(\mathrm{End}^0(A_{x_1},\lambda_{x_1}))(\mathbb{Q}_p)\cong\mathrm{GL}_g(\mathbb{Q}_p)$.

Let X_* be the cocharacter group of the Serre-Tate formal torus $\mathcal{A}_{g,n}^{/x_1}$, and let Y_* be the cocharacter group of the formal subtorus $Z(x)^{/x_1}$. Both X_* and X_*/Y_* are free $\mathbb{Z}_p$-modules. It is easy to see that the restriction to $\mathrm{SL}_g(\mathbb{Q}_p)$ of the linear action of $\mathrm{SU}(\mathrm{End}^0(A_{x_1}), *_{x_1})(\mathbb{Q}_p) \cong \mathrm{GL}_g(\mathbb{Q}_p)$ on $X_* \otimes_{\mathbb{Z}} \mathbb{Q}_p$ is isomorphic to the second symmetric product of the standard representation of $\mathrm{SL}_g(\mathbb{Q}_p)$. It is well-known that the latter is an absolutely irreducible representation of $\mathrm{SL}_g(\mathbb{Q}_p)$. Since the prime-to-$p$ Hecke orbit of x is infinite, $Y_* \neq (0)$, hence $Y_* = X_*$. In other words $Z(x)^{/x_1} = \mathcal{A}_{g,n}^{/x_1}$. Hence $Z(x) = \mathcal{A}_{g,n}$ because $\mathcal{A}_{g,n}$ is irreducible. ∎

Remark 9.11. We mentioned at the beginning of this section that there is an alternative argument for Step 4 of the proof of Theorem 9.2, which uses [**19**] instead of Theorem 9.7, and therefore is independent of [**67**]. We sketch the idea here; see §8 of [**7**] for more details.

We keep the notation in Step 3 of the proof of 9.2. Assume that $S \neq \Sigma_{F,p}$. Consider the universal $\mathcal{O}_F$-linear abelian scheme $(A \to Z_F^{\mathrm{ord}}, \iota)$ and the $(\mathcal{O}_F \otimes \mathbb{Z}_p)$-linear p-divisible group $(A \to Z_F^{\mathrm{ord}}, \iota)[p^\infty]$ over the base scheme Z_F^{ord}, which is smooth over $\mathbb{F}$. We have a canonical decomposition of $X_\wp := A[p^\infty] \to Z_F^{\mathrm{ord}}$ as the fiber product over Z_F^{ord} of $\mathcal{O}_\wp$-linear p-divisible groups $A[\wp^\infty] \to Z_F^{\mathrm{ord}}$, where $\wp$ runs through the finite set $\Sigma_{F,p}$ of all places of F above p. Let $X_1 \to Z_F^{\mathrm{ord}}$ (resp. $X_2 \to Z_F^{\mathrm{ord}}$) be the fiber product over Z_F^{ord} of those $X_\wp$'s with $\wp \in S$ (resp. with $\wp \notin S$), so that we have $A[p^\infty] = X_1 \times_{Z_F^{\mathrm{ord}}} X_2$.

We know that for every closed point s of Z_F^{ord}, the restriction to the formal completion $Z_F^{/s}$ of the $\left(\prod_{\wp \notin S} \mathcal{O}_\wp\right)$-linear p-divisible group $X_2 \to Z_F^{\mathrm{ord}}$ is constant. This means that $X_2 \to Z_F^{\mathrm{ord}}$ is the twist of a constant $\left(\prod_{\wp \notin S} \mathcal{O}_\wp\right)$-linear p-divisible group by a character

$$\chi : \pi_1^{\text{ét}}(Z_F^{\mathrm{ord}}) \to \prod_{\wp \notin S} \mathcal{O}_\wp^\times .$$

More precisely, one twists the étale part and toric part of the constant p-divisible group by χ and χ^{-1} respectively. Consequently $\mathrm{End}_{\prod_{\wp \notin S} \mathcal{O}_\wp}(X_2) \supseteq \prod_{\wp \notin S} (\mathcal{O}_\wp \times \mathcal{O}_\wp)$.

By the main results in [**19**], we have an isomorphism

$$\begin{array}{rcl} \mathrm{End}_{\mathcal{O}_F}(A/Z_F^{\mathrm{ord}}) \otimes_{\mathbb{Z}} \mathbb{Z}_p & \xrightarrow{\sim} & \mathrm{End}_{\mathcal{O}_F \otimes_{\mathbb{Z}} \mathbb{Z}_p}(A[p^\infty]/Z_F^{\mathrm{ord}}) \\ & & \| \\ & & \mathrm{End}_{\prod_{\wp \in S} \mathcal{O}_\wp}(X_1) \times \mathrm{End}_{\prod_{\wp \notin S} \mathcal{O}_\wp}(X_2) . \end{array}$$

Since $\mathrm{End}_{\prod_{\wp \notin S} \mathcal{O}_\wp}(X_2) \supseteq \prod_{\wp \notin S} (\mathcal{O}_\wp \times \mathcal{O}_\wp)$, we conclude that $\mathrm{End}_{\mathcal{O}_F}(A/Z_F^{\mathrm{ord}}) \otimes_{\mathbb{Z}} \mathbb{Q}$ is either a totally imaginary quadratic extension field of F or a central quaternion algebra over F. This implies that the abelian scheme $A \to Z_F^{\mathrm{ord}}$ admits smCM (see §10.15), therefore it is isotrivial. We have arrived at a contradiction because $\dim(Z_F^{\mathrm{ord}}) > 0$ by 9.8. Therefore $S = \Sigma_{F,p}$.

10. Notations and some results used

10.1. Abelian varieties. For the definition of an abelian variety and an abelian scheme, see [**54**], II.4, [**55**], 6.1. The dual of an abelian scheme $A \to S$ will be denoted by $A^t \to S$. We avoid the notation $\hat{A}$ as in [**55**], 6.8 for the dual abelian scheme, because of possible confusion with the formal completion (of a ring, of a scheme at a subscheme).

An isogeny $\varphi : A \to B$ of abelian schemes is a finite, surjective homomorphism. It follows that $\mathrm{Ker}(\varphi)$ is finite and flat over the base, [**55**], Lemma 6.12. This defines a dual isogeny $\varphi^t : B^t \to A^t$. And see 10.11.

The dimension of an abelian variety we will usually denote by g. If $m \in \mathbb{Z}_{>1}$ and A is an abelian variety we write $A[m]$ for the group scheme of m-torsion points. Note that if $m \in \mathbb{Z}_{>0}$ is invertible on the base scheme S, then $A[m]$ is a group scheme finite étale over S; if moreover $S = \mathrm{Spec}(K)$, in this case it is uniquely determined by the Galois module $A[m](k)$. See 10.5 for details. If the characteristic p of the base field divides m, then $A[m]$ is a group scheme which is not reduced.

A divisor D on an abelian scheme A/S defines a morphism $\varphi_D : A \to A^t$, see [**54**], theorem on page 125, see [**55**], 6.2. A *polarization* on an abelian scheme $\mu : A \to A^t$ is an isogeny such that for every geometric point $s \in S(\Omega)$ there exists an ample divisor D on A_s such that $\lambda_s = \varphi_D$, see [**54**], Application 1 on page 60, and [**55**], Definition 6.3. Note that a polarization is *symmetric* in the sense that

$$\left(\lambda : A \to A^t\right) \quad = \quad \left(A \xrightarrow{\kappa} A^{tt} \xrightarrow{\lambda^t} A^t\right),$$

where $\kappa : A \to A^{tt}$ is the canonical isomorphism.

Writing $\varphi : (B, \mu) \to (A, \lambda)$ we mean that $\varphi : A \to A$ and $\varphi^*(\lambda) = \mu$, i.e.,

$$\mu = \left(B \xrightarrow{\varphi} A \xrightarrow{\lambda} A^t \xrightarrow{\varphi^t} B^t\right).$$

10.2. Warning. Most recent papers distinguish between an abelian variety defined over a field K on the one hand, and $A \otimes_K K'$ over $K' \supsetneqq K$ on the other hand. The notation $\mathrm{End}(A)$ stands for the ring of endomorphisms of A over K. This is the way Grothendieck taught us to choose our notation.

In pre-Grothendieck literature and in some modern papers there is confusion between on the one hand A/K and "the same" abelian variety over any extension field. Often it is not clear what is meant by "a point on A"; the notation $\mathrm{End}_K(A)$ can stand for the "endomorphisms defined over K", but then sometimes $\mathrm{End}(A)$ can stand for the "endomorphisms defined over $\overline{K}$".

Please adopt the Grothendieck convention that a scheme $T \to S$ is what it is, and any scheme obtained by base extension $S' \to S$ is denoted by $T \times_S S' = T_{S'}$, etc. For an abelian scheme $X \to S$ write $\mathrm{End}(X)$ for the endomorphism ring of $X \to S$ (old terminology "endomorphisms defined over S"). Do not write $\mathrm{End}_T(X)$ but $\mathrm{End}(X \times_S T)$.

10.3. Moduli spaces. We try to classify isomorphism classes of polarized abelian varieties (A, μ). This is described by the theory of moduli spaces; see [**55**]. In particular see Chapter 5 of this book, where the notions of coarse and fine moduli scheme are described. We adopt the notation of [**55**]. By $\mathcal{A}_g \to \mathrm{Spec}(\mathbb{Z})$ we denote the coarse moduli scheme of polarized abelian varieties of dimension g. Note that for an algebraically closed field k there is a natural identification of $\mathcal{A}_g(k)$ with the set of isomorphism classes of (A, μ) defined over k, with $\dim(A) = g$. We write $\mathcal{A}_{g,d}$ for the moduli space of polarized abelian varieties (A, μ) with $\deg(\mu) = d^2$. Note that $\mathcal{A}_g = \bigsqcup_d \mathcal{A}_{g,d}$. Given positive integers g, d, n, denote by $\mathcal{A}_{g,d,n} \to \mathrm{Spec}(\mathbb{Z}[1/dn])$ the moduli space considering polarized abelian varieties with a symplectic level-n structure; in this case it is assumed that we have chosen and fixed an isomorphism from the constant group scheme $\mathbb{Z}/n\mathbb{Z}$ to μ_n over k, so

that symplectic level-n structure makes sense. According to these definitions we have $\mathcal{A}_{g,d,1} = \mathcal{A}_{g,d} \times_{\mathrm{Spec}(\mathbb{Z})} \mathrm{Spec}(\mathbb{Z}[1/d])$.

Most of the considerations in this course are over fields of characteristic p. Working over a field of characteristic p we should write $\mathcal{A}_g \otimes \mathbb{F}_p$ for the moduli space under consideration; however, in case it is clear what the base field K or the base scheme is, instead we will $\mathcal{A}_g$ instead of the notation $\mathcal{A}_g \otimes K$; we hope this will not cause confusion.

10.4. The Cartier dual. The group schemes considered will be assumed to be *commutative*. If G is a finite abelian group, and S is a scheme, we write $\underline{G}_S$ for the constant group scheme over S with fiber equal to G.

Let $N \to S$ be a finite locally free commutative group scheme. Let $\underline{\mathrm{Hom}}(N, \mathbb{G}_m)$ be the functor on the category of all S-schemes, whose value at any S-scheme T is $\mathrm{Hom}_T(N \times_S T, \mathbb{G}_m \times_{\mathrm{Spec}(\mathbb{Z})} T)$. This functor is a sheaf for the fpqc topology, and is representable by a flat locally free scheme N^D over S, see [**61**], I.2. This group scheme $N^D \to S$ is called the *Cartier dual* of $N \to S$, and it can be described explicitly as follows. If $N = \mathrm{Spec}(E) \to S = \mathrm{Spec}(R)$ we write $E^D := \mathrm{Hom}_R(E, R)$. The multiplication map on E gives a comultiplication on E^D, and the commutative comultiplication on E provides E^D with the structure of a commutative ring. With the inverse map they give E^D a structure of a commutative and cocommutative bialgebra over R, and make $\mathrm{Spec}(E^D)$ into a commutative group scheme. This commutative group scheme $\mathrm{Spec}(E^D)$ is naturally isomorphic to the Cartier dual N^D of N. It is a basic fact, easy to prove, that the natural homomorphism $N \to (N^D)^D$ is an isomorphism for every finite locally free group scheme $N \to S$.

Examples. The constant group schemes $\underline{\mathbb{Z}/n\mathbb{Z}}$ and $\mu_n := \mathrm{Ker}\,([n]_{\mathbb{G}_m})$ are Cartier dual to each other, over any base scheme. More generally, a finite commutative group scheme $N \to S$ is étale if and only if its Cartier dual $N^D \to S$ is *of multiplicative type*, i.e., there exists an étale surjective morphism $g : T \to S$, such that $N^D \times_S T$ is isomorphic to a direct sum of group schemes μ_{n_i} for suitable positive integers n_i. The above morphism $g : T \to S$ can be chosen to be finite étale surjective.

For every field $K \supset \mathbb{F}_p$, the group scheme α_p is self-dual. Recall that α_p is the kernel of the Frobenius endomorphism Fr_p on $\mathbb{G}_a$.

10.5. Étale finite group schemes as Galois modules. (Any characteristic.) Let K be a field, and let $G = \mathrm{Gal}(K^{\mathrm{sep}}/K)$. The main theorem of Galois theory says that there is an equivalence between the category of finite étale K-algebras and the category of finite sets with a continuous G-action. Taking group objects on both sides we arrive at:

Theorem. *There is an equivalence between the category of commutative finite étale group schemes over K and the category of finite continuous G-modules.*

See [**84**], 6.4. Note that an analogous equivalence holds in the case of not necessarily commutative group schemes.

This is a special case of the following. Let S be a connected scheme, and let $s \in S$ be a geometric base point; let $\pi = \pi_1(S, s)$. *There is an equivalence between the category of étale finite schemes over S and the category of finite continuous π-sets.* Here $\pi_1(S, s)$ is the algebraic fundamental group defined by Grothendieck in SGA 1; see [**33**].

Hence the definition of $T_\ell(A)$ for an abelian variety over a field K with $\ell \neq \mathrm{char}(K)$ can be given as:

$$T_\ell(A) = \varprojlim_i A[\ell^i](K^{\mathrm{sep}}),$$

considered as a continuous $\mathrm{Gal}(K^{\mathrm{sep}}/K)$-module.

Definition 10.6. Let S be a scheme. A *p-divisible group* over S is an inductive system $X = (X_n, \iota_n)_{n\in\mathbb{N}_{>0}}$ of finite locally free commutative group schemes over S satisfying the following conditions.

(i) X_n is killed by p^n for every $n \geq 1$.
(ii) Each homomorphism $\iota_n : X_n \to X_{n+1}$ is a closed embedding.
(iii) For each $n \geq 1$ the homomorphism $[p]_{X_{n+1}} : X_{n+1} \to X_{n+1}$ factors through $\iota_n : X_n \to X_{n+1}$, such that the resulting homomorphism $X_{n+1} \to X_n$ is faithfully flat. In other words there is a faithfully flat homomorphism $\pi_n : X_{n+1} \to X_n$ such that such that $\iota_n \circ \pi_n = [p]_{X_{n+1}}$. Here $[p]_{X_{n+1}}$ is the endomorphism "multiplication by p" on the commutative group scheme X_{n+1}.

Sometimes one writes $X[p^n]$ for the finite group scheme X_n. Equivalent definitions can be found in [**34**, Chapter III] and [**49**, Chapter I] and [**38**]; these are basic references to p-divisible groups.

Some authors use the terminology "Barsotti-Tate group", a synonym for "p-divisible group".

A p-divisible group $X = (X_n)$ over S is said to be *étale* (resp. *toric*) if every X_n is finite étale over S (resp. of multiplicative type over S).

For any p-divisible group $X \to S$, there is a locally constant function $h : S \to \mathbb{N}$, called the *height* of X, such that $\mathcal{O}_{X_n}$ is a locally free $\mathcal{O}_S$-algebra of rank p^h for every $n \geq 1$.

Example.

(1) Over any base scheme S we have the constant p-divisible group $\underline{\mathbb{Q}_p/\mathbb{Z}_p}$ of height 1, defined as the inductive limit of the constant groups $\underline{p^{-n}\mathbb{Z}/\mathbb{Z}}$ over S.
(2) Over any base scheme S, the p-divisible group $\mu_{p^\infty} = \mathbb{G}_m[p^\infty]$ is the inductive system $(\mu_{p^n})_{n\geq 1}$, where $\mu_{p^n} := \mathrm{Ker}\,([p^n]_{\mathbb{G}_m})$.
(3) Let $A \to S$ be an abelian scheme. For every i we write $G_i = A[p^i]$. The inductive system $G_i \subset G_{i+s} \subset A$ defines a p-divisible group of height $2g$. We shall denote this by $X = A[p^\infty]$ (although of course "p^∞" strictly speaking is not defined). A homomorphism $A \to B$ of abelian schemes defines a morphism $A[p^\infty] \to B[p^\infty]$ of p-divisible groups.

10.7. The Serre dual of a p-divisible group. Let $X = (X_n)_{n\in\mathbb{Z}_{>0}}$ be a p-divisible group over a scheme S. The *Serre dual* of X is the p-divisible group $X^t = \left(X_n^{\mathrm{D}}\right)_{n\geq 1}$ over S, where $X_n^{\mathrm{D}} := \underline{\mathrm{Hom}}_S(X_n, \mathbb{G}_m)$ is the Cartier dual of X_n, the embedding $X_n^{\mathrm{D}} \to X_{n+1}^{\mathrm{D}}$ is the Cartier dual of the faithfully flat homomorphism $\pi_n : X_{n+1} \to X_n$, and the faithfully flat homomorphism $X_{n+1}^{\mathrm{D}} \to X_n^{\mathrm{D}}$ is the Cartier dual of the embedding $\iota_n : X_n \to X_{n+1}$.

As an example, over any base scheme S the p-divisible group μ_{p^∞} is the Serre dual of the constant p-divisible group $\underline{\mathbb{Q}_p/\mathbb{Z}_p}$, because μ_{p^n} is the Cartier dual of $\underline{p^{-n}\mathbb{Z}/\mathbb{Z}}$. Below are some basic properties of Serre duals.

(1) The height of X^t is equal to the height of X.
(2) The Serre dual of a short exact sequence of p-divisible groups is exact.
(3) The Serre dual of X^t is naturally isomorphic to X.
(4) A p-divisible group $X = (X_n)$ is *toric* if and only if its Serre dual $X^t = (X_n^{\mathrm{D}})$ is étale. If this is the case, then the sheaf $\underline{\mathrm{Hom}}(X, \mu_{p^\infty})$ of *characters* of X is the projective limit of the étale sheaves X_n^{D}, where the transition map $X_{n+1}^{\mathrm{D}} \to X_n^{\mathrm{D}}$ is the Cartier dual of the embedding $\iota_n : X_n \to X_{n+1}$.
(5) Let $A \to S$ be an abelian scheme, and let $A^t \to S$ be the dual abelian scheme. Then the Serre dual of the p-divisible group $A[p^\infty]$ attached to the abelian scheme $A \to S$ is the p-divisible group $A^t[p^\infty]$ attached to the dual abelian scheme $A^t \to S$; see 10.11.

10.8. Discussion. Over any base scheme S (in any characteristic) for an abelian scheme $A \to S$ and for a prime number ℓ invertible in $\mathcal{O}_S$ one can define $T_\ell(A/S)$ as follows. For $i \in \mathbb{Z}_{>0}$ one chooses $N_i := A[\ell^i]$, regarded as a smooth étale sheaf of free $\mathbb{Z}/\ell^i\mathbb{Z}$-modules of rank $2\dim(A)$, and we have surjective maps $[\ell] : N_{i+1} \to N_i$ induced by multiplication by ℓ. The projective system of the N_i's "is" a smooth étale sheaf of $\mathbb{Z}_\ell$-modules of rank $2\dim(A)$, called the *ℓ-adic Tate module* of A/S, denoted by $T_\ell(A/S)$. Alternatively, we can consider $T_\ell(A/S)$ as a projective system

$$T_\ell(A/S) \quad = \quad \varprojlim_{i\in\mathbb{N}} A[\ell^i]$$

of the finite étale group schemes $A[\ell^i]$ over S. This projective system we call the *Tate ℓ-group* of A/S. Any geometric fiber of $T_\ell(A/S)_{\bar{s}}$ is constant, hence the projective limit of $T_\ell(A/S)_{\bar{s}}$ is isomorphic to $(\underline{\mathbb{Z}_\ell}_s)^{2g}$. If S is the spectrum of a field K, the Tate ℓ-group can be considered as a $\mathrm{Gal}(K^{\mathrm{sep}}/K)$-module on the group $\mathbb{Z}_\ell^{2g}$, see 10.5. *One should like to have an analogous concept for this notion in case p is not invertible on S.* This is precisely the role of $A[p^\infty]$ defined above. Historically a Tate ℓ-group is defined as a projective system, and the p-divisible group as an inductive system; it turns out that these are the best ways of handling these concepts (but the way in which direction to choose the limit is not very important). We see that the p-divisible group of an abelian variety should be considered as the natural substitute for the Tate ℓ-group.

