Algorithmic Number Theory
MSRI Publications
Volume **44**, 2008

Algorithmic theory of zeta functions over finite fields

DAQING WAN

ABSTRACT. We give an introductory account of the general algorithmic theory of the zeta function of an algebraic set defined over a finite field.

CONTENTS

1. Introduction

Let $\mathbb{F}_q$ be a finite field of q elements and p its characteristic. Let X be an algebraic set defined over $\mathbb{F}_q$. For each positive integer k, let N_k denote the number of $\mathbb{F}_{q^k}$-rational points on X. The zeta function $Z(X)$ of X is the generating function

$$Z(X) = Z(X,T) = \exp\left(\sum_{k=1}^{\infty} \frac{N_k}{k} T^k\right).$$

The author is partially supported by the NSF and the NSFC.

The zeta function contains important arithmetic and geometric information concerning X. It has been studied extensively in connection with the celebrated Weil conjectures [1949].

Both practical applications and theoretical investigations make a good understanding of the zeta function from an algorithmic point of view increasingly important. The aim of this paper is to present a brief introductory account of the various fundamental problems and results in the emerging algorithmic theory of zeta functions. We shall focus on general properties rather than on results that are restricted to special cases. In particular, in most of this paper we do not assume X to be smooth and projective, although in that case one can often say more.

The contents are organized as follows. In Section 2 we review general properties of zeta functions from an algorithmic point of view. A naive effective algorithm for computing the zeta function is given. If the characteristic p is small, one can use Dwork's p-adic method to obtain a polynomial time algorithm for computing the zeta function in the case that the numbers of variables and defining equations for X are fixed.

In Section 3, we show that the general case of algebraic sets can be reduced in various ways to the case that X is a hypersurface. A more detailed discussion of that crucial case is given in Section 4, with emphasis on the smooth projective case. In Section 5 we consider the complex pure weight decomposition. Using the LLL factorization algorithm and Deligne's main theorem, we show that, when the zeta function is given, one can compute in polynomial time how many zeros and poles with a given complex absolute value it has. In Section 6, which is devoted to the p-adic pure slope decomposition, we use the theory of Newton polygons to obtain a similar result for the number of zeros and poles with a given p-adic absolute value.

We conclude the paper by giving, in Section 7, an algorithm for the simpler problem of computing the zeta function modulo p. This algorithm shares several characteristic features with the general p-adic method for computing the full zeta function that is presented in [Lauder and Wan 2008] in this volume. Section 7 may thus serve as an introduction to that article.

All algorithms in this paper are deterministic. Probabilistic algorithms will not be discussed. Time is measured in bit operations.

2. Generalities on computing zeta functions

Let X be an algebraic set defined over a finite field $\mathbb{F}_q$ of q elements of characteristic p. For computational purposes, we may assume that X is affine, i.e., that it is the subset of affine n-space $\mathbb{A}^n$ defined by a system of polynomial equations:

$$\begin{cases} f_1(x_1, \ldots, x_n) = 0, \\ \quad \vdots \\ f_m(x_1, \ldots, x_n) = 0, \end{cases}$$

where $f_i \in \mathbb{F}_q[x_1, \ldots, x_n]$. Let

$$X(\mathbb{F}_q) = \{x = (x_1, \ldots, x_n) \in \mathbb{F}_q^n \mid f_1(x) = \ldots = f_m(x) = 0\}$$

be the finite set of $\mathbb{F}_q$-rational points on X. It is clear that the cardinality $\#X(\mathbb{F}_q)$ is effectively computable.

For algorithmic purposes, "giving" $\mathbb{F}_q$ means specifying p as well as an irreducible polynomial h in one variable over $\mathbb{F}_p$ that defines $\mathbb{F}_q$, so that q equals $p^{\deg h}$; elements of $\mathbb{F}_q$ are then represented as polynomials of degree less than $\deg h$ in a formal zero of h, with coefficients from $\mathbb{F}_p$. Giving X means specifying a system of m defining polynomials f_i in n variables with coefficients in $\mathbb{F}_q$. Let d be the maximum of the total degrees of the polynomials f_i. Then the dense input size for X is $O(m\binom{d+n}{n}\log q)$, which is $O(m(d+1)^n \log q)$. Our first fundamental problem is the following.

PROBLEM 2.1. *Given $\mathbb{F}_q$ and X, compute the number $\#X(\mathbb{F}_q)$ in time polynomial in the dense input size $O(m(d+1)^n \log q)$.*

This problem is trivial if q is fixed, so we may assume that q is large. In theory, the problem of counting $X(\mathbb{F}_q)$ can be reduced to the zero-dimensional case. Namely, let Y be the zero-dimensional algebraic set defined by

$$\{f_1 = \ldots = f_m = 0,\ x_1^q - x_1 = \ldots = x_n^q - x_n = 0\}.$$

Then it is clear that

$$\#X(\mathbb{F}_q) = \#Y(\mathbb{F}_q).$$

Following a suggestion of Eisenbud and Sturmfels, one may now compute a Gröbner basis for Y, and its cardinality equals $\#Y(\mathbb{F}_q)$. However, as q gets large, the cases where the Gröbner basis computation can be done efficiently are likely to become increasingly exceptional. (See [Eisenbud 1995] for Gröbner bases.)

Let $\bar{\mathbb{F}}_q$ denote a fixed algebraic closure of $\mathbb{F}_q$. For each positive integer k, let $\mathbb{F}_{q^k}$ denote the unique subfield of $\bar{\mathbb{F}}_q$ with q^k elements. Let $\#X(\mathbb{F}_{q^k})$ denote the number of $\mathbb{F}_{q^k}$-rational points on X. The following problem is harder but more interesting than Problem 2.1.

PROBLEM 2.2. *Given $\mathbb{F}_q$ and X, compute the sequence of numbers $\#X(\mathbb{F}_{q^k})$ $(k = 1, 2, \ldots)$.*

It may not be clear how a finite algorithm can compute an infinite sequence of numbers, but this will be clarified below. As we shall see, one can encode the entire sequence in a suitably defined generating function, which turns out to be a rational function. This so-called *zeta function* of X is finite in nature and can thus be written down in a finite amount of time. Actually doing this for given X is the content of Problem 2.2.

A *geometric point* of X is an $\bar{\mathbb{F}}_q$-rational point of X. From the equality

$$\bigcup_{k=1}^{\infty} X(\mathbb{F}_{q^k}) = X(\bar{\mathbb{F}}_q)$$

we see that each geometric point of X will be counted somewhere in the sequence of numbers $\#X(\mathbb{F}_{q^k})$. This may explain why many of the subtle geometric invariants associated with an algebraic variety X can be read off from its zeta function, in addition to a wealth of arithmetic information.

DEFINITION 2.3. The zeta function of X is the generating function

$$Z(X) = Z(X, T) = \exp\biggl(\sum_{k=1}^{\infty} \frac{T^k}{k} \#X(\mathbb{F}_{q^k})\biggr).$$

The q-th power Frobenius map Frob_q is the permutation of the set $X(\bar{\mathbb{F}}_q)$ of geometric points of X defined by

$$\mathrm{Frob}_q\colon\ x = (x_1, \dots, x_n) \mapsto x^q = (x_1^q, \dots, x_n^q).$$

The *degree* of a geometric point x is defined to be the smallest positive integer d such that

$$\mathrm{Frob}_q^d(x) = x$$

or, equivalently, such that $x \in X(\mathbb{F}_{q^d})$. A *closed point* over $\mathbb{F}_q$ is the orbit of a geometric point under Frob_q. All geometric points belonging to a given closed point have the same degree, and this common degree is called the *degree* of the closed point. We denote by $|X|$ the set of closed points of X over $\mathbb{F}_q$, and, for each positive integer k, by $M_k(X)$ the number of closed points of X of degree k. Since each closed point of degree k consists of exactly k points in $X(\mathbb{F}_{q^k})$, one deduces

$$\#X(\mathbb{F}_{q^k}) = \sum_{d|k} dM_d(X).$$

Considering the logarithmic derivative of the zeta function, one finds the *Euler product* expansion

$$Z(X) = \prod_{k=1}^{\infty} \frac{1}{(1-T^k)^{M_k(X)}} = \prod_{x \in |X|} \frac{1}{1-T^{\deg(x)}} \in 1 + T\mathbb{Z}[\![T]\!].$$

As the Weil conjectures [1949] predict, the zeta function is a rational function. The first proof, given by Dwork [1960], used p-adic analysis. The second proof, given by Grothendieck, used the theory of ℓ-adic cohomology, where ℓ is a prime number different from p. These two proofs pioneered the general p-adic and ℓ-adic study of zeta functions over finite fields.

THEOREM 2.4. *The zeta function $Z(X)$ is a rational function, i.e., it belongs to $\mathbb{Q}(T)$. If we write*

$$Z(X,T)=\frac{R_1(X,T)}{R_2(X,T)},\quad (R_1,R_2)=1,\quad R_i\in 1+T\mathbb{Q}[T],$$

then we have $R_i \in 1+T\mathbb{Z}[T]$.

