KATHOLIEKE UNIVERSITEIT LEUVEN

Faculteit Wetenschappen

Departement Wiskunde

point counting on nondegenerate curves

Wouter Castryck

Promotor :
Prof. dr. Jan DENEF

Proefschrift ingediend tot het
behalen van de graad van Doctor
in de Wetenschappen

14 november 2006

Dankwoord

Boven alles en iedereen wil ik mijn promotor, prof. Jan Denef, bedanken. Zijn ontembare enthousiasme en zijn heldere manier van uitleggen maken het een waar voorrecht om met hem te mogen samenwerken. Hij bracht dit interessante thesisonderwerp aan en gaf me de kans om deel te nemen aan buitenlandse congressen, workshops en zomerscholen. Maar vooral de vele uren, dagen en weken die hij geïnvesteerd heeft in het napluizen van bewijzen, het aanbrengen van nieuwe ideeën en het geduldig luisteren naar mijn uiteenzettingen apprecieer ik enorm. Dat laatste geldt ook voor de inspanningen van dr. Frederik Vercauteren. Het voorgelegde thesismateriaal stelde zich bij momenten erg weerbarstig op, en zonder hun bijdrage en inzichten was ik er wellicht nooit uitgeraakt.

Ook de andere leden van de jury, prof. Arno Kuijlaars, prof. Tanja Lange, prof. Jan Stienstra, prof. Joost van Hamel en prof. Wim Veys, wil ik bedanken voor hun nuttige commentaar en voor de tijd en moeite die ze staken in het beoordelen van de thesis.

De afgelopen drie jaar had ik het geluk mijn bureau te mogen delen met Hendrik, die ook in het point-counting-schuitje vertoeft. Ik wil hem bedanken voor zijn hulp bij computerperikelen, voor het nauwgezet opvolgen en bijvullen van mijn persoonlijke agenda ('Hebt ge da verslag al gemaakt?', 'Hebt gij nu geen les?', ...), voor de vele ambiance-platen die nu mijn harde schijf sieren, maar vooral voor onze dagelijkse dosis wiskundige discussies. Veel van de inzichten die ik op die manier heb verworven zitten impliciet in deze thesis bevat.

Ook één of enkele deuren verder kon ik altijd terecht met mijn wiskundige en minder wiskundige vraagjes. Bedankt Jan, voor de spelletjesavonden, Ann, voor de babbels en de tiramisu, Dirk, voor de voorkant van deze thesis, Lise, voor de relativerende opmerkingen, Filip, als drijvende kracht achter de *marginalen bar*, Tim, het gewauwel tijdens je bierproefavonden heeft nog niet tot nieuwe stellingen geleid maar dat kan niet uitblijven, An, voor het gekoekoek in de gang, Wannes, voor de culturele noot, en Bart, om ons departement met de muurklimmicrobe te besmetten. Jullie zorgden ervoor dat het steeds aangenaam vertoeven was in blok B.

Langs de financiële kant wil ik het F.W.O. (Fonds voor Wetenschappelijk Onderzoek – Vlaanderen) bedanken, maar zeker ook Bea, onze secretaresse die vaak de strijd met Goliath moest aanbinden om alle onkostenvergoedingen geregeld te krijgen.

Ik wil van de gelegenheid gebruik maken om ook op niet-wiskundig vlak een aantal mensen in dit dankwoord te vermelden. De kans is klein dat ze deze thesis van naaldje tot draadje zullen lezen, maar toch hebben ook zij wezenlijk bijgedragen tot het totstandkomen ervan. In de eerste plaats heb ik het over mijn ouders, omdat het de afgelopen vier jaar steeds weer een warm thuiskomen was, en omdat ze me vrij laten in mijn keuzes en me altijd steunen in wat ik doe. Mijn zus wil ik bedanken voor de al even warme tussenstops in Gent.

Na de taak komt het vermaak: daar zorgden sinds het begin de uitwijkelingen van het Geestcollege voor, die intussen bijna allemaal Leuven inruilden voor een nieuwe thuishaven. Ik wil hen bedanken voor de gezellige avonden in Café Commerce of Bar del Sol, de zeldzame quiz-overwinningen en de technisch verfijnde voetbalpartijtjes.

Naast de ouderlijke warmte waren er tenslotte nog tal van andere redenen waarom ik in het weekend graag naar Ichtegem terugkeerde: de KSA en zijn afspinsels The Flying Ducks en VC The Eiciediecies, het jeugdhuis (en het revue-comité in het bijzonder), de ping-pongclub van Eernegem, maar vooral de grote groep fijne mensen die ik via deze verenigingen heb mogen leren kennen.

Wouter Castryck

Contents

Chapter 1

Introduction

In this introductory chapter, we gather some facts on the problem of algorithmically determining the number of solutions to a system of equations over a finite field $\mathbb{F}_q$. As we will see in Subsection 1.1.2, this number has a deep geometric interpretation that was first described by Weil in 1949, leading to the related problem of computing the zeta function of an algebraic variety over $\mathbb{F}_q$. In Section 1.2, we briefly sketch the state of the art anno 2006. Section 1.3 contains an even briefer overview of the practical and theoretical applications of point counting. We conclude by situating this thesis in the ongoing research process.

Throughout this chapter, if $\mathbb{F}$ is a field, $\overline{\mathbb{F}}$ denotes a fixed algebraic closure. For any $n \in \mathbb{N}\setminus\{0\}$, the symbols $x_1, \ldots, x_n$ form a set of formal variables. If $n = 2$, we will write x and y instead of x_1 and x_2. When dealing with characteristic polynomials or zeta functions, we will use the variable t. Complexity estimates are made using Landau's big-Oh symbol O and measure the number of bit operations (time) or bits needed to stock the intermediate results (space). Very often, we will use the soft-Oh notation $\widetilde{O}$, neglecting factors that are logarithmic in the input size of the algorithm for which the complexity estimates are made. We will often implicitly use that field or ring operations can be done in quasi-linear time (using e.g. the Schönhage-Strassen multiplication method [96]).

1.1 The point counting problem

1.1.1 Problem statement and first 'ad hoc' examples

Given a finite field $\mathbb{F}_q$ (with q elements) and a set of polynomials $f_1, f_2, \ldots, f_s \in \mathbb{F}_q[x_1, \ldots, x_n]$ (for some $s \in \mathbb{N}_0$), one can ask how many solutions in $\mathbb{F}_q^n$ there are to the system of equations

$$\mathcal{S} : \begin{cases} f_1 &= 0 \\ f_2 &= 0 \\ \vdots & \\ f_s &= 0. \end{cases}$$

This question, being interesting in its own right, naturally arises when dealing with certain important problems appearing in pure mathematics and computer science. We refer to Section 1.3 for some details on this.

Of course, for a concrete system $\mathcal{S}$ one can naively compute the number of solutions by checking all q^n possibilities. However, as q and n get bigger this soon becomes an impossible task, even for a computer. Therefore, one needs to come up with smarter methods. Along with this goes the more conceptual question of what the number of solutions is actually determined by. For almost 150 years, that is until Weil conjectured the existence of a decent cohomology theory for varieties over finite fields (see Subsection 1.1.2), this was not well understood; even nowadays, many questions remain unanswered.

Note that decompositions of the type

$$N(\{f_1, f_2\}) = N(\{f_1\}) + N(\{f_2\}) - N(\{f_1 f_2\}) \tag{1.1}$$

or alternatively

$$N(\{f_1, f_2\}) = N(\{f_1^2 - a f_2^2\}) \tag{1.2}$$

(where N denotes the number of common solutions in $\mathbb{F}_q^n$ and $a \in \mathbb{F}_q$ is non-square) in principle reduce the problem to the case $s = 1$. In practice, however, these reductions are rarely simplifications because (absolutely) reducible polynomials are in general not easy to deal with.

We begin with the following classical example.

1.1 Example (linear systems) If $\mathcal{S}$ is a system consisting of $m \in \mathbb{N} \setminus \{0\}$ linear equations, determining its number of solutions is very easy. Indeed, using Gaussian elimination one can always rewrite $\mathcal{S}$ as

$$\left\{\begin{array}{ccccccccccccc}
x_1 & + & a_{12}x_2 & + & \dots & + & a_{1k}x_k & + & \dots & + & a_{1n}x_n & = & b_1 \\
 & & x_2 & + & \dots & + & a_{2k}x_k & + & \dots & + & a_{2n}x_n & = & b_2 \\
 & & & & \ddots & & \vdots & & & & \vdots & & \vdots \\
 & & & & & & x_k & + & \dots & + & a_{kn}x_n & = & b_k \\
 & & & & & & & & & & 0 & = & b_{k+1} \\
 & & & & & & & & & & \vdots & & \vdots \\
 & & & & & & & & & & 0 & = & b_m
\end{array}\right.$$

where k is the rank of the system. If $b_{k+1} = \cdots = b_m = 0$ there are q^{n-k} solutions, otherwise there are no solutions. In total one needs $\widetilde{O}((\log q)\max\{m,n\}^3)$ steps to obtain this answer. Note that naive counting would take $\widetilde{O}(mnq^n)$ steps.

As a reverse application to the decomposition given in (1.1), we can compute the number of solutions to a product of m linearly independent forms

$$(a_{11}x_1 + \cdots + a_{1n}x_n) \cdots (a_{m1}x_1 + \cdots + a_{mn}x_n) = 0.$$

Using the above, we obtain that there are

$$mq^{n-1} - \binom{m}{2} q^{n-2} + \cdots + (-1)^{m-1} q^{n-m} = q^{n-m}(q^m - (q-1)^m)$$

solutions. ■

Unfortunately, this is already where the easy part of the story ends. Linear systems are – up to our knowledge – the only large class of systems for which a simple and efficient point counting algorithm is known[1]. As soon as $\mathcal{S}$ is allowed to contain higher degree equations, the situation becomes much more particular and (in most cases) complicated. We give some examples of 'ad hoc' techniques that are sometimes useful.

1.2 Example (manipulating the equations) Often, one can manipulate the system in a way that preserves the number of solutions (or at least so that one can control the gain or loss), such that the equations become easier to handle. For instance, consider the single equation

$$y^2 = x^2 + x + 1.$$

Then we can substitute $y \leftarrow y + x$ as to obtain

$$(2y - 1)x = 1 - y^2.$$

Now it is clear that for every value of y there is a unique corresponding value of x, except if $2y = 1$. We find that there are q solutions if q is even, $2q - 1$ solutions if q is a power of 3, and $q - 1$ solutions in the other cases.

A somewhat more advanced example is the following. We will use it in Example 1.4 to give a new proof of Gauss' quadratic reciprocity law [12]. For any odd $n \in \mathbb{N}$, let N_n denote the number of solutions in $\mathbb{F}_q^n$ to the alternating equation

$$x_1^2 - x_2^2 + x_3^2 - x_4^2 + \cdots + x_n^2 = 1.$$

Since the characteristic 2 case is straightforward (for every value of $x_1, \ldots, x_{n-1}$ there is a unique corresponding x_n-value, so the number of solutions is q^{n-1}), we suppose that q is odd. Then if we substitute $x_1 \leftarrow x_1 + x_2$, we find that N_n equals the number of solutions to

$$x_1^2 + x_3^2 - x_4^2 + \cdots + x_n^2 - 1 = -2x_1x_2.$$

Thus for any value of x_1 different from 0, we have q^{n-2} solutions. If $x_1 = 0$ there are no solutions except if $x_3^2 - x_4^2 + \cdots + x_n^2 = 1$ (which happens in N_{n-2} cases): then all q possible values of x_2 do the job. In conclusion:

$$N_n = q^{n-2}(q - 1) + qN_{n-2}.$$

One finds that $N_n = q^{n-1} + q^{\frac{n-1}{2}}(N_1 - 1) = q^{n-1} + q^{\frac{n-1}{2}}$. ■

[1] Of course, this statement highly depends on what is meant by 'large', 'efficient' and 'simple'.

1.3 Example (partitioning the ambient space) Consider a single equation of the form

$$y^2 = f(x), \qquad f(x) \in \mathbb{F}_q[x]$$

and suppose that $f(x)$ is an odd polynomial (that is, $f(-x) = -f(x)$). If $q \equiv 3 \bmod 4$, then the number of solutions is q. Indeed, we can write $\mathbb{F}_q$ as a disjoint union

$$\{a \mid f(a) = 0\} \sqcup \left\{a \;\middle|\; f(a) \in (\mathbb{F}_q)^2 \setminus \{0\}\right\} \sqcup \left\{a \;\middle|\; f(a) \in \mathbb{F}_q \setminus (\mathbb{F}_q)^2\right\}$$

where $(\mathbb{F}_q)^2$ denotes the set of squares in $\mathbb{F}_q$. Now $a \mapsto -a$ defines a bijection between the latter two sets because $f(x)$ is odd and -1 is not a square (since $q \equiv 3 \bmod 4$). Thus $q = \#\mathbb{F}_q$ equals

$$\#\{a \mid f(a) = 0\} \;+\; 2 \cdot \#\left\{a \;\middle|\; f(a) \in (\mathbb{F}_q)^2 \setminus \{0\}\right\},$$

which is precisely the number of solutions to $y^2 = f(x)$. ■

1.4 Example (using multiplicative characters) Until the 1920's, the most powerful point counting techniques made use of multiplicative characters. A typical example is the following. Let $\mathbb{F}_q$ be a finite field and consider the equation

$$x_1^2 + x_2^2 + x_3^2 = 1.$$

If the field characteristic is 2, then for every choice of x_1 and x_2 there is a corresponding x_3-value, and the number of solutions is q^2. So suppose q is odd. Let $\chi : \mathbb{F}_q \to \{-1, 0, 1\}$ be the quadratic character on $\mathbb{F}_q$. One then verifies that the number of solutions to the above equation equals

$$\sum_{t_1+t_2+t_3=1} N(\{x_1^2 - t_1\})N(\{x_2^2 - t_2\})N(\{x_3^2 - t_3\})$$

$$= \sum_{t_1+t_2+t_3=1} (1 + \chi(t_1))\,(1 + \chi(t_2))\,(1 + \chi(t_3))\,.$$

Using that $\sum_{t\in\mathbb{F}_q} \chi(t) = 0$ this simplifies to

$$q^2 + \sum_{t_1+t_2+t_3=1} \chi(t_1t_2t_3) = q^2 + \sum_{t_1,t_2\in\mathbb{F}_q} \chi\left(t_1t_2(1 - t_1 - t_2)\right)$$

$$= q^2 + \sum_{t_1\in\mathbb{F}_q} \sum_{t_2\neq 1-t_1} \chi\left(\frac{t_1t_2}{1 - t_1 - t_2}\right).$$

Now if $t_1 \neq 0, 1$ the map

$$\mathbb{F}_q \setminus \{1 - t_1\} \to \mathbb{F}_q \setminus \{-t_1\} : y \mapsto \frac{t_1 y}{1 - t_1 - y}$$

is a bijection, so again using that quadratic characters sum up to zero, the number of solutions is

$$q^2 + (q-1)\chi(-1) - \sum_{t_1 \neq 0,1} \chi(-t_1) = q^2 + (q-1)\chi(-1) + \chi(-1) = q^2 + q\chi(-1).$$

We conclude that there are $q^2 + (-1)^{\frac{q-1}{2}} q$ solutions to our equation.

Using higher degree characters, much of the above can be generalized to arbitrary diagonal equations in any number of variables. This leads to the classical theory of Gauss and Jacobi sums, we refer to the book of Ireland and Rosen [56] for more details.

As announced in Example 1.2, we conclude with a proof of the quadratic reciprocity law. Let $\mathbb{F}_q$ be a finite prime field with $q \neq 2$ elements. Let p be another odd prime number and consider the equation

$$x_1^2 - x_2^2 + x_3^2 - x_4^2 + \cdots + x_p^2 = 1.$$

From Example 1.2, we know that there are $q^{p-1} + q^{\frac{p-1}{2}}$ solutions, which is $\equiv 1 + \left(\frac{q}{p}\right) \bmod p$. On the other hand, using the same arguments as above, we know that there are

$$q^{p-1} + \sum_{t_1+\cdots+t_p=1} \left(\frac{-1}{q}\right)^{\frac{p-1}{2}} \left(\frac{t_1 t_2 \cdots t_p}{q}\right)$$

$$\equiv 1 + (-1)^{\frac{p-1}{2}\frac{q-1}{2}} \sum_{t_1+\cdots+t_p=1} \left(\frac{t_1 t_2 \cdots t_p}{q}\right) \mod p$$

solutions. Now modulo p, the only $(t_1, \ldots, t_p)$ contributing to the sum is where all t_i are equal: by shifting, one sees that the other tuples appear in groups of size p. Therefore, the number of solutions mod p equals $1 + (-1)^{\frac{p-1}{2}\frac{q-1}{2}} \left(\frac{p}{q}\right)$. The reciprocity law follows. ■

1.5 Example In this last example we consider a field $\mathbb{F}_p$ with p elements, where p is an odd prime, and an equation of the form

$$y^2 = x^3 + cx, \qquad c \in \mathbb{F}_p^\times.$$

If $p \equiv 3 \bmod 4$, this example fits the situation of Example 1.3, so the number of solutions is precisely p. If $p \equiv 1 \bmod 4$, a well-known conjecture by Fermat (first proven by Euler) states that we can write

$$p = a^2 + b^2, \qquad a, b \in \mathbb{Z}.$$

It is clear that we can always assume $a \equiv 1 \bmod 4$. Using that the ring of Gaussian integers $\mathbb{Z}[\mathbf{i}]$ is a UFD, one obtains that a is uniquely determined by this condition. A theorem by Gauss then states that

$$\binom{\frac{p-1}{2}}{\frac{p-1}{4}} \equiv 2a \bmod p$$

(see [54] for a proof).

Now let N denote the number of solutions to $y^2 = x^3 + cx$. Then one immediately checks that $N = \sum_{x=0}^{p-1} \left(1 + \left(\frac{x^3+cx}{p}\right)\right)$, hence

$$N \equiv \sum_{x=0}^{p-1} (x^3 + cx)^{\frac{p-1}{2}} \bmod p.$$

Using that

$$\sum_{x=0}^{p-1} x^j \equiv \begin{cases} 0 \mod p & \text{if } p-1 \nmid j \\ -1 \mod p & \text{if } p-1 \mid j \end{cases}$$

we conclude that N is congruent modulo p to minus the coefficient of x^{p-1} in the expansion of the polynomial $(x^3 + cx)^{\frac{p-1}{2}}$, which is

$$c^{\frac{p-1}{4}} \binom{\frac{p-1}{2}}{\frac{p-1}{4}}.$$

By Gauss' theorem we finally get that

$$N \equiv -2ac^{\frac{p-1}{4}} \bmod p.$$

The above can be turned into a very efficient point counting algorithm for equations of the type $y^2 = x^3 + cx$ over some big prime field $\mathbb{F}_p$. Indeed, if $p \equiv 3 \bmod 4$, just output p. If $p \equiv 1 \bmod 4$, we can use Hasse's bound: we know that the number of solutions must lie in the interval $[p - 2\sqrt{p}, p + 2\sqrt{p}]$ (see [100] for a proof; this also follows from Theorem 1.8 below). So if $p > 16$ it suffices to retrieve the number of solutions modulo p. Furthermore a can be found very rapidly by using Euclid's algorithm in the ring of Gaussian integers $\mathbb{Z}[\mathbf{i}]$. All of this is summarized in the following:

INPUT: an odd prime number p and a $c \in \{0, \ldots, p-1\}$
OUTPUT: $N = \#\{(x, y) \in \mathbb{F}_p^2 \mid E : y^2 = x^3 + cx\}$

1. **if** $p \bmod 4 = 3$ **then output** p
2. **else if** $p = 5$ or $p = 13$ **then output** N by naive calculation
3. **else**
4. **repeat** $z \leftarrow (\text{Random}\{2, \ldots, p-1\})^{\frac{p-1}{2}}$
5. **until** $z^2 \equiv -1 \bmod p$
6. $a + b \cdot \mathbf{i} \leftarrow \text{GCD}(z + \mathbf{i}, p)$ (using Euclid's algorithm in $\mathbb{Z}[\mathbf{i}]$)
7. $a \leftarrow$ unique element in $\{\pm a, \pm b\}$ that is $\equiv 1 \bmod 4$
8. $N := -2ac^{\frac{p-1}{4}} \bmod p$ $(0 \le N < p)$
9. **if** $N \ge p - 2\sqrt{p}$ **then output** N **else output** $N + p$

The estimated time complexity is $O(\log^2 p)$, but note that there is a probabilistic step (4-5). The space needed is $O(\log p)$.

A similar method can be used to treat the case $y^2 = x^3 + c$, using unique factorization in $\mathbb{Z}[\omega]$ where ω is a primitive 3^{rd} root of unity. In fact, both methods are special cases of Cornacchia's algorithm, which can be used to determine the number of rational points on an elliptic curve $\overline{E}$ over a finite field $\mathbb{F}_q$ whenever the endomorphism ring $\text{End}(\overline{E})$ is known. We refer to Schoof's survey article [98] for more details.

Finally, we note that the above algorithm actually computes the so-called Hasse-Witt-invariant of the elliptic curve defined by $y^2 = x^3 + cx$. There is a higher genus analogue of this invariant, called the Hasse-Witt-matrix [75]. A very efficient algorithm that counts the number of solutions to an equation of the form $y^2 = x^5 + cx$ over a prime field, based on a Hasse-Witt-matrix computation, is presented in [37] (see also [11]). ■

All examples given above were treated using 'ad hoc' methods, adapted to the particular system that was considered, but unapplicable to most other situations. In the next section, we will discuss Weil's more structural approach to the point counting problem. Nevertheless, we want to emphasize that 'ad hoc' does not mean that only trivial mathematics is used, this should be clear from the last example.

1.1.2 Weil cohomologies and the zeta function

The big breakthrough came when people started to attack the problem from a more geometric point of view. That is, the system $\mathcal{S}$ defines an affine variety $\overline{X} \subset \mathbb{A}^n_{\overline{\mathbb{F}}_q}$ and the number of solutions to $\mathcal{S}$ is given by the number of $\mathbb{F}_q$-rational points on $\overline{X}$ (which explains why we talk about *point counting*). A crucial observation is that these $\mathbb{F}_q$-rational points are precisely those points of $\overline{X}$ that stay fixed under the action of Frobenius

$$\overline{\mathcal{F}}_q : \overline{X} \to \overline{X} : (p_1, \ldots, p_n) \mapsto (p_1^q, p_2^q, \ldots, p_n^q).$$

So in conclusion, a fancy way to say that we are interested in the number of solutions to a system of equations $\mathcal{S}$ over $\mathbb{F}_q$, is to say that we want to determine the number of points on a certain variety over $\overline{\mathbb{F}}_q$ that stay fixed under the action of Frobenius.

For varieties over $\mathbb{C}$, the problem of determining the number of fixed points of a given endomorphism f has been extensively studied before. Many interesting theorems have been proven, but most among are 'analytic' in nature, i.e. they make explicit use of the completeness of $\mathbb{C}$, which makes them unlikely to adapt to the $\overline{\mathbb{F}}_q$-situation. However, there is a result in algebraic topology by Lefschetz that drew Weil's attention in the 1940's.

1.6 Theorem (Lefschetz fixed point theorem) *Let X be a compact and $\mathbb{C}$-oriented manifold and let $f : X \to X$ be any continuous map. Define the* Lefschetz number

$$\Lambda_f = \sum_{i=0}^{\dim X} (-1)^i \,\text{Trace}(f^* \mid H^i(X)),$$

where $f^* \mid H^i(X)$ is the induced action of f on the i^{th} singular cohomology space of X. If $\Lambda_f \neq 0$, then f has a fixed point. Moreover, if f has only finitely many fixed points $x_1, \ldots, x_r$ there is a purely local way of assigning an index $I(x_j) \in \mathbb{Z}$ to each point so that $\Lambda_f = \sum_{j=1}^{r} I(x_j)$.

PROOF. See [47]. ■

Here, $\dim X$ denotes the topological dimension of X, which in case of an algebraic variety equals twice the algebraic dimension. Note that $\Lambda_{\mathrm{Id}_X} = \chi(X)$, so the Lefschetz number can be looked at as a 'generalized Euler characteristic'.

In common situations, the indices $I(x_j)$ are just 1, so that Λ_f is exactly the number of fixed points. In that case, the Lefschetz fixed point theorem expresses the number of fixed points in purely algebraic terms. Weil's hope was that the same would work for varieties over finite fields. That is: to any variety $\overline{X}$ living over a finite field $\mathbb{F}_q$, one should be able to

- associate finite-dimensional vector spaces $H^i(\overline{X})$ (for $i = 0, \ldots, 2\dim \overline{X}$) over some characteristic zero field
- have an induced action of Frobenius $\overline{\mathcal{F}}_q^*$ on each of these spaces,

so that the number of $\mathbb{F}_q$-rational points on $\overline{X}$ is given by

$$\sum_{i=0}^{2\dim \overline{X}} (-1)^i \operatorname{Trace}(\overline{\mathcal{F}}_q^* \mid H^i(\overline{X})). \tag{1.3}$$

At first sight, the existence of a cohomology theory for varieties over finite fields is nothing more than just something which is nice to hope for. But Weil had indications to believe that such a theory exists.

The Hasse-Weil zeta function and the Weil conjecture

The main argument was given by the properties of the *zeta function*, which is a generating series encoding the sequence $(N_k)_{k\in\mathbb{N}_0}$ where N_k is the number of $\mathbb{F}_{q^k}$-rational points on $\overline{X}$ (which is now *any* algebraic variety defined over $\mathbb{F}_q$):

$$Z_{\overline{X}}(t) = \exp\left(\sum_{k\in\mathbb{N}_0} N_k \frac{t^k}{k}\right) \quad \in \mathbb{Q}[[t]]. \tag{1.4}$$

To get in touch with the flavour, we give some elementary examples.

1.7 Example Since $-\log(1-t) = \sum_{k\in\mathbb{N}_0} \frac{t^k}{k}$, the zeta function of an $\mathbb{F}_q$-rational point is given by $1/(1-t)$.

Next, consider the curve in $\mathbb{A}^2_{\overline{\mathbb{F}}_q}$ defined by the equation $y^2 = x^2+x+1$, that was studied in Example 1.2. Suppose that the field characteristic differs from 2 and 3. Then we found that $N_k = q^k - 1$. Hence the zeta function becomes

$$\exp\left(\sum_{k\in\mathbb{N}_0} q^k \frac{t^k}{k} - \sum_{k\in\mathbb{N}_0} \frac{t^k}{k}\right) = \frac{1-t}{1-qt}.$$

Finally, consider the hypersurface $\overline{X}$ in $\mathbb{A}^3_{\overline{\mathbb{F}}_q}$ defined by $x_1^2+x_2^3+x_3^2=1$, and again suppose that q is odd. Then from Example 1.4 we know that in this case

$$N_k = q^{2k} + (-1)^{\frac{q^k-1}{2}} q^k = q^{2k} + \left((-1)^{\frac{q-1}{2}} q\right)^k$$

and

$$Z_{\overline{X}}(t) = \frac{1}{(1-q^2t)(1-(-1)^{\frac{q-1}{2}}qt)}.$$

■

In its general form, the zeta function was introduced by Weil in his famous 1949 paper '*Numbers of solutions of equations in finite fields*' [107], based on previous work by Artin, Schmidt, Hasse and others. In that paper, Weil wrote down a list of properties that he observed in his examples and which he believed to hold in general. His presumption became known as the *Weil conjecture*, and soon found its way among the most important research subjects in mathematics. It was eventually settled by Deligne in 1973 [21], using the machinery of Grothendieck (ℓ-adic cohomology, see further on). In 1949, Weil himself was already able to prove the conjecture in the $\dim \overline{X} = 1$ case.

1.8 Theorem (Weil conjecture) *Let $\overline{X}$ be a complete, smooth variety of dimension d, defined over a finite field $\mathbb{F}_q$ with q elements. Then*

1. *the zeta function $Z_{\overline{X}}(t)$ is a* ***rational function****; more precisely we have*

$$Z_{\overline{X}}(t) = \frac{P_1(t)P_3(t)\cdots P_{2d-1}(t)}{P_0(t)P_2(t)P_3(t)\cdots P_{2d}(t)}$$

   with $P_0(t) = 1-t$ and $P_{2d}(t) = 1-q^dt$;

2. *when enumerated properly, the degrees β_i of the $P_i(t)$ behave like* ***Betti numbers*** *in the following sense: the Euler characteristic $\chi(\overline{X})$ (the intersection number of the diagonal with itself in $\overline{X}\times\overline{X}$) equals the alternating sum $\sum_i (-1)^i\beta_i$; moreover, if $\overline{X}$ is obtained from a complete and smooth d-dimensional variety X defined over a number field K by reducing modulo a prime ideal $\mathfrak{p}\subset\mathcal{O}_K$, then the Betti numbers of X coincide with the β_i.*

3. *$Z_{\overline{X}}(t)$ satisfies a* ***functional equation***

$$Z_{\overline{X}}\left(\frac{1}{q^dt}\right) = \pm q^{\chi(\overline{X})d/2}t^{\chi(\overline{X})}Z_{\overline{X}}(t);$$

   if d is odd, then the sign is $+$ (the converse is not true);

4. *$Z_{\overline{X}}(t)$ satisfies an analogue of the* ***Riemann hypothesis****: the reciprocal roots of the $P_i(t)$ ($i=1,\ldots,2d-1$) are algebraic integers of absolute value $q^{i/2}$.*

PROOF. See Milne's course notes [82] for a detailed sketch of the proof. ■

The analogy between *4.* and the classical Riemann hypothesis is explained below, in the paragraph containing formula (1.6).

We will now briefly illustrate why Weil's observations strongly support the existence of a cohomology theory for varieties over finite fields, in which a Lefschetz fixed point theorem holds. Suppose that there indeed are finite-dimensional vector spaces $H^i(\overline{X})$ (over some field containing $\mathbb{Q}$), together with an induced action of Frobenius $\overline{\mathcal{F}}_q^*$ for which formula (1.3) expresses the number of $\mathbb{F}_q$-rational points on $\overline{X}$. In fact, if we suppose that the correspondence $\overline{\mathcal{F}}_q \mapsto \overline{\mathcal{F}}_q^*$ is functorial, it is then natural that for any $k \in \mathbb{N}_0$ the expression

$$\sum_{i=0}^{2\dim \overline{X}} (-1)^i \operatorname{Trace}(\overline{\mathcal{F}}_q^{*k} \mid H^i(\overline{X})) \tag{1.5}$$

equals N_k, the number of $\mathbb{F}_{q^k}$-rational points on $\overline{X}$. Now we have the following classical lemma.

1.9 Lemma (Newton's determinant formula) *Let V be a finite-dimensional vector space over a field $\mathbb{F}$ and let $\varphi : V \to V$ be an endomorphism. Then* $\det(\mathbb{I} - \varphi t)$ *equals*

$$\exp\left(\sum_{k=1}^{\infty} \operatorname{Trace}(\varphi^k \mid V) \frac{t^k}{k}\right).$$

PROOF. By moving on to some field extension $\mathbb{F}' \supset \mathbb{F}$ if necessary, we may assume that φ is given by a matrix in upper triangular form. The rest of the proof is straightforward. ■

If we combine (1.5) with (1.4) and use the determinant formula, we obtain

$$Z_{\overline{X}}(t) = \frac{\prod_{i=0}^{\dim \overline{X}-1} \det\left(\mathbb{I} - \left(\overline{\mathcal{F}}_q^* \mid H^{2i+1}(\overline{X})\right) t\right)}{\prod_{i=0}^{\dim \overline{X}} \det\left(\mathbb{I} - \left(\overline{\mathcal{F}}_q^* \mid H^{2i}(\overline{X})\right) t\right)}.$$

In particular, $Z_{\overline{X}}(t)$ is a rational function (*Weil conjecture 1*). Moreover, the degrees of the polynomials appearing in the numerator and the denominator are the Betti numbers of our cohomology, which supports *Weil conjecture 2*. As for the functional equation (*Weil conjecture 3*), this follows from an analogue of the Poincaré duality theorem identifying the action of $\overline{\mathcal{F}}_q^*$ on $H^i(\overline{X})$ with the action of a dual morphism $\overline{\mathcal{F}}_{q,*}$ on $H^{2n-i}(\overline{X})$. Then the zeta function can be rewritten as

$$Z_{\overline{X}}(t) = \frac{\prod_{i=0}^{\dim \overline{X}-1} \det\left(\mathbb{I} - \left(\overline{\mathcal{F}}_{q,*} \mid H^{2i+1}(\overline{X})\right) t\right)}{\prod_{i=0}^{\dim \overline{X}} \det\left(\mathbb{I} - \left(\overline{\mathcal{F}}_{q,*} \mid H^{2i}(\overline{X})\right) t\right)}.$$

Since $\overline{\mathcal{F}}_q$ is a finite morphism of degree $q^{\dim \overline{X}}$, it is natural to have $\overline{\mathcal{F}}_q^* \circ \overline{\mathcal{F}}_{q,*} = q^{\dim \overline{X}}$. One obtains that $\left(Z_{\overline{X}}(t)\right)^2 =$

$$\frac{\prod_{i=0}^{\dim \overline{X}-1} \det\left(\mathbb{I} - \left(\overline{\mathcal{F}}_q^* \mid H^{2i+1}(\overline{X})\right) t\right) \det\left(\mathbb{I} - \left(q^{\dim \overline{X}} \overline{\mathcal{F}}_q^{*-1} \mid H^{2i+1}(\overline{X})\right) t\right)}{\prod_{i=0}^{\dim \overline{X}} \det\left(\mathbb{I} - \left(\overline{\mathcal{F}}_q^* \mid H^{2i}(\overline{X})\right) t\right) \det\left(\mathbb{I} - \left(q^{\dim \overline{X}} \overline{\mathcal{F}}_q^{*-1} \mid H^{2i+1}(\overline{X})\right) t\right)}.$$

Evaluating this in $1/(q^{\dim \overline{X}} t)$, clearing up denominators and assuming that indeed $\chi(\overline{X}) = \sum (-1)^i \dim H^i(\overline{X})$ results in a functional equation. A cohomological interpretation of the last statement (*Weil conjecture 4*) is less obvious.

In conclusion, the Weil conjecture really breathes the existence of a decent cohomology theory for varieties over finite fields. Although he does not say so in his paper, this is what Weil had in mind when he wrote down the conjecture.

ℓ-adic and p-adic cohomologies

As already mentioned, Weil's presumption was right: by now, two such cohomology theories have been developed.

The first and most famous one is étale cohomology, which was developed by Grothendieck, being assisted by Artin[2], Verdier, Deligne and others, mainly in the 1960's. This is a very general type of cohomology, which specializes to the usual singular cohomology for varieties over $\mathbb{C}$. When applied to a variety over a finite field $\mathbb{F}_q$, étale cohomology serves as a Weil cohomology if one takes coefficients in the field of ℓ-adic numbers $\mathbb{Q}_\ell$ (where ℓ is a prime different from $p = \mathrm{char}(\mathbb{F}_q)$). This was used by Deligne to give the first complete proof of the Weil conjecture (although it was only the Riemann hypothesis part that was left open by Grothendieck). It lies far beyond the scope of this thesis to even define these ℓ-adic cohomology spaces. Instead we refer to Milne's book [81] or course notes [82]. We only mention that in the case of a curve $\overline{X}$, the first ℓ-adic cohomology space $H^1_{\mathrm{et}}(\overline{X}, \mathbb{Q}_\ell)$ is canonically related to the Tate module

$$T_\ell(\mathrm{Jac}(\overline{X})) = \lim_{\leftarrow} \mathrm{Jac}(\overline{X})[\ell^k]$$

of the jacobian variety of $\overline{X}$. Here $\mathrm{Jac}(\overline{X})[\ell^k]$ is the group of ℓ^k-torsion points and the inverse limit is taken over $k \in \mathbb{N} \setminus \{0\}$. Then $T_\ell(\mathrm{Jac}(\overline{X}))$ naturally becomes a module over $\mathbb{Z}_\ell$, the valuation ring of $\mathbb{Q}_\ell$. The correspondence between $H^1_{\mathrm{et}}(\overline{X}, \mathbb{Q}_\ell)$ and $T_\ell(\mathrm{Jac}(\overline{X}))$ is the underlying reason for the following theorem:

1.10 Theorem *If $\overline{X}$ is a complete and smooth genus g curve over a finite field $\mathbb{F}_q$, then its zeta function is of the form*

$$Z_{\overline{X}}(t) = \frac{P(t)}{(1-t)(1-qt)}$$

where $P(t)$ is a degree $2g$ polynomial and $t^{2g} P(\frac{1}{t})$ is the characteristic polynomial of q^{th} power Frobenius acting on $T_\ell(\mathrm{Jac}(\overline{X})) \otimes_{\mathbb{Z}_\ell} \mathbb{Q}_\ell$. Moreover, the number of $\mathbb{F}_q$-rational points on $\mathrm{Jac}(\overline{X})$ equals $P(1)$.

[2]Michael Artin, the son of Emil Artin, the 'inventor' of the zeta function.

PROOF. See Mumford's book [88] or Milne's course notes [79]. ■

On the other hand, in 1959 already, Dwork [29] was able to prove the rationality of the zeta function by implicitly making use of another type of cohomology: p-adic cohomology. This was investigated in more detail by Dwork himself, Grothendieck, Monsky and many others, but it wasn't until the work of Berthelot in the late 1970's that the right theoretical frame was found: rigid cohomology. In 2002, Kedlaya gave a completely p-adic proof of the Weil conjecture [63]. For smooth affine varieties, rigid cohomology allows a very explicit description that was already developed by Monsky and Washnitzer around 1970 [84, 85, 86, 105]. This so-called Monsky-Washnitzer cohomology is the main tool in this thesis, it will be discussed in detail in Chapter 4.

Relation with other zeta functions

Since half-way the 19^{th} century, zeta functions have become popular tools to introduce analysis in the study of objects arising in combinatorics, number theory, algebraic geometry, ... The mother of all zeta functions was studied by Riemann in 1859. He considered the meromorphic continuation $\zeta : \mathbb{C} \to \mathbb{C}$ of the function

$$\{s \in \mathbb{C} \,|\, \mathrm{Re}(s) > 1\} \to \mathbb{C} : s \mapsto \sum_{i=1}^{\infty} \frac{1}{n^s} = \prod_{p \text{ prime}} \frac{1}{1 - p^{-s}},$$

which is still an object of great interest (the Riemann hypothesis, that states that all solutions to $\zeta(s) = 0$ with $\mathrm{Re}(s) > 0$ satisfy $\mathrm{Re}(s) = \frac{1}{2}$, is probably the most famous open problem in mathematics). Since then, the term *zeta function* has been used to denominate several generating series that contain arithmetic and/or geometric information, the main one of course being the Hasse-Weil zeta function (at least in the context of this thesis). Other interesting examples include the Dedekind zeta function, the Igusa zeta function, the Poincaré-Serre series and the topological zeta function. Some of these are discussed in [22].

A fascinating aspect about zeta functions is that they seem to share many non-trivial properties, which in its turn seems to reveal the existence of one or more 'parent' zeta functions, from which the above are specializations. This is still very much an open domain, and a detailed treatment goes beyond the scope of this thesis. Nevertheless, below we give two interesting examples of such generalizing zeta functions. The first one is in the spirit of the classical Riemann ζ-function, whereas the second example is much more geometric in nature.

Consider the following function, that can be attached to any scheme X of finite type over $\mathbb{Z}$. It is given by

$$\zeta_X : \{s \in \mathbb{C} \,|\, \mathrm{Re}(s) > \dim X\} \to \mathbb{C} : s \mapsto \prod_{\mathfrak{p}} \frac{1}{1 - N_{\mathfrak{p}}^{-s}},$$

where the product runs over all closed points $\mathfrak{p}$ of X and where $N_{\mathfrak{p}}$ is the number of elements in the residue field of $\mathfrak{p}$. One sees that the Riemann ζ-function is the

continuation of $\zeta_{\operatorname{Spec}\mathbb{Z}}$, and that the Dedekind zeta function of a number field K is given by $\zeta_{\operatorname{Spec}\mathcal{O}_K}$. If $\overline{X}$ is a variety over a finite field $\mathbb{F}_q$ with q elements, then one can check that $\zeta_{\overline{X}}(s) = Z_{\overline{X}}(q^{-s})$. Note that part 4 of the Weil conjecture (Theorem 1.8) states that $\zeta_{\overline{X}}$ has its poles in

$$\left\{ s \in \mathbb{C} \;\middle|\; \operatorname{Re}(s) \in \left\{ 0, 1, 2, \ldots, \dim \overline{X} \right\} \right\}$$

and its zeroes in

$$\left\{ s \in \mathbb{C} \;\middle|\; \operatorname{Re}(s) \in \left\{ \frac{1}{2}, \frac{3}{2}, \ldots, \frac{\dim \overline{X} - 1}{2} \right\} \right\}, \tag{1.6}$$

which explains why this part is referred to as the Riemann hypothesis for varieties over finite fields. It has been conjectured that $\zeta_{\overline{X}}$ can be meromorphically extended to the whole of $\mathbb{C}$.

The second generalizing zeta function that we mention is a so-called *motivic* zeta function. Let $\mathbb{F}$ be a field and let $K_0(\operatorname{Var}_{\mathbb{F}})$ be the Grothendieck ring of varieties over $\mathbb{F}$. This is the free abelian group generated by the isomorphism classes $[X]$ of varieties over $\mathbb{F}$ and the relations $[X] = [X'] + [X \setminus X']$ for X' closed in X. It is turned into a ring by the product $[X][X'] = [X \times X']$. The map $X \mapsto [X]$ is often called the *universal Euler characteristic*. Then to a variety X over $\mathbb{F}$ one can attach

$$Z_X^{\mathrm{mot}}(T) = \sum_{n=0}^{\infty} [X^{(n)}] T^n \quad \in K_0(\operatorname{Var}_{\mathbb{F}})[[T]],$$

where $X^{(n)}$ is the n^{th} symmetric product of X. For a finite field $\mathbb{F}_q$ the map $K_0(\operatorname{Var}_{\mathbb{F}_q}) \to \mathbb{Z} : [X] \mapsto \#X(\mathbb{F}_q)$ is a well-defined morphism and one can check that $Z_X^{\mathrm{mot}}(T)$ specifies to the Hasse-Weil zeta function under this map. Some rationality results on the motivic zeta function have been proven. We refer to Kapranov's original paper [58] and to the notes of Denef and Loeser [22] for more details.

1.1.3 Problem statement revisited

In literature the term 'point counting' mostly refers to the problem of computing the zeta function of a given algebraic variety $\overline{X}$ over $\mathbb{F}_q$. A priori, this is a harder problem than just determining the number of $\mathbb{F}_q$-rational points on $\overline{X}$. But very often, computing the zeta function is the most effective way to obtain the number of points, because of the computational advantage one can pour out of the Weil conjecture. Moreover, the amount of extra information that is contained in the zeta function is huge. For instance, if $\overline{X}$ is a curve, one can immediately trace back its genus, its number of $\mathbb{F}_{q^k}$-rational points for any $k \in \mathbb{N}$, the number of $\mathbb{F}_q$-rational points on its jacobian (Theorem 1.10), the p-rank of its jacobian (see [102]), and so on.

1.2 Overview of known point counting methods

In this section we give a short overview of the state of the art in point counting. All methods below make – at least in some sense – use of Weil cohomology, either ℓ-adic, either p-adic. We emphasize that this section does *not* contain an exhaustive list of all known algorithms and we apologize in advance for not mentioning some relevant work in the field.

1.2.1 Input and output size

To have an idea of the complexity estimates that can a priori be expected from an algorithm that computes zeta functions, we first give measures for the input and output size. We will only do this for plane affine curves; all algorithms below for which complexity estimates are given fit this situation.

Suppose that the curve is given by a single bivariate equation $f(x, y) = 0$ over a finite field $\mathbb{F}_q$ with q elements. Then a measure for its size is

$$\begin{array}{lll} \text{number of monomials} & \times & (\text{ space needed to represent coefficient} \\ & & \quad + \text{ space needed to represent exponent vector}) \end{array}$$

The space needed to represent an element of $\mathbb{F}_q$ is roughly $\log q$. If the set of exponent vectors appearing in the equation is convex[3], then *Baker's formula* [5, 8] states that the number of monomials can be estimated from below by the genus g of the curve. Finally, the space needed to represent an exponent vector is measured by $\log d$, where d is the degree of f. Very often however, the parameter d is not taken into account since mostly $\log d \sim \log g$. Therefore, $g \log q$ is a popular measure for the input size, though we remark that $\log d$ is in general unbounded for fixed g.

The output size is dominated by the numerator of the zeta function, which by the Weil conjecture is a degree $2g$ polynomial, all of whose roots have absolute value $\sqrt{q}$. Since every coefficient is the sum of $\binom{2g}{i}$ i-fold products of such roots (for some $i \in \{0, \ldots, 2g\}$), we see that the output size can be measured as

$$g \log \left(\binom{2g}{g} q^g \right) \leq g \log \left(2^{2g} q^g \right) = O(g^2 \log q).$$

In conclusion, an efficient generic point counting algorithm should have a good time and space dependence on g and $\log q$. The best one can possibly expect is a method that takes $O(g^2 \log q)$ time and space. But this is far from reality:

1.11 Open Problem *It is unknown whether or not there exists a deterministic algorithm to compute the zeta function of a plane genus g degree d curve over a finite field of size q, whose running time is bounded by a fixed polynomial expression in $g, \log d$ and $\log q$.*

[3] I.e. whenever some $(a, b) \in \mathbb{N}^2$ is a convex combination of exponent vectors, then (a, b) also appears as an exponent vector.

1.12 Remark If the set of exponent vectors contains a lot of 'gaps', then $g \log q$ is an overestimation of the input size. For instance, a better measure for the size of a hyperelliptic curve $y^2 = x^{2g+1} + 1$ over an appropriate finite field $\mathbb{F}_q$ is $\log q + \log g$. It is clear from the output size that no polynomial time algorithm can exist for this particular class of curves. ■

1.2.2 ℓ-adic methods

Elliptic curves: Schoof's method

Explicit computation in Grothendieck's abstractly defined ℓ-adic cohomology spaces seems impossible in general. However, in the case of curves, we have a more down-to-earth description using the Tate module of the jacobian, see Theorem 1.10.

Schoof [97] used this to give the first polynomial time algorithm for computing the number of rational points on an elliptic curve $\overline{E}$ defined over a finite field $\mathbb{F}_q$ with q elements, where $\overline{E}$ is supposed to be the projective completion of a given Weierstrass form. Note that the Weil conjecture predicts[4] that

$$Z_{\overline{E}}(t) = \frac{qt^2 - Tt + 1}{(1-t)(1-qt)},$$

where $T \in \mathbb{Z}$ satisfies $T^2 \leq 4q$. By Theorem 1.10, T is the trace of Frobenius acting on $T_\ell(\overline{E}) \otimes_{\mathbb{Z}_\ell} \mathbb{Q}_\ell$ for any prime ℓ different from the field characteristic p. The idea of Schoof is then to compute this trace modulo ℓ, using torsion points. If one repeats this for all primes $\ell \leq \ell_m$, where ℓ_m is the smallest prime satisfying $\prod_{\ell \leq \ell_m} \ell > 4\sqrt{q}$, one can recover T. The time complexity of Schoof's algorithm[5] is $\widetilde{O}((\log q)^5)$. This was improved by mainly Atkin and Elkies to obtain a heuristically estimated time complexity of $\widetilde{O}((\log q)^4)$. The space complexity of this so-called Schoof-Elkies-Atkin (SEA) algorithm is $\widetilde{O}((\log q)^2)$. More details can be found in [33, 87, 73].

Higher genus curves

The same idea can in principle be used to determine the zeta function of a higher genus curve, as is illustrated in Pila's paper [92]. But to that end one must explicitly compute in the jacobian of the curve, which with today's machinery is very time-consuming. Pila's methods were improved [2, 52], but the resulting algorithms are still exponential in the genus of the input curve (and hence in the input size). Anno 2006, only the genus 2 case seems practically accessible using ℓ-adic methods [39, 41, 42].

[4] Actually, for elliptic curves the Weil conjecture was proven to be true by Hasse in 1934 already.

[5] It was originally omitting the characteristic 2 or 3 case, which was treated later on by Couveignes [17].

Higher-dimensional varieties

Up to our knowledge, no one sees how one could ever develop an efficient algorithm for computing zeta functions of higher-dimensional varieties (that are not jacobians) using ℓ-adic cohomology.

1.2.3 p-adic methods

The big advantage of p-adic cohomology over ℓ-adic cohomology is that its construction is much more explicit, which makes it better-suited for computational purposes. In general, p-adic algorithms are faster and more widely applicable than their ℓ-adic variants. However, and this is the big disadvantage of p-adic methods, they only seem practical for small field characteristics p since their running times depend on p rather than on $\log p$. Therefore, in all complexity estimates below, the field characteristic p is considered to be a *fixed* number.

The general idea is always as follows. Let $\overline{X}$ be an *affine* algebraic variety over a finite field $\mathbb{F}_q$ with q elements. Let $\mathbb{Q}_q$ be the fraction field of a complete discrete valuation ring $\mathbb{Z}_q$ whose residue field can be identified with $\mathbb{F}_q$. In general, $\mathbb{Q}_q$ is just a suitable unramified extension of the field of p-adic numbers $\mathbb{Q}_p$ and $\mathbb{Z}_q$ is its valuation ring. Then one can take an affine scheme X over $\mathbb{Z}_q$ whose special fiber $X \otimes_{\mathbb{Z}_q} \mathbb{F}_q$ is precisely $\overline{X}$. If X is well-chosen, the de Rham cohomology (whatever this means for an arbitrary scheme) of the generic fiber $X \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$ serves as a Weil cohomology for $\overline{X}$. The idea is then to explicitly compute the induced action of Frobenius (again: whatever this means) on this cohomology, in order to recover the zeta function.