In order to carry this analogy further we investigate aspects of $T_\ell(A)$ and wonder whether these can be carried over to $A[p^\infty]$ in case of an abelian variety A in characteristic p. The Tate ℓ-group is a twist of a pro-group scheme defined over $\mathrm{Spec}(\mathbb{Z}[1/\ell])$. What can be said in analogy about $A[p^\infty]$ in the case of an abelian variety A in characteristic p? We will see that *up to isogeny* $A[p^\infty]$ is a twist of an ind-group scheme over $\mathbb{F}_p$; however, "twist" here should be understood not only in the sense of separable Galois theory, but also using inseparable aspects: the main idea of Serre-Tate parameters, to be discussed in Section 2.

10.9. Let X be a p-divisible group over an Artinian local ring R whose residue field is of characteristic p.

(1) There exists a largest étale quotient p-divisible group $X_{\text{ét}}$ of X over R, such that every homomorphism from X to an étale p-divisible group factors uniquely through $X_{\text{ét}}$. The kernel of $X \to X_{\text{ét}}$ is called the *neutral component* of X, or the maximal connected p-divisible subgroup of X, denoted by X_{conn}.
(2) The Serre dual of the maximal étale quotient $X^t_{\text{ét}}$ of X^t is called the *toric part* of X, denoted X_{tor}. Alternatively, $X_{\text{tor}}[p^n]$ is the maximal subgroup scheme in $X[p^n]$ of multiplicative type, for each $n \geq 1$.
(3) We have two short exact sequences of p-divisible groups

$$0 \to X_{\text{tor}} \to X_{\text{conn}} \to X_{\ell\ell} \to 0$$

and

$$0 \to X_{\text{conn}} \to X \to X_{\text{ét}} \to 0$$

over R, where $X_{\ell\ell}$ is a p-divisible group over $R\dim(A)$ with trivial étale quotient and trivial toric part. The closed fiber of the p^n-torsion subgroup $X_{\ell\ell}[p^n]$ of $X_{\ell\ell}$ is unipotent for every $n \geq 1$.
(4) The scheme-theoretic inductive limit of X_{conn} (resp. X_{tor}) is a finite dimensional commutative formal group scheme $X^\wedge_{\text{conn}}$ over R (resp. a finite dimensional formal torus $X^\wedge_{\text{tor}}$ over R), called the *formal completion* of X_{conn} (resp. X_{tor}). The endomorphism $[p^n]$ on $X^\wedge_{\text{conn}}$ (resp. $X^\wedge_{\text{tor}}$) is faithfully flat; its kernel is canonically isomorphic to $X_{\text{conn}}[p^n]$ (resp. $X_{\text{tor}}[p^n]$). In particular one can recover the p-divisible groups X_{conn} (resp. X_{tor}) from the smooth formal group $X^\wedge_{\text{conn}}$ (resp. $X^\wedge_{\text{tor}}$).
(5) If $X = A[p^\infty]$ is the p-divisible group attached to an abelian scheme A over R, then $X^\wedge_{\text{conn}}$ is canonically isomorphic to the formal completion of A along its zero section.

A p-divisible group X over an Artinian local ring R whose maximal étale quotient is trivial is often said to be *connected*. Note that $X[p^n]$ is connected, or equivalently, geometrically connected, for every $n \geq 1$. The formal completion of a connected p-divisible group over R is usually called a *p-divisible formal group*. It is not difficult to see that a smooth formal group over an Artinian local ring R is a p-divisible formal group if and only its closed fiber is.

More information about the infinitesimal properties of p-divisible groups can be found in [**49**] and [**38**]. Among other things one can define the Lie algebra of a p-divisible group $X \to S$ when p is locally nilpotent in $\mathcal{O}_S$; it coincides with the Lie algebra of the formal completion of X_{conn} when S is the spectrum of an Artinian local ring.

10.10. The following are equivalent conditions for a g-dimensional abelian variety A over an algebraically closed field $k \supset \mathbb{F}_p$; A is said to be *ordinary* if these conditions are satisfied.

(1) $\text{Card}(A[p](k)) = p^g$, i.e., the p-rank of A is equal to g.
(2) $A[p^n](k) \cong \mathbb{Z}/p^n\mathbb{Z}$ for some positive integer n.
(3) $A[p^n](k) \cong \mathbb{Z}/p^n\mathbb{Z}$ for every positive integer n.
(4) The formal completion $A^{/0}$ of A along the zero point is a formal torus.
(5) The p-divisible group $A[p^\infty]$ attached to A is an extension of an étale p-divisible group of height g by a toric p-divisible group of height g.
(6) The σ-linear endomorphism on $\text{H}^1(A, \mathcal{O}_A)$ induced by the absolute Frobenius of A is bijective, where σ is the Frobenius automorphism on k.

Note that for an ordinary abelian variety A over a field $K \supset \mathbb{F}_p$ the Galois group $\mathrm{Gal}(K^{\mathrm{sep}}/K)$ acts on $A[p]_{\mathrm{loc}}$ and on $A[p]_{\text{ét}} = A[p]/A[p]_{\mathrm{loc}}$, and these actions need not be trivial. Moreover if K is not perfect, the extension

$$0 \to A[p]_{\mathrm{loc}} \to A[p] \to A[p]_{\text{ét}} \to 0$$

need not be split; this is studied extensively in Section 2.

Reminder. It is a general fact that every finite group scheme G over a field K sits naturally in the middle of a short exact sequence $0 \to G_{\mathrm{loc}} \to G \to G_{\text{ét}} \to 0$ of finite group schemes over K, where $G_{\text{ét}}$ is étale and G_{loc} is connected. If the rank of G is prime to the characteristic of K, then G is étale over K, i.e., G_{loc} is trivial; e.g. see [**60**].

10.11. We recall the statement of a basic duality result for abelian schemes over an arbitrary base scheme.

Theorem. (Duality theorem for abelian schemes, see [**61**], Theorem 19.1) *Let $\varphi : B \to A$ be an isogeny of abelian schemes. We obtain an exact sequence*

$$0 \quad \to \quad \mathrm{Ker}(\varphi)^D \quad \longrightarrow \quad A^t \quad \xrightarrow{\varphi^t} \quad B^t \quad \to \quad 0.$$

An application. Let A be a g-dimensional abelian variety over a field $K \supset \mathbb{F}_p$, and let A^t be the dual abelian variety of A. Then $A[n]$ and $A^t[n]$ are dual to each other for every non-zero integer n. This natural duality pairing identifies the maximal étale quotient of $A[n]$ (resp. $A^t[n]$ with the Cartier dual of the maximal subgroup of $A^t[n]$ (resp. $A[n]$ of multiplicative type. This implies that the Serre dual of the p-divisible group $A[p^\infty]$ is isomorphic to $A^t[p^\infty]$. Since A and A^t are isogenous, we deduce that the maximal étale quotient of the p-divisible group $A[p^\infty]$ and the maximal toric p-divisible subgroup of $A[p^\infty]$ have the same height.

10.12. Endomorphism rings. Let A be an abelian variety over a field K, or more generally, an abelian scheme over a base scheme S. We write $\mathrm{End}(A)$ for the *endomorphism ring* of A. For every $n \in \mathbb{Z}_{>0}$, multiplication by n on A is an epimorphic morphism of schemes because it is faithfully flat, hence $\mathrm{End}(A)$ is torsion-free. In the case S is connected, $\mathrm{End}(A)$ is a free $\mathbb{Z}$-module of finite rank. We write $\mathrm{End}^0(A) = \mathrm{End}(A) \otimes_{\mathbb{Z}} \mathbb{Q}$ for the *endomorphism algebra* of A. By Wedderburn's theorem every central simple algebra is a matrix algebra over a division algebra. If A is K-simple the algebra $\mathrm{End}^0(A)$ is a division algebra; in that case we write:

$$\mathbb{Q} \quad \subset \quad L_0 \quad \subset \quad L := \mathrm{Centre}(D) \quad \subset \quad D = \mathrm{End}^0(A);$$

here L_0 is a totally real field, and either $L = L_0$ or $[L : L_0] = 2$ and in that case L is a CM-field. In case A is simple $\mathrm{End}^0(A)$ is one of the four types in the Albert classification (see below). We write:

$$[L_0 : \mathbb{Q}] = e_0, \quad [L : \mathbb{Q}] = e, \quad [D : L] = d^2.$$

10.13. Let $(A, \mu) \to S$ be a polarized abelian scheme. As μ is an isogeny, there exist μ' and $n \in \mathbb{Z}_{>0}$ such that $\mu' \cdot \mu = n$; think of μ'/n as the inverse of μ. We define the *Rosati involution* $\varphi \mapsto \varphi^\dagger$ by

$$\varphi \mapsto \varphi^\dagger := \frac{1}{n} \mu' \cdot \varphi^t \cdot \mu, \qquad \varphi \in D = \mathrm{End}^0(A).$$

The definition does not depend on the choice of μ' and n; it can be characterized by $\varphi^t \cdot \mu = \mu \cdot \varphi^\dagger$. This map $^\dagger : D \to D$ is an anti-involution on D.

The Rosati involution $^\dagger : D \to D$ is positive definite; for references see Proposition II in 3.10.

Definition. A *simple division algebra of finite degree over $\mathbb{Q}$ with a positive definite involution*, i.e., an anti-isomorphism of order two which is positive definite, is called an *Albert algebra*.

Applications to abelian varieties and the classification have been described by Albert, [**1**], [**4**] [**2**], [**3**].

10.14. Albert's classification. Any Albert algebra belongs to one of the following types.

Type I(e_0) Here $L_0 = L = D$ is a totally real field.

Type II(e_0) Here $d = 2$, $e = e_0$, $\mathrm{inv}_v(D) = 0$ for all infinite v, and D is an indefinite quaternion algebra over the totally real field $L_0 = L$.

Type III(e_0) Here $d = 2$, $e = e_0$, $\mathrm{inv}_v(D) \neq 0$ for all infinite v, and D is a definite quaternion algebra over the totally real field $L_0 = L$.

Type IV(e_0, d) Here L is a CM-field, $[F : \mathbb{Q}] = e = 2e_0$, and $[D : L] = d^2$.

10.15. smCM. We say that an abelian variety X over a field K *admits sufficiently many complex multiplications over K*, abbreviated by "smCM over K", if $\mathrm{End}^0(X)$ contains a commutative semi-simple subalgebra of rank $2{\cdot}\dim(X)$ over $\mathbb{Q}$. Equivalently: for every simple abelian variety Y over K which admits a non-zero homomorphism to X the algebra $\mathrm{End}^0(Y)$ contains a field of degree $2{\cdot}\dim(Y)$ over $\mathbb{Q}$. For other characterizations see [**21**], page 63, see [**53**], page 347.

Note that if a simple abelian variety X of dimension g over a field *of characteristic zero* admits smCM then its endomorphism algebra $L = \mathrm{End}^0(X)$ is a CM-field of degree $2g$ over $\mathbb{Q}$. We will use the terminology "CM-type" in the case of an abelian variety X over $\mathbb{C}$ which admits smCM, and where the type is given, i.e., the action of the endomorphism algebra on the tangent space $T_{X,0} \cong \mathbb{C}^g$ is part of the data.

Note however that there exist (many) abelian varieties A admitting smCM (defined over a field of positive characteristic), such that $\mathrm{End}^0(A)$ is not a field.

By Tate we know that an abelian variety over a finite field admits smCM, see 10.17. By Grothendieck we know that an abelian variety over an algebraically closed field $k \supset \mathbb{F}_p$ which admits smCM is isogenous to an abelian variety defined over a finite field, see 10.19.

Terminology. Let $\varphi \in \mathrm{End}^0(A)$. Then $d\varphi$ is a K-linear endomorphism of the tangent space. If the base field is $K = \mathbb{C}$, this is just multiplication by a complex matrix x, and every multiplication by a complex matrix x leaving invariant the lattice Λ, where $A(\mathbb{C}) \cong \mathbb{C}^g/\Lambda$, gives rise to an endomorphism of A. If $g = 1$, i.e., A is an elliptic curve, and $\varphi \notin \mathbb{Z}$ then $x \in \mathbb{C}$ and $x \notin \mathbb{R}$. Therefore an endomorphism of an elliptic curve over $\mathbb{C}$ which is not in $\mathbb{Z}$ is sometimes called "a complex multiplication". Later this terminology was extended to all abelian varieties.

Warning. Sometimes the terminology "an abelian variety with CM" is used, when one wants to say "admitting smCM". An elliptic curve E has $\mathrm{End}(E) \supsetneqq \mathbb{Z}$ if and only if it admits smCM. Note that it is easy to give an abelian variety A which "admits CM", meaning that $\mathrm{End}(A) \supsetneqq \mathbb{Z}$, such that A does not admit smCM. However we will use the terminology "a CM-abelian variety" for an abelian variety which admits smCM.

Exercise 10.16. Show there exists an abelian variety A over a field k such that $\mathbb{Z} \subsetneqq \mathrm{End}(A)$ and such that A does not admit smCM.

Theorem 10.17. (Tate) *Let A be an abelian variety over a finite field.*
(1) *The algebra $\mathrm{End}^0(A)$ is semi-simple. Suppose A is simple; the center of $\mathrm{End}^0(A)$ equals $L := \mathbb{Q}(\pi_A)$.*
(2) *Suppose A is simple; then*

$$2g \quad = \quad [L:\mathbb{Q}]\cdot\sqrt{[D:L]}\,,$$

where g is the dimension of A. Hence: every abelian variety over a finite field admits smCM. *See* 10.15. *Moreover we have*

$$f_A \quad = \quad (\mathrm{Irr}_{\pi_A})^{\sqrt{[D:L]}}\,.$$

Here f_Z is the characteristic polynomial of the Frobenius morphism $\mathrm{Fr}_{A,\mathbb{F}_q} : A \to A$, and Irr_{π_A} is the irreducible polynomial over $\mathbb{Q}$ of the element π_A in the finite extension $L/\mathbb{Q}$.
(3) *Suppose A is simple,*

$$\mathbb{Q} \quad \subset \quad L := \mathbb{Q}(\pi_A) \quad \subset \quad D = \mathrm{End}^0(A).$$

The central simple algebra D/L

- *does not split at every real place of L,*
- *does split at every finite place not above p,*
- *and for $v \mid p$ the invariant of D/L is given by*

$$\mathrm{inv}_v(D/L) = \frac{v(\pi_A)}{v(q)}\cdot[L_v:\mathbb{Q}_p] \mod \mathbb{Z},$$

where L_v is the local field obtained from L by completing at v.

See [**78**], [**79**].

Remark 10.18. An abelian variety over a field of characteristic zero which admits smCM is defined over a number field; e.g. see [**77**], Proposition 26 on page 109.

Remark 10.19. The converse of Tate's result 10.17 (2) is almost true. We have the following theorem of Grothendieck: *Let A be an abelian variety over a field K which admits* smCM; *then A_k is isogenous to an abelian variety defined over a finite extension of the prime field, where $k = \overline{K}$*; see [**62**].

It is easy to give an example of an abelian variety (over a field of characteristic p), with smCM which is not defined over a finite field.

Exercise 10.20. Give an example of a simple abelian variety A over a field K such that $A \otimes \overline{K}$ is not simple.

10.21. Algebraization.

(1) Suppose we are given a formal p-divisible group X_0 over k with $\mathcal{N}(X_0) = \gamma$ ending at (h, c). We write $\mathcal{D}^\wedge = \mathrm{Def}(X_0)$ for the universal deformation space in equal characteristic p. By this we mean the following. Formal deformation theory of X_0 is prorepresentable; we obtain a formal scheme $\mathrm{Spf}(R)$ and a prorepresenting family $\mathcal{X}' \to \mathrm{Spf}(R)$. However, "a finite group scheme over a formal scheme actually is already defined over an actual scheme". Indeed, by [**17**], Lemma 2.4.4 on page 23, we know that there is an equivalence of categories of p-divisible groups over $\mathrm{Spf}(R)$, respectively over $\mathrm{Spec}(R)$. We write $\mathcal{D}(X_0) = \mathrm{Spec}(R)$, and corresponding to the pro-universal family $\mathcal{X}' \to \mathrm{Spf}(R)$ we have a family $\mathcal{X} \to \mathcal{D}(X_0)$. We will say that $\mathcal{X} \to \mathrm{Spec}(R) = \mathcal{D}(X_0)$ is the universal deformation of X_0 if the corresponding $\mathcal{X}' \to \mathrm{Spf}(R) = \mathcal{D}^\wedge = \mathrm{Def}(X_0)$ prorepresents the deformation functor.

Note that for a formal p-divisible group $\mathcal{X}' \to \mathrm{Spf}(R)$, where R is moreover an integral domain, it makes sense to consider "the generic fiber" of $\mathcal{X}/\mathrm{Spec}(R)$.

(2) Let A_0 be an abelian variety. The deformation functor $\mathrm{Def}(A_0)$ is prorepresentable. We obtain the prorepresenting family $A \to \mathrm{Spf}(R)$, which is a formal abelian scheme. If $\dim(A_0) > 1$ this family is *not algebraizable*, i.e., it does not come from an actual scheme over $\mathrm{Spec}(R)$.

(3) Let (A_0, μ_0) be a polarized abelian variety. The deformation functor $\mathrm{Def}(A_0, \mu_0)$ is prorepresentable. We can use the Chow-Grothendieck theorem, see [**32**], III^1.5.4 (this is also called a theorem of "GAGA-type"): the formal polarized abelian scheme obtained is algebraizable, and we obtain the universal deformation as a polarized abelian scheme over $\mathcal{D}(A_0, \mu_0) = \mathrm{Spec}(R)$.

The notions mentioned in (1), (2) and (3) will be used without further mention, assuming the reader to be familiar with the subtle differences between $\mathcal{D}(-)$ and $\mathrm{Def}(-)$.

10.22. Fix a prime number p. Base schemes and base fields will be of characteristic p, unless otherwise stated. We write k or Ω for an algebraically closed field. For the rest of this section we are working in characteristic p.

10.23. The Frobenius morphism. For a scheme S over $\mathbb{F}_p$ (i.e., $p \cdot 1 = 0$ in all fibers of $\mathcal{O}_S$), we define the absolute Frobenius morphism $\mathrm{fr} : S \to S$; if $S = \mathrm{Spec}(R)$ this is given by $x \mapsto x^p$ in R.

For a scheme $A \to S$ we define $A^{(p)}$ as the fiber product of $A \to S \xleftarrow{\mathrm{fr}} S$. The morphism $\mathrm{fr} : A \to A$ factors through $A^{(p)}$. This defines $F_A : A \to A^{(p)}$, a morphism over S; this is called *the relative Frobenius morphism*. If A is a group scheme over S, the morphism $F_A : A \to A^{(p)}$ is a homomorphism of group schemes. For more details see [**35**], Exp. VII$_A$.4. The notation $A^{(p/S)}$ is (maybe) more correct.

Examples. Suppose $A \subset \mathbb{A}^n_R$ is given as the zero set of a polynomial $\sum_I a_I X^I$ (multi-index notation). Then $A^{(p)}$ is the zero set of $\sum_I a_I^p X^I$, and $A \to A^{(p)}$ is given, on coordinates, by raising these to the power p. Note that if a point

$(x_1, \cdots, x_n) \in A$ then indeed $(x_1^p, \ldots, x_n^p) \in A^{(p)}$, and $x_i \mapsto x_i^p$ describes $F_A : A \to A^{(p)}$ on points.

Let $S = \mathrm{Spec}(\mathbb{F}_p)$; for any $T \to S$ we have a canonical isomorphism $T \cong T^{(p)}$. In this case $F_T = \mathrm{fr} : T \to T$.

10.24. Verschiebung. Let A be a *commutative* group scheme flat over a characteristic p base scheme. In [**35**], Exp. VII$_A$.4 we find the definition of the "relative Verschiebung"

$$V_A : A^{(p)} \to A; \quad \text{we have:} \quad F_A \cdot V_A = [p]_{A^{(p)}}, \quad V_A \cdot F_A = [p]_A.$$

In case A is an abelian variety we see that F_A is a faithfully flat homomorphism, and $\mathrm{Ker}(F_A) \subset A[p]$. In this case we do not need the somewhat tricky construction of [**35**], Exp. VII$_A$.4: since the kernel of the isogeny $F_A : A \to A^{(p)}$ is killed by p, we can define V_A as the isogeny from $A^{(p)}$ to A such that $V_A \cdot F_A = [p]_A$, and the equality $F_A \cdot V_A = [p]_{A^{(p)}}$ follows from $F_A \cdot V_A \cdot F_A = [p]_{A^{(p)}} \cdot F_A$ because F_A is faithfully flat.