The rationality of $Z(X)$ has an interesting consequence for the numbers $\#X(\mathbb{F}_{q^k})$, as follows. Let β_i and γ_j denote the reciprocal zeros of $R_1(X)$ and $R_2(X)$, respectively, so that $R_1(X)=\prod_i(1-\beta_i T)$ and $R_2(X)=\prod_j(1-\gamma_j T)$. Then one finds

$$\sum_{k=1}^{\infty}\#X(\mathbb{F}_{q^k})T^k = T\frac{d\log Z(X,T)}{dT}=\sum_j\frac{\gamma_j T}{1-\gamma_j T}-\sum_i\frac{\beta_i T}{1-\beta_i T}.$$

This implies

$$\#X(\mathbb{F}_{q^k})=\sum_j\gamma_j^k-\sum_i\beta_i^k\quad\text{for all } k\ge 1.$$

As a corollary, one deduces that for each positive integer k one has

$$Z(X\otimes\mathbb{F}_{q^k})=\frac{\prod_i(1-\beta_i^k T)}{\prod_j(1-\gamma_j^k T)}.$$

The integrality of the coefficients of R_i can be deduced from the rationality of $Z(X)$ and the following elementary result.

LEMMA 2.5. *Let $f\in 1+T\mathbb{Z}[\![T]\!]$ be a rational function. Write*

$$f=\frac{f_1}{f_2},\quad (f_1,f_2)=1\quad f_i\in 1+T\mathbb{Q}[T].$$

Then $f_i\in 1+T\mathbb{Z}[T]$.

PROOF. This is usually derived from the so-called lemma of Fatou; see [Katz 1971]. Here we include two additional proofs.

One proof uses the Newton polygon or the Weierstrass factorization theorem. Suppose some prime number ℓ occurs in the common denominator of the coefficients of f_1. Then the theory of Newton polygons shows that f_1 has an ℓ-adic zero in the open unit disk $|T|_\ell<1$. But the power series $f\in 1+T\mathbb{Z}[\![T]\!]$

is clearly analytic and nonzero in the open unit disk $|T|_\ell < 1$. This gives the desired contradiction.

A second proof, suggested by Hendrik Lenstra, uses Gauss's lemma for power series. The *content* $\mathrm{cont}(g)$ of a nonzero power series $g = \sum_i a_i T^i \in \mathbb{Z}[\![T]\!]$ is defined to be the greatest common divisor of its coefficients a_i. Call g *primitive* if $\mathrm{cont}(g) = 1$ or, equivalently, if g is not in the kernel of the natural map $\mathbb{Z}[\![T]\!] \to \mathbb{F}_\ell[\![T]\!]$ for any prime number ℓ. Since the rings $\mathbb{F}_\ell[\![T]\!]$ are domains, the product of any two primitive power series is primitive. One deduces

$$\mathrm{cont}(g_1 g_2) = \mathrm{cont}(g_1)\mathrm{cont}(g_2),$$

which is Gauss's lemma for power series.

Lenstra's proof of Lemma 2.5 then proceeds as follow. Write $f = g_1/g_2$, where $g_i \in \mathbb{Z}[T]$ and $(g_1, g_2) = 1$ in $\mathbb{Q}[T]$. It is clear that $\mathrm{cont}(f) = 1$. The relation $g_1 = g_2 f$ implies that $\mathrm{cont}(g_1) = \mathrm{cont}(g_2)$. Cancelling this common factor, we may assume $\mathrm{cont}(g_1) = \mathrm{cont}(g_2) = 1$. Since g_1 and g_2 are relatively prime over $\mathbb{Q}$, there is a positive integer n such that

$$n \in g_1\mathbb{Z}[T] + g_2\mathbb{Z}[T] \subset g_2\mathbb{Z}[\![T]\!],$$

the last inclusion because $g_1 = g_2 f$. Write $n = hg_2$ with $h \in \mathbb{Z}[\![T]\!]$. Then

$$n = \mathrm{cont}(n) = \mathrm{cont}(h)\mathrm{cont}(g_2) = \mathrm{cont}(h).$$

Hence h is divisible by n. We conclude that $g_2(0) = \pm 1$. This implies 2.5. □

In order to actually compute the zeta function, it is useful to know an upper bound for the total degree $\deg R_1 + \deg R_2$ of the zeta function. The following explicit bound was proved by Bombieri [1978].

THEOREM 2.6. *The total degree of $Z(X)$ satisfies*

$$\deg R_1 + \deg R_2 < (4d+9)^{n+m},$$

where

$$d = \max_{1 \le j \le m} \deg(f_j).$$

Bombieri's bound is a general purpose bound. Its proof depends on Dwork's p-adic method. It can be improved in various ways, especially when one takes the Newton polytope of the defining polynomials f_i into account, as was done by Adolphson and Sperber [1988]. The bound is reasonably good as a function of d, but the dependence on m can probably be significantly improved.

An easy consequence is the following result.

COROLLARY 2.7. *The zeta function $Z(X)$ is effectively computable.*

This corollary is obvious in the special case that the zeta function $Z(X)$ is known to be a polynomial, or known to be the reciprocal of a polynomial, up to some trivial known factors. In the general case, one can deduce Corollary 2.7 from Theorem 2.4 and Theorem 2.6 in several ways, for instance by using the results of Berlekamp and Massey on linear recurring sequences (see [Blahut 1998] for more detail). Here we use a simple linear algebra argument explained to me by Hendrik Lenstra.

Let D_1 and D_2 be upper bounds for the degree of the numerator and the denominator of $Z(X)$, respectively. For instance, we can take $D_i = D = (4d+9)^{n+m}$, by Theorem 2.6. Compute the first $D_1 + D_2 + 1$ terms of the power series

$$Z(X) = 1 + z_1 T + z_2 T^2 + \cdots + z_{D_1+D_2} T^{D_1+D_2} + \cdots$$

by explicitly counting $X(\mathbb{F}_{q^k})$ for $k \le D_1 + D_2$. Write a_i, b_i for the coefficients of R_1 and R_2 to be determined:

$$R_1(X) = 1 + a_1 T + \cdots + a_{D_1} T^{D_1},$$
$$R_2(X) = 1 + b_1 T + \cdots + b_{D_2} T^{D_2}.$$

The congruence

$$R_2(X) Z(X) \equiv R_1(X) \pmod{T^{D_1+D_2+1}}$$

gives a system of linear equations in the a_i's and the b_i's. This system has at least one rational solution, and using linear algebra we can find one. Denote it by

$$(a'_1, \ldots, a'_{D_1}; b'_1, \ldots, b'_{D_2}).$$

Let

$$R'_1(X) = 1 + a'_1 T + \cdots + a'_{D_1} T^{D_1},$$
$$R'_2(X) = 1 + b'_1 T + \cdots + b'_{D_2} T^{D_2}.$$

Then, the congruence

$$R'_2(X) Z(X) \equiv R'_1(X) \pmod{T^{D_1+D_2+1}}$$

holds as well, and therefore

$$R_2(X) R'_1(X) \equiv R_2(X) R'_2(X) Z(X) \equiv R'_2(X) R_1(X) \pmod{T^{D_1+D_2+1}}.$$

Since each of $R_2 R'_1$ and $R'_2 R_1$ has degree at most $D_1 + D_2$, we deduce $R_2 R'_1 = R'_2 R_1$, so

$$\frac{R'_1(X)}{R'_2(X)} = \frac{R_1(X)}{R_2(X)} = Z(X).$$

Removing the greatest common factor of $R'_1(X)$ and $R'_2(X)$, one obtains the reduced form of $Z(X)$. The proof is complete.

The above effective algorithm immediately implies the following.

COROLLARY 2.8. *For fixed n, m, d, q, the number $\#X(\mathbb{F}_{q^k})$ can be computed in time bounded by a polynomial in k.*

We now estimate the output size of any algorithm computing $Z(X)$. Trivially, $\#X(\mathbb{F}_{q^k}) \leq \#\mathbb{A}^n(\mathbb{F}_{q^k}) = q^{nk}$. Hence $T \cdot d \log Z(X, T)/dT$ converges as a complex power series for $|T| < q^{-n}$, so its reciprocal poles β_i and γ_j are bounded by q^n in absolute value. Let D denote the total degree of $Z(X)$. Then $R_1(X) = \prod_i (1 - \beta_i T)$ and $R_2(X) = \prod_j (1 - \gamma_j T)$ have altogether $O(D)$ coefficients, each of which is $O(2^D q^{nD})$. Then, regardless of the input size for X, the output size of the algorithm is $O(nD^2 \log q)$. There is no general formula for the total degree D. However, for fixed m, Bombieri's degree bound $(4d + 9)^{n+m}$ is reasonably good, and it is comparable to the dense input size $O(m(d + 1)^n \log q)$.

We shall be concerned with the case that m is fixed (or small). If m is large, then the problem of computing $\#X(\mathbb{F}_q)$ is of a totally different, more combinatorial nature. The fundamental question that we consider is the following.