Elliptic curves: Satoh's method

The first one to come up with a p-adic method was Satoh, who developed a fast algorithm for computing the number of points on an elliptic curve over a finite field of small characteristic [95]. Although Satoh does not explicitly mention the word 'cohomology', the spirit of his algorithm fits the above general idea. Let $\overline{E}$ be an ordinary elliptic curve over $\mathbb{F}_q$, given by an affine Weierstrass model. A theorem by Deuring [26] states that one can always find an affine scheme E over $\mathbb{Z}_q$ for which $\overline{E} = E \otimes_{\mathbb{Z}_q} \mathbb{F}_q$, such that $E \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$ is an elliptic curve with $\mathrm{End}(E \otimes_{\mathbb{Z}_q} \mathbb{Q}_q) \cong \mathrm{End}(\overline{E})$. In particular, the Frobenius endomorphism $\overline{\mathcal{F}}_q$ on $\overline{E}$ naturally lifts to an endomorphism $\mathcal{F}_q$ on $E \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$. The curve $E \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$ is called the *canonical lift* of $\overline{E}$ and the zeta function of $\overline{E}$ is determined by the induced action of Frobenius $\mathcal{F}_q^*$ on the algebraic de Rham cohomology of $E \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$. In fact, it suffices to compute $\mathcal{F}_q^*(\omega)$, where ω is the invariant differential of $E \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$. In his paper, Satoh proves that if $\mathcal{F}_q^*(\omega) = c\omega$, then

$$Z_{\widetilde{E}}(t) = \frac{qt^2 - \left(c + \frac{q}{c}\right) t + 1}{(1-t)(1-qt)}$$

(where $\widetilde{E}$ is the projective completion of $\overline{E}$). Since we know that $\left(c + \frac{q}{c}\right)^2 \leq 4q$, it suffices to compute this modulo a finite p-adic precision. All of this can be

turned into an effective algorithm (though there are some technical problems to be circumvented). This results in an algorithm with running time $\widetilde{O}((\log q)^3)$, needing $O((\log q)^3)$ space. Recall that the field characteristic p is fixed. Initially, the method was restricted to the case $p \geq 5$, but it was soon extended to $p = 2, 3$ [36].

Mainly due to the applications of point counting in cryptography (see Section 1.3 below), Satoh's technique became a popular research subject and several variants were presented. An overview of the evolution is given in the Ph.D. thesis of Vercauteren [106], who himself reduced the space complexity to $O((\log q)^2)$. By now, the number of points on an elliptic curve over a finite field of small characteristic can be computed using $\widetilde{O}((\log q)^2)$ time and space, using Harley's algorithm [49, 106].

Hyperelliptic curves: Kedlaya's method

In 2001, Kedlaya found a way to 'generalize' Satoh's method to hyperelliptic curves of any genus over finite fields of odd characteristic [60]. Instead of the canonical lift, which only applies to elliptic curves, he considered a so-called rigid analytical lift. This was introduced by Monsky and Washnitzer (who just called it 'lift') and it applies to arbitrary non-singular affine varieties. Roughly sketched, Kedlaya's method goes as follows. Let $\overline{H}$ be a hyperelliptic curve of genus g over a finite field $\mathbb{F}_q$ with q elements (q odd). Suppose that it is given by an affine Weierstrass model

$$y^2 = \overline{Q}(x)$$

where $\overline{Q}(x) \in \mathbb{F}_q[x]$ is a degree $2g+1$ polynomial without multiple roots. Let $Q(x) \in \mathbb{Z}_q[x]$ be any degree $2g+1$ polynomial that reduces to $\overline{Q}(x)$ modulo the local parameter p. Then $H = \operatorname{Spec} A^\dagger$, where

$$A^\dagger = \frac{\mathbb{Z}_q\langle x, y\rangle^\dagger}{(y^2 - Q(x))},$$

is our rigid analytical lift. The precise definition of $\mathbb{Z}_q\langle x, y\rangle^\dagger$ will be given in Chapter 4, here we just say that it consists of formal power series whose coefficients tend fast enough to zero p-adically. In particular, $H \otimes_{\mathbb{Z}_q} \mathbb{F}_q = \overline{H}$. An important feature of the construction of H is that it is endowed with an action of Frobenius, in the sense that there exists a morphism $\mathcal{F}_q$ such that the following diagram commutes:

$$\begin{array}{ccc} H & \xrightarrow{\mathcal{F}_q} & H \\ \downarrow & & \downarrow \\ \overline{H} & \xrightarrow{\overline{\mathcal{F}}_q} & \overline{H} \end{array}$$

(the vertical arrows are reduction modulo p). Then, as was proven by Monsky and Washnitzer, the algebraic de Rham cohomology of $H \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$ serves as a

Weil cohomology for $\overline{H}$. It is denoted by $H_{MW}(\overline{H}/\mathbb{Q}_q)$ (one can prove that this does not depend on the choice of $Q(x) \in \mathbb{Z}_q[x]$). One finds[6]

$$Z_{\widetilde{H}}(t) = \frac{\det\left(\mathbb{I} - \mathcal{F}_q^* t \;\middle|\; H^1_{MW}(\overline{H}/\mathbb{Q}_q)\right)}{(1-t)(1-qt)},$$

where $\widetilde{H}$ is the projective completion of $\overline{H}$. Kedlaya's algorithm then (very) roughly consists of

- computing a basis for $H^1_{MW}(\overline{H}/\mathbb{Q}_q)$;
- computing the action of Frobenius on this basis;
- expressing the result in terms of the basis again, to obtain a matrix of Frobenius.

Again, thanks to the Weil conjecture, it suffices to do all computations modulo a certain p-adic precision. The resulting time complexity is $\widetilde{O}(g^4(\log q)^3)$ and there is $O(g^3(\log q)^3)$ space needed, where p is fixed. Note that Kedlaya's algorithm has polynomial running time in the genus of the input curve: this was a big breakthrough on its own.

Larger classes of curves using Kedlaya's method

A nice feature of Kedlaya's algorithm is that there are no obvious theoretical obstructions for generalizations to larger classes of curves. This observation soon resulted in a point counting algorithm for superelliptic curves $y^r = Q(x)$ (where $r \not\equiv 0 \bmod \mathrm{char}(\mathbb{F}_q)$), having the same complexity estimates as in the hyperelliptic curve case [40]. This was generalized in its turn by Suzuki to what he calls 'strongly telescopic' curves, which are in general non-plane [103]. In the meantime, Denef and Vercauteren extended Kedlaya's algorithm to the characteristic 2 case [23, 24] and to C_{ab} curves over fields of any (small) characteristic [25]. This last algorithm needs $\widetilde{O}(g^5(\log q)^3)$ time and $\widetilde{O}(g^3(\log q)^3)$ space, again with p fixed. In this thesis, we vastly generalize this to the class of *nondegenerate curves*, which contains almost all plane curves (see also [13]). The expected complexity is $\widetilde{O}(g^{6.5}(\log q)^3)$ time and $\widetilde{O}(g^4(\log q)^3)$ space, but when applied to a C_{ab} curve, our algorithm has the same estimates as the original C_{ab} curve algorithm.

Other p-adic point counting methods

It is worth remarking that Kedlaya's method is not the only p-adic point counting technique that is being investigated for higher genus. In 2002, Mestre adapted his so-called AGM method (which was initially used to speed up Satoh's

[6] For readers being familiar with the cohomology theory of Monsky and Washnitzer, this formula can be misleading. Instead of $\mathcal{F}_q^*$, it is actually the dual morphism $q\mathcal{F}_q^{*-1}$ that appears in the M-W trace formula [105, Formula (1.2)]. One is invited to check that this makes no difference, because $\widetilde{H} \setminus \overline{H}$ consists of a single point.

algorithm for elliptic curve point counting) to ordinary hyperelliptic curves of any genus over finite fields of characteristic two [78]; it has been optimized by Lercier and Lubicz [74], while Ritzenthaler extended it to non-hyperelliptic curves of genus three [94]. These algorithms have running time $\widetilde{O}(n^2)$ (for fixed p and g) but are exponential in the genus.

Another interesting p-adic approach is to use Dwork's deformation theory [30]. One starts from a curve over $\mathbb{F}_q$ on which point counting is easy, e.g. a curve whose actual field of definition is a small subfield of $\mathbb{F}_q$. Then one slightly deforms this curve along a 1-parameter family to the curve of which one actually wants to compute the zeta function. Dwork's theory then predicts how the action of Frobenius will alter under this deformation. This was first proposed by Lauder [70] and has been studied in more detail by Gerkmann [45] and Hubrechts, who recently obtained a memory efficient version of Kedlaya's original algorithm [53]. Independently, Tsuzuki used similar ideas for computing certain one-dimensional Kloosterman sums [104].

Higher-dimensional varieties

In contrast with the ℓ-adic story, p-adic cohomology should in principle be suitable for computing zeta functions of varieties of any dimension.

In fact, the initial idea of Lauder was to use deformation theory for point counting on higher-dimensional varieties, again to be obtained from a variety on which point counting is easy, e.g. a diagonal form. The main advantage to be expected is a good time and space dependence on the dimension of the variety, as one avoids computing in multivariate polynomial rings. Lauder himself studied this in more detail in [71] to obtain a point counting algorithm for smooth projective hypersurfaces whose time dependence is singly exponential in the dimension.

Another approach was proposed by Gerkmann in his Ph.D. thesis [44], where smooth complete intersections were studied. He provided explicit methods to compute in the cohomology spaces of such varieties, but this did not result in a point counting algorithm yet.

Next, many of the ideas presented in our paper [13] apply to nondegenerate hypersurfaces of any dimension. Kedlaya spent some time on this [62] but temporarily postponed the project. Some of his ideas are written down in [1].

Finally, we mention that Lauder and Wan [72] developed a polynomial time algorithm for computing zeta functions of *arbitrary* varieties over finite fields of small characteristic, using Dwork's trace formula (which is an essential ingredient in Dwork's proof of the rationality of the zeta function [29]). Unfortunately, this algorithm seems not useful in practice. Note that, due to the p-adic nature, Lauder and Wan did not solve Open Problem 1.11: the running time of their algorithm is exponential in $\log p$.

1.2.4 The practical state of the art: a brief sketch

Several of the above algorithms have not been submitted to a careful complexity analysis yet; and even if they have, it should always be seen in practice how effective they are. The implementation results below are mainly collected from the corresponding authors' articles and were obtained using (different) home computers. Their only aim is to give a rough idea of the effectively measured running times.

By now, point counting on genus 1 curves has reached a certain degree of maturity; especially over fields of small characteristic this works very fast and memory efficient. E.g. using Schoof's (optimized) algorithm, the number of points on an elliptic curve over a prime field containing approximately 2^{80} elements can be computed in roughly 0.1 seconds. If the field size becomes of magnitude 2^{160}, this takes approximately 1.5 seconds. Computing the number of points on an elliptic curve over $\mathbb{F}_{2^{160}}$ using Harley's algorithm takes less than 0.1 seconds [106].

The genus 2 case is already more fragmentary. Over prime fields, the above sizes lie at the borderline of what is feasible: Gaudry and Schost managed to compute the number of points on a jacobian of size $\approx 2^{164}$ within 1 week [42]. Over fields of small characteristic, the same sizes can be treated in less than 1 minute using Kedlaya's technique. Over fields of characteristic 2, this can even be done in a couple of seconds using Mestre's AGM method.

In genus 3, only fields of small characteristic can be dealt with (up to our knowledge). For hyperelliptic curves, the above jacobian sizes can be treated in roughly 1 minute, again using Kedlaya's algorithm. When moving on to jacobians of size $\approx 2^{300}$, this takes about 10 minutes. In characteristic 2, this is comparable to the time needed by the AGM method. Over $\mathbb{F}_{2^{100}}$, Ritzenthaler computed the number of points on the jacobian of a genus 3 non-hyperelliptic curve in a similar amount of time [94].

For high genus, only Kedlaya's method and its generalizations seem applicable. Vercauteren computed the zeta function of a genus 160 hyperelliptic curve over $\mathbb{F}_2$ in 3.2 hours, and the zeta function of a $C_{3,5}$ curve over $\mathbb{F}_{2^{288}}$ in 12.5 hours [106].

Our algorithm for computing the zeta function of a nondegenerate curve has not been implemented yet. It is to be expected that a straightforward, naive implementation will try the user's patience. However, our belief is that using some smart tricks and optimization techniques, nondegenerate curves should become treatable in practice.

1.3 Applications of point counting

In this section, we briefly discuss some concrete applications of zeta function computation. In Appendix A, we shed a more philosophical light on the use of point counting.

1.3.1 Public key cryptography

Suppose Agnetha and Benny want to communicate with each other over the internet or some other unsecure channel. If they want to prevent others from reading their messages, they will have to encrypt these. Therefore, they must agree on a secret key. But how can they exchange this secret key safely over the internet? Using another key? In this way, one of course rolls into a vicious circle.

A surprising solution was offered by Diffie and Hellman in 1976 [27], thereby founding public key cryptography. Their method works as follows. Let Agnetha and Benny agree on a commutative group $G, +$ and an element $g \in G$. Let them choose integer numbers n_A and n_B privately. Agnetha computes

$$n_A \cdot g \quad = \quad \underbrace{g + g + \cdots + g}_{n_A \text{ times}}$$

and sends it to Benny. Similarly, Benny computes $n_B \cdot g$ and sends it to Agnetha. Both of them are now able to compute

$$\mathcal{K} = (n_A n_B) \cdot g.$$

Indeed, Agnetha computes it as $n_A \cdot (n_B \cdot g)$ and Benny computes it as $n_B \cdot (n_A \cdot g)$. They can then use $\mathcal{K}$ as a key for their secret communication.

An eavesdropper trying to obtain this key, will have to solve the following problem: given $g \in G$, $n_A \cdot g$, $n_B \cdot g$, find $(n_A n_B) \cdot g$. This is known as the Diffie-Hellman problem. It turns out to be very hard if G and g are well-chosen. Popular choices are

- $G = \mathbb{F}_q^\times, \cdot$ (the multiplicative group of a finite field $\mathbb{F}_q$) and
- $\mathrm{Jac}(\overline{H})(\mathbb{F}_q), +$ (the group of rational points on the jacobian of a genus 1,2 or 3 hyperelliptic curve $\overline{H}$ over a finite field $\mathbb{F}_q$).

The latter is considered to be the safest: no subexponential time algorithm to solve the Diffie-Hellman problem in the jacobian of a low genus hyperelliptic curve is known. However, some particular situations should be omitted, we refer to [15, Section 23.3] for more details on this. Here we just mention that many of these conditions concern the group size

$$\#\mathrm{Jac}(\overline{H})(\mathbb{F}_q)$$

(for instance, this should not factor into small prime numbers [15, Section 19.3]). Therefore, in order to know whether a curve is suitable for cryptographic purposes, one must be able to determine the number of rational points on the jacobian. This is fully determined by the zeta function, see Theorem 1.10.

By now, many other applications of the Diffie-Hellman scheme have shown up, such as encryption itself (instead of just key exchange) and digital signatures. For more on this and other types of public key cryptography we refer to [77].

1.3.2 Open mathematical problems

Many fascinating problems concerning the number of points on varieties over finite fields have not yet been solved. Fast point counting algorithms can serve in providing heuristics, detecting patterns, or simply speeding up work.

Distribution of Frobenius eigenvalues

Despite the Weil conjecture, which gives a big structural insight in the problem, the number of rational points on a given variety still just seems to be randomly picked within a certain range. However, not all outcomes have the same probability. The distribution of this number of points has become a subject of great interest.

The practical motivation to study this problem again stems from cryptography. As mentioned above, for a hyperelliptic curve to be suitable for cryptographic purposes, the number of points on its jacobian must satisfy certain smoothness conditions. It is interesting to know with what probability these smoothness conditions will be satisfied if the curve is chosen at random. See [66, 38] for some results and conjectures in the elliptic curve case.

In a 1989 experiment [90], Odlyzko empirically observed a remarkable connection between the distribution of the zeroes of the Riemann zeta function along the line

$$\left\{ s \in \mathbb{C} \;\middle|\; \mathrm{Re}(s) = \frac{1}{2} \right\}$$

and the distribution of certain eigenvalues appearing in random matrix theory. The analogy between the classical Riemann zeta function and the Hasse-Weil zeta function inspired Katz and Sarnak to connect this to how the eigenvalues of Frobenius acting on curves of fixed genus are distributed. Their results are presented in [59], and many interesting new questions showed up.

It is obvious that fast point counting algorithms are of great value for investigating this statistical behavior.

Jacobians over number fields

Many among the most intriguing open problems in number theory concern the arithmetic properties of abelian varieties over number fields. For instance, just to mention one, what is the maximal possible rank of the Mordell-Weil group? In a 1996 survey paper [93], Poonen listed some possible applications of point counting in the study of jacobians over number fields. E.g. one can try to bound the torsion part of the Mordell-Weil group of $\mathrm{Jac}(C)$, where C is a curve over a number field K. A large part of this torsion subgroup maps injectively into $\mathrm{Jac}(\overline{C})$ under reduction modulo a prime ideal $\mathfrak{p} \subset \mathcal{O}_K$. By computing $\#\mathrm{Jac}(\overline{C})(\mathcal{O}_K/\mathfrak{p})$ for different (good) primes, one can get an upper bound on the size of the torsion part of $\mathrm{Jac}(C)(K)$. A similar idea can be used to bound the rank of the endomorphism ring, which maps injectively into the endomorphism ring of the jacobian of the reduction of the curve (again, at a prime of good reduction).

Curves with many points

For any $g \in \mathbb{N}$ and prime power q, let $N_q(g)$ denote the maximal number of points one can find on a smooth and complete genus g curve over $\mathbb{F}_q$. Then the Weil conjecture implies that

$$N_q(g) \leq q + 1 + 2g\sqrt{q}.$$

For some time, it was believed that this bound is essentially sharp, but this turned out to be false. As g gets big when compared to q, the bound can be substantially improved. A theorem by Drinfel'd and Vlăduţ [28] (building on previous work of Ihara) states that for any q

$$\limsup_{g\to\infty} N_q(g)/g \leq \sqrt{q} + 1.$$

If q is a square, then the converse inequality holds [55]. A big open problem in this field is to prove the same for non-square q. There is not even a single such value of q for which this has been done. A related problem is to consider

$$\liminf_{g\to\infty} N_q(g)/g.$$

It was only recently proven that this number is always strictly bigger than 0, that is: there is a $c_q \in \mathbb{R}_0^+$ such that for *any* $g \in \mathbb{N} \setminus \{0\}$ there exists a genus g curve with at least $c_q g$ points [34].

Apart from these theoretical considerations, there is also a practical need for curves with many points: in the early 1980's, Goppa [46] proposed to use algebraic curves over finite fields for error-correction. The idea of error-correction is to provide a message with redundant information, so that minor errors occurring during transmission can be detected and corrected. Roughly sketched the methods goes as follows: let $\overline{C}$ be a smooth curve over a finite field $\mathbb{F}_q$, let $\mathcal{P} = \{P_1, \ldots, P_n\}$ be a set of $\mathbb{F}_q$-rational points on $\overline{C}$ and let G be an $\mathbb{F}_q$-rational divisor whose support is disjoint from $\mathcal{P}$. Let $f_1, \ldots, f_k$ be a basis for the Riemann-Roch space

$$\mathcal{L}(G) = \left\{ f \in \mathbb{F}_q(\overline{C})^\times \;\middle|\; (f) + G \geq 0 \right\} \cup \{0\}.$$

Then the coding works using the map

$$\mathbb{F}_q^k \to \mathbb{F}_q^n : (a_1, \ldots, a_k) \mapsto (f(P_1), \ldots, f(P_n)) \text{ with } f = \sum a_i f_i.$$

Using the Riemann-Roch theorem, one can analyze the error-correcting capacities of this code, which turns out to be very good if n is big. Therefore, one must take $\overline{C}$ such that it has many rational points. We note that other types of varieties have been proposed for error-correcting purposes and that an application of p-adic Frobenius computation to this has recently appeared in [1].

The point counting problem in its own right

Even from a purely theoretical point of view, Open Problem 1.11 is intriguing. If someone would come up with a polynomial running time algorithm for computing zeta functions of a plane curve, this would be a great conceptual breakthrough, no matter how practical or impractical the algorithm is.

1.4 This thesis

In this thesis, we present a Kedlaya-style algorithm that computes the zeta function of a curve in $\left(\mathbb{A}^1_{\mathbb{F}_q} \setminus \{0\}\right)^2$ defined by a Laurent polynomial

$$\overline{f} \in \mathbb{F}_q[x^{\pm 1}, y^{\pm 1}]$$

that is *nondegenerate with respect to its Newton polytope*. Here, $\mathbb{F}_q$ is any finite field. The precise definition of this nondegeneracy condition will be given in Chapter 2. Here we already mention that it is *almost always* satisfied, in the following sense: the probability that a randomly chosen Laurent polynomial with prescribed Newton polytope is nondegenerate, is ≈ 1 (if q is relatively big). In particular, any C_{ab} curve – and hence any elliptic, hyperelliptic or superelliptic curve – allows a nondegenerate model[7]. Therefore, our algorithm is a vast generalization of what was previously known. Moreover, when applied to a C_{ab} curve, it needs $\widetilde{O}(g^5(\log q)^3)$ time and $\widetilde{O}(g^3(\log q)^3)$ space, which are the same asymptotics as in the original algorithm of Denef and Vercauteren [25]. Recall that $p = \mathrm{Char}(\mathbb{F}_q)$ is fixed.

In general, our main result can be stated as:

1.13 Theorem *There exists a deterministic algorithm to compute the zeta function of a genus g nondegenerate curve over $\mathbb{F}_q$ that requires $\widetilde{O}((\log q)^3\Psi_t)$ bit-operations and $\widetilde{O}((\log q)^3\Psi_s)$ space for p fixed. Here, Ψ_t and Ψ_s are parameters that depend on the Newton polytope of the input curve only; for 'most common' Newton polytopes, $\Psi_t = \widetilde{O}(g^{6.5})$ and $\Psi_s = \widetilde{O}(g^4)$.*

For explicit formulas for Ψ_t and Ψ_s we refer to Theorem 6.8 and Theorem 6.9. The notion 'most common' is not intended to be made mathematically exact. It just means that the Newton polytope should not be shaped too exotically. Again we refer to Chapter 6 for more details.

To obtain the above result, we proved a number of new theoretical results that are interesting in their own right. They are presented in the remainder of this thesis, which is organized as follows.

In **Chapter 2 : Nondegenerate curves**, we introduce our main objects of study. We analyze their geometric properties and prove that the condition of

[7] In fact, this may fail if $q \sim g$ (see Lemma 2.21). But in that case the naive point counting method is fast enough.

being nondegenerate is Zariski-open. We give an explicit Riemann-Roch theorem as well as a sparse description of the cohomology (in the characteristic zero case). We end with a discussion on why nondegenerate curves are so well-suited for Kedlaya's approach.

Chapter 3 : The effective Nullstellensatz problem for discrete valuation rings is a self-contained chapter with a sparse effective Nullstellensatz as main result. We use this to prove some crucial properties of nondegenerate hypersurfaces, and investigate what happens if the condition of being nondegenerate is dropped.

In **Chapter 4 : Monsky-Washnitzer cohomology of nondegenerate curves**, we develop the main tools for our algorithm. We constructively show that the Monsky-Washnitzer cohomology of a nondegenerate curve over a finite field is isomorphic to the de Rham cohomology of a well-chosen lift (again a nondegenerate curve). We also give a new constructive proof of the fact that the Frobenius endomorphism can be lifted. Both proofs will play a key role in the development of our algorithm in Chapter 6. Moreover, the lift of Frobenius turns out to satisfy a very natural convergence criterion, which allows us to give a sparse description of the Monsky-Washnitzer cohomology of nondegenerate curves.

Chapter 5 : Linear algebra algorithms over p-adic rings is a very short chapter in which two classical linear algebra problems are reconsidered: a new technique for system solving is presented, and the classical algorithm for computing characteristic polynomials based on reduction to the Hessenberg form is put to a careful analysis.

In **Chapter 6 : Point counting on nondegenerate curves**, all this is put together in our point counting algorithm, culminating in Theorem 1.13 above. A detailed complexity analysis is given.

Finally, **Appendix A: Point counting for the non-mathematician** is an attempt to explain to the non-mathematical reader what this thesis is about, and **Appendix B: Nederlandse samenvatting** contains a summary in Dutch.

We want to emphasize that all proofs given below are either proofs of new results, either new proofs of known results that contain additional information, playing an essential role in the construction of our algorithm. The only exceptions to this are some new, elementary proofs of well-known facts[8] in Section 2.3.

Finally, we remark that the algorithm has not yet been implemented, besides some small subroutines whose primary objective was to double-check the correctness of our methods. Therefore, at no point in this thesis concrete running

[8]... or direct consequences of well-known facts (such as Theorem 2.18).

times will be given.

The main results were obtained together with Jan Denef and Frederik Vercauteren, and will be published in the International Mathematics Research Notices [13].

Chapter 2

Nondegenerate curves

This chapter introduces the main objects of study of this thesis: curves living in $(\mathbb{A}^1 \setminus \{0\})^2$ that are defined by a bivariate Laurent polynomial that is *nondegenerate with respect to its Newton polytope* Γ. Below, we define what this means and prove that the condition of being nondegenerate is generically satisfied. Next, we recall the process of toric resolution of singularities and prove that a wealth of geometric information is contained in Γ. We proceed with a very explicit Riemann-Roch theorem, a sparse description of the cohomology of nondegenerate curves, and some further properties. In the last section, we interpret the collected material from the point counting viewpoint.

Throughout, x and y are fixed formal variables. For any domain R and any subset $\mathcal{S} \subset \mathbb{R}^2$, we denote by $R[\mathcal{S}]$ the ring[1]

$$R[x^i y^j \mid (i,j) \in \mathcal{S} \cap \mathbb{Z}^2].$$

For instance, $R[\mathbb{N}^2]$ is just the polynomial ring $R[x,y]$ and $R[\mathbb{Z}^2]$ is the Laurent polynomial ring $R[x^{\pm 1}, y^{\pm 1}]$. For any $f \in R[\mathbb{Z}^2]$, we will interchange the notations $\frac{\partial f}{\partial x}$ (resp. $\frac{\partial f}{\partial y}$) and f_x (resp. f_y). Finally, if $\mathbb{F}$ is a field, $\overline{\mathbb{F}}$ will denote an algebraic closure.

2.1 A generic condition

Let $\mathbb{F}$ be an arbitrary field and denote with $\mathbb{T}^2_{\mathbb{F}} := (\mathbb{A}^1_{\mathbb{F}} \setminus \{0\})^2 \cong \operatorname{Spec} \mathbb{F}[\mathbb{Z}^2]$ the two-dimensional algebraic torus over $\mathbb{F}$. Any $f \in \mathbb{F}[\mathbb{Z}^2]$ allows a representation

$$f = \sum_{(i,j) \in \mathcal{S}} f_{i,j} x^i y^j$$

for some finite subset $\mathcal{S} \subset \mathbb{Z}^2$ and $f_{i,j} \in \mathbb{F} \setminus \{0\}$. $\mathcal{S}$ is called the *support* of f. The convex hull of $\mathcal{S}$ in $\mathbb{R}^2$ is called the *Newton polytope* of f and will be

[1] In case $\mathcal{S}$ is a (semi-)group, this should not be confused with the (semi-)group R-algebra generated by $\mathcal{S}$, for which the notation $R[\mathcal{S}]$ is usually preserved. In this thesis, $R[\mathcal{S}]$ will always denote the ring $R[x^i y^j \mid (i,j) \in \mathcal{S} \cap \mathbb{Z}^2]$ of Laurent polynomials.

denoted by $\Gamma(f)$ or just Γ. The boundary of Γ is denoted by $\partial\Gamma$. The faces of Γ can be subdivided according to their dimension. If $\dim\Gamma = 2$, we have *vertices*, *edges* and Γ itself. To an edge γ, we can associate its *arithmetic length* $\ell(\gamma) = \#(\gamma \cap \mathbb{Z}^2) - 1$. Note that if γ is the edge connecting (a, b) and (c, d), its arithmetic length is given by $\gcd(a - c, b - d) \in \mathbb{N} \setminus \{0\}$. Also note that

$$\#(\partial\Gamma \cap \mathbb{Z}^2) \;=\; \sum_{\gamma \text{ edge of } \Gamma} \ell(\gamma).$$

If σ is any subset of $\mathbb{R}^2$, we write f_σ for $\sum_{(i,j)\in\mathcal{S}\cap\sigma} f_{i,j}x^iy^j$.

2.1 Definition We say that a Laurent polynomial $f \in \mathbb{F}[\mathbb{Z}^2]$ is *nondegenerate with respect to its Newton polytope* Γ if for all faces γ of Γ (including Γ itself) the system of equations

$$f_\gamma = x\frac{\partial f_\gamma}{\partial x} = y\frac{\partial f_\gamma}{\partial y} = 0$$

has no solutions in $\mathbb{T}^2_{\mathbb{F}}$ (that is, there are no solutions in $\left(\overline{\mathbb{F}} \setminus \{0\}\right)^2$).

In the next section, we will discuss the geometric meaning of this notion. A first result in this thesis is that a sufficiently generic Laurent polynomial with given Newton polytope is nondegenerate [13]. This is easy and well-known in the characteristic 0 case. Over fields of finite characteristic however, things get more subtle, and the statement is no longer true in higher dimensions (see Remark 2.5 and Remark 3.15).

2.2 Lemma *Let $\Gamma \subset \mathbb{R}^2$ be the convex hull of a set of points in $\mathbb{Z}^2$. Consider the map*

$$\varphi : \mathbb{Z}^2 \to \mathbb{A}^2_{\mathbb{F}} : (i, j) \mapsto (i, j).$$

Then the dimension of the affine subspace of $\mathbb{A}^2_{\mathbb{F}}$ spanned by $\varphi(\Gamma \cap \mathbb{Z}^2)$ equals $\dim\Gamma$.

Proof. This is trivial if $\mathbb{F}$ is of characteristic 0, so assume that $\mathrm{char}(\mathbb{F}) = p > 0$. We may also assume that $\mathbb{F}$ is a prime field, i.e. $\mathbb{F} = \mathbb{Z}/(p)$. As the $\dim\Gamma = 0$ case is obvious, we first suppose that $\dim\Gamma = 1$. Take points $q_1 \neq q_2 \in \Gamma \cap \mathbb{Z}^2$ and suppose that $\varphi(q_1) = \varphi(q_2)$. Then we must have that $q_2 = q_1 + p^e v$ for some $e \in \mathbb{N}_0$ and some nonzero $v \in \mathbb{Z}^2$ that is not divisible by p. Because Γ is convex, it also contains $q_1 + v$, and definitely $\varphi(q_1) \neq \varphi(q_1 + v)$.

Now suppose $\dim\Gamma = 2$. Take points $q_1, q_2 \in \Gamma \cap \mathbb{Z}^2$ such that $\varphi(q_1) \neq \varphi(q_2)$. Take a $q_3 \in \Gamma \cap \mathbb{Z}^2$ that is not on the line through q_1 and q_2, but suppose $\varphi(q_3)$ *is* in the span of $\varphi(q_1)$ and $\varphi(q_2)$, say

$$q_3 = q_1 + k(q_2 - q_1) + p^e v$$

for some $e \in \mathbb{N}_0$ and some nonzero $v \in \mathbb{Z}^2$ that is not divisible by p and linearly independent of $q_2 - q_1$. Note that although this expansion is far from unique,

there is a natural upper bound for e, so that we may assume that it is maximal. Indeed, if we write $q_3 - q_1 = (a_1, a_2)$ and $q_2 - q_1 = (b_1, b_2)$, then it is not hard to see that $p^e | b_2 a_1 - a_2 b_1 \neq 0$. As a consequence, $\varphi(v)$ and $\varphi(q_2 - q_1)$ are linearly independent, since otherwise this would contradict the maximality of e.

Next, we may suppose that $0 \leq k < p^e$ by repeatedly replacing $p^e v \leftarrow p^e v \pm p^e(q_2 - q_1)$ if necessary. We may even suppose that $k \neq 0$, since otherwise we can proceed as in the $\dim \Gamma = 1$ case. Now define

$$q = \frac{k-1}{p^e} q_1 + \frac{p^e - k}{p^e} q_2 + \frac{1}{p^e} q_3 = q_2 + v.$$

The first equality shows that $q \in \Gamma$, the second one shows that $q \in \mathbb{Z}^2$. Finally, $\varphi(q)$ is not on the line through $\varphi(q_1)$ and $\varphi(q_2)$. ■

2.3 Proposition *Let Γ be a convex polytope in $\mathbb{R}^2$ with integer vertex coordinates and write $\mathcal{S} = \Gamma \cap \mathbb{Z}^2$. Then the set of points*

$$(f_{i,j})_{(i,j)\in\mathcal{S}} \in \mathbb{A}_{\mathbb{F}}^{\#\mathcal{S}}$$

for which $f = \sum f_{i,j} x^i y^j$ is not *nondegenerate with respect to its Newton polytope is contained in an algebraic set of codimension ≥ 1. Moreover, this algebraic set is defined over the prime subfield of $\mathbb{F}$.*

Proof. Let γ be a face of Γ. Suppose for now that it is two-dimensional. Let X_γ be the algebraic set in $\mathbb{A}_{\mathbb{F}}^{\#\mathcal{S}} \times (\mathbb{A}_{\mathbb{F}} \setminus \{0\})^2$ defined by the equations

$$\sum_{(i,j)\in\gamma\cap\mathbb{Z}^2} f_{i,j} x^i y^j = 0, \qquad \sum_{(i,j)\in\gamma\cap\mathbb{Z}^2} i f_{i,j} x^i y^j = 0, \qquad \sum_{(i,j)\in\gamma\cap\mathbb{Z}^2} j f_{i,j} x^i y^j = 0.$$

It has codimension 3. Indeed, for every $a, b \in \mathbb{F} \setminus \{0\}$ the above equations define a linear codimension 3 subspace of $\mathbb{A}_{\mathbb{F}}^{\#\mathcal{S}} \times \{x = a, y = b\}$. Here we used that there are no $a, b, c \in \mathbb{F}$ such that $a + bi + cj = 0$ for all $(i, j) \in \varphi(\gamma \cap \mathbb{Z}^2)$, where φ is the map from the foregoing lemma. Let Y_γ be the projection of X_γ on $\mathbb{A}_{\mathbb{F}}^{\#\mathcal{S}}$. It has codimension at least 1 and consists exactly of those $(f_{i,j})_{(i,j)\in\mathcal{S}}$ that correspond to a Laurent polynomial for which the nondegeneracy condition with respect to γ is not satisfied.

If γ has dimension < 2, one can again construct such a Y_γ using an appropriate change of variables so that f_γ becomes a univariate Laurent polynomial, or a constant.

Then the Zariski closure of $\cup_\gamma Y_\gamma$ is the requested algebraic set. We remark that $\cup_\gamma Y_\gamma$ may contain points that correspond to Laurent polynomials that *are* nondegenerate with respect to their Newton polytope: this will be the case whenever they have a Newton polytope that lies strictly inside Γ. ■

For our needs, the following corollary is very important. As usual, if q is a power of a prime number, $\mathbb{F}_q$ denotes the finite field with q elements.

2.4 Corollary *Let Γ be a convex polytope in $\mathbb{R}^2$ with integer vertex coordinates and let p be a prime number. Let P_n be the probability that a randomly chosen $\overline{f} \in \mathbb{F}_{p^n}[\mathbb{Z}^2]$ with support inside Γ is nondegenerate with respect to its Newton polytope. Then $P_n \to 1$ as $n \to \infty$.*

To sum up, the class of curves we are considering is very large. If the field size gets big, any randomly chosen bivariate Laurent polynomial with given Newton polytope will be nondegenerate. To convince the reader, we include some small experimental results obtained using the `MAGMA` computer system. For the Newton polytopes $\Gamma_1 = \mathrm{Conv}\{(-1,3),(-3,1),(1,-3),(2,-1),(2,2)\}$ and $\Gamma_2 = \mathrm{Conv}\{(0,0),(2,0),(5,3)\}$ and for fields of growing sizes, 100 random Laurent polynomials were chosen. The table below shows how many among them were nondegenerate.

	Γ_1	Γ_2		Γ_1	Γ_2
$\mathbb{F}_2$	19	30	$\mathbb{F}_{2^5}$	94	89
$\mathbb{F}_3$	31	47	$\mathbb{F}_{2^{10}}$	98	100
$\mathbb{F}_4$	46	56	$\mathbb{F}_{2^{15}}$	100	100
$\mathbb{F}_5$	51	47	$\mathbb{F}_{2^{20}}$	100	100
$\mathbb{F}_7$	58	69	$\mathbb{F}_{2^{25}}$	100	100
$\mathbb{F}_8$	53	76	$\mathbb{F}_{2^{30}}$	100	100
$\mathbb{F}_9$	65	78	$\mathbb{F}_{2^{35}}$	100	100

2.5 Remark Note that Proposition 2.3 is false if the condition of $\mathcal{S}$ being the entire $\mathbb{Z}^2$-part of a convex set is omitted: $\mathcal{S} = \{(0,0),(p,0),(0,p)\}$ is an easy counterexample (where $p > 0$ is the field characteristic). Another important note is that it is impossible to generalize the above to hypersurfaces of arbitrary dimension, this is discussed in Remark 3.15. ■

2.6 Convention Although we defined the notion of nondegeneracy for arbitrary $f \in \mathbb{F}[\mathbb{Z}^2]$, it does not behave well if the embedding space is 'too big', i.e. $\dim \Gamma(f) < 2$. Indeed, we intend to study *curves* in $\left(\mathbb{A}^1_{\mathbb{F}} \setminus \{0\}\right)^2$ that are defined by a nondegenerate Laurent polynomial. If $\dim \Gamma = 0$, f defines the empty set. If $\dim \Gamma = 1$, although it is true that f defines an algebraic set of dimension 1, it will in general not be irreducible. Therefore, from now on we will always implicitly assume that $\dim \Gamma = 2$.

2.2 Toric resolution of nondegenerate curve singularities

From the definition, it is immediate that a Laurent polynomial that is nondegenerate with respect to its Newton polytope Γ defines a nonsingular curve in $\mathbb{T}^2_{\mathbb{F}} = \left(\mathbb{A}^1_{\mathbb{F}} \setminus \{0\}\right)^2$. However, when considering its Zariski-closure in $\mathbb{P}^2_{\mathbb{F}}$, singularities of any kind may appear. In this section, we will replace $\mathbb{P}^2_{\mathbb{F}}$ with

another easy-to-describe compactification of $\mathbb{T}^2_{\mathbb{F}}$, the construction of which is closely related to Γ and in which the closure of our curve becomes nonsingular.

The material in this section is based on the theory of toric varieties (see [18] for a more detailed introduction). It is explained in a self-contained way. However, for the reader having some familiarity with the subject, we note the following in order to avoid confusion. Below, we are only interested in toric varieties associated to fans that are coming from two-dimensional polytopes. For this particular case, the classical construction associating a fan to a polytope and then associating a toric variety to this fan using dual cones can be shortcut. Therefore, no fans nor dual cones will be used.

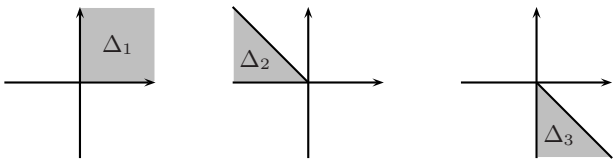
2.2.1 Toric surfaces

As announced, the theory of toric surfaces deals with certain compactifications of the two-dimensional algebraic torus $\mathbb{T}^2_{\mathbb{F}}$. The most famous compactification is of course the projective plane $\mathbb{P}^2_{\mathbb{F}}$, so let us revisit its construction as a motivational example.

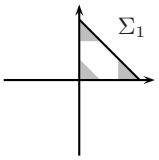
2.7 Example $\mathbb{P}^2_{\mathbb{F}}$ is obtained from $\mathbb{T}^2_{\mathbb{F}}$ by adjoining three lines: the x-axis, the y-axis and the line at infinity. In practice however, one prefers to adjoin only *two* lines at a time, resulting in an affine plane $\mathbb{A}^2_{\mathbb{F}}$ (which is easier to work in). $\mathbb{P}^2_{\mathbb{F}}$ is then obtained by gluing together three such $\mathbb{A}^2_{\mathbb{F}}$'s

$$\begin{array}{lllll} \mathbb{T}^2_{\mathbb{F}} & \cup\ x\text{-axis} & \cup\ y\text{-axis} & = & \operatorname{Spec}\mathbb{F}[x,y] \\ \mathbb{T}^2_{\mathbb{F}} & \cup\ y\text{-axis} & \cup\ \text{line at infinity} & = & \operatorname{Spec}\mathbb{F}[yx^{-1},x^{-1}] \\ \mathbb{T}^2_{\mathbb{F}} & \cup\ x\text{-axis} & \cup\ \text{line at infinity} & = & \operatorname{Spec}\mathbb{F}[xy^{-1},y^{-1}]. \end{array}$$

Note that we can rewrite these as $\operatorname{Spec}\mathbb{F}[\Delta_1]$, $\operatorname{Spec}\mathbb{F}[\Delta_2]$ and $\operatorname{Spec}\mathbb{F}[\Delta_3]$, where the Δ_i are the following cones in $\mathbb{R}^2$.



These cones are generated by the angles of (any multiple of) the standard 2-simplex Σ_1.



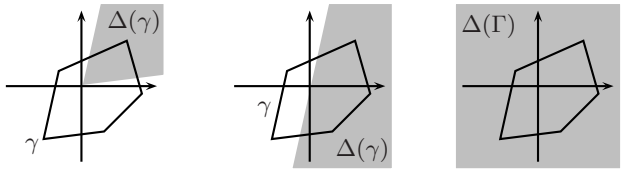
Thus $\mathbb{P}^2_{\mathbb{F}}$ can be constructed by gluing together three affine surfaces, one for each angle of Σ_1. In the language that will be described below, this means that $\mathbb{P}^2_{\mathbb{F}}$ is the *toric surface associated to* Σ_1, denoted $\mathbb{P}_{\mathbb{F},\Sigma_1}$ or shortly $\mathbb{P}_{\Sigma_1}$.

Note that the slopes of the edges of Σ_1 correspond to the 'degree functions' associated to the lines that are adjoined to $\mathbb{T}^2_{\mathbb{F}}$. Indeed, the degree of a Laurent polynomial $h = \sum_{i,j} h_{i,j} x^i y^j \in \mathbb{F}[\mathbb{Z}^2]$ can be measured

$$\begin{array}{rrcl} \text{with respect to the } x\text{-axis:} & \deg_x h = \max\{ & i & \mid h_{i,j} \neq 0\}, \\ \text{with respect to the } y\text{-axis:} & \deg_y h = \max\{ & j & \mid h_{i,j} \neq 0\} \\ \text{and with respect to the line at infinity:} & \deg_\infty h = \max\{ & i+j & \mid h_{i,j} \neq 0\}, \end{array}$$

while the edges of Σ_1 lie on the lines $i = 0$, $j = 0$ and $i + j = 1$. So another way to look at $\mathbb{P}_{\Sigma_1}$ is by adjoining to $\mathbb{T}^2_{\mathbb{F}}$ three lines along with a 'homogenization degree' function, one for each edge of Σ_1. ■

Now, let Γ be any convex polytope in $\mathbb{R}^2$ with integer vertex coordinates. To each face $\gamma \subset \Gamma$, we can associate the cone $\Delta(\gamma)$ generated (that is, obtained by taking linear combinations with coefficients in $\mathbb{R}^+$) by all vectors in $\{x - p \mid x \in \Gamma,\, p \in \gamma\}$.



cone at a vertex — cone at an edge — cone at Γ itself

By Gordan's lemma (see e.g. [64]), $\mathbb{F}[\Delta(\gamma)]$ is a finitely generated $\mathbb{F}$-algebra, so $\mathbb{A}_{\mathbb{F},\Delta(\gamma)} := \operatorname{Spec} \mathbb{F}[\Delta(\gamma)]$ is an affine variety over $\mathbb{F}$. It is called the *affine toric surface* associated to γ.

If $\gamma \subset \tau$ with τ another face of Γ, then $\Delta(\gamma) \subset \Delta(\tau)$ and $\mathbb{F}[\Delta(\tau)]$ is obtained from $\mathbb{F}[\Delta(\gamma)]$ by adjoining the inverse of each monomial $x^i y^j \in \mathbb{F}[\Delta(\gamma)]$ for which $(i, j) \in \operatorname{Lin}(\tau)$. Here, $\operatorname{Lin}(\tau)$ is the linear subspace of $\mathbb{R}^2$ generated by the differences of vectors in τ. Thus, $\operatorname{Spec} \mathbb{F}[\Delta(\tau)]$ is obtained from $\operatorname{Spec} \mathbb{F}[\Delta(\gamma)]$ by cutting away some zero locus. Otherwise said: $\mathbb{A}_\gamma$ contains $\mathbb{A}_\tau$ as a Zariski-open subvariety. Note that $\mathbb{A}_\Gamma = \mathbb{T}^2_{\mathbb{F}}$, so the algebraic torus is canonically an open subvariety of each $\mathbb{A}_\gamma$.

Now two such affine toric surfaces $\mathbb{A}_{\gamma_1}$ and $\mathbb{A}_{\gamma_2}$ can be glued together along their common open subvariety $\mathbb{A}_\tau$, where τ is the smallest face containing both γ_1 and γ_2. Gluing all these $\mathbb{A}_\gamma$'s together then precisely results in $\mathbb{P}_{\mathbb{F},\Gamma}$ (or shortly $\mathbb{P}_\Gamma$), the *toric surface associated to* Γ.

2.8 Theorem *$\mathbb{P}_\Gamma$ is a well-defined projective surface. If $S^{\mathbb{F}}_\Gamma$ is the graded ring*

$$\bigoplus_{d \in \mathbb{N}} \langle t^d x^i y^j \mid (i, j) \in d\Gamma \cap \mathbb{Z}^2 \rangle_{\mathbb{F}}$$

(where t is a formal variable), then $\mathbb{P}_\Gamma \cong \operatorname{Proj} S^{\mathbb{F}}_\Gamma$.

Proof. See for instance [7]. ■

Now, to every face $\gamma \subset \Gamma$ we can associate the algebraic torus

$$\mathbb{T}_\gamma := \operatorname{Spec} \mathbb{F}[\operatorname{Lin}(\gamma)].$$

Since $\operatorname{Lin}(\gamma) \subset \Delta(\gamma)$, we obtain a canonical surjective homomorphism from $\mathbb{F}[\Delta(\gamma)]$ to $\mathbb{F}[\operatorname{Lin}(\gamma)]$ by mapping the monomials $x^i y^j$ with $(i,j) \in \Delta(\gamma) \setminus \operatorname{Lin}(\gamma)$ to zero and the other monomials to themselves. This identifies $\mathbb{T}_\gamma$ with a closed subvariety of $\mathbb{A}_\gamma$. Note that $\dim \mathbb{T}_\gamma = \dim \gamma$ and that

$$\mathbb{P}_\Gamma = \bigsqcup_{\gamma \text{ face of } \Gamma} \mathbb{T}_\gamma,$$

i.e. $\mathbb{P}_\Gamma$ can be decomposed into a number of algebraic tori, one for each face of Γ. Furthermore, the closure of $\mathbb{T}_\gamma$ in $\mathbb{P}_\Gamma$ is

$$\bigsqcup_{\tau \text{ face of } \gamma} \mathbb{T}_\tau.$$

If $\mathbb{F} = \mathbb{F}_q$ is a finite field with q elements, this decomposition implies that the number of rational points on $\mathbb{P}_{\mathbb{F}_q,\Gamma}$ equals $(q-1)^2 + rq$, where r is the number of edges (or vertices) of Γ.

We then have the following theorem.

2.9 Theorem *$\mathbb{P}_\Gamma$ is a normal variety. In particular, it is nonsingular outside the tori associated to the vertices of Γ.*

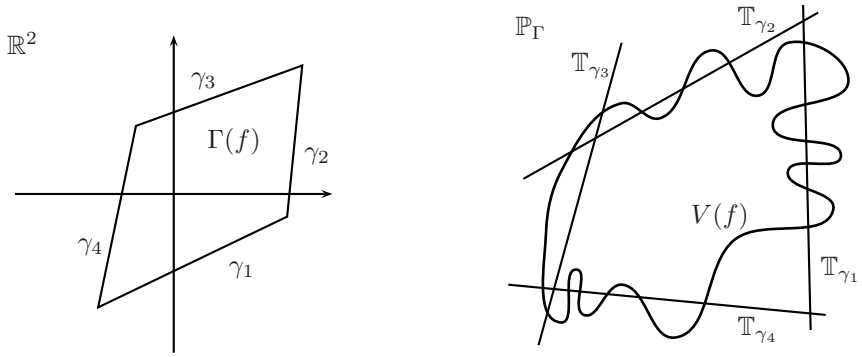
Proof. See [89, Proposition 1.2]. ■

In general, the zero-dimensional tori associated to the vertices of Γ will only rarely be nonsingular points. In fact, one can show that $\mathbb{P}_\Gamma$ is smooth if and only if every $\Delta(\gamma)$ is generated by linearly independent vectors (see [18, Theorem 2.1]).

Summarizing, the toric surface $\mathbb{P}_\Gamma$ can be looked at as being obtained from the two-dimensional algebraic torus $\mathbb{T}^2_\mathbb{F}$ by adjoining as many lines as there are edges of Γ. The points in which these lines intersect may be singular, but elsewhere $\mathbb{P}_\Gamma$ is smooth.