Remark 10.25. We use covariant Dieudonné module theory. The Frobenius on a group scheme G defines the Verschiebung on $\mathbb{D}(G)$; this we denote by $\mathcal{V}$, in order to avoid possible confusion. In the same way as "$\mathbb{D}(F) = \mathcal{V}$" we have "$\mathbb{D}(V) = \mathcal{F}$". See [**67**], 15.3.

Theorem 10.26. **BB** (Irreducibility of moduli spaces) *Let K be a field, and consider $\mathcal{A}_{g,1,n} \otimes K$ the moduli space of principally polarized abelian varieties over K-schemes, where $n \in \mathbb{Z}_{>0}$ is prime to the characteristic of K. This moduli scheme is geometrically irreducible.*

For fields of characteristic zero this follows by complex uniformization. For fields of positive characteristic this was proved by Faltings in 1984, see [**27**], at the same time for $p > 2$ by Chai in his Harvard PhD thesis, see [**8**]; also see [**28**], IV.5.10. For a pure characteristic-p-proof see [**67**], 1.4.

11. A remark and some questions

11.1. In 1.13 we have seen that the closure of the *full* Hecke orbit equals the related Newton polygon stratum. That result finds its origin in the construction of two *foliations*, as in [**68**]: Hecke-prime-to-p actions "move" a point in a central leaf, and Hecke actions only involving compositions of isogenies with kernel isomorphic to α_p "move" a point in an isogeny leaf, called $\mathcal{H}_\alpha$-actions; as an open Newton polygon stratum, up to a finite map, is equal to the product of a central leaf and an isogeny leaf, the result 1.13 for an irreducible component of a Newton polygon stratum follows if we show that $\mathcal{H}_\ell(x)$ is dense in the central leaf passing through x.

In the case of ordinary abelian varieties the central leaf is the whole open Newton polygon stratum. As the Newton polygon goes up central leaves get smaller. Finally, for supersingular points a central leaf is finite (see Lemma 1.14) and an isogeny leaf of a supersingular point is the whole supersingular locus.

In order to finish a proof of 1.13 one shows that Hecke-α actions act transitively on the set of geometric components of the supersingular locus, and that any Newton polygon stratum in $\mathcal{A}_{g,1}$ which is not supersingular is geometrically irreducible, see [**14**].

11.2. Let D be an Albert algebra; i.e., D is a division algebra, it is of finite rank over $\mathbb{Q}$, and it has a positive definite anti-involution $\dagger : D \to D$. Suppose a characteristic is given. There exists a field k of that characteristic, and an abelian variety A over k such that $\mathrm{End}^0(A) \cong D$, and such that $\dagger$ is the Rosati involution given by a polarization on A. This was proved by Albert, and by Shimura over $\mathbb{C}$ (see [**76**], Theorem 5). In general this was proved by Gerritzen [**30**]; for more references see [**64**].

One can ask which possibilities we have for $\dim(A)$, once D is given. This question is completely settled in characteristic zero. From properties of D one can derive some restrictions on $\dim(A)$. However the question which dimensions $\dim(A)$ can appear for a given D in positive characteristic is not yet completely settled.

Also, there is not yet a complete criterion for which endomorphism algebras can appear in positive characteristic.

11.3. In Section 5, in particular see the proofs of 5.10 and 5.16, we have seen a natural way of introducing coordinates in the formal completion at a point x where $a \leq 1$ on an (open) Newton polygon stratum:

$$(\mathcal{W}_\xi(\mathcal{A}_{g,1,n}))^{/x} = \mathrm{Spf}(B_\xi),$$

see the proof of 5.19. It would be nice to have a better understanding and interpretation of these "coordinates".

As in [**58**] we write

$$\triangle(\xi;\xi^*) := \{(x,y) \in \mathbb{Z} \mid (x,y) \prec \xi, \quad (x,y) \gneqq \xi^*, \quad x \leq g\}.$$

We write

$$B_{(\xi;\xi^*)} = k[[Z_{x,y} \mid (x,y) \in \triangle(\xi;\xi^*)]].$$

The inclusion $\triangle(\xi;\xi^*) \subset \triangle(\xi)$ defines $B_\xi \twoheadrightarrow B_{(\xi;\xi^*)}$ by equating to zero those elements $Z_{x,y}$ with $(x,y) \notin \Delta(\xi;\xi^*)$. Hence $\mathrm{Spf}(B_{(\xi;\xi^*)}) \subset \mathrm{Spf}(B_\xi)$. We also have the inclusion $\mathcal{C}(x) \subset \mathcal{W}_\xi(\mathcal{A}_{g,1,n})$.

Question. Does the inclusion $\triangle(\xi,\xi^*) \subset \triangle(\xi)$ define the inclusion $(\mathcal{C}(x))^{/x} \subset (\mathcal{W}_\xi(\mathcal{A}_{g,1,n}))^{/x}$?

A positive answer would give more insight in these coordinates, also along a central leaf, and perhaps a new proof of results in [**58**].

References

1. A. A. Albert, *On the construction of Riemann matrices. I*, Ann. of Math. (2) **35** (1934), no. 1, 1–28. MR 1503140
2. ———, *A solution of the principal problem in the theory of Riemann matrices*, Ann. of Math. (2) **35** (1934), no. 3, 500–515. MR 1503176
3. ———, *Involutorial simple algebras and real Riemann matrices*, Ann. of Math. (2) **36** (1935), no. 4, 886–964. MR 1503260
4. ———, *On the construction of Riemann matrices. II*, Ann. of Math. (2) **36** (1935), no. 2, 376–394. MR 1503230
5. P. Berthelot, L. Breen, and W. Messing, *Théorie de Dieudonné cristalline. II*, Lecture Notes in Mathematics, vol. 930, Springer-Verlag, Berlin, 1982. MR 667344 (85k:14023)
6. L. Breen, *Rapport sur la théorie de Dieudonné*, Journées de Géométrie Algébrique de Rennes (Rennes, 1978), Vol. I, Astérisque, vol. 63, Soc. Math. France, Paris, 1979, pp. 39–66. MR 563459 (81j:14025)
7. C.-L. Chai, *Families of ordinary abelian varieties: canonical coordinates, p-adic monodromy, Tate-linear subvarieties and Hecke orbits*, http://www.math.upenn.edu/~chai/.

8. ———, *Compactification of Siegel moduli schemes*, London Mathematical Society Lecture Note Series, vol. 107, Cambridge University Press, Cambridge, 1985. MR 853543 (88b:32074)
9. ———, *Every ordinary symplectic isogeny class in positive characteristic is dense in the moduli*, Invent. Math. **121** (1995), no. 3, 439–479. MR 1353306 (96f:11082)
10. ———, *Hecke orbits on Siegel modular varieties*, Geometric methods in algebra and number theory, Progr. Math., vol. 235, Birkhäuser Boston, Boston, MA, 2005, pp. 71–107. MR 2159378 (2006f:11050)
11. ———, *Monodromy of Hecke-invariant subvarieties*, Pure Appl. Math. Q. **1** (2005), no. 2, 291–303, Special issue: in memory of Armand Borel. MR 2194726 (2006m:11084)
12. ———, *A rigidity result for p-divisible formal groups*, Asian J. Math. **12** (2008), no. 2, 193–202.
13. C.-L. Chai and F. Oort, *Hecke orbits*, In preparation.
14. ———, *Monodromy and irreducibility of leaves*, Conference on abelian varieties, Amsterdam May 2006, http://www.math.upenn.edu/~chai/, http://www.math.uu.nl/people/oort/.
15. ———, *Hypersymmetric abelian varieties*, Pure Appl. Math. Q. **2** (2006), no. 1, 1–27. MR 2217565 (2007a:14052)
16. G. Cornell and J. H. Silverman (eds.), *Arithmetic geometry*, Springer-Verlag, New York, 1986, Papers from the conference held at the University of Connecticut, Storrs, Connecticut, July 30–August 10, 1984. MR 861969 (89b:14029)
17. A. J. de Jong, *Crystalline Dieudonné module theory via formal and rigid geometry*, Inst. Hautes Études Sci. Publ. Math. (1995), no. 82, 5–96 (1996). MR 1383213 (97f:14047)
18. ———, *Barsotti-Tate groups and crystals*, Proceedings of the International Congress of Mathematicians, Vol. II (Berlin, 1998), no. Extra Vol. II, 1998, pp. 259–265. MR 1648076 (99i:14051)
19. ———, *Homomorphisms of Barsotti-Tate groups and crystals in positive characteristic*, Invent. Math. **134** (1998), no. 2, 301–333, Erratum **138** (1999), 225. MR 1650324 (2000f:14070a)
20. A. J. de Jong and F. Oort, *Purity of the stratification by Newton polygons*, J. Amer. Math. Soc. **13** (2000), no. 1, 209–241. MR 1703336 (2000m:14050)
21. P. Deligne, *Hodge cycles on abelian varieties*, Hodge cycles, motives, and Shimura varieties, Lecture Notes in Mathematics, vol. 900, Springer-Verlag, Berlin, 1982, Notes by J.S. Milne, pp. 9–100. MR 654325 (84m:14046)
22. P. Deligne and G. Pappas, *Singularités des espaces de modules de Hilbert, en les caractéristiques divisant le discriminant*, Compositio Math. **90** (1994), no. 1, 59–79. MR 1266495 (95a:11041)
23. P. Deligne and K. A. Ribet, *Values of abelian L-functions at negative integers over totally real fields*, Invent. Math. **59** (1980), no. 3, 227–286. MR 579702 (81m:12019)
24. M. Demazure, *Lectures on p-divisible groups*, Lecture Notes in Mathematics, vol. 302, Springer-Verlag, Berlin, 1972. MR 0344261 (49 #9000)
25. M. Demazure and P. Gabriel, *Groupes algébriques. Tome I: Géométrie algébrique, généralités, groupes commutatifs*, Masson & Cie, Éditeur, Paris, 1970, Avec un appendice ıt Corps de classes local par Michiel Hazewinkel. MR 0302656 (46 #1800)
26. G. Faltings, *Endlichkeitssätze für abelsche Varietäten über Zahlkörpern*, Invent. Math. **73** (1983), no. 3, 349–366. MR 718935 (85g:11026a)
27. ———, *Arithmetische Kompaktifizierung des Modulraums der abelschen Varietäten*, Workshop Bonn 1984 (Bonn, 1984), Lecture Notes in Mathematics, vol. 1111, Springer, Berlin, 1985, pp. 321–383. MR 797429 (87c:14050)
28. G. Faltings and C.-L. Chai, *Degeneration of abelian varieties*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3), vol. 22, Springer-Verlag, Berlin, 1990, With an appendix by David Mumford. MR 1083353 (92d:14036)
29. G. Faltings and G. Wüstholz (eds.), *Rational points*, Aspects of Mathematics, E6, Friedr. Vieweg & Sohn, Braunschweig, 1984, Papers from the seminar held at the Max-Planck-Institut für Mathematik, Bonn, 1983/1984. MR 766568 (87h:14016)
30. L. Gerritzen, *On multiplication algebras of Riemann matrices*, Math Ann **194** (1971), 109–122. MR 0288141 (44 #5339)
31. E. Z. Goren and F. Oort, *Stratifications of Hilbert modular varieties*, J. Algebraic Geom. **9** (2000), no. 1, 111–154. MR 1713522 (2000g:14034)

32. A. Grothendieck, *Éléments de géométrie algébrique. III. Étude cohomologique des faisceaux cohérents. I*, Inst. Hautes Études Sci. Publ. Math. (1961), no. 11, 167. MR 0163910 (29 #1209)
33. ———, *Revêtements étales et groupe fondamental*, Lecture Notes in Mathematics, vol. 224, Springer-Verlag, Berlin, 1971, Séminaire de Géométrie Algébrique du Bois Marie 1960–1961 (SGA 1), Dirigé par A. Grothendieck. Augmenté de deux exposés de M. Raynaud. MR 0354651 (50 #7129)
34. ———, *Groupes de Barsotti-Tate et cristaux de Dieudonné*, Les Presses de l'Université de Montréal, Montreal, Que., 1974, Séminaire de Mathématiques Supérieures, No. 45 (Été, 1970). MR 0417192 (54 #5250)
35. A. Grothendieck and M. Demazure, *Schémas en groupes. I: Propriétés générales des schémas en groupes*, Lecture Notes in Mathematics, vol. 151, Springer-Verlag, Berlin, 1970, Séminaire de Géométrie Algébrique du Bois Marie 1962/64 (SGA 3). Dirigé par M. Demazure et A. Grothendieck. MR 0274458 (43 #223a)
36. M. Hazewinkel, *Formal groups and applications*, Pure and Applied Mathematics, vol. 78, Academic Press Inc. [Harcourt Brace Jovanovich Publishers], New York, 1978. MR 506881 (82a:14020)
37. T. Honda, *Isogeny classes of abelian varieties over finite fields*, J. Math. Soc. Japan **20** (1968), 83–95. MR 0229642 (37 #5216)
38. L. Illusie, *Déformations de groupes de Barsotti-Tate (d'après A. Grothendieck)*, Astérisque (1985), no. 127, 151–198, Séminaire sur les pinceaux arithmétiques: la conjecture de Mordell (L. Szpiro), Paris, 1983/84. MR 801922
39. T. Katsura and F. Oort, *Supersingular abelian varieties of dimension two or three and class numbers*, Algebraic geometry, Sendai, 1985, Adv. Stud. Pure Math., vol. 10, North-Holland, Amsterdam, 1987, pp. 253–281. MR 946242 (90c:14027)
40. 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, pp. 113–163. MR 563463 (81i:14014)
41. ———, *Appendix to Exposé V: Cristaux ordinaires et coordonnées canoniques*, Algebraic surfaces (Orsay, 1976–78), Lecture Notes in Mathematics, vol. 868, Springer, Berlin, 1981, Appendix to an article of P. Deligne, pp. 127–137. MR 638599 (83k:14039a)
42. ———, *Serre-Tate local moduli*, Algebraic surfaces (Orsay, 1976–78), Lecture Notes in Mathematics, vol. 868, Springer, Berlin, 1981, pp. 138–202. MR 638600 (83k:14039b)
43. H. Kraft, *Kommutative algebraische p-Gruppen (mit Anwendungen auf p-divisible Gruppen und abelsche Varietäten)*, Sonderforsch. Bereich Bonn, September 1975, 86 pp. manuscript.
44. M. Lazard, *Commutative formal groups*, Lecture Notes in Mathematics, vol. 443, Springer-Verlag, Berlin, 1975. MR 0393050 (52 #13861)
45. K.-Z. Li and F. Oort, *Moduli of supersingular abelian varieties*, Lecture Notes in Mathematics, vol. 1680, Springer-Verlag, Berlin, 1998. MR 1611305 (99e:14052)
46. J. Lubin, J-P. Serre, and J. Tate, *Elliptic curves and formal groups*, Lecture notes prepared in connection with the seminars held at the Summer Institute on Algebraic Geometry, Whitney Estate, Woods Hole, Massachusetts, July 6-July 31, 1964, http://www.ma.utexas.edu/users/voloch/lst.html.
47. J. Lubin and J. Tate, *Formal moduli for one-parameter formal Lie groups*, Bull. Soc. Math. France **94** (1966), 49–59. MR 0238854 (39 #214)
48. Yu. I. Manin, *Theory of commutative formal groups over fields of finite characteristic*, Uspehi Mat. Nauk **18** (1963), no. 6 (114), 3–90, Translated in Russ. Math. Surveys **18** (1963), 1–80. MR 0157972 (28 #1200)
49. W. Messing, *The crystals associated to Barsotti-Tate groups: with applications to abelian schemes*, Lecture Notes in Mathematics, vol. 264, Springer-Verlag, Berlin, 1972. MR 0347836 (50 #337)
50. S. Mochizuki, *The local pro-p anabelian geometry of curves*, Invent. Math. **138** (1999), no. 2, 319–423. MR 1720187 (2000j:14037)
51. L. Moret-Bailly, *Pinceaux de variétés abéliennes*, Astérisque (1985), no. 129, 266. MR 797982 (87j:14069)
52. S. Mori, *On Tate conjecture concerning endomorphisms of abelian varieties*, Proceedings of the International Symposium on Algebraic Geometry (Kyoto Univ., Kyoto, 1977) (Tokyo), Kinokuniya Book Store, 1978, pp. 219–230. MR 578861 (82d:14026)

53. D. Mumford, *A note of Shimura's paper "Discontinuous groups and abelian varieties"*, Math. Ann. **181** (1969), 345–351. MR 0248146 (40 #1400)
54. ———, *Abelian varieties*, Tata Institute of Fundamental Research Studies in Mathematics, No. 5, Published for the Tata Institute of Fundamental Research, Bombay, 1970. MR 0282985 (44 #219)
55. D. Mumford, J. Fogarty, and F. Kirwan, *Geometric invariant theory*, third ed., Ergebnisse der Mathematik und ihrer Grenzgebiete (2), vol. 34, Springer-Verlag, Berlin, 1994. MR 1304906 (95m:14012)
56. P. Norman and F. Oort, *Moduli of abelian varieties*, Ann. of Math. (2) **112** (1980), no. 3, 413–439. MR 595202 (82h:14026)
57. 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, pp. 595–621. MR 578876 (81f:14023)
58. F. Oort, *Foliations in moduli spaces of abelian varieties and dimension of leaves*, to appear.
59. ———, *Purity reconsidered*, Talk, Amsterdam XII-2002, ms 9 pp., http://www.math.uu.nl/people/oort/.
60. ———, *Algebraic group schemes in characteristic zero are reduced*, Invent. Math. **2** (1966), 79–80. MR 0206005 (34 #5830)
61. ———, *Commutative group schemes*, Lecture Notes in Mathematics, vol. 15, Springer-Verlag, Berlin, 1966. MR 0213365 (35 #4229)
62. ———, *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), 399–408. MR 0330175 (48 #8513)
63. ———, *Lifting algebraic curves, abelian varieties, and their endomorphisms to characteristic zero*, Algebraic geometry, Bowdoin, 1985, Part 2 (S. J. Bloch, ed.), Proc. Sympos. Pure Math., vol. 46, Amer. Math. Soc., Providence, RI, 1987, pp. 165–195. MR 927980 (89c:14069)
64. ———, *Endomorphism algebras of abelian varieties*, Algebraic geometry and commutative algebra, Vol. II, Kinokuniya, Tokyo, 1988, pp. 469–502. MR 977774 (90j:11049)
65. ———, *Newton polygons and formal groups: conjectures by Manin and Grothendieck*, Ann. of Math. (2) **152** (2000), no. 1, 183–206. MR 1792294 (2002e:14075)
66. ———, *Newton polygon strata in the moduli space of abelian varieties*, Moduli of abelian varieties (Texel Island, 1999) (F. Oort C. Faber, G. van der Geer, ed.), Progr. Math., vol. 195, Birkhäuser, Basel, 2001, pp. 417–440. MR 1827028 (2002c:14069)
67. ———, *A stratification of a moduli space of abelian varieties*, Moduli of abelian varieties (Texel Island, 1999) (F. Oort C. Faber, G. van der Geer, ed.), Progr. Math., vol. 195, Birkhäuser, Basel, 2001, pp. 345–416. MR 1827027 (2002b:14055)
68. ———, *Foliations in moduli spaces of abelian varieties*, J. Amer. Math. Soc. **17** (2004), no. 2, 267–296 (electronic). MR 2051612 (2005c:14051)
69. ———, *Minimal p-divisible groups*, Ann. of Math. (2) **161** (2005), no. 2, 1021–1036. MR 2153405 (2006i:14042)
70. V. Platonov and A. Rapinchuk, *Algebraic groups and number theory*, Pure and Applied Mathematics, vol. 139, Academic Press Inc., Boston, MA, 1994, Translated from the 1991 Russian original by Rachel Rowen. MR 1278263 (95b:11039)
71. M. Rapoport, *Compactifications de l'espace de modules de Hilbert-Blumenthal*, Compositio Math. **36** (1978), no. 3, 255–335. MR 515050 (80j:14009)
72. M. Rapoport and Th. Zink, *Period spaces for p-divisible groups*, Annals of Mathematics Studies, vol. 141, Princeton University Press, Princeton, NJ, 1996. MR 1393439 (97f:14023)
73. M. Raynaud, *"p-torsion" du schéma de Picard*, Journées de Géométrie Algébrique de Rennes (Rennes, 1978), Vol. II, Astérisque, vol. 64, Soc. Math. France, Paris, 1979, pp. 87–148. MR 563468 (81f:14026)
74. K. A. Ribet, *p-adic interpolation via Hilbert modular forms*, Algebraic geometry (Humboldt State Univ., Arcata, Calif., 1974), Proc. Sympos. Pure Math., vol. 29, Amer. Math. Soc., Providence, R. I., 1975, pp. 581–592. MR 0419414 (54 #7435)
75. L. Schneps and P. Lochak (eds.), *Geometric Galois actions. 1*, London Mathematical Society Lecture Note Series, vol. 242, Cambridge University Press, Cambridge, 1997, Around Grothendieck's "Esquisse d'un programme". MR 1483106 (98e:14003)
76. G. Shimura, *On analytic families of polarized abelian varieties and automorphic functions*, Ann. of Math. (2) **78** (1963), 149–192. MR 0156001 (27 #5934)

