Chapter One

Introduction, main results, context

by Bas Edixhoven

1.1 STATEMENT OF THE MAIN RESULTS

As the final results in this book are about fast computation of coefficients of modular forms, we start by describing the state of the art in this subject.

A convenient way to view modular forms and their coefficients in this context is as follows, in terms of Hecke algebras. For N and k positive integers, let $S_k(\Gamma_1(N))$ be the finite dimensional complex vector space of cusp forms of weight k on the congruence subgroup $\Gamma_1(N)$ of $\mathrm{SL}_2(\mathbb{Z})$. Each f in $S_k(\Gamma_1(N))$ has a power series expansion $f = \sum_{n\geq 1} a_n(f)q^n$, a complex power series converging on the open unit disk. These $a_n(f)$ are the coefficients of f that we want to compute, in particular for large n. For each positive integer n we have an endomorphism T_n of $S_k(\Gamma_1(N))$, and we let $\mathbb{T}(N,k)$ denote the sub-$\mathbb{Z}$-algebra of $\mathrm{End}(S_k(\Gamma_1(N))$ generated by them. The $\mathbb{T}(N,k)$ are commutative, and free $\mathbb{Z}$-modules of rank the dimension of $S_k(\Gamma_1(N))$, which is of polynomially bounded growth in N and k. For each N, k, and n one has the identity $a_n(f) = a_1(T_n f)$. The $\mathbb{C}$-valued pairing between $S_k(\Gamma_1(N))$ and $\mathbb{T}(N,k)$ given by $(f,t) \mapsto a_1(tf)$ identifies $S_k(\Gamma_1(N))$ with the space of $\mathbb{Z}$-linear maps from $\mathbb{T}(N,k)$ to $\mathbb{C}$, and we can write $f(T_n)$ for $a_n(f)$. All together this means that the key to the computation of coefficients of modular forms is the computation of the Hecke algebras $\mathbb{T}(N,k)$ and their elements T_n. A modular form f in $S_k(\Gamma_1(N))$ is determined by the $f(T_i)$ with $i \leq k\cdot[\mathrm{SL}_2(\mathbb{Z}) : \Gamma_1(N)]/12$, hence if T_n is known as a $\mathbb{Z}$-linear combination of these T_i, then $f(T_n)$ can be computed as the same $\mathbb{Z}$-linear combination of the $f(T_i)$.

The state of the art in computing the algebras $\mathbb{T}(N,k)$ can now be summarized as follows.

> *There is a deterministic algorithm that on input positive integers N and $k \geq 2$, computes $\mathbb{T}(N,k)$: it gives a $\mathbb{Z}$-basis and the multiplication table for this basis, in running time polynomial in N and k. Moreover, the Hecke operator T_n can be expressed in this $\mathbb{Z}$-basis in deterministic polynomial time in N, k and n.*

We do not know a precise reference for this statement, but it is rather obvious from the literature on calculations with modular forms for which we refer to William Stein's book [Ste2], and in particular to Section 8.10.2 of it. The algorithms alluded to above use that $S_k(\Gamma_1(N))$, viewed as $\mathbb{R}$-vector space, is naturally isomorphic to the $\mathbb{R}$-vector space obtained from the so-called "cuspidal subspace":

$$\mathrm{H}^1(\Gamma_1(N), \mathbb{Z}[x,y]_{k-2})_{\mathrm{cusp}} \subset \mathrm{H}^1(\Gamma_1(N), \mathbb{Z}[x,y]_{k-2})$$

in group cohomology. Here, $\mathbb{Z}[x,y]_{k-2}$ is the homogeneous part of degree $k-2$ of the polynomial ring $\mathbb{Z}[x,y]$ on which $\mathrm{SL}_2(\mathbb{Z})$ acts via its standard representation on $\mathbb{Z}[x,y]_1$. In this way, $\mathrm{H}^1(\Gamma_1(N), \mathbb{Z}[x,y]_{k-2})_{\mathrm{cusp}}$, modulo its torsion subgroup, is a free $\mathbb{Z}$-module of finite rank that is a faithful $\mathbb{T}(N,k)$-module, and the action of the T_n is described explicitly. The algorithms then use a presentation of $\mathrm{H}^1(\Gamma_1(N), \mathbb{Z}[x,y]_{k-2})_{\mathrm{cusp}}$ in terms of so-called "modular symbols" and we call them therefore modular symbols algorithms. The theory of modular symbols was developed by Birch, Manin, Shokurov, Cremona, Merel, ... It has led to many algorithms, implementations, and calculations, which together form the point of departure for this book.

The computation of the element T_n of $\mathbb{T}(N,k)$, using modular symbols algorithms, involves sums in which the number of terms grows at least linearly in n. If one computes such sums by evaluating and adding the terms one by one, the computation of T_n, for N and k fixed, will take time at least linear in n, and hence exponential in $\log n$. The same is true for other methods for computing T_n that we know of: computations with q-expansions that involve multiplication of power series, using linear combinations of

theta series, the "graph method" of Mestre and Oesterlé, and the Lefschetz trace formula for correspondences, holomorphic or not. Efforts to evaluate the encountered sums more quickly seem to lead, in each case, to the problem of computing coefficients of modular forms. For example, the graph method leads to the problem of quickly computing representation numbers of integer quadratic forms in 4 variables. In the case of the trace formula, there are maybe only $O(\sqrt{n})$ terms, but they contain class numbers of imaginary quadratic orders, these numbers being themselves directly related to coefficients of modular forms of half integral weight.

Let us now state one of the main results in this book, Theorem 15.2.1.

> *Assume that the generalized Riemann hypothesis (GRH) for zeta functions of number fields holds. There exists a deterministic algorithm that on input positive integers n and k, together with the factorization of n into prime factors, computes the element T_n of $\mathbb{T}(1,k)$ in running time polynomial in k and* $\log n$.

The restriction to modular forms of level 1 in this result is there for a technical reason. The result will certainly be generalized to much more general levels; see the Epilogue. The condition that the factorization of n into primes must be part of the input is necessary because we do not have a polynomial time algorithm for factoring integers. Conversely, see Remark 2.2.4 for evidence that factoring is not harder than computing coefficients of modular forms.

Let us describe how the computation of Galois representations is used for the computation of T_n. Standard identities express T_n in terms of the T_p for p dividing n. These T_p are computed, via the LLL basis reduction algorithm, from sufficiently many of their images under morphisms f from $\mathbb{T}(1,k)$ to finite fields, analogously to Schoof's algorithm for counting points of an elliptic curve over a finite field. Indeed, for such an f from $\mathbb{T}(1,k)$ to $\mathbb{F}$, with p not the characteristic, l, say, of $\mathbb{F}$, the image $f(T_p)$ is equal to the trace of $\rho_f(\mathrm{Frob}_p)$, where $\rho_f\colon \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(\mathbb{F})$ is the Galois representation attached to f, and $\rho_f(\mathrm{Frob}_p)$ a Frobenius element at p. The representation ρ_f is characterized by the following three conditions: it is semisimple, it is unramified outside l, and for all prime numbers $p \neq l$

one has

$$\text{trace}(\rho_f(\text{Frob}_p)) = f(T_p) \quad \text{and} \quad \det(\rho_f(\text{Frob}_p)) = p^{k-1} \quad \text{in } \mathbb{F}.$$

It is *the* main result of this book, Theorem 14.1.1, plus some standard computational number theory, that enables us to compute $\rho_f(\text{Frob}_p)$ in time polynomial in k, $\#\mathbb{F}$, and $\log p$ (note the log!). Under GRH, existence of sufficiently many maximal ideals of small enough index is guaranteed. We partly quote Theorem 14.1.1:

> *There is a deterministic algorithm that on input a positive integer k, a finite field $\mathbb{F}$, and a surjective ring morphism f from $\mathbb{T}(1,k)$ to $\mathbb{F}$ such that the associated Galois representation $\rho_f\colon \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \text{GL}_2(\mathbb{F})$ is reducible or has image containing $\text{SL}_2(\mathbb{F})$, computes ρ_f in time polynomial in k and $\#\mathbb{F}$.*

By "computing ρ_f" we mean the following. Let $K_f \subset \overline{\mathbb{Q}}$ be the finite Galois extension such that ρ_f factors as the natural surjection from $\text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ to $\text{Gal}(K_f/\mathbb{Q})$, followed by an injection into $\text{GL}_2(\mathbb{F})$. Then to give ρ_f means to give K_f as $\mathbb{Q}$-algebra, in terms of a multiplication table with respect to a $\mathbb{Q}$-basis, together with a list of all elements of $\text{Gal}(K_f/\mathbb{Q})$, as matrices with coefficients in $\mathbb{Q}$, and, for each σ in $\text{Gal}(K_f/\mathbb{Q})$, to give the corresponding element $\rho_f(\sigma)$ of $\text{GL}_2(\mathbb{F})$.