PROBLEM 2.9. *Given X with fixed m, compute the zeta function $Z(X)$ in time bounded by a polynomial in $(d + 1)^n \log q$.*

Remarks. If X has a sizable automorphism group, then one can often speed up the computation of $Z(X)$ by using a suitable equivariant theory. Examples include diagonal hypersurfaces, certain modular varieties, and certain Calabi–Yau hypersurfaces. In this paper, we do not assume that X is given with any additional structure of this sort.

Currently, the theory of zeta functions over finite fields comprises only two types of methods that are powerful enough to prove the rationality of the zeta function in the general case. These are the ℓ-adic method and the p-adic method, where ℓ denotes a prime number different from the characteristic p of the finite field $\mathbb{F}_q$. It is thus natural to try and exploit these general methods for algorithmic purposes.

In the ℓ-adic method one attempts to compute $Z(X) \bmod \ell^k$ using a suitable ℓ-adic trace formula. The zeta function $Z(X)$ can be recovered from its reduction modulo a single large prime power ℓ^k, or from its reductions modulo many small primes ℓ via the Chinese remainder theorem. Unfortunately, the available ℓ-adic trace formula is in the general case not yet effective. Thus the use of the ℓ-adic method is currently restricted to special varieties, such as curves and abelian varieties. In the cases in which it can be used, the ℓ-adic method usually results in a polynomial time algorithm if d, m, and n are fixed; see [Schoof 1985; Elkies 1998; Poonen 1996] for the first examples. It is an important open problem to make the ℓ-adic method effective in the general case.

In the p-adic method one attempts to compute $Z(X)$ mod p^k using a p-adic trace formula, where p^k is chosen so large that one can recover the zeta function $Z(X)$ from its reduction modulo p^k. There are many p-adic trace formulas. All of them can be made effective, although not all of them result in efficient algorithms. The general feeling is that the p-adic method is quite efficient if the characteristic p is suitably small, regardless of the size of the field $\mathbb{F}_q$ of definition and regardless of the degree of X. In [Lauder and Wan 2008] in this volume, we present a p-adic algorithm that proves the following theorem.

THEOREM 2.10. *There is an algorithm that, given X, computes the zeta function $Z(X)$ in time bounded by a polynomial in $2^m d^{n^2} p^n (\log q)^n$.*

For fixed m and n, and small p—say, $p = O((d \log q)^c)$ for some positive constant c—the algorithm of Theorem 2.10 runs in polynomial time, and it thus provides a partial solution to Problem 2.9. In Section 7 below, we illustrate some of the basic ideas of the p-adic method by treating an algorithm for the easier problem of computing $Z(X)$ mod p. For more details see [Lauder and Wan 2008].

3. Reduction to hypersurfaces

For the computation of the zeta function, the general case of an affine algebraic set X defined by a system

$$f_1(x) = \ldots = f_m(x) = 0$$

of m polynomial equations in n variables can be reduced to the case of an affine hypersurface defined by one single equation. In the present section we discuss various ways in which this reduction can be accomplished; not all of them are very efficient.

The quickest theoretical reduction depends on the observation that any algebraic set is birational to an affine hypersurface. One then continues by induction on the dimension. As it stands, this method is not very explicit. It may be of interest to make it both explicit and efficient.

A second method, which is explicit, exploits the inclusion-exclusion principle, as follows. For a subset $I \subset \{1, 2, \ldots, m\}$, let $H(I)$ be the affine hypersurface defined by

$$\prod_{i \in I} f_i = 0,$$

and let $H(I)^c$ be the complement of $H(I)$ in $\mathbb{A}^n$. Thus,

$$H(I) = \bigcup_{i \in I} \{f_i = 0\}, \qquad H(I)^c = \bigcap_{i \in I} \{f_i \neq 0\}.$$

In particular, we have $H(\varnothing) = \varnothing$ and $H(\varnothing)^c = \mathbb{A}^n$. The inclusion-exclusion principle implies that

$$Z(X) = \prod_{I \subset \{1,2,\ldots,m\}} Z(H(I)^c, T)^{(-1)^{\#I}}.$$

Since

$$Z(H(I)^c, T) = \frac{1}{(1-q^n T) Z(H(I), T)},$$

we conclude that for $m > 0$ we have

$$Z(X) = \prod_{\substack{I \subset \{1,2,\ldots,m\} \\ I \neq \varnothing}} Z(H(I), T)^{(-1)^{\#I-1}},$$

where the factor 1 corresponding to $I = \varnothing$ has been dropped. Each factor of the above product is now the zeta function of an affine hypersurface. Note that this reduction uses $2^m - 1$ hypersurfaces, but they are all in the original affine space $\mathbb{A}^n$.

If we apply Theorem 2.6 to each factor in the above identity, then we find that the total degree of $Z(X)$ is bounded by

$$\sum_{k=1}^{m} \binom{m}{k} (4kd + 9)^{n+1} < 2^m (4md + 9)^{n+1}.$$

In some cases where m is large this is better than what one obtains by applying Theorem 2.6 directly.

If one is willing to work with the slightly more general situation of L-functions of exponential sums, then one can use the single polynomial

$$g(x, y) = y_1 f_1(x) + \cdots + y_m f_m(x)$$

in $n + m$ variables. Let ζ_p denote a fixed primitive p-th root of unity in an extension field of $\mathbb{Q}$. For each positive integer k, define the exponential sum

$$S_k(g) = \sum_{x_i, y_j \in \mathbb{F}_{q^k}} \zeta_p^{\mathrm{Tr}_k(g(x,y))},$$

where Tr_k denotes the absolute trace from $\mathbb{F}_{q^k}$ to the prime field $\mathbb{F}_p$. The L-function associated to g is defined to be

$$L(g, T) = \exp\left(\sum_{k=1}^{\infty} \frac{S_k(g)}{k} T^k \right).$$

It is straightforward to check that

$$\#X(\mathbb{F}_{q^k}) = \frac{1}{q^{mk}} S_k(g).$$

This gives the desired reduction

$$Z(X) = L\Big(g, \frac{1}{q^m}T\Big).$$

Replacing f_i by af_i, one also deduces that

$$Z(X) = L\Big(ag, \frac{1}{q^m}T\Big)$$

for each nonzero $a \in \mathbb{F}_q$.

One can avoid the L-function in the above reduction by using the zeta function of the following Artin–Schreier hypersurface in $\mathbb{A}^{m+n+1}$:

$$Y\colon\ z^p - z = y_1 f_1(x) + \cdots + y_m f_m(x).$$

In fact, a direct calculation gives that

$$\#Y(\mathbb{F}_{q^k}) = \sum_{a\in\mathbb{F}_p} S_k(ag) = q^{(m+n)k} + \sum_{a\in\mathbb{F}_p^*} S_k(ag).$$

It follows that

$$Z(Y, q^{-m}T) = \frac{1}{1-q^nT} \cdot \prod_{a\in\mathbb{F}_p^*} L(ag, q^{-m}T) = \frac{1}{1-q^nT} Z(X)^{p-1}.$$

We obtain the formula

$$Z(X) = \big((1-q^nT)\cdot Z(Y, q^{-m}T)\big)^{1/(p-1)}.$$

For large p this reduction is not likely to be very efficient.

One can also use the affine hypersurface $H \subset \mathbb{A}^{m+n}$ defined by

$$H\colon\ g(x, y) = y_1 f_1(x) + \cdots + y_m f_m(x) = 0.$$

As S. Gao observed, one has

$$\#H(\mathbb{F}_q) = q^{m+n-1} + \#X(\mathbb{F}_q)(q^m - q^{m-1}).$$

This shows that for the purpose of counting rational points, we can work with a single hypersurface in the affine space $\mathbb{A}^{m+n}$. In terms of zeta functions, Gao's formula says that

$$Z(H, T) = \frac{Z(X, q^mT)}{(1-q^{m+n-1}T)Z(X, q^{m-1}T)}.$$

One can inductively solve for $Z(X)$ in terms of $Z(H, T)$. Doing this from the complex point of view, one gets the infinite complex product

$$Z(X) = \prod_{k=0}^{\infty}(1-q^{n-1-k}T)\cdot \prod_{k=0}^{\infty} Z(H, q^{-m-k}T).$$

Doing it from the p-adic point of view, one gets the infinite p-adic product

$$\frac{1}{Z(X)}=\prod_{k=0}^{\infty}(1-q^{n+k}T)\cdot\prod_{k=0}^{\infty}Z(H,q^{1-m+k}T).$$

There is a standard manner, as in Hilbert's tenth problem [Matiyasevich 1993], to define X by means of a system of equations of degree two. For example, if one of the equations defining X has a term $ax_1x_3^3x_4^2$, with $a\in\mathbb{F}_q$, then one can introduce new variables $x_{1,3}$, $x_{3,3}$, $x_{4,4}$, $x_{1,3,3,3}$, as well as new equations

$$x_{1,3}=x_1x_3,\qquad x_{3,3}=x_3^2,\qquad x_{4,4}=x_4^2,\qquad x_{1,3,3,3}=x_{1,3}x_{3,3},$$

and replace the term $ax_1x_3^3x_4^2$ by $ax_{1,3,3,3}x_{4,4}$; and one can proceed similarly with other terms. If one next applies Gao's reduction, then one obtains a hypersurface H that is defined by a cubic polynomial.