77. 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. MR 0125113 (23 #A2419)
78. J. Tate, *Endomorphisms of abelian varieties over finite fields*, Invent. Math. **2** (1966), 134–144. MR 0206004 (34 #5829)
79. ———, *Classes d'isogénie de variétés abéliennes sur un corps fini (d'après T. Honda)*, Séminaire Bourbaki. Vol. 1968/69: Exposés 347–363, Lecture Notes in Mathematics, vol. 179, Springer-Verlag, Berlin, 1971, Exp. 352, pp. 95–110. MR 0272579 (42 #7460)
80. G. van der Geer, *Hilbert modular surfaces*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3), vol. 16, Springer-Verlag, Berlin, 1988. MR 930101 (89c:11073)
81. G. van der Geer and B. Moonen, *Abelian varieties*, In preparation, http://staff.science.uva.nl/~bmoonen/boek/BookAV.html.
82. A. Vasiu, *Crystalline boundedness principle*, Ann. Sci. École Norm. Sup. (4) **39** (2006), no. 2, 245–300. MR 2245533 (2007d:14081)
83. ———, *Reconstructing p-divisible groups from their truncation of smaller level*, 2006, arXiv:math/0607268.
84. W. C. Waterhouse, *Introduction to affine group schemes*, Graduate Texts in Mathematics, vol. 66, Springer-Verlag, New York, 1979. MR 547117 (82e:14003)
85. W. C. Waterhouse and J. S. Milne, *Abelian varieties over finite fields*, 1969 Number Theory Institute, State Univ. New York, Stony Brook, N.Y., 1969), Proc. Sympos. Pure Math., vol. 20, Amer. Math. Soc., Providence, R.I., 1971, pp. 53–64. MR 0314847 (47 #3397)
86. A. Weil, *Sur les courbes algébriques et les variétés qui s'en déduisent*, Actualités Sci. Ind., no. 1041 = Publ. Inst. Math. Univ. Strasbourg **7** (1945), Hermann et Cie., Paris, 1948. MR 0027151 (10,262c)
87. ———, *Variétés abéliennes et courbes algébriques*, Actualités Sci. Ind., no. 1064 = Publ. Inst. Math. Univ. Strasbourg 8 (1946), Hermann & Cie., Paris, 1948. MR 0029522 (10,621d)
88. C.-F. Yu, *On reduction of Hilbert-Blumenthal varieties*, Ann. Inst. Fourier (Grenoble) **53** (2003), no. 7, 2105–2154. MR 2044169 (2005h:14062)
89. ———, *On the supersingular locus in Hilbert-Blumenthal 4-folds*, J. Algebraic Geom. **12** (2003), no. 4, 653–698. MR 1993760 (2004g:14044)
90. Ju. G. Zarhin, *Isogenies of abelian varieties over fields of finite characteristic*, Mat. Sb. (N.S.) **95(137)** (1974), 461–470, 472, English translation: *Math. USSR Sbornik* **24** (1974), 451–461. MR 0354685 (50 #7162b)
91. ———, *A remark on endomorphisms of abelian varieties over function fields of finite characteristic*, Izv. Akad. Nauk SSSR Ser. Mat. **38** (1974), 471–474, English translation: *Math. USSR-Izv.* **8** (1974), 477–480.
92. T. Zink, *de Jong-Oort purity for p-divisible groups*, preprint, http://www.mathematik.uni-bielefeld.de/~zink/z_publ.html.
93. ———, *Cartiertheorie kommutativer formaler Gruppen*, Teubner-Texte zur Mathematik [Teubner Texts in Mathematics], vol. 68, BSB B. G. Teubner Verlagsgesellschaft, Leipzig, 1984, With English, French and Russian summaries. MR 767090 (86j:14046)
94. ———, *On the slope filtration*, Duke Math. J. **109** (2001), no. 1, 79–95. MR 1844205 (2003d:14055)

Department of Mathematics, University of Pennsylvania, Philadelphia, PA 19104-6395, U.S.A.
E-mail address: chai@math.upenn.edu
URL: http://www.math.upenn.edu/~chai/

Mathematisch Instituut, Budapestlaan 6, NL - 3584 CD TA Utrecht, The Netherlands
Current address: Postbus 80010,NL - 3508 TA Utrecht
E-mail address: F.Oort@uu.nl
URL: http://www.math.uu.nl/people/oort/

Clay Mathematics Proceedings
Volume 8, 2009

Cartier isomorphism and Hodge Theory in the non-commutative case

Dmitry Kaledin

Abstract. These lectures attempt to give an elementary introduction to my recent paper "Non-commutative Hodge-to-de Rham degeneration via the method of Deligne-Illusie".

Contents

1. Introduction

One of the standard ways to compute the cohomology groups of a smooth complex manifold X is by means of the de Rham theory: the de Rham cohomology groups

$$H^{\bullet}_{DR}(X) = \mathbb{H}^{\bullet}(X, \Omega^{\bullet}_{DR}) \tag{1.1}$$

are by definition the hypercohomology groups of X with coefficients in the (holomorphic) de Rham complex $\Omega^{\bullet}_{DR}$, and since, by the Poincaré Lemma, $\Omega^{\bullet}_{DR}$ is a resolution of the constant sheaf $\mathbb{C}$, we have $H^{\bullet}_{DR}(X) \cong H^{\bullet}(X, \mathbb{C})$. If X is in fact algebraic, then $\Omega^{\bullet}_{DR}$ can also be defined algebraically, so that the right-hand side in (1.1) can be understood in two ways: either as the hypercohomology of an analytic space, or as the hypercohomology of a scheme equipped with the Zariski topology. One can show that the resulting groups $H^{\bullet}_{DR}(X)$ are the same (for compact X, this is just the GAGA principle; in the non-compact case this is a difficult but true fact established by Grothendieck [**Gro66**]).

2000 *Mathematics Subject Classification.* Primary 19D55, Secondary 18G60, 14F40.
Partially supported by CRDF grant RUM1-2694.

Of course, an algebraic version of the Poincaré Lemma is false, since the Zariski topology is not fine enough—no matter how small a Zariski neighborhood of a point one takes, it usually has non-trivial higher de Rham cohomology. However, the Lemma survives on the formal level: the completion $\widehat{\Omega^{\bullet}_{DR}}$ of the de Rham complex near a closed point $x \in X$ is quasi-isomorphic to $\mathbb{C}$ placed in degree 0.

Assume now that our X is a smooth algebraic variety over a perfect field k of characteristic $p > 0$. Does the de Rham cohomology still make sense?

The de Rham complex $\Omega^{\bullet}_{DR}$ itself is well-defined: Ω^1 is just the sheaf of Kähler differentials, which makes sense in any characteristic and comes equipped with the universal derivation $d : \mathcal{O}_X \to \Omega^1$, and $\Omega^{\bullet}_{DR}$ is its exterior algebra, which is also well-defined in characteristic p. However, the Poincaré Lemma breaks down completely—the homology of the de Rham complex remains large even after taking completion at a closed point.

In degree 0, this is actually very easy to see: for any local function f on X, we have $df^p = pf^{p-1}df = 0$, so that all the p-th powers of functions are closed with respect to the de Rham differential. Since we are in characteristic p, these powers form a subsheaf of algebras in $\mathcal{O}_X$ which we denote by $\mathcal{O}^p_X \subset \mathcal{O}_X$. This is a large subsheaf. In fact, if we denote by $X^{(1)}$ the scheme X with $\mathcal{O}^p_X$ as the structure sheaf, then $X \cong X^{(1)}$ as abstract schemes, with the isomorphism given by the Frobenius map $f \mapsto f^p$. Fifty years ago P. Cartier proved that in fact *all* the functions in $\mathcal{O}_X$ closed with respect to the de Rham differential are contained in $\mathcal{O}^p_X$, and moreover, one has a similar description in higher degrees: there exist natural isomorphisms

$$C : \mathcal{H}^{\bullet}_{DR} \cong \Omega^{\bullet}_{X^{(1)}}, \tag{1.2}$$

where on the left we have the homology sheaves of the de Rham complex, and on the right we have the sheaves of differential forms on the scheme $X^{(1)}$. These isomorphisms are known as *Cartier isomorphisms*.

The Cartier isomorphism has many applications, but one of the most unexpected was discovered in 1987 by P. Deligne and L. Illusie: one can use the Cartier isomorphism to give a purely algebraic proof of the following purely algebraic statement, which is normally proved by the highly transcendental Hodge Theory.

THEOREM 1.1 ([**DI87**]). *Assume given a smooth proper variety X over a field K of characteristic 0. Then the Hodge-to-de Rham spectral sequence*

$$H^{\bullet}(X, \Omega^{\bullet}) \Rightarrow H^{\bullet}_{DR}(X)$$

associated to the stupid filtration on the de Rham complex $\Omega^{\bullet}$ degenerates at the first term.

The proof of Deligne and Illusie was very strange, because it worked by reduction to positive characteristic, where the statement is not true for a general X. What they proved is that if one imposes two additional conditions on X, then the Cartier isomorphisms can be combined together into a quasi-isomorphism

$$\Omega^{\bullet}_{DR} \cong \bigoplus_i \mathcal{H}^i_{DR}[-i] \cong \bigoplus_i \Omega^i_{X^{(1)}}[-i] \tag{1.3}$$

in the derived category of coherent sheaves on $X^{(1)}$. The degeneration follows from this immediately for dimension reasons. The additional conditions are:

(i) X can be lifted to a smooth scheme over $W_2(k)$, the ring of second Witt vectors of the perfect field k (e.g. if $k = \mathbb{Z}/p\mathbb{Z}$, X has to be liftable to $\mathbb{Z}/p^2\mathbb{Z}$), and
(ii) we have $p > \dim X$.

To deduce Theorem 1.1, one finds by the standard argument a proper smooth model X_R of X defined over a finitely generated subring $R \subset K$, one localizes R so that it is unramified over $\mathbb{Z}$ and all its residue fields have characteristic greater than $\dim X$, and one deduces that all the special fibers of X_R/R satisfy the assumptions above; hence the differentials in the Hodge-to-de Rham spectral sequence vanish at all closed points of $\operatorname{Spec} R$, which means they are identically 0 by Nakayama.

The goal of these lectures is to present in a down-to-earth way the results of two recent papers [**Kal05**], [**Kal08**], where the story summarized above has been largely transferred to the setting of *non-commutative geometry*.

To explain what I mean by this, let us first recall that a non-commutative version of differential forms has been known for quite some time now. Namely, assume given an associative unital algebra A over a field k, and an A-bimodule M. Then its *Hochschild homology* $HH_\bullet(A, M)$ of A with coefficients in M is defined as

$$HH_\bullet(A) = \operatorname{Tor}^{\bullet}_{A^{opp}\otimes A}(A, M), \tag{1.4}$$

where $A^{opp} \otimes A$ is the tensor product of A and the opposite algebra A^{opp}, and the A-bimodule M is treated as a left module over $A^{opp} \otimes A$. Hochschild homology $HH_\bullet(A)$ is the Hochschild homology of A with coefficients in itself.

Assume for a moment that A is in fact commutative, and $\operatorname{Spec} A$ is a smooth algebraic variety over K. Then it has been proved back in 1962 in the paper [**HKR62**] that we have canonical isomorphisms $HH_i(A) \cong \Omega^i(A/k)$ for any $i \geq 0$. Thus for a general A, one can treat Hochschild homology classes as a replacement for differential forms.

Moreover, in the early 1980s it was discovered by A. Connes [**Con83**], J.-L. Loday and D. Quillen [**LQ83**], and B. Feigin and B. Tsygan [**FT83**], that the de Rham differential also makes sense in the general non-commutative setting. Namely, these authors introduced a new invariant of an associative algebra A called *cyclic homology*; cyclic homology, denoted $HC_\bullet(A)$, is related to the Hochschild homology $HH_\bullet(A)$ by a spectral sequence

$$HH_\bullet(A)[u^{-1}] \Rightarrow HC_\bullet(A), \tag{1.5}$$

which in the smooth commutative case reduces to the Hodge-to-de Rham spectral sequence (here u is a formal parameter of cohomological degree 2, and $HH_\bullet(A)[u^{-1}]$ is shorthand for "polynomials in u^{-1} with coefficients in $HH_\bullet(A)$").

It has been conjectured for some time now that the spectral sequence (1.5), or a version of it, degenerates under appropriate assumptions on A (which imitate the assumptions of Theorem 1.1). Following [**Kal08**], we will attack this conjecture by the method of Deligne and Illusie. To do this, we will introduce a certain non-commutative version of the Cartier isomorphism, or rather, of the "globalized" isomorphism (1.3) (in the process of doing it, we will need to introduce some conditions on A which precisely generalize the conditions (i), (ii) above). Then we prove a version of the degeneration conjecture as stated by M. Kontsevich and Ya.

Soibelman in [**KS06**] (we will have to impose an additional technical assumption which, fortunately, is not very drastic).

The paper is organized as follows. In Section 2 we recall the definition of the cyclic homology and some versions of it needed for the Cartier isomorphism (most of this material is quite standard; the reader can find good expositions in [**Lod98**] or [**FT87**]). One technical result needed in the main part of the paper has been separated into Section 3. In Section 4, we construct the Cartier isomorphism for an algebra A equipped with some additional piece of data which we call the *quasi-Frobenius map*. It exists only for special classes of algebras—e.g. for free algebras, or for the group algebra $k[G]$ of a finite group G—but the construction illustrates nicely the general idea. In Section 5, we show what to do in the general case. Here the conditions (i), (ii) emerge, and in a somewhat surprising way—as it turns out, they essentially come from algebraic topology, and the whole theory has a distinctly topological flavor. Finally, in Section 6 we show how to apply our generalized Cartier isomorphism to the Hodge-to-de Rham degeneration. The exposition in Sections 2-4 is largely self-contained. In the rest of the paper, we switch to a more descriptive style, with no proofs, and not many precise statements; this part of the paper should be treated as a companion to [**Kal08**].

Acknowledgments. This paper is a write-up (actually quite an enlarged write-up) of two lectures given in Göttingen in August 2006, at a summer school organized by Yu. Tschinkel and funded by the Clay Institute. I am very grateful to all concerned for making it happen, and for giving me an opportunity to present my results. In addition, I would like to mention that a large part of the present paper is written in overview style; many, if not most, of the things overviewed are certainly *not* my results. This especially concerns Section 2, on the one hand, and Section 6, on the other hand. Given the chosen style, it is difficult to provide exact attributions; however, I should at least mention that I've learned much of this material from A. Beilinson, A. Bondal, M. Kontsevich, B. Toën and B. Tsygan.

2. Cyclic homology package

2.1. Basic definitions. The fastest and most down-to-earth way to define cyclic homology is by means of an explicit complex. Namely, assume given an associative unital algebra A over a field k. To compute its Hochschild homology with coefficients in some bimodule M, one has to find a flat resolution of M. One such is the *bar resolution*—it is rather inconvenient in practical computations, but it is completely canonical, and it exists without any assumptions on A and M. The terms of this resolution are of the form $A^{\otimes n} \otimes M$, $n \geq 0$, and the differential $b' : A^{\otimes n+1} \otimes M \to A^{\otimes n} \otimes M$ is given by

$$b' = \sum_{0 \leq i \leq n} (-1)^i \, \mathsf{id}^{\otimes i} \otimes m \otimes \mathsf{id}^{\otimes n-i}, \tag{2.1}$$

where $m : A \otimes A \to A$, $m : A \otimes M \to M$ are the multiplication maps. Substituting this resolution into (1.4) gives a complex which computes $HH_\bullet(A, M)$; its terms are also $A^{\otimes i} \otimes M$, but the differential is given by

$$b = b' + (-1)^{n+1} t, \tag{2.2}$$

with the correction term t being equal to

$$t(a_0 \otimes \cdots \otimes a_{n+1} \otimes m) = a_1 \otimes \cdots \otimes a_{n+1} \otimes m a_0$$

for any $a_0, \dots, a_{n+1} \in A$, $m \in M$. Geometrically, one can think of the components $a_0, \dots, a_{n-1}, m$ of some tensor in $A^{\otimes n} \otimes M$ as having been placed at $n+1$ points on the unit interval $[0,1]$, including the endpoints $0, 1 \in [0,1]$; then each of the terms in the differential b' corresponds to contracting an interval between two neighboring points and multiplying the components sitting at its endpoints. To visualize the differential b in a similar way, one has to take $n+1$ points placed on the unit circle S^1 instead of the unit interval, including the point $1 \in S^1$, where we put the component m.

In the case $M = A$, the terms in the bar complex are just $A^{\otimes n+1}$, $n \geq 0$, and they acquire an additional symmetry: we let $\tau : A^{\otimes n+1} \to A^{\otimes n+1}$ be the cyclic permutation multiplied by $(-1)^n$. Note that in spite of the sign change, we have $\tau^{n+1} = \mathsf{id}$, so that it generates an action of the cyclic group $\mathbb{Z}/(n+1)\mathbb{Z}$ on every $A^{\otimes n+1}$. The fundamental fact here is the following.

LEMMA 2.1 ([**FT87**],[**Lod98**]). *For any n, we have*

$$(\mathsf{id} - \tau) \circ b' = -b \circ (\mathsf{id} - \tau),$$
$$(\mathsf{id} + \tau + \dots + \tau^{n-1}) \circ b = -b' \circ (\mathsf{id} + \tau + \dots + \tau^n)$$

as maps from $A^{\otimes n+1}$ to $A^{\otimes n}$.

Proof. Denote $m_i = \mathsf{id}^i \otimes m \otimes \mathsf{id}^{n-i} : A^{\otimes n+1} \to A^{\otimes n}$, $0 \leq i \leq n-1$, so that $b' = m_0 - m_1 + \dots + (-1)^{n-1} m_{n-1}$, and let $m_n = t = (-1)^n (b - b')$. Then we obviously have

$$m_{i+1} \circ \tau = \tau \circ m_i$$

for $0 \leq i \leq n-1$, and $m_0 \circ \tau = (-1)^n m_n$. Formally applying these identities, we conclude that

$$\begin{aligned} \sum_{0 \leq i \leq n} (-1)^i m_i \circ (\mathsf{id} - \tau) &= \sum_{0 \leq i \leq n} (-1)^i m_i - m_0 - \sum_{1 \leq i \leq n} (-1)^i \tau \circ m_{i-1} \\ &= -(\mathsf{id} - \tau) \circ \sum_{0 \leq i \leq n-1} (-1)^i m_i, \end{aligned} \tag{2.3}$$

$$\begin{aligned} b' \circ (\mathsf{id} + \tau + \dots + \tau^n) &= \sum_{0 \leq i \leq n-1} \sum_{0 \leq j \leq n} (-1)^i m_i \circ \tau^j \\ &= \sum_{0 \leq j \leq i \leq n-1} (-1)^i \tau^j \circ m_{i-j} + \sum_{1 \leq i \leq j \leq n} (-1)^{i+n} \tau^{j-1} \circ m_{n+i-j} \\ &= -(\mathsf{id} + \tau + \dots + \tau^{n-1}) \circ b, \end{aligned} \tag{2.4}$$

which proves the claim. □

As a corollary, the following diagram is in fact a bicomplex.

$$
\begin{array}{ccccccc}
\cdots & \longrightarrow & A & \xrightarrow{\mathsf{id}} & A & \xrightarrow{0} & A \\
 & & \uparrow{\scriptstyle b} & & \uparrow{\scriptstyle b'} & & \uparrow{\scriptstyle b} \\
\cdots & \longrightarrow & A\otimes A & \xrightarrow{\mathsf{id}+\tau} & A\otimes A & \xrightarrow{\mathsf{id}-\tau} & A\otimes A \\
 & & \uparrow{\scriptstyle b} & & \uparrow{\scriptstyle b'} & & \uparrow{\scriptstyle b} \\
\vdots & & \vdots & & \vdots & & \vdots \\
 & & \uparrow{\scriptstyle b} & & \uparrow{\scriptstyle b'} & & \uparrow{\scriptstyle b} \\
\cdots & \longrightarrow & A^{\otimes n} & \xrightarrow{\mathsf{id}+\tau+\cdots+\tau^{n-1}} & A^{\otimes n} & \xrightarrow{\mathsf{id}-\tau} & A^{\otimes n} \\
 & & \uparrow{\scriptstyle b} & & \uparrow{\scriptstyle b'} & & \uparrow{\scriptstyle b} \\
 & & \vdots & & \vdots & & \vdots
\end{array}
\tag{2.5}
$$

Here it is understood that the whole thing extends indefinitely to the left, all the even-numbered columns are the same, all odd-numbered columns are the same, and the bicomplex is invariant with respect to the horizontal shift by 2 columns. The total homology of this bicomplex is called the *cyclic homology* of the algebra A, and denoted by $HC_\bullet(A)$.