Before we describe in more detail, in the next sections, some history and context concerning our main results, we give one example and make some brief remarks. Many of these remarks are treated with more detail farther on.

The first nontrivial example is given by $k = 12$. The space of cuspidal modular forms of level one and weight 12 is one-dimensional, generated by the *discriminant modular form* Δ, whose coefficients are given by Ramanujan's τ-function:

$$\Delta = q\prod_{n\geq 1}(1-q^n)^{24} = \sum_{n\geq 1}\tau(n)q^n = q - 24q^2 + 252q^3 + \cdots \quad \text{in } \mathbb{Z}[[q]].$$

In this case, the Hecke algebra $\mathbb{T}(1,12)$ is the ring $\mathbb{Z}$, and, for each n in $\mathbb{Z}_{>0}$, we have $T_n = \tau(n)$. The results above mean that

> *for p prime, Ramanujan's $\tau(p)$ can be computed in time polynomial in* $\log p$.

For l prime, let ρ_l denote the Galois representation to $\mathrm{GL}_2(\mathbb{F}_l)$ attached to Δ. It was proved by Swinnerton-Dyer that for l not in $\{2, 3, 5, 7, 23, 691\}$ the image of ρ_l contains $\mathrm{SL}_2(\mathbb{F}_l)$. This means that for all l not in this short list the representation ρ_l has nonsolvable image, and so cannot be computed using computational class field theory. The classical congruences for Ramanujan's τ-function correspond to the l in the list above. Our results provide a generalization of these congruences in the sense that the number fields K_l that give the ρ_l "encode" the $\tau(p) \bmod l$ in such a way that $\tau(p) \bmod l$ can be computed in time polynomial in l and $\log p$, that is, just the same complexity as in the case where one has explicit congruences.

More generally, we hope that nonsolvable global field extensions whose existence and local properties are implied by the Langlands program can be made accessible to computation and so become even more useful members of the society of mathematical objects. Explicit descriptions of these fields make the study of global properties such as class groups and groups of units possible. Certainly, if we only knew the maximal Abelian extension of $\mathbb{Q}$ as described by general class field theory, then roots of unity would be very much welcomed.

The natural habitat for Galois representations such as the ρ_f above is that of higher degree étale cohomology with $\mathbb{F}_\ell$-coefficients of algebraic varieties over $\overline{\mathbb{Q}}$, together with the action of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. Our results provide some evidence that, also in interesting cases, such objects can be computed in reasonable time. We stress that this question is not restricted to varieties related to modular forms or automorphic forms. In fact, thinking of elliptic curves, over $\mathbb{Q}$, say, knowing that these are modular does not help for computing their number of points over finite fields: Schoof's algorithm uses algebraic geometry, not modularity.

The problem of computing étale cohomology with Galois action is clearly related to the question of the existence of polynomial time algorithms for computing the number of solutions in $\mathbb{F}_p$ of a fixed system of polynomial equations over $\mathbb{Z}$, when p varies. Our results treat this problem for the 11-dimensional variety that gives rise to Δ; see Section 1.5 for more details and also for an explicit variety of dimension 19 related to this.

The Epilogue describes a striking application of a generalization of our results to the problem of computing representation numbers of the $\mathbb{Z}^{2k}$

equipped with their standard inner products. This again is an example where there are explicit formulas only for small k, but where in general there (surely) exists an algorithm that computes such numbers as quickly as if such formulas did exist. Hence, from a computational perspective, such algorithms form a natural generalization of the finite series of formulas.

We very briefly describe the method by which we compute the ρ_f. Their duals occur in the higher degree étale cohomology of certain higher dimensional varieties, but no-one seems to know how to compute with this directly.

Via some standard methods in étale cohomology (the Leray spectral sequence, and passing to a finite cover to trivialize a locally constant sheaf of finite dimensional $\mathbb{F}_l$-vector spaces), or from the theory of congruences between modular forms, it is well known that the ρ_f are realized by subspaces V_f in the l-torsion $J_l(\overline{\mathbb{Q}})[l]$ of the Jacobian variety J_l of some modular curve X_l defined over $\mathbb{Q}$. The field K_f is then the field generated by suitable "coordinates" of the points $x \in V_f \subset J_l(\overline{\mathbb{Q}})[l]$. We are now in the more familiar situation of torsion points on Abelian varieties. But the price that we have paid for this is that the Abelian variety J_l depends on l, and that its dimension, equal to the genus of X_l, that is, equal to $(l-5)(l-7)/24$, grows quadratically with l. This makes it impossible to directly compute the $x \in V_l$ using computer algebra: known algorithms for solving systems of nonlinear polynomial equations take time exponential in the dimension, that is, exponential in l.

Instead of using computer algebra directly, Jean-Marc Couveignes suggested that we use approximations and height bounds. In its simplest form, this works as follows. Suppose that x is a rational number, $x = a/b$, with a and b in $\mathbb{Z}$ coprime. Suppose that we have an upper bound M for $\max(|a|, |b|)$. Then x is determined by any approximation $y \in \mathbb{R}$ of x such that $|y - x| < 1/2M^2$, simply because for all $x' \neq x$ with $x' = a'/b'$, where a' and b' in $\mathbb{Z}$ satisfy $\max(|a'|, |b'|) < M$, we have $|x' - x| = |(a'b - ab')/bb'| \geq 1/M^2$.

For the computation of K_f, we consider the minimal polynomial P_f in $\mathbb{Q}[T]$ of a carefully theoretically constructed generator α of K_f. We use approximations of all Galois conjugates of α, that is, of all roots of P_f. Instead of working directly with torsion points of J_l, we work with divisors on the curve X_l. Using this strategy, the problem of showing that P_f can

be computed in time polynomial in k and $\#\mathbb{F}$ is divided into two different tasks. First, to show that the number of digits necessary for a good enough approximation of P_f is bounded by a fixed power of $\#\mathbb{F}$. Second, to show that, given f and n, the coefficients of P_f can be approximated with a precision of n digits in time polynomial in $n \cdot \#\mathbb{F}$. The first problem is dealt with in Chapters 9, 10, and 11, using Arakelov geometry. The second problem is solved in Chapters 12 and 13, in two ways: complex approximations (numerical analysis), and approximations in the sense of reductions modulo many small primes, using exact computations in Jacobians of modular curves over finite fields. These five chapters form the technical heart of this book. The preceding chapters are meant as an introduction to them, or motivation for them, and the two chapters following them give the main results as relatively straightforward applications.

Chapters 6 and 7 stand a bit apart, as they are concerned with some real computations of Galois representations attached to modular forms. They use the method by complex approximations, but do not use a rigorously proved bound for sufficient accuracy. Instead, the approximations provide good *candidates* for polynomials P_f. These candidates have the correct Galois group and the right ramification properties. Recent modularity results by Khare, Wintenberger and Kisin (see [Kh-Wi1], [Kh-Wi2], [Kis1], and [Kis2]) are then applied to *prove* that the candidates do indeed give the right Galois representations.

1.2 HISTORICAL CONTEXT: SCHOOF'S ALGORITHM

The computation of Hecke operators from Galois representations and congruences can be viewed as a generalization of Schoof's method to count points on elliptic curves over finite fields; see [Sch2] and [Sch3]. René Schoof gave an algorithm to compute, for E an elliptic curve over a finite field $\mathbb{F}_q$, the number $\#E(\mathbb{F}_q)$ of $\mathbb{F}_q$-rational points in a time $O((\log q)^{5+\varepsilon})$. His algorithm works as follows.

The elliptic curve is embedded, as usual, in the projective plane $\mathbb{P}^2_{\mathbb{F}_q}$ as the zero locus of a Weierstrass equation, which, in inhomogeneous coordinates,

is of the form

$$y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6,$$

with the a_i in $\mathbb{F}_q$. We let $\mathbb{F}_q \to \overline{\mathbb{F}}_q$ be an algebraic closure. We let $\mathrm{F}_q\colon E \to E$ denote the so-called q-Frobenius. It is the endomorphism of E with the property that for all (a, b) in the affine part of $E(\overline{\mathbb{F}}_q)$ given by the Weierstrass equation above, we have $\mathrm{F}_q((a, b)) = (a^q, b^q)$. The theory of elliptic curves over finite fields says:

1. there is a unique integer a, called the *trace* of F_q, such that in the endomorphism ring of E one has $\mathrm{F}_q^2 - a\mathrm{F}_q + q = 0$;
2. $\#E(\mathbb{F}_q) = 1 - a + q$;
3. $|a| \leq 2q^{1/2}$.

So, computing $\#E(\mathbb{F}_q)$ is equivalent to computing this integer a. Schoof's idea is now to compute a modulo l for small prime numbers l. If the product of the prime numbers l exceeds $4q^{1/2}$, the length of the interval in which we know a to lie, then the congruences modulo these l determine a uniquely. Analytic number theory tells us that it will be sufficient to take all primes l up to approximately $(\log q)/2$.

