A differential approach to computing zeta functions over finite fields

Kiran S. Kedlaya

Department of Mathematics, Massachusetts Institute of Technology

AMS-SMM Joint Meeting
Zacatecas, May 25, 2007

Contents

Contents

The Riemann zeta function

For $\mathrm{Real}(s) > 1$, put $\zeta(s) = \sum_{n=1}^{\infty} n^{-s} = \prod_p (1 - p^{-s})^{-1}$. (E.g., by Euler, $\zeta(2) = \pi^2/6$.)

The Riemann zeta function

For $\mathrm{Real}(s) > 1$, put $\zeta(s) = \sum_{n=1}^{\infty} n^{-s} = \prod_p (1-p^{-s})^{-1}$. (E.g., by Euler, $\zeta(2) = \pi^2/6$.)

Theorem (Riemann, Hadamard, de la Vallée Poussin)

The function $\zeta(s)$ extends to a meromorphic function on $\mathbb{C}$, with a simple pole at $s = 1$ and no other poles. Moreover, $\zeta(s) \neq 0$ for $\mathrm{Real}(s) \geq 1$.

This implies the prime number theorem:

$$\{\#\text{ of primes} \leq x\} \sim \frac{x}{\log x}.$$

The Riemann zeta function

For $\mathrm{Real}(s) > 1$, put $\zeta(s) = \sum_{n=1}^{\infty} n^{-s} = \prod_p (1-p^{-s})^{-1}$. (E.g., by Euler, $\zeta(2) = \pi^2/6$.)

Theorem (Riemann, Hadamard, de la Vallée Poussin)

The function $\zeta(s)$ extends to a meromorphic function on $\mathbb{C}$, with a simple pole at $s = 1$ and no other poles. Moreover, $\zeta(s) \neq 0$ for $\mathrm{Real}(s) \geq 1$.

This implies the prime number theorem:

$$\{\#\text{ of primes} \leq x\} \sim \frac{x}{\log x}.$$

Conjecture (Riemann)

Other than $s = -2, -4, \ldots$, the zeroes of ζ occur on the line $\mathrm{Real}(s) = 1/2$.

Counting solutions modulo p: an unrelated problem?

Given a system of polynomial equations with integer coefficients, one may ask how many solutions it has modulo p.

Counting solutions modulo p: an unrelated problem?

Given a system of polynomial equations with integer coefficients, one may ask how many solutions it has modulo p.

Example

For every prime $p > 2$, the equation $x^2 - y^2 \equiv 1 \pmod{p}$ has $p - 1$ solutions.

Counting solutions modulo p: an unrelated problem?

Given a system of polynomial equations with integer coefficients, one may ask how many solutions it has modulo p.

Example

For every prime $p > 2$, the equation $x^2 - y^2 \equiv 1 \pmod{p}$ has $p - 1$ solutions.

Example

The number of solutions of $x^3 + y^3 \equiv 1 \pmod{p}$ was found by Gauss; for $p \equiv 1 \pmod{3}$, it can be expressed in terms of a solution of $a^2 + 3b^2 = p$.

Zeta functions of algebraic varieties

Definition (Weil)

For X an algebraic variety over $\mathbb{F}_p$, its *zeta function* is the formal power series

$$\zeta_X(t) = \exp\left(\sum_{n=1}^{\infty} \#X(\mathbb{F}_{p^n})\frac{t^n}{n}\right),$$

where $X(\mathbb{F}_{p^n})$ is the set of points of X with coordinates in the finite field $\mathbb{F}_{p^n}$.

More generally, we can start with a variety over $\mathbb{F}_q$ for q a power of p, then count points over $\mathbb{F}_{q^n}$ for all n. (Note that $\mathbb{F}_q \neq \mathbb{Z}/q\mathbb{Z}$ if $q \neq p$; that would give the *Igusa zeta function* instead.)