We see right away that the first, the third, and so on column when counting from the right is the bar complex which computes $HH_\bullet(A)$, and the second, the fourth, and so on column is acyclic (the top term is A, and the rest is the bar resolution for A). Thus the spectral sequence for this bicomplex has the form given in (1.5) (modulo obvious renumbering). On the other hand, the rows of the bicomplex are just the standard 2-periodic complexes which compute the cyclic group homology $H_\bullet(\mathbb{Z}/n\mathbb{Z}, A^{\otimes n})$ (with respect to the $\mathbb{Z}/n\mathbb{Z}$-action on $A^{\otimes n}$ given by τ).

Shifting (2.5) to the right by 2 columns gives the *periodicity map*

$$u : HC_{\bullet+2}(A) \to HC_\bullet(A),$$

which fits into an exact triangle

$$HH_{\bullet+2} \longrightarrow HC_{\bullet+2}(A) \longrightarrow HC_\bullet(A) \longrightarrow , \tag{2.6}$$

known as the *Connes exact sequence*. One can also invert the periodicity map—in other words, extend the bicomplex (2.5) not only to the left, but also to the right. This gives the *periodic cyclic homology* $HP_\bullet(A)$. Since the bicomplex for $HP_\bullet(A)$ is infinite in both directions, there is a choice involved in taking the total complex: we can take either the product, or the sum of the terms. We take the product. In characteristic 0, the sum is actually acyclic (because so is every row).

If A is commutative, $X = \mathrm{Spec}(A)$ is smooth, and $\mathsf{char}\, k$ is either 0 or greater than $\dim X$, then the only non-trivial differential in the Hodge-to-de Rham spectral sequence (1.5) is the first one, and it is the de Rham differential. Consequently, we have $HP_\bullet(A) = H^\bullet_{DR}(X)((u))$ (where as before, u is a formal variable of cohomological degree 2).

2.2. The p-cyclic complex. All of the above is completely standard; however, we will also need to use another way to compute $HC_\bullet(A)$, which is less

standard. Namely, fix an integer $p \geq 2$, and consider the algebra $A^{\otimes p}$. Let $\sigma : A^{\otimes p} \to A^{\otimes p}$ be the cyclic permutation, and let $A^{\otimes p}_{\sigma}$ be the diagonal $A^{\otimes p}$-bimodule with the bimodule structure twisted by σ—namely, we let

$$a \cdot b \cdot c = ab\sigma(c)$$

for any $a, b, c \in A^{\otimes p}$.

LEMMA 2.2. *We have* $HH_\bullet(A^{\otimes p}, A^{\otimes p}_{\sigma}) \cong HH_\bullet(A)$.

Proof. Induction on p. We may compute the tensor product in (1.4) over each of the factors A in $A^{\otimes p}$ in turn; this shows that

$$HH_\bullet(A^{\otimes p}, A^{\otimes p}_{\sigma}) \cong \mathrm{Tor}^{\bullet}_{(A^{\otimes(p-1)})^{opp} \otimes A^{\otimes(p-1)}} \left(A^{\otimes(p-1)}, \mathrm{Tor}^{\bullet}_{A^{opp}\otimes A}(A, A^{\otimes p}_{\sigma}) \right),$$

and one checks easily that as long as $p \geq 2$, so that $A^{\otimes p}_{\sigma}$ is flat over $A^{opp} \otimes A$, $\mathrm{Tor}^{i}_{A^{opp}\otimes A}(A, A^{\otimes p}_{\sigma})$ is naturally isomorphic to $A^{\otimes(p-1)}_{\sigma}$ if $i = 0$, and trivial if $i \geq 1$. □

By virtue of this Lemma, we can use the bar complex for the algebra $A^{\otimes p}$ to compute $HH_\bullet(A)$. The resulting complex has terms $A^{\otimes pn}$, $n \geq 0$. The differential $b'_p : A^{\otimes p(n+2)} \to A^{\otimes p(n+1)}$ is given by essentially the same formula as (2.1):

$$b'_p = \sum_{0 \leq i \leq n} (-1)^i m^p_i = \sum_{0 \leq i \leq n} (-1)^i \, \mathsf{id}^{\otimes pi} \otimes m^{\otimes p} \otimes \mathsf{id}^{\otimes p(n-i)},$$

where we decompose $A^{\otimes p(n+1)} = (A^{\otimes p})^{\otimes(n+1)}$. The correction term $t_p = m^p_{n+1}$ in (2.2) is given by $m_0 \circ \tau$ (where, as before, $\tau : A^{\otimes p(n+2)}$ is the cyclic permutation of order $p(n+2)$ twisted by a sign). Geometrically, the component m^p_i of the Hochschild differential b_p corresponds to contracting simultaneously the i-th, $(i+p)$-th, $(i+2p)$-th, and so on, intervals in the unit circle, divided into $p(n+2)$ intervals by $p(n+2)$ points. On the level of bar complexes, the comparison isomorphism $HH_\bullet(A^{\otimes p}, A^{\otimes p}_{\sigma}) \cong HH_\bullet(A)$ of Lemma 2.2 is represented by the map

$$M = m \circ (id \otimes m) \circ (\mathsf{id}^{\otimes 2} \otimes m) \circ \cdots \circ (\mathsf{id}^{\otimes pn-2} \otimes m) : A^{\otimes pn} \to A^{\otimes n}; \tag{2.7}$$

explicitly, we have

$$M(a_{1,1} \otimes a_{2,1} \otimes \cdots \otimes a_{n,1} \otimes a_{1,2} \otimes a_{2,2} \otimes \cdots \otimes a_{n,2} \otimes \cdots \otimes a_{1,p} \otimes a_{2,p} \otimes \cdots \otimes a_{n,p})$$
$$= a_{1,1} \otimes a_{2,1} \otimes \cdots \otimes a_{n-1,1} \otimes \left(a_{n,1} \cdot \prod_{2 \leq j \leq p} \prod_{1 \leq i \leq n} a_{i,j} \right)$$

for any $a_{1,1} \otimes a_{2,1} \otimes \cdots \otimes a_{n,1} \otimes a_{1,2} \otimes a_{2,2} \otimes \cdots \otimes a_{n,2} \otimes \cdots \otimes a_{1,p} \otimes a_{2,p} \otimes \cdots \otimes a_{n,p} \in A^{\otimes pn}$—in other words, $M : A^{\otimes pn} \to A^{\otimes n}$ leaves the first $n-1$ terms in the tensor product intact and multiplies the remaining $pn - n + 1$ terms. We leave it to the interested reader to check explicitly that $M \circ b_p = b \circ M$.

LEMMA 2.3. *For any n, we have*

$$(\mathsf{id} - \tau) \circ b'_p = -b_p \circ (\mathsf{id} - \tau),$$
$$(\mathsf{id} + \tau + \cdots + \tau^{pn-1}) \circ b_p = -b'_p \circ (\mathsf{id} + \tau + \cdots + \tau^{p(n+1)-1})$$

as maps from $A^{\otimes p(n+1)}$ to $A^{\otimes pn}$.

Proof. One immediately checks that, as in the proof of Lemma 2.1, we have

$$m^p_{i+1} \circ \tau = -\tau \circ m^p_i$$

for $0 \leq i \leq n$, and we also have $m^p_0 \circ \tau = m^p_{n+1}$. Then the first equality follows from (2.3), and (2.4) gives

$$(\mathsf{id} + \tau + \cdots + \tau^{n-1}) \circ b_p = -b'_p \circ (\mathsf{id} + \tau + \cdots + \tau^n)$$

(note that the proof of these two equalities does *not* use the fact that $\tau^{n+1} = \mathsf{id}$ on $A^{\otimes(n+1)}$). To deduce the second equality of the Lemma, it suffices to notice that

$$\mathsf{id} + \tau + \cdots + \tau^{p(n+1)-1} = (\mathsf{id} + \tau + \cdots + \tau^n) \circ (\mathsf{id} + \sigma + \cdots + \sigma^{p-1}),$$

and σ commutes with all the maps m^p_i. □

Using Lemma 2.3, we can construct a version of the bicomplex (2.5) for $p > 1$:

$$\begin{array}{ccccccc}
\cdots & \longrightarrow & A^{\otimes p} & \xrightarrow{\mathsf{id}+\tau+\cdots+\tau^{p-1}} & A^{\otimes p} & \xrightarrow{\mathsf{id}-\tau} & A^{\otimes p} \\
 & & \uparrow{\scriptstyle b_p} & & \uparrow{\scriptstyle b'_p} & & \uparrow{\scriptstyle b_p} \\
\cdots & \longrightarrow & A^{\otimes 2p} & \xrightarrow{\mathsf{id}+\cdots+\tau^{2p-1}} & A^{\otimes 2p} & \xrightarrow{\mathsf{id}-\tau} & A^{\otimes 2p} \\
 & & \uparrow{\scriptstyle b_p} & & \uparrow{\scriptstyle b'_p} & & \uparrow{\scriptstyle b_p} \\
\vdots & & \vdots & & \vdots & & \vdots \\
 & & \uparrow{\scriptstyle b_p} & & \uparrow{\scriptstyle b'_p} & & \uparrow{\scriptstyle b_p} \\
\cdots & \longrightarrow & A^{\otimes pn} & \xrightarrow{\mathsf{id}+\tau+\cdots+\tau^{pn-1}} & A^{\otimes pn} & \xrightarrow{\mathsf{id}-\tau} & A^{\otimes pn} \\
 & & \uparrow{\scriptstyle b_p} & & \uparrow{\scriptstyle b'_p} & & \uparrow{\scriptstyle b_p} \\
 & & \vdots & & \vdots & & \vdots
\end{array} \tag{2.8}$$

By abuse of notation, we denote the homology of the total complex of this bicomplex by $HC_\bullet(A^{\otimes p}, A^{\otimes p}_\sigma)$. (This is really abusive, since in general one *cannot* define cyclic homology with coefficients in a bimodule—unless the bimodule is equipped with additional structure, such as e.g. in [**Kal07**], which lies beyond the scope of this paper.) As for the usual cyclic complex, we have the periodicity map, the Connes exact sequence, and we can form the periodic cyclic homology $HP_\bullet(A^{\otimes p}, A^{\otimes p}_\sigma)$.

2.3. Small categories. Unfortunately, this is as far as the down-to-earth approach takes us. While it is true that the isomorphism $HH_\bullet(A^{\otimes p}, A^{\otimes p}_\sigma) \cong HH_\bullet(A)$ given in Lemma 2.2 can be extended to an isomorphism

$$HC_\bullet(A^{\otimes p}, A^{\otimes p}_\sigma) \cong HC_\bullet(A),$$

it is not possible to realize this extended isomorphism by an explicit map of bicomplexes. Indeed, already in degree 0 the comparison map M of (2.7) which realized the isomorphism

$$HH_0(A^{\otimes p}, A^{\otimes p}_\sigma) \to HH_0(A)$$

on the level of bar complexes is given by the multiplication map $A^{\otimes p} \to A$, and to define this multiplication map, one has to break the cyclic symmetry of the product $A^{\otimes p}$. The best one can obtain is a map between total complexes computing $HC_\bullet(A^{\otimes p}, A^{\otimes p}_\sigma)$ and $HC_\bullet(A)$ which preserves the filtration, but not the second

grading; when one tries to write the map down explicitly, the combinatorics quickly gets completely out of control.

For this reason, in [**Kal05**] and [**Kal08**] one follows [**Con83**] and uses a more advanced approach to cyclic homology which is based on the technique of *homology of small categories* (see e.g. [**Lod98**, Section 6]). Namely, for any small category Γ and any base field k, the category $\mathrm{Fun}(\Gamma, k)$ of functors from Γ to k-vector spaces is an abelian category, and the direct limit functor $\varinjlim_{\Gamma}$ is right exact. Its derived functors are called *homology functors* of the category Γ and denoted by $H_\bullet(\Gamma, E)$ for any $E \in \mathrm{Fun}(\Gamma, k)$. For instance, if Γ is a groupoid with one object with automorphism group G, then $\mathrm{Fun}(\Gamma, k)$ is the category of k-representations of the group G; the homology $H_\bullet(\Gamma, -)$ is then tautologically the same as the group homology $H_\bullet(G, -)$. Another example is the category Δ^{opp}, the opposite to the category Δ of finite non-empty totally ordered sets. It is not difficult to check that for any simplicial k-vector $E \in \mathrm{Fun}(\Delta^{opp}, k)$, the homology $H_\bullet(\Delta^{opp}, E)$ can be computed by the standard chain complex of E.

For applications to cyclic homology, one introduces special small categories Λ_∞ and Λ_p, $p \geq 1$. The objects in the category Λ_∞ are numbered by the positive integers and denoted $[n]$, $n \geq 1$. For any $[n], [m] \in \Lambda_\infty$, the set of maps $\Lambda_\infty([n], [m])$ is the set of all maps $f : \mathbb{Z} \to \mathbb{Z}$ such that

$$f(a) \leq f(b) \quad \text{whenever } a \leq b, \qquad f(a+n) = f(a) + m, \tag{2.9}$$

for any $a, b \in \mathbb{Z}$. For any $[n] \in \Lambda_\infty$, denote by $\sigma : [n] \to [n]$ the endomorphism given by $f(a) = a + n$. Then σ commutes with all maps in Λ_∞. The category Λ_p has the same objects as Λ_∞, and the set of maps is

$$\Lambda_p([n], [m]) = \Lambda_\infty([n], [m])/\sigma^p$$

for any $[n], [m] \in \Lambda_p$. The category Λ_1 is denoted simply by Λ; this is the original cyclic category introduced by A. Connes in [**Con83**]. By definition, we have projections $\Lambda_\infty \to \Lambda_p$ and $\pi : \Lambda_p \to \Lambda$.

If we only consider those maps in (2.9) which send $0 \in \mathbb{Z}$ to 0, then the resulting subcategory in Λ_∞ is equivalent to Δ^{opp}. This gives a canonical embedding $j : \Delta^{opp} \to \Lambda_\infty$, and consequently, embeddings $j : \Delta^{opp} \to \Lambda_p$.

The category Λ_p conveniently encodes the maps m_i^p and τ between various tensor powers $A^{\otimes pn}$ used in the complex (2.8): m_i^p corresponds to the map $f \in \Lambda_p([n+1], [n])$ given by

$$f(a(n+1)+b) = \begin{cases} an + b, & b \leq i, \\ an + b - 1, & b > i, \end{cases}$$

where $0 \leq b \leq n$, and τ is the map $a \mapsto a + 1$, twisted by the sign (alternatively, one can say that m_i^p are obtained from face maps in Δ^{opp} under the embedding $\Delta^{opp} \subset \Lambda_p$). The relations between these maps which we used in the proof of Lemma 2.3 are encoded in the composition laws of the category Λ_p. Thus for any object $E \in \mathrm{Fun}(\Lambda_p, k)$—they are called *p-cyclic objects*—one can form the

bicomplex of the type (2.8) (or (2.5), for $p = 1$):

$$\begin{array}{ccccc}
\cdots & \longrightarrow & E([1]) & \xrightarrow{\mathsf{id}+\tau+\cdots+\tau^{p-1}} & E([1]) & \xrightarrow{\mathsf{id}-\tau} & E([1]) \\
 & & \uparrow{\scriptstyle b_p} & & \uparrow{\scriptstyle b'_p} & & \uparrow{\scriptstyle b_p} \\
\cdots & \longrightarrow & E([2]) & \xrightarrow{\mathsf{id}+\cdots+\tau^{2p-1}} & E([2]) & \xrightarrow{\mathsf{id}-\tau} & E([2]) \\
 & & \uparrow{\scriptstyle b_p} & & \uparrow{\scriptstyle b'_p} & & \uparrow{\scriptstyle b_p} \\
\vdots & & \vdots & & \vdots & & \vdots \\
 & & \uparrow{\scriptstyle b_p} & & \uparrow{\scriptstyle b'_p} & & \uparrow{\scriptstyle b_p} \\
\cdots & \longrightarrow & E([n]) & \xrightarrow{\mathsf{id}+\tau+\cdots+\tau^{pn-1}} & E([n]) & \xrightarrow{\mathsf{id}-\tau} & E([n]) \\
 & & \uparrow{\scriptstyle b_p} & & \uparrow{\scriptstyle b'_p} & & \uparrow{\scriptstyle b_p} \\
 & & \vdots & & \vdots & & \vdots
\end{array} \tag{2.10}$$

Just as for the complex (2.8), we have periodicity, the periodic version of the complex, and the Connes exact sequence (2.6) (the role of Hochschild homology is played by the standard chain complex of the simplicial vector space $j^*E \in \mathrm{Fun}(\Delta^{opp}, k)$).

LEMMA 2.4. *For any $E \in \mathrm{Fun}(\Lambda_p, k)$, the homology $H_\bullet(\Lambda_p, E)$ can be computed by the bicomplex* (2.10).

Proof. The homology of the total complex of (2.10) is obviously a homological functor from $\mathrm{Fun}(\Lambda_p, k)$ to k (that is, short exact sequences in $\mathrm{Fun}(\Lambda_p, k)$ give long exact sequences in homology). Therefore it suffices to prove the claim for a set of projective generators of the category $\mathrm{Fun}(\Lambda_p, k)$. For instance, it suffices to consider all the representable functors E_n, $n \geq 1$—that is, the functors given by

$$E_n([m]) = k\left[\Lambda_p([n], [m])\right],$$

where in the right-hand side we take the k-linear span. Then on one hand, for general tautological reasons—essentially by the Yoneda Lemma—$H_\bullet(\Lambda_p, E_n)$ is k in degree 0 and 0 in higher degrees. On the other hand, the action of the cyclic group $\mathbb{Z}/pm\mathbb{Z}$ generated by $\tau \in \Lambda_p([m], [m])$ on $\Lambda_p([n], [m])$ is obviously free, and we have

$$\Lambda_p([n], [m])/\tau \cong \Delta^{opp}([n], [m])$$

—every $f : \mathbb{Z} \to \mathbb{Z}$ can be uniquely decomposed as $f = \tau^j \circ f_0$, where $0 \leq j < pm$, and f_0 sends 0 to 0. The rows of the complex (2.10) compute

$$H_\bullet(\mathbb{Z}/pm\mathbb{Z}, E_n([m])) \cong k\left[\Delta^{opp}([n], [m])\right],$$

and the first term in the corresponding spectral sequence is the standard complex for the simplicial vector space $E_n^\Delta \in \mathrm{Fun}(\Delta^{opp}, k)$ represented by $[n] \in \Delta^{opp}$. Therefore this complex computes $H_\bullet(\Delta^{opp}, E_n^\Delta)$, which is again k. □

The complex (2.5) is the special case of (2.10) for $p = 1$ and the following object $A_\# \in \mathrm{Fun}(\Lambda, k)$: we set $A_\#([n]) = A^{\otimes n}$, where the factors are numbered by

elements in the set $V([n]) = \mathbb{Z}/n\mathbb{Z}$, and any $f \in \Lambda([n],[m])$ acts by

$$A_{\#}(f)\left(\bigotimes_{i\in V([n])} a_i\right) = \bigotimes_{j\in V([m])} \prod_{i\in f^{-1}(j)} a_i,$$

(if $f^{-1}(i)$ is empty for some $i \in V([n])$, then the right-hand side involves a product numbered by the empty set; this is defined to be the unity element $1 \in A$). To obtain the complex (2.8), we note that for any p, we have a functor $i : \Lambda_p \to \Lambda$ given by $[n] \mapsto [pn]$, $f \mapsto f$. Then (2.10) applied to $i^*A_{\#} \in \mathrm{Fun}(\Lambda_p, k)$ gives (2.8). By Lemma 2.4 we have

$$HC_\bullet(A) \cong H_\bullet(\Lambda, k),$$
$$HC_\bullet(A^{\otimes p}, A_\sigma^{\otimes p}) \cong H_\bullet(\Lambda_p, k).$$

LEMMA 2.5 ([**Kal08**, Lemma 1.12]). *For any $E \in \mathrm{Fun}(\Lambda, k)$, we have a natural isomorphism*

$$H_\bullet(\Lambda_p, i^*E) \cong H_\bullet(\Lambda, E),$$

which is compatible with the periodicity map and with the Connes exact sequence (2.6). □

Thus $HC_\bullet(A^{\otimes p}, A_\sigma^{\otimes p}) \cong HC_\bullet(A)$. The proof of this Lemma is not difficult. First of all, a canonical comparison map $H_\bullet(\Lambda_p, i^*E) \to H_\bullet(\Lambda, E)$ exists for tautological adjunction reasons. Moreover, the periodicity homomorphism for $H_\bullet(\Lambda_p, -)$ is induced by the action of a canonical element $u_p \in H^2(\Lambda_p, k) = \mathrm{Ext}^2(k, k)$, where k means the constant functor $[n] \mapsto k$ from Λ_p to k. One checks explicitly that $i^*u = u_p$, so that the comparison map is indeed compatible with periodicity, and then it suffices to prove that the comparison map

$$H_\bullet(\Delta^{opp}, i^*E) \to H_\bullet(\Delta^{opp}, E)$$

is an isomorphism. When E is of the form $A_{\#}$, this is Lemma 2.2; in general, one shows that $\mathrm{Fun}(\Lambda, k)$ has a projective generator of the form $A_{\#}$. For details, we refer the reader to [**Kal08**].

