Trying to understand Deligne's proof of the Weil conjectures

(A tale in two parts)

January 29, 2008

(Part I : Introduction to étale cohomology)

1 Introduction

These notes are an attempt to convey some of the ideas, if not the substance or the details, of the proof of the Weil conjectures by P. Deligne [De1], as far as I understand them, which is to say somewhat superficially – but after all J.-P. Serre (see [Ser]) himself acknowledged that he didn't check everything...

What makes this possible is that this proof still contains some crucial steps which are beautiful in themselves and can be stated and even (almost) proved independently of the rest. The *context* of the proof has to be accepted as given, by analogy with more elementary cases already known (elliptic curves, for instance). Although it is possible to present various motivations for the introduction of the étale cohomology which is the main instrument, getting beyond hand waving is the matter of very serious work, and the only reasonable hope is that the analogies will carry enough weight. What I wish to emphasize is how much the classical study of manifolds was a guiding principle throughout the history of this wonderful episode of mathematical invention – until the Riemann hypothesis itself, that is, when Deligne found something completely different...

2 Statements

The Weil conjectures, as stated in [Wei], are a natural generalization to higher dimensional algebraic varieties of the case of curves that we have been studying this semester. So let X_0 be a variety of dimension d over a finite field $\mathbf{F}_q$. Following Deligne's notations, we write X_n for the variety obtained by looking at X_0 as defined over the extension field $\mathbf{F}_{q^n}$, and X when we consider X_0 over the algebraic closure $\overline{\mathbf{F}}_q$ of $\mathbf{F}_q$, symbolically

$$X_n = X_0 \times \mathbf{F}_{q^n}$$

and

$$X = X_0 \times \overline{\mathbf{F}}_q.$$

In general, notions and objects defined over the algebraic closure are called "geometric", while those defined over a finite field are "arithmetic"; subscripts will vary accordingly.

By definition, the zeta function of X_0 is the formal power series with coefficients in $\mathbf{Q}$ given by

$$Z(X_0) = \exp\Bigl(\sum_{n\geqslant 1} |X_0(\mathbf{F}_{q^n})|\frac{T^n}{n}\Bigr). \tag{1}$$

Looking at the field of definition of points in $\overline{\mathbf{F}}_q$, it is not hard to rewrite this as an "Euler product"

$$Z(X_0, q^{-s}) = \prod_{x\in|X_0|} (1 - N(x)^{-s})^{-1} \tag{2}$$

where $N(x) = q^{\deg(x)}$, and $|X_0|$ is the set of closed points of X_0, which is the same as the set of orbits of points on X for the action of the Galois group of $\overline{\mathbf{F}}_q$ over $\mathbf{F}_q$. This shows the analogy with the Riemann zeta function (the computation is very similar to that done for the zeta function of the Gauss or Kloosterman sums; it will be re-done later in section 6).

Examples. (1) If $X_0 = \mathbf{P}^d_{\mathbf{F}_q}$, then we easily find, by exact counting, that

$$Z(X_0) = \frac{1}{(1-T)(1-qT)\dots(1-q^dT)}. \tag{3}$$

While this can be defined in general, Weil made the following conjectures if X_0 is *smooth* and *projective*:

- $Z(X_0)$ is actually a rational function.
- There exists an integer χ and $\varepsilon = \pm 1$ such that $Z(X_0)$ satisfies a functional equation
$$Z\bigl(X_0, \frac{1}{q^dT}\bigr) = \varepsilon q^{d\chi/2}T^{\chi}Z(X_0, T). \tag{4}$$
- We can write $Z(X_0)$ as a rational function of the form
$$Z(X_0) = \frac{P_1 \dots P_{2d-1}}{P_0 \dots P_{2d}} \tag{5}$$
where $P_0,\dots,P_{2d}$ are polynomials with integer coefficients, satisfying
$$P_0 = 1 - T,\ P_{2d} = 1 - q^dT$$
and all the zeros of P_i, $i = 0, \dots, 2d$, are of modulus $q^{-i/2}$.

  This last statement, of course, is the analogue of the Riemann hypothesis. Note that this defines P_i uniquely when writing $Z(X_0)$ as a rational function, by looking at poles or zeros of modulus $q^{-i/2}$.

A few remarks: first, we can check quickly that it works for $\mathbf{P}^n_{\mathbf{F}_q}$, by looking at (3). In the functional equation we have $\chi = d + 1$ and $\varepsilon = (-1)^d$.

Moreover, the Riemann hypothesis can not be reduced to simply counting the number of points on $X_0(\mathbf{F}_q)$, and not even on $X_0(\mathbf{F}_{q^n})$ asymptotically when n tends to infinity, as was the case for elliptic curves or for curves in general, *unless* all but one of the "non-trivial" P_i ($1 \leqslant i \leqslant 2d-1$) are zero. Indeed, from (5) we get

$$|X_0(\mathbf{F}_{q^n})| = \sum_{0 \leqslant i \leqslant 2d} (-1)^i \sum_j \alpha_{i,j}^n$$

where $P_i = \prod_j (1 - \alpha_{i,j} T)$, which is

$$|X_0(\mathbf{F}_{q^n})| = q^{nd} + 1 + \sum_{1 \leqslant i \leqslant 2d-1} (-1)^i \sum_j \alpha_{i,j}^n$$

and the only obvious estimate for the error term from the Riemann hypothesis is $O(q^{n(d-1/2)})$.

3 The framework: étale cohomology

3.1 Ordinary cohomology

Imagine that Weil had known nothing about any kind of cohomology theory (maybe he slept during all the 1920's and 30's), and that he had made those conjectures, based on the evidence of the known, explicitly computable, cases. Then, I claim, formulae such as (3) above, alone, would have put people on the track.

To explain, I will recall the definition of the cohomology (with complex coefficients) of a topological space T. There are actually many different definitions, with various advantages and drawbacks, but maybe the simplest for computational purposes and the best suited to introduce the generalization to étale cohomology is what is called Čech cohomology.

The definition is motivated by the important concept of local and global properties. Given T, we might feel that we understand it well enough locally (as in the case of differentiable manifold where the local theory is that of differential calculus on open subsets of $\mathbf{R}^n$), and that some of the most interesting questions we may ask about T are global versions of those local notions. For instance, an ordinary differential equation always has a local solution, but when does it have a global one? The natural idea is to take all local solutions f_i, defined on some open subsets U_i, and compare them on the intersections, hoping that they will coincide.

This is what the Čech cohomology groups will make precise. They are defined with respect to an open covering $\mathcal{U} = (U_i)_{i \in J}$ of T. For any multi-index $I = (i_0, \ldots, i_n)$ of elements of J, we will write U_I for the intersection of the corresponding open subsets:

$$U_I = U_{i_0} \cap \ldots \cap U_{i_n}$$

and $\mathbf{C}_I$ for the vector space of complex valued functions on the set of connected components of U_I (equivalently, locally constant continuous functions on U_I).

Then we define the vector space of n-cochains, for $n \geqslant 0$, to be

$$C^n(\mathcal{U}, \mathbf{C}) = \prod_{I \in J^{n+1}} \mathbf{C}_I.$$

If $f \in C^n(\mathcal{U}, \mathbf{C})$, we will write $f(I)$ for its component at I; as a locally constant function, it is the map $U_I \to \mathbf{C}$, which takes $x \in U_I$ to the complex number associated to the connected component it lies in. For instance if every U in the covering is connected, an element of $C^0(\mathcal{U}, \mathbf{C})$ associates a complex number $c(U)$ to each of the open subsets of the given covering; to compare those numbers we can take the differences $c(U) - c(V)$: those are naturally associated to the intersections $U \cap V$. In general, we get a linear map

$$d^0 \,:\, C^0(\mathcal{U}, \mathbf{C}) \to C^1(\mathcal{U}, \mathbf{C})$$

such that $d^0(f)(i,j) = f(i) - f(j)$ (as functions on $U_{i,j}$, so $f(i)$ really means $f(i)$ *restricted to* $U_{i,j}$). Then $d^0(f) = 0$ simply means that f is constant: in other words, we have just one complex number $f(T)$ which is "globally" associated to f. This is the kind of things we wanted to study.

But we do not stop with d^0: we can clearly generalize it by defining linear maps d^n, for all $n \geqslant 0$

$$d^n \,:\, C^n(\mathcal{U}, \mathbf{C}) \to C^{n+1}(\mathcal{U}, \mathbf{C})$$

by

$$d^n(f)(i_0, \ldots, i_{n+1}) = \sum_{0 \leqslant j \leqslant n+1} (-1)^j f(i_0, \ldots, i_{j-1}, i_{j+1}, \ldots, i_{n+1}).$$

For instance:

$$d^1(f)(i,j,k) = f(i,j) - f(i,k) + f(j,k).$$

Now comes the most important fact: for any $n \geqslant 0$, we have $d^{n+1} \circ d^n = 0$, or more succinctly $d \circ d = 0$. This is the basic equation of all cohomology theories.

There is no mystery in this: take for instance $d^1 \circ d^0$, then by the formulae above

$$\begin{aligned} d^1 \circ d^0(f)(i,j,k) &= d^0(f)(i,j) - d^0(f)(i,k) + d^0(f)(j,k) \\ &= f(i) - f(j) - f(i) + f(k) + f(j) - f(k) \\ &= 0 \end{aligned}$$

voilà!

Armed with the formula $d^{n+1} \circ d^n = 0$ we can define the space of n-cocycles, $Z^n(\mathcal{U}, \mathbf{C})$, and the space of n-coboundaries, $B^n(\mathcal{U}, \mathbf{C})$:

$$Z^n(\mathcal{U}, \mathbf{C}) = \mathrm{Ker}(d^n \,:\, C^n(\mathcal{U}, \mathbf{C}) \to C^{n+1}(\mathcal{U}, \mathbf{C}))$$

$$B^n(\mathcal{U}, \mathbf{C}) = \mathrm{Im}(d^{n-1} \,:\, C^{n-1}(\mathcal{U}, \mathbf{C}) \to C^n(\mathcal{U}, \mathbf{C}))$$

with the effect that

$$B^n(\mathcal{U}, \mathbf{C}) \subset Z^n(\mathcal{U}, \mathbf{C}).$$

This means that we can finally define the n-th Čech cohomology space (or group) of T (with respect to the covering $\mathcal{U}$) as

$$H^n(\mathcal{U}, \mathbf{C}) = Z^n(\mathcal{U}, \mathbf{C}) / B^n(\mathcal{U}, \mathbf{C}).$$

(As a matter of convention, we define $C^{-1} = 0$ and $d^{-1} = 0$, so this definition applies equally well to H^0.)