An amusing application of Gao's formula is the reduction of the Hasse–Weil meromorphy conjecture to the case of a cubic hypersurface. Let the polynomials f_i have integer coefficients, and let the affine algebraic sets X and H be defined as above, but now over $\mathbb{Z}$. Let

$$\zeta(X,z)=\prod_{p\text{ prime}}Z(X\otimes\mathbb{F}_p,p^{-z})$$

be the global complex Hasse–Weil zeta function of X; it is defined for complex numbers z whose real part is sufficiently large. The Hasse–Weil conjecture asserts that $\zeta(X)$ can be extended to a meromorphic function on all of $\mathbb{C}$. Gao's formula implies that

$$\zeta(H,z)=\frac{\zeta(X,z-m)\zeta(z+1-m-n)}{\zeta(X,z-m+1)},$$

where $\zeta(z)$ is the Riemann zeta function. From this relation, one deduces by induction or iteration that if the Hasse–Weil conjecture is valid for all cubic hypersurfaces H, then it is valid for all algebraic sets X. (By contrast, for Hilbert's tenth problem on diophantine equations over the integers, one only obtains a reduction to quartic hypersurfaces.) If we write

$$\zeta(X,z)=\sum_{k=1}^{\infty}\frac{a_k(X)}{k^z}$$

as a Dirichlet series, then Corollary 2.7 shows that each coefficient $a_k(X)$ is effectively computable.

4. Hypersurface examples

In this section, we focus on the crucial case of hypersurfaces. We discuss them by increasing dimension.

EXAMPLE 1. Let X be the zero-dimensional hypersurface defined by $f(x)=0$, where $f(x)$ is a nonconstant monic polynomial over $\mathbb{F}_q$ in one variable. Write

$$f(x)=P_1(x)^{k_1}\cdots P_e(x)^{k_e},$$

where the $P_i(x)$ are pairwise distinct monic irreducible polynomials in $\mathbb{F}_q[x]$ and the k_i are positive integers. Then the Euler product reads

$$Z(X)=\prod_{j=1}^{e}\frac{1}{1-T^{\deg P_j(x)}}.$$

It is not hard to show that one can compute $Z(X)$ in polynomial time using the Frobenius map. Note that if one factors the polynomial $f(x)$ first, one does not get a polynomial time algorithm for computing $Z(X)$. This is because there is currently no known (deterministic) polynomial time algorithm for factoring univariate polynomials over $\mathbb{F}_q$ if p is large, see [Wan 1999] for a further discussion and for the close relation of $Z(X)$ to various algorithms for factoring univariate polynomials over finite fields. This example also occurred as an exercise ascribed to Lenstra in [Cohen 1993, Chap. 6, Exerc. 8].

EXAMPLE 2. Let $f(x_1,x_2)\in\mathbb{F}_q[x_1,x_2]$ be of degree d, and suppose that the homogenization of $f(x_1,x_2)$ defines a smooth projective plane curve C_d over $\mathbb{F}_q$. The genus g of the curve C_d is well known to be

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

From the Riemann–Roch theorem one can deduce (see [Monsky 1970])

$$Z(C_d,T)=\frac{P(C_d,T)}{(1-T)(1-qT)},$$

where $P(C_d,T)$ is a polynomial of degree $2g$. Weil proved that we further have

$$P(C_d,T)=\prod_{j=1}^{2g}(1-\alpha_jT),\quad |\alpha_j|=\sqrt{q},\quad \alpha_j\alpha_{2g+1-j}=q.$$

The Riemann–Roch theorem also shows that the special value $P(C_d,1)$ has the following arithmetic meaning:

$$P(C_d,1)=\#J(C_d)(\mathbb{F}_q)\in\mathbb{Z}_{>0},$$

where $J(C_d)$ is the Jacobian variety of C_d, which is a g-dimensional abelian variety over $\mathbb{F}_q$. All these results hold for any smooth projective geometrically irreducible curve over $\mathbb{F}_q$, not just for plane curves.

The ℓ-adic method can be made effective in the case of curves and abelian varieties. One can then use the Chinese remainder theorem as mentioned before. In this way, one obtains an algorithm for computing $Z(C_d, T)$ with running time $O\big((\log q)^{\Delta_d}\big)$, where Δ_d is in general an exponential function of d; see [Schoof 1985; Pila 1990]. Thus, for fixed d, the ℓ-adic method computes the zeta function $Z(C_d, T)$ in polynomial time, although the algorithm is still doubly exponential in d. For hyperelliptic curves, the exponent Δ_d has been improved to a polynomial in d; see [Adleman and Huang 1996]. On the other hand, using the p-adic algorithm in [Lauder and Wan 2008], one can compute the zeta function $Z(C_d, T)$ in time $(dp \log q)^{O(1)}$, which is polynomial in d but exponential in $\log p$. In particular, the zeta function $Z(C_d, T)$ can be computed in polynomial time if $p = O\big((d \log q)^{O(1)}\big)$. This example is important because of its many applications in number theory and cryptography; see [Koblitz 1989; Blake et al. 2000]. For special types of curves in small characteristic, more practical versions of various p-adic algorithms have been designed by a number of authors; see [Satoh 2000; Kedlaya 2001; Lauder and Wan 2002; Denef and Vercauteren 2002], and the references listed in those papers. Restricting Problem 2.9 to plane curves, we obtain the following.

PROBLEM 4.1. *Given a smooth projective plane curve C_d over $\mathbb{F}_q$, compute $Z(C_d, T)$ in time $O\big((d \log q)^c\big)$, where c is an explicit absolute positive constant.*

EXAMPLE 3. Let $f(x_1, \dots, x_n) \in \mathbb{F}_q[x_1, \dots, x_n]$ be of degree d, and suppose that the homogenization of f defines a smooth projective hypersurface H_d of dimension $n-1$ over $\mathbb{F}_q$. Then by the Weil conjectures [Deligne 1974] we can write

$$Z(H_d, T) = \frac{P(H_d, T)^{(-1)^n}}{\prod_{j=0}^{n-1}(1 - q^j T)},$$

where $P(H_d, T) \in 1 + T\mathbb{Z}[T]$ is a polynomial of degree

$$D = \frac{1}{d}\{(d-1)^{n+1} + (-1)^{n+1}(d-1)\}$$

(see [Monsky 1970]) and

$$P(H_d, T) = \prod_{j=1}^{D}(1 - \alpha_j T), \quad |\alpha_j| = \sqrt{q}^{n-1}, \quad \alpha_j \alpha_{D+1-j} = q^{n-1}.$$

This higher dimensional example is undoubtedly more difficult than the previous example of curves, and it has attracted a smaller amount of attention. Both theoretically and computationally, and from an applied point of view, our understanding of this case leaves a great deal to be desired. In particular, the ℓ-adic method has not been made effective for higher dimensional smooth projective hypersurfaces. We do know that by the p-adic algorithm in [Lauder and Wan 2008], the zeta function $Z(H_d, T)$ can be computed in time $(d^n p \log q)^{O(n)}$ for any $(n-1)$-dimensional hypersurface H_d, not necessarily smooth or projective. This gives a polynomial time algorithm for small p and fixed n. In the smooth projective case, it should be possible to improve the exponent $O(n)$ by finer p-adic cohomological methods. In fact, Lauder [2004b; 2004a] has gone further and used the deformation method on the cohomology space to show that the zeta function can be computed in time $(d^n p \log q)^{O(1)}$ for suitable smooth projective $(n-1)$-dimensional hypersurface H_d. It would be interesting to explore possible applications of higher dimensional hypersurfaces. As a special case of Problem 2.9, we have the following.

PROBLEM 4.2. *Given a smooth projective $(n-1)$-dimensional hypersurface H_d defined over $\mathbb{F}_q$, compute $Z(H_d, T)$ in time bounded by a polynomial in $(d+1)^n \log q$.*

Lauder's recent results solve this problem if p is small.

5. Pure weight decomposition

In this section, we consider the problem of computing the numbers of zeros and poles of $Z(X)$ with a given complex absolute value, for any algebraic set X defined over $\mathbb{F}_q$. Let $R(X, T)$ be the numerator or the denominator of the zeta function $Z(X)$.

Over the complex numbers $\mathbb{C}$, we can write

$$R(X, T) = \prod_i (1 - \alpha_i T),$$

where each α_i is a nonzero algebraic integer. Define the *weight* of a nonzero complex number α by

$$w(\alpha) = \log_q(\alpha\bar{\alpha}),$$

where $\bar{\alpha}$ denotes the complex conjugate of α. An elementary archimedean estimate shows that

$$\#X(\mathbb{F}_{q^k}) = O(q^{k \dim X}),$$

where the implied constant depends on the degree of X. As in Section 2, one deduces the elementary estimate

$$|\alpha_i| \le q^{\dim X}, \quad w(\alpha_i) \le 2\dim X.$$

The following much deeper result is a consequence of Deligne's main theorem on the Weil conjectures [Deligne 1980].