2.2.2 Resolution of nondegenerate curve singularities

Toric surfaces are a natural environment for the nonsingular models of curves that are defined by a bivariate Laurent polynomial that is nondegenerate with respect to its Newton polytope. Indeed, let f be such a polynomial and let $V(f)$ denote the closure in $\mathbb{P}_{\Gamma(f)}$ of the locus of f in the torus $\mathbb{T}^2_\mathbb{F}$.



Clearly, the nondegeneracy condition with respect to $\Gamma(f)$ itself states that $V(f) \cap \mathbb{T}_{\mathbb{F}}^2$ is nonsingular. Now let γ be an edge of Γ. Then $V(f) \cap \mathbb{T}_\gamma$ equals the locus of $x^{-i}y^{-j}f_\gamma$ in $\mathbb{T}_\gamma$, where (i,j) is some $\mathbb{Z}^2$-point on γ. It is then easily checked that $V(f)$ intersects the torus $\mathbb{T}_\gamma$ transversally if and only if $f_\gamma, x\frac{\partial f_\gamma}{\partial x}, y\frac{\partial f_\gamma}{\partial y}$ have no common zero in $\mathbb{T}_{\mathbb{F}}^2$. Similarly, the nondegeneracy condition with respect to a vertex guarantees that $V(f)$ does not contain the corresponding zero-dimensional torus. In conclusion, $V(f)$ is a nonsingular curve that intersects the one-dimensional tori corresponding to the edges of Γ transversally, and that does not contain the zero-dimensional tori corresponding to the vertices of Γ. For this reason, $V(f)$ is called the *toric resolution* of the affine curve defined by $f = 0$ on $\mathbb{T}_{\mathbb{F}}^2$.

2.3 An explicit Riemann-Roch theorem

Throughout this section, assume that $\mathbb{F}$ is a *perfect* field. Let $f \in \mathbb{F}[\mathbb{Z}^2]$ be a Laurent polynomial that is nondegenerate with respect to its (two-dimensional) Newton polytope Γ and let $C = V(f) \subset \mathbb{P}_\Gamma$ be the toric resolution of the curve defined by f on $\mathbb{T}_{\mathbb{F}}^2$. Enumerate the vertices $p_1, \ldots, p_r$ of Γ clockwise and let t_k be the edge connecting p_k with p_{k+1} (where $p_{r+1} = p_1$). Let $P_k := \mathbb{T}_{p_k} \subset \mathbb{P}_\Gamma$ be the zero-dimensional torus corresponding to p_k and let $T_k := \mathbb{T}_{t_k} \subset \mathbb{P}_\Gamma$ be the one-dimensional torus corresponding to t_k.

The following two divisors will play a lead role in this section (and the rest of the thesis):

$$D_{C,\Gamma} = -\sum_{k=1}^{r} N_k(T_k \cap C) \quad \text{and} \quad W_C = \sum_{k=1}^{r}(T_k \cap C).$$

Here, the N_k are defined as follows. Let e_k be the vector $(a_k, b_k) \in \mathbb{Z}^2$ with $\gcd(a_k, b_k) = 1$ which is perpendicular to t_k and points from t_k towards the interior of Γ. Then $N_k = p_k \cdot e_k$. Note that instead of p_k we could have taken any vertex on t_k, since the difference is perpendicular to e_k. The notation $D_{C,\Gamma}$ emphasizes that this divisor not only depends on C, but also on Γ. For instance, if we replace f by $x^i y^j f$ for some $(i,j) \in \mathbb{Z}^2$, then Γ is replaced by $\Gamma + (i,j)$,

but C remains the same. However, if from the context it is clear what Γ is, we will write D_C instead of $D_{C,\Gamma}$.

For any subset $\mathcal{S} \subset \mathbb{R}^2$, denote with $L_{\mathcal{S}}$ the $\mathbb{F}$-vector space generated by $x^i y^j$ with $(i,j) \in \mathcal{S} \cap \mathbb{Z}^2$. If D is a divisor on C which is defined over $\mathbb{F}$, then $\mathcal{L}(D)$ denotes the corresponding Riemann-Roch space

$$\left\{ f \in \mathbb{F}(C)^\times \;\middle|\; (f) + D \geq 0 \right\} \cup \{0\}.$$

Note that $D_{C,\Gamma}$ and W_C are defined over $\mathbb{F}$. If $\mathbb{F} \subset \mathbb{F}'$ is a field extension, we write

$$\mathcal{L}_{\mathbb{F}'}(D) = \left\{ f \in \mathbb{F}'(C)^\times \;\middle|\; (f) + D \geq 0 \right\} \cup \{0\}.$$

Since $\mathbb{F}$ is perfect, $\mathcal{L}_{\overline{\mathbb{F}}}(D)$ is generated by $\mathcal{L}(D)$ and $\dim_{\mathbb{F}} \mathcal{L}(D) = \dim_{\overline{\mathbb{F}}} \mathcal{L}_{\overline{\mathbb{F}}}(D)$. A proof of this can be found in [100], for a slight variant we refer to Lemma 4.3. We will often make abuse of notation and write things as $L_{\mathcal{S}} \subset \mathcal{L}(D)$, though the latter is defined as a subspace of the function field $\mathbb{F}(C)$.

The following lemmata (2.10 – 2.13) are well-known and implicitly contained in [19].

2.10 Lemma *Let $k \in \{1, \ldots, r\}$, $m \in \mathbb{N}_0$ and $(i,j) \in \mathbb{Z}^2$. If $(i,j) \in m\Gamma$, then $e_k \cdot (i,j) \geq mN_k$ with equality if and only if $(i,j) \in mt_k$.*

PROOF. This is straightforward. ■

2.11 Lemma *Let $g \in \mathbb{F}[\mathbb{Z}^2]$ have support inside $m\Gamma$ for some $m \in \mathbb{N}_0$. Let P be a point in $C \setminus \mathbb{T}^2_{\mathbb{F}}$ and denote with t_k the edge of Γ such that $P \in T_k$. Then we have:*

1. $\mathrm{ord}_P(g) \geq -\mathrm{ord}_P(mD_C)$;
2. *if $g_{mt_k} = f_{t_k} = 0$ has no solutions in $\mathbb{T}^2_{\mathbb{F}}$, then equality holds. Conversely, if equality holds for all $P \in T_k$, then $g_{mt_k} = f_{t_k} = 0$ has no solutions in $\mathbb{T}^2_{\mathbb{F}}$.*

PROOF. Let $p_k + \alpha$ be the integral point on t_k that is closest (but not equal) to p_k. Let $e_k = (a_k, b_k)$ be as above. Then $\alpha = (-b_k, a_k)$, since the vertices are enumerated clockwise. Choose a vector $\beta = (c,d)$ such that

$$\det \begin{pmatrix} -b_k & a_k \\ c & d \end{pmatrix} = -1.$$

Note that the cone $\Delta(t_k)$ is generated by $\alpha, -\alpha$ and β, so that

$$\mathbb{A}_{t_k} = \mathrm{Spec}\, \mathbb{F}[x', x'^{-1}, y'] \cong \mathbb{A}^2_{\mathbb{F}} \setminus \mathbb{A}^1_{\mathbb{F}}.$$

Here

$$\begin{array}{rcl} x' & = & x^{-b_k} y^{a_k} \\ y' & = & x^c y^d. \end{array} \tag{2.1}$$

Note that T_k corresponds to the locus of $y' = 0$ minus the origin. Since C intersects T_k transversally, we have that y' is a local parameter for C at P. Also note that x' is a unit in the local ring at P. The inverse transformation is given by

$$\begin{aligned} x &= x'^{-d} y'^{a_k} \\ y &= x'^{c} y'^{b_k} \end{aligned} \tag{2.2}$$

so that, using the notation $e'_k = (-d, c)$,

$$x^i y^j = x'^{e'_k \cdot (i,j)} y'^{e_k \cdot (i,j)}. \tag{2.3}$$

Using Lemma 2.10, we conclude that

$$g(x,y) = y'^{mN_k} (g_{mt_k}(x'^{-d}, x'^{c}) + y'(\ldots)). \tag{2.4}$$

The assertions follow. Indeed, $f_{t_k}(x'^{-d}, x'^{c})$ vanishes at P because (2.4) also holds for g replaced by f and $m = 1$. ■

2.12 Corollary *For $k = 1, \ldots, r$, we have that $\ell(t_k) = \#(T_k \cap C)$. In particular, $C \setminus \mathbb{T}^2_{\mathbb{F}}$ consists of $\#(\partial\Gamma \cap \mathbb{Z}^2)$ points.*

Proof. From the above proof, it follows that the points of $C \cap T_k$ correspond to the zeroes of $f_{t_k}(x'^{-d}, x'^{-c})$. Now the latter can be written as a power of x' times a degree $\ell(t_k)$ polynomial in x' with non-zero constant term and without multiple roots. ■

2.13 Corollary *For $(i,j) \in \mathbb{Z}^2$, we have that*

$$\mathrm{Div}_C(x^i y^j) = \sum_{k=1}^{r} (i,j) \cdot e_k(T_k \cap C),$$

which implies that $L_{m\Gamma} \subset \mathcal{L}(mD_C)$ for any $m \in \mathbb{N}_0$.

Proof. This follows immediately from Lemma 2.10 and equality (2.3). ■

An apparently new observation is that the canonical divisor class of C has an easy to describe representative.

2.14 Lemma *$D_C - W_C$ is a canonical divisor. More precisely*

$$D_C - W_C = \mathrm{Div}_C\left(\frac{dx}{xyf_y}\right).$$

In particular, $dx/(xyf_y)$ has no poles, nor zeroes on $C \cap \mathbb{T}^2_{\mathbb{F}}$.

PROOF. First, let P be a point of $C \setminus \mathbb{T}^2_{\mathbb{F}}$. We have to prove that

$$\mathrm{ord}_P \frac{dx}{xyf_y} = \mathrm{ord}_P D_C - 1\,.$$

With the notation as in Lemma 2.11, we have that $f_{t_k}(x'^{-d}(P), x'^c(P)) = 0$, where k is such that $P \in T_k$. Thus, because of the nondegeneracy of f:

$$\left(x\frac{\partial f_{t_k}}{\partial x}\right)(x'^{-d}(P), x'^c(P)) \neq 0 \quad \text{or} \quad \left(y\frac{\partial f_{t_k}}{\partial y}\right)(x'^{-d}(P), x'^c(P)) \neq 0\,.$$

We may suppose that the second condition holds. Indeed, the first case is treated analogously using that $dx/xyf_y = -dy/xyf_x$. Moreover, $\mathrm{ord}_P x$ is not a multiple of the characteristic p of $\mathbb{F}$. Indeed if it would, then from formulas (2.2) and the material above it, $a_k \equiv 0 \bmod p$ and $\alpha \equiv (-b_k, 0) \bmod p$ (if $p = 0$, these congruences become exact equalities). Hence f_{t_k} has a special form: it equals a monomial with exponent p_k times a Laurent polynomial with all exponents of y divisible by p. This Laurent polynomial vanishes at $(x'^{-d}(P), x'^c(P))$, because x' is a unit at P. This contradicts the assumed second condition on $\frac{\partial f_{t_k}}{\partial y}$.

Now apply Lemma 2.11 (and its proof) with g replaced by yf_y to find that $\mathrm{ord}_P yf_y = -\mathrm{ord}_P(D_C)$. Since $\mathrm{ord}_P x$ is not divisible by p, we have that $\mathrm{ord}_P dx/x = -1$ and the result follows.

Next, take $P \in C \cap \mathbb{T}^2_{\mathbb{F}}$. Write $P = (p_x, p_y)$. Because of the nondegeneracy we have that $\frac{\partial f}{\partial x}(P) \neq 0$ or $\frac{\partial f}{\partial y}(P) \neq 0$. In particular, $dx/xyf_y = -dy/xyf_x$ can have no pole at P. For the same reason, $x - p_x$ or $y - p_y$ must be local parameters at P so that for instance $dx/xyf_y = d(x-p_x)/xyf_y$ can have no zero at P. ■

2.15 Corollary $\deg D_C - \deg W_C = 2g - 2$, *where g is the genus of C.*

PROOF. From the Riemann-Roch theorem it follows that the degree of a canonical divisor is $2g - 2$. ■

The above observation allows us to give an elementary proof of the following well-known fact. See [65] for much more general theorems on this matter.

2.16 Corollary $g = \#((\Gamma \setminus \partial\Gamma) \cap \mathbb{Z}^2)$.

PROOF. By Corollary 2.12 and Corollary 2.15 we know that $\deg D_C = 2g - 2 + \#(\partial\Gamma \cap \mathbb{Z}^2)$. On the other hand, Pick's theorem [48] states that

$$\mathrm{Vol}(\Gamma) = \#((\Gamma \setminus \partial\Gamma) \cap \mathbb{Z}^2) + \frac{\#(\partial\Gamma \cap \mathbb{Z}^2)}{2} - 1. \tag{2.5}$$

Therefore, it suffices to prove that $\deg D_C = 2\mathrm{Vol}(\Gamma)$. For every edge t_k, consider the triangle Δ_k determined by the two vertices of t_k and the origin. If the origin happens to be one of the vertices, this is just a line segment. Then

$$\mathrm{Vol}(\Gamma) = \sum_k -\mathrm{sgn}(N_k)\mathrm{Vol}(\Delta_k).$$

Now Δ_k is a triangle with base $\ell(t_k)\|e_k\|$ (the length of t_k) and height $|p_k \cdot e_k|/\|e_k\|$, so that its volume equals $\ell(t_k)|N_k|/2$. The result follows. ■

We note that the inequality $g \leq \#((\Gamma \setminus \partial\Gamma) \cap \mathbb{Z}^2)$ holds in any case, i.e. without the nondegeneracy condition. This is *Baker's formula*, over $\mathbb{C}$ it was known already in 1893 [5]. A proof of the general case can be found in [8].

2.17 Corollary $\chi(C \cap \mathbb{T}_{\mathbb{F}}^2) = -2\mathrm{Vol}(\Gamma)$, *where* χ *is the Euler-Poincaré characteristic.*

Proof. Using Corollary 2.12 and Corollary 2.16, this is just a reformulation of Pick's theorem (2.5). ■

We conclude this section with the following theorem. It is a consequence of the fact that $H^i(X_\Gamma, \mathcal{E}) = 0$ for any $i \geq 1$ and any invertible sheaf $\mathcal{E}$ on X_Γ which is generated by its global sections (see [19, Corollary 7.3 and Proposition 6.7]). But for the convenience of the reader we will give an elementary proof.

2.18 Theorem *For any* $m \in \mathbb{N}_0$, *the Riemann-Roch space* $\mathcal{L}(mD_C)$ *is precisely given by* $L_{m\Gamma}$.

Proof. For this proof, the abuse of notation mentioned at the beginning of this section is a bit annoying. Therefore, we temporarily introduce the notation $\mathcal{A}_m$, which denotes the image of $L_{m\Gamma}$ *inside* the function field $\mathbb{F}(C)$. Note that the actual statement of the theorem should then be: $\mathcal{L}(mD_C) = \mathcal{A}_m$.

Corollary 2.13 states that $\mathcal{A}_m \subset \mathcal{L}(mD_C)$, so it suffices to prove that the dimensions are equal. Because $\deg D_C > 2g-2$ (Corollary 2.15), the Riemann-Roch theorem implies that $\dim \mathcal{L}(mD_C) = m \deg D_C + 1 - g$ for $m \in \mathbb{N}_0$. Note that this is a polynomial of degree 1 in m. Now consider the maps

$$r_m : L_{(m-1)\Gamma} \to L_{m\Gamma} : w \mapsto wf, \quad m \geq 1.$$

We claim that coker $r_m \cong \mathcal{A}_m$. Indeed, we will show that the natural map

$$\text{coker } r_m \to \mathcal{A}_m$$

is injective. Let $v \in L_{m\Gamma}$ be such that $v = 0$ in the function field. Then there is a unique Laurent polynomial q such that $v = fq$. Now for any $k \in \{1, \ldots, r\}$, we have that

$$\mathrm{ord}_{T_k} v = \mathrm{ord}_{T_k} f + \mathrm{ord}_{T_k} q.$$

Here, ord_{T_k} is the valuation at T_k in $\mathbb{P}_\Gamma$ (which is nonsingular in codimension one). From formula (2.4) one deduces that $\mathrm{ord}_{T_k} v \geq mN_k$ (indeed, y' is a local parameter at T_k). Similarly, we have that $\mathrm{ord}_{T_k} f = N_k$. Therefore, $\mathrm{ord}_{T_k} q \geq (m-1)N_k$. By a similar argument, now using (2.3), we conclude that $q \in L_{(m-1)\Gamma}$, which proves the claim. Now by a well-known result of Ehrhart

[32], $\dim L_{m\Gamma}$ is a quadratic polynomial in m with leading coefficient $\mathrm{Vol}(\Gamma)$ for $m \geq 0$. As a consequence,

$$\dim \mathcal{A}_m = \dim \operatorname{coker} r_m = \dim L_{m\Gamma} - \dim L_{(m-1)\Gamma}$$

is just like $\dim \mathcal{L}(mD_C)$ a linear polynomial in m for $m \geq 1$. Therefore it suffices to prove equality for $m = 1$ and $m \to \infty$.

The case $m = 1$ follows from the foregoing observations. Indeed,

$$\dim \mathcal{L}(D_C) = \deg D_C + 1 - g = 2g - 2 + \#(\partial\Gamma \cap \mathbb{Z}^2) + 1 - g = \#(\Gamma \cap \mathbb{Z}^2) - 1\,,$$

which is precisely $\dim \mathcal{A}_1$.

For the case $m \to \infty$ it suffices to prove that

$$\deg D_C = \lim_{m\to\infty} \frac{\dim L_{m\Gamma} - \dim L_{(m-1)\Gamma}}{m}\,.$$

Since $\dim L_{m\Gamma} = \mathrm{Vol}(\Gamma)m^2 + \ldots$, the right hand side is $2\mathrm{Vol}(\Gamma)$ which is indeed

$$2\#(\Gamma \setminus \partial\Gamma \cap \mathbb{Z}^2) + (\partial\Gamma \cap \mathbb{Z}^2) - 2 = 2g - 2 + \deg W_C$$

according to Pick's theorem [48], Corollary 2.12 and Corollary 2.16. ■

2.19 Corollary *$V(f)$ has a projective embedding for which the Hilbert polynomial $h(m)$ is precisely given by $\dim_{\mathbb{F}} \mathcal{L}(mD_C)$ (for $m \geq 1$).*

Proof. In the spirit of Theorem 2.8, one can actually show that $V(f)$ is isomorphic to Proj S, where

$$S = \frac{S_\Gamma^{\mathbb{F}}}{(tf)}.$$

See again [7] for more details. The result then follows from the above theorem (and its proof). ■

2.4 Cohomology of nondegenerate curves

Throughout this section, assume that $\mathbb{F}$ **is of characteristic** 0. Let $f \in \mathbb{F}[\mathbb{Z}^2]$ be nondegenerate with respect to its Newton polytope Γ, and **suppose that** Γ **contains** $(0,0)$ or equivalently: suppose that $D_{C,\Gamma} \geq 0$. This can always by achieved by multiplying f with an appropriate Laurent monomial. We will consider the *algebraic de Rham cohomology* of the coordinate ring $A = \frac{\mathbb{F}[\mathbb{Z}^2]}{(f)}$ of the affine curve defined by f. That is, consider the complex of $\mathbb{F}$-vector spaces

$$D^{-1}(A) \xrightarrow{d} D^0(A) \xrightarrow{d} D^1(A) \xrightarrow{d} D^2(A) \xrightarrow{d} \ldots$$

where $D^{-1}(A) = 0$, $D^0(A) = A$, $D^1(A)$ is the universal space of differentials of A over $\mathbb{F}$, $D^i(A) = \wedge_A^i D^1(A)$ for $i = 2, 3, \ldots$ and d is the usual exterior derivation. Then

$$H^i_{DR}(f/\mathbb{F}) := \frac{\ker d : D^i(A) \to D^{i+1}(A)}{\operatorname{im} d : D^{i-1}(A) \to D^i(A)} \quad \text{for } i = 0, 1, 2, \ldots$$

One can show that $H^0_{DR}(f/\mathbb{F}) = \mathbb{F}$, $H^i_{DR}(f/\mathbb{F}) = 0$ for $i \geq 2$ and

$$\dim_{\mathbb{F}} H^1_{DR}(f/\mathbb{F}) = 2g + R - 1,$$

where $g = \#((\Gamma \setminus \partial\Gamma) \cap \mathbb{Z}^2)$ is the genus of $V(f)$ and $R = \#((\partial\Gamma) \cap \mathbb{Z}^2)$ is the number of points on $V(f) \setminus \mathbb{T}^2_{\mathbb{F}}$. By Pick's theorem, we conclude that $\dim_{\mathbb{F}} H^1_{DR}(f/\mathbb{F}) = 2\mathrm{Vol}(\Gamma) + 1$.

The aim of this section is to prove the following 'sparse' description of $H^1_{DR}(f/\mathbb{F})$. A related description, for nondegenerate hypersurfaces of any dimension, is contained in [6, Corollary 6.10 and Theorem 7.13] and [7, Theorem 11.5]. The methods in these papers are very different from what we use in the simple proof below, which only works for curves.

2.20 Theorem *Let $D : \mathbb{F}[\mathbb{Z}^2] \to \mathbb{F}[\mathbb{Z}^2]$ be the operator $xy\left(\frac{\partial f}{\partial y}\frac{\partial}{\partial x} - \frac{\partial f}{\partial x}\frac{\partial}{\partial y}\right)$. Then we have a natural map*

$$\frac{L_{2\Gamma}}{fL_\Gamma + D(L_\Gamma)} \to H^1_{DR}(f/\mathbb{F})$$

which is in fact an isomorphism.

Proof. First note that D satisfies the Leibniz rule and that $Df = 0$, so that D is well-defined on A. Consider the map

$$\Lambda : A \to D^1(A) : h \mapsto h\frac{dx}{xyf_y}.$$

Let $\alpha, \beta \in A$ be such that $1 = \alpha f_x + \beta f_y$. Then $dx/f_y = \beta dx - \alpha dy$, which shows that Λ is well-defined. It is clearly injective, and since for any $g_1, g_2 \in A$ we have that $\Lambda(xy(f_y g_1 - f_x g_2)) = g_1 dx + g_2 dy$, it is in fact a bijection. One can check that $\Lambda \circ D = d$, so that

$$\frac{A}{D(A)} \xrightarrow{\Lambda} \frac{D^1(A)}{d(A)} \tag{2.6}$$

is a well-defined isomorphism. We will show that the canonical map

$$\frac{L_{2\Gamma}}{fL_\Gamma + D(L_\Gamma)} \to \frac{A}{D(A)} \tag{2.7}$$

is also an isomorphism, so that the composition of both is the requested natural map. We first prove injectivity. Suppose that the image of $h \in L_{2\Gamma}$ in A (which we also denote with h) equals $D(g)$ for some $g \in A$. Then

$$\mathrm{Div}_C dg = \mathrm{Div}_C \Lambda(h) \geq -2D_C + D_C - W_C = -D_C - W_C$$

by Theorem 2.18 and Lemma 2.14. Since $D_C \geq 0$, we conclude that $\mathrm{Div}_C g \geq -D_C$ and hence, again by Theorem 2.18, that $g \in L_\Gamma$ (modulo f). Finally, using the convexity of Γ it is easy to see that two polynomials in $L_{2\Gamma}$ that differ by a multiple of f actually differ by an element of fL_Γ.

Next, we prove surjectivity. We need to show that every $h \in A$ is equivalent modulo $D(A)$ with an $h' \in L_{2\Gamma}$. Using Theorem 2.18, it is then sufficient to show that the well-defined map

$$\mathcal{L}(D_C + E) \xrightarrow{D} \frac{\mathcal{L}(2D_C + E)}{\mathcal{L}(2D_C)} \tag{2.8}$$

is surjective for any effective divisor E having support in $C \setminus \mathbb{T}^2_{\mathbb{F}}$ which is defined over $\mathbb{F}$. Finally, using the perfectness of $\mathbb{F}$, it suffices[2] to show that the map

$$\mathcal{L}_{\overline{\mathbb{F}}}(D_C + E) \xrightarrow{D} \frac{\mathcal{L}_{\overline{\mathbb{F}}}(2D_C + E)}{\mathcal{L}_{\overline{\mathbb{F}}}(2D_C)} \tag{2.9}$$

of $\overline{\mathbb{F}}$-vector spaces is surjective. Write $D_C = \sum_{k=1}^r a_k P_k$ (with all $a_k \geq 0$). Note that $\deg D_C = \sum_{k=1}^r a_k > 2g - 2$ due to Corollary 2.15. Suppose $h \in \mathcal{L}_{\overline{\mathbb{F}}}(2D_C + E)$ has a pole of order $b_k > 2a_k$ at some place P_k. Then $\Lambda(h)$ has a pole of order $b'_k > a_k + 1$ at P_k (where $b'_k = b_k - a_k + 1$). Because of the Riemann-Roch theorem, we can find a function h_0 in

$$\mathcal{L}_{\overline{\mathbb{F}}}(a_1P_1 + \cdots + (b'_k - 1)P_k + \cdots + a_rP_r) \setminus \mathcal{L}_{\overline{\mathbb{F}}}(a_1P_1 + \cdots + (b'_k - 2)P_k + \cdots + a_rP_r).$$

Since we are working in characteristic 0 and since $b'_k - 1 > a_k \geq 0$, this means that dh_0 has a pole of order b'_k at P_k and a pole of order at most a_i+1 at the other places P_i. Applying Λ^{-1} then gives that Dh_0 has a pole of order b_k at P_k and a pole of order at most $2a_i$ at the other places P_i. Subtracting from h a suitable multiple of Dh_0 thus reduces the pole order at P_k. Note that $h_0 \in \mathcal{L}_{\overline{\mathbb{F}}}(D_C + E)$. Continuing in this way eventually results in an $h_1 \in \mathcal{L}_{\overline{\mathbb{F}}}(D_C + E)$ such that $h' = h - Dh_1 \in \mathcal{L}_{\overline{\mathbb{F}}}(2D_C)$. ■

2.5 Further properties

We present three other new stand-alone properties of nondegenerate Laurent polynomials. The latter two will be indispensable in our treatment of p-adic cohomology for nondegenerate curves (Chapter 4) and play a crucial role in the development of our point counting algorithm; the proofs will be given in Subsection 3.4.2.

[2] This can be seen by using that $\mathcal{L}_{\overline{\mathbb{F}}}(D_C + E)$ is generated by $\mathcal{L}(D_C + E)$ and by comparing dimensions.

2.21 Lemma *Any C_{ab} curve*[3] *C over a field $\mathbb{F}$ with at least $(a+1)(b+1)$ elements has a nondegenerate model with $\Gamma = \mathrm{Conv}\{(0,0),(0,a),(b,0)\}$ as Newton polytope.*

PROOF. We may suppose that C is given by a Weierstrass equation

$$f(x,y) = \alpha_{0,a}y^a + \alpha_{b,0}x^b + \sum_{ai+bj<ab} \alpha_{i,j}x^i y^j \quad \in \mathbb{F}[\mathbb{N}^2].$$

We want to find x_0 and y_0 in $\mathbb{F}$ such that the curve in $\mathbb{A}^2_{\mathbb{F}}$ defined by $f'(x,y) = f(x-x_0, y-y_0) = 0$ is non-tangent to both coordinate axes and does not contain the origin. Indeed, the latter condition guarantees that the Newton polytope of f' is Γ. Being non-tangent to the coordinate axes then precisely corresponds to the nondegeneracy conditions with respect to the edges adjacent to $(0,0)$. The nondegeneracy condition with respect to the edge spanned by $(0,a)$ and $(b,0)$ is automatically fulfilled.

Recall that the genus of a C_{ab} curve is given by

$$g = \frac{(a-1)(b-1)}{2}.$$

Using Hurwitz's theorem [50, Corollary 2.4] one can then check that the number of points with a vertical tangent is bounded by $(a-1)(b+1)$. Therefore we can find an x_0 such that the curve defined by $f(x-x_0, y)$ is non-tangent to the y-axis. Similarly, there are at most $(a+1)(b-1)$ points with a horizontal tangent line; moreover there are at most a points with x-coordinate x_0. In conclusion, if $\#\mathbb{F} \geq (a+1)(b+1)$, we can find the requested x_0 and y_0. ■

As a consequence, the class of nondegenerate curves may be looked at as a generalization of the class of C_{ab} curves, and hence of the class of hyperelliptic curves. This is definitely true from the point counting viewpoint, where the condition $\#\mathbb{F} \geq (a+1)(b+1)$ is always satisfied in practice (if $\#\mathbb{F} < (a+1)(b+1)$ then the claimed running time can be attained using brute force counting).

2.22 Lemma (Effective Nullstellensatz) *Let R be a field or a discrete valuation ring, and denote its residue field with k. Take $f \in R[\mathbb{Z}^2]$ and suppose that f and its reduction $\overline{f} \in k[\mathbb{Z}^2]$ have the same Newton polytope Γ, which we suppose to be two-dimensional and to contain the origin. If $\overline{f}$ is nondegenerate with respect to Γ, there exist $\alpha, \beta, \gamma \in R[\mathbb{Z}^2]$ such that $\Gamma(\alpha), \Gamma(\beta), \Gamma(\gamma) \subset 2\Gamma(f)$ and*

$$1 = \alpha f + \beta x \frac{\partial f}{\partial x} + \gamma y \frac{\partial f}{\partial y}.$$

[3] C_{ab} curves ($a, b \in \mathbb{N}_0$) are curves having at least one rational point P for which the monoid

$$\{ -\mathrm{ord}_P(f) \mid \exists m \in \mathbb{N} \text{ such that } f \in \mathcal{L}(mP)\}$$

equals $a\mathbb{N} + b\mathbb{N}$. They were introduced by Miura [83] and generalize the class of hyperelliptic curves (which are $C_{2,2g+1}$ curves). See [76] and [106] for short surveys and more references.

PROOF. See Lemma 3.16. ■

2.23 Lemma (Lifting preserves nondegeneracy) *Let R be a discrete valuation ring with residue field k and fraction field $\mathbb{K}$. Let $f \in R[\mathbb{Z}^2]$ and suppose that f and its reduction $\overline{f} \in k[\mathbb{Z}^2]$ have the same Newton polytope. If $\overline{f}$ is nondegenerate with respect to its Newton polytope, then so is f (when considered over $\mathbb{K}$).*

PROOF. See Lemma 3.17. ■

If $\overline{f}$ is *not* nondegenerate, one could wonder wether similar statements hold. For instance, is a Newton polytope preserving lift of a nonsingular curve still nonsingular? The following example shows that the answer is no.

2.24 Example Let p be a prime number and consider $\mathbb{Z}_p$, the valuation ring of $\mathbb{Q}_p$, the field of p-adic numbers. Denote its residue field with $\mathbb{F}_p$. Take $\alpha, \beta \in \mathbb{Z}_p \setminus p\mathbb{Z}_p$ such that $\alpha + \beta = p^p$. Consider

$$f = \alpha(x^p - x^{p-1}) + \beta(y^p - y^{p-1}) + (p-1)^{p-1}.$$

Then the curve defined by f has a singular point at $\left(\frac{p-1}{p}, \frac{p-1}{p}\right)$, though its reduction mod p defines a nonsingular curve (even on $\mathbb{A}^2_{\mathbb{F}_p}$). ■

2.6 Nondegenerate curves and Kedlaya's method

In this informal section, we explain why nondegenerate curves are particularly well-suited objects for point counting methods à la Kedlaya.

First, let us take a look at the hyperelliptic curve case. Let $y^2 - \overline{Q}(x)$ define a genus g hyperelliptic curve over a finite field $\mathbb{F}_q$ of odd characteristic. Here $\overline{Q}(x) \in \mathbb{F}_q[x]$ is a degree $2g+1$ polynomial without multiple roots. Let $\mathbb{Z}_q$ be a complete discrete valuation ring with residue field $\mathbb{F}_q$ and fraction field $\mathbb{Q}_q$ and let $Q \in \mathbb{Z}_q[X]$ be a degree preserving lift of $\overline{Q}$. Then Kedlaya effectively proves that the canonical maps

$$H^i_{DR}(y^2 - Q(x)/\mathbb{Q}_q) \longrightarrow H^i_{MW}(y^2 - \overline{Q}(x)/\mathbb{Q}_q) \tag{2.10}$$

are isomorphisms, which allows him to compute[4] in $H^1_{DR}(y^2 - Q(x)/\mathbb{Q}_q)$. This is a crucial aspect of Kedlaya's algorithm.

An essential role in this is played by the geometric correspondence between $y^2 - \overline{Q}(x)$ and $y^2 - Q(x)$: the latter also defines a genus g hyperelliptic curve

[4] In fact, he works with $\overline{C}'$ (resp. C'), which is obtained from $\overline{C} : y^2 = \overline{Q}(x)$ (resp. $C' : y^2 = Q(x)$) by removing the Weierstrass points. But this is only for technical reasons (to make the Frobenius endomorphism easier to lift).

with one Weierstrass point at ∞. Indeed, this follows from the fact that the discriminant of Q is non-zero since its reduction to $\mathbb{F}_q$ is precisely the discriminant of $\overline{Q}$.

Lemma 2.23 should be looked at as a generalization of this fact: any Newton polytope preserving lift $f \in \mathbb{Z}_q[\mathbb{Z}^2]$ of a nondegenerate polynomial $\overline{f} \in \mathbb{F}_q[\mathbb{Z}^2]$ is again nondegenerate. Now the geometry of nondegenerate curves is largely determined by the Newton polytope: the genus equals the number of interior lattice points (Corollary 2.16), the number of places at infinity equals the number of lattice points on the boundary (Corollary 2.12), and so on. So there is a deep geometric correspondence between $\overline{f}$ and f that is directed by the Newton polytope.

This allows us to apply the following very general theorem [61, Theorem 1].

2.25 Theorem *Let Y be a smooth proper $\mathbb{Z}_q$-scheme, let $Z \subset Y$ be a relative normal crossings divisor and let $X = Y \setminus Z$. If X is affine, then for any $i \in \mathbb{N}$ there exists a canonical isomorphism*

$$H^i_{DR}(X \otimes_{\mathbb{Z}_q} \mathbb{Q}_q/\mathbb{Q}_q) \to H^i_{MW}(X \otimes_{\mathbb{Z}_q} \mathbb{F}_q/\mathbb{Q}_q)\,. \tag{2.11}$$

Indeed, this applies in our situation with $X = \operatorname{Spec} \frac{\mathbb{Z}_q[\mathbb{Z}^2]}{(f)}$ and Y its closure in the toric scheme associated to $\Gamma(f)$ (which is constructed exactly as in Section 2.2, with $\mathbb{F}$ replaced by $\mathbb{Z}_q$). This is a smooth proper scheme and by the above observations $Z = Y \setminus X$ is indeed a relative normal crossings divisor. In Chapter 4 we will give an alternative proof of this theorem (in the case of nondegenerate curves). Along with this, we will prove an explicit bound on the p-adic denominators that are introduced during differential reduction, thereby generalizing Kedlaya's result to nondegenerate curves.

2.26 Remark What happens if the polynomial $\overline{f}$ is *not* nondegenerate? In fact, the original purpose of our research was to treat general plane curves, but this proved to be more difficult than originally expected. Indeed, because of Corollary 2.4, a randomly chosen lift $f \in \mathbb{Z}_q[\mathbb{Z}^2]$ is very likely to be nondegenerate. Since the geometry of the latter is largely determined by the Newton polytope (while the geometry of $\overline{f}$ is not), both curves may have completely different geometric properties, which makes the cohomology spaces harder to describe[5]. At the moment, it is unclear to us whether or not one can always find a carefully chosen lift f such that the canonical map

$$H^1_{DR}(f/\mathbb{Q}_q) \to H^1_{MW}(\overline{f}/\mathbb{Q}_q)$$

is an isomorphism (and if so, whether this lift is effectively computable). ■

Another feature of nondegenerate curves concerns the lifting of the Frobenius endomorphism $\overline{\mathcal{F}}_q$ on $\frac{\mathbb{F}_q[\mathbb{Z}^2]}{(\overline{f})}$ to the weak completion of $\frac{\mathbb{Z}_q[\mathbb{Z}^2]}{(f)}$. In [25], Denef and

[5] Even if the lift is not nondegenerate, the type of nondegeneracy may be completely different (see Example 2.24).

Vercauteren present a method that works in principle for arbitrary nonsingular plane curves (even for arbitrary hypersurfaces). The rate of convergence of this lift is related to the bounds appearing in the effective Nullstellensatz, which for nondegenerate curves take a particularly nice form (Lemma 2.22). As a result, we obtain a lift $\mathcal{F}_q$ of $\overline{\mathcal{F}}_q$ satisfying a good convergence rate in which the Newton polytope plays a very natural role. We refer to Chapter 4 for the details.

Finally, as explained in Section 2.4, we use the nondegeneracy of f to describe its first algebraic de Rham cohomology as a space of *functions* instead of *differential forms*. The translation between these two worlds is given by the map Λ appearing in the proof of Theorem 2.20. It will turn out that the natural convergence properties of $\mathcal{F}_q$ behave well under this translation, again due to the effective Nullstellensatz (Lemma 2.22). This will allow us to prove a 'sparse Lefschetz fixed point theorem' (Corollary 4.19), on which our point counting method is based.

Chapter 3

The effective Nullstellensatz problem for discrete valuation rings

In this self-contained chapter, we will treat a number of topics that are related to the effective Nullstellensatz problem. We give a very brief overview of what is known in the field case and introduce the analogous problem for discrete valuation rings, where the situation turns out to be more tricky. The key result is a sparse effective Nullstellensatz for fields and local rings [13] that seems new even over $\mathbb{C}$. The reader trying to situate this chapter in the scope of the whole thesis should keep the fields $\mathbb{F}_q$ and $\mathbb{Q}_q$ and the discrete valuation ring $\mathbb{Z}_q$ in mind. The results presented in Section 3.3 will then mainly be used to construct an effective lift of the Frobenius endomorphism in Chapter 4.

We note that a slightly related effective Nullstellensatz problem has also been studied for rings of algebraic integers in a number field, such as $\mathbb{Z}$. This is often called the *arithmetic* Nullstellensatz problem. We refer to [91] and [68] for more details. The results presented there will be used in Section 3.2 to prove a restricted effective Nullstellensatz over $\mathbb{Z}_q$.

Throughout this chapter, let r and n be fixed nonzero natural numbers and let $x_1, \ldots, x_n$ be formal variables. For any ring A, we will often use the abbreviations $A[\mathbb{N}^n] := A[x_1, \ldots, x_n]$ and $A[\mathbb{Z}^n] := A[x_1^{\pm 1}, \ldots, x_n^{\pm 1}]$. Also, we will use vector exponent notation: $x^{(i_1,\ldots,i_n)}$ stands for $x_1^{i_1} \cdots x_n^{i_n}$. Finally, if $\mathbb{F}$ is a field, $\overline{\mathbb{F}}$ will denote an algebraic closure.

3.1 Introduction

A cornerstone in algebraic geometry is Hilbert's Nullstellensatz, which essentially states the following.

3.1 Theorem (Nullstellensatz for fields) *Let $\mathbb{F}$ be a field and suppose that $f_1, \ldots, f_r \in \mathbb{F}[\mathbb{N}^n]$ have no common solution in $\overline{\mathbb{F}}^n$. Then there exist polynomials $g_1, \ldots, g_r \in \mathbb{F}[\mathbb{N}^n]$ such that $1 = g_1 f_1 + \cdots + g_r f_r$.*

PROOF. See for instance [4, Exercise 7.14]. ■

Of course, the g_i provided by the above theorem are generally non-unique and for many applications, both practical and theoretical, one would like to take the smallest possible ones. Here, 'small' could refer to the degrees of the g_i, or to their Newton polytopes (see Section 3.3), or (if this makes sense) to the number of bits needed to represent their coefficients, and so on. Is it possible to give a *general* bound, predicting the sizes of these smallest g_i in terms of the sizes of the f_i only? Are there interesting subclasses of systems $f_1, \ldots, f_r$ that allow a significantly better bound? Questions of this type are collected in what is called the *effective Nullstellensatz problem*, and are often harder than one would expect at first sight. Note that the classical proofs of Theorem 3.1 are all non-constructive and do not provide any additional information about the g_i.

The first positive answer to a question of the above type dates back to 1926. In her Ph. D. thesis [51], Hermann proved that we can always choose the g_i such that

$$\max_i \deg g_i f_i \leq 2(2d)^{2^{n-1}},$$

where $d = \max_i \deg f_i$. This bound is fairly good for $n = 1$ and $n = 2$, but for large n it becomes very rough. In the late 1980's, a renewed interest in the subject led to dramatical improvements, culminating in the result of Kollár [67] (which was independently proven by Fitchas and Galligo [35]):

$$\max_i \deg g_i f_i \leq \max\{3, d\}^n. \tag{3.1}$$

This bound is sharp whenever $d \geq 3$, as is illustrated by the following famous example [67, Example 2.3]. In the $d = 2$ case, Sombra [101] improved Kollár's bound to 2^{n+1}.

3.2 Example (Kollár) Consider the system $f_1 := x_1^d, f_2 := x_1 x_n^{d-1} - x_2^d, f_3 := x_2 x_n^{d-1} - x_3^d, \ldots, f_{n-1} := x_{n-2} x_n^{d-1} - x_{n-1}^d, f_n := x_{n-1} x_n^{d-1} - 1$, which has clearly no solutions. Take a Nullstellensatz expansion $1 = g_1 f_1 + \cdots + g_n f_n$ and consider the ideal $I \subset \mathbb{F}[x_1, \ldots, x_n, z]$ generated by the homogenizations of the f_i with respect to a new variable z. It is easily seen that $\max_i \deg g_i f_i$ is bounded from below by the least power of z that is contained in I. Now consider the image of I under the map

$$\mathbb{F}[x_1, \ldots, x_n, z] \to \mathbb{F}[x_1, \ldots, x_{n-1}, z] : f(x_1, \ldots, x_n, z) \mapsto f(x_1, \ldots, x_{n-1}, 1, z),$$

which is

$$(x_1^d, x_1 - x_2^d, x_2 - x_3^d, \ldots, x_{n-2} - x_{n-1}^d, x_{n-1} - z^d).$$

Clearly no power of z less than z^{d^n} can be contained in this ideal, and a fortiori the same holds for I. ■

Although Kollár's bound is optimal, many refinements have been made by giving bounds in terms of the Newton polytopes of the f_i. See for instance [101] for results of this type and for more references.

3.2 The DVR analogue

Now, let R be a discrete valuation ring with local parameter t. Denote by $k = \frac{R}{(t)}$ its residue field and by $\mathbb{K}$ its fraction field.

3.3 Proposition (Nullstellensatz for DVR's) *Suppose that $f_1, \ldots, f_r \in R[\mathbb{N}^n]$ are such that they have no common zero in $\overline{\mathbb{K}}^n$. Suppose moreover that their reductions mod t have no common zero in $\overline{k}^n$. Then there exist polynomials $g_1, \ldots, g_r \in R[\mathbb{N}^n]$ for which $1 = g_1 f_1 + \cdots + g_r f_r$.*

Though this is well-known[1], we give the following inductive proof (using the field case).

Proof. There are no common solutions in $\overline{\mathbb{K}}^n$, so we can find polynomials $\widetilde{g}_1, \ldots, \widetilde{g}_r \in \mathbb{K}[\mathbb{N}^n]$ such that

$$1 = \widetilde{g}_1 f_1 + \cdots + \widetilde{g}_r f_r. \tag{3.2}$$

Similarly, since the reductions $\overline{f}_1, \ldots, \overline{f}_r$ have no common solution in $\overline{k}^n$, we can find polynomials $g_1, \ldots, g_r \in R[\mathbb{N}^n]$ such that

$$1 \equiv g_1 f_1 + \cdots + g_r f_r \mod t. \tag{3.3}$$

Clearing denominators in (3.2) yields

$$t^m = g_1' f_1 + \cdots + g_r' f_r \tag{3.4}$$

for some $m \in \mathbb{N}$ and $g_1', \ldots, g_r' \in S$. If $m = 0$, we are done. If not, we can reduce m inductively as follows: rewrite (3.3) as $1 + Qt = g_1 f_1 + \cdots + g_r f_r$ for some $Q \in S$, multiply this equation with t^{m-1}, multiply (3.4) with Q, and subtract. ■

So, as far as we are concerned with the *existence* of the g_i only, the Nullstellensatz for DVR's is no more than a simple corollary to the classical field version. However, the *effective* Nullstellensatz problem suddenly becomes a lot harder to tackle. The following simple example[2] shows that it is *impossible* to

[1] In fact, even a much more general statement holds: *if A is a commutative ring and $f_1, \ldots, f_r \in A[\mathbb{N}^n]$ are such that for every prime ideal $\mathfrak{m} \subset A$ the reductions $\overline{f}_1, \ldots, \overline{f}_r \in (A/\mathfrak{m})[\mathbb{N}^n]$ have no common zero in $\overline{\mathrm{Frac}(A/\mathfrak{m})}$, then $(f_1, \ldots, f_r) = A[\mathbb{N}^n]$.* Indeed, suppose this were not true, then $(f_1, \ldots, f_r)$ is contained in some maximal ideal $M \subset A[\mathbb{N}^n]$. Let $\mathfrak{m} = M \cap A$, this is a prime ideal of A. Then the reductions $\overline{f}_1, \ldots, \overline{f}_r \in (A/\mathfrak{m})[\mathbb{N}^n]$ clearly share a solution in the extension field $A[\mathbb{N}^n]/M \supset \mathrm{Frac}(A/\mathfrak{m})$. Using the classical Nullstellensatz we conclude that they share a solution in $\overline{\mathrm{Frac}(A/\mathfrak{m})}$.

[2] It is inspired by an example in [3].

give a general Nullstellensatz bound that depends on the degrees of the f_i, the number of polynomials, the number of variables and the t-adic valuations of the coefficients only.

3.4 Example Consider the system $1 - tx_1, 1 - (t^m + t)x_1 \in R[x_1]$, where m is an arbitrarily large natural number. It has no solutions, neither over the fraction field, nor over the residue field. Take an expansion

$$1 = g_1(1 - tx_1) + g_2(1 - (t^m + t)x_1)$$

and reduce modulo t^m to obtain the identity

$$1 = (\overline{g}_1 + \overline{g}_2)(1 - tx_1)$$

in $R/(t^m)[x_1]$. Since the inverse of $1 - tx_1$ is $1 + tx_1 + t^2x_1^2 + \cdots + t^{m-1}x_1^{m-1}$, we conclude that

$$\max\{\deg g_1, \deg g_2\} \geq \deg(g_1 + g_2) \geq \deg(\overline{g}_1 + \overline{g}_2) \geq m - 1.$$

■

The missing parameter to give a general Nullstellensatz bound is the t-adic valuation of a certain polynomial evaluated in the coefficients appearing in $f_1, \ldots, f_r$. In the following, by Σ_d we mean

$$\{(i_1, \ldots, i_n) \in \mathbb{N}^n \mid i_1 + \cdots + i_n \leq d\}.$$

3.5 Theorem *Let R be a DVR with local parameter t. For every $n, r, d \in \mathbb{N} \setminus \{0\}$ there exists a non-zero polynomial $G_{n,r,d} \in R[c_{k,i}]_{k=1,\ldots,r;i\in\Sigma_d}$ of degree*

$$\leq r\binom{\max\{3,d\}^n + n}{n}$$

for which the following holds. If

$$f_1 = \sum_{i\in\Sigma_d} C_{1,i}x^i, \quad \ldots, \quad f_r = \sum_{i\in\Sigma_d} C_{r,i}x^i$$

are polynomials in $R[\mathbb{N}^n]$ that have no common solution, neither over the fraction field of R, nor over its residue field, then there exist polynomials $g_1, \ldots, g_r \in R[\mathbb{N}^n]$ for which $1 = g_1f_1 + \cdots + g_rf_r$ and

$$\max_i \deg g_if_i \leq \max\{3,d\}^n(1 + \operatorname{ord}_t G_{n,r,d}(C_{k,i})).$$

Proof. Given such a system $f_1, \ldots, f_r$, we know from (3.1) that there exist $g'_1, \ldots, g'_r \in \mathbb{K}[\mathbb{N}^n]$ (where $\mathbb{K} = \operatorname{Frac}(R)$) of degree $\leq \max\{3,d\}^n$ for which $1 = g'_1f_1 + \cdots + g'_rf_r$. In other words, the formula

$$1 = \left(\sum_{i\in\Sigma_{\max\{3,d\}^n}} g'_{1,i}x^i\right) f_1 + \ldots$$

gives rise to a system $\mathcal{S}_{f_1,\dots,f_r}$ of linear equations in

$$u = r\binom{\max\{3,d\}^n + n}{n}$$

unknowns $g'_{k,i}$ that is solvable over $\mathbb{K}$. Let $\rho := \max \operatorname{rank}(\mathcal{S}_{f_1,\dots,f_r}) \leq u$, where the maximum is taken over all $f_1,\dots,f_r$ having degree $\leq d$. Let $f_{0k} = \sum C_{0,k,i}x^i$ $(k = 1,\dots,r)$ be a set of polynomials for which this rank is actually obtained. Then $\mathcal{S}_{f_{01},\dots,f_{0r}}$ has a non-zero $(\rho\times\rho)$-minor, which is a degree ρ polynomial expression in the $C_{0,k,i}$. Let $G_{n,r,d}(c_{k,i}) \in R[c_{k,i}]$ be the corresponding polynomial.