Then the question is how one computes a modulo l. This should be done in time polynomial in $\log q$ and l. The idea is to use the elements of order dividing l in $E(\overline{\mathbb{F}}_q)$. We assume now that l does not divide q, that is, we avoid the characteristic of $\mathbb{F}_q$. For each l, the kernel $E(\overline{\mathbb{F}}_q)[l]$ of multiplication by l on $E(\overline{\mathbb{F}}_q)$ is a two-dimensional vector space over $\mathbb{F}_l$. The map F_q gives an endomorphism of $E(\overline{\mathbb{F}}_q)[l]$, and it follows that the image of a in $\mathbb{F}_l$ is the unique element of $\mathbb{F}_l$, also denoted a, such that for each v in $E(\overline{\mathbb{F}}_q)[l]$ we have $a\mathrm{F}_q v = \mathrm{F}_q^2 v + qv$. We remark that the image of a in $\mathbb{F}_l$ is the trace of the endomorphism of $E(\overline{\mathbb{F}}_q)[l]$ given by F_q, but this is not really used at this point.

To find this element a of $\mathbb{F}_l$, one proceeds as follows. We suppose that $l \neq 2$. There is a unique monic element ψ_l of $\mathbb{F}_q[x]$ of degree $(l^2-1)/2$, whose roots in $\overline{\mathbb{F}}_q$ are precisely the x-coordinates of the l^2-1 nonzero elements in $E(\overline{\mathbb{F}}_q)[l]$ (the rational function x on E is a degree two map to $\mathbb{P}^1_{\overline{\mathbb{F}}_q}$,

which as such is the quotient for the multiplication by -1 map on E). One then lets A_l be the $\mathbb{F}_q$-algebra obtained as

$$A_l := \mathbb{F}_q[x, y]/(y^2 + a_1xy + a_3y - x^3 - a_2x^2 - a_4x - a_6, \psi_l(x)).$$

The dimension of A_l as $\mathbb{F}_q$-vector space is l^2-1. An equivalent description of A_l is to say that it is the affine coordinate ring of the subscheme of points of order l of E. By construction of A_l, there is a tautological A_l-valued point v in $E(A_l)$ (its coordinates are the images of x and y in A_l). Now to find the element a of $\mathbb{F}_l$ that we are looking for, we then try one by one the elements i in $0, \pm 1, \ldots, \pm(l-1)/2$ until $i\mathrm{F}_q v = \mathrm{F}_q^2 v + qv$; then $i = a \bmod l$.

It is easy to see that the computation of the integer a can be done in time $O((\log q)^{5+\varepsilon})$ (using fast arithmetic for the elementary operations, for example, a multiplication in A_l costs about $(l^2(\log q))^{1+\varepsilon}$ time; $l^2(\log q)$ is the number of bits needed to store one element of A_l).

For the sake of completeness, let us mention that shortly after the appearance of Schoof's algorithm, Atkin and Elkies added some improvements to it, making it possible in certain cases to reduce the dimension of the $\mathbb{F}_q$-algebra from l^2-1 to linear in $l+1$ or $l-1$. This improvement, called the Schoof-Atkin-Elkies (SEA) algorithm, is important mainly for implementations. Its (average) complexity is $O((\log q)^{4+\varepsilon})$; for details, the reader is referred to [Sch3].

1.3 SCHOOF'S ALGORITHM DESCRIBED IN TERMS OF ÉTALE COHOMOLOGY

In order to describe Schoof's algorithm in the previous section, we referred to the theory of elliptic curves over finite fields. But there is a more general framework for getting information on the number of rational points of algebraic varieties over finite fields: cohomology, and Lefschetz's trace formula. Cohomology exists in many versions. The version directly related to Schoof's algorithm is étale cohomology with coefficients in $\mathbb{F}_l$. Standard references for étale cohomology are [SGA4], [SGA4.5], [SGA5], [Mil1], [Fr-Ki]. The reader is referred to these references for the notions that we will use below. We also recommend Appendix C of [Hart].

For the sake of precision, let us say that we define the notion of *algebraic variety over a field k* to mean *k-scheme that is separated and of finite type*. Attached to an algebraic variety X over a field k there are *étale cohomology groups with compact supports* $\mathrm{H}^i_{\mathrm{c}}(X_{\mathrm{et}}, \mathbb{F}_l)$, for all $i \geq 0$ and for all prime numbers l. Actually, the coefficients $\mathbb{F}_l$ can be replaced by more general objects, sheaves of Abelian groups on the étale site X_{et} of X, but we do not need this now. If X is a proper k-scheme, then the $\mathrm{H}^i_{\mathrm{c}}(X_{\mathrm{et}}, \mathbb{F}_l)$ are equal to the étale cohomology groups $\mathrm{H}^i(X_{\mathrm{et}}, \mathbb{F}_l)$ without condition on supports.

If k is separably closed, then the $\mathrm{H}^i_{\mathrm{c}}(X_{\mathrm{et}}, \mathbb{F}_l)$ are finite dimensional $\mathbb{F}_l$-vector spaces, zero for $i > 2\dim(X)$. In that case, they are the analog of the more easily defined cohomology groups $\mathrm{H}^i_{\mathrm{c}}(X, \mathcal{F})$ for complex analytic varieties: the derived functors of the functor that associates to a sheaf $\mathcal{F}$ of $\mathbb{Z}$-modules on X equipped with its Archimedean topology its $\mathbb{Z}$-module of global sections whose support is compact.

The construction of the $\mathrm{H}^i_{\mathrm{c}}(X_{\mathrm{et}}, \mathbb{F}_l)$ is functorial for proper morphisms: a proper morphism $f\colon X \to Y$ of algebraic varieties over k induces a pull-back morphism f^* from $\mathrm{H}^i_{\mathrm{c}}(Y_{\mathrm{et}}, \mathbb{F}_l)$ to $\mathrm{H}^i_{\mathrm{c}}(X_{\mathrm{et}}, \mathbb{F}_l)$.

Let now X be an algebraic variety over $\mathbb{F}_q$. Then we have the q-Frobenius morphism F_q from X to itself, and, by extending the base field from $\mathbb{F}_q$ to $\overline{\mathbb{F}}_q$, from $X_{\overline{\mathbb{F}}_q}$ to itself. This morphism F_q is proper, hence induces maps

$$\mathrm{F}_q^*\colon \mathrm{H}^i_{\mathrm{c}}(X_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mathbb{F}_l) \longrightarrow \mathrm{H}^i_{\mathrm{c}}(X_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mathbb{F}_l).$$

Hence, for each i in $\mathbb{Z}$, the trace $\mathrm{trace}(\mathrm{F}_q^*, \mathrm{H}^i_{\mathrm{c}}(X_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mathbb{F}_l))$ of the map is defined, and it is zero for $i < 0$ and $i > 2\dim(X)$. The set of fixed points of F_q on $X(\overline{\mathbb{F}}_q)$ is precisely the subset $X(\mathbb{F}_q)$. The *Lefschetz trace formula* then gives the following identity in $\mathbb{F}_l$:

$$\#X(\mathbb{F}_q) = \sum_i (-1)^i \mathrm{trace}(\mathrm{F}_q^*, \mathrm{H}^i_{\mathrm{c}}(X_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mathbb{F}_l)). \tag{1.3.1}$$

We can now say how Schoof's algorithm is related to étale cohomology. We consider again an elliptic curve E over a finite field $\mathbb{F}_q$. We assume that l does not divide q. Then, as for any smooth proper geometrically connected curve, $\mathrm{H}^0(E_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mathbb{F}_l) = \mathbb{F}_l$ and F_q^* acts on it as the identity, and $\mathrm{H}^2(E_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mathbb{F}_l)$ is one-dimensional and F_q^* acts on it by multiplication by q,

the degree of F_q. According to the trace formula (1.3.1), we have

$$\#E(\mathbb{F}_q) = 1 - \mathrm{trace}(\mathrm{F}_q^*, \mathrm{H}^1(E_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mathbb{F}_l)) + q.$$

It follows that for the integer a of the previous section, the trace of Frobenius, we have, for all l not dividing p the identity in $\mathbb{F}_l$,

$$a = \mathrm{trace}(\mathrm{F}_q^*, \mathrm{H}^1(E_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mathbb{F}_l)).$$

This identity is explained by the fact that there is a natural isomorphism, compatible with the action of F_q,

$$\mathrm{H}^1(E_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mathbb{F}_l) = E(\overline{\mathbb{F}}_q)[l].$$