THEOREM 5.1. *The weights of the reciprocal roots α_i of $R(X,T)$ are integers in the interval* $[0, 2\dim X]$. *That is*, *for each α_i*, *there is an integer w_i with* $0 \le w_i \le 2\dim X$ *such that*

$$\alpha_i \bar{\alpha}_i = q^{w_i}.$$

In particular, *each α_i is an ℓ-adic unit for all prime numbers $\ell \ne p$. Furthermore*, *each α_i and its Galois conjugates have the same weight*.

The last part of the theorem can be deduced from the first part in an elementary manner, as shown in the following result of Lenstra.

LEMMA 5.2. *Let $f \in \mathbb{Q}[T]$ be an irreducible polynomial. Suppose α and β are two complex roots of f with $\alpha\bar{\alpha} \in \mathbb{Q}$ and $\beta\bar{\beta} \in \mathbb{Q}$. Then $\alpha\bar{\alpha} = \beta\bar{\beta}$.*

PROOF. Let $\alpha\bar{\alpha} = a$, $\beta\bar{\beta} = b$, and $\deg(f) = n > 0$. We need to show that $a = b$. Since there is a field isomorphism $\mathbb{Q}(\alpha) \cong \mathbb{Q}(\beta)$ that sends α to β, we have

$$N_{\mathbb{Q}(\alpha)/\mathbb{Q}}(\alpha) = N_{\mathbb{Q}(\beta)/\mathbb{Q}}(\beta),$$

where N denotes the norm map. From $a \in \mathbb{Q}$ and $\mathbb{Q}(\alpha) = \mathbb{Q}(\bar{\alpha})$ one deduces

$$a^n = N_{\mathbb{Q}(\alpha)/\mathbb{Q}}(a) = N_{\mathbb{Q}(\alpha)/\mathbb{Q}}(\alpha) N_{\mathbb{Q}(\bar{\alpha})/\mathbb{Q}}(\bar{\alpha}) = N_{\mathbb{Q}(\alpha)/\mathbb{Q}}(\alpha)^2.$$

Similarly,

$$b^n = N_{\mathbb{Q}(\beta)/\mathbb{Q}}(\beta)^2.$$

Putting the above together, one deduces that $a^n = b^n$. Since $a \ge 0$ and $b \ge 0$, we conclude that $a = b$. □

For an integer w with $0 \le w \le 2\dim X$, let

$$R(w, X, T) = \prod_{w(\alpha_i)=w} (1 - \alpha_i T).$$

This is called the *pure weight w part* of $R(X,T)$. By the above theorem of Deligne, each $R(w,X,T)$ is a polynomial in $1 + T\mathbb{Z}[T]$. The *pure weight decomposition* is

$$R(X,T) = \prod_{w=0}^{2\dim X} R(w,X,T), \quad R(w,X,T) \in 1 + T\mathbb{Z}[T].$$

THEOREM 5.3. *Given the zeta function $Z(X)$, one can compute all pure parts $R(w, X, T)$ in polynomial time.*

PROOF. By the LLL factorization algorithm [Lenstra et al. 1982], the polynomial $R(X, T)$ can be factored as a product of irreducible polynomials in polynomial time:

$$R(X, T) = \prod_i g_i(T),$$

where each factor

$$g_i \in 1 + T\mathbb{Z}[T]$$

is irreducible over $\mathbb{Q}$. Write

$$g_i(T) = 1 + a_{i1}T + \cdots + a_{ie_i}T^{e_i} = \prod_{j=1}^{e_i}(1 - \beta_{ij}T), \quad a_{ie_i} \neq 0.$$

By Deligne's theorem 5.1, each β_{ij} is pure of some integer weight w_i. Hence $a_{ie_i}^2 = \prod_j \beta_{ij}\bar{\beta}_{ij} = q^{e_i w_i}$, and therefore

$$a_{ie_i} = \pm\sqrt{q}^{w_i e_i} \in \mathbb{Z}.$$

One can recover the integer weight w_i from

$$w_i = 2\frac{\log_q |a_{ie_i}|}{e_i}.$$

The pure weight w part of $Z(X)$ is then

$$R(w, X, T) = \prod_{w_i = w} g_i(T). \qquad \square$$

Clearly, for each irreducible factor $g_i(T)$ of $R(X, T)$, the map $\beta_{ij} \mapsto \bar{\beta}_{ij} = q^{w_i}/\beta_{ij}$ permutes the reciprocal roots of $g_i(T)$. This gives the following functional equation.

COROLLARY 5.4. *For each w, the pure weight part $R(w, X, T)$ satisfies*

$$R(w, X, 1/(q^w T)) = \pm q^{-wd(w,R)/2} T^{-d(w,R)} R(w, X, T),$$

where $d(w, R)$ denotes the degree of $R(w, X, T)$.

The following is also immediate from Theorem 5.3.

COROLLARY 5.5. *Given the zeta function, the degrees $d(w, R)$ of all pure weight parts $R(w, X, T)$ can be computed in polynomial time.*

The pure degrees $d(w, R)$ contain important geometric information about the variety X. For instance, if R_2 denotes the denominator of $Z(X)$, then

$$d(2\dim X, R_2)$$

is the number of top-dimensional components of $X \otimes \bar{\mathbb{F}}_q$. If X is a geometrically irreducible curve and R_1 denotes the numerator of $Z(X)$, then $d(1, R_1)$ is twice the genus of the nonsingular model of X. If X is smooth and projective, the pure degrees $d(w, R)$ are geometric invariants, that is, they depend only on $X \otimes \bar{\mathbb{F}}_q$. In fact, in this smooth projective case, they are given by the ℓ-adic Betti numbers, as shown in the proof of the following result.

COROLLARY 5.6. *For any smooth projective variety X defined over $\mathbb{F}_q$, the ℓ-adic Betti numbers of X can be effectively computed, where $\ell \neq p$.*

PROOF. For a smooth projective variety X over $\mathbb{F}_q$, the ℓ-adic trace formula states

$$Z(X) = \prod_{i=0}^{2\dim X} \det(I - (\mathrm{Frob}|H^i(X \otimes \bar{\mathbb{F}}_q, \mathbb{Q}_\ell))T)^{(-1)^{i-1}},$$

where Frob denotes the geometric Frobenius map. By the Weil conjectures as proved in [Deligne 1974], each (complex) eigenvalue of Frob acting on H^i has weight equal to i. Thus, the i-th Betti number

$$B_i(X, \ell) = \dim_{\mathbb{Q}_\ell} H^i(X \otimes \bar{\mathbb{F}}_q, \mathbb{Q}_\ell)$$

is given by the formula

$$B_i(X, \ell) = \begin{cases} d(i/2, R_1) & \text{if } i \text{ is odd,} \\ d(i/2, R_2) & \text{if } i \text{ is even,} \end{cases}$$

where R_1 and R_2 are the numerator and denominator of $Z(X)$, respectively. The corollary follows. □

Let Y be a smooth projective scheme over $\mathbb{Z}$ and let $X = Y \otimes \mathbb{F}_p$ be the reduction modulo p of Y. The cohomological comparison theorem shows that for all large primes p, the pure degrees $d(w, R(X, T))$ depend only on the geometry of $Y \otimes \bar{\mathbb{Q}}$, not on the chosen large prime p.

If X is singular or open, the pure degrees $d(w, R)$ may not be geometric invariants, as it may happen that a root and a pole of $Z(X)$ have a quotient that is a root of unity. On the other hand, it is clear that the pure difference degrees $d(w, R_1) - d(w, R_2)$ are geometric invariants, where we recall that R_1 and R_2 denote the numerator and the denominator of $Z(X)$, respectively.

It would be interesting to know if the pure degrees $d(w, R)$ can be computed in polynomial time without the zeta function being given. Even for a singular plane curve this seems to be unknown. In the case of a smooth projective complete intersection there is a well-known formula showing that the pure degrees can be computed from the total degree of $Z(X)$ and the dimension of X; see Example 3 of Section 4 for the case of a smooth projective hypersurface and [Deligne and Katz 1973, pp. 39–61] for the general smooth projective complete intersection case.

For arbitrary X over $\mathbb{F}_q$, not necessarily smooth or projective, Grothendieck [1968] has shown that for $\ell \neq p$ there is a similar ℓ-adic formula in terms of ℓ-adic cohomology with compact support:

$$Z(X) = \prod_{i=0}^{2\dim X} \det(I - (\mathrm{Frob}|H_c^i(X \otimes \bar{\mathbb{F}}_q, \mathbb{Q}_\ell))T)^{(-1)^{i-1}}.$$

But in this generality, even the conjectured independence of the ℓ-adic Betti numbers on ℓ is unknown. The following result of Katz [2001] provides weak evidence in this direction. It gives an explicit upper bound for the ℓ-adic Betti numbers with compact support that is independent of ℓ. It is obtained by means of an inductive reduction to the Bombieri–Adolphson–Sperber degree bound for $Z(X)$, which in turn is p-adic in nature.