3. One vanishing result

For our construction of the Cartier map, we will need one vanishing-type result on periodic cyclic homology in prime characteristic—we want to claim that the periodic cyclic homology $HP_\bullet(E)$ of a p-cyclic object E vanishes under some assumptions on E.

First, consider the cyclic group $\mathbb{Z}/np\mathbb{Z}$ for some $n, p \geq 1$, with the subgroup $\mathbb{Z}/p\mathbb{Z} \subset \mathbb{Z}/pn\mathbb{Z}$ and the quotient $\mathbb{Z}/n\mathbb{Z} = (\mathbb{Z}/pn\mathbb{Z})/(\mathbb{Z}/p\mathbb{Z})$. It is well-known that for any representation V of the group $\mathbb{Z}/pn\mathbb{Z}$, we have the Hochschild-Serre spectral sequence

$$H_\bullet(\mathbb{Z}/n\mathbb{Z}, H_\bullet(\mathbb{Z}/p\mathbb{Z}, V)) \Rightarrow H_\bullet(\mathbb{Z}/pn\mathbb{Z}, -).$$

To see it explicitly, one can compute the homology $H_\bullet(\mathbb{Z}/np\mathbb{Z}, V)$ by a complex which is slightly more complicated than the standard one. Namely, write down the

diagram

$$
\begin{array}{ccccc}
\xrightarrow{\mathsf{id}-\sigma} & V & \xrightarrow{d_\sigma} & V & \xrightarrow{\mathsf{id}-\sigma} V \\
& \uparrow{\scriptstyle \mathsf{id}-\tau} & & \uparrow{\scriptstyle \mathsf{id}-\tau} & \uparrow{\scriptstyle \mathsf{id}-\tau} \\
\xrightarrow{\mathsf{id}-\sigma} & V & \xrightarrow{d_\sigma} & V & \xrightarrow{\mathsf{id}-\sigma} V \\
& \uparrow{\scriptstyle d_\tau} & & \uparrow{\scriptstyle d_\tau} & \uparrow{\scriptstyle d_\tau} \\
\xrightarrow{\mathsf{id}-\sigma} & V & \xrightarrow{d_\sigma} & V & \xrightarrow{\mathsf{id}-\sigma} V \\
& \uparrow{\scriptstyle \mathsf{id}-\tau} & & \uparrow{\scriptstyle \mathsf{id}-\tau} & \uparrow{\scriptstyle \mathsf{id}-\tau}
\end{array}
\tag{3.1}
$$

where τ is the generator of $\mathbb{Z}/pn\mathbb{Z}$, $\sigma = \tau^n$ is the generator of $\mathbb{Z}/p\mathbb{Z} \subset \mathbb{Z}/pn\mathbb{Z}$, and $d_\sigma = \mathsf{id} + \sigma + \cdots + \sigma^p$, $d_\tau = \mathsf{id} + \tau + \cdots + \tau^{n-1}$. This is not quite a bicomplex since the vertical differential squares to $\mathsf{id} - \sigma$, not to 0; to correct this, we add to the total differential the term $\mathsf{id} : V \to V$ of bidegree $(-1, 2)$ in every term in the columns with odd numbers (when counting from the right). The result is a filtered complex which computes $H_\bullet(\mathbb{Z}/pn\mathbb{Z}, V)$, and the Hochschild-Serre spectral sequence appears as the spectral sequence of the filtered complex (3.1).

One feature which is apparent in the complex (3.1) is that it has two different periodicity endomorphisms: the endomorphism which shifts the diagram to the left by two columns (we will denote it by u), and the endomorphism which shifts the diagram downwards by two rows (we will denote it by u').

Assume now given a field k and a p-cyclic object $E \in \mathrm{Fun}(\Lambda_p, k)$, and consider the complex (2.10). Its n-th row is the standard periodic complex which computes $H_\bullet(\mathbb{Z}/pn\mathbb{Z}, E([n]))$, and we can replace all these complexes by the corresponding complex (3.1). By virtue of Lemma 2.3, the result is a certain filtered bicomplex of the form

$$
\begin{array}{ccccc}
\xrightarrow{\mathsf{id}-\sigma} & C_\bullet(E) & \xrightarrow{\mathsf{id}+\sigma+\cdots+\sigma^{p-1}} & C_\bullet(E) & \xrightarrow{\mathsf{id}-\sigma} C_\bullet(E) \\
& \uparrow{\scriptstyle B} & & \uparrow{\scriptstyle B} & \uparrow{\scriptstyle B} \\
\xrightarrow{\mathsf{id}-\sigma} & C'_\bullet(E) & \xrightarrow{\mathsf{id}+\sigma+\cdots+\sigma^{p-1}} & C'_\bullet(E) & \xrightarrow{\mathsf{id}-\sigma} C'_\bullet(E) \\
& \uparrow{\scriptstyle B} & & \uparrow{\scriptstyle B} & \uparrow{\scriptstyle B} \\
\xrightarrow{\mathsf{id}-\sigma} & C_\bullet(E) & \xrightarrow{\mathsf{id}+\sigma+\cdots+\sigma^{p-1}} & C_\bullet(E) & \xrightarrow{\mathsf{id}-\sigma} C_\bullet(E), \\
& \uparrow{\scriptstyle B} & & \uparrow{\scriptstyle B} & \uparrow{\scriptstyle B}
\end{array}
\tag{3.2}
$$

with id of degree $(-1, 2)$ added to the total differential, where $C_\bullet(E)$, resp. $C'_\bullet(E)$, is the complex with terms $E([n])$ and the differential b_p, resp. b'_p, and B is the horizontal differential in the complex (2.10) written down for $p = 1$. The complex $C_\bullet(E)$ computes the Hochschild homology $HH_\bullet(E)$, the complex $C'_\bullet(E)$ is acyclic, and the whole complex (3.2) computes the cyclic homology $HC_\bullet(E)$.

We see that the cyclic homology of the p-cyclic object E actually admits two periodicity endomorphisms: u and u'. The horizontal endomorphism u is the usual periodicity map; the vertical map u' is something new. However, we have the following.

LEMMA 3.1. *In the situation above, assume that* $p = \mathsf{char}\, k$. *Then the vertical periodicity map* $u' : HC_\bullet(E) \to HC_{\bullet-2}(E)$ *is equal to* 0.

Sketch of a proof. It might be possible to write explicitly a contracting homotopy for the map u', but this is very complicated; instead, we will sketch the "scientific" proof which uses small categories (for details, see [**Kal08**]). For any small category Γ, its *cohomology* $H^\bullet(\Gamma, k)$ is defined as

$$H^\bullet(\Gamma, k) = \mathrm{Ext}^\bullet_{\mathrm{Fun}(\Gamma,k)}(k, k),$$

where k in the right-hand side is the constant functor. This is an algebra which obviously acts on $H_\bullet(\Gamma, E)$ for any $E \in \mathrm{Fun}(\Gamma, k)$.

The cohomology $H^\bullet(\Lambda, k)$ of the cyclic category Λ is the algebra of polynomials in one generator u of degree 2, $u \in H^2(\Lambda, k)$; the action of this u on the cyclic homology $HC_\bullet(-)$ is the periodicity map. The same is true for the m-cyclic categories Λ_m for all $m \geq 1$.

Now, recall that we have a natural functor $\pi : \Lambda_p \to \Lambda$, so that there are two natural elements in $H^2(\Lambda_p, k)$—the generator u and the preimage $\pi^*(u)$ of the generator $u \in H^2(\Lambda, k)$. The action of u gives the horizontal periodicity endomorphism of the complex (3.2), and the action of $\pi^*(u)$ gives the vertical periodicity endomorphism u'. We have to prove that if $\mathsf{char}\, k = p$, then $\pi^*(u) = 0$.

To do this, one uses a version of the Hochschild-Serre spectral sequence associated to π—namely, we have a spectral sequence

$$H^\bullet(\Lambda) \otimes H^\bullet(\mathbb{Z}/p\mathbb{Z}, k) \Rightarrow H^\bullet(\Lambda_p, k).$$

If $\mathsf{char}\, k = p$, then the group cohomology algebra $H^\bullet(\mathbb{Z}/p\mathbb{Z}, k)$ is the polynomial algebra $k[u, \varepsilon]$ with two generators: an even generator $u \in H^2(\mathbb{Z}/p\mathbb{Z}, k)$ and an odd generator $\varepsilon \in H^1(\mathbb{Z}/p\mathbb{Z}, k)$. Since $H^\bullet(\Lambda_p, k) = k[u]$, the second differential d_2 in the spectral sequence must send ε to $\pi^*(u)$, so that indeed, $\pi^*(u) = 0$ in $H^2(\Lambda_p, k)$. □

Consider now the version of the complex (3.2) which computes the periodic cyclic homology $HP_\bullet(E)$—to obtain it, one has to extend the diagram to the right by periodicity. The rows of the extended diagram then become the standard complexes which compute the *Tate homology* $\check{H}_\bullet(\mathbb{Z}/p\mathbb{Z}, C_\bullet(E))$. We remind the reader that the Tate homology $\check{H}_\bullet(G, -)$ is a certain homological functor defined for any finite group G which combines together homology $H_\bullet(G, -)$ and cohomology $H^\bullet(G, -)$, and that for a cyclic group $\mathbb{Z}/m\mathbb{Z}$ with generator σ, the Tate homology $\check{H}_\bullet(\mathbb{Z}/m\mathbb{Z}, W)$ with coefficients in some representation W may be computed by the 2-periodic standard complex

$$\cdots \xrightarrow{d_-} W \xrightarrow{d_+} W \xrightarrow{d_-} W \xrightarrow{d_+} \cdots \tag{3.3}$$

with $d_+ = \mathsf{id} + \sigma + \cdots + \sigma^{m-1}$ and $d_- = \mathsf{id} - \sigma$.

If W is a free module over the group algebra $k[G]$, then the Tate homology vanishes in all degrees, $\check{H}_\bullet(G, W) = 0$. When $G = \mathbb{Z}/m\mathbb{Z}$, this means that the standard complex is acyclic. If m is prime and equal to the characteristic of the base field k, the converse is also true—$\check{H}_\bullet(\mathbb{Z}/m\mathbb{Z}, W) = 0$ if and only if W is free over $k[\mathbb{Z}/m\mathbb{Z}]$. We would like to claim a similar vanishing for Tate homology $\check{H}_\bullet(\mathbb{Z}/p\mathbb{Z}, W_\bullet)$ with coefficients in some *complex* $W_\bullet$ of $k[\mathbb{Z}/p\mathbb{Z}]$-modules; however, this is not possible unless we impose some finiteness conditions on $W_\bullet$.

Definition 3.2. A complex $W_\bullet$ of $k[\mathbb{Z}/p\mathbb{Z}]$-modules is *effectively finite* if it is chain homotopic to a complex of finite length. A p-cyclic object $E \in \mathrm{Fun}(\Lambda_p, k)$ is *small* if its standard complex $C_\bullet(E)$ is effectively finite.

Now we can finally state our vanishing result for periodic cyclic homology.

Proposition 3.3. *Assume that $p = \mathsf{char}\, k$. Assume that a p-cyclic object $E \in \mathrm{Fun}(\Lambda_p, k)$ is small, and that $E([n])$ is a free $k[\mathbb{Z}/p\mathbb{Z}]$-module for every object $[n] \in \Lambda_p$. Then $HP_\bullet(E) = 0$.*

Proof. To compute $HP_\bullet(E)$, let us use the periodic version of the complex (3.2). We then have a long exact sequence of cohomology

$$HP_{\bullet-1}(E) \longrightarrow \check{H}_\bullet(\mathbb{Z}/p\mathbb{Z}, C_\bullet(E)) \longrightarrow HP_\bullet(E) \xrightarrow{u'} ,$$

where $\check{H}_\bullet(\mathbb{Z}/p\mathbb{Z}, C_\bullet(E))$ is computed by the total complex of the bicomplex

$$\begin{array}{ccccccc}
\cdots & \xrightarrow{\mathsf{id}+\sigma+\cdots+\sigma^{p-1}} & E([1]) & \xrightarrow{\mathsf{id}-\sigma} & E([1]) & \xrightarrow{\mathsf{id}+\sigma+\cdots+\sigma^{p-1}} & \cdots \\
 & & \uparrow{\scriptstyle b_p} & & \uparrow{\scriptstyle b_p} & & \\
\cdots & \xrightarrow{\mathsf{id}+\sigma+\cdots+\sigma^{p-1}} & E([2]) & \xrightarrow{\mathsf{id}-\sigma} & E([2]) & \xrightarrow{\mathsf{id}+\sigma+\cdots+\sigma^{p-1}} & \cdots \\
 & & \uparrow{\scriptstyle b_p} & & \uparrow{\scriptstyle b_p} & & \\
 & & \vdots & & \vdots & & \\
 & & \uparrow{\scriptstyle b_p} & & \uparrow{\scriptstyle b_p} & & \\
\cdots & \xrightarrow{\mathsf{id}+\sigma+\cdots+\sigma^{p-1}} & E([n]) & \xrightarrow{\mathsf{id}-\sigma} & E([n]) & \xrightarrow{\mathsf{id}+\sigma+\cdots+\sigma^{p-1}} & \cdots \\
 & & \uparrow{\scriptstyle b_p} & & \uparrow{\scriptstyle b_p} & & \\
 & & \vdots & & \vdots & &
\end{array} \tag{3.4}$$

By Lemma 3.1, the connecting differential in the long exact sequence vanishes, so that it suffices to prove that $\check{H}_\bullet(\mathbb{Z}/p\mathbb{Z}, C_\bullet(E)) = 0$. Since $E([n])$ is free, all the rows of the bicomplex (3.4) are acyclic. But since E is small, $C_\bullet(E)$ is effectively finite; therefore the spectral sequence of the bicomplex (3.4) converges, and we are done. □

4. Quasi-Frobenius maps

We now fix a perfect base field k of characteristic $p > 0$, and consider an associative algebra A over k. We want to construct a cyclic homology version of the Cartier isomorphism (1.2) for A. In fact, we will construct a version of the inverse isomorphism C^{-1}; it will be an isomorphism

$$C^{-1} : HH_\bullet(A)((u))^{(1)} \longrightarrow HP_\bullet(A), \tag{4.1}$$

where, as before, $HH_\bullet(A)((u))$ in the left-hand side means "Laurent power series in one variable u of degree 2 with coefficients in $HH_\bullet(A)$".

If A is commutative and $X = \mathrm{Spec}\, A$ is smooth, then $HH_\bullet(A) \cong \Omega^\bullet(X)$, $HP_\bullet(A) \cong H^\bullet_{DR}(X)((u))$, and (4.1) is obtained by inverting (1.2) (and repeating the resulting map infinitely many times, once for every power of the formal variable u).

It is known that the commutative inverse Cartier map is induced by the Frobenius isomorphism; thus to generalize it to non-commutative algebras, it is natural to start the story with the Frobenius map.

At first glance, the story thus started ends immediately: the map $a \mapsto a^p$ is not an algebra endomorphism of A unless A is commutative (in fact, the map is not even additive, $(x+y)^p \neq x^p + y^p$ for general non-commuting x and y). So, there is no Frobenius map in the non-commutative world.

However, to analyze the difficulty, let us decompose the usual Frobenius into two maps:

$$A \xrightarrow{\varphi} A^{\otimes p} \xrightarrow{M} A,$$

where φ is given by $\varphi(a) = a^{\otimes p}$, and M is the multiplication map, $M(a_1 \otimes \cdots \otimes a_p) = a_1 \cdots a_p$. The map φ is very bad (e.g. not additive), but this is the same both in the commutative and in the general associative case. It is the map M which creates the problem: it is an algebra map if and only if A is commutative.

In general, it is not possible to correct M so that it becomes an algebra map. However, even not being an algebra map, it can be made to act on Hochschild homology, and we already saw how: we can take the map (2.7) of Subsection 2.2.

As for the very bad map φ, fortunately, it turns out that it can be perturbed quite a bit. In fact, the only property of this map which is essential is the following one.

LEMMA 4.1. *Let V be a vector space over k, and let the cyclic group $\mathbb{Z}/p\mathbb{Z}$ act on its p-th tensor power $V^{\otimes p}$ by cyclic permutation. Then the map $\varphi : V \to V^{\otimes p}$, $v \mapsto v^{\otimes p}$ sends V into the kernel of either of the differentials d_+, d_- of the standard complex* (3.3) *and induces an isomorphism*

$$V^{(1)} \to \check{H}_i(\mathbb{Z}/p\mathbb{Z}, V^{\otimes p})$$

both for odd and even degrees i.

Proof. The map φ is compatible with the multiplication by scalars, and its image is σ-invariant, so that it indeed sends V into the kernel of either of the differentials $d_-, d_+ : V^{\otimes p} \to V^{\otimes p}$. We claim that it is additive "modulo $\operatorname{Im} d_\pm$", and that it induces an isomorphism $V^{(1)} \cong \operatorname{Ker} d_\pm / \operatorname{Im} d_\mp$. Indeed, choose a basis in V, so that $V \cong k[S]$, the k-linear span of a set S. Then $V^{\otimes p} = k[S^p]$ decomposes as $k[S^p] = k[S] \oplus k[S^p \setminus \Delta]$, where $S \cong \Delta \subset S^p$ is the diagonal. This decomposition is compatible with the differentials $d_\pm$, which actually vanish on the first summand $k[S]$. The map φ, accordingly, decomposes as $\varphi = \varphi_0 \oplus \varphi_1$, $\varphi_0 : V^{(1)} \to k[S]$, $\varphi_1 : V^{(1)} \to k[S^p \setminus \Delta]$. The map φ_0 is obviously additive and an isomorphism; therefore it suffices to prove that the second summand of (3.3) is acyclic. Indeed, since the $\mathbb{Z}/p\mathbb{Z}$-action on $S^p \setminus \Delta$ is free, we have $\check{H}^\bullet(\mathbb{Z}/p\mathbb{Z}, k[S^p \setminus \Delta]) = 0$. □

DEFINITION 4.2. A *quasi-Frobenius map* for an associative unital algebra A over k is a $\mathbb{Z}/p\mathbb{Z}$-equivariant algebra map $F : A^{(1)} \to A^{\otimes p}$ which induces the isomorphism $\check{H}^\bullet(\mathbb{Z}/p\mathbb{Z}, A^{(1)}) \to \check{H}^\bullet(\mathbb{Z}/p\mathbb{Z}, A^{\otimes p})$ of Lemma 4.1.

Here the $\mathbb{Z}/p\mathbb{Z}$-action on $A^{(1)}$ is trivial, and the algebra structure on $A^{\otimes p}$ is the obvious one (all the p factors commute). We note that since $\check{H}^i(\mathbb{Z}/p\mathbb{Z}, k) \cong k$ for every i, we have $\check{H}^i(\mathbb{Z}/p\mathbb{Z}, A^{(1)}) \cong A^{(1)}$, so that a quasi-Frobenius map must be injective. Moreover, since the Tate homology $\check{H}^\bullet(\mathbb{Z}/p\mathbb{Z}, A^{\otimes p}/A^{(1)})$ vanishes, the cokernel of a quasi-Frobenius map must be a free $k[\mathbb{Z}/p\mathbb{Z}]$-module.

In this Section, we will construct a Cartier isomorphism (4.1) for algebras which admit a quasi-Frobenius map (and satisfy some additional assumptions). In the interest of full disclosure, we remark right away that quasi-Frobenius maps are very rare—in fact, we know only two examples:

(i) A is the tensor algebra $T^\bullet V$ of a k-vector space V—it suffices to give F on the generators, where it exists by Lemma 4.1.
(ii) $A = k[G]$ is the group algebra of a (discrete) group G—a quasi-Frobenius map F is induced by the diagonal embedding $G \subset G^p$.