An example

Example

If $p > 2$, and X is defined in the plane by the equation $x^2 - y^2 = 1$, then $\#X(\mathbb{F}_{p^n}) = p^n - 1$, so

$$\zeta_X(t) = \exp\left(\sum_{n=1}^{\infty} \frac{(p^n - 1)t^n}{n}\right) = \frac{1-t}{1-pt}.$$

Relationship with Riemann's construction

To better see the analogy with Riemann, rewrite

$$\zeta_X(p^{-s}) = \prod_x (1 - p^{-n(x)s})^{-1},$$

where x runs over Galois orbits of $\overline{\mathbb{F}_p}$-rational points of X, and $n(x)$ is the smallest n such that x is defined over $\mathbb{F}_{p^n}$.

Relationship with Riemann's construction

To better see the analogy with Riemann, rewrite

$$\zeta_X(p^{-s}) = \prod_x (1 - p^{-n(x)s})^{-1},$$

where x runs over Galois orbits of $\overline{\mathbb{F}_p}$-rational points of X, and $n(x)$ is the smallest n such that x is defined over $\mathbb{F}_{p^n}$.

Handy corollary: if X is the disjoint union of Y and Z, then

$$\zeta_X(t) = \zeta_Y(t)\zeta_Z(t).$$

Zeta functions of algebraic varieties (contd.)

The following is analogous to Riemann's theorem.

Theorem (Dwork, Grothendieck)

The series $\zeta_X(t)$ represents a rational function of t with integer coefficients.

There is also an analogue of the Riemann hypothesis, but in this case it is a theorem of Deligne.

Contents

Abelian groups in cryptography

There are several techniques in cryptography based on the use of a "generic" abelian group G. For such a group, it should be easy to write a computer program to compute $A+B$ (and $-A$) from A, B, but it should be hard to take *discrete logarithms*: if $B = nA$ for some integer n, it should be hard to recover n from A, B.

Abelian groups in cryptography

There are several techniques in cryptography based on the use of a "generic" abelian group G. For such a group, it should be easy to write a computer program to compute $A+B$ (and $-A$) from A, B, but it should be hard to take *discrete logarithms*: if $B = nA$ for some integer n, it should be hard to recover n from A, B.

Example (Diffie-Hellman)

Alice and Bob wish to agree on a secret password, but have no way to communicate securely. They agree (in public) on an abelian group G and an element $P \in G$. Alice and Bob secretly pick random numbers a, b, and reveal (in public) aP, bP. The secret password is then abP, but an onlooker only sees P, aP, bP.

Suitability of groups for cryptography

Suitability of groups for cryptography

If $\#G = rs$ and $\gcd(r,s) = 1$, we can reduce discrete logarithms in G to discrete logarithms in two groups, of orders r and s. So for best results, the order of G should be almost prime, i.e., it should have a large prime factor.

Suitability of groups for cryptography

If $\#G = rs$ and $\gcd(r,s) = 1$, we can reduce discrete logarithms in G to discrete logarithms in two groups, of orders r and s. So for best results, the order of G should be almost prime, i.e., it should have a large prime factor.

A bad example would be the additive group $\mathbb{F}_p$; one can take discrete logarithms by Euclid's algorithm. A better example is the multiplicative group $\mathbb{F}_p^*$, but it is not ideal either; there is a better than exhaustive algorithm for finding discrete logarithms (number field sieve).

Algebraic curves and cryptography

Instead, let C be a smooth plane cubic curve (an *elliptic curve*) over $\mathbb{F}_q$, e.g.,

$$y^2 = x^3 + x + 1.$$

(The right side could instead be any cubic polynomial with no repeated roots.) Then the set of $\mathbb{F}_q$-rational points of C (in the projective plane) forms a group.

Algebraic curves and cryptography

Instead, let C be a smooth plane cubic curve (an *elliptic curve*) over $\mathbb{F}_q$, e.g.,

$$y^2 = x^3 + x + 1.$$

(The right side could instead be any cubic polynomial with no repeated roots.) Then the set of $\mathbb{F}_q$-rational points of C (in the projective plane) forms a group.

More generally, if C is a smooth, projective, geometrically irreducible curve over $\mathbb{F}_q$, there is a natural group variety containing C, the *Jacobian* $J(C)$, whose $\mathbb{F}_q$-rational points form a group. In the previous case, this coincides with C itself.