THEOREM 5.7. *Let the polynomials $f_1, \dots, f_m$ form a system of defining equations for the affine algebraic set X, and put $d = \max_i \deg(f_i)$. Then, for every prime number $\ell \neq p$, we have*

$$\sum_{i \geq 0} \dim_{\mathbb{Q}_\ell} H_c^i(X \otimes \bar{\mathbb{F}}_q, \mathbb{Q}_\ell) \leq 2^{m+2}(md+3)^{n+1}.$$

Turning to global zeta functions, it may be of interest to point out that the real parts of the zeros of the classical Riemann zeta function $\zeta(z)$ are not known to be effectively computable in a finite region. To make this precise, denote, for real numbers $w \in (0,1)$ and $t > 0$, by $d(w, \zeta; t)$ the number of zeros of $\zeta(z)$ that lie on the line segment

$$\mathrm{Re}(z) = w, \quad 0 \leq \mathrm{Im}(z) \leq t.$$

For any given t, there are only finitely many w for which $d(w, \zeta; t) > 0$. The following computational problem is now analogous to finding the pure weight decomposition: given t, determine the finitely many positive integers among the numbers $d(w, \zeta; t)$, for $0 \leq w \leq 1$, as well as approximations to the corresponding numbers w. No effective algorithm for doing this is currently available. The main difficulty is caused by the possibility of a multiple zero or zeros that are very close to each other, see [Odlyzko 1994]. Of course, the Riemann hypothesis says that $d(w, \zeta; t) = 0$ for all t and all $w \neq 1/2$.

6. Pure slope decomposition

In the previous section, we considered the problem of purity decomposition of $Z(X)$ from the complex point of view. One can also consider the purity decomposition from a nonarchimedean point of view. If ℓ is a prime number different from p, then Deligne's Theorem 5.1 shows that $Z(X)$ is already pure from an ℓ-adic point of view. Thus, we will consider the remaining p-adic case in this section. Let $R(X,T)$ again denote either the numerator or the denominator of $Z(X)$.

Let $\mathbb{C}_p$ be the completion of a fixed algebraic closure of $\mathbb{Q}_p$. Over $\mathbb{C}_p$, we can write

$$R(X,T) = \prod_i (1-\alpha_i T),$$

where each α_i is a nonzero algebraic integer in $\mathbb{C}_p$. Define the slope of a nonzero element $\alpha \in \mathbb{C}_p$ by

$$s(\alpha) = \mathrm{ord}_q(\alpha) = -\log_q |\alpha|_p,$$

where $|\cdot|_p$ denotes the p-adic absolute value, normalized such that $|p| = 1/p$. The slopes $s(\alpha_i)$ are nonnegative rational numbers since the α_i are algebraic integers. One immediately derives the bound

$$0 \le s(\alpha_i) \le s(\alpha_i \bar{\alpha}_i) = w(\alpha_i) \le 2\dim X.$$

This bound can be improved somewhat. Deligne's integrality theorem [1973, pp. 384–400] states that $q^{\dim X}/\alpha_i$ is an algebraic integer. We deduce the following result.

THEOREM 6.1. *The slopes $s(\alpha_i)$ are rational numbers in* $[0, \dim X]$.

Other than in the complex absolute value case, α_i and its Galois conjugates over $\mathbb{Q}$ may have different slopes. Of course, each α_i and its Galois conjugates over $\mathbb{Q}_p$ do have the same slope. The slopes $s(\alpha_i)$ are not integers (or half integers) in general. They are merely rational numbers. It would be interesting to get good bounds for the denominators of the slopes $s(\alpha_i)$.

For a rational number s with $0 \le s \le \dim X$, let

$$R(s,X,T) = \prod_{s(\alpha_i)=s} (1-\alpha_i T).$$

This is called the pure slope s part of $R(X,T)$. We have the p-adic purity decomposition

$$R(X,T) = \prod_{s\in\mathbb{Q}} R(s,X,T).$$

The question is then to understand each pure slope part $R(s,X,T)$.

Note that in general, the pure slope parts $R(s, X, T)$ do not have coefficients in $\mathbb{Z}$ any more. The theory of Newton polygons implies that $R(s, X, T)$ is a polynomial in $1 + T(\bar{\mathbb{Q}} \cap \mathbb{Z}_p)[T]$, as we shall see now.

Write

$$R(X, T) = 1 + a_1 T + \cdots + a_e T^e = \prod_{i=1}^{e} (1 - \alpha_i T), \quad a_e \neq 0.$$

DEFINITION 6.2. The Newton polygon $\mathrm{NP}(R)$ of $R(X, T)$ is the lower convex hull in the plane $\mathbb{R}^2$ of the points

$$(k, \mathrm{ord}_q(a_k)), \quad k = 0, 1, \ldots, e.$$

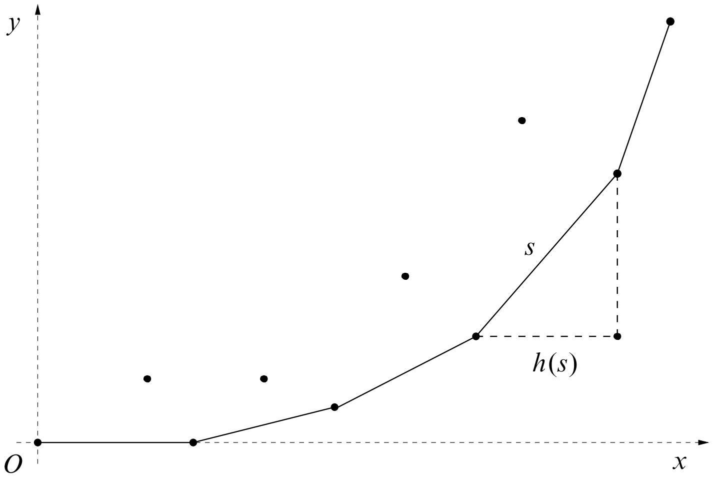

Figure 1. Newton polygon

A basic property of the Newton polygon is the following result; see [Koblitz 1984].

THEOREM 6.3. *The polynomial* $R(X, T) \in 1 + T\mathbb{Z}[T]$ *has exactly* h *reciprocal zeros* α_i *with slope* $\mathrm{ord}_q(\alpha_i) = s$ (*counting multiplicities*) *if and only if* $\mathrm{NP}(R)$ *has a side of slope* s *and horizontal length* h. *Furthermore*, *the coefficients of* $R(s, X, T)$ *are in* $\mathbb{Z}_p$ *for each* s.

The following is an immediate consequence.

COROLLARY 6.4. *Let* $d(s, R)$ *denote the degree of* $R(s, X, T)$. *Then*, $d(s, R)$ *is the horizontal length of the slope* s *side of* $\mathrm{NP}(R)$. *In particular*, *when* $Z(X)$ *is given*, *the* p*-adic pure degrees* $d(s, R)$ *can be computed in polynomial time.*

It would be interesting to know if the p-adic pure degrees $d(s, R)$ can be computed in polynomial time without the zeta function being given. This amounts to computing the Newton polygons of the numerator and the denominator of the zeta function. Even in a very well-behaved situation such as the smooth projective hypersurface case, we do not have a complete answer in general; see [Wan 2004] for a theoretical introduction to Newton polygons for zeta functions and L-functions. In the smooth projective case, the p-adic pure degrees $d(s, R)$ are geometric invariants.

If X is the good reduction modulo p of some smooth projective scheme over $\mathbb{Z}$, the p-adic pure degrees $d(s, R)$ depend not just on the geometry of the generic fibre $Y \otimes \bar{\mathbb{Q}}$, but also on the chosen prime p. Thus, the p-adic pure degrees $d(s, R)$ contain arithmetic information on Y. They are related to but much deeper than the topological Hodge numbers of $Y \otimes \mathbb{Q}$ defined in terms of the De Rham cohomology of $Y \otimes \mathbb{Q}$; see [Mazur 1972] for an introductory account. Describing the variation of $d(s, R)$ as p varies is a very subtle arithmetic problem, already in the special case that $Y \otimes \mathbb{Q}$ is an elliptic curve.

As a polynomial with coefficients in $\mathbb{Z}_p$, the pure slope part $R(s, X, T)$ cannot be written down in a finite amount of time. However, given $R(X, T)$, one can compute the pure slope parts $R(s, X, T)$ modulo any given power of p in polynomial time using the Newton polygon and Hensel lifting. Note that here we do not factor $R(X, T)$ into a product of irreducible factors over $\mathbb{Q}_p$, which is a harder problem.

7. Zeta functions modulo p

In [Lauder and Wan 2008] a p-adic algorithm for computing the zeta function $Z(X)$ is given, where p is the characteristic of $\mathbb{F}_q$. In this final section, we describe some of the basic ideas behind that algorithm by showing how $Z(X)$ may be computed modulo p. The method expands the outline given in [Wan 1999].