Now, using Cramer's rule, we can find a solution to $S_{f_{01},\dots,f_{0r}}$ such that the valuations of the denominators appearing in this solution are bounded by $\operatorname{ord}_t G_{n,r,d}(C_{0,k,i})$. In fact, this statement holds in general: for *any* $f_1,\dots,f_r$ as in the énoncé, we can find a solution to $S_{f_1,\dots,f_r}$ whose denominators are bounded by $\operatorname{ord}_t G_{n,r,d}(C_{k,i})$. Indeed, either $G_{n,r,d}(C_{k,i})$ equals zero, or it is a non-zero minor of maximal dimension of $S_{f_1,\dots,f_r}$. Now using the induction procedure described in the proof of Proposition 3.3 and again using (3.1) (but now over the residue field), we get the desired result. ■

3.6 Example A $G_{1,2,1}$ is given by $c_{1,0}c_{2,1} - c_{2,0}c_{1,1}$. Applying this to Example 3.4 gives the bound $3(m+1)$. In fact it gives $m+1$ since the Kollár bound $\max\{3,d\}^n$ can be replaced by 1 in the univariate linear case. Note that we already proved $m-1+1 = m$ to be a lower bound, so that at least in this case the bound presented in the theorem is fairly sharp. ■

Although the proof is more or less constructive, Theorem 3.5 is only of theoretical interest: if n and d get bigger, the polynomials $G_{n,r,d}$ soon become huge objects that are impossible to compute. In many situations however, there is no need to do so, since much better bounds hold. This is particularly the case if the reductions of $f_1,\dots,f_r \bmod t$ satisfy some generic condition that will be explained in Section 3.3. Another interesting case is when R is the valuation ring of a finite extension of the field of p-adic numbers $\mathbb{Q}_p$ (for some prime number p). In that case one can hope for the existence of a number field K such that $f_1,\dots,f_r \in \mathcal{O}_K[\mathbb{N}^n]$ and then make use of an arithmetic Nullstellensatz (for instance [68, Theorem 3.6] – the statement of this theorem will be given in the proof of the lemma below). An example of such an application is the following.

Let $\rho \in \mathbb{N}_0$ and consider $\mathbb{Q}_q = \mathbb{Q}_{p^\rho}$, the unique degree ρ unramified extension of $\mathbb{Q}_p$. Denote by $\mathbb{Z}_q$ its valuation ring. Let $\mathbb{F}_q$ be its residue field and let $\overline{r}(X)$ be a (monic) defining polynomial for $\mathbb{F}_q$ over $\mathbb{F}_p = \mathbb{Z}/(p)$, thus $\deg \overline{r}(X) = \rho$. Let $r(X) \in \mathbb{Z}[X]$ be a monic degree ρ polynomial that reduces to $\overline{r}(X)$ modulo (p). Let $\theta \in \mathbb{C}$ be a root of $r(X)$ and define $K = \mathbb{Q}(\theta)$. Let $\mathcal{O}_K$ be its ring of algebraic integers. Then one can easily check that $\mathfrak{p} = (p) \subset \mathcal{O}_K$ is a prime ideal and that $\mathbb{Z}_q$ can be identified with the $\mathfrak{p}$-adic completion of $\mathcal{O}_K$.

3.7 Lemma *Let $f_1, \ldots, f_r \in \mathbb{Z}_q[\mathbb{N}^n]$ be polynomials of maximal degree d and suppose they have no common solutions, neither over the algebraic closure of the fraction field $\mathbb{Q}_q$, nor over the algebraic closure of the residue field $\mathbb{F}_q$. Denote with $\mathcal{C}$ the set of non-zero coefficients appearing in the f_i and suppose that $\mathcal{C} \subset \mathbb{Z}[\theta] \subset \mathcal{O}_K$. Then all $c \in \mathcal{C}$ have a representation*

$$c = \sum_{i=0}^{\rho-1} a_{c,i}\theta^i, \quad a_{c,i} \in \mathbb{Z}.$$

Let $B \in \mathbb{N}_0$ be an upper bound for the absolute values of the $a_{c,i}$. Then there exist $g_1, \ldots, g_r \in \mathbb{Z}_q[\mathbb{N}^n]$ for which $1 = g_1 f_1 + \cdots + g_r f_r$ and

$$\begin{aligned} \max_i \deg f_i g_i &\leq 4nd^{n+1} + 4n(n+1)d^n\,[\rho \cdot \log(\rho \cdot B \cdot B') \\ &\quad + \log r + (n+7)\log(n+1)d] \cdot \max\{3,d\}^n \\ &= O(d^{2n+1}n^3\rho \log(n \cdot \rho \cdot B \cdot B')) \end{aligned}$$

where $B' \in \mathbb{N}_0$ is a strict upper bound for the absolute values of the coefficients of $r(X)$.

Proof. Let M_K^{arch} be the set of norms on K that extend the ordinary absolute value on $\mathbb{Q}$ and let $M_K^{\text{n-arch}}$ consist of all norms extending a π-adic norm (for some prime number π) on $\mathbb{Q}$. Let $M_K = M_K^{\text{arch}} \cup M_K^{\text{n-arch}}$. Consider the so-called *global height function*: if $\mathcal{A}$ is a finite subset of K, then the height of $\mathcal{A}$ is defined as

$$h(\mathcal{A}) = \frac{1}{\rho} \sum_{|\cdot| \in M_K} N_{|\cdot|} \max(\{\log|a| \mid a \in \mathcal{A}\} \cup \{0\}). \tag{3.5}$$

Here $N_{|\cdot|}$ equals the extension degree $[K_{|\cdot|} : \mathbb{Q}_{|\cdot|}]$ (where the lower index stands for completion).

Then by [68, Theorem 3.6] there exist an $a \in \mathcal{O}_K \setminus \{0\}$ and polynomials $g_1, \ldots, g_r \in \mathcal{O}_K[\mathbb{N}^n]$ of degree $\leq 4nd^n$ for which $a = g_1 f_1 + \cdots + g_r f_r$. Furthermore we have

$$h(a) \leq 4n(n+1)d^n(h(\mathcal{C}) + \log r + (n+7)\log(n+1)d). \tag{3.6}$$

Now since $\mathcal{C} \subset \mathcal{O}_K$, the only norms contributing to $h(\mathcal{C})$ are the archimedean ones, which are in correspondence with the complex roots of $r(X)$. By Cauchy's bound, the complex norm of any of these roots is bounded by B'. Therefore, any archimedean norm of any $c \in \mathcal{C}$ is bounded by $\rho \cdot B \cdot B'^{\rho}$, so

$$h(\mathcal{C}) \leq \rho \cdot \log(\rho \cdot B \cdot B'). \tag{3.7}$$

The next step is to bound $h(a)$ from below. We will make use of the following well-known product formula:

$$\prod_{|\cdot| \in M_K} |a|^{N_{|\cdot|}} = 1.$$

Denote $\mathrm{ord}_p a$ (i.e. its valuation in $\mathbb{Z}_q$) with s. We then have

$$\begin{aligned} h(a) &= \frac{1}{\rho} \sum_{|\cdot| \in M_K^{\mathrm{arch}}} N_{|\cdot|} \max\{\log|a|, 0\} \\ &\geq \frac{1}{\rho} \log \prod_{|\cdot| \in M_K^{\mathrm{arch}}} |a|^{N_{|\cdot|}} \\ &= \frac{1}{\rho} \log \left[\left(\prod_{|\cdot| \in M_K} |a|^{N_{|\cdot|}} \right) / \left(\prod_{|\cdot| \in M_K^{\text{n-arch}}} |a|^{N_{|\cdot|}} \right) \right] \\ &= \frac{-1}{\rho} \log \left(\prod_{|\cdot| \in M_K^{\text{n-arch}}} |a|^{N_{|\cdot|}} \right) \\ &\geq s. \end{aligned}$$

The last inequality holds because $[K_{|\cdot|_\mathfrak{p}} : \mathbb{Q}_{|\cdot|_p}] = [\mathbb{Q}_q : \mathbb{Q}_p] = \rho$.

Combining the bounds for $h(a)$ and $h(\mathcal{C})$ with (3.6) and moving on to $\mathbb{Z}_q$ yields that there are $g_1, \ldots, g_r \in \mathbb{Z}_q[\mathbb{N}^n]$ of degrees $\leq 4nd^n$ and an $s \in \mathbb{N}$ that satisfies

$$s \leq 4n(n+1)d^n(\rho \cdot \log(\rho \cdot B \cdot B') + \log r + (n+7)\log(n+1)d)$$

for which

$$p^s = g_1 f_1 + \cdots + g_r f_r.$$

Now we can apply the induction procedure mentioned in the proof of Proposition 3.3, together with Kollár's bound (3.1) (over the residue field) to get the desired result. ■

The situation that is sketched above may seem particular, but it occurs a lot in practice. Often, the polynomials $f_1, \ldots, f_r$ are obtained from given reductions mod p by choosing arbitrary lifts. Then these lifts can always be chosen in $\mathcal{O}_K[\mathbb{N}^n]$.

A slight variant of this is the following. Let $\overline{f} \in \mathbb{F}_q[\mathbb{N}^n]$ define a nonsingular affine hypersurface in $\mathbb{A}^2_{\mathbb{F}_q}$. A canonical way to lift $\overline{f}$ to a polynomial $f \in \mathbb{Z}_q[\mathbb{N}^n]$ is by taking each coefficient

$$\overline{a}_1[X]^{\rho-1} + \overline{a}_2[X]^{\rho-2} + \cdots + \overline{a}_\rho \quad \in \mathbb{F}_q$$

to

$$a_1\theta^{\rho-1} + a_2\theta^{\rho-2} + \cdots + a_\rho \quad \in \mathbb{Z}[\theta]$$

where the $a_i \in \{0, \ldots, p-1\}$. If f happens to define a nonsingular hypersurface (over $\mathbb{Q}_q$), then we can apply the above lemma with $r = n+1$, $B' = p$ (if $r(X)$ is chosen appropriately), $d = \deg f$ and $B = p \cdot d$ to find polynomials $\alpha, \beta_1, \ldots, \beta_n \in \mathbb{Z}_q[\mathbb{N}^n]$ with

$$\begin{aligned} \deg \alpha f, \max_i \deg \beta_i \tfrac{\partial f}{\partial x_i} \quad \leq \quad & 4nd^{n+1} + 4n(n+1)d^n\,[2\rho \cdot \log(\rho p d) \\ & + (n+8)\log(n+1)d] \cdot \max\{3, d\}^n \end{aligned}$$

such that $1 = \alpha f + \beta_1 \frac{\partial f}{\partial x_1} + \cdots + \beta_n \frac{\partial f}{\partial x_n}$.

3.3 A sparse effective Nullstellensatz

In this section, it is more natural to work with *Laurent* polynomials instead of ordinary polynomials. The effective Nullstellensatz problem then becomes: given $f_1, \ldots, f_r \in R[\mathbb{Z}^n]$ such that they have no common solution in $(\overline{\mathbb{K}} \setminus \{0\})^n$ and such that their reductions $\overline{f}_1, \ldots, \overline{f}_r \in k[\mathbb{Z}^n]$ have no common solution in $(\overline{k} \setminus \{0\})^n$, find the 'smallest' possible $g_1, \ldots, g_r \in R[\mathbb{Z}^n]$ such that $1 = g_1 f_1 + \cdots + g_r f_r$ (the existence of such an expansion is an easy corollary to Proposition 3.3). It will turn out that an elegant answer exists whenever $r = n + 1$ and $\overline{f}_1, \ldots, \overline{f}_r$ satisfy some 'generic condition'.

3.8 Definition Let A be a ring and let $h \in A[\mathbb{Z}^n]$. The *support* of h is the subset of $\mathbb{Z}^n$ consisting of all $z = (z_1, \ldots, z_n)$ for which $x^z := x_1^{z_1} \cdots x_n^{z_n}$ has a non-zero coefficient in h. For any subset $\sigma \subset \mathbb{R}^n$, we denote by h_σ the Laurent polynomial obtained from h by setting all terms corresponding to exponent vectors that lie outside of σ equal to zero.

3.9 Condition Let $\mathbb{F}$ be a field and let $\Gamma \subset \mathbb{R}^n$ be a convex polytope with vertices in $\mathbb{Z}^n$. We say that $f_1, \ldots, f_r \in \mathbb{F}[\mathbb{Z}^n]$ satisfy *Condition 3.9 with respect to* Γ if the supports of $f_1, \ldots, f_r$ are contained in Γ and if for all faces γ (including Γ itself) the system

$$f_{1\gamma} = f_{2\gamma} = \cdots = f_{r\gamma} = 0$$

has no solution in $(\mathbb{F} \setminus \{0\})^n$.

Calling this condition 'generic' is justified by the following proposition.

3.10 Proposition *Let Γ be a convex polytope in $\mathbb{R}^n$ with integer vertex coordinates and write $\mathcal{S} = \Gamma \cap \mathbb{Z}^n$. Let r be a natural number $\geq n + 1$. Then the set of points*

$$((f_{1,e})_{e \in \mathcal{S}}, \ldots, (f_{r,e})_{e \in \mathcal{S}}) \in \mathbb{A}_{\mathbb{F}}^{(\#\mathcal{S}) \times r}$$

for which $f_1 := \sum f_{1,e} x^e$, …, $f_r := \sum f_{r,e} x^e$ satisfy Condition 3.9 with respect to Γ is contained in an algebraic set of codimension ≥ 1. Moreover, this algebraic set is defined over the prime subfield of $\mathbb{F}$.

Proof. This can be proved using a technique similar to the one used for Proposition 2.3. ■

3.11 Theorem *Let Γ be a convex polytope in $\mathbb{R}^n$ with vertices in $\mathbb{Z}^n$ and suppose that $\dim \Gamma = n$. Let R be a DVR with maximal ideal $\mathfrak{m}$. Let $f_0, f_1, \ldots, f_n \in R[\mathbb{Z}^n]$ have supports inside Γ and suppose that their reductions mod $\mathfrak{m}$ satisfy Condition 3.9 with respect to Γ. Then for any $g \in R[\mathbb{Z}^n]$ with support inside $(n+1)\Gamma$, there exist $h_0, \ldots, h_n \in R[\mathbb{Z}^n]$ with support in $n\Gamma$ such that $g = h_0 f_0 + \cdots + h_n f_n$.*

PROOF. Write $k = R/\mathfrak{m}$. Let S_Γ^k be the graded ring consisting of all k-linear combinations of terms of the form

$$t^d x^e, \text{with } d \in \mathbb{N} \text{ and } e \in d\Gamma \cap \mathbb{Z}^2.$$

The degree of such a term is by definition equal to d. Similarly, let S_Γ^R consist of the R-linear combinations.

Let Δ be the cone in $\mathbb{R}^{n+1}$ generated by all vectors (d, e) with $d \in \mathbb{N}$ and $e \in d\Gamma$. Clearly $S_\Gamma^k = k[\Delta]$. Because the systems $\overline{f}_{0\gamma} = \cdots = \overline{f}_{n\gamma} = 0$ have no common solution in $\mathbb{T}_k^n$, the locus in $\operatorname{Spec}(S_\Gamma^k)$ of $(t\overline{f}_0, \ldots, t\overline{f}_n)$ consists of only one point. This is easily verified considering the restrictions of the locus of the $t\overline{f}_i$ to the tori that partition $\operatorname{Spec}(S_\Gamma^k)$. Hence

$$\frac{S_\Gamma^k}{(t\overline{f}_0, \ldots, t\overline{f}_n)}$$

has noetherian dimension zero. On the other hand S_Γ^k is a Cohen-Macaulay ring by a well-known result of Hochster (see e.g. [19, Theorem 3.4]) that states that $k[C]$ is Cohen-Macaulay for any cone C. So $t\overline{f}_0, \ldots, t\overline{f}_n$ is a regular sequence. This means that we have exact sequences

$$0 \to \left(\frac{S_\Gamma^k}{(t\overline{f}_0, \ldots, t\overline{f}_i)}\right)_{d-1} \to \left(\frac{S_\Gamma^k}{(t\overline{f}_0, \ldots, t\overline{f}_i)}\right)_d \to \left(\frac{S_\Gamma^k}{(t\overline{f}_0, \ldots, t\overline{f}_{i+1})}\right)_d \to 0,$$

where the second arrow is multiplication by $t\overline{f}_{i+1}$ and where $(\cdots)_d$ denotes the homogeneous part of degree d. Thus

$$\dim_k \left(\frac{S_\Gamma^k}{(t\overline{f}_0, \ldots, t\overline{f}_{i+1})}\right)_d = \dim_k \left(\frac{S_\Gamma^k}{(t\overline{f}_0, \ldots, t\overline{f}_i)}\right)_d - \dim_k \left(\frac{S_\Gamma^k}{(t\overline{f}_0, \ldots, t\overline{f}_i)}\right)_{d-1}.$$

By a result of Ehrhart [32] $\dim_k(S_\Gamma^k)_d$, the number of lattice points in $d\Gamma$, is a polynomial function in d for all $d \geq 0$. We obtain that

$$\dim_k \left(\frac{S_\Gamma^k}{(t\overline{f}_0, \ldots, t\overline{f}_n)}\right)_d$$

is a polynomial function in d for all $d \geq n+1$. Since the noetherian dimension is zero, this polynomial must be zero as well.

In particular, we have that the k-linear map

$$W_k : \bigoplus_{i=0}^n (S_\Gamma^k)_n \to (S_\Gamma^k)_{n+1} : (t^n\overline{h}_0, \ldots, t^n\overline{h}_n) \mapsto t^{n+1}(\overline{h}_0\overline{f}_0 + \cdots + \overline{h}_n\overline{f}_n)$$

is surjective. But then necessarily the corresponding R-map

$$W_R : \bigoplus_{i=0}^n (S_\Gamma^R)_n \to (S_\Gamma^R)_{n+1} : (t^n h_0, \ldots, t^n h_n) \mapsto t^{n+1}(h_0 f_0 + \cdots + h_n f_n)$$

is surjective. Indeed, let M be the matrix of W_R. Then its reduction modulo $\mathfrak{m}$ is the matrix of W_k, so it has a minor of maximal dimension with non-zero determinant. But this means that M itself has a minor of maximal dimension whose determinant is a unit in R. ■

The above proof is inspired by an argument in [69], see also [6, Section 4]. Note that the same proof works if R is an arbitrary local ring.

3.4 The case of a polynomial and its derivatives

3.4.1 Negative results

Very often, the systems $f_1, \ldots, f_r$ to be considered consist of a polynomial f and its partial derivatives $\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}$. One could hope that this case allows better effective Nullstellensätze. The examples below show that this is unlikely. We include them in the thesis so that future researchers either avoid putting effort in the same matter or can contrast in some conclusions (we will do this ourselves in the next subsection, where we treat the special case of nondegenerate hypersurfaces).

3.12 Example Our first example deals with the field case and shows that Kollár's bound is still asymptotically sharp (for fixed n). Its definition is inspired by Example 3.2 given in the introduction of this chapter. Let $\mathbb{F}$ be a field of characteristic $p > 0$ and let $d \geq 2p$ be a multiple of p. Consider the degree $d+1$ polynomial

$$f = x_n^{d+1} + x_1^{d-p}x_2^p + 1 + \sum_{i=3}^{n} \left(x_i^p x_{i-1} x_1^{d-p} + x_{i-1}^{d-p+1} x_1^p \right).$$

It defines an irreducible, nonsingular hypersurface in $\mathbb{A}^n_{\mathbb{F}}$. Take polynomials $\alpha, \beta_1, \ldots, \beta_n \in \mathbb{F}[\mathbb{N}^n]$ such that

$$1 = \alpha f + \beta_1 \frac{\partial f}{\partial x_1} + \cdots + \beta_n \frac{\partial f}{\partial x_n}$$

and let $\lambda = \max_i\{\deg \alpha f, \deg \beta_i \frac{\partial f}{\partial x_i}\}$. Now proceed as in Example 3.2, using the map

$$\mathbb{F}[x_1, \ldots, x_n, z] \to \mathbb{F}[x_2, \ldots, x_n, z] : f(x_1, x_2, \ldots, x_n, z) \mapsto f(1, x_2, \ldots, x_n, z),$$

to conclude that

$$\lambda \geq \frac{d^2(d-p)^{n-2}}{p^{n-1}},$$

which is $O((\deg f)^n)$ for fixed n. We did not succeed in constructing a similar example in the characteristic 0 case. ■

3.13 Example This second example deals with the DVR case. It shows that a general bound in terms of $\deg f$, n and the valuations of the coefficients of f is still impossible. The argument is the same as the one used in Example 3.4. So let R be a DVR with local parameter t. Define $f = x_2 - tx_1x_2 + (t^m + t^2)x_1^2 - 1 \in R[x_1, x_2]$ for some big natural number m. Since

$$\left\{ \begin{array}{rcl} f & = & x_2 - tx_1x_2 + (t^m + t^2)x_1^2 - 1 \\ \frac{\partial f}{\partial x_1} & = & -tx_2 + 2(t^m + t^2)x_1 \\ \frac{\partial f}{\partial x_2} & = & 1 - tx_1 \end{array} \right. \begin{array}{rcl} = & 0 \\ = & 0 \\ = & 0 \end{array}$$

has no solutions, neither over the fraction field, nor over the residue field, there exist polynomials $\alpha, \beta_1, \beta_2 \in R[x_1, x_2]$ that satisfy

$$1 = \alpha f + \beta_1 \frac{\partial f}{\partial x_1} + \beta_2 \frac{\partial f}{\partial x_2}.$$

Putting $x_2 = 1 + tx_1$ and reducing modulo t^m gives the following identity in $R/(t^m)[x_1]$:

$$\begin{aligned} 1 &= \overline{\alpha}((1 - tx_1)(1 + tx_1) + t^2x_1^2 - 1) + \overline{\beta}_1(-t(1 + tx_1) + 2t^2x_1) + \overline{\gamma}(1 - tx_1) \\ &= -t\overline{\beta}(1 - tx_1) + \overline{\gamma}(1 - tx_1) = (\overline{\gamma} - t\overline{\beta})(1 - tx_1). \end{aligned}$$

Now proceed as in Example 3.4. ■

3.4.2 Nondegenerate hypersurfaces

When applied to a Laurent polynomial and its partial derivatives, Condition 3.9 gets a nice geometric meaning.

3.14 Definition Let $\mathbb{F}$ be a field. The *Newton polytope* $\Gamma(f)$ of a Laurent polynomial $f \in \mathbb{F}[\mathbb{Z}^n]$ is the convex hull of its support. We say that f is *nondegenerate with respect to its Newton polytope* if $f, x_1\frac{\partial f}{\partial x_1}, \ldots, x_n\frac{\partial f}{\partial x_n}$ satisfy Condition 3.9 with respect to $\Gamma(f)$.

One can show that such a nondegenerate polynomial f defines a hypersurface in $(\mathbb{F} \setminus \{0\})^n$ that has a natural complete, nonsingular model in $\mathbb{P}_{\Gamma(f)}$, the toric variety associated to $\Gamma(f)$. For $n = 2$, this is described in full detail in Chapter 2. For larger n, we refer to [18].

3.15 Remark Despite Proposition 3.10, the condition of being nondegenerate is no longer generic if $\mathbb{F}$ is of finite characteristic p. Indeed, let Γ be the polytope in $\mathbb{R}^3$ spanned by $(0,0,0)$, $(1,0,0)$, $(0,1,0)$ and $(1,1,p)$. A randomly chosen polynomial f with $\Gamma(f) = \Gamma$ will be of the form

$$f = a + bx_1 + cx_2 + dx_1x_2x_3^p.$$

If

$$(a,b,c,d) \in Z = \left\{(a,b,c,d) \in \mathbb{P}^3_{\mathbb{F}} \,|\, a,b,c,d \neq 0\right\}$$

then the hypersurface defined by f has a singular point $\left(-\frac{a}{b}, -\frac{a}{c}, \sqrt[p]{\frac{bc}{ad}}\right) \in (\overline{\mathbb{F}} \setminus \{0\})^3$.

However, the condition of being nondegenerate *is* generic if $n = 2$. This is precisely what Proposition 2.3 states. ■

The main results of this section are the following easy but interesting corollaries to Theorem 3.11. They were already stated in Section 2.5 in the case $n = 2$. The first result will be essential in devising a sharp bound for the rate of convergence of a lift of the Frobenius endomorphism in Chapter 4. The second one has at first sight nothing to do with Hilbert's Nullstellensatz: it states that an arbitrary Newton polytope preserving lift of a nondegenerate polynomial will again be nondegenerate.

3.16 Lemma *Let $\mathbb{F}$ be a field, let $f \in \mathbb{F}[\mathbb{Z}^n]$ and suppose that $\Gamma(f)$ is an n-dimensional polytope that contains the origin. If f is nondegenerate with respect to its Newton polytope, there exist $\alpha, \beta_1, \ldots, \beta_n \in \mathbb{F}[\mathbb{Z}^n]$ such that*

$$1 = \alpha f + \beta_1 x_1 \frac{\partial f}{\partial x_1} + \cdots + \beta_n x_n \frac{\partial f}{\partial x_n}$$

with $\Gamma(\alpha), \Gamma(\beta_1), \ldots, \Gamma(\beta_n) \subset n\Gamma(f)$.

Proof. Apply Theorem 3.11 to $f, x_1 \frac{\partial f}{\partial x_1}, \ldots, x_n \frac{\partial f}{\partial x_n}$. ■

3.17 Lemma *Let R be a DVR with residue field k and let $f \in R[\mathbb{Z}^n]$. Suppose f and its reduction $\overline{f}$ have the same Newton polytope. If $\overline{f}$ is nondegenerate with respect to its Newton polytope, then so is f (when considered over the fraction field $\mathbb{K}$ of R).*

Proof. Write $\Gamma = \Gamma(f) = \Gamma(\overline{f})$. Let γ be any face of Γ. If $\overline{f}$ is nondegenerate with respect to Γ, then so is $\overline{f}_\gamma$ with respect to γ. Using an appropriate change of variables[3] (mapping γ in some $\mathbb{R}^{\dim \gamma} \times \{0\} \times \cdots \times \{0\} \subset \mathbb{R}^n$), we can apply Theorem 3.11 to find a Laurent monomial $x_1^{r_1} \ldots x_n^{r_n}$ and Laurent polynomials $g_0, g_1, \ldots, g_n \in R[\mathbb{Z}^n]$ such that

$$x_1^{r_1} \ldots x_n^{r_n} = g_0 f_\gamma + g_1 \frac{\partial f_\gamma}{\partial x_1} + \cdots + g_n \frac{\partial f_\gamma}{\partial x_n}.$$

In particular, the system $f_\gamma = x_1 \frac{\partial f_\gamma}{\partial x_1} = \cdots = x_n \frac{\partial f_\gamma}{\partial x_n} = 0$ can have no solutions in $(\overline{\mathbb{K}} \setminus \{0\})^n$. ■

[3]This should be of the type $x^v \mapsto x^{Av+b}$ where $b \in \mathbb{Z}^n$ and $A \in \mathbb{Z}^{n \times n}$ has determinant ± 1. It is easily seen that such a change of variables preserves nondegeneracy.

Chapter 4

Monsky-Washnitzer cohomology of nondegenerate curves

In this chapter we will develop the cohomology theory of Monsky and Washnitzer [84, 85, 86] for nondegenerate curves. Our main reference is the survey by van der Put [105], but some of the new proofs given below are more constructive and focus on the computational aspect. This holds in particular for the finiteness of the first cohomology space (Section 4.2) and the existence of a lift of the Frobenius endomorphism (which is dealt with in Section 4.3). We conclude this chapter with a new sparse description of the first Monsky-Washnitzer cohomology space of a nondegenerate curve.

As in the foregoing chapters, we fix some notation. Throughout, x and y are fixed formal variables. For any integral domain R and any subset $\mathcal{S} \subset \mathbb{R}^2$, we denote by $R[\mathcal{S}]$ the ring $R[x^i y^j \mid (i,j) \in \mathcal{S} \cap \mathbb{Z}^2]$. If R is a complete DVR with local parameter t, and if $R[\mathcal{S}]$ is a finitely generated R-algebra, we denote its t-adic completion by $R\langle\mathcal{S}\rangle$ and its *weak completion* (this will be defined below) by $R\langle\mathcal{S}\rangle^\dagger$. Finally, if $\mathbb{F}$ is any field, $\overline{\mathbb{F}}$ denotes a fixed algebraic closure.

4.1 Definition

Let $\mathbb{F}_q$ be a finite field with $q = p^n$ elements (p prime). Let $\overline{f} \in \mathbb{F}_q[\mathbb{Z}^2]$ be a Laurent polynomial that is nondegenerate with respect to its Newton polytope Γ. Then, as explained in the introductory chapter, the aim is to find a so-called "Weil cohomology" for $\overline{C} = V(\overline{f}) \cap \mathbb{T}^2_{\overline{\mathbb{F}}_q}$, i.e. to associate to $\overline{C}$ certain cohomology spaces $H^i(\overline{C})$ and an induced action of Frobenius $\overline{\mathcal{F}}_q^*$ for which a Lefschetz fixed point formula

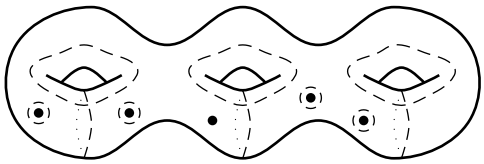
$$\#\overline{C}(\mathbb{F}_{q^k}) = \sum_i (-1)^i \operatorname{Trace}(\overline{\mathcal{F}}_q^{*k} \mid H^i(\overline{C})) \tag{4.1}$$

can be proven. The theory of Monsky and Washnitzer turns out to provide such a Weil cohomology. In this section, we describe these Monsky-Washnitzer cohomology spaces (in the case of nondegenerate curves), with emphasis on *why* they are defined the way they are.

During the search for a good definition of $H^i(\overline{C})$ ($i \in \mathbb{N}$), the following 'trivial' conditions should be kept in mind. First, it should be possible to define $\overline{\mathcal{F}}_q^*$, i.e. the spaces should be endowed with a natural action of Frobenius. Next, the spaces should be finite-dimensional, so that we can take traces[1]. In fact, in the spirit of the Weil conjecture (Theorem 1.8) we even want that

$$\dim H^0_{\text{Weil}}(\overline{C}) = 1, \quad \dim H^1_{\text{Weil}}(\overline{C}) = 2\text{Vol}(\Gamma)+1, \quad \dim H^i_{\text{Weil}}(\overline{C}) = 0 \text{ for } i \geq 2.$$

Indeed, $\overline{C}$ is obtained from $\widetilde{C} = V(\overline{f})$, which is a smooth complete genus $g = \#((\Gamma \setminus \partial\Gamma) \cap \mathbb{Z}^2)$ curve (Corollary 2.16), by taking away $R = \#(\partial\Gamma \cap \mathbb{Z}) \geq 1$ points (Corollary 2.12). Over $\mathbb{C}$ this would mean that $\overline{C}$ is a genus g Riemann surface missing R points,



genus 3 curve over $\mathbb{C}$ with 5 points missing
(together with $2 \cdot 3 + 5 - 1 = 10$ generators of $H_1^{\text{sing}}(\overline{C}, \mathbb{Z})$)

for which it is well-known that $\dim H^0_{\text{sing}}(\overline{C}, \mathbb{C}) = 1$, $\dim H^2_{\text{sing}}(\overline{C}, \mathbb{C}) = 0$ and

$$\dim H^1_{\text{sing}}(\overline{C}, \mathbb{C}) = 2g + R - 1$$

which is indeed $2\text{Vol}(\Gamma) + 1$ according to Pick's theorem. By analogy, these would then also be the Betti numbers to be expected for our $H^i(\overline{C})$. Another argument follows from the shape of the zeta function. As mentioned in the introductory chapter, Weil proved that the zeta function of the complete model equals

$$Z_{\widetilde{C}}(t) = \frac{P(t)}{(1-t)(1-qt)}$$

for a degree $2g$ polynomial $P(t)$. Now one has that

$$Z_{\widetilde{C}}(t) = Z_{\overline{C}}(t) \cdot \exp\left(\sum_{k=1}^{\infty} \frac{t^{\kappa_1 k}}{k} + \sum_{k=1}^{\infty} \frac{t^{\kappa_2 k}}{k} + \cdots + \sum_{k=1}^{\infty} \frac{t^{\kappa_t k}}{k}\right)$$

[1] Of course, being defined on a finite-dimensional vector space is *not* strictly necessary for an operator to have a trace. If the field of coefficients is complete, e.g. $\mathbb{R}$ or $\mathbb{Q}_p$, it suffices that the 'diagonal elements' add up to a convergent series. We refer to Remark 4.15 for more comments on this.

where the κ_i are the cardinalities of the $\mathrm{Gal}(\overline{\mathbb{F}}_q, \mathbb{F}_q)$-conjugate classes of points in $\widetilde{C} \setminus \mathbb{T}^2_{\mathbb{F}_q}$. Note that $\kappa_1 + \cdots + \kappa_t = R$. Hence

$$Z_{\overline{C}}(t) = \frac{P(t)(1 + t + \cdots + t^{\kappa_1 - 1})(1 - t^{\kappa_2}) \cdots (1 - t^{\kappa_t})}{(1 - qt)} \tag{4.2}$$

is a degree $2g + R - 1 = 2\mathrm{Vol}(\Gamma) + 1$ polynomial divided by a degree 1 polynomial. This matches exactly with what follows from (4.1) if the $H^i(\overline{C})$ have the dimensions given above.

4.1.1 The way towards the definition: trial and error

A first naive idea would be to work with the algebraic de Rham cohomology[2] of the coordinate ring

$$\overline{A} = \frac{\mathbb{F}_q[\mathbb{Z}^2]}{(\overline{f})},$$

but its finite characteristic causes the resulting spaces to be way too big. For example, if $\overline{f} = x$, one would like $H^1_{DR}(\overline{f}/\mathbb{F}_q)$ to be one-dimensional, but it turns out to be infinite-dimensional: all differentials of the form $y^{\lambda p - 1} dy$ ($\lambda \in \mathbb{Z}$) are non-exact. The reason is of course that expressions of the form

$$\frac{y^{\lambda p}}{\lambda p}$$

make no sense in characteristic p.

Now consider $\mathbb{Z}_q$, the valuation ring of $\mathbb{Q}_q$, a degree n unramified extension of the field of p-adic numbers $\mathbb{Q}_p$. Identify its residue field with $\mathbb{F}_q$. Then one could try to solve the above problem by taking an $f \in \mathbb{Z}_q[\mathbb{Z}^2]$ that reduces to $\overline{f}$ mod p, in order to work with the algebraic de Rham cohomology of

$$A = \frac{\mathbb{Z}_q[\mathbb{Z}^2]}{(f)}.$$

Then $H^1_{DR}(f/\mathbb{Q}_q)$ has better properties, but two new problems emerge.

1. The curve in $\mathbb{T}^2_{\mathbb{Q}_q}$ defined by f might have completely different geometric properties than the curve in $\mathbb{T}^2_{\mathbb{F}_q}$ defined by $\overline{f}$. In particular, $H^1_{DR}(f/\mathbb{Q}_q)$ might again have the 'wrong' dimension.

2. It is not clear how the Frobenius endomorphism $\overline{\mathcal{F}}_q : \overline{A} \to \overline{A} : \overline{a} \mapsto \overline{a}^q$ induces an action on $H^1_{DR}(f/\mathbb{Q}_q)$. To this end, it would be natural to have a $\mathbb{Z}_q$-algebra endomorphism $\mathcal{F}_q$ that *lifts* $\overline{\mathcal{F}}_q$, in the sense that it makes the following diagram commute:

$$\begin{array}{ccc} A & \xrightarrow{\mathcal{F}_q} & A \\ \downarrow & & \downarrow \\ \overline{A} & \xrightarrow{\overline{\mathcal{F}}_q} & \overline{A} \end{array}$$

[2]Recall its construction from Section 2.4, where it was introduced for fields of characteristic 0; the same construction works for the situations that are considered below.

(the vertical arrows are reduction mod p). But this seems to be impossible in general.

In the context of this thesis, the solution to the first problem is intuitively clear: simply force f to have the same Newton polytope as $\overline{f}$. Then from Lemma 2.23 it follows that f is nondegenerate as well. Since the geometry of nondegenerate curves is largely determined by the Newton polytope (cf. the material in Section 2.3), f and $\overline{f}$ share a lot of important properties. In particular, $H^1_{DR}(f/\mathbb{Q}_q)$ has the right dimension.

The second problem can be solved by moving on to the p-adic completion of A:

$$A^\infty = \frac{\mathbb{Z}_q\langle\mathbb{Z}^2\rangle}{(f)},$$

where

$$\mathbb{Z}_q\langle\mathbb{Z}^2\rangle = \left\{ \sum_{(i,j)\in\mathbb{Z}^2} a_{ij}x^iy^j \;\middle|\; |a_{ij}|_p \to 0 \text{ as } |i|+|j| \to \infty \right\}.$$

The elements of $\mathbb{Z}_q\langle\mathbb{Z}^2\rangle$ are called *strictly convergent power series* and $\mathbb{Z}_q\langle\mathbb{Z}^2\rangle$ is often referred to as the *Tate algebra*. Then a lift of Frobenius exists because of Hensel's lemma (this will be explained in detail in Section 4.3). But once again we run into trouble: since the integral of a strictly convergent power series is not necessarily strictly convergent itself, the dimension of the first cohomology space will again be too big. For instance, if $f = x$, all differentials of the form

$$\sum_{j\in\mathbb{N}} p^{\lambda j} y^{p^{\lambda j}-1} dy \quad (\lambda \in \mathbb{N}\setminus\{0\})$$

are non-exact.

Monsky and Washnitzer remedy this by working with a ring of power series that converge fast enough for their integrals to converge as well, but in which Hensel's lemma still holds:

$$A^\dagger = \frac{\mathbb{Z}_q\langle\mathbb{Z}^2\rangle^\dagger}{(f)},$$

where

$$\mathbb{Z}_q\langle\mathbb{Z}^2\rangle^\dagger = \left\{ \sum_{(i,j)\in\mathbb{Z}^2} a_{ij}x^iy^j \;\middle|\; \exists c \in]0,1[\text{ for which } \frac{|a_{ij}|_p}{c^{|i|+|j|}} \to 0 \text{ as } |i|+|j| \to \infty \right\}.$$

$\mathbb{Z}_q\langle\mathbb{Z}^2\rangle^\dagger$ is called the *weak p-adic completion* of $\mathbb{Z}_q[\mathbb{Z}^2]$ and its elements are called *overconvergent power series*. Then at last, the algebraic de Rham cohomology of $A^\dagger$ turns out to have the right properties. It has the expected Betti numbers and there is a natural action of Frobenius: this will be explained below.

4.1.2 Definition

The *Monsky-Washnitzer cohomology* of a nondegenerate Laurent polynomial $\overline{f} \in \mathbb{F}_q[\mathbb{Z}^2]$ is defined as follows. From now on, we will **always assume that** $\Gamma = \Gamma(f)$ **contains** $(0,0)$. This does not affect the generality of the main results below, since we can always shift Γ by multiplying $\overline{f}$ with a suitable Laurent monomial. Note that we made the same assumption in Section 2.4.

Let $f \in \mathbb{Z}_q[\mathbb{Z}^2]$ be a Newton polytope preserving lift of $\overline{f}$ and consider $A^\dagger = \frac{\mathbb{Z}_q\langle\mathbb{Z}^2\rangle^\dagger}{(f)}$ as above. Let $D^1(A^\dagger)$ be the universal $\mathbb{Z}_q$-module of differentials of $A^\dagger$ and let $D^2(A^\dagger)$ be its exterior square. More concretely,

$$D^1(A^\dagger) = \frac{A^\dagger dx + A^\dagger dy}{\left(\frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy\right)A^\dagger}, \qquad D^2(A^\dagger) = D^1(A^\dagger) \bigwedge_{A^\dagger} D^1(A^\dagger)$$

(where dx and dy may be looked at as formal symbols). Consider the $\mathbb{Z}_q$-morphisms[3]

$$d_0 : A^\dagger \to D^1(A^\dagger) : h \mapsto \frac{\partial h}{\partial x}dx + \frac{\partial h}{\partial y}dy,$$

$$d_1 : D^1(A^\dagger) \to D^2(A^\dagger) : h_1 dx + h_2 dy \mapsto \left(\frac{\partial h_1}{\partial y} - \frac{\partial h_2}{\partial x}\right) dx \wedge dy,$$

both of which we will often simply denote by d. Then $H^0_{MW}(\overline{f}/\mathbb{Q}_q) := \ker d_0 \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$,

$$H^1_{MW}(\overline{f}/\mathbb{Q}_q) := \frac{\ker d_1}{d_0(A^\dagger)} \otimes_{\mathbb{Z}_q} \mathbb{Q}_q \quad \text{and} \quad H^2_{MW}(\overline{f}/\mathbb{Q}_q) := \frac{D^2(A^\dagger)}{d_1(D^1(A^\dagger))} \otimes_{\mathbb{Z}_q} \mathbb{Q}_q.$$

We note that these spaces are well-defined, i.e. they do not depend on the choice of f. We refer the interested reader to [105, Theorem 2.4.4(i)] for a proof.

4.1.3 First properties

We can immediately verify that the spaces $H^0_{MW}(\overline{f}/\mathbb{Q}_q)$ and $H^2_{MW}(\overline{f}/\mathbb{Q}_q)$ have the expected dimensions. Indeed, using the continuity of d, it is easily seen that $H^0_{MW}(\overline{f}/\mathbb{Q}_q) = \mathbb{Q}_q$. As for $H^2_{MW}(\overline{f}/\mathbb{Q}_q)$, we will show that $D^2(A^\dagger) := \bigwedge^2_{A^\dagger} D^1(A^\dagger) = 0$. Indeed, from the Nullstellensatz for DVR's (Proposition 3.3) it follows that there exist $\beta, \gamma \in \mathbb{Z}_q[\mathbb{Z}^2]$ such that

$$1 = \beta\frac{\partial f}{\partial x} + \gamma\frac{\partial f}{\partial y} \quad (\text{in } A^\dagger).$$

In particular

$$h dx \wedge dy = h\left(\beta\frac{\partial f}{\partial x} + \gamma\frac{\partial f}{\partial y}\right) dx \wedge dy = -h\left(\beta\frac{\partial f}{\partial y}dy \wedge dy - \gamma\frac{\partial f}{\partial x}dx \wedge dx\right) = 0$$

[3] As one could guess, $\frac{\partial}{\partial x}$ and $\frac{\partial}{\partial y}$ are the unique linear and continuous operators that extend the formal partial derivations on $A = \frac{\mathbb{Z}_q[\mathbb{Z}^2]}{(f)}$.

for any $h \in A^\dagger$. Note that this allows us to rewrite the definition of $H^1_{MW}(\overline{f}/\mathbb{Q}_q)$ as follows:

$$H^1_{MW}(\overline{f}/\mathbb{Q}_q) = \frac{D^1(A^\dagger)}{d(A^\dagger)} \otimes_{\mathbb{Z}_q} \mathbb{Q}_q. \tag{4.3}$$

Verifying that $\dim H^1(\overline{f}/\mathbb{Q}_q) = 2\mathrm{Vol}(\Gamma) + 1$ is more difficult. A proof using the Lefschetz fixed point formula will be given in Section 4.4. We first show that $\dim H^1_{MW}(\overline{f}/\mathbb{Q}_q) < \infty$.

4.2 Finiteness of $\dim H^1_{MW}(\overline{f}/\mathbb{Q}_q)$

In full generality, the finiteness of Monsky-Washnitzer cohomology was proven in 1997 by Berthelot [9], see also Remark 4.15. For curves however, this was known from the beginning: it follows for instance from the Monsky-Washnitzer trace formula [105, Formula (1.2)], when combined with the Weil conjecture. Below, we give a new proof of this fact, for the case of nondegenerate curves. The big advantage of our proof is that it entails explicit bounds on the p-adic denominators that are introduced during differential reduction (see Subsection 6.1.5), which is a crucial part of our point counting algorithm.

Consider $C = V(f) \subset \mathbb{P}_{\mathbb{Q}_q,\Gamma}$ and $\widetilde{C} = V(\overline{f}) \subset \mathbb{P}_{\mathbb{F}_q,\Gamma}$ (we preserve the notation $\overline{C}$ for $V(\overline{f}) \cap \mathbb{T}^2_{\mathbb{F}_q}$). The main consequence of our assumption that Γ contains $(0,0)$ **is that $D_C = D_{C,\Gamma}$ is an effective divisor**. Let $\{t_1, \ldots, t_r\}$ be the edges of Γ, let $\{T_1, \ldots, T_r\} \subset \mathbb{P}_{\mathbb{Q}_q,\Gamma}$ be the corresponding tori over $\mathbb{Q}_q$ and let $\{\overline{T}_1, \ldots, \overline{T}_r\} \subset \mathbb{P}_{\mathbb{F}_q,\Gamma}$ be the corresponding tori over $\mathbb{F}_q$. The reductions mod p of the points in $T_k \cap C$ are precisely the points of $\overline{T}_k \cap \widetilde{C}$. For every $k = 1, \ldots, r$, we can find $c_k, b_k \in \mathbb{Z}$ such that

$$x^{c_k} y^{b_k} \tag{4.4}$$

defines a local parameter, at both P and $\overline{P}$, for each $P \in T_k \cap C$. Indeed, these assertions follow from the proof of Lemma 2.11: c_k, b_k depend only on the geometry of Γ. If in what follows we say '*local parameter over* $\mathbb{Z}_q$', actually any $t \in \mathbb{Z}_q[\mathbb{Z}^2]/(f)$ for which both t and its reduction mod p are local parameters at P resp. $\overline{P}$ will work.

Let $\mathbb{Q}_q^{\mathrm{ur}}$ be the maximal unramified extension of $\mathbb{Q}_q$. Denote by $\mathbb{Z}_q^{\mathrm{ur}}$ the valuation ring of $\mathbb{Q}_q^{\mathrm{ur}}$. Note that all places $P \in C \setminus \mathbb{T}^2_{\mathbb{Q}_q}$ are defined over $\mathbb{Q}_q^{\mathrm{ur}}$.

4.1 Definition

1. Let $L^{(0)} = \mathbb{Z}_q^{\mathrm{ur}}[\mathbb{Z}^2]$, then for any set S of Laurent polynomials, define $S^{(0)} = S \cap L^{(0)}$.

2. Let $L^{(1)}$ be the subset of $L^{(0)}$ consisting of those h for which the following holds. For every $P \in C \setminus \mathbb{T}^2_{\mathbb{Q}_q}$, take a local parameter t over $\mathbb{Z}_q$. Then the condition is that
$$\frac{t}{dt}\Lambda(h) = \sum_{i=v}^{\infty} a_i t^i \quad (a_i \in \mathbb{Z}_q^{\mathrm{ur}})$$

satisfies $\mathrm{ord}_p a_i \geq \mathrm{ord}_p i$ (alternative notation: $i|a_i$) for all $i < 0$. For any set S of Laurent polynomials let $S^{(1)} = S \cap L^{(1)}$.

We remark that the above definitions are vulnerable to notational abuses. For instance, if S consists of *cosets* of Laurent polynomials, then $S^{(0)}$ consists of those Laurent polynomials having a representative in $L^{(0)}$, and so on.

The set $L^{(1)}$ appears naturally[4] when we apply the operator

$$D = xy\left(\frac{\partial f}{\partial y}\frac{\partial}{\partial x} - \frac{\partial f}{\partial x}\frac{\partial}{\partial y}\right)$$

(that was introduced in the proof of Theorem 2.20), to an element in $L^{(0)}$. Indeed, let $P \in C \setminus \mathbb{T}^2_{\mathbb{Q}_q}$ and let t be a local parameter over $\mathbb{Z}_q$ at P. By the definition of D, we have

$$\frac{t}{dt}\Lambda(Dh) = \frac{t}{dt}dh,$$

where $\Lambda : A \to D^1(A) : h \mapsto h\frac{dx}{xyf_y}$. Write $h = \sum_{i=v}^{\infty} b_i t^i$, then $\frac{t}{dt}dh = \sum_{i=v}^{\infty} i b_i t^i \in L^{(1)}$. The main result of this section is the following (we refer the reader to Section 2.3 for notational conventions on Riemann-Roch spaces).

4.2 Theorem *Let E be an effective divisor which is defined over $\mathbb{Q}_q$ and whose support is contained in $C \setminus \mathbb{T}^2_{\mathbb{Q}_q}$, then the map*

$$\mathcal{L}^{(0)}(D_C + E) \xrightarrow{D} \frac{\mathcal{L}^{(1)}(2D_C + E)}{\mathcal{L}^{(1)}(2D_C)}$$

is surjective.

Note that the surjectivity of (2.8) (over $\mathbb{Q}_q$) is a corollary to the above. In fact, the above theorem will lead to an 'integral' version of Theorem 2.20. This will not only allow us to prove the finiteness of $H^1_{MW}(\overline{f}/\mathbb{Q}_q)$, it will also be the key ingredient of our differential reduction algorithm in Chapter 6. We postpone the proof of Theorem 4.2 to the end of this section. First we need a few more technical results.

4.3 Lemma *Let D be a divisor on C which is defined over $\mathbb{Q}_q$ and which has support in $C \setminus \mathbb{T}^2_{\mathbb{Q}_q}$. Then the $\mathbb{Z}_q^{\mathrm{ur}}$-module $\mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(D)$ is generated by the elements of $\mathcal{L}^{(0)}(D)$.*

Proof. We make use of a classical argument in perfect field theory. Take $h \in \mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(D)$ and let $\mathbb{Q}_{q^s} \supset \mathbb{Q}_q$ be a finite unramified degree s extension of $\mathbb{Q}_q$ over which h is defined. Let $\{\alpha_1, \ldots, \alpha_s\}$ be a $\mathbb{Z}_q$-basis of $\mathbb{Z}_{q^s}$ (the valuation ring of $\mathbb{Q}_{q^s}$) and write $\mathrm{Gal}(\mathbb{Q}_{q^s}, \mathbb{Q}_q) = \{\sigma_1 = \mathrm{id}, \sigma_2, \ldots, \sigma_s\}$. Then $A = (\sigma_j(\alpha_i))_{i,j}$ is invertible over $\mathbb{Z}_{q^s}$ since its determinant is a p-adic unit[5]: its

[4] In [31, Proposition 5.3.1], Edixhoven uses a similar set.