Let us describe how one constructs this isomorphism. On E_{et} we have the short exact sequence of sheaves called the Kummer sequence:

$$0 \longrightarrow \mu_l \longrightarrow \mathbb{G}_{\mathrm{m}} \longrightarrow \mathbb{G}_{\mathrm{m}} \longrightarrow 0$$

where the map on $\mathbb{G}_{\mathrm{m}}$ is multiplication by l in the group law of $\mathbb{G}_{\mathrm{m}}$, that is, taking lth powers. This short exact sequence gives an exact sequence of cohomology groups after pullback to $E_{\overline{\mathbb{F}}_q,\mathrm{et}}$:

$$\begin{aligned}\{1\} \longrightarrow \mu_l(\overline{\mathbb{F}}_q) \longrightarrow \overline{\mathbb{F}}_q^{\times} \longrightarrow \overline{\mathbb{F}}_q^{\times} \longrightarrow \mathrm{H}^1(E_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mu_l) \\ \longrightarrow \mathrm{H}^1(E_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mathbb{G}_{\mathrm{m}}) \longrightarrow \mathrm{H}^1(E_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mathbb{G}_{\mathrm{m}}) \longrightarrow \cdots\end{aligned}$$

Just as for any scheme, one has

$$\mathrm{H}^1(E_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mathbb{G}_{\mathrm{m}}) = \mathrm{Pic}(E_{\overline{\mathbb{F}}_q}).$$

It follows that

$$\mathrm{H}^1(E_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mu_l) = \mathrm{Pic}(E_{\overline{\mathbb{F}}_q})[l].$$

Finally, using the exact sequence

$$0 \longrightarrow \mathrm{Pic}^0(E_{\overline{\mathbb{F}}_q}) \longrightarrow \mathrm{Pic}(E_{\overline{\mathbb{F}}_q}) \xrightarrow{\deg} \mathbb{Z} \longrightarrow 0$$

and the fact that E is its own Jacobian variety, that is, $\mathrm{Pic}^0(E_{\overline{\mathbb{F}}_q}) = E(\overline{\mathbb{F}}_q)$, we obtain

$$\mathrm{H}^1(E_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mu_l) = \mathrm{Pic}(E_{\overline{\mathbb{F}}_q})[l] = \mathrm{Pic}^0(E_{\overline{\mathbb{F}}_q})[l] = E(\overline{\mathbb{F}}_q)[l].$$

The choice of an isomorphism between $\mu_l(\overline{\mathbb{F}}_q)$ and $\mathbb{F}_l$ gives us the desired isomorphism between $\mathrm{H}^1(E_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mathbb{F}_l)$ and $E(\overline{\mathbb{F}}_q)[l]$. In fact, we note that by using the Weil pairing from $E(\overline{\mathbb{F}}_q)[l] \times E(\overline{\mathbb{F}}_q)[l]$ to $\mu_l(\overline{\mathbb{F}}_q)$, we get an isomorphism,

$$\mathrm{H}^1(E_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mathbb{F}_l) = E(\overline{\mathbb{F}}_q)[l]^\vee,$$

that is more natural than the one used above; in particular, it does not depend on the choice of an isomorphism $\mathbb{F}_l \to \mu_l(\overline{\mathbb{F}}_q)$.

1.4 SOME NATURAL NEW DIRECTIONS

We have seen that the two-dimensional $\mathbb{F}_l$-vector spaces that are used in Schoof's algorithm for elliptic curves can also be seen as étale cohomology groups. A natural question that arises is then the following.

> *Are there other interesting cases where étale cohomology can be used to construct polynomial time algorithms for counting rational points of varieties over finite fields?*

A more precise question is the following.

> *Let n and m be in $\mathbb{Z}_{\geq 0}$, and let $f_1, \ldots, f_m$ be in $\mathbb{Z}[x_1, \ldots, x_n]$. Is there an algorithm that on input a prime number p computes $\#\{a \in \mathbb{F}_p^n \mid \forall i : f_i(a) = 0\}$ in time polynomial in $\log p$?*

We believe that the answer to this question is yes, and that the way in which such an algorithm can work is to compute étale cohomology.

1.4.1 Curves of higher genus

The first step in the direction of this question was taken by Jonathan Pila. In [Pil] he considered principally polarized Abelian varieties of a fixed dimension, and curves of a fixed genus, and showed that in those cases polynomial time algorithms for computing the number of rational points over finite fields exist. In these cases, the only relevant cohomology groups are in degree one, that is, they are of the form $\mathrm{H}^1(X_{\overline{\mathbb{F}}_q,\mathrm{et}}, \mathbb{F}_l)$ with X a smooth proper curve, or an Abelian variety, over the field $\mathbb{F}_q$. As in Schoof's algorithm, the way to deal with these cohomology groups is to view them as

$J(\overline{\mathbb{F}}_q)[l]$, the kernel of multiplication by l on the Abelian variety J. In the case where X is a curve, one lets J be the Jacobian variety of X.

As Pila makes use of explicit systems of equations for Abelian varieties, his algorithm has a running time that is at least exponential in the dimension of the Abelian variety, and hence, in the case of curves, as a function of the genus of the curve.

The current state of the art concerning the question of counting the rational points of curves over finite fields seems still to be the same: algorithms have a running time that is exponential in the genus. As an illustration, let us mention that in [Ad-Hu] Adleman and Huang give an algorithm that computes $\#X(\mathbb{F}_q)$ in time $(\log q)^{O(g^2 \log g)}$, where X is a hyperelliptic curve over $\mathbb{F}_q$ and g is the genus of X.

Recent progress in the case where the characteristic of the finite fields $\mathbb{F}_q$ is fixed, using so-called p-adic methods, will be discussed in Section 1.6 below. In that case, there are algorithms whose running time is polynomial in g and $\log q$.

1.4.2 Higher degree cohomology, modular forms

Another direction in which one can try to generalize Schoof's algorithm is to varieties of higher dimension, where nontrivial cohomology groups of degree higher than one are needed. In this context, we would call the degree two cohomology group of a curve trivial, because the trace of F_q on it is q.

More generally speaking, cohomology groups, but now with l-adic coefficients, that are of dimension one are expected to have the property that the trace of F_q can only be of the form $q^n\zeta$, with n an integer greater than or equal to zero, and ζ a root of unity. This means that one-dimensional cohomology groups are not so challenging. Indeed, it is the fact that for elliptic curves over $\mathbb{F}_p$ all integers in the interval $[p+1-2p^{1/2}, p+1+2p^{1/2}]$ can occur that makes the problem of point counting very different from point counting on nonsingular quadric surfaces in $\mathbb{P}^3_{\mathbb{F}_q}$, for example, where the outcome can only be q^2+2q+1 or q^2+1.

It follows that the simplest case to consider is cohomology groups of dimension two, in degree at least two, on which the action of F_q is not given by a simple rule as in the one-dimensional case. Such cohomology groups

are provided by modular forms, as we will explain later in Section 2.2. Let us just say for the moment that there is a direct relation with elliptic curves, via the concept of *modularity* of elliptic curves over $\mathbb{Q}$, that we will now sketch.

Let E be an elliptic curve over $\mathbb{Q}$, given by some Weierstrass equation. Such a Weierstrass equation can be chosen to have its coefficients in $\mathbb{Z}$. A Weierstrass equation for E with coefficients in $\mathbb{Z}$ is called *minimal* if the absolute value of its *discriminant* is minimal among all Weierstrass equations for E with coefficients in $\mathbb{Z}$; this discriminant then only depends on E and will be denoted $\mathrm{discr}(E)$. In fact, two minimal Weierstrass equations define isomorphic curves in $\mathbb{P}^2_{\mathbb{Z}}$, the projective plane over $\mathbb{Z}$. In other words, E has a Weierstrass minimal model over $\mathbb{Z}$, which will be denoted by $E_{\mathbb{Z}}$. For each prime number p, we let $E_{\mathbb{F}_p}$ denote the curve over $\mathbb{F}_p$ given by reducing a minimal Weierstrass equation modulo p; it is the fiber of $E_{\mathbb{Z}}$ over $\mathbb{F}_p$. The curve $E_{\mathbb{F}_p}$ is smooth if and only if p does not divide $\mathrm{discr}(E)$. The possible singular fibers have exactly one singular point: an ordinary double point with rational tangents, or with conjugate tangents, or an ordinary cusp. The three types of reduction are called split multiplicative, nonsplit multiplicative, and additive, respectively, after the type of group law that one gets on the complement of the singular point. For each p we then get an integer a_p by requiring the following identity:

$$p + 1 - a_p = \#E(\mathbb{F}_p).$$

This means that for all p, a_p is the trace of F_p on the degree one étale cohomology of $E_{\overline{\mathbb{F}}_p}$, with coefficients in $\mathbb{F}_l$, or in $\mathbb{Z}/l^n\mathbb{Z}$ or in the l-adic numbers $\mathbb{Z}_l$. For p not dividing $\mathrm{discr}(E)$ we know that $|a_p| \leq 2p^{1/2}$. If $E_{\mathbb{F}_p}$ is multiplicative, then $a_p = 1$ or -1 in the split and nonsplit case. If $E_{\mathbb{F}_p}$ is additive, then $a_p = 0$. We also define for each p an element $\varepsilon(p)$ in $\{0, 1\}$ by setting $\varepsilon(p) = 1$ for p not dividing $\mathrm{discr}(E)$ and setting $\varepsilon(p) = 0$ for p dividing $\mathrm{discr}(E)$. The *Hasse-Weil L-function* of E is then defined as

$$L_E(s) = \prod_p L_{E,p}(s), \qquad L_{E,p}(s) = \left(1 - a_p p^{-s} + \varepsilon(p)\, p p^{-2s}\right)^{-1},$$

for s in $\mathbb{C}$ with $\Re(s) > 3/2$ (indeed, the fact that $|a_p| \leq 2p^{1/2}$ implies that the product converges for such s). To explain this function more conceptu-

ally, we note that for all p and for all $l \neq p$ we have the identity

$$1 - a_p t + \varepsilon(p) p t^2 = \det(1 - t \mathrm{F}_p^*, \mathrm{H}^1(E_{\overline{\mathbb{F}},\mathrm{et}}, \mathbb{Q}_l)).$$

The reader should notice that now we use étale cohomology with coefficients in $\mathbb{Q}_l$, the field of l-adic numbers, and not in $\mathbb{F}_l$. The reason for this is that we want the last identity above to be an identity between polynomials with integer coefficients, and not with coefficients in $\mathbb{F}_l$.