Without loss of generality we may restrict to hypersurfaces. Thus, let X be the affine hypersurface defined by a polynomial $f(x_1, \dots, x_n)$ over $\mathbb{F}_q$ of total degree d in n variables. Let A_d be the $\mathbb{F}_q$-vector space of polynomials in x_1, …, x_n of total degree at most d that are divisible by the product $x_1 \cdots x_n$:

$$A_d = (x_1 \cdots x_n \mathbb{F}_q[x_1, \dots, x_n])_{\leq d}.$$

It has a row basis $\vec{e}$ consisting of monomials

$$\vec{e} = \{x^u | u = (u_1, \dots, u_n),\ u_i \geq 1,\ |u| \leq d\},$$

where

$$x^u = x_1^{u_1} \cdots x_n^{u_n}, \quad |u| = u_1 + \cdots + u_n.$$

One computes that

$$\dim_{\mathbb{F}_q} A_d = \binom{d}{n}.$$

DEFINITION 7.1. Let σ be the p-th power Frobenius map acting on $\mathbb{F}_q$:

$$\sigma(a) = a^p, \quad a \in \mathbb{F}_q.$$

Let ψ_p be the σ^{-1}-linear operator on the $\mathbb{F}_q$-vector space $\mathbb{F}_q[x_1, \dots, x_n]$ defined by

$$\psi_p\biggl(\sum_u a_u x^u\biggr) = \sum_u \sigma^{-1}(a_u)\psi_p(x^u),$$

where

$$\psi_p(x^u) = \begin{cases} x^{u/p}, & \text{if } p|u, \\ 0, & \text{otherwise.} \end{cases}$$

Define

$$\psi_q = \psi_p^r = \psi_p \circ \cdots \circ \psi_p$$

to be the r-th iterate of ψ_p, where $q = p^r$.

Since σ^r is the identity on $\mathbb{F}_q$, the operator ψ_q is actually $\mathbb{F}_q$-linear, although ψ_p is only σ^{-1}-linear. The operator ψ_q is a left inverse of the q-th power Frobenius map on $\mathbb{F}_q[x] = \mathbb{F}_q[x_1, \dots, x_n]$. For any polynomial $g \in \mathbb{F}_q[x]$, multiplication by g is also an $\mathbb{F}_q$-linear map from $\mathbb{F}_q[x]$ to itself.

LEMMA 7.2. *The $\mathbb{F}_q$-linear composed operator $\psi_q \circ f^{q-1}$ maps the finite dimensional $\mathbb{F}_q$-subspace A_d of $\mathbb{F}_q[x]$ to itself*; *here f^{q-1} denotes multiplication by f^{q-1}. Similarly, the σ^{-1}-linear composed operator $\psi_p \circ f^{p-1}$ maps A_d to itself*, *and we have the relation*

$$\psi_q \circ f^{q-1} = (\psi_p \circ f^{p-1})^r.$$

PROOF. Let $h \in A_d$. Then h has degree at most d, so $f^{q-1}h$ has degree at most $d(q-1) + d = dq$. Thus, the degree of $\psi_q(f^{q-1}h)$ is at most d. Furthermore, if h is divisible by $x_1 \cdots x_n$, then $\psi_q(f^{q-1}h)$ is also divisible by $x_1 \cdots x_n$. This proves that $\psi_q(f^{q-1}h) \in A_d$. The proof of the second part of the lemma is the same. To prove the last part, we write

$$q - 1 = (p-1) + p(p-1) + \cdots + p^{r-1}(p-1).$$

This gives

$$f(x)^{q-1} = \prod_{i=0}^{r-1} f^{\sigma^i}(x^{p^i})^{p-1},$$

where $f^{\sigma^i}(x)$ denotes the polynomial obtained by applying σ^i to the coefficients of f. Using the easily checked relation

$$\psi_p \circ g^{\sigma}(x^p) = g(x) \circ \psi_p,$$

one deduces

$$\psi_q \circ f^{q-1} = \psi_p^r \circ \prod_{i=0}^{r-1} f^{\sigma^i}(x^{p^i})^{p-1} = (\psi_p \circ f^{p-1})^r. \qquad \square$$

THEOREM 7.3. *Let X be the affine hypersurface defined by a polynomial $f(x_1, \dots, x_n)$ over $\mathbb{F}_q$ of total degree d in n variables. Then we have the congruence formula*

$$(Z(X)^{(-1)^n} \bmod p) = \det(I - (\psi_q \circ f^{q-1}|A_d)T).$$

Before we prove this result, we first recall some facts on L-functions in characteristic p. Let $g \in \mathbb{F}_q[x] = \mathbb{F}_q[x_1, \dots, x_n]$. For a geometric point $x \in \mathbb{A}^n(\bar{\mathbb{F}}_q)$, the product $g(x)g(x^q)\cdots g(x^{q^{\deg(x)-1}})$ is an element of $\mathbb{F}_q$ that clearly depends only on the orbit (the closed point) of x under the q-th power Frobenius map. We define the L-function of g by

$$L(g, T) = \prod_x \frac{1}{1 - g(x)g(x^q)\cdots g(x^{q^{\deg(x)-1}})T^{\deg(x)}} \in 1 + T\mathbb{F}_q[\![T]\!],$$

where x runs over the set of closed points of the affine space $\mathbb{A}^n$ over $\mathbb{F}_q$.

For a real number c, write A_c for the finite dimensional $\mathbb{F}_q$-subspace of $\mathbb{F}_q[x]$ generated by the monomials of total degree at most c that are divisible by the product $x_1 \cdots x_n$. If the total degree of g is at most e, one checks as in the above lemma that the operator $\psi_q \circ g$ maps the subspace $A_{e/(q-1)}$ to itself. Furthermore, the (matrix of the) induced map $\psi_q \circ g$ on the quotient vector space $x_1 \cdots x_n \mathbb{F}_q[x]/A_{e/(q-1)}$ is strictly triangular with respect to the monomial basis $\{x^u | u_i \geq 1, |u| > e/(q-1)\}$. Thus, the composed operator $\psi_q \circ g$ acting on the infinite dimensional $\mathbb{F}_q$-vector space $x_1 \cdots x_n \mathbb{F}_q[x]$ has a well defined characteristic power series $\det(I - (\psi_q \circ g | x_1 \cdots x_n \mathbb{F}_q[x])T)$, which is given by the polynomial

$$\det(I - (\psi_q \circ g | x_1 \cdots x_n \mathbb{F}_q[x])T) = \det(I - (\psi_q \circ g | A_{e/(q-1)})T).$$

The characteristic p version of Dwork's trace formula for the affine space $\mathbb{A}^n$, as given in [Wan 1996], implies that

$$L(g, T)^{(-1)^{n-1}} = \det(I - (\psi_q \circ g | x_1 \cdots x_n \mathbb{F}_q[x])T)$$

and thus

$$L(g, T)^{(-1)^{n-1}} = \det(I - (\psi_q \circ g | A_{e/(q-1)})T).$$

The reader is referred to [Lauder and Wan 2008] to see the Dwork trace formula over the n-torus, which is cleaner than the Dwork trace formula over the affine n-space $\mathbb{A}^n$.

We now return to the proof of the theorem. Taking $g = f^{q-1}$ and $e = d(q-1)$, we deduce that $e/(q-1) = d$ and

$$L(f^{q-1}, T)^{(-1)^{n-1}} = \det(I - (\psi_q \circ f^{q-1} | A_d)T).$$

For a geometric point $x \in \mathbb{A}^n(\bar{\mathbb{F}}_q)$, one has

$$g(x)g(x^q)\cdots g(x^{q^{\deg(x)-1}}) = f(x)^{q^{\deg(x)}-1},$$

which is 0 or 1 according as $x \in X(\bar{\mathbb{F}}_q)$ or not. It follows that $L(f^{q-1}, T)$ is the reduction modulo p of the zeta function of the complement of X in $\mathbb{A}^n$. Hence for $n > 0$ we obtain

$$L(f^{q-1}, T) = (1/Z(X) \bmod p).$$

Substituting this into the above formula for $L(f^{q-1}, T)^{(-1)^{n-1}}$, we obtain the theorem.

COROLLARY 7.4. *The zeta function $Z(X)$ modulo p can be computed in time bounded by a polynomial in $p\binom{d}{n}\log q$.*

PROOF. Recall that $q = p^r$. Let M_1 be the matrix of the σ^{-1}-linear map $\psi_p \circ f^{p-1}$ with respect to the row monomial basis $\vec{e}$ of A_d. That is,

$$(\psi_p \circ f^{p-1})(\vec{e}) = \vec{e}M_1.$$

Then, by the σ^{-1}-linearity of $\psi_p \circ f^{p-1}$, we deduce

$$(\psi_p \circ f^{p-1})^2(\vec{e}) = (\psi_p \circ f^{p-1})(\vec{e}M_1) = \vec{e}M_1 M_1^{\sigma^{-1}},$$

where $M_1^{\sigma^{-1}}$ is obtained from M_1 by applying σ^{-1} to each entry (and similarly with $M_1^{\sigma^{-i}}$ below). By iteration, one finds that the matrix of the $\mathbb{F}_q$-linear map

$$\psi_q \circ f^{q-1} = (\psi_p \circ f^{p-1})^r$$

with respect to the row basis $\vec{e}$ is given by

$$M_r = M_1 M_1^{\sigma^{-1}} \cdots M_1^{\sigma^{-(r-1)}}.$$

The matrix M_1 can be written down in time $(p\binom{d}{n}\log q)^{O(1)}$. It follows that the matrix M_r can also be computed in time $(p\binom{d}{n}\log q)^{O(1)}$. The zeta function modulo p is essentially just the characteristic polynomial of the matrix M_r:

$$(Z(X)^{(-1)^n} \bmod p) = \det(I - M_r T).$$

The corollary is proved. Alternatively, applying σ^{r-1} to the above congruence formula, we have

$$(Z(X)^{(-1)^n} \bmod p) = \det(I - M_1^{\sigma^{r-1}} \cdots M_1^{\sigma} M_1 T).$$

This formula may be slightly more efficient from a computational point of view.