Of course, those spaces depend on the covering $\mathcal{U}$; however, given two coverings $\mathcal{U}$ and $\mathcal{V}$, there is a third $\mathcal{W}$ refining both (take the intersections), and this gives maps

$$H^n(\mathcal{U}, \mathbf{C}) \to H^n(\mathcal{W}, \mathbf{C})$$

and

$$H^n(\mathcal{V}, \mathbf{C}) \to H^n(\mathcal{W}, \mathbf{C})$$

which allows us to think of the "union" of the cohomology groups over all coverings, which is then independent of any particular one. Or, more simply, if you think of local-global problems as above, you may take a covering by open subsets on which the local problem is known and argue with it. Actually, both ways are more or less the same in many cases: for instance, in the case of manifolds, it can be shown that for any covering with contractible open subsets (or subsets homeomorphic to convex subsets of $\mathbf{R}^n$), the cohomology spaces are the same. In any case, we will write $H^n(T, \mathbf{C})$ for the resulting cohomology groups, which are intrinsic invariants of the space T only.

Once you take this for granted, the Čech cohomology groups have many advantages. First, they can be actually computed in many cases: if the covering happens to be finite for instance, each C^n is finite dimensional, which makes it obvious that H^n is then also finite dimensional.

As another example, take the covering of a space T by its connected components $(U_i)_{i \in \pi_0(T)}$. Then $U_{(i,j)} = \emptyset$ for $i \neq j$ so $d^0 = 0$ and $Z^0 = C^0$; as $d^{-1} = 0$, we get (abusing notation: this particular covering is sufficient in this case, though not in all, of course)

$$H^0(T, \mathbf{C}) = \mathbf{C}^{\pi_0(T)}$$

namely a $\mathbf{C}$-vector space of dimension equal to the number of connected components of T. This shows already that the cohomology groups carry some information on the topology of T (although this might seem too obvious to be of any use, Hartshorne [Har], III-11-3, contains a proof of an important theorem of algebraic geometry based on such a result about another H^0).

Another very important remark is that there is nothing sacred here about $\mathbf{C}$ in the definition of C^n: we might replace it by $\mathbf{R}$, $\mathbf{Q}$, $\mathbf{Z}$, or even by more general vector spaces associated to the covering. For differential equations, it would be easier to compare local solutions if we could take instead of $\mathbf{C}_I$ the space of solutions defined on U_I; similarly for C^1 we can take the solutions on the intersection (0 if it's empty), so that we give a meaning to $d^0(f)(i,j) = f(i) - f(j)$ simply by taking the solutions over U_i and U_j and restricting them to $U_i \cap U_j$ before computing the difference.

This simple idea has had enormous consequences: the abstraction of the principle gives the notion of a *sheaf* on T (another French invention associated with the happy effects of jails on mathematical creativity, since *faisceaux* were invented by Jean Leray during World War II when he was prisoner in Germany). This is a rule $\mathcal{F}$ associating a vector space $\mathcal{F}(U)$ to every open subset $U \subset T$ so that

- $\mathcal{F}(\emptyset) = 0$.
- A restriction map $\mathcal{F}(U) \to \mathcal{F}(V)$ is defined for any $V \subset U$, and is transitive: restricting to $W \subset V \subset U$ directly from U or going through V has the same result.

- $\mathcal{F}$ satisfies a *locality* axiom, meaning that an element of $\mathcal{F}(U)$ is determined uniquely by its restrictions to a covering of U, and conversely, given elements of $\mathcal{F}(U_i)$ for a covering (U_i) of U, which coincide on the intersections, there is an element in $\mathcal{F}(U)$ which restricts to those. This is of course true for functions and anything related to functions with the usual restriction.[1]

With a sheaf $\mathcal{F}$ we then obtain the cohomology spaces of T with coefficients in $\mathcal{F}$, denoted by $H^n(T, \mathcal{F})$. Computing $H^0(T, \mathcal{F})$ is still an easy thing: by the definition of the local nature of elements of $\mathcal{F}(U)$ (called sections of $\mathcal{F}$ over U), an element of $H^0(\mathcal{U}, \mathcal{F})$ gives sections $s_i \in \mathcal{F}(U_i)$ whose restrictions coincide on the intersections $U_i \cap U_j$, and therefore must patch together to give an element of $\mathcal{F}(T)$ (a "global" section), so that

$$H^0(T, \mathcal{F}) = \mathcal{F}(T).$$

But let us come back to the Weil conjectures. Where has this digression brought us? The point is that as people were defining and getting to know cohomology, they computed the cohomology groups of many classically defined topological spaces, and one of the first was n-dimensional projective space over $\mathbf{C}$. Here is the answer in this case (which is not difficult to get with the help of Čech cohomology, using the standard covering of $\mathbf{P}^n_{\mathbf{C}}$ with $n+1$ subsets homeomorphic to $\mathbf{C}^n$, see [Har] III, 5, for similar computations): one finds first that $H^k(\mathbf{P}^n_{\mathbf{C}}, \mathbf{C}) = 0$ for all $k > 2n$, which reflects the general result that $H^k(T, \mathbf{C}) = 0$ if T is a real manifold and $k > \dim(T)$, and then in the interesting range

$$\begin{array}{rcl} H^{2i}(\mathbf{P}^n_{\mathbf{C}}, \mathbf{C}) & = & \mathbf{C}, \text{ if } 0 \leqslant i \leqslant n. \\ H^{2i+1}(\mathbf{P}^n_{\mathbf{C}}, \mathbf{C}) & = & 0, \text{ if } 0 \leqslant i < n. \end{array} \tag{6}$$

And now, as Hamlet would say, "Look here, upon this picture (3), and on this (6)". Comparing with the conjectural form of $Z(\mathbf{P}^d_{\mathbf{F}_q})$ (5), we see that there is a factor of degree 1 for each index $2i$ such that $H^{2i} = \mathbf{C}$, and no odd index factors where $H^{2i+1} = 0$. This coincidence, extended to other cases where X_0 appeared as the "reduction mod p" of a well-known variety $X_{\mathbf{C}}$ over $\mathbf{C}$, and where computations of both $Z(X_0)$ and $H^i(X_{\mathbf{C}}, \mathbf{C})$ were available, is actually the motivation behind this precise expression of $Z(X_0)$ as a rational function: Weil already added the precision that P_i would be a polynomial of degree $\dim H^i(X_{\mathbf{C}}, \mathbf{C})$ in this case.

As another confirmation of this, consider the case of curves. We will see by Bombieri's adaptation of Stepanov's method that for $C_0/\mathbf{F}_q$ smooth and projective there exists a polynomial $P_1 \in \mathbf{Z}[T]$ of degree $2g$, g being the genus of C, such that

$$Z(C_0) = \frac{P_1}{(1-T)(1-qT)}$$

while it was known since the beginning of the century that the cohomology of a Riemann surface C is given by

$$\begin{array}{rcl} H^0(C, \mathbf{C}) & = & \mathbf{C} \\ H^1(C, \mathbf{C}) & = & \mathbf{C}^{2g} \\ H^2(C, \mathbf{C}) & = & \mathbf{C} \end{array} \tag{7}$$

[1] This locality axiom explains the intervention of connected components in the definition of the "constant sheaf $\mathbf{C}$" which gave the first definition: "constant" means in fact "locally constant".

(and the others vanish).

This provides a hint that some cohomological explanation of the Weil conjectures exists. But the analogy actually goes well beyond what could be a coincidence, to give very convincing evidence that it should be the key to a proof.

Indeed, algebraic topologists had established the following fundamental facts about the cohomology of (say) compact, connected, and oriented manifolds of dimension d over $\mathbf{C}$:

- "Cohomological dimension": if X is of dimension d, then $H^i(X, \mathbf{C}) = 0$ for all $i > 2d$.

- Finiteness: for any $i \geqslant 0$, $H^i(X, \mathbf{C})$ is a finite dimensional $\mathbf{C}$-vector space.

- "Functoriality in X": if $f : X \to Y$ is a continuous map then there are associated maps in cohomology for all $i \geqslant 0$
$$f^\star : H^i(Y, \mathbf{C}) \to H^i(X, \mathbf{C}).$$
In particular, a map $X \to X$ induces endomorphisms of the finite dimensional $\mathbf{C}$-vector spaces $H^i(X, \mathbf{C})$.

- Poincaré duality: there is an isomorphism (depending on the choice of an orientation of X)
$$\mathrm{tr} : H^{2d}(X, \mathbf{C}) \simeq \mathbf{C}$$
and for all $i \leqslant d$ natural bilinear forms ("cup product")
$$H^i(X, \mathbf{C}) \times H^{2d-i}(X, \mathbf{C}) \to H^{2d}(X, \mathbf{C}) \simeq \mathbf{C}$$
which are perfect pairings of finite dimensional vector spaces; in particular, H^i is dual to H^{2d-i}, and they have the same dimension.

- Künneth theorem: for any $i \geqslant 0$, there are canonical isomorphisms
$$H^i(X \times Y, \mathbf{C}) = \bigoplus_{j+k=i} H^j(X, \mathbf{C}) \otimes H^k(X, \mathbf{C}).$$

- Lefschetz trace formula: let $f : X \to X$ be a differentiable map with isolated fixed points, and $L(f)$ the algebraic number of those (i.e. they are counted with multiplicity, positive if f induces an orientation preserving map on the tangent space, negative if it induces an orientation reversing map). Then we have the equality
$$L(f) = \sum_{i=0}^{2d} (-1)^i \mathrm{Tr}(f^\star \mid H^i(X, \mathbf{C})). \tag{8}$$

Remarks. Of these, the first and second are not too hard in Čech cohomology, the third is a simple exercise since a covering $\mathcal{U}$ of Y gives a covering $f^\star(\mathcal{U}) = (f^{-1}(U_i))$ of X. Poincaré duality is deeper. The Lefschetz trace formula, arguably the most surprising among those properties since it gives a very concrete link between properties of the space X and its cohomology[2], is actually a formal consequence of the others (including some which are omitted here). You might get some feeling by checking for a zero-dimensional space, namely a finite set of points.