The function L_E was conjectured to have a holomorphic continuation over all of $\mathbb{C}$, and to satisfy a certain precisely given functional equation relating the values at s and $2-s$. In that functional equation appears a certain positive integer N_E called the *conductor* of E, composed of the primes p dividing $\mathrm{discr}(E)$ with exponents that depend on the behavior of E at p, that is, on $E_{\mathbb{Z}_p}$. This conjecture on continuation and functional equations was proved for semistable E (that is, E such that there is no p where E has additive reduction) by Wiles and Taylor-Wiles, and in the general case by Breuil, Conrad, Diamond, and Taylor; see [Edi2] for an overview of this. In fact, the continuation and functional equation are direct consequences of the modularity of E that was proved by Wiles, Taylor-Wiles and so on (see below). The weak Birch and Swinnerton-Dyer conjecture says that the dimension of the $\mathbb{Q}$-vector space $\mathbb{Q} \otimes E(\mathbb{Q})$ is equal to the order of vanishing of L_E at 1. Anyway, the function L_E gives us integers a_n for all $n \geq 1$ as follows:

$$L_E(s) = \sum_{n \geq 1} a_n n^{-s} \quad \text{for } \Re(s) > 3/2.$$

From these a_n one can then consider the function

$$f_E : \mathbb{H} = \{z \in \mathbb{C} \mid \Im(z) > 0\} \to \mathbb{C}, \quad z \mapsto \sum_{n \geq 1} a_n e^{2\pi i n z}.$$

Equivalently, we have

$$f_E = \sum_{n \geq 1} a_n q^n, \quad \text{with} \quad q : \mathbb{H} \to \mathbb{C}, \quad z \mapsto e^{2\pi i z}.$$

A more conceptual way to state the relation between L_E and f_E is to say that L_E is obtained, up to elementary factors, as the *Mellin transform* of f_E:

$$\int_0^\infty f_E(it) t^s \frac{dt}{t} = (2\pi)^{-s} \Gamma(s) L_E(s) \quad \text{for } \Re(s) > 3/2.$$

After all these preparations, we can finally state what the modularity of E means:

> *f_E is a modular form of weight two for the congruence subgroup $\Gamma_0(N_E)$ of $\mathrm{SL}_2(\mathbb{Z})$.*

For some more details on the concept of modular forms we refer to Section 2.2. At this moment, we just want to say that the last statement means that f_E has, as Mazur says in Singh's BBC documentary on Wiles's proof of Fermat's Last Theorem, an enormous amount of symmetry. This symmetry is with respect to the action of $\mathrm{GL}_2(\mathbb{Q})^+$, the group of invertible 2-by-2 matrices with coefficients in $\mathbb{Q}$ whose determinant is positive, on the upper half-plane $\mathbb{H}$. This symmetry gives, by Mellin transformation, the functional equation of L_E. Conversely, it had been proved in [Wei1] by Weil that if sufficiently many twists of L_E by Dirichlet characters satisfy the conjectured holomorphic continuation and functional equation, then f_E is a modular form of the type mentioned.

We now remark that Schoof's algorithm implies that, for p prime, the coefficient a_p in the q-expansion of $f_E = \sum_{n\geq 1} a_n q^n$ can be computed in time polynomial in $\log p$. One of the aims of this book is to generalize this last fact to certain modular forms of higher weight. Before we give precise definitions in Section 2.2, we will discuss a typical case in the next section.

1.5 MORE HISTORICAL CONTEXT: CONGRUENCES FOR RAMANUJAN'S τ-FUNCTION

References for this section are the articles [Ser2], [Swi], and [Del1] by Serre, Swinnerton-Dyer, and Deligne.

A typical example of a modular form of weight higher than two is the discriminant modular form, usually denoted Δ. One way to view Δ is as the holomorphic function on the upper half-plane $\mathbb{H}$ given by

$$\Delta = q \prod_{n\geq 1} (1-q^n)^{24}, \tag{1.5.1}$$

where q is the function from $\mathbb{H}$ to $\mathbb{C}$ given by $z \mapsto \exp(2\pi i z)$. The coeffi-

cients in the power series expansion

$$\Delta = \sum_{n\geq 1} \tau(n) q^n \tag{1.5.2}$$

define the famous *Ramanujan τ-function*.

To say that Δ is a modular form of weight 12 for the group $\mathrm{SL}_2(\mathbb{Z})$ means that for all elements $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$ of $\mathrm{SL}_2(\mathbb{Z})$ the following identity holds for all z in $\mathbb{H}$:

$$\Delta\left(\frac{az+b}{cz+d}\right) = (cz+d)^{12}\Delta(z), \tag{1.5.3}$$

which is equivalent to saying that the multidifferential form $\Delta(z)(dz)^{\otimes 6}$ is invariant under the action of $\mathrm{SL}_2(\mathbb{Z})$. As $\mathrm{SL}_2(\mathbb{Z})$ is generated by the elements $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, it suffices to check the identity in (1.5.3) for these two elements. The fact that Δ is q times a power series in q means that Δ is a *cusp form*: it vanishes at "$q = 0$". It is a fact that Δ is the first example of a nonzero cusp form for $\mathrm{SL}_2(\mathbb{Z})$: there is no nonzero cusp form for $\mathrm{SL}_2(\mathbb{Z})$ of weight smaller than 12, that is, there are no nonzero holomorphic functions on $\mathbb{H}$ satisfying (1.5.3) with the exponent 12 replaced by a smaller integer, whose Laurent series expansion in q is q times a power series. Moreover, the $\mathbb{C}$-vector space of such functions of weight 12 is one-dimensional, and hence Δ is a basis of it.

The one-dimensionality of this space has as a consequence that Δ is an eigenform for certain operators on this space, called *Hecke operators*, that arise from the action on $\mathbb{H}$ of $\mathrm{GL}_2(\mathbb{Q})^+$, the subgroup of $\mathrm{GL}_2(\mathbb{Q})$ of elements whose determinant is positive. This fact explains that the coefficients $\tau(n)$ satisfy certain relations that are summarized by the following identity of Dirichlet series (converging for $\Re(s) \gg 0$, for the moment, or just formal series, if one prefers that):

$$L_\Delta(s) := \sum_{n\geq 1} \tau(n) n^{-s} = \prod_p (1 - \tau(p)p^{-s} + p^{11}p^{-2s})^{-1}, \tag{1.5.4}$$

where the product is over all prime numbers. These relations

$$\begin{aligned} &\tau(mn) = \tau(m)\tau(n) && \text{if } \gcd(m,n) = 1, \\ &\tau(p^n) = \tau(p^{n-1})\tau(p) - p^{11}\tau(p^{n-2}) && \text{if } p \text{ is prime and } n \geq 2, \end{aligned} \tag{1.5.5}$$

were conjectured by Ramanujan, and proved by Mordell. Using these identities, $\tau(n)$ can be expressed in terms of the $\tau(p)$ for p dividing n.

As L_Δ is the Mellin transform of Δ, L_Δ is holomorphic on $\mathbb{C}$, and satisfies the functional equation (Hecke)

$$(2\pi)^{-(12-s)}\Gamma(12-s)L_\Delta(12-s) = (2\pi)^{-s}\Gamma(s)L_\Delta(s).$$

The famous *Ramanujan conjecture* states that for all primes p one has the inequality

$$|\tau(p)| < 2p^{11/2}, \tag{1.5.6}$$

or, equivalently, that the complex roots of the polynomial $x^2 - \tau(p)x + p^{11}$ are complex conjugates of each other, and hence are of absolute value $p^{11/2}$. This conjecture was proved by Deligne as a consequence of [Del1] and his proof of the analog of the Riemann hypothesis in the Weil conjectures in [Del2].