[5] This is the main reason why we work over $\mathbb{Q}_q^{\mathrm{ur}}$ instead of $\overline{\mathbb{Q}}_q$.

reduction modulo p is the discriminant of the basis $\{\overline{\alpha}_1, \ldots, \overline{\alpha}_s\}$ of $\mathbb{F}_{q^s}$ over $\mathbb{F}_q$. Define $(g_1, \ldots, g_s) := A(\sigma_1(f), \ldots, \sigma_s(f))$. Then the g_i are in $\mathcal{L}^{(0)}(D)$ because they are the traces of the $\alpha_i f$ and because D is defined over $\mathbb{Q}_q$. Since A is invertible, we can write $f = \sigma_1(f)$ as a $\mathbb{Z}_{q^s}$-linear combination of elements in $\mathcal{L}^{(0)}(D)$. ■

4.4 Theorem (Integral version of Theorem 2.18) *For every $m \in \mathbb{N}_0$, the module $\mathcal{L}^{(0)}(mD_C)$ is precisely given by $L^{(0)}_{m\Gamma}$.*

Proof. Take an element of $\mathcal{L}^{(0)}_{m\Gamma}$, represented by some $h \in \mathbb{Z}_q[\mathbb{Z}^2]$. By Theorem 2.18, there is an $\alpha \in \mathbb{Q}_q[\mathbb{Z}^2]$ such that $h + \alpha f$ has support in $m\Gamma$. Write $\alpha = \alpha_1 + \alpha_2$, where all coefficients of α_1 are integral and all coefficients of α_2 are non-integral. We claim that $h + \alpha_1 f$ has support in $m\Gamma$. Indeed, suppose this were not true, then $\alpha_2 f$ has a non-zero term with support outside $m\Gamma$. This implies that α_2 has a non-zero term with support outside $(m-1)\Gamma$. Let $a_{ij}x^i y^j$ be such a term. Then Γ has an edge spanning a line $dX + eY = c$ (with $\Gamma \subset \{(r,s) \mid dr + es \leq c\}$) such that $di + ej > (m-1)c$. Consider the following monomial order:

$$\begin{aligned} x^r y^s \prec x^k y^\ell \qquad & \text{if} \quad dr + es < dk + e\ell \\ & \text{or if} \quad dr + es = dk + e\ell \text{ and } r < k \\ & \text{or if} \quad dr + es = dk + e\ell,\ r = k \text{ and } s < \ell \end{aligned}$$

(where the last line is only of use if $e = 0$). We may suppose that $x^i y^j$ is maximal with respect to $\prec$. Take the term $b_{rs}x^r y^s$ of f that is maximal with respect to $\prec$ (in particular, $dr + es = c$). Then $a_{ij}b_{rs}x^{i+r}y^{j+s}$ is a term of $\alpha_2 f$ with support outside $m\Gamma$. Because $h + \alpha_1 f + \alpha_2 f$ has support in $m\Gamma$ and $h + \alpha_1 f \in \mathbb{Z}_q[\mathbb{Z}^2]$, this implies that $a_{ij}b_{rs}$ is integral. But this is impossible, since a_{ij} is non-integral and b_{rs} is a p-adic unit. ■

Note that the modules $\mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(mD_C)$ are then given by $L^{(0)}_{m\Gamma} \otimes_{\mathbb{Z}_q} \mathbb{Z}_q^{\mathrm{ur}}$. This follows from the above theorem together with Lemma 4.3 (although a straightforward proof works as well).

4.5 Lemma *Let D be an effective divisor on C with support contained in $C \setminus \mathbb{T}^2_{\mathbb{Q}_q}$ and let $P \in \mathrm{Supp}(D)$. Assume that $\deg D \geq 2g$, then there exists an $h \in \mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(D)$ such that:*

1. *h has a pole at P of multiplicity $\mathrm{ord}_P(D)$.*
2. *Let t be a local parameter over $\mathbb{Z}_q$ at P. Then h has an expansion $\sum_{i=v}^{\infty} a_i t^i$, with all $a_i \in \mathbb{Z}_q^{\mathrm{ur}}$ and a_v a unit in $\mathbb{Z}_q^{\mathrm{ur}}$.*

Proof. First suppose that Γ contains at least one interior lattice point. Then the modules $\mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(D)$ and $\mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(D-P)$ are free. This can be seen by using that it

are submodules of $\mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(mD_{C,\Gamma'})$ for some big enough m and a shift Γ' of Γ that contains the origin as an interior lattice point (indeed, in that case $D_{C,\Gamma'} > 0$, i.e. the coefficient at every place $P \in C \setminus \mathbb{T}^2_{\mathbb{Q}_q}$ is > 0). Above we proved that $\mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(mD_{C,\Gamma'})$ is finitely generated. Then every submodule is finitely generated as well: this follows from a well-known theorem on modules over noetherian rings. But it is also well-known that every finitely generated and torsion-free module over a principal ideal domain is free.

Now consider the following diagram where the vertical arrows are the natural reduction modulo p maps:

$$\begin{array}{ccc} \mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(D-P) & \xrightarrow{\subset} & \mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(D) \\ \downarrow\!\!\downarrow & & \downarrow\!\!\downarrow \\ \mathcal{L}_{\overline{\mathbb{F}}_q}(\overline{D}-\overline{P}) & \xrightarrow{\subsetneq} & \mathcal{L}_{\overline{\mathbb{F}}_q}(\overline{D}). \end{array}$$

The vertical maps are surjective, since after tensoring with $\mathbb{F}_q$ they become clearly injective and hence surjective because both have the same dimension by Riemann-Roch (here we used that $\mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(D)$ and $\mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(D-P)$ are free). Since $\deg D = \deg \overline{D} \geq 2g$, there is a function $\overline{h} \in \mathcal{L}_{\overline{\mathbb{F}}_q}(\overline{D}) \setminus \mathcal{L}_{\overline{\mathbb{F}}_q}(\overline{D}-\overline{P})$. Then any $h \in \mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(D)$ that reduces to $\overline{h}$ mod p will do the job.

If Γ does not have an interior lattice point, the above proof is not valid. But using that this implies that the genus of C (and of $\overline{C}$) is 0 (Corollary 2.16), one can get around this problem as follows. Let $T_k \subset \mathbb{P}_{\mathbb{Q}_q,\Gamma}$ be the one-dimensional torus containing P and let Γ' be a shift of Γ (that still contains the origin) such that the edge t'_k corresponding to T_k is not adjacent to the origin. Then $D_{C,\Gamma'}$ is an effective divisor with a non-zero coefficient at P. Therefore, $\mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}((\mathrm{ord}_P D-1)\cdot P)$ and $\mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}((\mathrm{ord}_P D)\cdot P)$ are contained in $\mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(mD_{C,\Gamma'})$ for some big enough $m \in \mathbb{N}_0$, so again we conclude that both modules are free. Since the genus is 0, we can find a function $\overline{h}$ in $\mathcal{L}_{\overline{\mathbb{F}}_q}((\mathrm{ord}_P D)\cdot\overline{P}) \setminus \mathcal{L}_{\overline{\mathbb{F}}_q}((\mathrm{ord}_P D-1)\cdot\overline{P})$ and by the same argument as above we can find a pre-image h that does the job. ■

4.6 Lemma *Suppose that all places $P \in C \setminus \mathbb{T}^2_{\mathbb{Q}_q}$ are defined over $\mathbb{Q}_q$. Let E be an effective divisor on C with support in $C \setminus \mathbb{T}^2_{\mathbb{Q}_q}$ and let $P \in C \setminus \mathbb{T}^2_{\mathbb{Q}_q}$. Suppose that $\deg E > 2g-2$, then the map*

$$\mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(E+P) \xrightarrow{D} \frac{\mathcal{L}^{(1)}_{\mathbb{Q}_q^{\mathrm{ur}}}(E+D_C+P)}{\mathcal{L}^{(1)}_{\mathbb{Q}_q^{\mathrm{ur}}}(E+D_C)}$$

is surjective.

PROOF. Note that this map is well-defined since E is effective. Let $h \in \mathcal{L}^{(1)}_{\mathbb{Q}_q^{\mathrm{ur}}}(E+D_C+P) \setminus \mathcal{L}^{(1)}_{\mathbb{Q}_q^{\mathrm{ur}}}(E+D_C)$, then by Corollary 2.14 we have $\mathrm{Div}\Lambda(h) = \mathrm{Div}h + D_C - W_C$. Let t be a local parameter over $\mathbb{Z}_q$ at P, then

$$\begin{aligned}\mathrm{ord}_P\left(\frac{t\Lambda(h)}{dt}\right) &= \mathrm{ord}_P(h) + \mathrm{ord}_P(D_C)\\ &= -\mathrm{ord}_P(E+D_C+P) + \mathrm{ord}_P(D_C)\\ &= -\mathrm{ord}_P(E) - 1 = -n\,,\end{aligned}$$

with $n = \mathrm{ord}_P(E+P)$. Therefore we can write

$$\frac{t\Lambda(h)}{dt} = b_0 t^{-n} + b_1 t^{-n+1} + \cdots\,,$$

with $n|b_0$. Using Lemma 4.5 with the divisor D replaced by $E+P$ gives an $h_0 \in \mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(E+P)$ with power series expansion at P:

$$h_0 = a_0 t^{-n} + a_1 t^{-n+1} + \cdots\,,$$

and with a_0 a p-adic unit. Define $h_1 = h + \frac{b_0}{na_0} D(h_0) \in \mathcal{L}^{(1)}_{\mathbb{Q}_q^{\mathrm{ur}}}(E+D_C+P)$, then we have

$$\frac{t\Lambda(h_1)}{dt} = \frac{t\Lambda(h)}{dt} + \frac{b_0}{na_0}\frac{tdh_0}{dt} = 0\cdot t^{-n} + \cdots\,,$$

and thus $\mathrm{ord}_P\left(\frac{t\Lambda(h_1)}{dt}\right) \geq -n+1$. Note that

$$\mathrm{ord}_P\left(\frac{t\Lambda(h_1)}{dt}\right) = \mathrm{ord}_P(h_1) + \mathrm{ord}_P(D_C)\,,$$

since $\mathrm{Div}\Lambda(h_1) = \mathrm{Div}h_1 + D_C - W_C$. Hence we see that $\mathrm{ord}_P(h_1) \geq -n+1-\mathrm{ord}_P(D_C) = 1-\mathrm{ord}_P(E+D_C+P)$, thus $h_1 \in \mathcal{L}^{(1)}_{\mathbb{Q}_q^{\mathrm{ur}}}(E+D_C)$ which finishes the proof. ■

We are now ready for the proof of the main theorem.

PROOF OF THEOREM 4.2. We can repeatedly apply Lemma 4.6 (using that $D_C \geq 0$) to obtain that the $\mathbb{Z}_q^{\mathrm{ur}}$-linear map

$$\mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(D_C+E) \xrightarrow{D} \frac{\mathcal{L}^{(1)}_{\mathbb{Q}_q^{\mathrm{ur}}}(2D_C+E)}{\mathcal{L}^{(1)}_{\mathbb{Q}_q^{\mathrm{ur}}}(2D_C)} \tag{4.5}$$

is surjective. If we tensor up with $\mathbb{Q}_q^{\mathrm{ur}}$ and then use a similar dimension argument as in the proof of Theorem 2.20 (or simply use the result presented there), we obtain that

$$\mathcal{L}(D_C+E) \xrightarrow{D} \frac{\mathcal{L}(2D_C+E)}{\mathcal{L}(2D_C)} \tag{4.6}$$

is surjective. Let $\{v_1,\dots,v_s\}$ be a $\mathbb{Z}_q$-basis of $\mathcal{L}^{(0)}(D_C)$ that extends to a $\mathbb{Z}_q$-basis $\{v_1,\dots,v_t\}$ of $\mathcal{L}^{(0)}(D_C+E)$. This is also a $\mathbb{Q}_q$-basis of $\mathcal{L}(D_C+E)$ and a $\mathbb{Z}_q^{\mathrm{ur}}$-basis of $\mathcal{L}^{(0)}_{\mathbb{Q}_q^{\mathrm{ur}}}(D_C+E)$ by Lemma 4.3.

Take $h\in\mathcal{L}^{(1)}(2D_C+E)$. By surjectivity of (4.5) and (4.6), we can find

$$\lambda_{s+1}v_{s+1}+\cdots+\lambda_t v_t \quad\text{and}\quad \mu_{s+1}v_{s+1}+\cdots+\mu_t v_t$$

in the inverse image, with $\lambda_i\in\mathbb{Z}_q^{\mathrm{ur}}$ and $\mu_i\in\mathbb{Q}_q$. Thus if we write $h_0=(\lambda_{s+1}-\mu_{s+1})v_{s+1}+\cdots+(\lambda_t-\mu_t)v_t$, then $\mathrm{Div}_C D(h_0)\geq -2D_C$, from which $\mathrm{Div}_C dh_0\geq -D_C-W_C$ and hence $\mathrm{Div}h_0\geq -D_C$ (recall that $D_C\geq 0$). By choice of $v_1,\dots,v_t$ this implies that $h_0=0$. Hence for $i=s+1,\dots,t$, the coefficients $\lambda_i=\mu_i$ are in $\mathbb{Z}_q^{\mathrm{ur}}\cap\mathbb{Q}_q=\mathbb{Z}_q$. So h has an inverse image in $\mathcal{L}^{(0)}(D_C+E)$. ■

4.7 Corollary *The canonical map*

$$H^1_{DR}(f/\mathbb{Q}_q)\to H^1_{MW}(\overline{f}/\mathbb{Q}_q) \tag{4.7}$$

is surjective. In particular $\dim H^1_{MW}(\overline{f}/\mathbb{Q}_q)\leq 2Vol(\Gamma)+1$.

Proof. One can extend the action of D and Λ by linearity and continuity to the entire dagger ring. It is then easy to see that it suffices to prove the surjectivity of the canonical map

$$\frac{A}{D(A)}\otimes_{\mathbb{Z}_q}\mathbb{Q}_q\to\frac{A^\dagger}{D(A^\dagger)}\otimes_{\mathbb{Z}_q}\mathbb{Q}_q.$$

For $m\in\mathbb{N}_0$, consider $\Diamond_m=\{(i,j)\in\mathbb{Z}^2\mid |i|+|j|\leq m\}$. Then $\Diamond_m=m\Diamond_1$. Take a $\mathbb{Q}_q$-rational divisor E with support in $C\setminus\mathbb{T}^2_{\mathbb{Q}_q}$ such that $L_{\Diamond_1}\subset\mathcal{L}(E)$. Then $L_{\Diamond_m}\subset\mathcal{L}(mE)$.

Take a $h=\sum_{(i,j)\in\mathbb{Z}^2}a_{ij}x^iy^j\in\mathbb{Z}_q\langle\mathbb{Z}^2\rangle^\dagger$ and let $\delta\in\mathbb{R}_0^+,\varepsilon\in\mathbb{R}^+$ be such that

$$\mathrm{ord}_p a_{ij}\geq\delta(|i|+|j|)-\varepsilon$$

for all $(i,j)\in\mathbb{Z}^2$. Write h as an infinite sum $h=\sum_{m\in\mathbb{N}_0}h_m$ of Laurent polynomials such that the support of h_m is contained in $\Diamond_m\setminus\Diamond_{m-1}$ (where $\Diamond_0$ is the empty set). Then the Gauss norm of h_m is at least $\delta m-\varepsilon$, so we can as well write

$$h=\sum_{m\in\mathbb{N}_0}p^{\mu(m)}h'_m$$

for polynomials $h'_m\in\mathbb{Z}_q[\mathbb{Z}^2]$. Here $\mu(m)=\min\{0,\lfloor m\delta-\varepsilon\rfloor\}$.

Write $E=\sum_{i=1}^r a_P P$ and let $a=\max_P a_P$. Since $h'_m\in\mathcal{L}^{(0)}(mE)$, we have that

$$p^{\lfloor\log_p(am)\rfloor}h'_m\in\mathcal{L}^{(1)}(mE)\subset\mathcal{L}^{(1)}(2D_C+mE).$$

Thus by Theorem 4.2 we can find a $g_m\in A$ such that $p^{\lfloor\log_p(am)\rfloor}h'_m=r_m+D(g_m)$ with $r_m\in\mathcal{L}^{(1)}(2D_C)\subset L_{2\Gamma}$. We conclude that

$$h=\sum_{m\in\mathbb{N}_0}p^{\mu(m)-\lfloor\log_p(am)\rfloor}r_m+D\left(\sum_{m\in\mathbb{N}_0}p^{\mu(m)-\lfloor\log_p(am)\rfloor}g_m\right)$$

(both sums converge). Since $\sum_{m\in\mathbb{N}_0} p^{\mu(m)-\lfloor \log_p(am)\rfloor} r_m \in A \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$, this concludes the proof. ■

4.3 Lifting Frobenius

Below, we show that there exists a $\mathbb{Z}_p$-algebra endomorphism $\mathcal{F}_p : A^\dagger \to A^\dagger$ that *lifts* $\overline{\mathcal{F}}_p : \overline{A} \to \overline{A} : \overline{a} \mapsto \overline{a}^p$ in the sense that $\overline{\mathcal{F}}_p \circ \pi = \pi \circ \mathcal{F}_p$, where π is reduction modulo p. Then $\mathcal{F}_q := \mathcal{F}_p \circ \cdots \circ \mathcal{F}_p$ is a $\mathbb{Z}_q$-algebra morphism that lifts $\overline{\mathcal{F}}_q$. Note that splitting up $\mathcal{F}_q$ into n copies of $\mathcal{F}_p$ (p small) dramatically improves the running time of the algorithms described in Chapter 6: this is the main reason why p-adic point counting algorithms are impractical for large values of p.

Again, the existence of such a lift was already known, but our new proof is constructive and gives explicit bounds on the rate of convergence, which is of great importance for the algorithm presented in Chapter 6: together with the material in the foregoing section, it allows us to bound the size of the objects we are dealing with and to determine the p-adic precision that is required during computation.

Moreover, the Newton polytope Γ turns out to play a very natural role in this convergence rate. This results in a 'sparse' description of the cohomology theory of Monsky and Washnitzer in Section 4.5.

4.3.1 A generalized Hensel's lemma

4.8 Theorem *Let Γ be a convex polygon in $\mathbb{R}^2$ with vertices in $\mathbb{Z}^2$. Take $a, b \in \mathbb{N}$ (not both zero) and let $H(Z) = \sum h_k Z^k \in \mathbb{Z}_q[\mathbb{Z}^2][Z]$ satisfy*

1. *$\Gamma(h_k) \subset (ak+b)\Gamma$ for all $k \in \mathbb{N}$;*
2. *$h_0 \equiv 0 \mod p$;*
3. *$h_1 \equiv 1 \mod p$.*

Then there exists a unique solution $Z_0 = \sum_{(i,j)\in\mathbb{Z}^2} a_{i,j} x^i y^j \in (p) \subset \mathbb{Z}_q\langle\mathbb{Z}^2\rangle$ to $H(Z) = 0$. Moreover, if $m \in \mathbb{N}$ and $(r,s) \in \mathbb{Z}^2$ are such that $(r,s) \notin m\Gamma$, then $\mathrm{ord}_p a_{r,s} \geq \frac{m}{2(a+b)}$.

4.9 Remark Note that we implicitly force Γ to contain the origin: this follows from conditions 1 and 3. If $\Gamma = \{(0,0)\}$, Theorem 4.8 is just Hensel's lemma over $\mathbb{Z}_q$. Finally, remark that if (r,s) is not contained in any multiple of Γ, the above lemma implies that $a_{r,s}$ equals 0. ■

Proof. The existence and uniqueness of Z_0 follow immediately from Hensel's lemma, applied over $\mathbb{Z}_q\langle\mathbb{Z}^2\rangle$. Therefore we only need to prove the convergence bound. Let $(r,s) \in \mathbb{Z}^2$ and $m \in \mathbb{N}$ be such that $(r,s) \notin m\Gamma$. Then there exists an edge spanning a line $eX + fY = c$ ($e, f, c \in \mathbb{Z}$), where $\Gamma \subset \{(i,j) \in$

$\mathbb{Z}^2 \mid ei + fj \le c\}$, such that $er + fs > mc$. Using a transformation of variables of the type used in Lemma 2.11, we may assume that $e = 0, f = 1$ and $c \ge 0$. Thus $s > mc$.

Now, replace in $H(Z)$ all occurrences of y^{-1} with a new variable t. We get

$$H_{\text{repl}}(Z) = \sum h_{k,\text{repl}}(x, y, t)Z^k \in \mathbb{Z}_q[x^{\pm 1}, y, t][Z]$$

with $\deg_y h_{k,\text{repl}} \le (ak + b)c$. Note that the conditions for Hensel's lemma are still satisfied. So there exists a unique

$$Z_{0,\text{repl}} = \sum_{(i,j,k)\in\mathbb{Z}\times\mathbb{N}^2} b_{i,j,k}x^i y^j t^k \quad \in (p) \subset \mathbb{Z}_q\langle x^{\pm 1}, y, t\rangle$$

satisfying $H_{\text{repl}}(Z_{0,\text{repl}}) = 0$. If we substitute y^{-1} for t, we get precisely Z_0, due to the uniqueness statement in Hensel's lemma. Henceforth

$$a_{r,s} = \sum_{j-k=s} b_{r,j,k}. \tag{4.8}$$

Let K be a suitably ramified extension of $\mathbb{Q}_q$ and denote by R its valuation ring. Consider

$$H'_{\text{repl}}(Z') = \sum p^{\mu_2(k-1)}h_k(x, p^{-\mu_1}y', t)Z'^k \quad \in K[x^{\pm 1}, y', t][Z']$$

obtained from $H_{\text{repl}}(Z)$ by substituting $y \leftarrow p^{-\mu_1}y'$, $Z \leftarrow p^{\mu_2}Z'$ and multiplying everything with $p^{-\mu_2}$. Here μ_1, μ_2 are positive rational numbers to be determined later. We know that if

$$\mu_2 + j\mu_1 < 1 \quad \forall j \le bc, \tag{4.9}$$

$$j\mu_1 < 1 \quad \forall j \le (a + b)c, \tag{4.10}$$

and

$$(k - 1)\mu_2 \ge j\mu_1 \quad \forall j \le (ak + b)c \tag{4.11}$$

for $k = 2, \ldots, \deg H$, then H'_{repl} has integral coefficients and $H'_{\text{repl}}(0) \equiv 0$ and $\frac{dH'_{\text{repl}}}{dZ'}(0) \equiv 1 \bmod P$. Here P is the maximal ideal of R. In that case, Hensel's lemma implies that there is a unique $Z'_{0,\text{repl}} \in P \cdot R\langle x^{\pm 1}, y', t\rangle$ such that $H'_{\text{repl}}(Z'_{0,\text{repl}}) = 0$. Write

$$Z'_{0,\text{repl}} = \sum_{(i,j,k)\in\mathbb{Z}\times\mathbb{N}^2} b'_{i,j,k}x^i y'^j t^k$$

and perform reverse substitution to obtain that

$$\sum_{(i,j,k)\in\mathbb{Z}\times\mathbb{N}^2} p^{\mu_2}p^{j\mu_1}b'_{i,j,k}x^i y^j t^k$$

is a solution to $H_{\mathrm{repl}}(Z) = 0$ in $P \cdot R\langle x^{\pm 1}, y, t\rangle$. Again using the uniqueness statement in Hensel's lemma we conclude that this is precisely $Z_{0,\mathrm{repl}}$. As a consequence

$$\mathrm{ord}_p b_{i,j,k} \geq j\mu_1 + \mu_2.$$

Using (4.8) we find that $\mathrm{ord}_p a_{r,s} \geq s\mu_1 + \mu_2 > mc\mu_1 + \mu_2$. This gives the desired result, since we can take $\mu_2 = \frac{2(a+b)c-bc}{2(a+b)c+\varepsilon}$ and $\mu_1 = \frac{1}{2(a+b)c+\varepsilon}$ for any $\varepsilon \in \mathbb{Q}_{>0}$. ■

4.10 Remark We note that in a subsequent paper [62] to ours [13], Kedlaya gives a related result, actually with a more elegant proof (instead of using Hensel's lemma in $\mathbb{Z}_q\langle\mathbb{Z}^2\rangle$ and then proving some rate of convergence, he immediately works in a complete subring consisting of Laurent series already satisfying the convergence rate; an example of such a ring is given in Remark 4.12 below). ■

4.3.2 The construction of $\mathcal{F}_p$

We are now ready to describe the construction of $\mathcal{F}_p$. In doing so, we will systematically make a notational distinction between power series g and the cosets $[g]$ (modulo f) they represent (something which is usually not done in order to simplify notation). We will use a technique that was first described in [25]. Suppose we can find a $Z_0 \in \mathbb{Z}_q\langle\mathbb{Z}^2\rangle^\dagger$ and polynomials $\delta_x, \delta_y \in \mathbb{Z}_q[\mathbb{Z}^2]$ such that

$$[f^\sigma(x^p(1+\delta_x Z_0), y^p(1+\delta_y Z_0))] = [0] \quad \text{in } A^\dagger,$$

where f^σ is obtained from f by applying Frobenius substitution[6] to the coefficients. Then

$$\mathcal{F}_p : A^\dagger \to A^\dagger : \begin{cases} [x] & \mapsto & [x^p(1+\delta_x Z_0)] \\ [y] & \mapsto & [y^p(1+\delta_y Z_0)] \end{cases}$$

(acting on $\mathbb{Z}_q$ by Frobenius substitution and extended by linearity and continuity) is a well-defined $\mathbb{Z}_p$-algebra morphism that lifts $\overline{\mathcal{F}}_p$.

Take $\overline{\beta}, \overline{\beta}_x, \overline{\beta}_y \in \mathbb{F}_q[\mathbb{Z}^2]$ with support in 2Γ for which

$$1 = \overline{\beta}\,\overline{f} + \overline{\beta}_x x\frac{\partial \overline{f}}{\partial x} + \overline{\beta}_y y\frac{\partial \overline{f}}{\partial y}$$

(this is possible due to Corollary 3.16). Let $\delta, \delta_x, \delta_y$ be arbitrary Newton polytope preserving lifts of $\overline{\beta}^p, \overline{\beta}_x^p$ resp. $\overline{\beta}_y^p$. Then clearly $\Gamma(\delta), \Gamma(\delta_x), \Gamma(\delta_y) \subset 2p\Gamma$.

Now let $x^a y^b$ be any monomial such that $g(x,y) = x^a y^b f(x,y)$ has support in $\mathbb{N}^2$ and define $G(Z) = x^{-pa}y^{-pb}g^\sigma(x^p(1+\delta_x Z), y^p(1+\delta_y Z)) \in \mathbb{Z}_q[\mathbb{Z}^2][Z]$, where g^σ is again obtained from g by applying Frobenius substitution to the coefficients. Since

$$G(0) \equiv f^p \quad \text{and} \quad \frac{dG}{dZ}(0) \equiv 1 + (a\delta_x + b\delta_y - \delta)f^p \mod p$$

[6] By Frobenius substitution we mean the map $\mathbb{Z}_q \to \mathbb{Z}_q : \sum_{i=0}^\infty \pi_i p^i \mapsto \sum_{i=0}^\infty \pi_i^p p^i$, where the π_i are Teichmüller representatives.

we see that $[G(Z)] = [0]$ has a unique solution $[Z_1]$ that is congruent to 0 mod p in the Henselian ring $A^\dagger$. However, Hensel's lemma does not provide any information on the convergence rate of Z_1 (or any other representative of $[Z_1]$). To solve this problem, define

$$H(Z) = G(Z) - (a\delta_x + b\delta_y - \delta)f^p Z - f^p.$$

Then clearly $[G(Z)] = [H(Z)]$, but now the conditions of Hensel's lemma are satisfied over the base ring, so that there exists a unique $Z_0 \in (p) \subset \mathbb{Z}_q\langle\mathbb{Z}^2\rangle$ for which $H(Z_0) = 0$. We have that $[Z_0] = [Z_1]$. Note that if we expand

$$H(Z) = \sum_{k=0}^{\deg H} h_k(x,y)Z^k,$$

one easily checks that

$$\Gamma(h_k) \subset (2k+1)p\Gamma. \tag{4.12}$$

Therefore, we can apply Lemma 4.8 and conclude that $Z_0 = \sum_{(i,j)\in\mathbb{Z}^2} a_{i,j}x^i y^j$ where the $a_{i,j}$ satisfy:

$$\forall i,j \in \mathbb{Z}, m \in \mathbb{N} : \ (i,j) \notin m\Gamma \ \Rightarrow \mathrm{ord}_p a_{i,j} \geq \frac{m}{6p}. \tag{4.13}$$

Our next step is to investigate what the convergence rate of Z_0 tells us about the convergence rate of $Z_x = 1 + \delta_x Z_0$ and $Z_y = 1 + \delta_y Z_0$. Write $Z_x = \sum_{(i,j)\in\mathbb{Z}^2} b_{i,j}x^i y^j$. We claim that

$$\forall i,j \in \mathbb{Z}, m \in \mathbb{N} : \ (i,j) \notin m\Gamma \ \Rightarrow \mathrm{ord}_p b_{i,j} \geq \frac{m}{8p}.$$

Indeed, since $Z_0 \equiv 0 \bmod p$, this statement is definitely true for $m < 8p$. If $m \geq 8p$, then $\frac{m-2p}{6p} \geq \frac{m}{8p}$. Now suppose $(i,j) \notin m\Gamma$. Write $\delta_x = \sum_{(i,j)\in 2p\Gamma} d_{i,j}x^i y^j$. We know that

$$b_{i,j} = \sum_{k+r=i,\, \ell+s=j} d_{k,\ell}a_{r,s}$$

and since $(k,\ell) \in 2p\Gamma$, we know that all (r,s) appearing in the above expansion are not contained in $(m-2p)\Gamma$. Therefore

$$\mathrm{ord}_p b_{i,j} \geq \frac{m-2p}{6p} \geq \frac{m}{8p}.$$

These observations allow us to state the main result of this section.

4.11 Theorem *There exist units $Z_x, Z_y \in \mathbb{Z}_q\langle\mathbb{Z}^2\rangle^\dagger$ such that*

$$\mathcal{F}_p : A^\dagger \to A^\dagger : \begin{cases} [x] & \mapsto & [x^p Z_x] \\ [y] & \mapsto & [y^p Z_y] \end{cases}$$

(extended by linearity and continuity and acting on $\mathbb{Z}_q$ by Frobenius substitution) is a well-defined $\mathbb{Z}_p$-algebra morphism that lifts $\overline{\mathcal{F}}_p$. Moreover Z_x, Z_y, Z_x^{-1}, Z_y^{-1} satisfy the following convergence criterion: if $(i,j) \in \mathbb{Z}^2, m \in \mathbb{N}$ are such that $(i,j) \notin m\Gamma$, then the coefficient of $x^i y^j$ has p-order $> \frac{m}{9p}$.

PROOF. It only remains to show that Z_x^{-1} and Z_y^{-1} satisfy the convergence criterion. By analogy, it suffices to show this for Z_x^{-1}. Write

$$Z_x = \sum_{(i,j)\in\mathbb{Z}^2} a_{ij}x^i y^j \quad \text{and} \quad Z_x^{-1} = \sum_{(i,j)\in\mathbb{Z}^2} d_{ij}x^i y^j.$$

Let $(r,s) \in \mathbb{Z}^2$ and $m \in \mathbb{N}$ be such that $(r,s) \notin m\Gamma$. As in the proof of Theorem 4.8, we may suppose that $Y = c$ (for some $c \geq 0$) is an edge of Γ and that $s > mc$ (and that $\Gamma \subset \{(x,y) \in \mathbb{R}^2 \,|\, y \leq c\}$). Now replace in Z_x all occurrences of y^{-1} with a new variable t to get

$$Z_{x,\mathrm{repl}} = \sum_{(i,j,k)\in\mathbb{Z}\times\mathbb{N}^2} b_{ijk}x^i y^j t^k.$$

Note that $Z_{x,\mathrm{repl}}$ is still invertible in $\mathbb{Z}_q\langle x^{\pm 1}, y, t\rangle$, i.e. it is of the form $1+p\cdot(\dots)$. Now replace y by $p^{-\mu}y'$ (for some $\mu \in \mathbb{Q}_{>0}$ that is to be determined later) to obtain

$$Z'_{x,\mathrm{repl}} = \sum_{(i,j,k)\in\mathbb{Z}\times\mathbb{N}^2} p^{-j\mu}b_{ijk}x^i y'^j t^k \quad \in R\langle x, y', t\rangle$$

for some suitably ramified p-adic ring R. Now suppose that

$$Z'_{x,\mathrm{repl}} = 1 + \pi\cdot(\dots) \tag{4.14}$$

for some π in P, the maximal ideal of R. Then it is invertible, hence we can write

$$Z'^{-1}_{x,\mathrm{repl}} = \sum_{(i,j,k)\in\mathbb{Z}\times\mathbb{N}^2} c_{ijk}x^i y'^j t^k \quad \in R\langle x, y', t\rangle.$$

Replacing y' by $p^{\mu}y$ yields

$$Z^{-1}_{x,\mathrm{repl}} = \sum_{(i,j,k)\in\mathbb{Z}\times\mathbb{N}^2} p^{j\mu}c_{ijk}x^i y^j t^k,$$

so that $\mathrm{ord}_p d_{r,s} \geq \mu s$, since

$$d_{r,s} = \sum_{j-k=s} p^{j\mu}c_{rjk}.$$

It remains to specify μ. We claim that we can take

$$\mu = \frac{1}{(8p+1)c + \varepsilon}$$

for some arbitrarily small $\varepsilon \in \mathbb{Q}_{>0}$. Since $Z_{x,\mathrm{repl}} = 1+p\cdot(\dots)$, $p^{-j\mu}b_{ijk}$ definitely has strictly positive valuation if $j \leq (8p+1)c$. If $j > (8p+1)c$, we can use the following argument. Since Z_x satisfies the convergence criterion mentioned in the construction of $\mathcal{F}_p$, for any $\lambda \in \mathbb{N}$ we have that

$$j \geq \lambda c \quad \Rightarrow \quad \mathrm{ord}_p b_{ijk} \geq \frac{\lambda}{8p}.$$

In particular, for any j one has $\mathrm{ord}_p b_{ijk} \geq \frac{\frac{j}{c}-1}{8p} = \frac{j-c}{8pc}$ if $c > 0$ (if $c = 0$ then $b_{ijk} = 0$ for all $j > 0$ so there is no problem). One can check that if $j > (8p+1)c$, the inequality

$$\frac{j-c}{8pc} > \frac{j}{(8p+1)c+\varepsilon}$$

holds.

In conclusion, $\mathrm{ord}_p d_{rs} > \frac{mc}{(8p+1)c+\varepsilon}$. Since this holds for any ε, the result follows. ■

4.12 Remark The bigger denominator ($9p$ instead of $8p$) is a small price we have to pay during inversion, but it also allows us to write down a strict inequality ($>$ instead of $\geq$). In this form, the convergence criterion is closed under multiplication, i.e.

$$\left\{ \sum_{(i,j)\in\mathbb{Z}^2} a_{i,j}x^i y^j \in \mathbb{Z}_q\langle\mathbb{Z}^2\rangle \;\middle|\; \forall m \in \mathbb{N}, (i,j) \in \mathbb{Z}^2 : (i,j) \notin m\Gamma \Rightarrow \mathrm{ord}_p a_{i,j} > \frac{m}{9p} \right\}$$

is a ring. We will use this in Section 4.5. ■

4.4 The Lefschetz fixed point formula

In this section, we will state the Lefschetz fixed point formula for nondegenerate curves. As before, our main reference is [105].

The $\mathbb{Z}_q$-algebra morphism $\mathcal{F}_q = \mathcal{F}_p \circ \cdots \circ \mathcal{F}_p : A^\dagger \to A^\dagger$ induces a $\mathbb{Z}_q$-module morphism $\mathcal{F}_q^* : D^1(A^\dagger) \to D^1(A^\dagger)$ through

$$g_1 dx + g_2 dy \mapsto \mathcal{F}_q(g_1) d\mathcal{F}_q(x) + \mathcal{F}_q(g_2) d\mathcal{F}_q(y).$$

Using that $\mathcal{F}_q$ is a morphism of *algebra's*, one can easily check that $\mathcal{F}_q^* \circ d = d \circ \mathcal{F}_q$. Therefore, $\mathcal{F}_q^*$ is well-defined on $H^1_{MW}(\overline{f}/\mathbb{Q}_q)$, on which it acts bijectively (see [105, (3.2)] for a proof).

4.13 Theorem *The number of solutions to $\overline{f} = 0$ in $\left(\mathbb{F}_{q^k} \setminus \{0\}\right)^2$ is given by*

$$q^k - \mathrm{Trace}\left(q^k \mathcal{F}_q^{*-k} \;\middle|\; H^1_{MW}(\overline{f}/\mathbb{Q}_q)\right). \tag{4.15}$$

Proof. See [105, (4.1)]. ■

4.14 Corollary $\dim H^1_{MW}(\overline{f}/\mathbb{Q}_q) = 2\mathrm{Vol}(\Gamma) + 1$.

Proof. From Newton's determinant formula (1.9), we conclude that the zeta function equals

$$Z_{\overline{f}}(t) = \frac{P(t)}{1-qt}$$

for some degree $\dim H^1_{MW}(\overline{f}/\mathbb{Q}_q)$ polynomial $P(t)$. On the other hand, from (4.2) we know that $\deg P(t) = 2\mathrm{Vol}(\Gamma) + 1$. ■

4.15 Remark In our proof, $\dim H^1_{MW}(\overline{f}/\mathbb{Q}_q) = 2\mathrm{Vol}(\Gamma) + 1$ is presented as an 'a posteriori' result: it follows from the Lefschetz fixed point formula, using the Weil conjecture. This is in contradiction with the philosophy behind Weil cohomology: one should deduce the Weil conjecture from properties of the cohomology spaces, and not the other way round.

But in fact, a straightforward analysis of $\dim H^1_{MW}(\overline{f}/\mathbb{Q}_q)$ is highly nontrivial. For higher dimensional varieties, the finiteness of the Monsky-Washnitzer cohomology spaces has even been an open problem for many years. The main obstacle was the absence of a Hironaka-like resolution of singularities theorem for varieties over fields of finite characteristic. It was only in 1997 that Berthelot [9] was able to prove that the Monsky-Washnitzer cohomology spaces of any affine variety are finite-dimensional, making use of de Jong's alteration theorem [20, Theorem 4.1].

Nevertheless, Monsky and Washnitzer were able to prove a Lefschetz fixed point theorem, without needing that the involved spaces are finite-dimensional. Indeed, they showed that the induced action of Frobenius is *nuclear*; this implies that the diagonal elements of any matrix add up to a convergent series, so that one still has a well-defined trace map. ■

An important corollary is the following.

4.16 Corollary *The canonical map $H^1_{DR}(f/\mathbb{Q}_q) \to H^1_{MW}(\overline{f}/\mathbb{Q}_q)$ is an isomorphism.*

Proof. Corollary 4.7 states that this map is surjective and by the above, the dimensions are equal. ■

4.5 Sparse description of $H^1_{MW}(\overline{f}/\mathbb{Q}_q)$

Let us return to the isomorphism

$$\frac{A^\dagger}{D(A^\dagger)} \xrightarrow{\Lambda} H^1_{MW}(\overline{f}/\mathbb{Q}_q)$$

that was used in Section 4.2 to prove the finiteness of the Betti numbers. As shown, $H^1_{MW}(\overline{f}/\mathbb{Q}_q)$ is endowed with a natural action of the Frobenius endomorphism, which we denoted by $\mathcal{F}_q^*$. This translates to an action on $A^\dagger/D(A^\dagger)$ via

$$\mathcal{G}_q : A^\dagger \to A^\dagger : h \mapsto \left(\Lambda^{-1} \circ \mathcal{F}_q^* \circ \Lambda\right)(h).$$

Now consider the following submodule of $A^\dagger$:

$$A^\dagger_\Gamma = \left\{ \sum_{(i,j)\in\mathbb{Z}^2} a_{ij}x^iy^j \in A^\dagger \,\middle|\, a_{ij} = 0 \text{ whenever } (i,j) \notin \bigcup_{m\in\mathbb{N}_0} m\Gamma \right\}.$$

Similarly, let $A_\Gamma = \mathbb{Z}_q\left[\bigcup_{m\in\mathbb{N}_0} m\Gamma\right]/(f) \subset A$. Then it is easily seen that D respects these submodules. The following lemma is the key observation of this section.

4.17 Lemma *$\mathcal{G}_q : A^\dagger \to A^\dagger$ respects the submodule $A^\dagger_\Gamma$*

Proof. According to Lemma 2.22, we can find $\alpha, \beta, \gamma \in \mathbb{Z}_q[\mathbb{Z}^2]$ that are supported in 2Γ such that $1 = \gamma f + \alpha x\frac{\partial f}{\partial x} + \beta y\frac{\partial f}{\partial y}$. Then one can check that $\mathcal{G}_q(h)$ is given by

$$\begin{aligned} & yf_y\left(\mathcal{F}_q(h)\mathcal{F}_q(\beta)\frac{xd\mathcal{F}_q(x)}{\mathcal{F}_q(x)dx} - \mathcal{F}_q(h)\mathcal{F}_q(\alpha)\frac{xd\mathcal{F}_q(y)}{\mathcal{F}_q(y)dx}\right) \\ - \quad & xf_x\left(\mathcal{F}_q(h)\mathcal{F}_q(\beta)\frac{yd\mathcal{F}_q(x)}{\mathcal{F}_q(x)dy} - \mathcal{F}_q(h)\mathcal{F}_q(\alpha)\frac{yd\mathcal{F}_q(y)}{\mathcal{F}_q(y)dy}\right), \end{aligned}$$

from which the lemma follows. ∎

4.18 Lemma *The canonical maps*

$$\frac{A_\Gamma}{D(A_\Gamma)} \to \frac{A}{D(A)} \quad \textit{and} \quad \frac{A^\dagger_\Gamma}{D(A^\dagger_\Gamma)} \to \frac{A^\dagger}{D(A^\dagger)}$$

are isomorphisms.

Proof. Consider the natural commutative diagram of maps

$$\begin{array}{ccc} A_\Gamma/D(A_\Gamma) & \longrightarrow & A/D(A) \\ \downarrow & & \downarrow \\ A^\dagger_\Gamma/D(A^\dagger_\Gamma) & \longrightarrow & A^\dagger/D(A^\dagger). \end{array}$$

In Corollary 4.7 we showed that $A/D(A) \otimes_{\mathbb{Z}_q} \mathbb{Q}_q \to A^\dagger/D(A^\dagger) \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$ is an isomorphism of $\mathbb{Q}_q$-vector spaces. In fact, working a bit more carefully, the same proof shows that $A/D(A) \to A^\dagger/D(A^\dagger)$ is an isomorphism of $\mathbb{Z}_q$-modules. Replacing $\Diamond_1$ by Γ and E by D_C, the same technique works to prove that $A_\Gamma/D(A_\Gamma) \to A^\dagger_\Gamma/D(A^\dagger_\Gamma)$ is an isomorphism. Therefore, it suffices to prove that

$$\frac{A_\Gamma}{D(A_\Gamma)} \to \frac{A}{D(A)}$$

is an isomorphism.

First, suppose that $h = D(g)$ for some $h \in A_\Gamma$ and $g \in A$. Then $\Lambda(h) = dg$, hence

$$\mathrm{Div}_C dg \geq -(m-1)D_C - W_C$$

for some $m \in \mathbb{N}_0$. Since $D_C \geq 0$, we conclude that $g \in \mathcal{L}^{(0)}((m-1)\Gamma)$, so $g \in A_\Gamma$ by Theorem 4.4. Thus the map is injective.

Next, take any $h \in A$ and let E be an effective divisor such that $\mathrm{Div}_C h \geq -E$. By Theorem 4.2 there is a $g \in \mathcal{L}^{(1)}(2D_C) \subset A_\Gamma$ such that $h - g \in D(A)$. Thus the map is surjective. ■

4.19 Corollary (Sparse Lefschetz fixed point theorem) *The number of solutions in* $\left(\mathbb{F}_{q^k} \setminus \{0\}\right)^2$ *to* $\overline{f} = 0$ *is given by*

$$q^k - \mathrm{Trace}\left(q^k \mathcal{G}_q^{-k} \mid A_\Gamma^\dagger / D(A_\Gamma^\dagger) \otimes_{\mathbb{Z}_q} \mathbb{Q}_q\right). \tag{4.16}$$

4.6 Summary and point counting strategy

Let us have a look at where we are now, from the point counting viewpoint. Using the above and the results in Sections 2.4, 4.2 and 4.4 we know that

	sparse		**functions**		**differentials**
algebraic	$\frac{A_\Gamma}{D(A_\Gamma)} \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$	$\longrightarrow$	$\frac{A}{D(A)} \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$	$\xrightarrow{\Lambda}$	$H^1_{DR}(f/\mathbb{Q}_q)$
	$\downarrow$		$\downarrow$		$\downarrow$
M-W	$\frac{A_\Gamma^\dagger}{D(A_\Gamma^\dagger)} \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$	$\longrightarrow$	$\frac{A^\dagger}{D(A^\dagger)} \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$	$\xrightarrow{\Lambda}$	$H^1_{MW}(\overline{f}/\mathbb{Q}_q)$

is a commutative diagram of isomorphisms. The top row is the algebraic de Rham side of the story: these are the spaces we actually want to compute in. The second row is the Monsky-Washnitzer side: this is where the Frobenius endomorphism lives. In course of this chapter, we replaced the classical cohomology description using *differential forms* (the right-most column) by a description using *functions* (the middle column), to eventually obtain a sparse description (the left-most column). The Frobenius endomorphism $\mathcal{F}_q^*$ on $H^1_{MW}(\overline{f}/\mathbb{Q}_q)$ translates to an endomorphism $\mathcal{G}_q$ acting on $A_\Gamma^\dagger / D(A_\Gamma^\dagger) \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$.

The main idea behind our point counting algorithm is then to use the sparse Lefschetz fixed point theorem (Corollary 4.19). More generally it suffices to compute the characteristic polynomial of $\mathcal{G}_q$ acting on $A_\Gamma^\dagger / D(A_\Gamma^\dagger) \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$, from which the zeta function and hence the number of points can be deduced (we refer to the introductory chapter for more details). Thanks to the Weil conjecture it suffices to do this modulo a certain finite p-adic precision.

More concretely, take a basis of $L_{2\Gamma}/D(L_\Gamma)$, which by the material in Section 2.4 is also a basis of $A_\Gamma/D(A_\Gamma) \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$ and hence of $A_\Gamma^\dagger / D(A_\Gamma^\dagger) \otimes_{\mathbb{Z}_q} \mathbb{Q}_q$. Compute the action of the $\mathbb{Q}_p$-morphism $\mathcal{G}_p = \Lambda^{-1} \circ \mathcal{F}_p^* \circ \Lambda$ on these basis elements modulo a well-chosen p-adic precision, and express the results again in

terms of the basis, as such one gets a matrix of $\mathcal{G}_p$. Deduce from this a matrix of $\mathcal{G}_q = \mathcal{G}_p \circ \cdots \circ \mathcal{G}_p$ and compute the characteristic polynomial. The concrete implementation details are given in Chapter 6.

Chapter 5

Linear algebra algorithms over p-adic rings

In this very short chapter we present methods to tackle two typical problems arising when doing linear algebra over $\mathbb{Z}_q = \mathbb{Z}_{p^n}$, for use in Chapter 6. We deal with system solving and characteristic polynomial computation.

As in the introductory chapter, all time and space estimates are made using the Landau symbol O and measure the bit complexity. The Soft-Oh $\widetilde{O}$ neglects factors that are logarithmic in the input size. We will use that operations in $\mathbb{Z}_q/(p^N)$ can be done in quasi-linear time (using e.g. the Schönhage-Strassen multiplication method [96]).

5.1 System solving

The problem of system solving is of course very classical and has been extensively studied before in much more general contexts. But the situation we are interested in is particular: the required p-adic precision grows with the dimensions of the matrices that are involved (see Chapter 6). Below, we present a new technique – basically Newton iteration, where in each step there is a loss of precision to be taken care of – that exploits this behavior. It often works better than the classical divide and conquer algorithms.

Let $r, s \in \mathbb{N}_0$ and consider a matrix $A \in \mathbb{Z}_q^{r \times s}$ and a vector $b \in \mathbb{Z}_q^r$. Let $N \in \mathbb{N}_0$ denote the required p-adic precision. The aim is to find an $\mathbf{x} \in \mathbb{Z}_q^s$ such that $A \cdot \mathbf{x} \equiv b \bmod p^N$. Note that this is slightly weaker than finding the reduction mod p^N of an $\mathbf{x} \in \mathbb{Z}_q^s$ such that $A \cdot \mathbf{x} = b$ (exact equality over $\mathbb{Z}_q$), but only slightly: from Lemma 5.1 below it follows that it suffices to increase the precision in order to solve this.