If p is small, that is, $p = (\binom{d}{n} \log q)^{O(1)}$, the above corollary gives a polynomial time algorithm for computing $Z(X)$ modulo p. If p is large, we do not get a polynomial time algorithm. □

Acknowledgment

It is a pleasure to thank S. Gao, H.W. Lenstra, Jr., and A.M. Odlyzko for several interesting discussions. Special thanks are due to the referee for many helpful suggestions.

References

[Adleman and Huang 1996] L. M. Adleman and M.-D. A. Huang, "Counting rational points on curves and abelian varieties over finite fields", pp. 1–16 in *Algorithmic number theory* (ANTS-II) (Talence, 1996), edited by H. Cohen, Lecture Notes in Comput. Sci. **1122**, Springer, Berlin, 1996.

[Adolphson and Sperber 1988] A. Adolphson and S. Sperber, "On the degree of the L-function associated with an exponential sum", *Compositio Math.* **68**:2 (1988), 125–159.

[Blahut 1998] R. E. Blahut, "Decoding of cyclic codes and codes on curves", pp. 1569–1633 in *Handbook of coding theory*, vol. II, edited by V. S. Pless and W. C. Huffman, North-Holland, Amsterdam, 1998.

[Blake et al. 2000] I. F. Blake, G. Seroussi, and N. P. Smart, *Elliptic curves in cryptography*, London Mathematical Society Lecture Note Series **265**, Cambridge University Press, Cambridge, 2000.

[Bombieri 1978] E. Bombieri, "On exponential sums in finite fields, II", *Invent. Math.* **47**:1 (1978), 29–39.

[Cohen 1993] H. Cohen, *A course in computational algebraic number theory*, Graduate Texts in Mathematics **138**, Springer, Berlin, 1993.

[Deligne 1974] P. Deligne, "La conjecture de Weil, I", *Inst. Hautes Études Sci. Publ. Math.* no. 43 (1974), 273–307.

[Deligne 1980] P. Deligne, "La conjecture de Weil, II", *Inst. Hautes Études Sci. Publ. Math.* no. 52 (1980), 137–252.

[Deligne and Katz 1973] P. Deligne and N. Katz, *Groupes de monodromie en géométrie algébrique* (SGA 7 II), Lecture Notes in Math. **340**, Springer, Berlin, 1973.

[Denef and Vercauteren 2002] J. Denef and F. Vercauteren, "An extension of Kedlaya's algorithm to Artin–Schreier curves in characteristic 2", pp. 308–323 in *Algorithmic number theory* (ANTS-V) (Sydney, 2002), edited by C. Fieker and D. R. Kohel, Lecture Notes in Comput. Sci. **2369**, Springer, Berlin, 2002.

[Dwork 1960] B. Dwork, "On the rationality of the zeta function of an algebraic variety", *Amer. J. Math.* **82** (1960), 631–648.

[Eisenbud 1995] D. Eisenbud, *Commutative algebra with a view toward algebraic geometry*, Graduate Texts in Math. **150**, Springer, New York, 1995.

[Elkies 1998] N. D. Elkies, "Elliptic and modular curves over finite fields and related computational issues", pp. 21–76 in *Computational perspectives on number theory* (Chicago, IL, 1995), edited by D. A. Buell and J. T. Teitelbaum, AMS/IP Stud. Adv. Math. **7**, Amer. Math. Soc., Providence, RI, 1998.

[Grothendieck 1968] A. Grothendieck, "Formule de Lefschetz et rationalité des fonctions L", pp. 279:1–15 in *Séminaire Bourbaki* 1964/65, North Holland, Amsterdam, 1968. Reprinted Soc. Math. France, Paris, 1995 (with 1965/66 volume).

[Katz 1971] N. M. Katz, "On a theorem of Ax", *Amer. J. Math.* **93** (1971), 485–499.

[Katz 2001] N. M. Katz, "Sums of Betti numbers in arbitrary characteristic", *Finite Fields Appl.* **7**:1 (2001), 29–44.

[Kedlaya 2001] K. S. Kedlaya, "Counting points on hyperelliptic curves using Monsky–Washnitzer cohomology", *J. Ramanujan Math. Soc.* **16**:4 (2001), 323–338. Errata in **18**:4 (2003), 417–418.

[Koblitz 1984] N. Koblitz, *p-adic numbers, p-adic analysis, and zeta-functions*, 2nd ed., Graduate Texts in Mathematics **58**, Springer, New York, 1984.

[Koblitz 1989] N. Koblitz, *Hyperelliptic cryptosystems*, vol. 1, 1989.

[Lauder 2004a] A. G. B. Lauder, "Counting solutions to equations in many variables over finite fields", *Found. Comput. Math.* **4**:3 (2004), 221–267.

[Lauder 2004b] A. G. B. Lauder, "Deformation theory and the computation of zeta functions", *Proc. London Math. Soc.* (3) **88**:3 (2004), 565–602.

[Lauder and Wan 2002] A. Lauder and D. Wan, "Computing zeta functions of Artin-Schreier curves over finite fields", pp. 34–55 , 2002.

[Lauder and Wan 2008] A. Lauder and D. Wan, "Counting points on varieties over finite fields of small characteristic", pp. 579–612 in *Surveys in algorithmic number theory*, edited by J. P. Buhler and P. Stevenhagen, Math. Sci. Res. Inst. Publ. **44**, Cambridge University Press, New York, 2008.

[Lenstra et al. 1982] A. K. Lenstra, H. W. Lenstra, Jr., and L. Lovász, "Factoring polynomials with rational coefficients", *Math. Ann.* **261**:4 (1982), 515–534.

[Matiyasevich 1993] Y. V. Matiyasevich, "Hilbert's Tenth Problem", (1993).

[Mazur 1972] B. Mazur, "Frobenius and the Hodge filtration", *Bull. Amer. Math. Soc.* **78** (1972), 653–667.

[Monsky 1970] P. Monsky, *p-adic analysis and zeta functions*, Lectures in Mathematics **4**, Kinokuniya, Tokyo, 1970.

[Odlyzko 1994] A. M. Odlyzko, "Analytic computations in number theory", pp. 451–463 in *Mathematics of computation 1943–1993: a half-century of computational mathematics* (Vancouver, 1993), edited by W. Gautschi, Proc. Sympos. Appl. Math. **48**, Amer. Math. Soc., Providence, RI, 1994.

[Pila 1990] J. Pila, "Frobenius maps of abelian varieties and finding roots of unity in finite fields", *Math. Comp.* **55**:192 (1990), 745–763.

[Poonen 1996] B. Poonen, "Computational aspects of curves of genus at least 2", pp. 283–306 in *Algorithmic number theory* (ANTS-II) (Talence, 1996), edited by H. Cohen, Lecture Notes in Comput. Sci. **1122**, Springer, Berlin, 1996.

[Satoh 2000] T. Satoh, "The canonical lift of an ordinary elliptic curve over a finite field and its point counting", *J. Ramanujan Math. Soc.* **15**:4 (2000), 247–270.

[Schoof 1985] R. Schoof, "Elliptic curves over finite fields and the computation of square roots mod p", *Math. Comp.* **44**:170 (1985), 483–494.

[Wan 1996] D. Wan, "Meromorphic continuation of L-functions of p-adic representations", *Ann. of Math.* (2) **143**:3 (1996), 469–498.

[Wan 1999] D. Wan, "Computing zeta functions over finite fields", pp. 131–141 in *Finite fields: theory, applications, and algorithms* (Waterloo, ON, 1997), edited by R. C. Mullin and G. L. Mullen, Contemp. Math. **225**, Amer. Math. Soc., Providence, RI, 1999.

[Wan 2004] D. Wan, "Variation of p-adic Newton polygons for L-functions of exponential sums", *Asian J. Math.* **8**:3 (2004), 427–471.

[Weil 1949] A. Weil, "Numbers of solutions of equations in finite fields", *Bull. Amer. Math. Soc.* **55** (1949), 497–508.

DAQING WAN
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA
IRVINE, CA 92697-3875
UNITED STATES
dwan@math.uci.edu