All the above, in any case, are much simpler than their analogues below for étale cohomology, and it is important to stress that all were known and available at the time Weil made his conjectures, and it was not necessary to rediscover them.

This Lefschetz trace formula however springs to our attention for the problem of counting points on varieties over finite fields: we know that the points of X_0 with coordinates in $\mathbf{F}_{q^n}$ are just the fixed points of the n-th power F^n of the Frobenius morphism $F : X \to X$. Therefore let us be formal and apply (8):

$$|X_0(\mathbf{F}_{q^n})| = \sum_{i=0}^{2d} (-1)^d \mathrm{Tr}((F^n)^\star \mid H^i(X))$$

gives

$$\sum_{n\geqslant 1} |X_0(\mathbf{F}_{q^n})| \frac{T^n}{n} = \sum_{i=0}^{2d} (-1)^i \sum_{n\geqslant 1} \mathrm{Tr}((F^n)^\star \mid H^i(X)) \frac{T^n}{n}$$

but it is merely a linear algebra computation to check that any endomorphism f of a finite dimensional vector space V satisfies the formal power series identity

$$\sum_{n\geqslant 1} \mathrm{Tr}(f^n \mid V) \frac{T^n}{n} = -\log\det(1 - fT)$$

(if $\dim V = 1$, that's just the power series of the logarithm), so that we get

$$Z(X_0) = \prod_{i=0}^{2d} \det(1 - F^\star T \mid H^i(X))^{(-1)^{i+1}}. \tag{9}$$

This is exactly of the conjectured form (5) with polynomials

$$P_i = \det(1 - F^\star T \mid H^i(X))$$

in particular the degree of P_i is indeed the dimension of $H^i(X)$!

But here arises the complication: the astute reader will have noticed the absence of the coefficient of the cohomology in the computation above: $H^i(X, ?)$, that is the question. For this was not for mere notational brevity: it is the crux of the matter.

Indeed, the proof of the properties that we have stated depends crucially on the fact that X is a manifold; similar results fail for more general topological spaces, when using the Čech cohomology with coefficients in $\mathbf{C}$. And we are now dealing with algebraic varieties, defined over a field of characteristic p. A cohomology theory of those varieties is required, and during the 50's and 60's, much thought was given to the matter.

[2] It easily implies, for instance, such results as Brouwer's fixed point theorem: every continuous map from the unit ball in $\mathbf{R}^n$ to itself has a fixed point.

3.2 Enter étale cohomology

The first idea, simply to use the Zariski topology, fails: this is a rather coarse topology, and although it is possible to define groups $H^i(X, \mathbf{C})$ by the formal definition above, it doesn't give anything useful (you might want to try some computation to discover that the spaces obtained do not coincide with the classical ones, especially for the dimension). Serre, however, showed that this topology was good enough provided one took as coefficients some objects better suited to its nature: those are sheaves, as defined above, satisfying a further condition called "coherence". For instance, the basic sheaf on a variety X/k is the so-called structural sheaf $\mathcal{O}_X$ defined by $\mathcal{O}_X(U) = k[U]$, the ring of regular functions on the open set $U \subset X$. This theory has a life of its own ([Har], III), but has limited usefulness for the Weil conjectures: indeed, if X_0 is defined over $\mathbf{F}_q$, all coherent sheaves on X_0 will be characteristic p objects and their cohomology will give at best vector spaces over $\mathbf{F}_q$: traces of operators acting on cohomology, for instance, will be elements of $\mathbf{F}_q$, and couldn't give an expression for the Zeta function which belongs to $\mathbf{Q}[[T]]$.

So the search was extended further, and the solution was found by Alexander Grothendieck around 1960: he invented étale cohomology, and was thus able to justify the formal reasoning above leading to the rationality of the Zeta function by means of a Lefschetz trace formula. The complete treatment required years of work in the form of yearly *Séminaires de Géométrie Algébrique* held in Orsay, ultimately producing the infamous S.G.A notes (about 4000 pages) which contained all necessary details and much more.

The idea is simply brilliant. Looking back at Čech cohomology, remember what the motivation was: we felt that we knew enough about the local structure of our manifolds (since they are locally the same as $\mathbf{C}^d$) to assume we can solve our problems restricted to some sufficiently small covering and use Čech cohomology with respect to this covering in order to compare our various results and deduce global statements. But Zariski topology, even taking coverings by affine varieties, simply *doesn't* look locally any simpler than it does globally for arithmetic problems, such as counting points in our preferred field, be it $\mathbf{Q}$ or $\mathbf{F}_q$ (what good will it do to remove a point from a curve?).

Is there then a process which produces from X/k a variety which we may fairly call arithmetically simpler, and consider as a "localization" of X? Grothendieck's idea is that looking at X over extension fields of k, and especially over the separable closure of k, is such a process. This can create points on the variety, and as we know from many examples, most problems get really easier: for example, quadratic forms are classified by one invariant only, their rank, all polynomials have roots... There is a real feeling that we can consider those problems over $\bar{k}$ as known.

So we will change the idea of localizing to mean just something like that; in other words, we replace the notion of open subset by something more general, which still satisfies some formal properties sufficient to introduce Čech cohomology in a way analogous to the classical definition. This is called a Grothendieck topology; in our case, we are introducing the étale topology on a variety X/k defined over a field k. [3]

[3] Here the language of schemes [Har] is almost indispensable, since it makes it possible to talk of a variety over a field without any reference to points in an algebraic closure, and to define maps between such objects, even if there are no points in the variety with coefficients

A last way of thinking of this might be the following: considering an open subset $U_f = \{x \in X \mid f(x) \neq 0\}$ in the Zariski topology means algebraically that we allow ourselves to use the inverse of a function f, which wasn't defined before. In the étale case, we further allow ourselves the use of solutions of (separable) polynomial equations in f: its square root, for instance.

Formally, the notion of an étale morphism $f : X \to Y$ between varieties is introduced first. The precise definition (f is flat and unramified, if you really must know), we leave aside. Such a map is however always open, so $U = f(X)$ is a Zariski open subset of Y and X is then more or less a finite unramified covering of U. In particular, if k'/k is a separable field extension, then extending scalars from k to k' always gives an étale morphism $X \times k' \to X$. This adds solutions of algebraic equations in the base field. Of course not all étale maps are of this type. Others (as mentioned) give roots of equations whose coefficients are functions on U: they have Y of the form

$$Y = \{(x, y) \in U \times k \mid y^d + a_1(x)y^{d-1} + \ldots + a_d(x) = 0\}$$

with $Y \to X$ the projection; here the a_i are regular functions on U, and the equation has to be separable.

In particular, f is not injective in general, although its fibers are finite, so that in particular $\dim X = \dim Y$.

Étale morphisms will be considered then as "open" subsets in the étale topology. To justify this, we need to be able to take unions and intersections of them, and this is not very hard.

First, the union of any family $(X_i \to X)$ is just the obvious map from the disjoint union of the X_i's to X.

Second, if $f_1 : X_1 \to X$ and $f_2 : X_2 \to X$ are two étale maps, then let $Z = \{(x_1, x_2) \in X_1 \times X_2 \mid f_1(x_1) = f_2(x_2) \in X\}$, and $f : Z \to X$ be given by $f(x_1, x_2) = f_1(x_1) = f_2(x_2)$: we call Z the "intersection" of X_1 and X_2. You can check that if f_1 and f_2 are the injections of some open subsets into X, then this coincides with the usual notion.

Moreover, the notion of étale maps is transitive, which makes it possible to speak of the restriction of an étale map: if $Z \to Y \to X$ are two étale maps, the composite is still étale, and we say that it is the restriction to Z of the "open subset" $Y \to X$. For instance, there are as usual two restrictions from the intersection Z defined above to X_1 and X_2.

And finally an étale covering of X is just a family $(X_i \to X)$ of étale maps to X such that their images cover X:

$$X = \bigcup_i f_i(X_i).$$

This is now enough to rewrite verbatim the notion of a sheaf (of abelian groups) over X for the étale topology: it is a rule $\mathcal{F}$ associating an abelian group $\mathcal{F}(Y)$ to each étale map $Y \to X$, which is compatible with restriction, and which is local, where local is meant with respect to étale coverings and étale intersections.

With a sheaf $\mathcal{F}$ and a covering $\mathcal{U}$, we can then define the Čech cohomology groups $H^i(\mathcal{U}, \mathcal{F})$. Again, locality will imply

$$H^0(\mathcal{U}, \mathcal{F}) = \mathcal{F}(X)$$

in the base field.

(id : $X \to X$ is étale, so it is defined).

Yet again, general results show that it is possible to take the "union" over all étale coverings, getting cohomology groups associated to the variety X/k, written

$$H^i_{\text{ét}}(X, \mathcal{F}).$$

Some examples are in order here. The only obvious sheaves in this topology are the constant ones, similar to $\mathbf{C}$ in the original definition of Čech cohomology; as before, constant means *locally* constant, so if M is an abelian group, the constant sheaf $\underline{M}$ is defined by

$$\underline{M}(Y) = \text{Map}(Y, M),$$

the group of continuous maps $Y \to M$, where M has the discrete topology, so any $f \in \underline{M}(Y)$ is constant on each connected component of Y. In this case, the cohomology groups are simply denoted by $H^i_{\text{ét}}(X, M)$. In particular, if n is any positive integer there is the constant sheaf $\underline{\mathbf{Z}/n\mathbf{Z}}$ on X/k, which will be very important.

Another important sheaf is the "multiplicative group" $\mathbf{G}_m$, where $\mathbf{G}_m(Y)$, for any étale map $Y \to X$, is the group of invertible regular functions on Y. For any n prime to the characteristic of k, there is a subgroup μ_n of $\mathbf{G}_m$ consisting of regular functions which are n-th roots of unity. Notice that by extending k to $k[\xi]$, where $\xi \in \bar{k}$ is a primitive n-th root of unity, which means restricting to the "open" subset $X \times k[\xi] \to X$, μ_n becomes constant, isomorphic to $\mathbf{Z}/n\mathbf{Z}$. We thus see a case of a sheaf which is locally constant for the étale topology, whereas it certainly isn't, in general, in the Zariski topology. Also, n is assumed prime to the characteristic so that the polynomial $X^n - 1$ is separable over k.