Using Gaussian elimination, where in each step the pivot is taken to have minimal p-adic valuation, one can find invertible matrices $N_1 \in \mathbb{Z}_q^{r \times r}, N_2 \in \mathbb{Z}_q^{s \times s}$ such that

$$N_1 \cdot A \cdot N_2$$

is a diagonal matrix whose diagonal elements are called the *invariant factors* of A. We then have the following lemma (the proof is immediate).

5.1 Lemma *Let $\theta \in \mathbb{N}$ be an upper bound for the p-adic valuations of the non-zero invariant factors of A and let $N \geq \theta$. Let $\mathbf{x}_0 \in \mathbb{Z}_q^s$ satisfy*

$$A \cdot \mathbf{x}_0 \equiv b \bmod p^N.$$

If there is an $\mathbf{x} \in \mathbb{Z}_q^s$ such that

$$A \cdot \mathbf{x} = b,$$

then $\mathbf{x}$ can be chosen to satisfy $\mathbf{x} \equiv \mathbf{x}_0 \bmod p^{N-\theta}$.

The method works as follows. First, precompute the invariant factors and the matrices N_1 and N_2 (and their inverses) modulo $p^{2\theta}$. In total, we need $\widetilde{O}(d^3 n\theta)$ time to do this, where $d = \max\{r, s\}$ is the dimension of A.

Now suppose we have an $\mathbf{x}_0$ such that $A \cdot \mathbf{x}_0 \equiv b \bmod p^N$ for some $N \geq \theta$. By Lemma 5.1, we can find an $\mathbf{x}$ of the form $\mathbf{x}_0 + \mathbf{t}p^{N-\theta}$ such that $A \cdot \mathbf{x} \equiv b \bmod p^{2N}$. To this end, we have to find a $\mathbf{t}$ such that

$$A \cdot \mathbf{t} \equiv \frac{b - A \cdot \mathbf{x}_0}{p^{N-\theta}} \bmod p^{N+\theta}.$$

Let $T(N)$ denote the time needed to solve a linear system (with fixed linear part A) up to precision N, assuming it has a p-adic solution. Then

$$T(2N) = T(N) + T(N + \theta) + \widetilde{O}(d^2 nN).$$

Here, the first term comes from the time needed to compute $\mathbf{x}_0$. The last term is dominated by the computation of $A \cdot \mathbf{x}_0$ modulo p^{2N}. The second term comes from the time needed to compute $\mathbf{t}$, given $(b - A \cdot \mathbf{x}_0)/p^{N-\theta} \bmod p^{2N}$. Similarly, $T(N + \theta) = T(N) + T(2\theta) + \widetilde{O}(d^2 nN)$. Using our precomputation and the fact that $\theta \leq N$, we have that $T(2\theta) = \widetilde{O}(d^2 nN)$. In conclusion,

$$T(2N) = 2T(N) + \widetilde{O}(d^2 nN).$$

It is obvious that this recurrence relation still holds if $N < \theta$ (again using our precomputation). From a well-known observation in complexity theory (see for instance [43, Lemma 8.2.]) we conclude that

$$T(N) = \widetilde{O}(d^2 nN).$$

Together with our precomputation this results in $\widetilde{O}(d^2 nN + d^3 n\theta)$ bit-operations.

Let us do a quick comparison with the classical divide and conquer algorithms, iterating on the dimension of the system instead of the p-adic precision. Suppose for ease of exposition that we have a square system of dimension d, all of whose invariant factors are p-adic units. Then, following for instance [57,

Theorem 5.1 and Section 6], we can solve this using $O(d^{2.70})$ ring operations, resulting in a time complexity of $\widetilde{O}(d^{2.70}N)$ (we fix n). The above method needs $\widetilde{O}(d^2N + d^3)$ bit-operations, so it is asymptotically better as soon as $N \sim d^{0.30}$. In most systems that we will be considering in Chapter 6, we will have $N \sim d^{0.67}$ (the precision will be $\sim g$, where g is the genus of the input curve, and the system dimensions will be $\sim g^{1.5}$).

5.2 Characteristic polynomial computation

In this section we review the classical algorithm based on reduction to the Hessenberg form. When applied over $\mathbb{Z}_q$ instead of a field, some caution is needed.

First, let R be any ring. A matrix $A \in R^{d\times d}$ is said to be in (upper) Hessenberg-form if it is of the following almost upper triangular shape

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & \dots & a_{1,d-1} & a_{1d} \\ a_{21} & a_{22} & a_{23} & a_{24} & \dots & a_{2,d-1} & a_{2d} \\ 0 & a_{32} & a_{33} & a_{34} & \dots & a_{3,d-1} & a_{3d} \\ 0 & 0 & a_{43} & a_{44} & \dots & a_{4,d-1} & a_{4d} \\ 0 & 0 & 0 & a_{54} & \dots & a_{5,d-1} & a_{5d} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & a_{d,d-1} & a_{dd} \end{bmatrix}$$

Let A_i $(i = 1, \dots, d)$ be the matrix obtained from A by deleting the $i-1$ first rows and columns (thus $A_1 = A$), and let $\chi_i(t) = \det(A_i - \mathbb{I}t) \in R[t]$ be its characteristic polynomial. Note that all A_i are again in Hessenberg form. One can then check the following recurrence relation:

$$\chi_i(t) = (a_{ii} - t)\chi_{i+1}(t) + \sum_{j=i+1}^{d} (-1)^{j-i} \left(\prod_{k=i}^{j-1} a_{k+1,k} \right) a_{i,j}\chi_{j+1}(t)$$

where $i = 1, \dots, d$ and $\chi_{d+1}(t) = 1$. This can be used to compute $\chi(t) = \chi_1(t)$ using $O(d^2)$ operations in $R[t]$. Using that the degrees of the polynomials that are involved are bounded by d, we find that the characteristic polynomial of a $d \times d$ Hessenberg matrix can be found using d^3 ring operations.

Now suppose we want to compute the characteristic polynomial $\chi(t)$ modulo some precision p^N, $N \in \mathbb{N}_0$, of *any* matrix $A \in \mathbb{Z}_q^{d\times d}$ (whose entries are also given modulo p^N). So in fact we work in the ring $R = \mathbb{Z}_q/(p^N)$. Below we show that there exists an R-invertible matrix $E \in R^{d\times d}$ such that $E \cdot A \cdot E^{-1}$ is in Hessenberg form. Note that A and $E \cdot A \cdot E^{-1}$ have the same characteristic polynomial.

This Hessenberg reduction is just Gaussian row elimination, but now the pivot is taken to be an entry on the lower diagonal. The idea is that performing the corresponding column operation does not affect the reduction process, exactly because we are working below the diagonal. Since R is not a field, the

pivot should be chosen carefully: it should be a non-zero entry having minimal p-adic valuation[1] among the lower diagonal elements of the column that is considered (if there is no non-zero entry, no reduction is necessary). All this is summarized in the following pseudo-code (ν_p is the valuation function).

INPUT: a matrix $A = (a_{ij})_{i,j}$ in $R^{d\times d}$
OUTPUT: an equivalent Hessenberg matrix

1. **for** $j = 1, \ldots, d-2$ **do** (running through the columns)
2. **if** $\{a_{j+1,j}, \ldots, a_{d,j}\} \neq \{0\}$ **then**
3. take k in $\{j+1, \ldots, d\}$ for which $\nu_p(a_{k,j})$ is minimal (selecting pivot)
4. (Row $j+1$) $\leftrightarrow$ (Row k) and (Col $j+1$) $\leftrightarrow$ (Col k)
5. **for** $\ell = j+2, \ldots, d$ **do** (running through the rows)
6. take $\alpha \in R$ such that $a_{\ell,j} = \alpha \cdot a_{j+1,j}$
7. (Row ℓ) $\leftarrow$ (Row ℓ) $- \alpha \cdot$ (Row $j+1$)
8. (Col $j+1$) $\leftarrow$ (Col $j+1$) $+ \alpha \cdot$ (Col ℓ)

The Hessenberg reduction needs $O(d^3)$ operations in R.

In conclusion, the total time needed to compute the characteristic polynomial takes $O(d^3)$ operations in $R = \mathbb{Z}_q/(p^N)$, resulting in a complexity estimate of $\widetilde{O}(d^3 nN)$. The space needed is $O(d^2 nN)$.

[1] I.e. the p-adic valuation of an element in the corresponding equivalence class in $\mathbb{Z}_q$.

Chapter 6

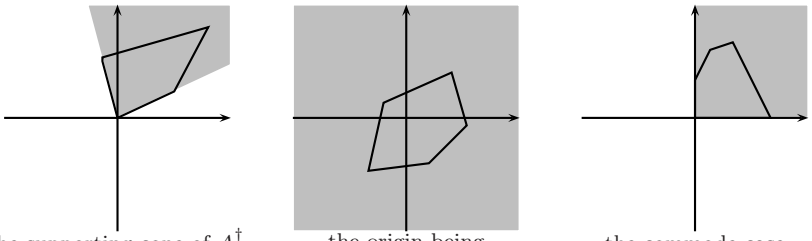
Point counting on nondegenerate curves

In this chapter we describe two algorithms that make use of the theory presented in Chapter 4 (and, in particular, of the ideas given at the end of Section 4.5). They compute the characteristic polynomial[1] $\chi(t) \in \mathbb{Z}[t]$ of Frobenius $\mathcal{F}_q^*$ acting on the first Monsky-Washnitzer cohomology space $H^1_{MW}(\overline{f}/\mathbb{Q}_q)$ of a given nondegenerate Laurent polynomial $\overline{f} \in \mathbb{F}_q[\mathbb{Z}^2]$. Combining Newton's determinant formula (1.9), the trace formula (Theorem 4.13) and the Weil conjecture (Theorem 1.8), one can check that the zeta function is then given by

$$Z_{V(\overline{f})\cap\mathbb{T}^2_{\mathbb{F}_q}}(t) = \frac{\frac{1}{q^{g+R-1}}\,\chi(qt)}{(1-qt)}$$

where R is the number of lattice points on the boundary of $\Gamma = \Gamma(\overline{f})$ (which equals the number of points on $V(\overline{f}) \setminus \mathbb{T}^2_{\mathbb{F}_q}$ because of Corollary 2.12).

Without loss of generality, we may assume that Γ contains the origin. Then the splitting up into different algorithms is based on the shape of $\bigcup_{m\in\mathbb{N}_0} m\Gamma$, the supporting cone of $A^\dagger_\Gamma$.



the supporting cone of $A^\dagger_\Gamma$ | the origin being an interior point | the commode case

[1] To avoid confusion: throughout this thesis, the characteristic polynomial of an operator M is given by $\det(M - \mathbb{I}t)$.

The first algorithm treats the case where the origin is an *interior* point of the Newton polytope Γ, i.e. the supporting cone of $A_\Gamma^\dagger$ is all of $\mathbb{R}^2$. Note that this is a very general case: whenever the genus of our nondegenerate curve is ≥ 1, we can shift its Newton polytope (by multiplying our equation with a suitable monomial) so that the origin becomes an interior point. Since point counting on rational curves is easy, this algorithm in fact treats nondegenerate curves in full generality.

The second algorithm will be discussed only briefly and deals with the commode case, this is when the supporting cone of $A_\Gamma^\dagger$ is $\mathbb{R}_+^2$. In this case, there are some simplifications. For technical reasons, we will impose the extra condition that $\overline{f}$ is monic in y, but nevertheless it is still the case that occurs most in practice.

Similar algorithms could be developed for the case where the supporting cone of $A_\Gamma^\dagger$ is the upper half plane, or the right half plane, and so on.

As before, all time and space estimates below measure the bitwise complexity. The Soft-Oh $\widetilde{O}$ neglects factors that are logarithmic in the input size. We will often implicitly use that field or ring operations can be done in quasi-linear time (using e.g. the Schönhage-Strassen multiplication method [96]).

6.1 The genus ≥ 1 case

6.1.1 Input and output analysis

We remark that a similar analysis has already been made in the introductory chapter (Subsection 1.2.1). Since the situation here is slightly different, it is done again. As input our algorithm expects an $\overline{f} \in \mathbb{F}_q[\mathbb{Z}^2]$ ($q = p^n$, p prime) that is nondegenerate with respect to its Newton polytope Γ, which we suppose to contain at least one interior lattice point. A good measure for the input size is

$$\text{number of monomials} \quad \times \quad \begin{array}{l}(\text{ space needed to represent coefficient} \\ \quad + \text{ space needed to represent exponent vector})\end{array}$$

which is $\sim \#(\Gamma \cap \mathbb{Z}^2) \cdot (\log q + \log \delta)$, where δ is the *degree* of $\overline{f}$, that is

$$\max\{|i| + |j| \,|\, (i,j) \in \Gamma\}.$$

Denote as before the genus of $C(\overline{f})$ (which equals $\#(\Gamma \setminus \partial\Gamma \cap \mathbb{Z}^2)$ by Corollary 2.16) with g. From a result by Scott [99], that states that $\#(\Gamma \cap \mathbb{Z}^2) \leq 3g+7$ whenever $g \geq 1$, it follows that $\#(\Gamma \cap \mathbb{Z}^2)$ is asymptotically equivalent with g.

As output our algorithm gives the characteristic polynomial of Frobenius acting on $H^1_{MW}(\overline{f}/\mathbb{Q}_q)$, which we denote with $\chi(t) \in \mathbb{Z}[t]$. A measure for its size follows easily from the Weil conjecture. Indeed, its degree equals $2\mathrm{Vol}(\Gamma) + 1$ and $2g$ of its roots have absolute value $q^{1/2}$. The other roots correspond to $\#(\partial\Gamma \cap \mathbb{Z}^2) - 1$ places lying on $V(f) \setminus \mathbb{T}^2_{\mathbb{F}_q}$ and have absolute value q. Now, since the i^{th} coefficient of $\chi(t)$ is the sum of $\binom{2\mathrm{Vol}(\Gamma)+1}{i}$ i-fold products of such roots,

we conclude that an upper bound for the absolute values of the coefficients is given by

$$\binom{2\mathrm{Vol}(\Gamma)+1}{\mathrm{Vol}(\Gamma)} q^{g+R-1} \leq 2^{2\mathrm{Vol}(\Gamma)+1} q^{g+R-1}.$$

Recall that R is the number of lattice points on the boundary of Γ, which is $\leq 2g+7$ by Scott's result. Therefore, the number of bits needed to represent $\chi(t)$ is

$$O\left((2\mathrm{Vol}(\Gamma)+1)\cdot \log(2^{2\mathrm{Vol}(\Gamma)+1} q^{g+R-1})\right) = O(ng^2)$$

for p fixed.

6.1.2 Remarks on curve representations and computing with polytopes

As mentioned above, we suppose that our input curve $\overline{f} \in \mathbb{F}_q[\mathbb{Z}^2]$ is given as an array of coefficients, together with the corresponding exponent vectors[2]. E.g. an elliptic curve $\overline{f} = y^2 - x^3 - \overline{a}x - \overline{b} = 0$ over $\mathbb{F}_q$ can be represented as

$$[\{1;(0,2)\},\{-1;(3,0)\},\{-\overline{a};(1,0)\},\{-\overline{b};(0,0)\}].$$

Below we will implicitly assume that we can select coefficients (that is

$$\texttt{input:}\ \overline{f} \text{ and } (i,j) \qquad \texttt{output:}\ \text{pointer to coefficient of } \overline{f} \text{ at } x^i y^j)$$

in $\widetilde{O}(1)$ time. This can be done if the array representing $\overline{f}$ is sorted with respect to the second components. So if necessary, one first needs to apply a sorting algorithm such as quicksort, taking $\widetilde{O}(g^2 \log \delta)$ time.

The Newton polytope of $\overline{f}$ can be easily computed using Graham's algorithm [16, Chapter 35] in $\widetilde{O}(g \log \delta)$ time. The output of this algorithm is the set of vertices

$$v_1 = (a_1, b_1), \ldots, v_r = (a_r, b_r)$$

of $\Gamma = \Gamma(\overline{f})$ (enumerated clockwise). One can then check that the number of points on the boundary is given by (let $(a_{r+1}, b_{r+1}) := (a_1, b_1)$)

$$R = \sum_{i=1}^{r} \gcd(a_{i+1} - a_i, b_{i+1} - b_i),$$

that the volume is given by

$$\mathrm{Vol}(\Gamma) = \sum_{i=2}^{r-1} \left| \frac{(a_i - a_1)(b_{i+1} - b_1) - (a_{i+1} - a_1)(b_i - b_1)}{2} \right|$$

and that the genus of $\overline{f}$ is

$$g = \mathrm{Vol}(\Gamma) - \frac{R}{2} + 1$$

by Pick's theorem [48]. All of these can be computed in $\widetilde{O}(g \log \delta)$ time.

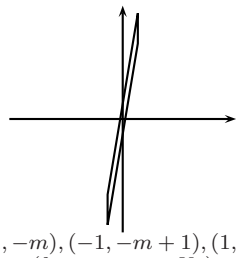
[2]The same representation will be used for the lift $f \in \mathbb{Z}_q[\mathbb{Z}^2]$.

6.1.3 Preliminary step: optimizing the Newton polytope

In this step, we will modify the equation of $\overline{f}$ in a way that preserves the zeta function and the property of being nondegenerate, such that:

1. the origin is an interior point of $\Gamma(\overline{f})$ – so that we can apply the theory of Section 4.5;
2. $\Gamma(\overline{f})$ has a unique top vertex (c_t, d_t) and a unique bottom vertex (c_b, d_b) – so that the set $\mathcal{B} = \{x^i y^j \mid (i,j) \in \mathbb{Z}^2,\ d_b \leq j < d_t\}$ is an $\mathbb{F}_q$-basis for $\frac{\mathbb{F}_q[\mathbb{Z}^2]}{(\overline{f})}$, a $\mathbb{Z}_q$-basis for $\frac{\mathbb{Z}_q[\mathbb{Z}^2]}{(f)}$ and a $\mathbb{Q}_q$-basis for $\frac{\mathbb{Q}_q[\mathbb{Z}^2]}{(f)}$ (where f is – as before – a Newton polytope preserving lift of $\overline{f}$); this will play a crucial role in the reduction algorithm (Subsection 6.1.5);
3. the width and height of $\Gamma(\overline{f})$ are bounded by some fixed polynomial expression in g – this allows us to get rid of the parameter δ during complexity analysis.

The third condition is the hardest. At first sight, one might think that it is superfluous. Indeed, for 'most common' Newton polytopes (we will come back to this informal notion in the next section) that satisfy condition 1, one expects that $\delta \sim \sqrt{g}$. But in general δ is unbounded for fixed g, as shown in the following picture.



$\Gamma = \mathrm{Conv}\{(-1,-m),(-1,-m+1),(1,m),(1,m-1)\}$
(for some $m \in \mathbb{N}_0$)

In this example, δ grows linearly with m, while g stays 1.

$\mathbb{Z}$-affine maps

The transformations we will consider are given by $\mathbb{Z}$-*affine maps*, these are just affine maps $\mathbb{A}^2_{\mathbb{Z}} \to \mathbb{A}^2_{\mathbb{Z}}$, i.e. given by

$$\varphi : (i,j) \mapsto A \cdot (i,j) + (c,d)$$

for some $(c,d) \in \mathbb{A}^2_{\mathbb{Z}}$ and some invertible $A \in \mathbb{Z}^{2\times 2}$. Such a φ naturally induces maps

$$\mathbb{R}^2 \to \mathbb{R}^2 : (i,j) \mapsto A \cdot (i,j) + (c,d)$$

and (for a domain R)

$$R[\mathbb{Z}^2] \to R[\mathbb{Z}^2] : \sum r_{ij}(x,y)^{(i,j)} \mapsto \sum r_{ij}(x,y)^{\varphi(i,j)}$$

(where $(x,y)^{(i,j)}$ abbreviates $x^i y^j$). We will again denote both maps by φ: it will always be clear from the context which map is concerned. On the combinatorial side, the most important property of $\mathbb{Z}$-affine maps is that they preserve volumes. On the algebraic side, the most important property is that they are isomorphisms (so that a lot of geometric features, including the zeta function, stay unchanged) that preserve nondegeneracy. We leave the details to the reader.

Bounding δ

We begin with the third condition.

6.1 Lemma *Let Γ be a two-dimensional convex polytope in $\mathbb{R}^2$ with integer vertex coordinates. Then there exists a $\mathbb{Z}$-affine map φ such that the distance between any two vertices of $\varphi(\Gamma)$ is bounded by*

$$(r-2)\frac{48}{\pi}\mathrm{Vol}(\varphi(\Gamma)),$$

where r is the number of vertices of Γ.

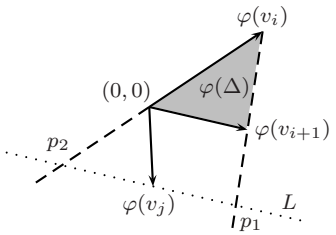
Proof. By translating, we may suppose that the origin is a vertex of Γ. Enumerate the other vertices $v_1, \ldots, v_{r-1}$ clockwise. Take $i \in \{1, \ldots, r-2\}$ such that the triangle Δ spanned by 0, v_i and v_{i+1} has volume $\geq \frac{\mathrm{Vol}(\Gamma)}{r-2}$. Because of Minkowski's theorem there exists a linear $\mathbb{Z}$-affine map φ such that

$$\|\varphi(v_i)\| \cdot \|\varphi(v_{i+1})\| \leq \frac{8}{\pi}\mathrm{Vol}(\Delta).$$

Recall that $\mathbb{Z}$-affine maps preserve volumes. Let $j \in \{1, \ldots, r-1\}$. We will show that $\|\varphi(v_j)\| \leq (r-2)\frac{24}{\pi}\mathrm{Vol}(\Delta)$ from which the result follows. Without loss of generality we may assume that $j > i+1$. The case $j < i$ is similar and $j = i, j = i+1$ are trivial. Then because of the convexity of $\varphi(\Gamma)$ and because the origin is a vertex of Γ (hence of $\varphi(\Gamma)$), $\varphi(v_j)$ lies in the cone with top $\varphi(v_i)$ and spanned by

$$-\varphi(v_i) \quad \text{and} \quad \varphi(v_{i+1}) - \varphi(v_i).$$

Let L be the line through $\varphi(v_j)$ that is parallel to $\varphi(v_{i+1})$ and let p_1 and p_2 be the intersection points with the edges of the cone described above. All of this is illustrated in the picture below.



Now combining the estimates

$$\|\varphi(v_j)\| \leq \|\varphi(v_{i+1})\| + \|p_1 - \varphi(v_{i+1})\| + \|\varphi(v_j) - p_1\|,$$
$$\|\varphi(v_j)\| \leq \|p_2\| + \|\varphi(v_j) - p_2\|$$

and

$$\|p_1 - p_2\| \leq \|p_2\| + \|\varphi(v_{i+1})\| + \|p_1 - \varphi(v_{i+1})\|$$

we obtain that

$$\|\varphi(v_j)\| \leq \|\varphi(v_{i+1})\| + \|p_1 - \varphi(v_{i+1})\| + \|p_2\|.$$

Now note that the triangle spanned by 0, $\varphi(v_{i+1})$ and $\varphi(v_j)$ has the same volume as the one spanned by 0, $\varphi(v_{i+1})$ and p_2, which is $\frac{\|p_2\|}{\|\varphi(v_i)\|}\mathrm{Vol}(\Delta)$. This gives us the estimate

$$\|p_2\| \leq \frac{\mathrm{Vol}(\Gamma) - \mathrm{Vol}(\Delta)}{\mathrm{Vol}(\Delta)}\|\varphi(v_i)\| \leq (r-3)\|\varphi(v_i)\|.$$

Similarly, one obtains

$$\|p_1 - \varphi(v_{i+1})\| \leq (r-3)\|\varphi(v_{i+1}) - \varphi(v_i)\|.$$

We conclude that

$$\|\varphi(v_j)\| \leq \|\varphi(v_{i+1})\| + (r-3)\|\varphi(v_i)\| + (r-3)\|\varphi(v_{i+1}) - \varphi(v_i)\|$$
$$\leq (r-2)\|\varphi(v_{i+1})\| + 2(r-2)\|\varphi(v_i)\|.$$

The result follows. ■

Note that $r \leq R \leq 2g+7$, so we have our desired bound for δ in terms of g.

Unique top and bottom vertex

The next step is to fulfill condition 2, i.e. to transform our polytope such that it has a unique top and bottom vertex.

6.2 Lemma *Let Γ be a two-dimensional convex polytope in $\mathbb{R}^2$ with integer vertex coordinates. Then there exists a $\mathbb{Z}$-affine map φ such that $\varphi(\Gamma)$ has a unique top vertex and a unique bottom vertex. Moreover, if $B \in \mathbb{R}^+$ is such that the distance between any two vertices of Γ is bounded by B, then the distance between any two vertices of $\varphi(\Gamma)$ is bounded by $4\sqrt{2}B^2$.*

Proof. Take vertices v_t and v_b such that $B := \|v_t - v_b\|$ is maximal. By translating if necessary, we may assume that v_b coincides with the origin. Take coprime $\alpha, \beta \in \mathbb{Z}^2$ such that (α, β) is perpendicular to v_t (and such that $(\beta, -\alpha)$ points in the direction of v_t). Choose $c, d \in \mathbb{Z}$ such that $\alpha d - \beta c = -1$. Then

$$\varphi : \mathbb{R}^2 \to \mathbb{R}^2 : \begin{pmatrix} X \\ Y \end{pmatrix} \mapsto \begin{pmatrix} -d & c \\ \beta & -\alpha \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \end{pmatrix}$$

is an invertible $\mathbb{Z}$-linear map such that $\varphi(\Gamma)$ has a unique top vertex $\varphi(v_t)$ and a unique bottom vertex $\varphi(v_b)$. Note that $\|(\alpha,\beta)\| \leq \|v_t\| = B$ and that we can take $|c|, |d| \leq \max\{|\alpha|, |\beta|\} \leq B$. Therefore, the norm of any vertex $\varphi(v_i)$ of $\varphi(\Gamma)$ is bounded by

$$2\sqrt{2}B^2.$$

The distance between any two vertices of $\varphi(\Gamma)$ is then bounded by $4\sqrt{2}B^2$. ■

Full optimization

As the first condition can be satisfied by simply translating the Newton polytope, we are now ready to describe and analyze the full procedure.

`STEP I.` Choose a vertex $v = (a, b)$ of Γ and set $\overline{f} \leftarrow x^{-a}y^{-b}\overline{f}$, $\Gamma \leftarrow \Gamma(\overline{f}) = \Gamma - (a, b)$.

`STEP II.` Choose adjacent vertices $v_i, v_{i+1} \neq (0, 0)$ such that $(r-2)\mathrm{Vol}(\Delta) \geq \mathrm{Vol}(\Gamma)$, where r is the number of vertices of Γ and Δ is the triangle spanned by v_i, v_{i+1} and $(0, 0)$. Apply Euclid's algorithm [10, Algorithm 3.1] to find shortest vectors w_i and w_{i+1} of the lattice spanned by v_i and v_{i+1}, together with an invertible $A \in \mathbb{Z}^{2\times 2}$ such that $Av_i = w_i$ and $Av_{i+1} = w_{i+1}$. Set $\overline{f} = \sum f_{ij}(x, y)^{(i,j)} \leftarrow \sum f_{ij}(x, y)^{A(i,j)}$ and $\Gamma \leftarrow \Gamma(\overline{f})$.

`STEP III.` Choose vertices v_t and v_b such that $\|v_t - v_b\|$ is maximal and set $\overline{f} \leftarrow (x, y)^{-v_b}\overline{f}$, $\Gamma \leftarrow \Gamma - v_b$, $v_t \leftarrow v_t - v_b$.

`STEP IV.` Write $v_b = r(\beta, -\alpha)$ for some coprime α, β and some $r \in \mathbb{N}$. Compute c, d such that $\alpha d - \beta c = -1$ and such that $|c|, |d| \leq \max\{|\alpha|, |\beta|\}$. Let

$$A = \begin{pmatrix} -d & c \\ \beta & -\alpha \end{pmatrix} \cdot \begin{pmatrix} X \\ Y \end{pmatrix}$$

and set $\overline{f} = \sum f_{ij}(x, y)^{(i,j)} \leftarrow \sum f_{ij}(x, y)^{A(i,j)}$ and $\Gamma \leftarrow \Gamma(\overline{f})$.

`STEP V.` Choose a point (a, b) in $(\Gamma \setminus \partial\Gamma) \cap \mathbb{Z}^2$ and set $\overline{f} \leftarrow x^{-a}y^{-b}\overline{f}$, $\Gamma \leftarrow \Gamma(\overline{f}) = \Gamma - (a, b)$.

The time complexity of this optimization procedure is dominated by `STEP II` and amounts to $\widetilde{O}(g \log^2 \delta)$. The space needed is $O(g \log \delta)$.

6.1.4 Asymptotic estimates of some parameters

From now on, we suppose that $\overline{f} \in \mathbb{F}_q[\mathbb{Z}^2]$ is such that the conditions mentioned at the beginning of the foregoing subsection are satisfied. Then one can introduce a set of new parameters in terms of which we will describe the time and space complexity of our algorithm. All of these parameters can in principle be

bounded by some polynomial expression in the genus g of $\overline{f}$, so that we truly obtain a deterministic algorithm with polynomial running time in the input size. But since these bounds are often far too pessimistic, it is more meaningful to keep these new parameters in our complexity estimates.

In the following, we will often state that some property holds for *most common polytopes*: this is not intended to be made mathematically exact. It just means that the Newton polytope should not be shaped too exotically. But for instance, the statement will always be true if the Newton polytope has a unique right-most vertex and a unique left-most vertex that lie on the x-axis, and a unique top vertex and a unique bottom vertex that lie on the y-axis.

We first note that the genus g, which equals the number of interior lattice points of $\Gamma = \Gamma(\overline{f})$ can be interchanged with the volume of Γ or with its total number of lattice points, as they are all asymptotically equivalent. Indeed, this follows from Scott's result mentioned above, together with Pick's theorem:

$$g \leq \#(\Gamma \cap \mathbb{Z}^2) \leq 3g + 7$$

$$g \leq \mathrm{Vol}(\Gamma) \leq 2g + 3.$$

(given $g \geq 1$).

Another parameter is δ, as defined above. We will also make use of the width w, i.e. the maximal difference between the first coordinates of two points of Γ, and the height h, i.e. $d_t - d_b$. Of course, $h, w \leq 2\delta \leq 2w + 2h$. For most common polytopes, wh will behave like g. In general we can use the rough bound

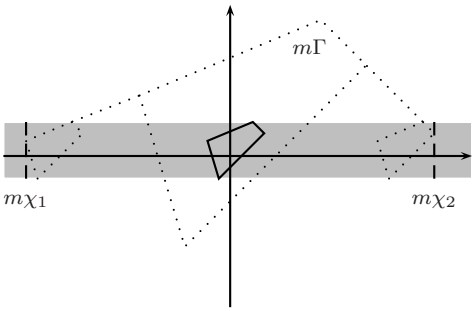
$$w, h \leq 2\sqrt{2}(r-2)^2 \frac{48^2}{\pi^2} \mathrm{Vol}(\Gamma)^2 = O(g^4)$$

that was proven in the foregoing subsection.

Next, we need some parameters that at first sight seem to depend on the geometry of $\widetilde{C} = V(\overline{f})$ and $C = V(f)$ (where f is a Newton polytope preserving lift of $\overline{f}$). But in fact they can be bounded in terms of Γ. First, we need $\chi_1, \chi_2 \in \mathbb{Z}$ such that

$$\mathcal{L}(mD_C) \subset S_{[m\chi_1, m\chi_2]}$$

for all $m \in \mathbb{N}_0$, where $S_{[r,s]}$ is the $\mathbb{Q}_q$-vector space generated by the subset $\{x^i y^j \mid i \in [r,s],\ d_b \leq j < d_t\}$ $(r, s \in \mathbb{Z})$ of the basis $\mathcal{B}$ that was introduced in condition 2 at the beginning of the foregoing subsection. By Theorem 2.18, $m\chi_1$ and $m\chi_2$ in fact measure the space needed to represent an element of $L_{m\Gamma}$.



As can be seen on the picture, χ_1 and χ_2 are determined by the slopes of the top and bottom edges of Γ. Denote as before the top vertex with (c_t, d_t) and let (a, b) be the clockwise-next vertex. Suppose that $a \geq c_t$. Then it is not hard to see that Laurent polynomials with support in the upper half plane part of $m\Gamma$ reduce (modulo f) into $S_{[-\infty, m\tau]}$ where $\tau = c_t + \left\lfloor \frac{d_t(a-c_t)}{d_t-b} \right\rfloor$. Now

$$\frac{d_t(a-c_t)}{d_t-b} = (a-c_t) + \frac{b(a-c_t)}{d_t-b} \leq w + b(a-c_t) \leq w + 2\mathrm{Vol}(\Gamma) \leq 4g + w + 6.$$

The one but last inequality comes from the fact that the triangle with vertices $(0,0)$, (c_t, d_t), (a, b) is contained in Γ. Its volume equals

$$\frac{ad_t - c_t b}{2} \geq \frac{(a-c_t)b}{2}.$$

Therefore, $\tau \leq c_t + 4g + w + 6$. Using the same argument for the lower half plane, we conclude that $\mathcal{L}(mD_C) \subset S_{[-\infty, m(\max(c_t, c_b)+4g+w+6)]}$. This is definitely also true when $a < c_t$. By analogy, $\mathcal{L}(mD_C) \subset S_{[m(\min(c_t,c_b)-4g-w-6), +\infty]}$, which proves that we can take χ_1, χ_2 such that $\chi_2 - \chi_1 \leq 8g + 3w + 12$. For most common polytopes, $h(\chi_2 - \chi_1)$ is expected to be $O(g^{3/2})$.

Strongly related with the foregoing are optimal $\kappa_1, \kappa_2 \in \mathbb{Z}$ such that $\mathcal{L}(D_C + E_y + \mathrm{Div}_\infty(x) + \mathrm{Div}_0(x)) \subset S_{[\kappa_1, \kappa_2]}$, see Corollary 6.5 below. Note that

$$\pm \mathrm{ord}_P(x) \leq h \quad \text{and} \quad \pm \mathrm{ord}_P(y) \leq w$$

for any place $P \in C \setminus \mathbb{T}^2_{\mathbb{Q}_q}$: this follows from Corollary 2.13. Therefore, $\mathcal{L}(D_C + E_y + \mathrm{Div}_\infty(x) + \mathrm{Div}_0(x)) \subset \mathcal{L}((hw + 2h + 1)D_C)$. By the foregoing, we conclude that we can take $\kappa_2 - \kappa_1 = O(hw(\chi_2 - \chi_1)) = O(hw(g + w))$. But for most common polytopes, a much better bound holds: we can omit E_y (see Remark 6.3) and have that $\mathrm{Div}_\infty(x) + \mathrm{Div}_0(x) \leq 2D_C$. Therefore, we can use the same asymptotic as above, i.e. $O(g + w)$.

Finally, during complexity analysis, we will often make use of the trivial estimates $g \leq h(\chi_2 - \chi_1), h(\kappa_2 - \kappa_1)$ and $w \leq \chi_2 - \chi_1$.

6.1.5 Differential reduction

Let $f \in \mathbb{Z}_q[\mathbb{Z}^2]$ be a Newton polytope preserving lift of $\overline{f}$, and let $C = V(f)$. In this section we will describe a method to reduce elements of $A = \mathbb{Q}_q[\mathbb{Z}^2]/(f)$ modulo the operator D that was studied extensively in Chapter 4. There we proved the following (Theorem 4.2 with E replaced by $(m-2)D_C$):

For any $m \in \mathbb{N} \setminus \{0, 1\}$ the map

$$\mathcal{L}^{(0)}((m-1)D_C) \xrightarrow{D} \frac{\mathcal{L}^{(1)}(mD_C)}{\mathcal{L}^{(1)}(2D_C)}$$

is surjective.

This can be turned into a reduction algorithm and also provides a sharp bound for the loss of precision incurred during reduction. Indeed, since the Newton polytope Γ contains the origin as an interior point, any Laurent polynomial $h \in \mathbb{Z}_q[\mathbb{Z}^2]$ will be contained in an $L^{(0)}_{m\Gamma}$ with $m \in \mathbb{N}_0$ big enough. Let

$$\varepsilon = \left\lceil \log_p \max\{-\mathrm{ord}_P(h)\}_{P \in C \setminus \mathbb{T}^2_{\bar{\mathbb{Q}}_q}} \right\rceil ,$$

then clearly $p^\varepsilon h \in L^{(1)}_{m\Gamma}$. So we can as well assume that $h \in L^{(1)}_{m\Gamma}$. By Theorem 2.18, we have $\mathcal{L}(mD_C) = L_{m\Gamma}$ and applying the above shows that there exists a $g \in L^{(0)}_{(m-1)\Gamma}$ such that $h_r = h - D(g) \in \mathcal{L}^{(1)}(2D_C)$. Note that after multiplication with p^ε the entire reduction process is integral, so if we want to recover the result h_r modulo p^N, we need to compute h modulo $p^{N+\varepsilon}$. To finalize the computation, we need to express h_r on a basis for $H^1_{DR}(f/\mathbb{Q}_q)$, which could cause a further loss of precision, depending on the basis chosen. But clearly, as long as we choose a '$\mathbb{Z}_q$-module basis' for $H^1_{DR}(f/\mathbb{Q}_q)$, no further loss of precision will occur. More precisely, we mean the following. Consider the module

$$M_H = \frac{\mathcal{L}^{(0)}(2D_C)}{D(\mathcal{L}(D_C)) \cap L^{(0)}} ,$$

then M_H is a free $\mathbb{Z}_q$-module since it is finitely generated and torsion-free. Therefore, by Theorem 2.20 any $\mathbb{Z}_q$-basis for M_H forms a suitable basis for $H^1_{DR}(f/\mathbb{Q}_q)$, such that in the final reduction step, no further loss of precision is incurred.

In the above description, we used any representative for an element of the coordinate ring of C; in practice however, we would like to work with a unique representative. Given the Newton polytope Γ of f, there are many possibilities to choose a suitable basis $\mathcal{B}$ for $\mathbb{Q}_q[\mathbb{Z}^2]/(f)$. But the assumptions about Γ made in Subsection 6.1.3 already led to the following natural choice

$$\mathcal{B} = \{x^k y^l \mid k, l \in \mathbb{Z}, d_b \leq l < d_t\} ,$$

with (c_t, d_t) (resp. (c_b, d_b)) the unique highest (resp. lowest) point of Γ.

As before, let $S_{[m_1,m_2]}$ with $m_1 < m_2$ denote the space of Laurent polynomials with support in the rectangle $[m_1, m_2] \times [d_b, d_t[$, then the reduction process proceeds in two phases: the first phase reduces terms in $S_{[0,m]}$ with $m \in \mathbb{N}_0$ and the second phase reduces terms in $S_{[-m,0]}$ with $m \in \mathbb{N}_0$. Since both phases are so similar, we will focus mainly on the first phase and briefly mention the changes for the second phase.

Phase 1:

Any element $h \in S^{(0)}_{[0,m]}$ can be forced into $S^{(1)}_{[0,m]}$ by multiplying it with p^ε where

$$\varepsilon = \left\lceil \log_p(mM_x + \Delta) \right\rceil$$

with

$$M_x = \max\{-\mathrm{ord}_P(x)\}_{P \in C \setminus \mathbb{T}^2_{\bar{\mathbb{Q}}_q}}$$

and

$$\Delta = \max\{-\mathrm{ord}_P(y^{d_t-1}), -\mathrm{ord}_P(y^{d_b})\}_{P\in C\setminus\mathbb{T}^2_{\mathbb{Q}_q}}.$$

If we now want to apply Theorem 4.2 to an element $h \in S^{(1)}_{[0,m]}$, we need to find a divisor E over $\mathbb{Q}_q$ such that $S^{(1)}_{[0,m]} \subset \mathcal{L}^{(1)}(2D_C+E)$. Then by Theorem 4.2 there exists a $g \in \mathcal{L}^{(0)}(D_C+E)$ such that $h-D(g) \in \mathcal{L}^{(1)}(2D_C)$. In practice however, we do not want to work with explicit Riemann-Roch spaces; as such we want to find a divisor E (depending on m) and constants $c_1, c_2 \in \mathbb{Z}$ (independent of m) such that

$$S_{[0,m]} \subset \mathcal{L}(2D_C + E) \quad \text{and} \quad \mathcal{L}(D_C + E) \subset S_{[c_1,m+c_2]}.$$

The reduction algorithm then becomes very simple indeed: to reduce $h \in S^{(1)}_{[0,m]}$, we only need to find a $g \in S^{(0)}_{[c_1,m+c_2]}$ such that $h - D(g) \in \mathcal{L}^{(1)}(2D_C)$, using linear algebra.

Recall that the divisor of any function $h \in \mathbb{Q}_q(C)$ can be written as the difference of the zero divisor and the pole divisor, i.e. $\mathrm{Div}(h) = \mathrm{Div}_0(h) - \mathrm{Div}_\infty(h)$, $\mathrm{Div}_0(h) \geq 0$, $\mathrm{Div}_\infty(h) \geq 0$ and $\mathrm{Supp}(\mathrm{Div}_0(h)) \cap \mathrm{Supp}(\mathrm{Div}_\infty(h)) = \emptyset$. Furthermore, two trivial observations are that $h \in \mathcal{L}(\mathrm{Div}_\infty(h))$ and $\mathrm{Div}_\infty(h^{-1}) = \mathrm{Div}_0(h)$. Consider the divisor

$$E_m = -d_b\mathrm{Div}_0(y) + (d_t - 1)\mathrm{Div}_\infty(y) + m\mathrm{Div}_\infty(x)$$

then $E_m \geq 0$ and $S_{[0,m]} \subset \mathcal{L}(E_m) \subset \mathcal{L}(2D_C+E_m)$, so we can apply Theorem 4.2 with $E = E_m$. Note that E_m is indeed defined over $\mathbb{Q}_q$.

6.3 Remark It is clear that the choice for E_m is not entirely optimal, since we could subtract the contributions in $2D_C$ and still obtain the above inclusion. The most important simplification in practice is that 2Γ is 'likely' to contain the interval $[d_b, d_t - 1]$ on the y-axis and then E_m can be simply taken to be $m\mathrm{Div}_\infty(x)$. However, in general this need not be the case. ■

To determine the constants c_1 and c_2 we first prove the following lemma.

6.4 Lemma *Let E be a divisor on C which is defined over $\mathbb{Q}_q$ and with $\deg E > 2g-2$, and let $h \in \mathbb{Q}_q(C)$ be a function on C. Then for any $m \in \mathbb{N}_0$ the following map is an isomorphism:*

$$\frac{\mathcal{L}(E + Div_\infty(h))}{\mathcal{L}(E)} \xrightarrow{\cdot h^{m-1}} \frac{\mathcal{L}(E + mDiv_\infty(h))}{\mathcal{L}(E + (m-1)Div_\infty(h))}.$$

Proof. Since $\deg E > 2g - 2$ and $\mathrm{Div}_\infty(h) \geq 0$, the Riemann-Roch theorem implies that the dimensions of both vector spaces are equal to $\deg \mathrm{Div}_\infty(h)$, so it suffices to prove injectivity. Let $g \in \mathcal{L}(E + \mathrm{Div}_\infty(h))$ and assume that $h^{m-1}g \in \mathcal{L}(E + (m-1)\mathrm{Div}_\infty(h))$, i.e.

$$(m-1)\mathrm{Div}(h) + \mathrm{Div}(g) \geq -E - (m-1)\mathrm{Div}_\infty(h),$$

which implies that $\mathrm{Div}(g) \geq -E - (m-1)\mathrm{Div}_0(h)$. Since $g \in \mathcal{L}(E + \mathrm{Div}_\infty(h))$, i.e. $\mathrm{Div}(g) \geq -E - \mathrm{Div}_\infty(h)$ and the supports of $\mathrm{Div}_0(h)$ and $\mathrm{Div}_\infty(h)$ are disjoint, we conclude $\mathrm{Div}(g) \geq -E$ or $g \in \mathcal{L}(E)$. ■

In what follows, we will use the abbreviation $E_y = -d_b\mathrm{Div}_0(y) + (d_t - 1)\mathrm{Div}_\infty(y)$, so $E_m = E_y + m\mathrm{Div}_\infty(x)$. As in Subsection 6.1.4, choose integers $\kappa_1 \leq 0$ and $\kappa_2 \geq 0$ such that $\mathcal{L}^{(0)}(D_C + E_y + \mathrm{Div}_\infty(x) + \mathrm{Div}_0(x)) \subset S_{[\kappa_1,\kappa_2]}$. In particular, $\mathcal{L}^{(0)}(D_C + E_1) \subset S_{[\kappa_1,\kappa_2]}$. This can then be generalized to the following.

6.5 Corollary $\mathcal{L}(D_C + E_m) \subset S_{[\kappa_1, m-1+\kappa_2]}$.

Proof. Apply Lemma 6.4 with $E = D_C + E_y$ and $h = x$. ■

Thus, given $h \in S^{(1)}_{[0,m]}$ we find $g \in S^{(0)}_{[\kappa_1,m-1+\kappa_2]}$ such that $h - D(g) \in \mathcal{L}^{(1)}(2D_C)$ using linear algebra over $\mathbb{Z}_q$. However, for big m the linear systems involved get quite large, so we compute g in several steps: let $h_0 = h$ and choose a constant $c \in \mathbb{N}_0$, then in step $1 \leq i \leq t$ (where t will be determined later) we compute a g_i such that

$$h_i = h_{i-1} - D(g_i) \in S^{(1)}_{[0,m-ic]}\,.$$

In the last step, i.e. step $t+1$ we find a $g_{t+1} \in S^{(0)}_{[\kappa_1,m-tc-1+\kappa_2]}$ such that

$$h_{t+1} = h_t - D(g_{t+1}) \in \mathcal{L}^{(1)}(2D_C)\,.$$

We postpone this last step until after Phase 2, since it is better to treat the last steps of both phases at once. To determine which monomials appear in the g_i for $1 \leq i \leq t$ we prove the following lemma.

6.6 Lemma *For $m \in \mathbb{N}_0, k \in \mathbb{Z}$ with $d_b \leq k < d_t$, we have $D(x^m y^k) \in S^{(1)}_{[\kappa_1+m-1,\kappa_2+m-1]}$.*

Proof. By definition of D we have $D(x^m y^k) = x^m y^k(myf_y - kxf_x)$. Note that the support of $g = myf_y - kxf_x$ is contained in Γ and thus $g \in \mathcal{L}(D_C)$. Furthermore, by definition of E_y we have $y^k \in \mathcal{L}(E_y)$. Therefore, by definition of κ_1 and κ_2 we conclude that $D(x^m y^k) \in S^{(1)}_{[\kappa_1+m-1,\kappa_2+m-1]}$. ■

The above lemma finalizes the description of the algorithm: in step i it suffices to take g_i in $S_{[a_i,b_i]}$ with

$$a_i = m - ic - \kappa_2 + 2 \quad \text{and} \quad b_i = m - (i-1)c + \kappa_2 - 1\,,$$

and to work modulo x^{m-ic}. There are two natural conditions that t and c should satisfy. The first one is related to the fact that we want to work in $S_{[0,+\infty]}$ only. Therefore,

$$a_t \geq -\kappa_1 + 1 \quad \text{which is equivalent with} \quad tc \leq m + \kappa_1 - \kappa_2 + 1\,.$$

The second condition keeps track of the fact that something which is already in $\mathcal{L}^{(1)}(2D_C)$ cannot be reduced anymore. Let $\chi_1 \leq 0, \chi_2 \geq 0$ be such that $L_{m\Gamma} \subset L_{[m\chi_1, m\chi_2]}$ (as in Subsection 6.1.4). Then $\mathcal{L}^{(1)}(2D_C) \subset L_{[2\chi_1, 2\chi_2]}$ and it suffices to impose

$$tc \leq m - 2\chi_2.$$

The number of unknowns in the linear system of equations in step i is precisely the number of monomials in $S_{[a_i,b_i]}$, which equals $(d_t - d_b)(c + 2\kappa_2 - 2)$. Note that this also appears as a natural upper bound for the number of terms in $D(S_{[a_i,b_i]})$ modulo x^{m-ic}, so we obtain a system with at least as many unkowns as equations.

Phase 2:

Since the second phase is very similar to the first, we will only briefly mention the main differences. To force an element $h \in S^{(0)}_{[-m,0]}$ with $m \in \mathbb{N}_0$ into $S^{(1)}_{[-m,0]}$, we need to multiply with p^{ε} where

$$\varepsilon = \left\lceil \log_p(mM_{1/x} + \Delta) \right\rceil$$

with $M_{1/x} = \max\{-\mathrm{ord}_P(x^{-1})\}_{P \in C \setminus \mathbb{T}^2_{\mathbb{Q}_q}}$ and Δ as before, so from now on assume that $h \in S^{(1)}_{[-m,0]}$. The divisor E_m now becomes $E_m = E_y + m\mathrm{Div}_\infty(x^{-1})$ and applying Lemma 6.4 with $h = x^{-1}$ shows

$$\mathcal{L}(D_C + E_y + m\mathrm{Div}_\infty(x^{-1})) \subset S_{[-m+1+\kappa_1, \kappa_2]}\,,$$

where κ_1, κ_2 are chosen as in Phase 1. In step i we now compute a g_i such that $h_i = h_{i-1} - D(g_i) \in S^{(1)}_{[-m+ic,0]}$ for some constant $c \in \mathbb{N}_0$. An analogue of Lemma 6.6 (replace $S^{(1)}_{[\kappa_1+m-1,\kappa_2+m-1]}$ with $S^{(1)}_{[\kappa_1-m+1,\kappa_2-m+1]}$) finally leads to $g_i \in S^{(0)}_{[a_i,b_i]}$ with

$$a_i = -m + (i-1)c + \kappa_1 + 1 \quad \text{and} \quad b_i = -m + ic - \kappa_1 - 2\,.$$

The number of steps t is determined by the following inequalities:

$$tc \leq m + \kappa_1 - \kappa_2 + 1 \quad \text{and} \quad tc \leq m + 2\chi_1.$$

The systems to be solved have $(d_t - d_b)(c - 2\kappa_1 - 2)$ unknowns, that are related by at most the same number of equations.