Finally, Ramanujan conjectured congruences for the integers $\tau(p)$ with p prime, modulo certain powers of certain small prime numbers. In order to state these congruences we define, for $n \geq 1$ and $r \geq 0$:

$$\sigma_r(n) := \sum_{1 \leq d \mid n} d^r;$$

that is, $\sigma_r(n)$ is the sum of the rth powers of the positive divisors of n. We will now list the congruences that are given in the first pages of [Swi]:

$$\begin{aligned}
&\tau(n) \equiv \sigma_{11}(n) && \bmod 2^{11} && \text{if } n \equiv 1 \mod 8,\\
&\tau(n) \equiv 1217\sigma_{11}(n) && \bmod 2^{13} && \text{if } n \equiv 3 \mod 8,\\
&\tau(n) \equiv 1537\sigma_{11}(n) && \bmod 2^{12} && \text{if } n \equiv 5 \mod 8,\\
&\tau(n) \equiv 705\sigma_{11}(n) && \bmod 2^{14} && \text{if } n \equiv 7 \mod 8,
\end{aligned}$$

$$\begin{aligned}
&\tau(n) \equiv n^{-610}\sigma_{1231}(n) && \bmod 3^6 && \text{if } n \equiv 1 \mod 3,\\
&\tau(n) \equiv n^{-610}\sigma_{1231}(n) && \bmod 3^7 && \text{if } n \equiv 2 \mod 3,
\end{aligned}$$

$$\tau(n) \equiv n^{-30}\sigma_{71}(n) \qquad \bmod 5^3 \qquad \text{if } n \text{ is prime to } 5,$$

$$\tau(n) \equiv n\sigma_9(n) \quad \mod 7 \quad \text{if } n \equiv 0, 1, 2 \text{ or } 4 \mod 7,$$
$$\tau(n) \equiv n\sigma_9(n) \quad \mod 7^2 \quad \text{if } n \equiv 3, 5 \text{ or } 6 \mod 7,$$

$$\tau(p) \equiv 0 \quad \mod 23 \quad \text{if } p \text{ is prime and not a square mod } 23,$$
$$\tau(p) \equiv 2 \quad \mod 23 \quad \text{if } p \neq 23 \text{ is a prime of the form } u^2 + 23v^2,$$
$$\tau(p) \equiv -1 \quad \mod 23 \quad \text{for other primes } p \neq 23,$$

$$\tau(n) \equiv \sigma_{11}(n) \mod 691.$$

The reader is referred to [Swi] for the origin and for proofs of these congruences. There, Swinnerton-Dyer remarks that the proofs do little to explain why such congruences occur. Serre conjectured an explanation in [Ser2]. First of all, Serre conjectured the existence, for each prime number l, of a continuous representation

$$\rho_l : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \longrightarrow \mathrm{Aut}(V_l), \tag{1.5.7}$$

with V_l a two-dimensional $\mathbb{Q}_l$-vector space, such that ρ_l is unramified at all primes $p \neq l$, and such that for all $p \neq l$ the characteristic polynomial of $\rho_l(\mathrm{Frob}_p)$ is given by

$$\det(1 - x\mathrm{Frob}_p, V_l) = 1 - \tau(p)x + p^{11}x^2. \tag{1.5.8}$$

To help the reader, let us explain what unramified at p means, and what the Frobenius elements Frob_p are. For p prime, we let $\mathbb{Q}_p$ denote the topological field of p-adic numbers, and $\mathbb{Q}_p \to \overline{\mathbb{Q}}_p$ an algebraic closure. The action of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ on the set $\mathrm{Hom}(\overline{\mathbb{Q}}, \overline{\mathbb{Q}}_p)$ of embeddings of $\overline{\mathbb{Q}}$ into $\overline{\mathbb{Q}}_p$ is transitive, and each embedding induces an injection from $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$ into $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, the image of which is called a decomposition group of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ at p. The injections from $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$ into $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ and the corresponding decomposition groups at p obtained like this are all conjugated by the action of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. In order to go farther we need to say a bit about the structure of $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$. We let $\mathbb{Q}_p^{\mathrm{unr}}$ be the maximal unramified

extension of $\mathbb{Q}_p$ in $\overline{\mathbb{Q}}_p$, that is, the composite of all finite extensions K of $\mathbb{Q}_p$ in $\overline{\mathbb{Q}}_p$ such that p is a uniformizer for the integral closure O_K of $\mathbb{Z}_p$ in K. We let $\mathbb{Z}_p^{\mathrm{unr}}$ be the integral closure of $\mathbb{Z}_p$ in $\mathbb{Q}_p^{\mathrm{unr}}$; it is a local ring, and its residue field is an algebraic closure $\overline{\mathbb{F}}_p$ of $\mathbb{F}_p$. The subextension $\mathbb{Q}_p^{\mathrm{unr}}$ gives a short exact sequence:

$$I_p \hookrightarrow \mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p) \twoheadrightarrow \mathrm{Gal}(\overline{\mathbb{F}}_p/\mathbb{F}_p). \tag{1.5.9}$$

The subgroup I_p of $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$ is called the inertia subgroup. The quotient $\mathrm{Gal}(\overline{\mathbb{F}}_p/\mathbb{F}_p)$ is canonically isomorphic to $\hat{\mathbb{Z}}$, the profinite completion of $\mathbb{Z}$, by demanding that the element 1 of $\hat{\mathbb{Z}}$ corresponds to the Frobenius element Frob_p of $\mathrm{Gal}(\overline{\mathbb{F}}_p/\mathbb{F}_p)$ that sends x to x^p for each x in $\overline{\mathbb{F}}_p$.

Let now ρ_l be a continuous representation from $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ to $\mathrm{GL}(V_l)$ with V_l a finite dimensional $\mathbb{Q}_l$-vector space. Each embedding of $\overline{\mathbb{Q}}$ into $\overline{\mathbb{Q}}_p$ then gives a representation of $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$ on V_l. Different embeddings give isomorphic representations because they are conjugated by an element in the image of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ under ρ_l. We now choose one embedding, and call the representation $\rho_{l,p}$ of $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$ on V_l obtained like this the local representation at p attached to ρ_l. This being defined, ρ_l is then said to be unramified at a prime p if $\rho_{l,p}$ factors through the quotient $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p) \to \mathrm{Gal}(\overline{\mathbb{F}}_p/\mathbb{F}_p)$, that is, if I_p acts trivially on V_l. If ρ_l is unramified at p, then we get an element $\rho_l(\mathrm{Frob}_p)$ in $\mathrm{GL}(V_l)$. This element depends on our chosen embedding of $\overline{\mathbb{Q}}$ into $\overline{\mathbb{Q}}_p$, but its conjugacy class under $\rho_l(\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}))$ does not. In particular, we get a well defined conjugacy class in $\mathrm{GL}(V_l)$, and so the characteristic polynomial of $\rho_l(\mathrm{Frob}_p)$ is now defined if ρ_l is unramified at p.

Continuous representations such as ρ_l can be reduced modulo powers of l as follows. The compactness of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ implies that with respect to a suitable basis of V_l the representation ρ_l lands in $\mathrm{GL}_2(\mathbb{Z}_l)$, and hence gives representations to $\mathrm{GL}_2(\mathbb{Z}/l^n\mathbb{Z})$ for all $n \geq 0$. This reduction of ρ_l modulo powers of l is not unique, but the semisimplification of the reduction modulo l is well defined, that is, two reductions lead to the same Jordan-Hölder constituents. According to Serre, the congruences above would then be explained by properties of the image of ρ_l.

For example, if the image of the reduction modulo l of ρ_l is reducible, say an extension of two characters α and β from $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ to $\mathbb{F}_l^\times$, then one

has the identity in $\mathbb{F}_l$, for all $p \neq l$,

$$\tau(p) \equiv \alpha(\mathrm{Frob}_p) + \beta(\mathrm{Frob}_p). \tag{1.5.10}$$

The characters α and β are unramified outside l. By the Kronecker-Weber theorem, the maximal Abelian subextension of $\mathbb{Q} \to \overline{\mathbb{Q}}$ that is unramified outside l is the cyclotomic extension generated by all l-power roots of unity, with Galois group $\mathbb{Z}_l^\times$. It then follows that $\alpha = \chi_l^n$ and $\beta = \chi_l^m$ for suitable n and m, where χ_l is the character giving the action on the lth roots of unity in $\overline{\mathbb{Q}}$: for all σ in $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ and for all ζ in $\overline{\mathbb{Q}}^\times$ with $\zeta^l = 1$ one has $\sigma(\zeta) = \zeta^{\chi_l(\sigma)}$. The identity (1.5.10) in $\mathbb{F}_l$ above then takes the form

$$\tau(p) = p^n + p^m \mod l, \quad \text{for all } p \neq l, \tag{1.5.11}$$

which indeed is of the same form as the congruences mod l for $\tau(p)$ listed above. For example, the congruence mod 691 corresponds to the statement that the reduction modulo l of ρ_l gives the two characters 1 and χ_l^{11}.