Zeta functions and group orders

Form of the zeta function for curves

Let C be a smooth, projective, geometrically irreducible curve of genus g over $\mathbb{F}_q$. (For instance, an elliptic curve has genus 1.) Then

$$\zeta_C(t) = \frac{P(t)}{(1-t)(1-qt)}$$

with P a polynomial of degree $2g$, whose roots in $\mathbb{C}$ lie on the circle $|z| = q^{-1/2}$. The group $J(C)(\mathbb{F}_q)$ has order $P(1)$.

Zeta functions and group orders

Form of the zeta function for curves

Let C be a smooth, projective, geometrically irreducible curve of genus g over $\mathbb{F}_q$. (For instance, an elliptic curve has genus 1.) Then

$$\zeta_C(t) = \frac{P(t)}{(1-t)(1-qt)}$$

with P a polynomial of degree $2g$, whose roots in $\mathbb{C}$ lie on the circle $|z| = q^{-1/2}$. The group $J(C)(\mathbb{F}_q)$ has order $P(1)$.

Consequently, for a given C, if we can compute ζ_C, we can then tell whether $\#J(C)(\mathbb{F}_q)$ has a large prime factor.

Some strategies I won't discuss

There are several useful strategies for computing ζ_C that I won't have time to focus on in this talk, so I mention them now.

- Count $\#C(\mathbb{F}_{q^n})$ for $n = 1, \ldots, g$. This is only good for small examples.
- Shanks's method: fix an element P of $J(C)$ and try to find integers m, n such that $mP = nP$ ("birthday paradox").
- Schoof's method: compute ζ_C modulo a small auxiliary prime ℓ, by finding the ℓ-torsion points of $J(C)(\overline{\mathbb{F}_q})$. Repeat enough times, apply Chinese remainder theorem.
- Build a quantum computer. (This would also solve discrete logarithms.)

Contents

The problem at hand

For ease of exposition, I will restrict to the following class of examples. Assume $q = p \neq 2$, and let C be the curve

$$y^2 = P(x)$$

in the projective plane over $\mathbb{F}_p$, where $P(x)$ is a monic polynomial of degree $2g+1$ with no repeated roots. This is a *hyperelliptic curve* of genus g.

Cohomology and zeta functions

Let X be an algebraic variety over $\mathbb{F}_p$. One often studies ζ_X by constructing a *cohomology theory* associating to X some vector spaces $H^i(X)$ over some field K, each equipped with a linear transformation F such that

$$\#X(\mathbb{F}_{p^n}) = \sum_i (-1)^i \operatorname{Trace}(F^n, H^i(X)).$$

Then

$$\zeta_X(T) = \prod_i \det(1 - tF, H^i(X))^{(-1)^{i+1}}.$$

This is similar to the Lefschetz fixed point formula in topology (Weil's analogy): the points of X over $\mathbb{F}_{p^n}$ are the fixed points of the n-th power of the *Frobenius map* which on each coordinate acts as $x \mapsto x^p$.

Cohomology theories

The most famous cohomology theory that can be used to study zeta functions is *étale cohomology* (Grothendieck et al.). It is the most well-developed for theoretical purposes, but it is mostly useless for numerical computations.

Cohomology theories

The most famous cohomology theory that can be used to study zeta functions is *étale cohomology* (Grothendieck et al.). It is the most well-developed for theoretical purposes, but it is mostly useless for numerical computations.

We use *Monsky-Washnitzer (MW) cohomology*, a cohomology theory inspired by the cohomology of differential forms (de Rham cohomology). This is harder to develop in theory, but much easier to compute in practice.

p-adic numbers

The coefficient field of MW cohomology (when $q = p$) is the field $\mathbb{Q}_p$ of *p-adic numbers*, which are "left-infinite base p expansions". For instance, in $\mathbb{Q}_2$,

$$(\cdots 111) + 1 = 0.$$

Just like real numbers, you cannot manipulate true arbitrary p-adic numbers on a computer, because you can only keep finitely many digits.

p-adic numbers

The coefficient field of MW cohomology (when $q = p$) is the field $\mathbb{Q}_p$ of *p-adic numbers*, which are "left-infinite base p expansions". For instance, in $\mathbb{Q}_2$,

$$(\cdots 111) + 1 = 0.$$

Just like real numbers, you cannot manipulate true arbitrary p-adic numbers on a computer, because you can only keep finitely many digits.