Step $t+1$:

During Phase 1 and Phase 2, we reduced a given polynomial $h \in L^{(1)}$ modulo D to obtain a polynomial $h_t \in S^{(1)}_{[-n_1,n_2]}$, where $n_1 \in \mathbb{N}_0$ is roughly of size $\max\{-2\chi_1, \kappa_2 - \kappa_1\}$ and $n_2 \in \mathbb{N}_0$ is roughly of size $\max\{2\chi_2, \kappa_2 - \kappa_1\}$. In this last step, we reduce to a polynomial $h_{t+1} \in \mathcal{L}^{(1)}(2D_C)$ by brute force.

From Corollary 6.5 (and its Phase 2 analogue) we know that there is a $g_{t+1} \in S^{(0)}_{[-n_1+1+\kappa_1, n_2-1+\kappa_2]}$ such that

$$h_t - D(g_{t+1}) \in \mathcal{L}^{(1)}(2D_C),$$

so we can compute h_{t+1} by solving a system of at most $(d_t - d_b)(2(\kappa_2 - \kappa_1) + n_1 + n_2 - 3)$ equations in

$$(d_t - d_b)(\kappa_2 - \kappa_1 + n_1 + n_2 - 1) + \#(2\Gamma \cap \mathbb{Z}^2)$$

unknowns. Here, the latter term equals $4\mathrm{Vol}(\Gamma) + \#(\partial\Gamma \cap \mathbb{Z}^2) + 1$ by Ehrhart's theorem [32].

Bounding the non-zero invariant factors:

We conclude with a bound on the non-zero invariant factors of the systems that are considered above, which is necessary if we want to solve them using the algorithm from Section 5.1.

6.7 Lemma *Let $m \in \mathbb{N}_0$ be the level at which the reduction starts, i.e. suppose that the polynomial to be reduced is in $S^{(0)}_{[-m,m]}$. The p-adic valuations of the non-zero invariant factors of the matrices A appearing in our reduction algorithm are bounded by $\theta = \lceil \log_p((m + 2(\kappa_2 - \kappa_1 + 1))M + \Delta) \rceil$, where*

$$M = \max\{\pm \mathrm{ord}_P(x)\}_{P \in C \setminus \mathbb{T}^2_{\mathbb{Q}_q}}$$

and

$$\Delta = \max\{-\mathrm{ord}_P(y^{d_t - 1}), -\mathrm{ord}_P(y^{d_b})\}_{P \in C \setminus \mathbb{T}^2_{\mathbb{Q}_q}}.$$

Proof. We claim that A has the following property: *if $b \in p^\theta \mathbb{Z}_q^r$ is such that the system $A \cdot \boldsymbol{x} = b$ has a solution in $\mathbb{Q}_q^s$, then it has a solution in $\mathbb{Z}_q^s$.* Since N_1 and N_2 are invertible over $\mathbb{Z}_q$, this property then still holds for the matrix $N_1 \cdot A \cdot N_2$, from which the result easily follows.

For simplicity, we will only prove the claim in case A comes from the system that has to be solved during Step 1 of Phase 1. The other cases work similarly. Let $b \in p^\theta \mathbb{Z}_q^r$ be such that $A \cdot \mathbf{x} = b$ has a solution in $\mathbb{Q}_q^s$. Then b corresponds to a polynomial

$$h_b \in S^{(1)}_{[m-c+1, m+2\kappa_2-2]}$$

for which there exists a $g \in S_{[m-c-\kappa_2+2, m-1+\kappa_2]}$ such that

$$h_b - D(g) \in S_{[0, m-c]}.$$

By Theorem 4.2 and Corollary 6.5 (see the first sentence after the proof of Corollary 6.5), we can reduce this further to eventually obtain a $g \in S_{[\kappa_1, m-1+\kappa_2]}$ such that

$$h_b - D(g) \in \mathcal{L}(2D_C).$$

Now, let $\{v_1, \ldots, v_m\}$ be a $\mathbb{Q}_q$-basis for

$$\frac{\mathcal{L}(2D_C)}{D(\mathcal{L}(D_C))}.$$

As explained in Section 2.4, this is also a basis for $H^1_{DR}(f/\mathbb{Q}_q)$. In any case, we can find a $g_0 \in \mathcal{L}(D_C)$ such that $h_b - D(g) - D(g_0) = \lambda_1 v_1 + \cdots + \lambda_m v_m$ for some $\lambda_1, \ldots, \lambda_m \in \mathbb{Q}_q$.

On the other hand, since $h_b \in S^{(1)}_{[m-c+1, m+2\kappa_2-2]}$, we can find a Laurent polynomial $g' \in S^{(0)}_{[\kappa_1, m+3\kappa_2-3]}$ such that

$$h_b - D(g') \in \mathcal{L}(2D_C),$$

again by Theorem 4.2 and Corollary 6.5. Finally, we find a $g'_0 \in \mathcal{L}(D_C)$ for which $h_b - D(g') - D(g'_0) = \mu_1 v_1 + \cdots + \mu_m v_m$ for some $\mu_1, \ldots, \mu_m \in \mathbb{Q}_q$.

Using uniqueness, we conclude that $D(g+g_0) = D(g'+g'_0)$. Hence $d(g+g_0) = \Lambda D(g+g_0) = \Lambda D(g'+g'_0) = d(g'+g'_0)$ so that $g+g_0$ and $g'+g'_0$ only differ by a constant. In particular, $g' \in S^{(0)}_{[\kappa_1, m-1+\kappa_2]}$. This concludes the proof. ■

6.1.6 The algorithm

We are now ready to describe our point counting algorithm.

STEP 0: compute p-adic lift of $\overline{f}$

First note that we assume that $\mathbb{F}_p$ is represented as $\mathbb{Z}/(p)$ and that $\mathbb{F}_q$ is represented as $\mathbb{F}_p/(\overline{r}(X))$ for some monic irreducible degree n polynomial $\overline{r}(X)$. Take $r(X) \in \mathbb{Z}[X]$ such that it has coefficients in $\{0, \ldots, p-1\}$ and reduces to $\overline{r}(X)$ modulo (p). Then $\mathbb{Z}_q$ can be represented as $\mathbb{Z}_p/(r(X))$. Let

$$\overline{a}_{n-1}[X]^{n-1} + \cdots + \overline{a}_1[X] + \overline{a}_0$$

be any element of $\mathbb{F}_q$. By the *canonical*[3] *lift* to $\mathbb{Z}_q$, we mean

$$a_{n-1}[X]^{n-1} + \cdots + a_1[X] + a_0,$$

where the $a_j \in \{0, \ldots, p-1\}$ are the unique elements that reduce to $\overline{a}_j$ mod (p). Finally, if $\overline{f} = \sum_{(i,j)\in\mathbb{Z}^2\cap\Gamma} \overline{b}_{ij} x^i y^j$, define $f = \sum_{(i,j)\in\mathbb{Z}^2\cap\Gamma} b_{ij} x^i y^j$ where the b_{ij} are canonical lifts.

Complexity analysis. This step needs $\widetilde{O}(ng)$ time and space.

[3] Of course, from a mathematical point of view this lift is not very canonical (e.g. it depends on the choice of $r(X)$).

STEP I: determine p-adic precision

Assume that all calculations are done modulo p^N for some $N \in \mathbb{N}$. What conditions should N satisfy? From the material in Subsection 6.1.1, it follows that it suffices to compute $\chi(t)$ modulo $p^{\widetilde{N}}$, where

$$\widetilde{N} \geq \left\lceil \log_p \left(2 \binom{2\mathrm{Vol}(\Gamma)+1}{\mathrm{Vol}(\Gamma)} q^{g+R-1} \right) \right\rceil .$$

However, during the reduction process (**STEP V.II**) there is some loss of precision: to ensure that everything remains integral we need to multiply with p^{ε} where

$$\varepsilon = \left\lceil \log_p(mM + \Delta) \right\rceil$$

with

$$M = \max\{\pm \mathrm{ord}_P(x)\}_{P \in C \setminus \mathbb{T}^2_{\mathbb{Q}_q}},$$

$$\Delta = \max\{-\mathrm{ord}_P(y^{d_t-1}), -\mathrm{ord}_P(y^{d_b})\}_{P \in C \setminus \mathbb{T}^2_{\mathbb{Q}_q}}$$

and $m = \max\{|m_1|, |m_2|\}$ the level at which the reduction starts. Here, $m_1, m_2 \in \mathbb{Z}$ are such that the objects to be reduced are in $S_{[m_1,m_2]}$. From Corollary 2.13, it is immediate that $M \leq h$ and $\Delta \leq hw$. To see what m is bounded by, note that the objects to be reduced have support in $(9pN + 5p)\Gamma$ (when computed modulo p^N). Indeed, from **STEP V.I** we see that these objects are of the form

$$\begin{aligned} & yf_y \left(\mathcal{F}_p(x^i y^j)\mathcal{F}_p(\beta) \frac{x\partial \mathcal{F}_p(x)}{\mathcal{F}_p(x)\partial x} - \mathcal{F}_p(x^i y^j)\mathcal{F}_p(\alpha) \frac{x\partial \mathcal{F}_p(y)}{\mathcal{F}_p(y)\partial x} \right) \\ - \quad & xf_x \left(\mathcal{F}_p(x^i y^j)\mathcal{F}_p(\beta) \frac{y\partial \mathcal{F}_p(x)}{\mathcal{F}_p(x)\partial y} - \mathcal{F}_p(x^i y^j)\mathcal{F}_p(\alpha) \frac{y\partial \mathcal{F}_p(y)}{\mathcal{F}_p(y)\partial y} \right). \end{aligned}$$

where $(i, j) \in 2\Gamma$. Here $\alpha, \beta \in \mathbb{Z}_q[\mathbb{Z}^2]$ are Laurent polynomials with support in 2Γ for which $1 \equiv \alpha x f_x + \beta y f_y \bmod f$ (see Corollary 3.16). The bound then follows from Theorem 4.11 and the remark below it.

Since $L_{(9pN+5p)\Gamma} \subset S_{[(9pN+5p)\chi_1,(9pN+5p)\chi_2]}$, we obtain that

$$\varepsilon \leq \left\lceil \log_p((9pN + 5p)\max\{|\chi_1|, \chi_2\}h + hw) \right\rceil .$$

As a consequence, this is a natural bound on the valuations of the denominators appearing in the matrix of $\mathcal{F}_p^*$ (as computed in **STEP VII**). During **STEP VIII** and **STEP IX**, our denominators could grow up to $p^{n(2\mathrm{Vol}(\Gamma)+1)\varepsilon}$. In conclusion, it suffices to take N such that it satisfies $N \geq$

$$\begin{aligned} & \left\lceil \log_p \left(2 \tbinom{2\mathrm{Vol}(\Gamma)+1}{\mathrm{Vol}(\Gamma)} q^{g+R-1} \right) \right\rceil \\ & \qquad + n(2\mathrm{Vol}(\Gamma) + 1) \left\lceil \log_p((9pN + 5p)\max\{|\chi_1|, \chi_2\}h + hw) \right\rceil \end{aligned}$$

In particular, $N = \widetilde{O}(ng)$.

STEP II: compute effective Nullstellensatz expansion

In this step, one computes (up to precision p^N) polynomials $\alpha, \beta, \gamma \in \mathbb{Z}_q[\mathbb{Z}^2]$ with support in 2Γ such that

$$1 = \gamma f + \alpha x \frac{\partial f}{\partial x} + \beta y \frac{\partial f}{\partial y}.$$

This defines a linear system $A \cdot \mathbf{x} = B$ that can be solved using Gaussian elimination, in each step of which the pivot is taken to be a p-adic unit. This is possible since the linear map defined by A is surjective (by Theorem 3.11). In particular, there is no loss of precision. Note that instead of Gaussian elimination, one can use the method described in Chapter 5. In this way, one gains a factor g time. But for the overall complexity analysis this makes no difference.

Complexity analysis. Selecting the entries of A takes $\widetilde{O}(g^2)$ time (see Subsection 6.1.2). One then needs $\widetilde{O}(nNg^3) = \widetilde{O}(n^2g^4)$ time and $O(nNg^2) = \widetilde{O}(n^2g^3)$ space to solve the system.

STEP III: compute lift of Frobenius

Take lifts $\delta, \delta_x, \delta_y \in \mathbb{Z}_q[\mathbb{Z}^2]$ of $\overline{\gamma}^p$, $\overline{\alpha}^p$, $\overline{\beta}^p$ and compute a zero of the polynomial

$$H(Z) = (1+\delta_x Z)^a (1+\delta_y Z)^b f^\sigma(x^p(1+\delta_x Z), y^p(1+\delta_y Z))$$

(as described in Section 4.3.2) up to precision p^N, using Newton iteration and starting from the approximate solution 0. Reduce all intermediate calculations modulo f to the basis $\mathcal{B} = \{ x^i y^j \mid d_b \leq j < d_t \}$ (this is why the terms $-(a\delta_x + b\delta_y - \delta) f^p Z - f^p$, that were added for theoretical reasons, can be omitted in the formula for $H(Z)$). Finally, if we denote the result by Z_0, expand $Z_x := 1 + \delta_x Z_0$, $Z_y := 1 + \delta_y Z_0$ and compute their inverses up to precision p^N using Newton iteration (again reduce the intermediate calculations modulo f). Note that if we take a and b minimal, then $\deg H \leq w + h$.

Complexity analysis. Remark that it is better *not* to expand the polynomial $H(Z)$ (nor its derivative $\frac{dH}{dZ}(Z)$), but to leave it in the above compact representation. The reason is that the expanded versions of H and $\frac{dH}{dZ}$ are very space-consuming.

A similar complexity estimate has been made in [25]. The complexity is dominated by the last iteration step, which in its turn is dominated by $O(g)$ computations of terms of the form

$$(1+\delta_x Z')^i (1+\delta_y Z')^j$$

where $Z' \in S_{[6pN\chi_1, 6pN\chi_2]}$, $i \in \{0, \ldots, w\}$ and $j \in \{0, \ldots, h\}$ (because of (4.13)). Note that reducing a polynomial with support in $[6pN\chi_1, 6pN\chi_2] \times [-\lambda d_b, \lambda(d_t - 1)]$ (for some $\lambda \in \mathbb{N}_0$) to the basis mentioned above can be done in $\widetilde{O}(\lambda h N(\chi_2 - \chi_1) \cdot g \cdot nN) = \widetilde{O}(\lambda n^3 g^3 h(\chi_2 - \chi_1))$ time (at least if we know that all intermediate

results are supported in $[6pN\chi_1, 6pN\chi_2] \times \mathbb{Z}$ modulo p^N). Therefore, the overall time complexity of **STEP III** amounts to $\widetilde{O}(n^3 g^4 h(\chi_2 - \chi_1))$, whereas the space complexity is $\widetilde{O}(n^3 g^2 h(\chi_2 - \chi_1))$. Note that this indeed dominates the time and space needed to compute the Frobenius substitutions, each of which can be done in $\widetilde{O}(n \cdot nN)$ time (see e.g. [15, Section 12.5]).

The complexity of computing $Z_x, Z_y, Z_x^{-1}, Z_y^{-1}$ works similarly and is dominated by the above.

STEP IV: 'precompute' $\mathcal{F}_p^*(dx/xyf_y)$

Here, $\mathcal{F}_p^*$ is the $\mathbb{Q}_q$-vector space endomorphism of $\Omega_C(C \cap \mathbb{T}^2_{\mathbb{Q}_q})$ induced by $\mathcal{F}_p$. Note that $dx/f_y = \beta y dx - \alpha x dy$. Thus $\mathcal{F}_p^*(dx/xyf_y) =$

$$\mathcal{F}_p(\beta)\left(\frac{\partial \mathcal{F}_p(x)}{\mathcal{F}_p(x)\partial x}dx + \frac{\partial \mathcal{F}_p(x)}{\mathcal{F}_p(x)\partial y}dy\right) - \mathcal{F}_p(\alpha)\left(\frac{\partial \mathcal{F}_p(y)}{\mathcal{F}_p(y)\partial x}dx + \frac{\partial \mathcal{F}_p(y)}{\mathcal{F}_p(y)\partial y}dy\right).$$

Rearranging terms gives that this equals

$$\left(\mathcal{F}_p(\beta)\frac{\partial \mathcal{F}_p(x)}{\mathcal{F}_p(x)\partial x} - \mathcal{F}_p(\alpha)\frac{\partial \mathcal{F}_p(y)}{\mathcal{F}_p(y)\partial x}\right)dx + \left(\mathcal{F}_p(\beta)\frac{\partial \mathcal{F}_p(x)}{\mathcal{F}_p(x)\partial y} - \mathcal{F}_p(\alpha)\frac{\partial \mathcal{F}_p(y)}{\mathcal{F}_p(y)\partial y}\right)dy.$$

However, as will become clear in the following step, it is more natural to precompute

$$E := yf_y\left(\mathcal{F}_p(\beta)\frac{x\partial \mathcal{F}_p(x)}{\mathcal{F}_p(x)\partial x} - \mathcal{F}_p(\alpha)\frac{x\partial \mathcal{F}_p(y)}{\mathcal{F}_p(y)\partial x}\right) - xf_x\left(\mathcal{F}_p(\beta)\frac{y\partial \mathcal{F}_p(x)}{\mathcal{F}_p(x)\partial y} - \mathcal{F}_p(\alpha)\frac{y\partial \mathcal{F}_p(y)}{\mathcal{F}_p(y)\partial y}\right).$$

Furthermore, this object has nicer convergence properties, in the sense that it is supported modulo p^N in an easy to determine multiple of Γ ($(9pN + 3p)\Gamma$ to be precise). Therefore, we have a good control (in terms of χ_1 and χ_2) on the size of the objects we are computing with.

Complexity analysis. The complexity of this step is dominated by the computation of $O(g)$ expressions of the form $Z_x^i Z_y^j$, where $|i|$ and $|j|$ are $O(\delta)$. As before, this results in $\widetilde{O}(n^3 g^4 h(\chi_2 - \chi_1))$ time and $\widetilde{O}(n^3 g^2 h(\chi_2 - \chi_1))$ space.

STEP V: determine the action of Frobenius

For every $(i, j) \in 2\Gamma$, do the following two substeps.

SUBSTEP V.I: expand the Frobenius action on $x^i y^j$

In this step, one actually computes

$$\mathcal{G}_p(x^i y^j) = \Lambda^{-1}(\mathcal{F}_p^*(\Lambda(x^i y^j))).$$

Note that $\mathcal{F}_p^*(\Lambda(x^iy^j))$ is given by $\mathcal{F}_p(x^iy^j)\mathcal{F}_p^*(dx/xyf_y)$. To translate back, if

$$\mathcal{F}_p^*(\Lambda(x^iy^j)) = g_{ij,1}dx + g_{ij,2}dy$$

then

$$\Lambda^{-1}(\mathcal{F}_p^*(\Lambda(x^iy^j))) = xy(f_y g_{ij,1} - f_x g_{ij,2}).$$

Therefore, we output

$$\mathcal{F}_p(x^iy^j) \cdot E$$

where E is the expression that was precomputed during the foregoing step.

Complexity analysis. The complexity of the first substep can be estimated using a method similar to what we did in **STEP IV**, resulting in $\widetilde{O}(n^3g^3h(\chi_2 - \chi_1))$ time (per monomial) and $\widetilde{O}(n^3g^2h(\chi_2 - \chi_1))$ space.

SUBSTEP V.II: reduce modulo D

In this section we apply the method that was described in Subsection 6.1.5 to the output of the foregoing substep (after multiplying with p^ε) to obtain polynomials $r_{ij} \in \mathcal{L}^{(1)}(2D_C) \subset L^{(0)}_{2\Gamma}$. Note that we want our output r_{ij} to be supported in 2Γ: at this stage, we are no longer interested in the reduction to the basis $\mathcal{B} = \{\, x^iy^j \mid d_b \leq j < d_t\}$.

Complexity analysis. For the second substep, it suffices to analyze the complexity of Phase 1 and Step t+1, as described in Subsection 6.1.5. During Phase 1, one needs to solve systems of size $\sim h(2\kappa_2 + c)$. Therefore, it is optimal to choose $c = \kappa_2$. The number of systems to be solved is then bounded by $m/c = m/\kappa_2$. Using similar estimates for Phase 2 and using the analysis made in Section 5.1 (together with the bounds on the valuations of the non-zero invariant factors given in Lemma 6.7), this results in a use of

$$\widetilde{O}(h^2(\kappa_2 - \kappa_1)(\chi_2 - \chi_1)nN^2 + h^3(\kappa_2 - \kappa_1)^2(\chi_2 - \chi_1)nN)$$

time before proceeding to Step t+1. In this final step, one needs to solve a linear system of size $O(h\max\{\kappa_2 - \kappa_1, \chi_2 - \chi_1\})$, resulting in a time-cost of

$$\widetilde{O}(h^2\left(\max\{\kappa_2 - \kappa_1, \chi_2 - \chi_1\}\right)^2 nN + h^3\left(\max\{\kappa_2 - \kappa_1, \chi_2 - \chi_1\}\right)^3 n).$$

The extra space needed during Phase 1 and Step t+1 is

$$\widetilde{O}(h^2\left(\max\{\kappa_2 - \kappa_1, \chi_2 - \chi_1\}\right)^2 nN),$$

though this will in general be dominated by the space needed to store the polynomial h that is to be reduced, which is $\widetilde{O}(n^3g^2h(\chi_2 - \chi_1))$.

Overall complexity analysis. Since **SUBSTEP V.I** and **SUBSTEP V.II** have to be executed for $O(g)$ monomials, we obtain the following global estimates for **STEP V**: a time-cost of

$$\widetilde{O}(n^3g^3h^2\left(\max\{\kappa_2 - \kappa_1, \chi_2 - \chi_1\}\right)^2 + n^2g^2h^3\left(\max\{\kappa_2 - \kappa_1, \chi_2 - \chi_1\}\right)^3)$$

and a space-cost of $\widetilde{O}(n^3gh^2\left(\max\{\kappa_2-\kappa_1,\chi_2-\chi_1\}\right)^2)$.

Note that our time-estimate dominates the time needed to actually *compose* the systems that are to be solved.

STEP VI: compute a $\mathbb{Z}_q$-basis of $M_H = \frac{\mathcal{L}^{(0)}(2D_C)}{(D(\mathcal{L}(D_C)))^{(0)}}$

Note that from the proof of Lemma 4.4, we have that $\mathcal{L}^{(0)}(mD_C) = L^{(0)}_{m\Gamma}$ for any $m \in \mathbb{N}_0$. Therefore, we actually have to compute a $\mathbb{Z}_q$-basis of

$$\frac{L^{(0)}_{2\Gamma}}{(D(L_\Gamma)+fL_\Gamma)^{(0)}}.$$

Consider the module $D(L^{(0)}_\Gamma)+fL^{(0)}_\Gamma$ and express a vector $\mathcal{A}$ whose entries are the generators $\left\{D(x^iy^j), fx^iy^j\right\}_{(i,j)\in\Gamma\cap\mathbb{Z}^2}$ in terms of a vector $\mathcal{C}$ whose entries are $\{x^ry^s\}_{(r,s)\in 2\Gamma\cap\mathbb{Z}^2}$:

$$\mathcal{A} = E\cdot\mathcal{C}.$$

Now compute $\mathbb{Z}_q$-invertible matrices N_1 and N_2 (and their inverses) such that $N_1\cdot E\cdot N_2$ is a diagonal matrix. Its non-zero entries are the non-zero invariant factors of E and will be denoted by $d_1,\ldots,d_\ell$. If we write

$$N_1\cdot\mathcal{A} = N_1\cdot E\cdot N_2\cdot N_2^{-1}\cdot\mathcal{C},$$

we see that the entries of $N_2^{-1}\cdot\mathcal{C}$ form a basis $\{f_1,\ldots,f_k\}$ of $L^{(0)}_{2\Gamma}$ such that $\{d_1f_1,\ldots,d_\ell f_\ell\}$ is a basis of $D(L^{(0)}_\Gamma)+fL^{(0)}_\Gamma$. It is then easily seen that $\{f_1,\ldots,f_\ell\}$ is a basis of $(D(L_\Gamma)+fL_\Gamma)^{(0)}$. Finally, $\{f_{\ell+1},\ldots,f_k\}$ is a basis of M_H.

When computing modulo a finite precision, some caution is needed: to determine $f_{\ell+1},\ldots,f_k$ modulo p^N, it does not suffice to do the above computations modulo the same precision. During this step (and only during this step), we need to compute modulo p^{N+N_0}, where $N_0 = \lfloor \ell n\log_p(\ell whnp)\rfloor + 1 = O(N)$. Indeed, we claim that N_0 is a strict upper bound for the p-adic valuation of any non-zero $(\ell\times\ell)$-minor of E. As a consequence, the valuations of the non-zero invariant factors $d_1,\ldots,d_\ell$ are also strictly bounded by N_0. Therefore, we will be able to find invertible matrices $\widetilde{N}_1$ and $\widetilde{N}_2$ such that

$$\widetilde{N}_1\cdot E\cdot\widetilde{N}_2^{-1}$$

is congruent modulo p^{N+N_0} to the above diagonal matrix. The 'basis'

$$\{\widetilde{f}_{\ell+1},\ldots,\widetilde{f}_k\}$$

we find in this way corresponds modulo p^N to the basis mentioned above: if we would want to finalize the above diagonalization (which was only carried out modulo p^{N+N_0}), we would need to subtract from the $\widetilde{f}_i$ Laurent polynomials with coefficients divisible by $p^{(N+N_0)}/p^{N_0} = p^N$. Actually, one can check that

$\{\widetilde{f}_{\ell+1}, \ldots, \widetilde{f}_k\}$ is a basis itself, but we won't need this. If in **STEP VII** we write $f_{\ell+1}, \ldots, f_k$ and N_2, we actually mean the reductions mod p^N of $\widetilde{f}_{\ell+1}, \ldots, \widetilde{f}_k$ and $\widetilde{N}_2$ that were computed this way.

It remains to prove the claim, i.e. the p-adic valuation of any $(\ell \times \ell)$-minor of E is bounded by N_0. Let $r(X)$ be the polynomial from **STEP 0** and let $\theta \in \mathbb{C}$ be a root of it. Consider $K = \mathbb{Q}(\theta)$ and let $\mathcal{O}_K$ be its ring of algebraic integers. Then $\mathfrak{p} = (p) \subset \mathcal{O}_K$ is a prime ideal and the $\mathfrak{p}$-adic completion of K can be identified with $\mathbb{Q}_q$. Under this identification, E has entries

$$\sum_{i=0}^{n-1} a_i \theta^i \in \mathcal{O}_K$$

where the $a_i \in \mathbb{Z}$ satisfy $|a_i| \leq 2whp$. Since the complex norm of *any* root of $r(X)$ is bounded by p by Cauchy's bound, we conclude that the entries e of E satisfy

$$|e_{ij}|_K \leq nwhp^n \leq (whnp)^n$$

for *any* archimedean norm $|\cdot|_K$ on K that extends the classical absolute value on $\mathbb{Q}$. Since an $(\ell \times \ell)$-minor m is the sum of $\ell!$ ℓ-fold products of such entries, it follows that

$$|m|_K \leq (\ell whnp)^{\ell n}.$$

Since m is an algebraic integer, from the product formula we have

$$|m|_{\mathfrak{p}}^{-n} \leq \prod |m|_K \leq (\ell whnp)^{\ell n^2}$$

(if $m \neq 0$), where $|\cdot|_{\mathfrak{p}}$ is scaled such that $|p|_{\mathfrak{p}} = 1/p$ and where the product is over all archimedean norms $|\cdot|_K$ on K, to be counted twice if it comes from a non-real root of $r(X)$. From this we finally get that $\mathrm{ord}_p m \leq \ell n \log_p(\ell whnp)$.

Complexity analysis. This step needs $O(g^3)$ ring operations, each of which takes $\widetilde{O}(nN)$ time. Therefore, the time complexity of this step is $\widetilde{O}(n^2g^4)$ while the space complexity amounts to $\widetilde{O}(n^2g^3)$.

STEP VII: compute a matrix of p^{th} power Frobenius

From **STEP V**, we know that $p^\varepsilon x^i y^j$ is mapped to r_{ij}. Therefore, it is straightforward to compute the action of Frobenius on $f_{\ell+1}, \ldots, f_k$ and express it in terms of $\mathcal{C}$:

$$\Lambda^{-1} \mathcal{F}_p^* \Lambda \; p^\varepsilon \begin{pmatrix} f_{\ell+1} \\ \vdots \\ f_k \end{pmatrix} = F \cdot \mathcal{C}.$$

Since $F \cdot \mathcal{C} = F \cdot N_2 \cdot N_2^{-1} \cdot \mathcal{C}$, we obtain a matrix of Frobenius as $p^{-\varepsilon}$ times an appropriate submatrix M of $F \cdot N_2$.

Complexity analysis. The complexity of this step is dominated by the computation of $F \cdot N_2$, which takes $\widetilde{O}(n^2g^4)$ time and $\widetilde{O}(n^2g^3)$ space, and by $O(g^2)$ Frobenius substitutions, taking an extra $\widetilde{O}(g^2 \cdot n \cdot nN) = \widetilde{O}(n^3g^3)$ time.

STEP VIII: compute a matrix of q^{th} power Frobenius

The matrix $p^{-\varepsilon}M$ of the foregoing step is a matrix of $\mathcal{F}_p^*$, which is a $\mathbb{Q}_p$-vector space morphism acting on $H^1_{MW}(C \cap \mathbb{T}^2_{\mathbb{Q}_q})$. A matrix of $\mathcal{F}_q^*$ is then given by $p^{-n\varepsilon}\mathcal{M}_n$ where $\mathcal{M}_n = M^{\sigma^{n-1}} \cdot M^{\sigma^{n-2}} \cdot \ldots \cdot M^{\sigma} \cdot M$.

$\mathcal{M}_n$ can be computed using the following method that was presented by Kedlaya [60]: let $n = \mathfrak{n}_1\mathfrak{n}_2 \ldots \mathfrak{n}_k$ be the binary expansion of n and write $n' = \mathfrak{n}_1\mathfrak{n}_2 \cdots \mathfrak{n}_{k-1}$, then we have the formula

$$\mathcal{M}_n = \mathcal{M}_{n'}^{\sigma^{n'+\mathfrak{n}_k}} \cdot \mathcal{M}_{n'}^{\sigma^{\mathfrak{n}_k}} \cdot M^{\mathfrak{n}_k}$$

by means of which $\mathcal{M}_n$ can be computed recursively .

Complexity analysis. Applying some σ^i $(i \leq n)$ to a matrix of size $O(g)$ takes $\widetilde{O}(g^2 \cdot n \cdot nN) = \widetilde{O}(n^3g^3)$ time, if we precompute $[X]^{\sigma^i}$ as a root of the polynomial r that defines $\mathbb{Z}_q$, using Newton iteration and starting from the approximate solution $[X]^{p^i} \in \mathbb{F}_q$. The complexity of **STEP VIII** is then dominated by $O(\log n)$ matrix multiplications and $O(\log n)$ applications of some σ^i, resulting in $\widetilde{O}((n+g)n^2g^3)$ time. The space needed is $\widetilde{O}(n^2g^3)$.

STEP IX: output the characteristic polynomial of Frobenius

The characteristic polynomial $\widetilde{\chi}(t)$ of $\mathcal{M}_n$ can be computed using a careful implementation of the classical algorithm based on the reduction to the Hessenberg form, as it is explained in Section 5.2. Write

$$\widetilde{\chi}(t) = \sum_{i=0}^{2\mathrm{Vol}(\Gamma)+1} c_i t^i.$$

Then the characteristic polynomial of $\mathcal{F}_q^*$ (or of $p^{-n\varepsilon}\mathcal{M}_n$) is given by

$$\chi(t) = \sum_{i=0}^{2\mathrm{Vol}(\Gamma)+1} p^{(i-2\mathrm{Vol}(\Gamma)-1)n\varepsilon} c_i t^i \quad \in \mathbb{Z}[t].$$

This finalizes the description of the algorithm.

Complexity analysis. This needs $O(g^3 \cdot nN) = \widetilde{O}(n^2g^4)$ time and $\widetilde{O}(n^2g^3)$ space.

6.1.7 Conclusions

When we sum up, we obtain the following refinement of Theorem 1.13.

6.8 Theorem *There exists a deterministic algorithm to compute the zeta function of a bivariate Laurent polynomial $\overline{f} \in \mathbb{F}_{p^n}[\mathbb{Z}^2]$ that is nondegenerate with*

respect to its Newton polytope Γ*, given that the latter contains the origin and has a unique top and bottom vertex. Let* $g, h, \kappa_1, \kappa_2, \chi_1, \chi_2$ *be as above. Then for fixed* p*, it has running time*

$$\widetilde{O}(n^3 g^3 h^2 \left(\max\{\kappa_2 - \kappa_1, \chi_2 - \chi_1\}\right)^2 + n^2 g^2 h^3 \left(\max\{\kappa_2 - \kappa_1, \chi_2 - \chi_1\}\right)^3).$$

The space complexity amounts to

$$\widetilde{O}(n^3 g h^2 \left(\max\{\kappa_2 - \kappa_1, \chi_2 - \chi_1\}\right)^2).$$

The $\widetilde{O}$*-notation hides factors that are logarithmic in* n *and* g*. Furthermore, we assume that* p *is fixed. For 'most common' polytopes, the estimates* $h(\chi_2 - \chi_1) \approx h(\kappa_2 - \kappa_1) \approx g^{3/2}$ *hold, so that the algorithm needs* $\widetilde{O}(n^3 g^6 + n^2 g^{6.5})$ *time and* $\widetilde{O}(n^3 g^4)$ *space.*

Recall from Subsection 6.1.3 that the above conditions on Γ are not restrictive.

6.1.8 The zeta function of the complete model

As mentioned at the beginning of this chapter, the zeta function of $\overline{C} = V(\overline{f}) \cap \mathbb{T}^2_{\mathbb{F}_q}$ can be recovered as

$$Z_{\overline{C}}(t) = \frac{\frac{1}{q^{g+R-1}}\,\chi(qt)}{(1-qt)} = \frac{P'(t)}{(1-qt)}.$$

From this, one can easily compute the zeta function of the complete model $\widetilde{C} = V(\overline{f})$, by repeatedly trying to factor out $(1 - t^\kappa)$ from $P'(t)(1-t)$ for descending values of $\kappa = R-2, R-3, \ldots$. We refer to formula (4.2) to see why this works.

6.2 The commode case

Now, we briefly study the case where $\Gamma = \Gamma(\overline{f})$ satisfies the following conditions:

1. there are $a, b \in \mathbb{N}$ such that the line connecting $(0,0)$ and $(0,a)$ as well as the line connecting $(0,0)$ and $(b,0)$ are edges of Γ;
2. $(0,a)$ is a unique top vertex of Γ.

Below, we call the first condition being *commode* and the second condition being *monic*. Then the supporting cone of $A^\dagger_\Gamma$ is $\mathbb{R}^+_0$, and the set $\mathcal{B} = \{x^i y^j \mid (i,j) \in \mathbb{N}^2, j < a\}$ is an $\mathbb{F}_q$-basis for $\frac{\mathbb{F}_q[\mathbb{N}^2]}{(\overline{f})}$ and a $\mathbb{Z}_q$-basis for $\frac{\mathbb{Z}_q[\mathbb{N}^2]}{(f)}$, where f is a Newton polytope preserving lift of $\overline{f}$. Then, using the material in Chapter 4 (and especially the results in Section 4.5), we can do exactly the same as in the genus ≥ 1 case. But there are some simplifications concerning the parameters.

First, note that no optimization is necessary, since the triangle spanned by $(0,0)$, $(0,a)$ and $(\deg_x \overline{f},0)$ is contained in Γ. Therefore, $\mathrm{Vol}(\Gamma) \geq a \deg_x f/2$ and the estimate $wh \sim g$ is automatically fulfilled, where w and $h = a$ are the width resp. height of Γ and g is the number of interior lattice points.

Next, we can define χ such that $L_{m\Gamma} \subset S_{[0,m\chi]} = [0, m\chi] \times [0, a-1]$ for any $m \in \mathbb{N}_0$. As above, $\chi = O(g)$. Note that if $(b,0)$ is a unique right-most vertex of Γ, the estimate $a\chi = O(g^{3/2})$ can always be obtained by interchanging x and y if necessary.

Finally, let κ be such that $\mathcal{L}(2D_C + \mathrm{Div}_\infty(x)) \subset S_{[0,\kappa]} = [0,\kappa] \times [0, a-1]$. Then as above one has $\mathcal{L}(2D_C + m\mathrm{Div}_\infty(x)) \subset S_{[0,m+\kappa]}$ and $D(S_{[0,m]}) \subset S_{[0,m-1+\kappa]}$. Since $(1,0) \in \Gamma$, it suffices to take κ such that $\mathcal{L}(3D_C) \subset S_{[0,\kappa]}$. Hence we can take $\kappa = 3\chi$.

The algorithm itself then works completely analogous with the genus ≥ 1 case, leading to the following theorem.

6.9 Theorem *There exists a deterministic algorithm to compute the zeta function of a bivariate Laurent polynomial $\overline{f} \in \mathbb{F}_{p^n}[\mathbb{N}^2]$ that is nondegenerate with respect to its Newton polytope Γ, given that the latter is commode and monic. Let g, a, χ be as above. Then for fixed p, it has running time*

$$\widetilde{O}(n^3 g^3 a^2 \chi^2 + n^2 g^2 a^3 \chi^3).$$

The space complexity amounts to

$$\widetilde{O}(n^3 g a^2 \chi^2).$$

The $\widetilde{O}$-notation hides factors that are logarithmic in n and g. Furthermore, we assume that p is fixed. If Γ has a unique right-most vertex lying on the x-axis, the estimate $a\chi \approx g^{3/2}$ holds, so that the algorithm needs $\widetilde{O}(n^3 g^6 + n^2 g^{6.5})$ time and $\widetilde{O}(n^3 g^4)$ space.

Note that in the C_{ab} curve case, a better estimate for $a\chi = ab$ is g, yielding a time complexity of $\widetilde{O}(n^3 g^5)$ and a space complexity of $\widetilde{O}(n^3 g^3)$. This is the same as in the algorithm presented in [25].

As before, the zeta function of the complete model of $V(\overline{f}) \cap \mathbb{T}^2_{\mathbb{F}_q}$ can be easily derived from the zeta function of $V(\overline{f}) \cap \mathbb{T}^2_{\mathbb{F}_q}$ itself.

Toric compactification of $\mathbb{A}^2_{\mathbb{F}_q}$ instead of $\mathbb{T}^2_{\mathbb{F}_q}$

A first consequence of the assumption of commodeness is that $\mathbb{A}^2_{\mathbb{F}_q}$ is canonically embedded in $\mathbb{P}_{\mathbb{F}_q,\Gamma}$, the toric compactification of $\mathbb{T}^2_{\mathbb{F}_q}$ with respect to Γ. As such, we can consider $\mathbb{P}_{\mathbb{F}_q,\Gamma}$ as a compactification of the affine plane, instead of the torus. Therefore we can work with a notion of nondegeneracy that is slightly *weaker* than the one given in Chapter 2: it is no longer necessary to impose the nondegeneracy conditions with respect to the faces lying on the coordinate axes. However, we now should explicitly impose that $\overline{f}$ defines a nonsingular

curve in $\mathbb{A}^2_{\mathbb{F}_q}$. The main geometric difference with the old notion is that now we allow our curve to be tangent to the coordinate axes. Thus the conditions of Lemma 2.21 can be weakened: now any C_{ab} curve $\overline{C}$ that does not contain all $\mathbb{F}_q$-rational points in the plane, has a nondegenerate model in the new sense. Indeed, in that case it suffices to shift $\overline{C}$ such that it does not contain the origin. An example of a curve containing all rational points is the elliptic curve over $\mathbb{F}_2$ defined by

$$y^2 + y + x^3 + x = 0.$$

An important remark is that Lemma 2.22 and Corollary 2.23 still hold under this weaker condition. One way to see this is by adapting the proof of Theorem 3.11 to the above situation. Another way is as follows: by moving on to a field extension $\mathbb{F}_{q^r}$ if necessary, we can always find $x_0, y_0 \in \mathbb{F}_{q^r}$ such that $\overline{f}(x - x_0, y - y_0)$ is nondegenerate in the old sense. Note that this transformation does not affect the Newton polytope Γ since it is commode. As a consequence, we can find $\alpha, \beta, \gamma \in \mathbb{Q}_{q^r}[\mathbb{N}^2]$ that are supported in 2Γ, such that

$$\gamma f(x - x_0, y - y_0) + \alpha \frac{\partial f}{\partial x}(x - x_0, y - y_0) + \beta \frac{\partial f}{\partial y}(x - x_0, y - y_0) = 1.$$

Then by translating back and separating the $\mathbb{Q}_q$-part, we get the desired Nullstellensatz expansion.

In particular, if $f \in \mathbb{Z}_q[\mathbb{N}^2]$ is an arbitrary lift of $\overline{f}$ with the same Newton polytope Γ, then f is nondegenerate in our new sense. Let C denote the nonsingular curve obtained by taking the closure of the locus of f in $\mathbb{P}^2_{\mathbb{Q}_q,\Gamma}$. Then we will compute in $H^1_{DR}(f/\mathbb{Q}_q) = H^1_{DR}(C \cap \mathbb{A}^2_{\mathbb{Q}_q})$, instead of $H^1_{DR}(C \cap \mathbb{T}^2_{\mathbb{Q}_q})$. Note that the difference $C \cap (\mathbb{A}^2_{\mathbb{Q}_q} \setminus \mathbb{T}^2_{\mathbb{Q}_q})$ consists of $a + b$ nonsingular points, which by Theorem 2.25 implies that

$$\dim H^1_{DR}(C \cap \mathbb{A}^2_{\mathbb{Q}_q}) = \dim H^1_{MW}(\overline{C} \cap \mathbb{A}^2_{\mathbb{F}_q}) = 2\mathrm{Vol}(\Gamma) - a - b + 1\,.$$

Next, we have that Theorem 2.18 still holds, with the same definition for D_C (but note that it is now supported outside the x- and y-axis, because Γ is commode). Again this can be seen by first moving on to a finite field extension, then translating $\overline{f}$ such that the old notion of nondegeneracy is fulfilled, then using Theorem 2.18 and finally using that $\mathbb{Q}_q$ is perfect to return to the base field. The main difference with the general case is that Theorem 2.20 needs to be reformulated as follows.

6.10 Theorem *Let $D : \mathbb{Q}_q[\mathbb{N}^2] \to \mathbb{Q}_q[\mathbb{N}^2]$ be as before, i.e. the operator $xy\left(\frac{\partial f}{\partial y}\frac{\partial}{\partial x} - \frac{\partial f}{\partial x}\frac{\partial}{\partial y}\right)$. Then we have a natural map*

$$\frac{L^-_{2\Gamma}}{fL^-_\Gamma + D(L_\Gamma)} \to H^1_{DR}(C \cap \mathbb{A}^2_{\mathbb{F}_q})$$

which is in fact an isomorphism. Here L^-_S denotes the part of L_S that is supported in $S \cap \mathbb{N}^2_0$.

PROOF. The proof works exactly as in Theorem 2.20; the definition of Λ remains the same but now we have to restrict to $\mathcal{L}(-\mathrm{Div}_0(x) - \mathrm{Div}_0(y))$ to obtain a well-defined map from $A = \mathbb{Q}_q[\mathbb{N}^2]/(f)$ to $D^1(A)$. ■

All of this can be turned into an algorithm to compute the characteristic polynomial of Frobenius acting on $H^1_{MW}(V(\overline{f}) \cap \mathbb{A}^2_{\mathbb{F}_q})$ in exactly the same way as above. This results in the same asymptotics as in Theorem 6.9.

6.3 Overall conclusion

In this chapter, we presented a generalization of Kedlaya's algorithm to compute the zeta function of a nondegenerate curve over a finite field of small characteristic. As the condition of nondegeneracy is generic, the algorithm works for curves that are defined by a randomly chosen bivariate Laurent polynomial with given Newton polytope Γ. It requires $\widetilde{O}(n^3\Psi_t)$ amount of time and $\widetilde{O}(n^3\Psi_s)$ amount of space, where Ψ_t, Ψ_s are functions that depend on Γ only. For non-exotic choices of Γ, we have that $\Psi_t \sim g^{6.5}$ and $\Psi_s \sim g^4$, where g is the number of interior lattice points of Γ (which is precisely the geometric genus of the curve). In the case of a C_{ab} curve, we obtain the estimates $\Psi_t \sim g^5$ and $\Psi_s \sim g^3$, so that the algorithm works (at least asymptotically) as fast as the one presented in [25]. At this moment, the algorithm has not yet been implemented.

In order to develop the algorithm, we proved a number of theoretical results on nondegenerate curves that are interesting in their own right, for instance a linear effective Nullstellensatz for sparse Laurent polynomials in any number of variables. Also, we adapted the Frobenius lifting technique used in [25] to prove a convergence rate in which the Newton polytope Γ plays a very natural role. Furthermore, we gave a sparse description of the first Monsky-Washnitzer cohomology group and the action of Frobenius on it.

Appendix A

Point counting for the non-mathematician

The aim of this appendix is to explain to the non-mathematician what point counting is about. If the term 'finite field' does not ring a bell, one is strongly encouraged to read this section: it might (unfortunately) be the only accessible part of the thesis. This is not a matter of mathematics being extremely difficult, nor of mathematicians doing unnecessarily complicated. But abstract mathematics is a kind of science that is very much written in its own language, used to denote concepts and structures that one does not encounter in daily life. It takes some time to learn this language, and makes it hard to communicate.

Finite fields

In abstract mathematics, the notion of a *field* is very important. Roughly spoken, a field is a structure in which one can perform additions, subtractions, multiplications and divisions in a reasonable way.

For example, the set of *natural numbers* $\{0, 1, 2, 3, 4, \dots\}$ is *not* a field, since it is impossible to compute for instance $5 - 8$ or $3/4$ in this structure.

In the set of *integers* $\{0, 1, -1, 2, -2, 3, -3, \dots\}$ this problem is partly solved, but still it is impossible to compute $3/4$.

The set of *fractions* or *rational numbers* such as 0, -1, $\frac{1}{2}$ and $-\frac{11}{5}$ is the first actual *field* that you have encountered during your education. Fractions can be added, subtracted, multiplied and divided without any problem. Of course, it is impossible to divide by 0, but that is too much to ask for.

Another field that you are definitely familiar with is the set of *real numbers* such as 0, -0.5, $3.14159...$ and -1.3. Some of you may even remember what *complex numbers* are, like $-2 + 3\mathbf{i}$ and $1.2 - 8\mathbf{i}$: that is another example of a field.

All examples of structures given above have *infinitely* many elements. But there are also many *finite* structures in which one can compute. The most

common one is certainly the analog clock, consisting of twelve elements

$$\{0, 1, 2, 3, \ldots, 11\}.$$

How much is $8 + 7$ in such a clock?

Or in other words: if it is 8 o'clock, what time is it 7 hours later? The answer is 3. In the same manner one can wonder what $3 - 5$ should be: if it is 3 o'clock, what time was it 5 hours ago? The answer is 10. Also multiplication is possible: $5 \times 7 = 11$ (just compute $5 \times 7 = 35$ in the usual way and subtract 12 until one finds a number between 0 and 11).

Unfortunately, division is problematic in such a clock. For instance, what is $1/2$? Suppose this were some x in $\{0, 1, 2, 3, \ldots, 11\}$. Then we would have

$$\begin{array}{rcll} 1/2 & = & x & \\ 1 & = & 2 \times x & \text{(multiply both sides by 2)} \\ 6 & = & 0 \times x & \text{(multiply both sides by 6)} \\ 6 & = & 0, & \end{array}$$

but this is of course not true. Therefore, it is impossible to compute $1/2$ in a clock. The main reason for this is the presence of so-called *zero divisors*. For instance $2 \times 6 = 0$, while 2 nor 6 are equal to 0. Such strange things cannot happen in the set of rational numbers or in the set of real numbers.

The above problem stems from the fact that 12 is not a *prime number*. Recall that a prime number is a natural number different from 1 that is only divisible by 1 and itself. Indeed, $12 = 2 \times 6$ and that is the reason why $2 \times 6 = 0$ in an analog clock.

If our clock would have 13 indications, instead of 12, the problems mentioned above would be solved. For instance, $1/2$ then equals 7, as one can check that $2 \times 7 = 1$.

Similarly, all other divisions (except division by 0, of course) can be carried out properly. It takes some work to actually *prove* this, but we omit it here. In any case, a clock with 13 indications is our first example of a *finite field* and should be kept in mind throughout the rest of this appendix. We denote it with $\mathbb{F}_{13}$. Similarly, for any prime number p, we have a finite field $\mathbb{F}_p$.