Deligne, in [Del1], proved the existence of the ρ_l, as conjectured by Serre, by showing that they occur in the degree one l-adic étale cohomology of certain sheaves on certain curves, and in the degree 11 étale cohomology with $\mathbb{Q}_l$-coefficients of a variety of dimension 11. This last variety is, loosely speaking, the 10-fold fibered product of the universal elliptic curve. Deligne's constructions will be discussed in detail in Sections 2.2 and 2.4. It should be said that Shimura had already shown how to construct Galois representations in the case of modular forms of weight two; in that case one does not need étale cohomology, because torsion points of Jacobians of modular curves suffice, see [Shi1]. Blasius proved in [Bla] that the ρ_l cannot be obtained from the cohomology of Abelian varieties, using tensor constructions and subquotients.

At this point we give the following precise and simple statement, relating Ramanujan's τ-function to point counting on an algebraic variety C_{10} (more precisely, a quasi-projective scheme over $\mathbb{Z}$), for which one easily writes down a system of equations. Moreover, the statement relates the weight of Δ to the classical question in geometry on cubic plane curves passing through a given set of points: up to 9 points the situation is easy and the count is given by a polynomial. See also [Ber], especially Section 15.

1.5.12 Proposition *For $n \in \mathbb{Z}_{\geq 0}$, q a prime power and $\mathbb{F}_q$ a finite field with q elements, let $C_n(\mathbb{F}_q)$ be the set of $(C, P_1, \ldots, P_n)$, where C is a smooth cubic in $\mathbb{P}^2_{\mathbb{F}_q}$ and $P_i \in C(\mathbb{F}_q)$. Then there are $f_0, \ldots, f_{10} \in \mathbb{Z}[x]$ such that for all $\mathbb{F}_q$ and $n \leq 9$ one has $\#C_n(\mathbb{F}_q)/\#\mathrm{PGL}_3(\mathbb{F}_q) = f_n(q)$, and for all prime numbers p,*

$$\#C_{10}(\mathbb{F}_p) \,/\, \#\mathrm{PGL}_3(\mathbb{F}_p) = -\tau(p) + f_{10}(p).$$

Proof For n in $\mathbb{Z}_{\geq 0}$ and $\mathbb{F}_q$ a field with q elements, let $\mathcal{E}_n(\mathbb{F}_q)$ denote the category, and also its set of objects, of $(E/\mathbb{F}_q, P_1, \ldots, P_n)$, where $E/\mathbb{F}_q$ is an elliptic curve and $P_i \in E(\mathbb{F}_q)$; the morphisms are the isomorphisms $\phi\colon E \to E'$ such that $\phi(P_i) = P_i'$. For each n, the category $\mathcal{E}_n(\mathbb{F}_q)$ has only finitely many objects up to isomorphism, and one defines

$$\#\mathcal{E}_n(\mathbb{F}_q) = \sum_{x \in \mathcal{E}_n(\mathbb{F}_q)} \frac{1}{\#\,\mathrm{Aut}(x)},$$

where, in the sum, one takes one x per isomorphism class. It is well known (see [Del1], [Beh]) that for $n \leq 9$ the functions $q \mapsto \#\mathcal{E}_n(\mathbb{F}_q)$ are given by certain elements f_n in $\mathbb{Z}[x]$, and that there is an f_{10} in $\mathbb{Z}[x]$ such that for all prime numbers p one has $\#\mathcal{E}_{10}(\mathbb{F}_p) = -\tau(p) + f_{10}(p)$. In view of this, the claims in Proposition 1.5.12 are a consequence of the following equality, for all $n \in \mathbb{Z}_{\geq 0}$ and all prime powers q:

(1.5.13) for all $n \in \mathbb{Z}_{\geq 0}$ and all $\mathbb{F}_q$: $\quad \#C_n(\mathbb{F}_q) = \#\mathrm{PGL}_3(\mathbb{F}_q) \cdot \#\mathcal{E}_n(\mathbb{F}_q)$.

We prove (1.5.13) by comparing the subsets on both sides in which the underlying curves are fixed.

Let $n \in \mathbb{Z}_{\geq 0}$ and q a prime power. Let F be a nonsingular projective geometrically irreducible curve of genus one over $\mathbb{F}_q$, and let E_0 be its Jacobian. Then F is an E_0-torsor. By Lang's theorem, Theorem 2 of [Lan2], $F(\mathbb{F}_q)$ is not empty.

Let $C_n(\mathbb{F}_q)_F$ be the subset of $C_n(\mathbb{F}_q)$ consisting of the $(C, P_1, \ldots, P_n)$ with C isomorphic to F. The number of C in $\mathbb{P}^2_{\mathbb{F}_q}$ that are isomorphic to F is the number of embeddings $i\colon F \to \mathbb{P}^2_{\mathbb{F}_q}$, divided by $\#\,\mathrm{Aut}(F)$. Such embeddings are obtained from line bundles $\mathcal{L}$ of degree three on F, together with a basis, up to $\mathbb{F}_q^\times$, of $\mathcal{L}(F)$. Hence the number of embeddings is

$\#\mathrm{PGL}_3(\mathbb{F}_q) \cdot \#E_0(\mathbb{F}_q)$. The group $\mathrm{Aut}(F)$ has the subgroup of translations $E_0(\mathbb{F}_q)$, with quotient $\mathrm{Aut}(E_0)$. So we find

$$\#C_n(\mathbb{F}_q)_F = \#\mathrm{PGL}_3(\mathbb{F}_q) \cdot (\#E_0(\mathbb{F}_q))^n / \#\,\mathrm{Aut}(E_0).$$

On the other hand, let $\mathcal{E}_n(\mathbb{F}_q)_{E_0}$ be the full subcategory of $\mathcal{E}_n(\mathbb{F}_q)$ with objects the $(E_0, P_1, \ldots, P_n)$, with P_i in $E_0(\mathbb{F}_q)$. The group $\mathrm{Aut}(E_0)$ acts on the set of objects of $\mathcal{E}_n(\mathbb{F}_q)_{E_0}$, and this action *is* the set of morphisms in $\mathcal{E}_n(\mathbb{F}_q)_{E_0}$. This means that:

$$\#\mathcal{E}_n(\mathbb{F}_q)_{E_0} = (\#E_0(\mathbb{F}_q))^n / \#\,\mathrm{Aut}(E_0).$$

Summing over the isomorphism classes of F gives (1.5.13). □

1.5.14 Remark For those who like more technical statements, we mention (without further explanation) that, for all $n \in \mathbb{Z}_{\geq 1}$, we have a Zariski PGL_3-torsor $C_n \to \mathcal{E}_n$, sending $(C \subset \mathbb{P}^2_S, P_1, \ldots, P_n)$ to $((C, P_0), P_1, \ldots, P_n)$, where $P_0 \in C(S)$ is determined uniquely by the condition that, locally on S, $\mathcal{O}_{\mathbb{P}^2}(1)|_C$ is isomorphic to $\mathcal{O}_C(P_0 + 2P_1)$.

1.5.15 Remark The polynomials f_n mentioned in Proposition 1.5.12 have been computed by Carel Faber and Gerard van der Geer. Their result is

$$\begin{aligned}
f_0 &= x,\\
f_1 &= x^2 + x,\\
f_2 &= x^3 + 3x^2 + x - 1,\\
f_3 &= x^4 + 6x^3 + 6x^2 - 2x - 3,\\
f_4 &= x^5 + 10x^4 + 20x^3 + 4x^2 - 14x - 7,\\
f_5 &= x^6 + 15x^5 + 50x^4 + 40x^3 - 30x^2 - 49x - 15,\\
f_6 &= x^7 + 21x^6 + 105x^5 + 160x^4 - 183x^2 - 139x - 31,
\end{aligned}$$

$$
\begin{aligned}
f_7 &= x^8 + 28x^7 + 196x^6 + 469x^5 + 280x^4 - 427x^3 - 700x^2 \\
&\quad - 356x - 63, \\
f_8 &= x^9 + 36x^8 + 336x^7 + 1148x^6 + 1386x^5 - 406x^4 - 2436x^3 \\
&\quad - 2224x^2 - 860x - 127, \\
f_9 &= x^{10} + 45x^9 + 540x^8 + 2484x^7 + 4662x^6 + 1764x^5 - 6090x^4 \\
&\quad - 9804x^3 - 6372x^2 - 2003x - 255, \\
f_{10} &= x^{11} + 55x^{10} + 825x^9 + 4905x^8 + 12870x^7 + 12264x^6 \\
&\quad - 9240x^5 - 33210x^4 - 33495x^3 - 17095x^2 - 4553x - 511.
\end{aligned}
$$

We refer to Birch [Bir] for results on the distribution of the number of rational points on elliptic curves over finite fields, that also make $\tau(p)$ appear.