However, the general construction of the Cartier map given in Section 5 will be essentially the same—it is only the notion of a quasi-Frobenius map that we will modify.

PROPOSITION 4.3. *Assume given an algebra A over k equipped with a quasi-Frobenius map $F : A^{(1)} \to A^{\otimes p}$, and assume that the category A-bimod of A-bimodules has finite homological dimension. Then there exists a canonical isomorphism*

$$\varphi : HH_\bullet(A)((u)) \cong HP_\bullet(A).$$

Proof. Consider the functors $i, \pi : \Lambda_p \to \Lambda$ and the restrictions

$$\pi^* A^{(1)}_\#, i^* A_\# \in \mathrm{Fun}(\Lambda_p, k).$$

For any $[n] \in \Lambda_p$, the quasi-Frobenius map $F : A^{(1)} \to A^{\otimes p}$ induces a map

$$F^{\otimes n} : \pi^* A^{(1)}_\#([n]) = (A^{(1)})^{\otimes n} \to i^* A_\#([n]) = A^{\otimes pn}.$$

By the definition of a quasi-Frobenius map, these maps commute with the action of the maps $\tau : [n] \to [n]$ and $m^p_i : [n+1] \to [n]$, $0 \leq i < n$ (recall that $m^p_i = m^{\otimes p}_i$). Moreover, since $m^p_0 \circ \tau = m^p_{n+1}$, $F^{\otimes \bullet}$ also commutes with m^p_n. All in all, the collection of the tensor power maps $F^{\otimes \bullet}$ gives a map $F_\# : \pi^* A^{(1)}_\# \to i^* A_\#$ of objects in $\mathrm{Fun}(\Lambda_p, k)$. We denote by Φ the induced map

$$\Phi = HP_\bullet(F_\#) : HP_\bullet(\pi^* A^{(1)}_\#) \to HP_\bullet(\Lambda_p, i^* A_\#).$$

By Lemma 2.5, the right-hand side is precisely $HP_\bullet(A)$. As for the left-hand side, we note that σ is trivial on $\pi^* A_\#([n])$ for every $[n] \in \Lambda_p$; therefore the odd horizontal differentials

$$\begin{aligned} \mathsf{id} + \tau + \cdots + \tau^{pn-1} &= (\mathsf{id} + \tau + \cdots + \tau^{n-1}) \circ (\mathsf{id} + \sigma + \cdots + \sigma^{p-1}) \\ &= p(\mathsf{id} + \tau + \cdots + \tau^{n-1}) = 0 \end{aligned}$$

in (2.10) vanish, and we have

$$HP_\bullet(\pi^* A^{(1)}_\#) \cong HH_\bullet(A^{(1)})((u)).$$

Finally, to show that Φ is an isomorphism, we recall that the quasi-Frobenius map F is injective, and its cokernel is a free $k[\mathbb{Z}/p\mathbb{Z}]$-module. One deduces easily that the same is true for each tensor power $F^{\otimes n}$; thus $F_\#$ is injective, and its cokernel $\mathrm{Coker}\, F_\#$ is such that $\mathrm{Coker}\, F_\#([n])$ is a free $k[\mathbb{Z}/p\mathbb{Z}]$-module for any $[n] \in \Lambda_p$. To finish the proof, use the long exact sequence of cohomology and Proposition 3.3. The only thing left to check is that Proposition 3.3 is applicable—namely, that the p-cyclic object $i^* A_\#$ is small in the sense of Definition 3.2.

To do this, we have to show that the bar complex $C_\bullet(A^{\otimes p}, A_\sigma^{\otimes p})$ which computes $HH_\bullet(A^{\otimes p}, A_\sigma^{\otimes p})$ is effectively finite. It is here that we need to use the assumption of finite homological dimension on the category A-bimod. Indeed, to compute $HH_\bullet(A^{\otimes p}, A_\sigma^{\otimes p})$, we can choose any projective resolution $P_\bullet^p$ of the diagonal $A^{\otimes p}$-bimodule $A^{\otimes p}$. In particular, we can take any projective resolution $P_\bullet$ of the diagonal A-bimodule A, and use its p-th power. To obtain the bar complex $C_\bullet(A^{\otimes p}, A_\sigma^{\otimes p})$, one uses the bar resolution $C'_\bullet(A)$. However, all these projective resolutions $P_\bullet$ are chain homotopic to each other, so that the resulting complexes will be also chain homotopic *as complexes of $k[\mathbb{Z}/p\mathbb{Z}]$-modules*. By assumption, the diagonal A-bimodule A has a projective resolution $P_\bullet$ of finite length; using it gives a complex of finite length which is chain homotopic to $C_\bullet(A^{\otimes p}, A_\sigma^{\otimes p})$, just as required by Definition 3.2. □

5. Cartier isomorphism in the general case

5.1. Additivization. We now turn to the general case: we assume given a perfect field k of characteristic $p > 0$ and an associative k-algebra A, and we want to construct a Cartier-type isomorphism (4.1) without assuming that A admits a quasi-Frobenius map in the sense of Definition 4.2.

Consider again the non-additive map $\varphi : A \to A^{\otimes p}$, $a \mapsto a^{\otimes p}$, and let us change the domain of its definition: instead of A, let φ be defined on the k-vector space $k[A]$ spanned by A (where A is considered as a set). Then φ obviously uniquely extends to a k-linear additive map

$$\varphi : k[A] \to A^{\otimes p}. \tag{5.1}$$

Taking the k-linear span is a functorial operation: setting $V \mapsto k[V]$ defines a functor Span_k from the category of k-vector spaces to itself. The functor Span_k is non-additive, but it has a tautological surjective map $\mathsf{Span}_k \to \mathsf{Id}$ onto the identity functor, and one can show that Id is the maximal additive quotient of the functor Span_k. If $V = A$ is an algebra, then $\mathsf{Span}_k(A)$ is also an algebra, and the tautological map $\mathsf{Span}_k(A) \to A$ is an algebra map.

We note that in both examples (i), (ii) in Section 4 where an algebra A did admit a quasi-Frobenius map, what really happened was that the tautological surjective algebra map $\mathsf{Span}_k(A) \to A$ admitted a splitting $s : A \to \mathsf{Span}_k(A)$; the quasi-Frobenius map was obtained by composing this splitting map s with the canonical map (5.1).

Unfortunately, in general the projection $\mathsf{Span}_k(A) \to A$ does not admit a splitting (or at least, it is not clear how to construct one). In the general case, we will modify both sides of the map (5.1) so that splittings will become easier to come by. To do this, we use the general technique of *additivization* of non-additive functors from the category of k-vector spaces to itself.

Consider the small category $\mathcal{V} = k\text{-Vect}^{\mathsf{fg}}$ of finite-dimensional k-vector spaces, and consider the category $\mathrm{Fun}(\mathcal{V}, k)$ of *all* functors from $\mathcal{V}$ to the category k-Vect of all k-vector spaces. This is an abelian category. The category $\mathrm{Fun}_{add}(\mathcal{V}, k)$ of all *additive* functors from $\mathcal{V}$ to k-Vect is also abelian (in fact, an additive functor is completely defined by its value at the one-dimensional vector space k, so that $\mathrm{Fun}_{add}(\mathcal{V}, k)$ is equivalent to the category of modules over $k \otimes_{\mathbb{Z}} k$). We have the full embedding $\mathrm{Fun}_{add}(\mathcal{V}, k) \subset \mathrm{Fun}(\mathcal{V}, k)$, and it admits a left-adjoint functor—in other words, for any functor $F \in \mathrm{Fun}(\mathcal{V}, k)$ there is an additive functor F_{add} and a

map $F \to F_{add}$ which is universal with respect to maps to additive functors. This "universal additive quotient" is not very interesting. For instance, if F is the p-th tensor power functor, $V \mapsto V^{\otimes p}$, then its universal additive quotient is the trivial functor $V \mapsto 0$.

To obtain a useful version of this procedure, we have to consider the derived category $\mathcal{D}(\mathcal{V}, k)$ of the category $\mathrm{Fun}(\mathcal{V}, k)$ and the full subcategory $\mathcal{D}_{add}(\mathcal{V}, k) \subset \mathcal{D}(\mathcal{V}, k)$ spanned by complexes whose homology object lie in $\mathrm{Fun}_{add}(\mathcal{V}, k)$.

The category $\mathcal{D}_{add}(\mathcal{V}, k)$ is closed under taking cones, thus triangulated (this has to be checked, but this is not difficult), and it contains the derived category of the abelian category $\mathrm{Fun}_{add}(\mathrm{Fun}_k, k)$. However, $\mathcal{D}_{add}(\mathcal{V}, k)$ is much larger than this derived category. In fact, even for the identity functor $\mathsf{Id} \in \mathrm{Fun}_{add}(\mathcal{V}, k) \subset \mathrm{Fun}(\mathcal{V}, k)$, the natural map

$$\mathrm{Ext}^i_{\mathrm{Fun}_{add}(\mathcal{V},k)}(\mathsf{Id}, \mathsf{Id}) \to \mathrm{Ext}^i_{\mathrm{Fun}(\mathcal{V},k)}(\mathsf{Id}, \mathsf{Id})$$

is an isomorphism only in degrees 0 and 1. Already in degree 2, there appear extension classes which cannot be represented by a complex of additive functors.

Nevertheless, it turns out that just as for abelian categories, the full embedding $\mathcal{D}_{add}(\mathcal{V}, k) \subset \mathcal{D}(\mathcal{V}, k)$ admits a left-adjoint functor. We call it the *additivization functor* and denote by $\mathsf{Add}_\bullet : \mathcal{D}(\mathcal{V}, k) \to \mathcal{D}_{add}(\mathcal{V}, k)$. For any $F \in \mathrm{Fun}(\mathcal{V}, k)$, $\mathsf{Add}_\bullet(F)$ is a complex of functors from $\mathcal{V}$ to k with additive homology functors.

The construction of the additivization $\mathsf{Add}_\bullet$ is relatively technical; we will not reproduce it here and refer the reader to [**Kal08**, Section 3]. The end result is that first, additivization exists, and second, it can be represented explicitly, by a very elegant "cube construction" introduced fifty years ago by Eilenberg and MacLane. Namely, to any functor $F \in \mathrm{Fun}(\mathcal{V}, k)$ one associates a complex $Q_\bullet(F)$ of functors from $\mathcal{V}$ to k such that the homology of this complex consists of additive functors, and we have an explicit map $F \to Q_\bullet(F)$ which descends to a universal map in the derived category $\mathcal{D}(\mathcal{V}, k)$. In fact, the complex $Q_\bullet(F)$ is concentrated in non-negative homological degrees, and $Q_0(F)$ simply coincides with F, so that the universal map is the tautological embedding $F = Q_0(F) \to Q_\bullet(F)$. Moreover, assume that the functor F is *multiplicative* in the following sense: for any $V, W \in \mathcal{V}$, we have a map

$$F(V) \otimes F(W) \to F(V \otimes W),$$

and these maps are functorial and associative in the obvious sense. Then the complex $Q_\bullet(F)$ is also multiplicative. In particular, if we are given a multiplicative functor F and an associative algebra A, then $F(A)$ is an associative algebra; in this case, $Q_\bullet(F)$ is an associative DG algebra concentrated in non-negative degrees.

5.2. Generalized Cartier map. Consider again the canonical map (5.1). There are two non-additive functors involved: the k-linear span functor $V \mapsto k[V]$, and the p-th tensor power functor $V \mapsto V^{\otimes p}$. Both are multiplicative. We will denote by $Q_\bullet(V)$ the additivization of the k-linear span, and we will denote by $P_\bullet(V)$ the additivization of the p-th tensor power. Since additivization is functorial, the map (5.1) gives a map

$$\varphi : Q_\bullet(V) \to P_\bullet(V)$$

for any finite-dimensional k-vector space V; if $A = V$ is an associative algebra, then $Q_\bullet(A)$ and $P_\bullet(A)$ are associative DG algebras, and φ is a DG algebra map. We will need several small refinements of this construction.

(i) We extend both $Q_\bullet$ and $P_\bullet$ to arbitrary vector spaces and arbitrary algebras by taking the limit over all finite-dimensional subspaces.

(ii) The p-th power $V^{\otimes p}$ carries the permutation action of the cyclic group $\mathbb{Z}/p\mathbb{Z}$, and the map (5.1) is $\mathbb{Z}/p\mathbb{Z}$-invariant; by the functoriality of the additivization, $P_\bullet(V)$ also carries an action of $\mathbb{Z}/p\mathbb{Z}$, and the map φ is $\mathbb{Z}/p\mathbb{Z}$-invariant.

(iii) The map (5.1), while not additive, respects the multiplication by scalars, up to a Frobenius twist; unfortunately, the additivization procedure ignores this. From now on, we will assume that the perfect field k is actually finite, so that the group k^* of scalars is a finite group whose order is coprime to p. Then k^* acts naturally on $k[V]$, hence also on $Q_\bullet(V)$, and the map φ factors through the space $\overline{Q}_\bullet(V) = Q_\bullet(V)_{k^*}$ of covariants with respect to k^*.

The end result: in the case of a general algebra A, our replacement for a quasi-Frobenius map is the canonical map

$$\varphi : \overline{Q}_\bullet(A)^{(1)} \to P_\bullet(A), \tag{5.2}$$

which is a $\mathbb{Z}/p\mathbb{Z}$-invariant DG algebra map. We can now repeat the procedure of Section 4 replacing a quasi-Frobenius map F with this canonical map φ. This gives a canonical map

$$\Phi : HH_\bullet(\overline{Q}_\bullet(A)_\#)^{(1)}((u)) \to HP_\bullet(P_\bullet(A)_\#), \tag{5.3}$$

where $\overline{Q}_\bullet(A)_\#$ in the left-hand side is a complex of cyclic objects, and $P_\bullet(A)_\#$ in the right-hand side is the complex of p-cyclic objects. There is one choice to be made because both complexes are infinite; we agree to interpret the total complex which computes $HP_\bullet(E_\bullet)$ and $HH_\bullet(E_\bullet)$ for an infinite complex $E_\bullet$ of cyclic or p-cyclic objects as the *sum*, not the *product* of the corresponding complexes for the individual terms $HP_\bullet(E_i)$, $HH_\bullet(E_i)$.

To understand what (5.3) has to do with the Cartier map (4.1), we need some information on the structure of DG algebras $P_\bullet(A)$ and $\overline{Q}_\bullet(A)$.

The DG algebra $P_\bullet(A)$ has the following structure: $P_0(A)$ is isomorphic to the p-th tensor power $A^{\otimes p}$ of the algebra A, and all the higher terms $P_i(A)$, $i \geq 1$ are of the form $A^{\otimes p} \otimes W_i$, where W_i is a certain representation of the cyclic group $\mathbb{Z}/p\mathbb{Z}$. The only thing that will matter to us is that all the representations W_i are *free* $k[\mathbb{Z}/p\mathbb{Z}]$-modules. Consequently, $P_i(A)$ is free over $k[\mathbb{Z}/p\mathbb{Z}]$ for all $i \geq 1$. For the proofs, we refer the reader to [**Kal08**, Subsection 4.1]. As a corollary, we see that if A is such that A-bimod has finite homological dimension, then we can apply Proposition 3.3 to all the higher terms in the complex $P_\bullet(A)_\#$ and deduce that the right-hand of (5.3) is actually isomorphic to $HP_\bullet(A)$:

$$HP_\bullet(P_\bullet(A)_\#) \cong HP_\bullet(A).$$

We note that it is here that it matters how we define the periodic cyclic homology of an infinite complex (the complex $P_\bullet(A)$ is actually acyclic, so, were we to take the product and not the sum of individual terms, the result would be 0, not $HP_\bullet(A)$).

The structure of the DG algebra $\overline{Q}_\bullet$ is more interesting. As it turns out, the homology $H_i(\overline{Q}_\bullet(A))$ of this DG algebra in degree i is isomorphic to $A \otimes \mathsf{St}(k)_i$, where $\mathsf{St}(k)_\bullet$ is the dual to the *Steenrod algebra* known in Algebraic Topology—more precisely, $\mathsf{St}(k)_i^*$ is the algebra of stable cohomology operations with coefficients in k

of degree i. The proof of this is contained in [**Kal08**, Section 3]; [**Kal08**, Subsection 3.1] contains a semi-informal discussion of why this should be so, and what is the topological interpretation of all the constructions in this Section. The topological part of the story is quite large and well-developed—among other things, it includes the notions of *Topological Hochschild Homology* and *Topological Cyclic Homology* which have been the focus of much attention in Algebraic Topology in the last fifteen years. A reader who really wants to understand what is going on should definitely consult the sources, some of which are indicated in [**Kal08**]. However, within the scope of the present lectures, we will leave this subject completely alone. The only topological fact that we will need is the following description of the Steenrod algebra in low degrees:

$$\mathsf{St}_i(k) = \begin{cases} k, & i = 0, 1, \\ 0, & 1 < i < 2p - 2. \end{cases} \tag{5.4}$$

The proof can be easily found in any algebraic topology textbook.

Thus in particular, the 0-th homology of $\overline{Q}_\bullet(A)$ is isomorphic to A itself, so that we have an augmentation map $\overline{Q}_\bullet(A) \to A$ (this is actually induced by the tautological map $Q_0(A) = k[A] \to A$). However, there is also non-trivial homology in higher degrees. Because of this, the left-hand side of (5.3) is larger than the left-hand side of (4.1), and the canonical map Φ of (5.3) has no chance of being an isomorphism (for a topological interpretation of the left-hand side of (5.3), see [**Kal08**, Subsection 3.1]).

In order to get an isomorphism (4.1), we have to resort to splittings again, and it would seem that we gained nothing, since splitting the projection $\overline{Q}_\bullet(A) \to A$ is the same as splitting the projection $\overline{Q}_0(A) = k[A]_{k^*} \to A$. Fortunately, in the world of DG algebras we can get away with something less than a full splitting map. We note the following obvious fact: any *quasi-isomorphism* $f : A_\bullet \to B_\bullet$ of DG algebras induces an isomorphism $HH_\bullet(A_\bullet) \to HH_\bullet(B_\bullet)$ of their Hochschild homology. Because of this, it suffices to split the projection $\overline{Q}_\bullet(A) \to A$ "up to a quasi-isomorphism". More precisely, we introduce the following.

Definition 5.1. A *DG splitting* $\langle \overline{A}_\bullet, s \rangle$ of a DG algebra map $f : \widetilde{A}_\bullet \to A_\bullet$ is a pair of a DG algebra $\overline{A}_\bullet$ and a DG algebra map $s : \overline{A}_\bullet \to \widetilde{A}_\bullet$ such that the composition $f \circ s : \overline{A}_\bullet \to A_\bullet$ is a quasi-isomorphism.

Lemma 5.2. *Assume that the associative algebra A is such that A-*bimod *has finite homological dimension. For any DG splitting $\langle A_\bullet, s \rangle$ of the projection $\overline{Q}_\bullet(A) \to A$, the composition map*

$$\Phi \circ s : HH_\bullet(A)^{(1)}((u)) \cong HH_\bullet(A_\bullet)^{(1)}((u)) \to \\ \to HH_\bullet(\overline{Q}_\bullet(A))^{(1)}((u)) \to HP_\bullet(P_\bullet(A)_\#) \cong HP_\bullet(A)$$

is an isomorphism in all degrees.

The proof is not completely trivial but very straightforward; we leave it as an exercise (or see [**Kal08**, Subsection 4.1]). By virtue of this lemma, all we have to do to construct a Cartier-type isomorphism (4.1) is to find a DG splitting of the projection $\overline{Q}_\bullet(A) \to A$.

5.3. DG splittings. To construct DG splittings, we use obstruction theory for DG algebras, which turns out to be pretty much parallel to the usual obstruction theory for associative algebras. A skeleton theory sufficient for our purposes is given in [**Kal08**, Subsection 4.3]. Here are the main points.