Fortunately, this is no problem when computing ζ_X: in practice, you can give a bound on the size of any given coefficient of ζ_X. Given this bound, you can determine the coefficient from a sufficiently good p-adic approximation. (This is like computing a quantity known to be an integer, by computing it as a real number with an error of less than 0.5.)

p-adic cohomology and zeta functions

Note

MW cohomology is only defined for *smooth affine* varieties.

My original curve C is not affine, because it was defined in the projective plane. I need to take out a subvariety Z consisting of finitely many points. If X is what remains, then

$$\zeta_C = \zeta_X \zeta_Z.$$

Note that there is a unique point at infinity on C, with homogeneous coordinates $[0:1:0]$. I could take Z to consist of that point alone; however, it will be more convenient to take Z to consist of the point at infinity *plus* the points with y-coordinate zero.

Algebraic differential forms

The basic idea of Monsky-Washnitzer cohomology is to use algebraic differential forms. But this is a bad idea when working over a field where $p = 0$: e.g., in the polynomial ring $\mathbb{F}_p[x]$, you can't always solve the equation

$$\frac{df}{dx} = \sum_i c_i x^i$$

by setting

$$f = \sum_i \frac{c_i}{i+1} x^{i+1}.$$

Instead, we first pass from the original equation modulo p to an equation with integer coefficients, and work there.

Lifting the curve

We started with the curve $y^2 = P(x)$ over $\mathbb{F}_p$. Choose a lift $\tilde{P}$ of P to a monic polynomial of degree $2g+1$ over $\mathbb{Z}$. Then $y^2 = \tilde{P}(x)$ describes a new hyperelliptic curve $\tilde{C}$ over $\mathbb{Q}_p$, on which differential forms behave nicely.

Lifting the curve

We started with the curve $y^2 = P(x)$ over $\mathbb{F}_p$. Choose a lift $\tilde{P}$ of P to a monic polynomial of degree $2g+1$ over $\mathbb{Z}$. Then $y^2 = \tilde{P}(x)$ describes a new hyperelliptic curve $\tilde{C}$ over $\mathbb{Q}_p$, on which differential forms behave nicely.

Again, let $\tilde{X}$ be the affine curve obtained from $\tilde{C}$ by taking out the point at infinity and the points with y-coordinate 0. The ring of regular functions on $\tilde{X}$ is

$$R = \mathbb{Q}_p[x,y,z]/(y^2 - \tilde{P}(x), yz - 1).$$

Algebraic de Rham cohomology

Let Ω be the R-module generated by dx, dy modulo

$$2y\,dy - \tilde{P}'(x)\,dx.$$

Let $d : R \to \Omega$ be the $\mathbb{Q}_p$-linear derivation sending x, y to dx, dy. That is,

$$df(x,y) = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy.$$

Put $H^0(X) = \mathbb{Q}_p$. Let $H^1(X)$ be the quotient of Ω by the $\mathbb{Q}_p$-submodule generated by df for all $f \in R$.

Algebraic de Rham cohomology

Let Ω be the R-module generated by dx, dy modulo

$$2y\,dy - \tilde{P}'(x)\,dx.$$

Let $d : R \to \Omega$ be the $\mathbb{Q}_p$-linear derivation sending x, y to dx, dy. That is,

$$df(x,y) = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy.$$

Put $H^0(X) = \mathbb{Q}_p$. Let $H^1(X)$ be the quotient of Ω by the $\mathbb{Q}_p$-submodule generated by df for all $f \in R$.

Theorem

$H^1(X)$ is a vector space over $\mathbb{Q}_p$ with basis

$$\frac{x^i\,dx}{y} \quad (i = 0, \ldots, 2g-1), \qquad \frac{x^i\,dx}{y^2} \quad (i = 0, \ldots, 2g).$$

Computing in algebraic de Rham cohomology

Just having a basis for $H^1(X)$ is not enough. We must also be able to write any element of Ω as a linear combination of basis elements plus some df. Fortunately, this is not difficult.