We have not proved that $\mathbf{G}_m$, or μ_n, is a sheaf: indeed, because étale morphisms are not just injections, this is by no means obvious (see [Mil], chapter 2).

We can see what it involves in the simplest circumstance, where X is just a one point space, but defined over k. This means that although X is as trivial as you can get as a topological space, it isn't in the étale topology: you get many étale maps $Y \to X$ by taking Y to be a one point space defined over k', where k' is a finite separable extension of k. In this case, we have $\mu_n(Y) = \mu_n(k')$, the group of n-th roots of unity in k'.

Any such $Y \to X$ gives a covering of X since there is just one point in both spaces. However, the subtlety is that although there is just one map in the covering, the intersection of Y with itself is not Y! This is because Y has non-trivial automorphisms, namely (assuming k'/k to be Galois) the elements of the Galois group $\text{Gal}(k'/k)$. Translating correctly the definitions one indeed shows that the condition for μ_n (or $\mathbf{G}_m$) to be a sheaf over k is equivalent to the fundamental result of Galois theory: $x \in k'$ is in k if and only if $x^\sigma = x$ for all $\sigma \in \text{Gal}(k'/k)$.

This gives the first connection between this étale business and number theory, and the whole theory might be viewed as a generalization of Galois theory to higher dimensions. Indeed, pursuing the investigation of sheaves over a one point space over k shows that such an object is equivalent to an abelian group M with an action of the Galois group $\text{Gal}(k^s/k)$ of a separable closure of k, and the cohomology groups of the corresponding sheaf are the same as the Galois cohomology groups $H^i(k, M)$, which had been defined and extensively studied already, independently of any work on the Weil conjectures.

3.3 A concrete example: elliptic curves

There is yet another nice example, where we can actually compute a non-trivial étale cohomology group and so get a better feeling, and it involves the usual suspects: elliptic curves (see [Sil] for facts about elliptic curves). To simplify we consider elliptic curves over an algebraically closed field.

We have to consider all étale coverings $\mathcal{U}$ of E, compute the associated Čech cohomology groups and take the "union". The reason it is possible to do so explicitly in this case is that because E is an elliptic curves, all its étale coverings will be given by isogenies $E' \to E$, E' being another elliptic curve, and the theory of those is well-known.

Let $f : Y \to E$ be an étale map. Then Y is a curve over k, and it is smooth (this is a general property of étale morphisms). Because we are dealing with curves, it's not hard to reduce to Y projective. Assume first that it is connected. Then since E is an elliptic curve, it has genus one, and f is étale, therefore unramified, so the Hurwitz formula ([Sil], II 5.9)

$$2 - 2g(Y) = \deg(f)(2 - 2g(E)) = 0$$

implies that the genus of Y is again one: this is the special property of elliptic curves which is the main point. After choosing a suitable base point on Y, f is an isogeny between elliptic curves. We will write E' instead of Y for this elliptic curve. Since f is non-constant and therefore surjective, it defines an étale covering $\mathcal{U} = (E' \to E)$.

We now proceed to compute the Čech cochains for this étale covering and coefficient sheaf $\underline{M}$, for any abelian group M at this stage. Since E' is connected, we have

$$C^0(\mathcal{U}, \underline{M}) = M.$$

Next let E_1 be the intersection of $E' \to E$ with itself; as was the case with fields, this is not just E'. Let's compute:

$$E_1 = \{(x, y) \in E' \times E' \mid f(x) = f(y)\} \tag{10}$$

but f is an isogeny, therefore also a group morphism, and denoting $E'[f] = \mathrm{Ker}(f)$ (it is a finite subgroup of E of order $\deg f$ since f is separable) we get an isomorphism

$$\left\{ \begin{array}{ccc} E' \times E'[f] & \to & E_1 \\ (x, \sigma) & \mapsto & (x, x - \sigma) \end{array} \right. \tag{11}$$

(with inverse $(x, y) \mapsto (x, x - y)$.)

Thus E_1 is not connected, and we get

$$C^1(\mathcal{U}, \underline{M}) = \mathrm{Map}(E_1, M) = \mathrm{Map}(E'[f], M) \tag{12}$$

the last Map being simply the set theoretic applications $E'[f] \to M$, and the last equality is actually the bijection induced by the isomorphism (11), so if $c \in \mathrm{Map}(E_1, M)$, it corresponds to $\tilde{c}$ which has

$$\tilde{c}(\sigma) = f(0, \sigma).$$

Similarly, we compute E_2, the "triple intersection":

$$E_2 = \{(x,y,z) \in E' \times E' \times E' \mid f(x) = f(y) = f(z)\}$$

and $E_2 \simeq E' \times E'[f] \times E'[f]$ through the map

$$(x, \sigma, \tau) \mapsto (x, x - \sigma, x - \tau) \tag{13}$$

so that

$$C^2(\mathcal{U}, \underline{M}) = \mathrm{Map}(E_2, M) = \mathrm{Map}(E'[f] \times E'[f], M) \tag{14}$$

and this time $c \in \mathrm{Map}(E_2, M)$ corresponds to $\tilde{c}$ with

$$\tilde{c}(\sigma, \tau) = c(0, \sigma, \tau).$$

To compute $H^0(\mathcal{U}, M)$ and $H^1(\mathcal{U}, M)$ we have the sequence

$$0 \to C^0(\mathcal{U}, \underline{M}) \overset{d^0}{\to} C^1(\mathcal{U}, \underline{M}) \overset{d^1}{\to} C^2(\mathcal{U}, \underline{M})$$

which is therefore of the form

$$0 \to M \overset{d^0}{\to} \mathrm{Map}(E'[f], M) \overset{d^1}{\to} \mathrm{Map}(E'[f] \times E'[f], M) \tag{15}$$

and it remains to determine d^0 and d^1.

For d^0, it's easy: if r_1, r_2 are the two restrictions $E_1 \overset{\to}{\to} E$, namely the two projections in (10), and $c \in C^0(\mathcal{U}, \underline{M})$, then c is constant on E and $d^0(c) \in C^1(\mathcal{U}, \underline{M})$ is

$$d^0(f)(x,y) = f \circ r_1(x,y) - f \circ r_2(x,y) = f(x) - f(y) = 0$$

i.e. $d^0 = 0$, and

$$H^0(\mathcal{U}, M) = M.$$

For d^1, we have the three restrictions s_i $(1 \leqslant i \leqslant 3)$ from E_2 to E_1, which are, in the two descriptions of E_2 and E_1

$$\begin{array}{rlcrl} s_1 : & (x,y,z) \mapsto (x,y) & \text{or} & s_1 : & (x,\sigma,\tau) \mapsto (x,\sigma) \\ s_2 : & (x,y,z) \mapsto (x,z) & \text{or} & s_2 : & (x,\sigma,\tau) \mapsto (x,\tau) \\ s_3 : & (x,y,z) \mapsto (y,z) & \text{or} & s_3 : & (x,\sigma,\tau) \mapsto (x-\sigma, \tau-\sigma) \end{array} \tag{16}$$

which now gives the formula for d^1 in (15): for $\tilde{c} \in \mathrm{Map}(E'[f], M)$

$$\begin{array}{rcl} d^1(\tilde{c})(\sigma,\tau) & = & \tilde{c} \circ s_1(\sigma,\tau) - \tilde{c} \circ s_2(\sigma,\tau) + \tilde{c} \circ s_3(\sigma,\tau) \\ & = & \tilde{c}(\sigma) - \tilde{c}(\tau) + \tilde{c}(\tau - \sigma). \end{array}$$

Because $d^0 = 0$, we have $H^1(\mathcal{U}, M) = \mathrm{Ker}(d^1)$, and so we conclude exactly from this formula that

$$H^1(\mathcal{U}, M) = \mathrm{Hom}(E'[f], M) \tag{17}$$

the group of *group* homomorphisms from $E'[f]$ to M.

Now for the real $H^*_{\text{ét}}(E, M)$: any general covering $(U_i \to E)$ must contain a U_i as above, and so has a refinement by one of this form. Since the H^0 found above doesn't depend on $\mathcal{U}$, we conclude in every case that

$$H^0_{\text{ét}}(E, M) = M$$

(which is really a general property of cohomology and has nothing to do with E.)

On the other hand, for H^1, we can notice some non-obvious phenomena which throw much light on the general theory. First, since $E'[f]$ is a finite group, we have $H^1(\mathcal{U}, M) = 0$ if M is without torsion, for instance $M = \mathbf{Z}$, so $H^1_{\text{ét}}(E, \mathbf{Z}) = 0$. This is anomalous if we compare with the situation over $\mathbf{C}$, so shows that we are forced to consider torsion sheaves, such as $\underline{\mathbf{Z}/m\mathbf{Z}}$, if we want to get a good theory, and $H^1_{\text{ét}}(E, M)$ will be necessarily a torsion group. This might seem a big step backwards in our quest for a good theory for the Weil conjectures, but the point is that the torsion here is related to M and not to p: in a way similar to the construction of the p-adic fields, which are of characteristic zero, from the consideration of all $\mathbf{Z}/p^n\mathbf{Z}$, $n \geqslant 1$, we will be able to find a way out.

Indeed, assume now that $M = \mathbf{Z}/\ell^n\mathbf{Z}$ for $\ell \neq p$ a prime number. Then

$$\mathrm{Hom}(E'[f], \mathbf{Z}/\ell^n\mathbf{Z}) = \mathrm{Hom}(E'[f]_\ell, \mathbf{Z}/\ell^n\mathbf{Z})$$

and the ℓ-part $E'[f]_\ell$ of $E'[f]$ is the kernel of another étale cover $E'' \to E$, so in computing $H^1_{\text{ét}}(E, \mathbf{Z}/\ell^n\mathbf{Z})$ we can restrict our attention to the coverings f with degree a power of ℓ, say ℓ^m.