In [Swi], Swinnerton-Dyer gives results, partly resulting from his correspondence with Serre, in which the consequences of the existence of the ρ_l for congruences of $\tau(p)$ modulo l are explored. A natural question to ask is if there are primes l other than 2, 3, 5, 7, 23, and 691 modulo which there are similar congruences for $\tau(p)$.

For each $p \neq l$, $\tau(p)$ is the trace of $\rho_l(\mathrm{Frob}_p)$, and the determinant of $\rho_l(\mathrm{Frob}_p)$ equals p^{11}. Hence, a polynomial relation between $\tau(p)$ and p^{11}, valid modulo some l^n for all $p \neq l$, is a relation between the determinant and the trace of all $\rho_l(\mathrm{Frob}_p)$ in $\mathrm{GL}_2(\mathbb{Z}/l^n\mathbb{Z})$. But Chebotarev's theorem (see [Lan6] or [Ca-Fr], for example) implies that every element of the image of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ in $\mathrm{GL}_2(\mathbb{Z}/l^n\mathbb{Z})$ is of the form Frob_p for infinitely many p. Hence, such a polynomial relation is valid for all elements in the image of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ in $\mathrm{GL}_2(\mathbb{Z}/l^n\mathbb{Z})$. For this reason, the existence of nontrivial congruences modulo l^n as above for $\tau(p)$ depends on this image.

The image of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ in $\mathbb{Z}_l^\times$ under $\det \circ \rho_l$ is equal to the subgroup of 11th powers in $\mathbb{Z}_l^\times$. To explain this, we note that $\det \circ \rho_l$ is a continuous character from $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ to $\mathbb{Z}_l^\times$, unramified outside l, and such that Frob_p is mapped to p^{11} for all $p \neq l$; this implies that $\det \circ \rho_l$ is the 11th power of the l-adic cyclotomic character $\chi_l : \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathbb{Z}_l^\times$, defined by $\sigma(z) = z^{\chi_l(\sigma)}$, for all σ in $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ and all z in $\overline{\mathbb{Q}}^\times$ of l-power order.

In order to state the results in [Swi], one calls a prime number l *exceptional (for* Δ*)* if the image of ρ_l, taking values in $\mathrm{GL}_2(\mathbb{Z}_l)$, does *not* contain $\mathrm{SL}_2(\mathbb{Z}_l)$. For l not exceptional, that is, such that the image of ρ_l con-

tains $\mathrm{SL}_2(\mathbb{Z}_l)$, the image of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ in $\mathrm{GL}_2(\mathbb{Z}_l) \times \mathbb{Z}_l^\times$, under (ρ_l, χ_l), is the subgroup H of elements (g, t) such that $\det(g) = t^{11}$. This subgroup H maps surjectively to $\mathbb{F}_l \times \mathbb{F}_l^\times$ under $(g, t) \mapsto (\mathrm{trace}(g), t)$, and therefore there can be no congruence for $\tau(p)$ modulo l as above.

The Corollary to Theorem 4 in [Swi] states, among other results, that the list of primes which are exceptional for Δ is $\{2, 3, 5, 7, 23, 691\}$. The main tool that is used that we have not discussed, is the theory of modular forms modulo l, or, equivalently, the theory of congruences modulo l between modular forms. As a consequence, there are no similar congruences for $\tau(p)$ modulo primes other than the ones listed above. The special form of the congruences modulo 23 is explained by the fact that in that case the image of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ in $\mathrm{GL}_2(\mathbb{F}_{23})$ is dihedral; in the other cases the residual representation, that is, the representation to $\mathrm{GL}_2(\mathbb{F}_l)$, is reducible. In the case $l = 2$, Swinnerton-Dyer has determined the image of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ in $\mathrm{GL}_2(\mathbb{Z}_2)$ exactly: see the appendix in [Swi].

The direction in which we generalize Schoof's algorithm is to give an algorithm that computes for prime numbers l that are not exceptional for Δ the field extension $\mathbb{Q} \to K_l$ that corresponds to the representation of $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ to $\mathrm{GL}_2(\mathbb{F}_l)$ that comes from Δ. The field K_l is given in the form $\mathbb{Q}[x]/(f_l)$. The computation has a running time that is polynomial in l. It is fair to say that this algorithm makes the mod l Galois representations attached to Δ accessible to computation, at least theoretically. As the field extensions that are involved are nonsolvable, this should be seen as a step beyond computational class field theory, and beyond the case of elliptic curves, in the direction to make the results of the Langlands program accessible to computations.

As a consequence, one can compute $\tau(p)$ mod l in time polynomial in $\log p$ and l, by reducing f_l as above mod p and some more computations that will be described later (see Section 15.1). By doing this for sufficiently many l, just as in Schoof's algorithm, one then gets an algorithm that computes $\tau(p)$ in time polynomial in $\log p$.

In Section 15.2 the method used here is generalized to the case of modular forms for $\mathrm{SL}_2(\mathbb{Z})$ of arbitrary weight. The main result there is Theorem 15.2.1.

1.6 COMPARISON WITH p-ADIC METHODS

Before we start seriously with the theory of modular forms and the Galois representations attached to them in the next chapter, we compare our generalization of Schoof's algorithm with the so-called p-adic methods that have been developed since 2000 by Satoh [Sat], Kedlaya [Ked] (see also [Edi3]), Hubrechts, [Hub], Lauder and Wan [La-Wa1], [La-Wa2], [Lau1] and [Lau2], Fouquet, Gaudry, Gürel and Harley [Fo-Ga-Ha], [Ga-Gu], Denef and Vercauteren and Castryk [De-Ve], [Ca-De-Ve], Mestre, Lercier and Lubicz [Le-Lu], Carls, Kohel and Lubicz, [Ca-Ko-Lu], [Ca-Lu], and Gerkmann, [Ger1] and [Ger2]. Actually, we should notice that such a method was already introduced in [Ka-Lu] in 1982, but that this article seems to have been forgotten (we thank Fre Vercauteren for having drawn our attention to this article). An overview of the p-adic approach is given in [Ch].

In all these methods, one works with fields of small characteristic p, hence of the form $\mathbb{F}_q$ with $q = p^m$ and p fixed. All articles cited in the previous paragraph have the common property that they give algorithms for computing the number of $\mathbb{F}_q$-rational points on certain varieties X over $\mathbb{F}_q$, using, sometimes indirectly, cohomology groups with p-adic coefficients, whence the terminology "p-adic methods."

For example, Satoh [Sat] uses the canonical lift of ordinary elliptic curves and the action of the lifted Frobenius endomorphism on the tangent space, which can be interpreted in terms of the algebraic de Rham cohomology of the lifted curve. Kedlaya [Ked] uses Monsky-Washnitzer cohomology of certain affine pieces of hyperelliptic curves. In fact, all cohomology groups used here are de Rham type cohomology groups, given by complexes of differential forms on certain p-adic lifts of the varieties in question. Just as an example, let us mention that Kedlaya [Ked] gives an algorithm that for fixed $p \neq 2$ computes the zeta functions of hyperelliptic curves given by the equations

$$y^2 = f(x),$$

where f has arbitrary degree, in time $m^3 \deg(f)^4$. The running times of the other algorithms are similar, but all have in common that the running time grows at least linearly in p (or linear in $O(p^{1/2})$, in [Harv]), hence

exponentially in $\log p$. The explanation for this is that somehow in each case nonsparse polynomials of degree at least linear in p have to be manipulated.

Summarizing this recent progress, one can say that, at least from a theoretical point of view, the problem of counting the solutions of systems of polynomial equations over finite fields of a fixed characteristic p and in a fixed number of variables has been solved. If p is not bounded, then almost nothing is known about the existence of polynomial time algorithms.

A very important difference between this book, using étale cohomology with coefficients in $\mathbb{F}_l$, and the p-adic methods, is that the Galois representations on $\mathbb{F}_l$-vector spaces that we obtain are *global* in the sense that they are representations of the absolute Galois group of the global field $\mathbb{Q}$. The field extensions such as the $K_l = \mathbb{Q}[x]/(f_l)$ arising from Δ discussed in the previous section have the advantage that one can choose to do the required computations over the complex numbers, approximating f_l, or p-adically at some suitable prime p, or in $\mathbb{F}_p$ for sufficiently many small p. Also, as we have said already, being able to compute such field extensions K_l, that give mod l information on the Frobenius elements at all primes $p \neq l$, is very interesting. On the other hand, the p-adic methods force one to compute with p-adic numbers, or, actually, modulo some sufficiently high power of p, and it gives information only on the Frobenius at p. The main drawback of the étale cohomology with $\mathbb{F}_l$-coefficients seems to be that the degree of the field extensions as K_l to be dealt with grows exponentially in the dimension of the cohomology groups; for that reason, we do not know how to use étale cohomology to compute $\#X(\mathbb{F}_q)$ for X a curve of arbitrary genus in a time polynomial in $\log q$ and the genus of X. Nevertheless, for modular curves, see the end of Section 15.3.