(i) Given a DG algebra $A_\bullet$ and a DG $A_\bullet$-bimodule $M_\bullet$, one defines *Hochschild cohomology* $HH^\bullet_{\mathcal{D}}(A_\bullet, M_\bullet)$ as

$$HH^\bullet_{\mathcal{D}}(A_\bullet, M_\bullet) = \operatorname{Ext}^\bullet_{\mathcal{D}}(A_\bullet, M_\bullet),$$

where $A_\bullet$ in the right-hand side is the diagonal $A_\bullet$-bimodule, and $\operatorname{Ext}^\bullet_{\mathcal{D}}$ are the spaces of maps in the "triangulated category of $A_\bullet$-bimodules"—that is, the derived category of the abelian category of DG $A_\bullet$-bimodules localized with respect to quasi-isomorphisms. Explicitly, $HH^\bullet(A_\bullet, M_\bullet)$ can be computed by using the bar resolution of the diagonal bimodule $A_\bullet$. This gives a complex with terms $\operatorname{Hom}(A_\bullet^{\otimes n}, M_\bullet)$, where $n \geq 0$ is a non-negative integer, and a certain differential $\delta : \operatorname{Hom}(A_\bullet^{\otimes \bullet}, M_\bullet) \to \operatorname{Hom}(A_\bullet^{\otimes \bullet+1}, M_\bullet)$; the groups $HH^\bullet(A_\bullet, M_\bullet)$ are computed by the total complex of the bicomplex

$$M_\bullet \xrightarrow{\delta} \operatorname{Hom}(A_\bullet, M_\bullet) \xrightarrow{\delta} \cdots \xrightarrow{\delta} \operatorname{Hom}(A_\bullet^{\otimes \cdots}, M_\bullet) \xrightarrow{\delta} .$$

(ii) By a *square-zero extension* of a DG algebra $A_\bullet$ by a DG $A_\bullet$-bimodule we understand a DG algebra $\widetilde{A_\bullet}$ equipped with a surjective map $\widetilde{A_\bullet} \to A_\bullet$ whose kernel is identified with $M_\bullet$ (in particular, the induced $\widetilde{A_\bullet}$-bimodule structure on the kernel factors through the map $\widetilde{A_\bullet} \to A_\bullet$). Then square-zero extensions are classified up to a quasi-isomorphism by elements in the Hochschild cohomology group $HH^2_{\mathcal{D}}(A_\bullet, M_\bullet)$. A square-zero extension admits a DG splitting if and only if the corresponding class in $HH^2_{\mathcal{D}}(A_\bullet, M_\bullet)$ is trivial.

To apply this machinery to the augmentation map $\overline{Q}_\bullet(A) \to A$, we consider the *canonical filtration* $\overline{Q}_\bullet(A)_{\geq \bullet}$ on $\overline{Q}_\bullet(A)$ defined, as usual, by

$$\overline{Q}_i(A)_{\geq j} = \begin{cases} 0, & i \leq j, \\ \operatorname{Ker} d, & i = j+1, \\ \overline{Q}_i(A), & i > j+1, \end{cases}$$

where d is the differential in the complex $\overline{Q}_\bullet(A)$. We denote the quotients by $\overline{Q}_\bullet(A)_{\leq j}(A) = \overline{Q}_\bullet(A)/\overline{Q}_\bullet(A)_{\geq j}$, and we note that for any $j \geq 1$, $\overline{Q}_\bullet(A)_{\leq j}$ is a square-zero extension of $\overline{Q}_\bullet(A)_{\leq j-1}$ by a DG bimodule quasi-isomorphic to $A \otimes \mathsf{St}_j(k)[j]$ (here $\mathsf{St}_j(k)$ is the corresponding term of the dual Steenrod algebra, and $[j]$ means the degree shift). We use induction on j and construct a collection $\langle A^j_\bullet, s \rangle$ of compatible DG splittings of the surjections $\overline{Q}_\bullet(A)_{\leq j} \to A$. There are three steps.

Step 1. For $j = 0$, there is nothing to do: the projection $\overline{Q}_\bullet(A)_{\leq 0} \to A$ is a quasi-isomorphism.

Step 2. For $j = 1$, it tuns out that the projection $\overline{Q}_\bullet(A)_{\leq 1} \to A$ admits a DG splitting if and only if the k-algebra A admits a lifting to a flat algebra $\widetilde{A}$ over the ring $W_2(k)$ of second Witt vectors of the field k. In fact, even more is true: DG splittings are in some sense in a functorial one-to-one correspondence with such liftings; the reader will find precise statements and explicit detailed proofs in [**Kal08**, Subsection 4.2].

Step 3. We then proceed by induction. Assume given a DG splitting $A^j_\bullet$, $s : A^j_\bullet \to \overline{Q}_\bullet(A)_{\leq j}$ of the projection $\overline{Q}_\bullet(A)_{\leq j} \to A$. Form the "Baer sum" $\overline{A}^j_\bullet$ of the map s with the square-zero extension $p : \overline{Q}_\bullet(A)_{\leq j+1} \to \overline{Q}_\bullet(A)_{\leq j}$—that is, let

$$\overline{A}^j_\bullet \subset \overline{Q}_\bullet(A)_{\leq j+1} \oplus A^j_\bullet$$

be the subalgebra obtained as the kernel of the map

$$\overline{Q}_\bullet(A)_{\leq j+1} \oplus A^j_\bullet \xrightarrow{p\oplus(-s)} \overline{Q}_\bullet(A)_{\leq j}.$$

Then $\overline{A}^j_\bullet$ is a square-zero extension of $A^j_\bullet$ by a DG $A^j_\bullet$-bimodule $\operatorname{Ker} p$ which is quasi-isomorphic to $A \otimes \mathsf{St}_j(k)[j]$. Since $\mathcal{A}^j_\bullet$ is quasi-isomorphic to A, these are classified by elements in the Hochschild cohomology group

$$HH^2(A^j_\bullet, \operatorname{Ker} p) \cong HH^{3+j}(A, A) \otimes \mathsf{St}_{j+1}(k).$$

If $j < 2p-3$, this group is trivial by (5.4), so that a DG splitting $A^{j+1}_\bullet$ exists. In higher degrees, we have to impose conditions on the algebra A. Here is the end result.

Proposition 5.3. *Assume given an associative algebra A over a finite field k of characteristic p such that*

(i) *A lifts to a flat algebra over the ring $W_2(k)$ of second Witt vectors, and*
(ii) *A-bimod has finite homological dimension, and moreover, we have that $HH^j(A, A) = 0$ whenever $j \geq 2p$.*

Then there exists a DG splitting $A_\bullet$, $s : A_\bullet \to \overline{Q}_\bullet(A)$ of the augmentation map $\overline{Q}_\bullet(A) \to A$.

Proof. Construct a compatible system of DG splittings $A^j_\bullet$ as described above, and let $A_\bullet = \lim_{\leftarrow} A^j_\bullet$. □

Theorem 5.4. *Assume given an associative algebra A over a finite field k of characteristic p which satisfies the assumptions* (i), (ii) *of Proposition* 5.3. *Then there exists an isomorphism*

$$C^{-1} : HH_\bullet(A)((u))^{(1)} \longrightarrow HP_\bullet(A),$$

as in (4.1).

Proof. Combine Proposition 5.3 and Lemma 5.2. □

This is our generalized Cartier map. We note that the conditions (i), (ii) that we have to impose on the algebra A are completely parallel to the conditions (i), (ii) on page 539 which appear in the commutative case: (i) is literally the same, and as for (ii), note that if A is commutative, then the category of A-bimodules is equivalent to the category of quasicoherent sheaves on $X \times X$, where $X = \operatorname{Spec} A$. By a famous theorem of Serre, this category has finite homological dimension if and only if X is smooth, and this dimension is equal to $\dim(X \times X) = 2 \dim X$.

6. Applications to Hodge Theory

To finish the paper, we return to the original problem discussed in the Introduction: the degeneration of the Hodge-to-de Rham spectral sequence. On the surface of it, Theorem 5.4 is strong enough so that one can apply the method of Deligne and Illusie in the non-commutative setting. However, it has one fault. While in the commutative case we are dealing with an *algebraic variety* X, Theorem 5.4 is only valid for an associative *algebra*. In particular, were we to try to deduce the classical Cartier isomorphism (1.2) from Theorem 5.4, we would only get it for affine algebraic varieties. In itself, it might not be completely meaningless. However, the commutative Hodge-to-de Rham degeneration is only true for a *smooth* and *proper* algebraic variety X—and a variety of dimension ≥ 1 cannot be proper and affine at the same time. The general non-commutative degeneration statement also requires some versions of properness, and in the affine setting, this reduces to requiring that the algebra A is finite-dimensional over the base field. A degeneration statement for such algebras, while not as completely trivial as its commutative version, is not, nevertheless, very exciting.

Fortunately, the way out of this difficulty has been known for some time; roughly speaking, one should pass to the level of *derived categories*—after which all varieties, commutative and non-commutative, proper or not, become essentially affine.

More precisely, one first notices that Hochschild homology of an associative algebra A is *Morita-invariant*—that is, if B is a different associative algebra such that the category B-mod of B-modules is equivalent to the category A-mod of A-modules, then $HH_\bullet(A) \cong HH_\bullet(B)$. The same is true for cyclic and periodic cyclic homology, and for Hochschild cohomology $HH^\bullet(A)$. In fact, B. Keller has shown in [**Kel99**] how to construct $HC_\bullet(A)$ and $HH_\bullet(A)$ starting directly from the abelian category A-mod, without using the algebra A at all.

Moreover, Morita-invariance holds on the level of derived categories: if there exists a left-exact functor $F : A\text{-}\mathsf{mod} \to B\text{-}\mathsf{mod}$ such that its derived functor is an equivalence of the derived categories $\mathcal{D}(A\text{-}\mathsf{mod}) \cong \mathcal{D}(B\text{-}\mathsf{mod})$, then $HH_\bullet(A) \cong HH_\bullet(B)$, and the same is true for $HC_\bullet(-)$, $HP_\bullet(-)$, and $HH^\bullet(-)$.

Unfortunately, one cannot recover $HH_\bullet(A)$ and other homological invariants directly from the derived category $\mathcal{D}(A\text{-}\mathsf{mod})$ considered as a triangulated category—the notion of a triangulated category is too weak. One has to fix some "enhancement" of the triangulated category structure. At present, it is not clear what is the most convenient choice among several competing approaches. In practice, however, every "natural" way to construct a triangulated category $\mathcal{D}$ also allows one to equip it with all possible enhancements, so that the Hochschild homology $HH_\bullet(\mathcal{D})$ and other homological invariants can be defined.

As long as we work over a fixed field, probably the most convenient of those "natural" ways is provided by the DG algebra techniques. For every associative DG algebra $A^\bullet$ over a field k, one defines $HH_\bullet(A^\bullet)$, $HC_\bullet(A^\bullet)$, $HP_\bullet(A^\bullet)$, and $HH^\bullet(A^\bullet)$ in the obvious way, and one shows that if two DG algebras $A^\bullet$, $B^\bullet$ have equivalent triangulated categories $\mathcal{D}(A^\bullet\text{-}\mathsf{mod})$, $\mathcal{D}(B^\bullet\text{-}\mathsf{mod})$ of DG modules, then all their homological invariants such as $HH_\bullet(-)$ are isomorphic. Moreover, the DG algebra approach is versatile enough to cover the case of non-affine schemes. Namely, one can show that for every quasiprojective variety X over a field k, there

exists a DG algebra $A^\bullet$ over k such that $\mathcal{D}(A^\bullet\text{-}\mathsf{mod})$ is equivalent to the derived category of coherent sheaves on X. Then $HH_\bullet(A^\bullet)$ is the same as the Hochschild homology of the category of coherent sheaves on X, and the same is true for the other homological invariants—in particular, if X is smooth, we have

$$HH_i(A^\bullet) \cong \bigoplus_j H^j(X, \Omega_X^{i+j}),$$

and $HC_\bullet(A^\bullet)$ is similarly expressed in terms of the de Rham cohomology groups of X. It is in this sense that all the varieties become affine in the "derived non-commutative" world. We note that in general, although X is the usual commutative algebraic variety, one cannot ensure that the algebra $A^\bullet$ which appears in this construction is also commutative.

Thus for our statement on the Hodge-to-de Rham degeneration, we use the language of associative DG algebras. The formalism we use is mostly due to B. Toën; the reader will find a good overview in [**TV05**, Section 2], and also in B. Keller's talk [**Kel06**] at ICM Madrid.

Definition 6.1. Assume given a DG algebra $A^\bullet$ over a field k.

(i) $A^\bullet$ is *compact* if it is perfect as a complex of k-vector spaces.
(ii) $A^\bullet$ is *smooth* if it is perfect as the diagonal DG bimodule over itself.

By definition, a DG $B^\bullet$-module $M_\bullet$ over a DG algebra $B^\bullet$ is perfect if it is a compact object of the triangulated category $\mathcal{D}(B^\bullet)$ in the sense of category theory—that is, we have

$$Hom(M_\bullet, \varinjlim N_\bullet) = \varinjlim \mathrm{Hom}(M_\bullet, N_\bullet)$$

for any filtered inductive system $N_\bullet \in \mathcal{D}(B^\bullet)$. It is an easy exercise to check that compact objects in the category k-Vect are precisely the finite-dimensional vector spaces, so that a complex of k-vector spaces is perfect if and only if its homology is trivial outside of a finite range of degrees, and all the non-trivial homology groups are finite-dimensional k-vector spaces. In general, there is a theorem which says that a DG module $M_\bullet$ is perfect if and only if it is a retract—that is, the image of a projector—of a DG module $M'_\bullet$ which becomes a free finitely-generated $B^\bullet$-module if we forget the differential. We refer the reader to [**TV05**] for exact statements and proofs. We note only that if a DG algebra $A^\bullet$ describes an algebraic variety X—that is, $\mathcal{D}(A^\bullet) \cong \mathcal{D}(X)$—that $A^\bullet$ is compact if and only if X is proper, and $A^\bullet$ is smooth if and only if X is smooth (for smoothness, one uses Serre's Theorem mentioned in the end of Section 5).

Theorem 6.2. *Assume given an associative DG algebra $A^\bullet$ over a field K of characteristic* 0. *Assume that $A^\bullet$ is smooth and compact. Moreover, assume that $A^\bullet$ is concentrated in non-negative degrees. Then the Hodge-to-de Rham spectral sequence*

$$HH_\bullet(A^\bullet)[u] \Rightarrow HC_\bullet(A^\bullet)$$

of (1.5) *degenerates at first term.*

In this theorem, we have to require that $A^\bullet$ is concentrated in non-negative degrees. This is unfortunate but inevitable in our approach to the Cartier map, which in the end boils down to Lemma 4.1—whose statement is obviously incompatible with any grading one might wish to put on the vector space V. Thus our construction of the Cartier isomorphism does not work at all for DG algebras. We

circumvent this difficulty by passing from DG algebras to *cosimplicial* algebras—that is, associative algebras $\mathcal{A} \in \mathrm{Fun}(\Delta, K)$ in the tensor category $\mathrm{Fun}(\Delta, K)$—for which one can construct the Cartier map "pointwise" (it is the passage from DG to cosimplicial algebras which forces us to require $A^i = 0$ for negative i). This occupies the larger part of [**Kal08**, Subsection 5.2], to which we refer the reader. Here we will only quote the end result.

PROPOSITION 6.3. *Assume given a smooth and compact DG algebra $A^\bullet$ over a finite field k of characteristic $p = \mathsf{char}\, k$. Assume that $A^\bullet$ is concentrated in non-negative degrees, and that, moreover,*

(i) *$A^\bullet$ can be lifted to a flat DG algebra over the ring $W_2(k)$ of second Witt vectors of the field k, and*
(ii) *$HH^i(A, A) = 0$ when $i \geq 2p$.*

Then there exists an isomorphism

$$C^{-1} : HH_\bullet(A^\bullet)((u)) \cong HP_\bullet(A^\bullet),$$

and the Hodge-to-de Rham spectral sequence (1.5) *for the DG algebra $A^\bullet$ degenerates at first term.*

As in the commutative case of [**DI87**], degeneration follows immediately from the existence of the Cartier isomorphism C^{-1} for dimension reasons. The construction of the map C^{-1} essentially repeats what we did in Section 5 in the framework of cosimplicial algebras, with a lot of technical nuisance because of the need to ensure the convergence of various spectral sequences, see [**Kal08**, Subsection 5.3]. To deduce Theorem 6.2, one uses the standard technique of the reduction to positive characteristic, just as in the commutative case; this is made possible by the following beautiful theorem due to B. Toën [**Toë08**].

THEOREM 6.4 ([**Toë08**]). *Assume given a smooth and compact DG algebra $A^\bullet$ over a field K. Then there exists a finitely generated subring $R \subset K$ and a DG algebra $A^\bullet_R$, smooth and compact over R, such that $A^\bullet \cong A^\bullet \otimes_R K$.*

We note that this result does not require the algebra $A^\bullet$ to be concentrated in non-negative degrees. We expect that neither does our Theorem 6.2, but so far, we could not prove it—the technical difficulties seem to be much too severe.

References

[Con83] A. Connes, *Cohomologie cyclique et foncteurs* Ext^n, C. R. Acad. Sci. Paris Sér. I Math. **296** (1983), no. 23, 953–958. MR 777584 (86d:18007)

[DI87] P. Deligne and L. Illusie, *Relèvements modulo p^2 et décomposition du complexe de de Rham*, Invent. Math. **89** (1987), no. 2, 247–270. MR 894379 (88j:14029)

[FT83] B. L. Feĭgin and B. L. Tsygan, *Cohomology of Lie algebras of generalized Jacobi matrices*, Funktsional. Anal. i Prilozhen. **17** (1983), no. 2, 86–87, English translation in: Functional Anal. Appl. 17 (1983), no. 2, 153–155. MR 705056 (85c:17008)

[FT87] ———, *Additive K-theory*, K-theory, arithmetic and geometry (Moscow, 1984–1986), Lecture Notes in Math., vol. 1289, Springer, Berlin, 1987, pp. 67–209. MR 923136 (89a:18017)

[Gro66] A. Grothendieck, *On the de Rham cohomology of algebraic varieties*, Inst. Hautes Études Sci. Publ. Math. (1966), no. 29, 95–103. MR 0199194 (33 #7343)

[HKR62] G. Hochschild, B. Kostant, and A. Rosenberg, *Differential forms on regular affine algebras*, Trans. Amer. Math. Soc. **102** (1962), 383–408. MR 0142598 (26 #167)

[Kal05] D. Kaledin, *Non-commutative Cartier operator and Hodge-to-de Rham degeneration*, 2005, math.AG/0511665.

[Kal07] ———, *Cyclic homology with coefficients*, 2007, math.KT/0702068.
[Kal08] D. Kaledin, *Non-commutative Hodge-to-de Rham degeneration via the method of Deligne-Illusie*, Pure Appl. Math. Q. **4** (2008), no. 3, part 2, 785–875. MR 2435845
[Kel99] B. Keller, *On the cyclic homology of exact categories*, J. Pure Appl. Algebra **136** (1999), no. 1, 1–56. MR 1667558 (99m:18012)
[Kel06] ———, *On differential graded categories*, International Congress of Mathematicians. Vol. II, Eur. Math. Soc., Zürich, 2006, pp. 151–190. MR 2275593 (2008g:18015)
[KS06] M. Kontsevich and Y. Soibelman, *Notes on A-infinity algebras, A-infinity categories and non-commutative geometry, I*, 2006, math.RA/0606241.
[Lod98] J.-L. Loday, *Cyclic homology*, second ed., Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 301, Springer-Verlag, Berlin, 1998. MR 1600246 (98h:16014)
[LQ83] J.-L. Loday and D. Quillen, *Homologie cyclique et homologie de l'algèbre de Lie des matrices*, C. R. Acad. Sci. Paris Sér. I Math. **296** (1983), no. 6, 295–297. MR 695381 (85d:17010)
[Toë08] B. Toën, *Anneaux de définition des dg-algèbres propres et lisses*, Bull. Lond. Math. Soc. **40** (2008), no. 4, 642–650. MR 2441136
[TV05] B. Toën and M. Vaquié, *Moduli of objects in dg-categories*, 2005, math.AG/0503269.

Steklov Mathematical Institute, Gubkina str. 8, 119991, Moscow, Russia
E-mail address: kaledin@mi.ras.ru

This book is based on survey lectures given at the 2006 Clay Summer School on Arithmetic Geometry at the Mathematics Institute of the University of Göttingen. Intended for graduate students and recent Ph.D.'s, this volume will introduce readers to modern techniques and outstanding conjectures at the interface of number theory and algebraic geometry.

The main focus is rational points on algebraic varieties over non-algebraically closed fields. Do they exist? If not, can this be proven efficiently and algorithmically? When rational points do exist, are they finite in number and can they be found effectively? When there are infinitely many rational points, how are they distributed?

For curves, a cohesive theory addressing these questions has emerged in the last few decades. Highlights include Faltings' finiteness theorem and Wiles's proof of Fermat's Last Theorem. Key techniques are drawn from the theory of elliptic curves, including modular curves and parametrizations, Heegner points, and heights.

The arithmetic of higher-dimensional varieties is equally rich, offering a complex interplay of techniques including Shimura varieties, the minimal model program, moduli spaces of curves and maps, deformation theory, Galois cohomology, harmonic analysis, and automorphic functions. However, many foundational questions about the structure of rational points remain open, and research tends to focus on properties of specific classes of varieties.

ISBN 978-0-8218-4476-2
9 780821 844762
CMIP/8

www.ams.org
www.claymath.org