Now the theory of the dual isogeny ([Sil], III 6) furnishes another isogeny

$$\hat{f} \,:\, E \to E'$$

such that the composite

$$E \xrightarrow{\hat{f}} E' \xrightarrow{f} E$$

is the map $[\ell^m] \,:\, E \to E$. This is also étale, showing that $[\ell^m]$ refines $\mathcal{U}$, and we only have to consider those $[\ell^m]$. Then we know ([Sil], III 6-4) that

$$E[\ell^m] \simeq \mathbf{Z}/\ell^m\mathbf{Z} \times \mathbf{Z}/\ell^m\mathbf{Z}$$

and consequently

$$\begin{aligned} H^1(\mathcal{U}, \mathbf{Z}/\ell^n\mathbf{Z}) &= \mathrm{Hom}(\mathbf{Z}/\ell^m\mathbf{Z} \times \mathbf{Z}/\ell^m\mathbf{Z}, \mathbf{Z}/\ell^n\mathbf{Z}) \\ &= \begin{cases} \mathbf{Z}/\ell^m\mathbf{Z} \times \mathbf{Z}/\ell^m\mathbf{Z}, \text{ if } m \leqslant n. \\ \mathbf{Z}/\ell^n\mathbf{Z} \times \mathbf{Z}/\ell^n\mathbf{Z}, \text{ if } m \geqslant n. \end{cases} \end{aligned}$$

and taking all m's together gives finally

$$H^1_{\text{ét}}(E, \mathbf{Z}/\ell^n\mathbf{Z}) = \mathbf{Z}/\ell^n\mathbf{Z} \times \mathbf{Z}/\ell^n\mathbf{Z}$$

which is this time exactly the analogue of the classical result

$$H^1(E_{\mathbf{C}}, \mathbf{C}) = \mathbf{C} \times \mathbf{C}.$$

This is the general pattern: only cohomology with coefficients in torsion sheaves, with order prime to the characteristic of the field, will behave as expected, but taking all $\mathbf{Z}/\ell^n\mathbf{Z}$, $n \geqslant 1$, the corresponding groups $H^i_{\text{ét}}(X, \mathbf{Z}/\ell^n\mathbf{Z})$ will come together correctly so a theory with coefficients in the ring of ℓ-adic integers will be obtained by *defining*

$$H^i_{\text{ét}}(X, \mathbf{Z}_\ell) = \varprojlim H^i_{\text{ét}}(X, \mathbf{Z}/\ell^n\mathbf{Z})$$

(which are $\mathbf{Z}_\ell$-modules) and one over the ℓ-adic field by

$$H^i_{\text{ét}}(X,\mathbf{Q}_\ell) = H^i_{\text{ét}}(X,\mathbf{Z}_\ell)\otimes\mathbf{Q}_\ell$$

(and this is extended to a certain category of étale sheaves on X, called ℓ-adic constructible sheaves.)

For elliptic curves, our calculation shows that

$$H^1_{\text{ét}}(E,\mathbf{Z}_\ell) = \mathbf{Z}_\ell^2$$

and even more precisely gives an isomorphism

$$H^1_{\text{ét}}(E,\mathbf{Z}_\ell) \simeq \text{Hom}(T_\ell(E),\mathbf{Z}_\ell)$$

so $H^1_{\text{ét}}$ is simply the dual of the Tate module of E at ℓ. You can now compare the proof of the rationality of the zeta function of an elliptic curve in [Sil] with the one derived from the Lefschetz trace formula for ℓ-adic étale cohomology.

3.4 The fundamental results of étale cohomology

Working hard, Grothendieck and his collaborators succeeded in establishing the fundamental properties of this étale cohomology theory, for smooth projective varieties over an algebraically closed field of characteristic p (possibly $p=0$) and any prime $\ell\neq p$:

- "Cohomological dimension": if X is of dimension d, then $H^i_{\text{ét}}(X,\mathbf{Q}_\ell)=0$ for all $i>2d$.
- Finiteness: for any $i\geqslant 0$, $H^i_{\text{ét}}(X,\mathbf{Q}_\ell)$ is a finite dimensional $\mathbf{Q}_\ell$-vector space.
- "Functoriality in X": if $f : X\rightarrow Y$ is a morphism then there are associated maps in cohomology for all $i\geqslant 0$

  $$f^\star : H^i_{\text{ét}}(Y,\mathbf{Q}_\ell)\rightarrow H^i_{\text{ét}}(X,\mathbf{Q}_\ell).$$

  In particular, a map $X\rightarrow X$ induces endomorphisms of the finite dimensional $\mathbf{Q}_\ell$-vector spaces $H^i(X,\mathbf{Q}_\ell)$.
- Poincaré duality: there is an isomorphism

  $$\text{tr} : H^{2d}_{\text{ét}}(X,\mathbf{Q}_\ell)\simeq\mathbf{Q}_\ell$$

  and for all $i\leqslant d$ natural bilinear forms ("cup product")

  $$H^i_{\text{ét}}(X,\mathbf{Q}_\ell)\times H^{2d-i}_{\text{ét}}(X,\mathbf{Q}_\ell)\rightarrow H^{2d}_{\text{ét}}(X,\mathbf{Q}_\ell)\simeq\mathbf{Q}_\ell$$

  which are perfect pairings of finite dimensional vector spaces; in particular, $H^i_{\text{ét}}$ is dual to $H^{2d-i}_{\text{ét}}$, and they have the same dimension.
- Künneth theorem: for any $i\geqslant 0$, there are canonical isomorphisms

  $$H^i_{\text{ét}}(X\times Y,\mathbf{Q}_\ell) = \bigoplus_{j+k=i} H^j_{\text{ét}}(X,\mathbf{Q}_\ell)\otimes H^k_{\text{ét}}(X,\mathbf{Q}_\ell).$$

- Lefschetz trace formula: let $f : X \to X$ be a morphism map with isolated fixed points, satisfying a certain separability assumption on $1 - df$ acting on the tangent spaces at the fixed points, and $L(f)$ the number of those. Then we have the equality

$$L(f) = \sum_{i=0}^{2d} (-1)^i \mathrm{Tr}(f^\star \mid H^i_{\text{ét}}(X, \mathbf{Q}_\ell)). \tag{18}$$

- Comparison theorem: if X is smooth and projective over $\mathbf{C}$, then

$$H^i_{\text{ét}}(X, \mathbf{Q}_\ell) \otimes_{\mathbf{Q}_\ell} \mathbf{C} \simeq H^i(X_{\mathbf{C}}, \mathbf{C}).$$

This shows that the results for ordinary cohomology translate beautifully, and from those results most of the formal computations of cohomology groups can be adapted.

Even over $\mathbf{C}$, there is an important advantage in using étale cohomology groups instead of the usual ones: if a variety is defined over a subfield k of $\mathbf{C}$, then the automorphisms of $\mathbf{C}$ over k will act naturally on the étale groups, but not on the complex ones. This is easy to see in the description with Čech cohomology. For elliptic curves, this action is (dual to) the usual action of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ on the Tate module, whose importance is well-known.

Remarks. I don't know much about the details of the proofs. Every statement here is a hard theorem. Looking at the explanations in [De3], it seems that the strategy is consistently to find a proof of the classical statements which is sufficiently intrinsic to translate to the algebraic case in principle. Then every step is usually much harder, and after many (sometimes sophisticated) reductions, the proof is reduced to an arithmetic statement (which sometimes happens to have been already considered independently), which is proved more or less directly.

Another remark is that we have here introduced an auxiliary prime ℓ, with the outcome that we get infinitely many different cohomology theories, and tough questions of dependence, or independence, on ℓ will appear. It would be so much nicer to be able to do the same with $\ell = p$, but you can check with elliptic curves that it doesn't work: this is similar to the fact that the Tate module at p is not $\mathbf{Z}_p^2$, but sometimes $\mathbf{Z}_p$ and sometimes even 0 (for the "supersingular" elliptic curves).

This being done, we can come back to the formal proof of the rationality of the zeta function of X_0, for X_0 smooth and projective over $\mathbf{F}_q$, and see that is becomes perfectly justified. We obtain the formula

$$Z(X_0) = \prod_{i=0}^{2d} \det(1 - F^\star T \mid H^i_{\text{ét}}(X, \mathbf{Q}_\ell))^{(-1)^{i+1}}. \tag{19}$$

discovering how the right hand side which seems to be a rational function in $\mathbf{Q}_\ell(T)$ is actually one in $\mathbf{Q}(T)$, independent of ℓ (one needs a small lemma, to the effect that $\mathbf{Q}[[T]] \cap \mathbf{Q}_\ell(T) = \mathbf{Q}(T)$.)

The functional equation (4) follows also quite easily from Poincaré duality, with the extra information that the map $F^\star$ induced by the Frobenius map respects the duality pairings, and on $H^{2d}_{\text{ét}}(X, \mathbf{Q}_\ell) \simeq \mathbf{Q}_\ell$ is simply multiplication by q^n (see [Har], Appendix C, for the lemma from linear algebra which verifies this).

And the Riemann hypothesis, one suddenly wonders?

There was no obvious classical analogue to it, and the statements above do not seem to contain a clue. Grothendieck, however, identified a theorem of Hodge and Lefschetz whose analogue, if it could be proved, would nicely imply the Riemann hypothesis. This became known as one of the "standard conjectures", and it looked so much harder than the rest that a solution seemed as far away in 1970 as ten years before. Then Deligne found his proof which had nothing to do with this approach, giving a nice illustration of the irreducible unexpectedness of life.

So we have come all this long way to begin the real subject of these notes.

(Part II : Summary of the proof)

Summary of part I (if you missed the first episode). By a somber winter night in 1949, André Weil, reading dusty manuscripts of C.-F. Gauss, discovers the existence of a spectacular treasure lying somewhere in the deep jungles of arithmetic geometry. He manages to sketch a map to reach it but is unable to mount an expedition to embark on the perilous journey... Ten years later, the audacious explorator Alexander Grothendieck finds the secret passage that may lead to the spot marked with an ×, and advancing fearlessly with his small group of indefatigable geometers, attains the first points marked on Weil's map. However the trove where lies the priceless jewel stands behind a towering mountain which seems to defy hope. The clever Belgian, Deligne, then goes out on his own by a circuitous route...