But there are more finite fields. In fact, for every power of a prime number, such as 13^2, 7^9 or 2^{10000}, there is a finite field with that number of elements. They are at least as important as the 'prime-clocks' described above, but unfortunately their structure is a bit more complicated. To get an idea of the flavour, we will describe what for instance $\mathbb{F}_{3^2}$ looks like, but the reader who has had enough of finite fields already can immediately skip to the next paragraph. An element of $\mathbb{F}_{3^2}$ looks like

$$[aX + b],$$

where a and b are elements of $\mathbb{F}_3 = \{0, 1, 2\}$ and X is just a fixed symbol. Adding and subtracting is easy: simply add or subtract the a's and the b's separately. For instance, $[2X + 1] + [0X + 2] = [2X + 0]$, or $[0X + 2] + [1X + 2] = [1X + 1]$. Multiplication is done as if the $[aX + b]$'s were polynomials, but in the end every appearance of X^2 should be replaced by 2:

$$\begin{aligned} [2X + 1][1X + 2] &= [(2 \times 1)X^2 + (2 \times 2)X + (1 \times 1)X + (1 \times 2)] \\ &= [2X^2 + 2X + 2] \\ &= [(2 \times 2) + 2X + 2] = [1 + 2X + 2] = [2X + 0]. \end{aligned}$$

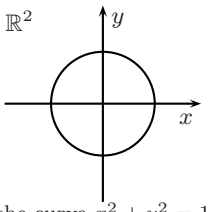
Again one can prove that division works properly in such a structure.

Curves over finite fields

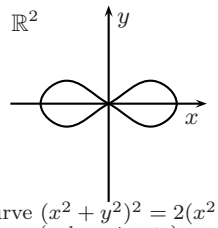
For a moment, let us return to the infinite case, in particular to the field of real numbers $\mathbb{R}$. Then we can take a look at an algebraic[1] equation in two variables x and y, such as

$$x^2 + y^2 = 1 \qquad \text{or} \qquad (x^2 + y^2)^2 = 2(x^2 - y^2).$$

If we plot all tuples (x, y) satisfying this equation, we obtain a *curve* in the real plane $\mathbb{R}^2$. For our examples we get the following.



the curve $x^2 + y^2 = 1$
(a *circle*)



the curve $(x^2 + y^2)^2 = 2(x^2 - y^2)$
(a *lemniscate*)

There is much to say about real curves: this is a very classical subject with still a lot of ongoing research. But from the viewpoint of this thesis, we only mention that almost all curves that are obtained in this way contain *infinitely* many points: there are infinitely many points on a circle, there are infinitely many points on a lemniscate, and so on.

[1] By 'algebraic' we mean that only additions, multiplications, ... should be involved. Expressions of the type $x^2 + \sin(xy) + e^y = 1$ are not taken into account, since these make no sense in abstract fields.

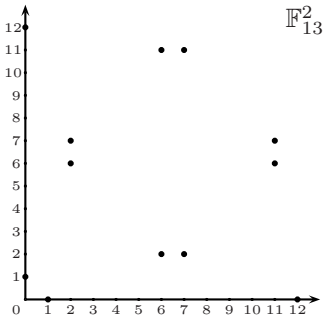
Is there a way of drawing curves over a finite field, say for instance $\mathbb{F}_{13}$ (a clock with 13 indications)? In principle, nothing stops us from doing the same as in $\mathbb{R}$. That is, we look at an equation

$$x^2 + y^2 = 1$$

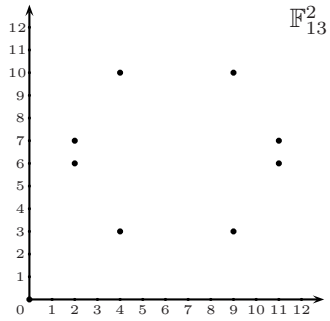
and then plot all tuples (x, y) (which now consist of elements of $\mathbb{F}_{13}$) that satisfy this equation. For instance, $(6, 2)$ is such a tuple since

$$6 \times 6 + 2 \times 2 = 1$$

in a clock with 13 indications (verify this!). Continuing in this way, we obtain graphs of the following type.



the 'circle' $x^2 + y^2 = 1$



the 'lemniscate' $(x^2 + y^2)^2 = 2(x^2 - y^2)$

These do not look very much like 'curves', yet we still call them curves because of the analogy with the real numbers. This analogy goes much further than just both being the set of solutions to some algebraic equation. Although this lies beyond the scope of this introductory text, we mention that it is for this deeper analogy that we really need that $\mathbb{F}_{13}$ is a *field*. Over a standard clock with 12 indications for instance, similar pictures could be made but much of the analogy would be spoiled.

An immediate feature of curves over finite fields is that they only contain a *finite* number of points. Of course, there are only finitely many points in the whole plane! In our examples, we find 12 points lying on the 'circle' and 9 points lying on the 'lemniscate'.

This thesis

We are now ready to have a second look at the title of this thesis: *'Point counting on nondegenerate curves'*. In fact, a better (but longer) title would be *'Point counting on nondegenerate plane curves over finite fields'*. If we forget about the word 'nondegenerate', all of this can be understood: we investigate methods that, given an algebraic equation in two variables x and y over a finite field, output the number of points on the corresponding curve. A naive method would be to explicitly write down all these points, but if the field size gets big (we mean really big, that is fields à la $\mathbb{F}_{10000000000000000000000067}$) this soon

becomes intractable, even for a computer. Therefore, one has to come up with smarter ideas. In 2001, the American mathematician Kedlaya presented such an idea, based on deep theoretical results that were obtained in the early 1970's. He worked this out for a special class of curves (so-called *hyperelliptic curves*); the resulting method works particularly fast for fields of the type $\mathbb{F}_{p^n}$, where p is small and n is big[2].

The original purpose of our research was to generalize Kedlaya's algorithm to work for arbitrary curves, but this turned out to be harder than expected. Nevertheless we managed to treat a very large class of curves. This is where the term 'nondegenerate' pops in, which is a weak technical condition that is almost always satisfied.

What's the use of this?

Asking for the use of their research makes many mathematicians feel a bit uncomfortable. Often, there are no *direct* applications and this might give the impression that they are just playing around in their own fantasy world.

Let us spend a few words on this. Mathematics is a kind of science that balances between philosophy and engineering, and this is reflected in the variety of answers one can get to the above question. The philosopher would say that mathematics is simply worth studying in its own right, and even asking for the use of it indicates some lack of respect for the queen of sciences. The practically oriented engineer would reply that it is unjustifiable that society pays for research from which no direct feedback is guaranteed. As always in this kind of stories, the truth lies in the middle. But since in today's world the philosopher needs more support than the engineer, we remark that this debate is not new. For instance, in the beginning of the 20^{th} century there lived this great engineer, Thomas Edison, being widely admired for his practical inventions including the light bulb and several telegraphic devices. At about the same time, scientists like Albert Einstein and Max Planck developed relativity theory and quantum physics, their only interest being to understand how nature works. They were sometimes criticized for this, but several decades later their better understanding resulted in important practical inventions such as the GPS system, the laser and the transistor, and especially quantum physics has become indispensable for modern electronics. Anyway, at least one conclusion of this endless discussion should be that society *does* benefit from research whose primary objective is to find out what the (mathematical or physical) world looks like, but perhaps only very indirectly and with some delay...

Back to the subject of this thesis: point counting. This takes a somewhat special position in the above perspective. First of all, it is about 'fast computation', which sounds very engineery. Indeed, in modern computer science, finite fields have become very popular structures to compute in. The main applications can be found in error-correction (e.g. used on CD's and DVD's) and in

[2]So unfortunately, the field $\mathbb{F}_{10000000000000000000000067}$ that was mentioned above *cannot* be dealt with using Kedlaya's method. In fact, up to our knowledge nobody knows how to deal with fields of this type (except for a restricted class of curves).

cryptography, the science dealing with secret messages and digital signatures. An important example of such an application is the so-called *elliptic curve cryptosystem*. This is a cryptographic method that makes use of curves defined by an equation of the form

$$y^2 + Axy + By = x^3 + Cx^2 + Dx + E$$

(over some finite field). How this precisely works is explained in Section 1.3, we skip it here. We only mention that not *all* equations of the above type give rise to a safe cryptographic method: there is a restriction on the number of points lying on the corresponding curve. So this is where the subject of this thesis comes into play: to know wether or not a given equation is suitable for cryptographic purposes, one must be able to count points. The faster, the better.

Now to be honest, the cryptographic relevance of this thesis should not be overestimated. For the above class of curves, efficient point counting algorithms were known already and work faster than our method, which is designed to treat a more general class of curves (at the price of being a bit slower). It is not very likely that these more general curves will be used in cryptography in the near future. But one never knows, and soon or late other practical uses of point counting will probably pop up: after all, it is a very natural problem.

The main use of our research lies at the philosopher's side of the story. There we have the more conceptual question of what the number of points on a curve is actually determined by. For any curve over any finite field, one can blindly compute this number, but what does it say? Is it just a random number? Or is there more structure behind it? This is a very old and natural question, dating back to the work of Gauss and Jacobi at the beginning of the 19^{th} century, and it turns out to be much harder and much more fascinating than one would expect at first sight. Some of recent history's cleverest mathematicians have spent years of their life to tackle this problem, and especially since the work of Weil half-way the 20^{th} century, some great progress has been made. Nevertheless, many important questions remain unanswered. Of course, these questions have proven to be very difficult and it will need a smarter mathematician than I am to solve them. But it is very likely that this mathematician will use a computer to check some hypotheses, to find certain patterns, to provide heuristic evidence, or simply to speed up his/her work. The more curves that can be treated, the better. The faster the point counting methods that are used, the more curves that can be treated. It is my hope that our results can serve in this.

Appendix B

Nederlandse samenvatting

Een belangrijk probleem in de computationele getaltheorie is het volgende: ontwerp een efficiënt algoritme dat bij invoer van een algebraïsche variëteit $\overline{X}$ over een eindig veld $\mathbb{F}_q$, het aantal $\mathbb{F}_q$-rationale punten op $\overline{X}$ als uitvoer geeft. Hierbij veronderstellen we dat $\overline{X}$ gegeven wordt door een aantal concrete vergelijkingen. Het is duidelijk dat de naïeve methode, die alle $\mathbb{F}_q$-rationale punten van de omgevende affiene of projectieve ruimte overloopt en kijkt of ze al dan niet op $\overline{X}$ liggen, *niet* efficiënt is. Als $\mathbb{F}_q$ groot wordt is dit onbegonnen werk, zelfs voor de snelste computers.

Tot op heden is dit probleem nog altijd verre van opgelost. De moeilijkheid is dat het aantal rationale punten van een variëteit over een eindig veld een grillig en schijnbaar lukraak gekozen getal is, dat veel van zijn wiskundige geheimen nog niet heeft prijsgegeven. Ons belangrijkste theoretisch inzicht hebben we te danken aan Weil [107], die halfweg de vorige eeuw vaststelde dat het aantal rationale punten bepaald wordt door de eigenwaarden van het Frobeniusendomorfisme

$$\mathcal{F} : \overline{X} \to \overline{X} : (x_1, \ldots, x_n) \mapsto (x_1^q, \ldots, x_n^q) \quad (\text{waarbij } q = \#\mathbb{F}_q),$$

wanneer werkend op nader te bepalen cohomologieruimten geassocieerd aan $\overline{X}$. Zo'n cohomologietheorieën werden in de jaren '60 en '70 daadwerkelijk gevonden. Eerst ontwikkelde Grothendieck de zogenaamde ℓ-adische cohomologie [81], die coëfficiënten aanneemt in het veld van de ℓ-adische getallen $\mathbb{Q}_\ell$, waarbij ℓ een priemgetal is verschillend van de karakteristiek p van $\mathbb{F}_q$. Wat later kwamen ook p-adische versies op de proppen, aanvankelijk eerder fragmentarisch maar nadien in één theoretisch jasje gestoken door Berthelot.

Op zich zijn de eigenwaarden van Frobenius natuurlijk even grillig van aard als het aantal rationale punten zelf, maar in sommige gevallen zijn ze wel snel te berekenen. Dit werd impliciet gebruikt door Schoof [97] (1985, ℓ-adisch) en Satoh [95] (2000, p-adisch) die elk een efficiënt algoritme beschreven om het aantal rationale punten te bepalen op een elliptische kromme in Weierstrassvorm. De rekentijd die (geoptimaliseerde versies van) deze algoritmes nodig hebben

kan worden afgeschat door

$$\widetilde{O}\left((\log q)^{4+\mu}\right) \quad \text{resp.} \quad \widetilde{O}\left(p(\log q)^2\right),$$

waarbij $\mu \leq 1$ heuristisch gewoon 0 is. Hierbij is $\widetilde{O}$ de Soft-Oh die termen die logaritmisch zijn in de invoergrootte verwaarloost. De geheugenkost van beide algoritmes is $\widetilde{O}\left((\log q)^2\right)$. Merk op dat Satoh's algoritme veel sneller is voor kleine waarden van p, maar erg traag wordt over velden van grote karakteristiek: dat is kenmerkend voor alle algoritmes die gebruik maken van p-adische technieken. We komen hier verder nog op terug.

Het eerste algoritme dat expliciet gebruik maakt van Weilcohomologie werd beschreven door Kedlaya [60] (2001, p-adisch), wiens methode het aantal rationale punten op een gegeven hyperelliptische kromme in Weierstrassvorm berekent[1]. De tijd die het algoritme nodig heeft[2] is $\widetilde{O}(pg^4 (\log q)^3)$ en de geheugenkost bedraagt $O(g^3 (\log q)^3)$, waarbij g het geslacht is van de ingevoerde kromme. Opnieuw merken we op dat de berekening traag wordt voor velden van grote karakteristiek. Maar voor kleine karakteristiek klopt Kedlaya's algoritme zelfs dat van Schoof, hoewel het veel algemener werkt. Bovendien heeft het algoritme een goede tijdsafhankelijkheid van g. Tenslotte is het theoretische kader van Kedlaya's methode erg flexibel, wat de weg opent naar algoritmes die een veel grotere klasse van variëteiten kunnen behandelen.

Vrijwel onmiddellijk buitten Gaudry en Gürel dit uit om het geval van superelliptische krommen te behandelen [40]. Later slaagden Denef en Vercauteren erin om dit op hun beurt te veralgemenen naar de klasse van C_{ab}-krommen [25]. Hun algoritme heeft een tijdskost van $\widetilde{O}(g^5 (\log q)^3)$ en een geheugenkost van $\widetilde{O}(g^3 (\log q)^3)$, waarbij p vast beschouwd wordt. In deze thesis gaan we een grote stap verder en presenteren we een algoritme dat het aantal rationale punten bepaalt op zogenaamde *niet-gedegenereerde krommen*, een heel algemene klasse die bijna alle vlakke krommen omvat. De resultaten werden bekomen in samenwerking met Jan Denef en Frederik Vercauteren en zullen worden gepubliceerd in 'International Mathematics Research Notices' [13].

Niet-gedegenereerde krommen

Zij $\mathbb{F}$ een veld en zij $f \in \mathbb{F}[\mathbb{Z}^2] = \mathbb{F}[x^{\pm 1}, y^{\pm 1}]$ een bivariate Laurentveelterm. De *drager* van f is de deelverzameling $\mathcal{D}$ van $\mathbb{Z}^2$ zodat

$$f = \sum_{i,j \in \mathcal{D}} f_{ij} x^i y^j$$

met $f_{i,j} \in \mathbb{F} \setminus \{0\}$. De kleinste convexe veelhoek in $\mathbb{R}^2$ die $\mathcal{D}$ omvat noemen we de Newtonpolytoop $\Gamma(f)$ van f. We zeggen dat f *niet-gedegenereerd* is *ten*

[1] Dit was aanvankelijk enkel over velden van oneven karakteristiek, het geval $p = 2$ werd later behandeld door Denef en Vercauteren [24].

[2] In de praktijk is de afhankelijkheid $\widetilde{O}(pg^3(\log q)^3)$, de factor g^4 duikt enkel op in sommige gevallen in karakteristiek 2.

opzichte van zijn Newtonpolytoop als de stelsels

$$\mathcal{S}_\gamma : f_\gamma = \frac{\partial f_\gamma}{\partial x} = \frac{\partial f_\gamma}{\partial y} = 0 \qquad \text{met } f_\gamma = \sum_{(i,j)\in\gamma} f_{ij} x^i y^j$$

voor γ eender welk hoekpunt of eender welke zijde van $\Gamma(f)$, of $\Gamma(f)$ zelf, geen oplossing hebben in $(\mathbb{F} \setminus \{0\})^2$. We noemen een kromme $C(f)$ in $\left(\mathbb{A}^1_{\mathbb{F}} \setminus \{0\}\right)^2$ kortweg *niet-gedegenereerd* als ze gedefinieerd wordt door een Laurentveelterm f die niet-gedegenereerd is ten opzichte van zijn Newtonpolytoop, op voorwaarde dat die Newtonpolytoop tweedimensionaal is.

We kunnen bewijzen dat het niet-gedegenereerd zijn van een Laurentveelterm ten opzichte van een gegeven Newtonpolytoop een Zariski-open conditie is, die gedefinieerd is over het priemveld van $\mathbb{F}$. Losjes uitgedrukt betekent dit dat 'bijna alle' Laurentveeltermen niet-gedegenereerd zijn ten opzichte van hun Newtonpolytoop. Voor eindige velden neemt dit de volgende concrete vorm aan: zij $P_{\Gamma,n}$ de kans dat een willekeurig gekozen $\overline{f} \in \mathbb{F}_{p^n}[\mathbb{Z}^2]$ met Γ als Newtonpolytoop niet-gedegenereerd is, dan is

$$\lim_{n\to\infty} P_{\Gamma,n} = 1.$$

De meetkundige betekenis van het niet-gedegenereerd zijn is dat het complete, niet-singuliere model van $C(f)$ op een natuurlijke manier kan ingebed worden in het torische oppervlak geassocieerd aan de Newtonpolytoop van f. Bijzonder interessant aan niet-gedegenereerde krommen is dat veel meetkundige eigenschappen een combinatorische interpretatie hebben. Het meetkundige geslacht van $C(f)$ is gelijk aan het aantal inwendige $\mathbb{Z}^2$-punten van $\Gamma(f)$, het aantal plaatsen dat aan $C(f)$ moet toegevoegd worden om het complete, niet-singuliere model te bekomen is gelijk aan het aantal $\mathbb{Z}^2$-punten op de rand van $\Gamma(f)$, de Euler-Poincarékarakteristiek van $C(f)$ is gelijk aan $-2\mathrm{Vol}(\Gamma(f))$, ...

De volgende eigenschap is cruciaal voor onze doeleinden. Zij $\mathbb{F}_q$ een eindig veld en zij $\mathbb{Z}_q$ een discrete-valuatiering met residuveld $\mathbb{F}_q$. Zij $\overline{f} \in \mathbb{F}_q[\mathbb{Z}^2]$ niet-gedegenereerd ten opzichte van zijn Newtonpolytoop Γ. Zij $f \in \mathbb{Z}_q[\mathbb{Z}^2]$ eender welke Laurentveelterm met Newtonpolytoop Γ die reduceert tot $\overline{f}$. Dan is f automatisch niet-gedegenereerd ten opzichte van zijn Newtonpolytoop (wanneer beschouwd over het breukenveld $\mathbb{Q}_q$ van $\mathbb{Z}_q$). Dit zorgt voor een diep meetkundig verband tussen $C(\overline{f})$ en $C(f)$, dat gedirigeerd wordt door de Newtonpolytoop. Inderdaad, beide krommen hebben hetzelfde geslacht, hetzelfde aantal 'gaten', dezelfde Euler-Poincarékarakteristiek. Maar de connectie gaat veel verder en leidt tot een *effectief* isomorfisme tussen de p-adische cohomologie van $\overline{f}$ en de de-Rhamcohomologie van f, wat een rechtstreekse veralgemening is van Kedlaya's lemma [60, Lemma's 2 en 3] voor hyperelliptische krommen en een essentieel ingrediënt van ons algoritme.

p-adische cohomologie

Het algoritme van Kedlaya is gebaseerd op de theorie van Monsky en Washnitzer, die een expliciete beschrijving geven van de p-adische cohomologie van

affiene variëteiten, voor ons geval als algebraïsche de-Rhamcohomologie van de ring van *overconvergente functies* op $C(f)$. Dit zijn in essentie machtreeksen

$$\sum_{(i,j)\in\mathbb{Z}^2} a_{ij}x^i y^j \qquad a_{ij}\in\mathbb{Q}_q$$

(modulo f) waarvoor $-\log|a_{ij}|_p$ minstens lineair groeit met $|i|+|j|$. Deze vormen een ring, de zogenaamde *dagger ring* van $C(\overline{f})$, en de algebraïsche de-Rhamcohomologie van deze ring (met coëfficiënten in $\mathbb{Q}_q$) is per definitie de Monsky-Washnitzercohomologie van $C(\overline{f})$. Men kan aantonen dat deze niet afhangt van de keuze van f en dat het Frobeniusmorfisme op een natuurlijke manier werkt op deze cohomologie [105]. Het aantal punten op $C(\overline{f})$ kan dan bekomen worden uit de karakteristieke veelterm van deze actie van Frobenius op de eerste cohomologieruimte $H^1_{MW}(\overline{f}/\mathbb{Q}_q)$.

Met een *effectief* isomorfisme

$$H^1_{DR}(f/\mathbb{Q}_q)\to H^1_{MW}(\overline{f}/\mathbb{Q}_q)$$

bedoelen we het volgende. Laten we vanaf nu veronderstellen dat $(0,0)\in\Gamma$. Dit kan altijd bekomen worden door $\overline{f}$ te vermenigvuldigen met een gepaste Laurentmonoom. Dan kunnen we op het complete model van $C(f)$ een divisor $D_C\geq 0$ definiëren, die combinatorische informatie over Γ bevat en waarvoor het volgende geldt: voor elke $m\in\mathbb{N}_0$ wordt de Riemann-Rochruimte $\mathcal{L}^{(0)}(mD_C)$ (de $^{(0)}$ wijst erop dat we ons beperken tot de $\mathbb{Z}_q$-module van functies die gedefinieerd zijn over $\mathbb{Z}_q$) precies beschreven door de verzameling Laurentveeltermen met coëfficiënten in $\mathbb{Z}_q$ en met drager in $m\Gamma$ (Stelling 4.4). Om deze expliciete beschrijving optimaal te kunnen uitbuiten, vertalen we de cohomologieruimten – die normaliter met behulp van *differentiaalvormen* uitgedrukt worden – naar ruimten van *functies*, via de bijectie

$$\Lambda:\mathbb{Q}_q(C(f))\to\Omega_{C(f)}(\mathbb{Q}_q):h\mapsto\frac{h}{xy\frac{\partial f}{\partial y}}dx.$$

Men kan nagaan dat exacte differentiaalvormen onder deze bijectie overeenkomen met het beeld onder $D=xy\left(\frac{\partial f}{\partial y}\frac{\partial}{\partial x}-\frac{\partial f}{\partial x}\frac{\partial}{\partial y}\right)$. Bijgevolg kan $H^1_{DR}(f/\mathbb{Q}_q)$ via Λ bekeken worden als ruimte van functies modulo D. Analoog kan $H^1_{MW}(\overline{f}/\mathbb{Q}_q)$ via Λ bekeken worden als ruimte van overconvergente functies modulo D.

Een belangrijk resultaat uit Hoofdstuk 4 is dan (zie Stelling 4.2):

Stelling *Voor elke $m\in\mathbb{N}_{\geq 2}$ is*

$$\mathcal{L}^{(0)}((m-1)D_C)\xrightarrow{D}\frac{\mathcal{L}^{(1)}(mD_C)}{\mathcal{L}^{(1)}(2D_C)}$$

een goed gedefinieerde, surjectieve afbeelding.

Hierbij is $\mathcal{L}^{(1)}(mD_C)$ de deelmodule van $\mathcal{L}^{(0)}(mD_C)$ bekomen door een bepaalde voorwaarde op de p-adische valuaties van de coëfficiënten te leggen (zie Definitie 4.1). Wat belangrijk is, is dat de voorwaarde voldaan is indien de valuaties van de coëfficiënten groter zijn dan

$$\log_p(mh + hw),$$

waarbij h respectievelijk w de hoogte respectievelijk de breedte van Γ voorstelt (zie **STEP I** van Subsectie 6.1.6).

Ten eerste zegt bovenstaande stelling dat *alle* functies met drager in een zekere $m\Gamma$ (voor m groot genoeg) gereduceerd kunnen worden modulo D naar functies met drager in 2Γ. Ten tweede zegt ze dat de p-adische valuaties van de noemers die tijdens dit reductieproces geïntroduceerd worden slechts logaritmisch kunnen stijgen met m, waardoor het reductieproces ook werkt voor *overconvergente* functies (waarvan de p-adische valuaties van de coëfficiënten minstens lineair stijgen).

Omdat we kunnen aantonen dat alle functies modulo D in $m\Gamma$ te passen zijn (voor voldoende grote m), geeft dit ons een isomorfisme tussen $H^1_{DR}(f/\mathbb{Q}_q)$ en $H^1_{MW}(\overline{f}/\mathbb{Q}_q)$: beide zijn isomorf met $\mathcal{L}(2D_C)$ mod D. Het expliciete karakter van bovenstaande afbeelding geeft ons bovendien een manier om functies modulo D te reduceren (zie verderop) en het verlies aan p-adische precisie goed af te schatten.

Actie van Frobenius

Om het aantal punten te berekenen met behulp van p-adische cohomologie, volstaat het de actie van Frobenius op $H^1_{MW}(\overline{f}/\mathbb{Q}_q)$ te kennen modulo een zekere precisie p^N. Dan kunnen we met behulp van de Weilconjectuur (Stelling 1.8) de zetafunctie (en dus het aantal oplossingen) exact terugvinden.

Cruciaal aan ons algoritme (en aan alle p-adische algoritmes) is dat de actie van de q^{de}-macht-Frobenius kan worden opgesplitst in $\log_p q$ toepassingen van de p^{de}-macht-Frobenius. Dit reduceert de tijdscomplexiteit aanzienlijk, en is de reden waarom ons algoritme net als alle andere p-adische methoden *niet* efficiënt is voor velden van grote karakteristiek.

Via Newtoniteratie kunnen we deze actie snel berekenen, met behulp van een methode die voor het eerst gepresenteerd werd in [25]. Dankzij een krachtige effectieve Nullstellensatz (zie Hoofdstuk 3) speelt de Newtonpolytoop Γ een erg natuurlijke rol in het convergentiegedrag van deze actie. Concreet is het beeld van eender welke functie uit $\mathcal{L}(2D_C)$ modulo p^N gedragen in

$$(9pN + 5p)\Gamma$$

(zie weer **STEP I** uit Subsectie 6.1.6) en worden hierbij geen noemers geïntroduceerd.

Het algoritme

Het algoritme bestaat er dus in om de karakteristieke veelterm van de q^{de}-macht-Frobenius te berekenen modulo p^N voor N groot genoeg (we tonen aan dat $N = \widetilde{O}(g(\log q))$ volstaat). Dit doen we door de p^{de}-macht-Frobenius te laten werken op $x^i y^j$ voor elke $(i,j) \in 2\Gamma$, en dan het resultaat modulo D terug te reduceren naar 2Γ. Achteraf zoeken we dan een $\mathbb{Z}_q$-modulebasis voor $\mathcal{L}(2D_C)$ modulo D om een matrix van de p^{de}-macht-Frobenius te bekomen, waaruit snel een matrix van de q^{de}-macht-Frobenius berekend kan worden.

Dit reduceren modulo D doen we stapsgewijs. We vertrekken van een functie g in $\mathcal{L}((9pN+5p)D_C)$, en we zoeken een h zodat $g - D(h) \in \mathcal{L}(rD_C)$ voor een op voorhand gekozen $r < 9pN + 5p$. Zo gaat het verder tot we in $\mathcal{L}(2D_C)$ uitkomen. Als we g op voorhand vermenigvuldigen met een voldoende grote macht van p, dan blijven we in de ruimtes $\mathcal{L}^{(1)}(rD_C)$ en weten we dat er omwille van Stelling 4.2 geen noemers hoeven te worden geïntroduceerd tijdens het reductieproces.

De functie h zoeken we met de methode van de onbepaalde coëfficiënten, dus door een stelsel op te lossen. Maar een probleem hierbij is dat we modulo f werken, waardoor het onmogelijk is om coëfficiënten te vergelijken. Daarom reduceren we eerst alles naar een $\mathbb{Z}_q$-basis voor $\frac{\mathbb{Z}_q[x^{\pm 1}, y^{\pm 1}]}{(f)}$. Als Γ een uniek hoogste hoekpunt (c_t, d_t) en een uniek laagste hoekpunt (c_b, d_b) heeft, dan is

$$\mathcal{B} = \left\{ x^i y^j \mid d_b \leq j < d_t \right\}$$

een natuurlijk keuze voor zo'n basis. De reductie naar deze basis zorgt er dan voor dat we de functie h nu wel kunnen vinden door het oplossen van een stelsel. Ze leidt bovendien tot een compactere representatie van de objecten waarmee we werken. Vanuit theoretisch standpunt is deze reductie echter niet zo elegant, omdat we de natuurlijke band met de Newtonpolytoop wat verliezen. Ze zorgt ervoor dat we de tijds- en ruimtecomplexiteit van ons algoritme moeten afschatten in termen van een aantal parameters die niet *intrinsiek* zijn, i.e. die duidelijk afhangen van de positie en de oriëntatie van Γ. Desalniettemin zijn al deze parameters op hun beurt (polynomiaal) af te schatten in termen van het geslacht g. Zie Subsectie 6.1.4.

Voor het oplossen van de stelsels gebruiken we een nieuwe methode die uitbuit dat de p-adische precisie modulo dewelke we rekenen groot is in vergelijking met de dimensies van de stelsels. Dit wordt beschreven in Sectie 5.1. De getallen r worden bekomen door de volgende afweging te maken: hoe kleiner r, hoe minder stelsels er moeten worden opgelost, maar hoe groter de stelsels zijn die optreden.

Uiteindelijk bekomen we het volgende resultaat.

Stelling *Er bestaat een deterministisch algoritme dat de zetafunctie van een niet-gedegenereerde kromme van geslacht g over $\mathbb{F}_q$ in $\widetilde{O}((\log q)^3 \Psi_t)$ tijd kan uitrekenen en waarbij $\widetilde{O}((\log q)^3 \Psi_s)$ geheugen gebruikt wordt. Hierbij is $p = \text{char}(\mathbb{F}_q)$ vast en zijn Ψ_t en Ψ_s parameters die enkel van de Newtonpolytoop van*

de invoerkromme afhangen. Voor 'de meeste' Newtonpolytopen is $\Psi_t = \widetilde{O}(g^{6.5})$ *en* $\Psi_s = \widetilde{O}(g^4)$.

Met 'de meeste' bedoelen we dat de Newtonpolytoop niet al te exotisch geschapen mag zijn, maar voorts is het niet de bedoeling dit begrip wiskundig exact te maken. In het geval van een C_{ab}-kromme gelden de betere afschattingen $\Psi_t = \widetilde{O}(g^5)$ en $\Psi_s = \widetilde{O}(g^3)$, waardoor ons algoritme dezelfde complexiteit heeft als dat van Denef en Vercauteren [25].

Bibliography

[1] Timothy Abbott, Kiran Kedlaya and David Roe, *Bounding Picard numbers of surfaces using p-adic cohomology*, to appear in proc. of "Arithmetic, geometry and coding theory (AGCT-10)", conference held at CIRM in 2005

[2] Leonard Adleman and Ming-Deh Huang, *Counting points on curves and abelian varieties over finite fields*, J. Symb. Comp. **32** (3), pp. 171-189 (2001)

[3] Matthias Asschenbrenner, *Ideal membership in polynomial rings over the integers*, J. Am. Math. Soc., electronically published, 34 pp., available at `http://arxiv.org/math.AC/0305172` (2004)

[4] Michael Atiyah and Ian Macdonald, *Introduction to Commutative Algebra*, Addison-Wesley, Reading, Mass. (1969)

[5] Henry Baker, *Examples of applications of Newton's polygon to the theory of singular points of algebraic functions*, Trans. Cambridge Phil. Soc. **15**, pp. 403-450 (1893)

[6] Victor Batyrev, *Variations of the mixed Hodge structure of affine hypersurfaces in algebraic tori*, Duke Math. J. **69** (2), pp. 349-409 (1993)

[7] Victor Batyrev and David Cox, *On the Hodge Structure of Projective Hypersurfaces in Toric Varieties*, Duke Math. J. **75** (2), pp. 293-338 (1994)

[8] Peter Beelen and Ruud Pellikaan, *The Newton polygon of plane curves with many rational points*, Des. Codes Cryptogr. **21**, pp. 41-67 (2000)

[9] Pierre Berthelot, *Finitude et pureté cohomologique en cohomologie rigide*, with an appendix in English by Aise Johan de Jong, Inv. Math. **128** (2), pp. 329-377 (1997)

[10] Frits Beukers, *Lattice Reduction*, in: Arjeh Cohen, Hans Cuypers and Hans Sterk (eds.), *Some Tapas of Computer Algebra*, Algorithms and Computation in Math. **4**, Springer (1999)

[11] Joe Buhler and Neal Koblitz, *Lattice Basis Reduction, Jacobi Sums and Hyperelliptic Cryptosystems*, Bull. Australian Math. Soc. **58**, pp. 147-154 (1998)

[12] Wouter Castryck, *A shortened classical proof of the quadratic reciprocity law*, preprint, 2 pp., available at `http://wis.kuleuven.be/algebra/artikels/reciprocity.pdf`

[13] Wouter Castryck, Jan Denef and Frederik Vercauteren, *Computing Zeta Functions of Nondegenerate Curves*, to appear in Int. Math. Research Notices, available at `http://arxiv.org/math.NT/0607308`, 41 pp.

[14] Henri Cohen, *A course in computational algebraic number theory*, GTM Springer-Verlag **138** (1991)

[15] Henri Cohen and Gerhard Frey (eds.), Roberto Avanzi, Christophe Doche, Tanja Lange, Kim Nguyen and Frederik Vercauteren, *Handbook of Elliptic an Hyperelliptic Curve Cryptography*, Discrete Math. Appl., Chapman & Hall (2006)

[16] Thomas Cormen, Charles Leierson, Ronald Rivest, *Introduction to algorithms*, MIT Press Cambr. (1990)

[17] Jean-Marc Couveignes, *Quelques calculs en théories des nombres*, thèse, Univ. de Bordeaux I (1994)

[18] David Cox, *Toric varieties and toric resolutions*, Resolution of singularities (Obergurgl, 1997), Progr. Math. **181**, Birkhuser, Basel, pp. 259-284 (2000)

[19] Vladimir Danilov, *The geometry of toric varieties*, Russian Math. Surveys **33** (2), pp. 97-154 (1978)

[20] Aise Johan de Jong, *Smoothness, semi-stability and alterations*, IHES Sci. Publ. Math. **83**, pp. 51-93 (1996)

[21] Pierre Deligne, *La conjecture de Weil. I.*, IHES Sci. Publ. Math. **43**, pp. 273-307 (1974)

[22] Jan Denef and François Loeser, *On some rational generating series occuring in arithmetic geometry*, Geometric Aspects of Dwork Theory **1**, de Gruyter, pp. 509-526 (2004)

[23] Jan Denef and Frederik Vercauteren, *An extension of Kedlaya's algorithm to Artin-Schreier curves in characteristic 2*, proceedings of ANTS-V, Springer LNCS **2369**, pp. 308-323 (2002)

[24] Jan Denef and Frederik Vercauteren, *An extension of Kedlaya's algorithm to hyperelliptic curves in characteristic 2*, J. Cryptology **19** (1), pp. 1-25 (2006)

[25] Jan Denef and Frederik Vercauteren, *Computing Zeta functions of C_{ab} curves using Monsky-Washnitzer cohomology*, to appear in Finite Fields and Their Applications, Elsevier (2005)

[26] Max Deuring, *Die Typen der Multiplikatorenringe elliptischer Funktionenkörper*, Abh. Math. Sem. Hansischen Univ. **14**, pp. 197-272 (1941)

[27] Whitfield Diffie and Martin Hellman, *New directions in cryptography*, IEEE Trans. Information Theory, IT-**22** (6), pp. 644-654 (1976)

[28] Vladimir Drinfel'd and Serge Vlăduţ, *The number of points on an algebraic curve*, Funktsional. Anal. i Prilozhen **17**, pp. 68-69 [Funct. Anal. Appl. **17** pp. 53-54] (1983)

[29] Bernard Dwork, *On the rationality of the Zeta function of an algebraic variety*, Amer. J. Math. **82**, pp. 631-648 (1960)

[30] Bernard Dwork, *A deformation theory for the zeta function of a hypersurface*, Proc. ICM Stockholm '62, pp. 247-259 (1962)

[31] Bas Edixhoven, *Point counting after Kedlaya*, notes of a EIDMA-Stieltjes Graduate course given in Leiden, 23 pp., available at `http://www.math.leidenuniv.nl/~edix/oww/mathofcrypt/carls_edixhoven/kedlaya.pdf` (2003)

[32] Eugène Ehrhart, *Sur un problème de géométrie diophantienne linéaire. I. Polyèdres et réseaux,* J. Reine Angew. Math. **226**, pp. 1-29 (1967)

[33] Noam Elkies, *Elliptic and modular curves over finite fields and related computational issues*, Computational perspectives on number theory, vol. 7 of AMS/IP Stud. Adv. Math., pp. 21-76, Am. Math. Soc. (1998)

[34] Noam Elkies, Everett Howe, Andrew Kresch, Bjorn Poonen, Joseph Wetherell and Michael Zieve, *Curves of every genus with many points, II: Asymptotically good families,* Duke Math. J. **122** (2), pp. 399-422 (2004)

[35] Noaï Fitchas and André Galligo, *Nullstellensatz effectif et conjecture de Serre (théorème de Quillen-Suslin) pour le calcul formel* (French), Math. Nachr. **149**, pp. 231-253 (1990)

[36] Mireille Fouquet, Pierrick Gaudry and Robert Harley, *An extension of Satoh's algorithm and its implementation*, J. Ramanujan Math. Soc. **15** (4), pp. 281-318 (2000)

[37] Eisaku Furukawa, Mitsuru Kawazoe and Tetsuya Takahashi, *Counting points for hyperelliptic curves of type $y^2 = x^5 + ax$ over finite prime fields*, Lecture Notes in Comp. Sci. **3006** (Selected Areas in Cryptography), pp. 26-41 (2004)

[38] Steven Galbraith and James McKee, *The probability that the number of points on an elliptic curve over a finite field is prime*, Journal of the London Math. Soc. **62** (3), pp. 671-684 (2000)

[39] Pierrick Gaudry, *Counting points on genus 2 curves over finite fields*, slides of a talk given at Durham, available at `http://www.lix.polytechnique.fr/Labo/Pierrick.Gaudry/papers.en.html` (2000)

[40] Pierrick Gaudry and Nicolas Gürel, *An extension of Kedlaya's point-counting algorithm to superelliptic curves*, proceedings of ASIACRYPT 2001, Springer LNCS **2248**, pp. 480-494 (2001)

[41] Pierrick Gaudry and Robert Harley, *Counting points on hyperelliptic curves over finite fields*, proceedings of ANTS-IV, Springer LNCS **1838**, pp. 313-332 (2000)

[42] Pierrick Gaudry and Éric Schost, *Construction of Secure Random Curves of Genus 2 over Prime Fields*, proceedings of EUROCRYPT 2004, Springer LNCS **3027**, pp. 239-256 (2004)

[43] Jürgen Gerhard and Joachim von zur Gathen, *Modern Computer Algebra*, Cambridge University Press, New York (1999)

[44] Ralf Gerkmann, *The p-adic Cohomology of Varieties over Finite Fields and Applications on the Computation of Zeta functions*, Ph. D. thesis, Univ. of Essen (2003)

[45] Ralf Gerkmann, *Relative rigid cohomology and point counting on families of elliptic curves*, preprint, available at `http://joguinf.informatik.uni-mainz.de/~gerkmann/legendre.pdf`, 25 pp.

[46] Valery Goppa, *Geometry and codes* (translated from Russian), Mathematics and its Applications (Soviet Series) **24**, Kluwer Ac. Publ. Group, Dordrecht (1988)

[47] Marvin Greenberg and John Harper, *Algebraic Topology: A First Course*, Mathematics Lecture Note Series **58**, Benjamin/Cummings Publishing Co. (1981)

[48] Branko Grünbaum and Geoffrey Shephard, *Pick's Theorem*, Amer. Math. Monthly **100**, pp. 150-161 (1993)

[49] Robert Harley, *Asymptotically optimal p-adic point-counting*, e-mail to NMBRTHRY list (December 2002)

[50] Robin Hartshorne, *Algebraic Geometry*, GTM Springer-Verlag **52** (1977)

[51] Grete Hermann, *Die Frage der endlich vielen Schritte in der Theorie der Polynomideale* (German), Math. Ann. **95** (1), pp. 736-788 (1926)

[52] Ming-Deh Huang and Doug Ierardi, *Counting points on curves over finite fields*, J. Symb. Comp. **25** (1), pp. 1-21 (1998)

[53] Hendrik Hubrechts, *Memory efficient hyperelliptic curve point counting*, preprint, available at `http://arxiv.org/math.NT/0609032`, 10 pp.

[54] Richard Hudson and Kenneth Williams, *Binomial coefficients and Jacobi sums*, Trans. Am. Math. Soc. **281** (2), pp. 431-505 (1984)

[55] Yasutaka Ihara, *Some remarks on the number of rational points of algebraic curves over finite fields*, J. Fac. Sci. Univ. Tokyo **28** (1981), pp. 721-724

[56] Kenneth Ireland and Michael Rosen, *A classical introduction to modern number theory*, Second edition, GTM **84**, Springer-Verlag, New York (1990)

[57] Erich Kaltofen and Gilles Villard, *On the complexity of computing determinants*, Computational Complexity **13**, pp. 91-130 (2004)

[58] Mikhail Kapranov, *The elliptic curve in the S-duality theory and Eisenstein series for Kac-Moody groups*, preprint, available at `http://arxiv.org/math.AG/0001005`, 41 pp.

[59] Nicholas Katz and Peter Sarnak, *Random Matrices, Frobenius Eigenvalues, and Monodromy*, Am. Math. Soc. Colloquium Publications **45** (1999)

[60] Kiran Kedlaya, *Counting points on hyperelliptic curves using Monsky-Washnitzer cohomology*, J. Ramunajan Math. Soc. **16** (4), pp. 323-338 (2001)

[61] Kiran Kedlaya, *Computing zeta functions via p-adic cohomology*, proceedings of ANTS-VI, Springer LNCS **3076**, pp. 1-17 (2004)

[62] Kiran Kedlaya, *Computing zeta functions of nondegenerate toric hypersurfaces*, draft

[63] Kiran Kedlaya, *Fourier transforms and p-adic "Weil II"*, to appear in Comp. Math., available at `http://arxiv.org/math.NT/0210149`, 57 pp.

[64] George Kempf, Finn Knudsen, David Mumford and Bernard Saint-Donat, *Toroidal embeddings. I.*, Lecture Notes in Math. **339**, Springer-Verlag, Berlin-New York, viii+209 pp. (1973)

[65] Askold Khovanskii, *Newton polyhedra, and the genus of complete intersections*, Funktsional. Anal. i Prilozhen. **12**, pp. 51-61 (1978), english translation in Funct. Anal. Appl. **12**, pp. 38-46 (1978)

[66] Neal Koblitz, *Primality of the number of points on an elliptic curve over a finite field*, Pacific J. Math. **131** (1), pp. 157-165

[67] János Kollár, *Sharp effective Nullstellensatz*, J. Amer. Math. Soc. **1** (4), pp. 963-975 (1988)

[68] Teresa Krick, Luis Pardo and Martín Sombra, *Sharp estimates for the arithmetic Nullstellensatz*, Duke Math. J. **109**, pp. 521-598 (2001)

[69] Anatoli Kushnirenko, *Newton polytopes and the Bezout theorem*, Funct. Anal. Appl. **10** (3), pp. 233-235 (1976)

[70] Alan Lauder, *Deformation theory and the computation of zeta functions*, Proc. London Math. Soc. **88** (3), pp. 565-602 (2004)

[71] Alan Lauder, *Counting solutions to equations in many variables over finite fields*, Found. of Comp. Math. **4** (3), pp. 221-267 (2004)

[72] Alan Lauder and Daqin Wan, *Counting points on varieties over finite fields of small characteristic*, to appear in "Algorithmic Number Theory: Lattices, Number Fields, Curves and Cryptography" (Mathematical Sciences Research Institute Publications), Cambr. Univ. Press

[73] Reynald Lercier, *Algorithmique des courbes elliptiques dans les corps finis*, Ph.D. thesis, Lab. d'Informatique de l'École polytechnique (LIX) (1997)

[74] Reynald Lercier and David Lubicz, *A quasi quadratic time algorithm for hyperelliptic curve point counting*, to appear in Ramanujan J., available at `http://medicis.polytechnique.fr/~lercier/file/LL05.pdf`, 19 pp.

[75] Juri Manin, *The Hasse-Witt matrix of an algebraic curve*, Amer. Math. Soc. Transl. Ser. **45**, pp. 245-264 (1965)

[76] Ryutaroh Matsumoto, *The C_{ab} curve*, note available at `http://www.rmatsumoto.org/cab.pdf`

[77] Alfred Menezes, Paul van Oorschot and Scott Vanstone, *Handbook of Applied Cryptography*, 5th printing, CRC Press (2001)

[78] Jean-François Mestre, *Algorithmes pour compter des points de courbes en petite charactéristique et en petit genre*, notes of a talk given at Rennes, available at `http://www.math.jussieu.fr/~mestre/` (2000)

[79] James Milne, *Abelian Varieties*, course notes, available at `http://www.jmilne.org`

[80] James Milne, *Algebraic Number Theory*, course notes, available at `http://www.jmilne.org`

[81] James Milne, *Etale Cohomology*, Princeton Math. Ser. **33**, Princeton Univ. Press (1980)

[82] James Milne, *Lectures on Etale Cohomology*, course notes, available at `http://www.jmilne.org`

[83] Shinji Miura, *Algebraic geometric codes on certain plane curves*, Trans. IEICE J75-A **11** (Japanese), pp. 1735-1745 (1992)

[84] Paul Monsky and Gerard Washnitzer, *Formal cohomology I*, Ann. Math. **88** (2), pp. 181-217 (1968)

[85] Paul Monsky, *Formal cohomology II: The cohomology sequence of a pair*, Ann. Math. **88** (2), pp. 218-238 (1968)

[86] Paul Monsky, *Formal cohomology III: Fixed point theorems*, Ann. Math. **93** (2), pp. 315-343 (1971)

[87] Volker Müller, *Ein Algorithmus zur Bestimmung der Punktanzahl elliptischer Kurven über endlichen Körpern der Charakteristik größer drei*, Ph.D. thesis, Tech. Fak. de Univ. des Saarlandes (1995)

[88] David Mumford, *Abelian varieties*, Tata Inst. of Fund. Res. Studies in Math. **5**, Oxford University Press (1970)

[89] Tadao Oda, *Convex bodies and algebraic geometry*, Springer-Verlag (1988)

[90] Andrew Odlyzko, *The 10^{20}th zero of the Riemann zeta function and 70 million of its neighbours*, ATT Bell Lab. preprint (1989)

[91] Patrice Philippon, *Dénominateurs dans le théorème des zéros de Hilbert*, Acta Arith. **58**, pp. 125 (1990)

[92] Jonathan Pila, *Frobenius maps of abelian varieties and finding roots of unity in finite fields*, Math. Comp. **55** (192), pp. 745-763 (1990)

[93] Bjorn Poonen, *Computational aspects of curves of genus at least 2*, in Algorithmic Number Theory, Springer LNCS **1122**, pp. 283-306 (1996)

[94] Christophe Ritzenthaler, *Point counting on genus 3 non hyperelliptic curves*, proceedings of ANTS-VI, Springer LNCS **3076**, pp. 379-394 (2004)

[95] Takakazu Satoh, *The canonical lift of an ordinary elliptic curve over a finite field and its point counting,* J. Ramanujan Math. Soc. **15** (4), pp. 247-270 (2000)

[96] Arnold Schönhage and Volker Strassen, *Schnelle Multiplikation grosser Zahlen*, Computing (Arch. Elektron. Rechnen) **7**, pp. 281-292 (1971)

[97] René Schoof, *Elliptic curves over finite fields and the computation of square roots mod p*, Math. Comp. **44** (170), pp. 483-494 (1985)

[98] René Schoof, *Counting points on elliptic curves over finite fields*, Les 18imes Journes Arithmtiques, Bordeaux 1993, J. Thorie de Nombres Bordeaux **7**, pp. 219-254 (1995)

[99] Paul Scott, *On convex lattice polygons*, Bull. Austr. Math. Soc. **15** (3), pp. 395-399 (1976)

[100] Joseph Silverman, *The arithmetic of elliptic curves*, GTM **106**, Springer-Verlag, New York (1986)

[101] Martín Sombra, *A sparse effective Nullstellensatz*, Adv. Appl. Math. **22**, pp. 271-295 (1999)

[102] Henning Stichtenoth, *Die Hasse-Witt-Invariante eines Kongruenzfunktionenkörpers*, Arch. Math. (Basel) **33** (4), pp. 357-360 (1979/80)

[103] Joe Suzuki, *Generalizing Kedlaya's order counting based on Miura theory*, preprint, available at `http://citeseer.ist.psu.edu/657428.html`, 14 pp.

[104] Nobuo Tsuzuki, *Bessel F-isocrystals and an algorithm for computing Kloosterman sums*, preprint

[105] Marius van der Put, *The cohomology of Monsky and Washnitzer*, in: Daniel Barsky et Philippe Robba (eds.), *Introductions aux cohomologies p-adiques (Luminy, 1984)*, Mém. Soc. France **23** (4), pp. 33-59 (1986)

[106] Frederik Vercauteren, *Computing zeta functions of curves over finite fields,* Ph. D. Thesis, K. U. Leuven (2003)

[107] André Weil, *Numbers of solutions of equations in finite fields*, Bull. Am. Math. Soc. **55**, pp. 497-508 (1949)