Computing in algebraic de Rham cohomology

Just having a basis for $H^1(X)$ is not enough. We must also be able to write any element of Ω as a linear combination of basis elements plus some df. Fortunately, this is not difficult.

Example

Start with $\frac{f(x)\,dx}{y^m}$ for some $m > 2$. Since $\tilde{P}$ has no repeated roots, we can write $f(x) = f_1(x)\tilde{P}(x) + f_2(x)\tilde{P}'(x)$ for some f_1, f_2. On one hand, in Ω,

$$\frac{f_1(x)\tilde{P}(x)\,dx}{y^m} = \frac{f_1(x)\,dx}{y^{m-2}}$$

in Ω; on the other hand, in $H^1(X)$, $d(2f_2(x)/((m-2)y^{m-2})) = 0$, so

$$\frac{f_2(x)\tilde{P}'(x)\,dx}{y^m} = \frac{2f_2(x)\,dy}{y^{m-1}} \equiv \frac{2f_2'(x)\,dx}{(m-2)y^{m-2}}.$$

The action of Frobenius

Remember that we need not just vector spaces $H^0(X), H^1(X)$, but also maps F on these. They come from lifting the Frobenius map on X.

The action of Frobenius

Remember that we need not just vector spaces $H^0(X), H^1(X)$, but also maps F on these. They come from lifting the Frobenius map on X.

To compute the action of F on a 1-form, substitute

$$x \mapsto x^p$$

$$y \mapsto y^p \left(1 + p\frac{\tilde{P}(x^p) - \tilde{P}(x)^p}{py^{2p}}\right)^{1/2},$$

where the last expression is expanded as an infinite series. Then rewrite each term of the series in terms of a basis of H^1.

The action of Frobenius

Remember that we need not just vector spaces $H^0(X), H^1(X)$, but also maps F on these. They come from lifting the Frobenius map on X.

To compute the action of F on a 1-form, substitute

$$x \mapsto x^p$$

$$y \mapsto y^p \left(1 + p\frac{\tilde{P}(x^p) - \tilde{P}(x)^p}{py^{2p}}\right)^{1/2},$$

where the last expression is expanded as an infinite series. Then rewrite each term of the series in terms of a basis of H^1.

This is an infinite process, but we only want finitely many digits of p-adic accuracy. One can truncate at a certain point without losing any of this accuracy.

The conclusion

Theorem (Monsky)

With $H^0(X), H^1(X), F$ defined as above,

$$\#X(\mathbb{F}_{p^n}) = \sum_{i=0}^{1} (-1)^i \operatorname{Trace}((pF^{-1})^n, H^i(X)).$$

So we can recover the zeta function of X, and hence of the original curve C.

Contents

Non-prime base fields

The above description required $q = p$, but cases $q \neq p$ are much more interesting in applications: the complexity of the calculation depends much more strongly on p than on q.

We also excluded $p = 2$, but a variation in that case is possible (Denef-Vercauteren).

For p small, it is reasonable to perform this calculation even if q has several hundred digits!

Even in the prime case...

In the case $q = p$, a straightforward implementation is no faster than counting points directly: both are $O(p)$.

Recent work of David Harvey improves this to $O(p^{1/2})$, so one can compute zeta functions in some examples where $p \sim 10^{15}$.

This may have some applications to computing L-functions associated to algebraic curves, in order to investigate, e.g., the Birch-Swinnerton-Dyer conjecture.

Other varieties

One can consider many other classes of curves (Castryck-Denef-Vercauteren), or even higher-dimensional varieties (Abbott-K-Roe, de Jong, Lauder).

Ideas from differential geometry (parallel transport) are also helpful (Gerkmann, Hubrechts, Lauder). This has attracted some interest among physicists interested in mirror symmetry.

The end

These slides will be available online at

`http://math.mit.edu/~kedlaya/papers`.