4 Outline of the proof

We have now come to the point where we can think of the Riemann Hypothesis. Recall its statement: let X_0 be a smooth projective variety of dimension d defined over a finite field $\mathbf{F}_q$ of characteristic $p > 0$, and $Z(X_0)$ its zeta function

$$Z(X_0) = \exp\Big(\sum_{n\geqslant 1} |X_0(\mathbf{F}_{q^n})| \frac{T^n}{n}\Big)$$

which we know to be actually a rational function.[4] Then there should exist polynomials $P_i(X_0)$ with integer coefficients such that

$$Z(X_0) = \frac{P_1(X_0)\ldots P_{2d-1}(X_0)}{P_0(X_0)\ldots P_{2d}(X_0)}$$

and all the complex zeros of P_i have modulus $q^{-i/2}$, in other words they can be written as

$$P_i = \prod_j (1 - \alpha_{i,j} T)$$

with $|\alpha_{i,j}| = p^{i/2}$.

Now Grothendieck's theory of ℓ-adic étale cohomology gives an expression of this shape, for any choice of a prime $\ell \neq p$, with polynomials given explicitly by

$$P_i(X_0, \ell) = \det(1 - F^\star T \,|\, H^i_{\text{ét}}(X, \mathbf{Q}_\ell)) \tag{20}$$

namely (almost) the characteristic polynomial of the endomorphism of the ℓ-adic étale cohomology spaces of $X = X_0 \times \overline{\mathbf{F}}_q$ induced by the Frobenius morphism $F : X \to X$.

Those polynomials, however, are seemingly in $\mathbf{Q}_\ell[T]$, not in $\mathbf{Z}[T]$, and might conceivably not be the ones whose existence is surmised in the Riemann Hypothesis. The following conjecture is therefore a refinement of the original one:

[4] This was actually proved by Dwork before Grothendieck completed his program to show it with étale cohomology; Dwork's proof was basically direct, but didn't seem to give any hope for the Riemann Hypothesis.

Conjecture 1 *The polynomials $P_i(X_0, \ell)$ have integral coefficients which do not depend on the choice of ℓ, and all their complex roots have absolute value $q^{-i/2}$ (equivalently, from (20), the eigenvalues of $F^\star$ on $H^i_{ét}(X, \mathbf{Q}_\ell)$ all have modulus $q^{i/2}$.)*

This is the statement that was proved by Deligne. However, because the statement about the modulus of the roots distinguishes uniquely the roots of P_i among the zeros or poles of the zeta function, independently of ℓ, it is easy to prove that this conjecture is equivalent to the following result:

Theorem 1 *(Deligne, [De1]).*
For all i and all $\ell \neq p$, the eigenvalues of $F^\star$ on $H^i_{ét}(X, \mathbf{Q}_\ell)$ are algebraic numbers all complex conjugates of which have modulus $q^{i/2}$.

The remainder of these notes will try to summarize the proof of this theorem.

Let us ponder how this might be approached. We know (or will know) that this is true for curves, as was proved by Weil himself. Moreover, one of the basic properties of étale cohomology, the Künneth theorem, describes the cohomology of a product $X \times Y$ in terms of that of X and Y

$$H^i_{\text{ét}}(X \times Y, \mathbf{Q}_\ell) \simeq \bigoplus_{j+k=i} H^j_{\text{ét}}(X, \mathbf{Q}_\ell) \otimes H^k_{\text{ét}}(X, \mathbf{Q}_\ell)$$

and this isomorphism is compatible with the action of the Frobenius, which implies that if the theorem is true for X and Y, it is also for $X \times Y$ (in terms of zeta functions, the Künneth theorem says that $Z(X_0 \times Y_0)$ is the "Rankin-Selberg convolution" of $Z(X_0)$ and $Z(Y_0)$, in the sense that its zeros (resp. poles) are all possible products of one of the first factor and one of the second). If any variety could be written as a product of a curve and another variety, we could therefore argue by induction on the dimension and call it a day.

It is of course not so, but the proof will take a cue from this and work by induction (although in the end it will not require the use of the previously known result for curves), and by trying to make the induction work by finding a map $f : X \to \mathbf{P}^1$ which looks as much as possible like the first projection would if X were isomorphic to $\mathbf{P}^1 \times Y$. This, in a sense, will fail only at a finite number of points, and a careful analysis (motivated, once again, by a classical approach developed by Lefschetz to compute the cohomology of complex varieties) of the relations between the cohomologies of the fibers of f at those points and the global cohomology of X will bring a dramatic reduction to a last irreducible case, where Deligne's reading of the paper of Rankin introducing the Rankin-Selberg convolution will contribute the spark which ultimately connects the last strands (block that metaphor!).

Until the moment comes for this crucial ingredient, the argument will be mostly geometric. Arithmetic enters the stage as a classical *deus ex machina*, which seems adequate to conclude this tale worthy of the greatest Greek tragedies.

5 Geometric reductions

As a matter of notation when discussing these reduction steps, it will be convenient to use the shorthand RH to allude to the full theorem 1, but also

$\mathrm{RH}(X_0)$ to mean its restriction to the single variety X_0, or $\mathrm{RH}^i(X_0)$ for the further restriction to the eigenvalues of $F^\star$ to the single étale cohomology group $H^i_{\text{ét}}(X, \mathbf{Q}_\ell)$, and even, if $V \subset H^i_{\text{ét}}(X, \mathbf{Q}_\ell)$ is stable by the action of $F^\star$, $\mathrm{RH}(V)$ for the restriction to the eigenvalues of $F^\star$ restricted to V.

The following lemma is elementary but nevertheless repeatedly used during the reduction steps.

Lemma 1 *Given two subspaces $V \subset H^i_{\acute{e}t}(X, \mathbf{Q}_\ell)$ and $W \subset H^j_{\acute{e}t}(Y, \mathbf{Q}_\ell)$ stable by $F^\star$, and a $\mathbf{Q}_\ell$-linear map $\varphi : V \to W$ which satisfies for all $v \in V$*

$$\varphi(F^\star v) = q^{(i-j)/2} F^\star \varphi(v)$$

(such maps will be called compatible with $F^\star$) the following implications hold:

(i) If φ is injective, then $\mathrm{RH}(W)$ *implies* $\mathrm{RH}(V)$*.*

(ii) If φ is surjective, then $\mathrm{RH}(V)$ *implies* $\mathrm{RH}(W)$*.*

The geometric reductions that will now be put into effect are based on the idea of studying the cohomology of a projective variety $X_{\mathbf{C}} \subset \mathbf{P}^n_{\mathbf{C}}$ by relating it to that of its hyperplane sections $Y = H \cap X$, for hyperplanes $H \subset \mathbf{P}^n$, and using induction on the dimension. The first step in this direction goes back to the classical italian geometers. Bertini proved (over $\mathbf{C}$) that for H generic, the hyperplane section Y is a smooth projective variety, of dimension $d - 1$. Over a finite field, this remains true, possibly after making a field extension $\mathbf{F}_{q^n}/\mathbf{F}_q$. But as in the proof of Weil's estimate for Kloosterman sums, such finite extensions do not matter for proving the Riemann Hypothesis, so we will assume that they are done without mentioned the need to.

Given such a smooth hyperplane section Y, Lefschetz showed first that Y "contained" all the information about the cohomology of X, except for the middle dimensional group $H^d(X, \mathbf{C})$. This is the weak Lefschetz theorem, which extended to étale cohomology says the following

Theorem 2 *For all $i \geqslant 2$ there are maps*

$$H^{i-2}_{\acute{e}t}(Y, \mathbf{Q}_\ell) \to H^i_{\acute{e}t}(X, \mathbf{Q}_\ell)$$

compatible with $F^\star$, and we have:

(i) For $i = d+1$, $H^{d-1}_{\acute{e}t}(Y, \mathbf{Q}_\ell) \to H^{d+1}_{\acute{e}t}(X, \mathbf{Q}_\ell)$ is surjective.

(ii) For $i > d+1$, $H^{i-2}_{\acute{e}t}(Y, \mathbf{Q}_\ell) \to H^i_{\acute{e}t}(X, \mathbf{Q}_\ell)$ is an isomorphism.

(In geometric terms over $\mathbf{C}$, those maps are "cup-product with the cohomology class of the hyperplane").

With this we can reduce RH to RH^d, the corresponding statement limited to the middle-dimensional cohomology group $H^d_{\text{ét}}$. Indeed, argue by induction on the dimension, starting from dimension 0 which is trivial. Assuming $\mathrm{RH}(Y_0)$ for all Y_0 of dimension $\leqslant d-1$, and given X_0 of dimension d, we first use Poincaré duality to show that $\mathrm{RH}^{2d-i}(X_0)$ is equivalent to $\mathrm{RH}^i(X_0)$. Then we take a smooth hyperplane section Y_0 of X_0 and apply the theorem (and the lemma): it follows that $\mathrm{RH}^{d+1}(X_0)$ is implied by $\mathrm{RH}^{d-1}(Y_0)$ (the middle-dimensional case for Y_0), and that for $i > d+1$, $\mathrm{RH}^i(X_0)$ is equivalent to $\mathrm{RH}^{i-2}(Y_0)$. But both are true by the induction hypothesis.

One advantage of this reduction is that we are left with only one cohomology group to consider for a given X_0. In a sense, this is the kind of situation where "analytic methods" may apply, in the same way as counting points could prove the Riemann Hypothesis for curves because only $H^1_{\text{ét}}$ was non-trivial. In this spirit, we now show that RH^d, which asserts that the eigenvalues have a precise modulus, is a consequence of a weaker result asserting only an inequality.

More precisely, let $\boldsymbol{rh}^d$ stand for this statement:

For any smooth projective and even dimensional variety X_0, all eigenvalues of $F^\star$ acting on $H^d_{\acute{e}t}(X, \mathbf{Q}_\ell)$ are algebraic numbers, and if α is a complex conjugate of one, then

$$q^{\frac{d}{2}-\frac{1}{2}} \leqslant |\alpha| \leqslant q^{\frac{d}{2}+\frac{1}{2}}.$$

Proposition 1 RH *is implied by* $\boldsymbol{rh}^d$.

Proof.

Let α be an eigenvalue of $F^\star$ acting of $H^d_{\text{ét}}(X_0, \mathbf{Q}_\ell)$, β a complex conjugate of α. From the Künneth formula it follows that for any $k \geqslant 1$, α^k is an eigenvalue of $F^\star$ on $H^{kd}_{\text{ét}}(X^k, \mathbf{Q}_\ell)$. If k is even, by $\boldsymbol{rh}^d$ applied to X_0^k we have

$$q^{\frac{kd}{2}-\frac{1}{2}} \leqslant |\alpha^k| \leqslant q^{\frac{kd}{2}+\frac{1}{2}}$$

and therefore

$$q^{\frac{d}{2}-\frac{1}{2k}} \leqslant |\alpha| \leqslant q^{\frac{d}{2}+\frac{1}{2k}}$$

which gives $\mathrm{RH}^d(X_0)$ when letting k go to infinity.

Then RH is a corollary of the previous reduction.

◇

Notations $\boldsymbol{rh}^d(X_0)\ldots$ will now be used as we proceed. Notice that lemma 1 is also valid with $\boldsymbol{rh}^d$ replacing RH, provided i and j are as required by the statement.

We now take an even dimensional X_0 as required. We have to investigate $H^d_{\text{ét}}(X, \mathbf{Q}_\ell)$. In the complex case, Lefschetz continued his use of hyperplane sections, but this time it is not enough to consider one only.

Choose an embedding $X \to \mathbf{P}^n$ for some $n \geqslant 1$. The nice geometric construction of Lefschetz is to take a linear subspace $A \subset \mathbf{P}^n$ of codimension 2, and consider together all the hyperplane sections $H \cap X$ where H contains A. The space D parametrizing those hyperplanes is isomorphic to $\mathbf{P}^1$, and we consider

$$\tilde{X} = \{(x, H) \in X \times D \,|\, x \in H\}.$$

This is again an algebraic variety and it comes with two natural maps

$$\begin{tikzcd} X & \tilde{X} \arrow[l, "\pi"'] \arrow[d, "f"] \\ & D \end{tikzcd}$$

Moreover, all this can be done over $\mathbf{F}_q$ by choosing A itself defined over $\mathbf{F}_q$, and we get corresponding $\tilde{X}_0$, D_0,...

If A is chosen so that $X \not\subset A$, then $A \cap X$ is a proper closed subset of X. For $x \in X \setminus A$, the conditions $x \in H$ and $H \supset A$ then determine H uniquely as the linear space spanned by x and A. Therefore π is an isomorphism over this open subset of X, showing that $\tilde{X}$ is birational to X. It is actually isomorphic to the blow-up of $X \cap A$ in $\tilde{X}$. From this, one can prove that there are injections (for all $i \geqslant 0$)

$$H^i_{\text{ét}}(X, \mathbf{Q}_\ell) \to H^i_{\text{ét}}(\tilde{X}, \mathbf{Q}_\ell)$$

and from lemma 1, we can restrict our attention to $\tilde{X}_0$. The main benefit from this operation is that we now have at our disposal the map

$$f_0 : \tilde{X}_0 \to D_0 \simeq \mathbf{P}^1_0$$

which tries to approximate $\tilde{X}_0$ as the product of a curve (namely, $\mathbf{P}^1_0$) and a smaller dimensional variety.

This map f is called a Lefschetz pencil. The name comes from the (obvious) fact that the fiber $\tilde{X}_t = f^{-1}(t)$ of f over a point $t \in \mathbf{P}^1$ corresponding to a hyperplane $H_t \subset \mathbf{P}^n$ is simply the section $H_t \cap X$.

We want the map f to be as nice as possible. From a picture, it is fairly clear that in general there will be points where f is singular: its differential vanishes. This corresponds to the fibers $\tilde{X}_s$ which are singular varieties. However, by geometric arguments, it is possible to establish the existence of a Lefschetz pencil $f : \tilde{X} \to D$ which has only a finite number of singular fibers, each having only one singular point which is an ordinary double point. (Over $\mathbf{C}$, this would mean that $\tilde{X}_s$ could be described in an affine neighborhood of x_s by a non-singular quadratic polynomial.)

The set of s for which $\tilde{X}_s$ is singular is denoted by S, and x_s, for $s \in S$, is the singular point in the fiber. Of course, $U = D \setminus S$ is an open subset of D.

The goal is now to understand the cohomology of $\tilde{X}$ with the help of f. At this point, the use of coefficient sheaves for étale cohomology, more general than $\mathbf{Q}_\ell$, becomes essential. The reason is that, as might be intuitively expected, the cohomology groups of $\tilde{X}$ will be related to the cohomology of the fibers of f (which are hyperplane sections of X, after all, so this goes well with the weak Lefschetz theorem), but in a way which takes into account the variation of those fibers, especially the presence of the singular ones (all the others being more or less the same). Thus there will be sheaves, called the higher direct images of f, denoted $R^i f_* \mathbf{Q}_\ell$, which describe this variation.

The precise relationship we seek is then expressed by means of a very technical algebraic device called the Leray spectral sequence of f. This is an iterative process, which – in this case – starts from the cohomology groups

$$E_2^{p,q} = H^p_{\text{ét}}(D, R^q f_* \mathbf{Q}_\ell)$$

(p, $q \geqslant 0$) and produces ultimately the sought-after

$$H^{p+q}_{\text{ét}}(\tilde{X}, \mathbf{Q}_\ell).$$

The main property of this process, which I have no intention to describe, is that it is also entirely compatible with the action of the Frobenius $F^\star$, which has the practical consequence that the eigenvalues of $F^\star$ at the end of the spectral sequence (on $H^d_{\text{ét}}(\tilde{X}, \mathbf{Q}_\ell)$) will satisfy $\boldsymbol{rh}^d$ if all those at the start on $E_2^{p,q}$ (with $p + q = d$) do, basically because of lemma 1 again.

Now, because D is a curve, for all "good" sheaves $\mathcal{F}$, and it can be proved that those higher direct images are good, the cohomology groups $H^i_{\text{ét}}(D, \mathcal{F})$ vanish for $i > 2$. So as far as $H^d_{\text{ét}}(\tilde{X}, \mathbf{Q}_\ell)$ is concerned, we have to investigate three terms only:

$$E_2^{0,d} = H^0_{\text{ét}}(D, R^d f_* \mathbf{Q}_\ell)$$

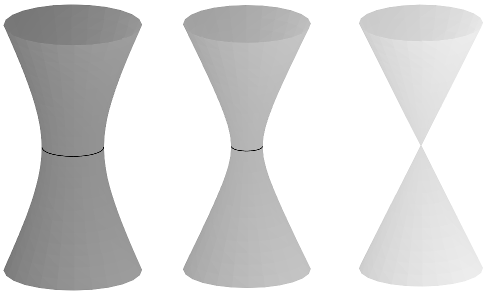

Figure 1: A vanishing cycle

$$E_2^{2,d-2} = H^2_{\text{ét}}(D, R^{d-2}f_*\mathbf{Q}_\ell)$$
$$E_2^{1,d-1} = H^1_{\text{ét}}(D, R^{d-1}f_*\mathbf{Q}_\ell).$$

As might be expected, the last one will be the troublesome one, since H^0 and H^2 of curves are much easier to deal with than H^1 in general. Notice also that $d-1$ is the dimension of the fibers of f.

Indeed, the analysis of those sheaves brings first the following proposition:

Proposition 2 *(i) For $i \neq d-1$, the sheaves $R^if_*\mathbf{Q}_\ell$ are constant, and they are isomorphic to the constant sheaf associated to the $\mathbf{Q}_\ell$- vector space $H^i_{\acute{e}t}(X_u, \mathbf{Q}_\ell)$, where u is any point in U.*

(ii) For $i = d-1$, the sheaf $R^{d-1}f_*\mathbf{Q}_\ell$ is determined by its restriction to U (in sheaf-theoretic notations, if $j : U \to D$ is the injection, we have $R^{d-1}f_*\mathbf{Q}_\ell = j_*j^*R^if_*\mathbf{Q}_\ell$).

The first point now takes care easily of the first two terms of the spectral sequence that we had to consider: for $E_2^{0,d}$, for instance, because $H^0_{\text{ét}}$ of a constant sheaf is the just the associated vector space, we deduce that

$$H^0_{\text{ét}}(D, R^df_*\mathbf{Q}_\ell) = H^d_{\text{ét}}(X_u, \mathbf{Q}_\ell)$$

but X_u is smooth and therefore admits a smooth hyperplane section Y; by the weak Lefschetz theorem 2, we have a surjection

$$H^{d-2}_{\text{ét}}(Y, \mathbf{Q}_\ell) \to H^d_{\text{ét}}(X_u, \mathbf{Q}_\ell)$$

compatible with $F^\star$ and we can apply an induction hypothesis to Y which is of even dimension $d-2$. The second term is dual to this one.

It remains to treat the third term $E_2^{1,d-1}$. Fix one $u \in U$; we now study the "fiber" of $R^{d-1}f_*\mathbf{Q}_\ell$ at u, which is simply the cohomology of the fiber, $H^{d-1}_{\text{ét}}(X_u, \mathbf{Q}_\ell)$. For $s \in S$, there is a map from the cohomology of the singular fiber X_s to that of X_u:

$$H^{d-1}_{\text{ét}}(X_s, \mathbf{Q}_\ell) \to H^{d-1}_{\text{ét}}(X_u, \mathbf{Q}_\ell).$$

This map is injective; in other words (Figure 1 tries to show it somehow – in the classical case –, but you need to know the other interpretation of cohomology as group of classes of "cycles", such as the circle in the figure, which disappears when the good fiber degenerates to the singular one), when we move from the "generic" fiber X_u to the "special" fiber X_s, some of the cohomology of X_u vanishes. The point is that it is possible to describe and study this vanishing part quite precisely, and it will prove to give enough information to complete the proof.

Namely, remember that because $d-1$ is the middle dimension for the cohomology of the fibers, Poincaré duality gives actually a bilinear form on $H^{d-1}_{\text{ét}}(X, \mathbf{Q}_\ell)$:

$$H^{d-1}_{\text{ét}}(X, \mathbf{Q}_\ell) \times H^{d-1}_{\text{ét}}(X, \mathbf{Q}_\ell) \to \mathbf{Q}_\ell$$

which is non-degenerate and, because $d-1$ is odd, is shown to be alternating.

We can therefore define the space of vanishing cycles at s to be the subspace

$$V_s = H^{d-1}_{\text{ét}}(X_s, \mathbf{Q}_\ell)^{\perp} \subset H^{d-1}_{\text{ét}}(X_u, \mathbf{Q}_\ell).$$

However, because the singularities of X_s are controlled, one proves that the vanishing part is small: actually, V_s is of dimension 1 exactly.

This is the local Lefschetz theory, for $s \in S$ fixed. Now considering all of S, we define the space of vanishing cycles ("cycles évanescents" in French) to be the subspace E of $H^{d-1}_{\text{ét}}(X_u, \mathbf{Q}_\ell)$ spanned by the V_s, $s \in S$.

We now want to extend this to all of U to get a subsheaf $\mathcal{E}$ over U describing the variation of the vanishing cycles, as $R^{d-1}f_*\mathbf{Q}_\ell$ describes that of the full cohomology group. This is not automatic; it depends on the action of a certain group, called the algebraic fundamental group of U and denoted $\pi_1(U, u)$, whose main property here is that to give a sheaf on U is the same as to give an abelian group on which $\pi_1(U, u)$ acts. This π_1 is to be thought of as a generalization of the absolute Galois group of a field. Indeed, for a smooth curve C, it is shown to be a quotient of the Galois group of the function field of C.

The abelian group here is E, and the theory of the fundamental group shows that it is generated by some subgroups I_s, one for each $s \in S$. Part of the local Lefschetz theory then studies the action of I_s on $H^{d-1}_{\text{ét}}(X_u, \mathbf{Q}_\ell)$ (the action of π_1 for which the associated sheaf is simply $R^{d-1}f_*\mathbf{Q}_\ell$.) The result is the Picard-Lefschetz formula:

Theorem 3 *One can choose a generator δ_s of V_s such that for all $\sigma \in I_s$ and $\gamma \in H^{d-1}_{\acute{e}t}(X_u, \mathbf{Q}_\ell)$*

$$\sigma(\gamma) = \gamma \pm \varepsilon_s(\sigma) < \delta_s, \gamma > \delta_s$$

where ε_s is a character of I_s and $< \cdot, \cdot >$ is the pairing defined above.

Since δ_s is a generator of $V_s \subset E$, we immediately deduce that E is indeed stable by the action of $\pi_1(U, u)$. Moreover, it is part of the basic theory that this action respects the alternating form $< \cdot, \cdot >$, so the orthogonal $E^{\perp}$ of E in $H^{d-1}_{\text{ét}}(X_u, \mathbf{Q}_\ell)$ is also stable, and we get two subsheaves $\mathcal{E}$ and $\mathcal{E}^{\perp}$ of $R^{d-1}f_*\mathbf{Q}_\ell$.

Once again, all those constructions can be done over $\mathbf{F}_q$ (after possibly a finite extension), and there are also sheaves $\mathcal{E}_0, \ldots$

But the Picard-Lefschetz formula also implies immediately that the orthogonal of E has another description as the subspace of $H^{d-1}_{\text{ét}}(X_u, \mathbf{Q}_\ell)$ invariant under $\pi_1(U, u)$: for $\sigma \in I_s$,

$$\sigma(\gamma) = \gamma$$

is equivalent to

$$< \delta_s, \gamma >= 0.$$

The outcome of this is that $\mathcal{E}^{\perp}$ is a constant sheaf.[5] But as $E \oplus E^{\perp} = H^{d-1}_{\text{ét}}(X_u, \mathbf{Q}_\ell)$, we also have

$$\mathcal{E} \oplus \mathcal{E}^{\perp} = R^{d-1} f_* \mathbf{Q}_\ell$$

(on U). Now if we go from U to D, using the fact that

$$R^{d-1} f_* \mathbf{Q}_\ell = j_* j^* R^{d-1} f_* \mathbf{Q}_\ell$$

and because the first cohomology group of a constant sheaf on $\mathbf{P}^1$ vanishes (![6]), ***rh*** will be satisfied for X_0 if we can prove the corresponding inequality for the eigenvalues of $F^\star$ acting on $H^1(D, j_*\mathcal{E})$ (here we use the fact that $\mathcal{E}$ comes from a sheaf $\mathcal{E}_0$ on U_0 to get this action of the Frobenius).

This is now (almost) the irreducible case. A last use of (a generalization of) Poincaré duality and another lemma show that it is enough to prove the one-sided inequality

$$|\alpha| \leqslant q^{\frac{d}{2}+\frac{1}{2}}$$

for the eigenvalues of $F^\star$ acting this time on another cohomology group $H^1_c(D, \mathcal{E})$ (cohomology with compact support on U; this is necessary because U is not projective).

6 The crucial step

Now we will do some arithmetic, at last. To make this section more understandable, it is necessary to provide a better explanation of the way we can associate L-functions to sheaves, which will be to the zeta function (associated with the constant sheaf $\mathbf{Q}_\ell$) as Artin L-functions are to the zeta function in algebraic number theory. A clarification of the Euler product form of the zeta function is also in order.

Let U_0 be an affine curve over $\mathbf{F}_q$ which is the complement in $\mathbf{P}^1_0$ of a finite set of points S, and $A = \mathbf{F}_q[U_0]$ the algebra of regular functions on U_0, over $\mathbf{F}_q$: this means that A is just the subring of $\mathbf{F}_q(T)$ formed by rational functions which are defined outside S. If $U_0 = \mathbf{A}^1_0$, for instance, we have $A = \mathbf{F}_q[T]$. The set of "closed points" of U_0, written $|U_0|$, is by definition the set of maximal ideals in A. Since A is principal they correspond to irreducible elements in A, for instance to irreducible polynomials in $\mathbf{F}_q[T]$ in the case of the affine line.

Then the Euler product for $Z(U_0)$ is

$$Z(U_0) = \prod_{x \in |U_0|} \frac{1}{1 - T^{\deg(x)}}.$$

Let us repeat the proof in the case of the affine line: taking the logarithmic derivative of the left-hand side we get

$$\begin{aligned} -T^{-1} \sum_{x \in |\mathrm{A}^1_0|} \frac{-\deg(x) T^{\deg(x)}}{1 - T^{\deg(x)}} &= T^{-1} \sum_{x} \sum_{n \geqslant 1} \deg(x) T^{n \deg(x)} \\ &= T^{-1} \sum_{n \geqslant 1} |X_0(\mathbf{F}_{q^n})| T^n \end{aligned}$$

[5] In the correspondence between sheaves and abelian groups with action of $\pi(U, u)$, the constant sheaf $\underline{M}$ corresponds, as seems natural, to the group M with the trivial action.

[6] Maybe the only case where an H^1 on a curve is simpler than even H^0 and H^2...

since every x of degree $\deg(x) \mid n$ defines as many conjugates, all in $\mathbf{F}_{q^n}$, by factoring into linear factors. This is the same expression as obtained from

$$Z(\mathrm{A}_0^1) = \exp\Big(\sum_{n\geqslant 1} |\mathrm{A}_0^1(\mathbf{F}_{q^n})| \frac{T^n}{n}\Big)$$

by taking the same logarithmic derivative.

We now have a sheaf $\mathcal{E}_0$ on a curve U_0 and we want to define its L-function by an Euler product. The easiest way is to consider the equivalent formulation as a finite dimensional $\mathbf{Q}_\ell$-vector space E_0 on which $\pi_1(U_0, u)$ acts continuously. Indeed, $\pi_1(U_0, u)$, as we mentioned, is related to the Galois group of the field of rational functions on U_0, which is simply $\mathbf{F}_q(T)$. Since $x \in |U_0|$ is a maximal ideal, we can consider the decomposition group at x, D_x, and the inertia subgroup $I_x \subset D_x$. It happens that π_1 has the property that the image of I_x inside is trivial. In other words, when we restrict to D_x the action on E_0, the group D_x/I_x acts, and as in ordinary algebraic number theory, this is just the Galois group of the finite residue field extension $\mathbf{F}_{q^{\deg(x)}}/\mathbf{F}_q$. It is generated by the Frobenius morphisms F_x and the local factor is

$$\det(1 - F_x T^{\deg(x)} \mid E_0)$$

(as expected from the case of Artin L-functions), giving the following L function for E_0 (or $\mathcal{E}_0$):

$$L(\mathcal{E}_0) = \prod_{x\in|U_0|} \det(1 - F_x T^{\deg(x)} \mid E_0)^{-1}. \tag{21}$$

For sheaves, Grothendieck developed a generalization of his trace formula

$$\sum_{x\in U_0(\mathbf{F}_{q^n})} \mathrm{Tr}(F_x^n \mid E_0) = \sum_{0\leqslant i\leqslant 2d} (-1)^i \mathrm{Tr}(F^\star \mid H^i_{\text{ét}}(U, \mathcal{E}))$$

which implies as before the rationality of the L-function and gives a cohomological formula

$$L(\mathcal{E}_0) = \prod_{0\leqslant i\leqslant 2d} \det(1 - F^\star T \mid H^i_{\text{ét}}(U, \mathcal{E}))^{(-1)^{i+1}}. \tag{22}$$

It now makes sense to ask about the eigenvalues of Frobenius at x, and about the poles of the local L-factor, which are of course the inverse of those eigenvalues. We say that a sheaf $\mathcal{E}_0$ is of weight β if for all $x \in |U_0|$, all those eigenvalues at x are algebraic numbers and all their conjugates are of absolute value $q^{\deg(x)\beta/2}$.

From the elementary property that an absolutely convergent Euler product doesn't vanish, we then get

Lemma 2 *Let $\mathcal{E}_0$ be a sheaf on E_0 of weight β such that $H^0_{\acute{e}t}(U, \mathcal{E})$ and $H^2_{\acute{e}t}(U, \mathcal{E})$ are zero. Then all eigenvalues of $F^\star$ acting on $H^1_{\acute{e}t}(U, \mathcal{E})$ are algebraic numbers and all their complex conjugates α satisfy*

$$|\alpha| \leqslant q^{\frac{\beta}{2}+1}.$$

Proof. The formula (22) is simply here

$$L(\mathcal{E}_0) = \det(1 - F^\star T \,|\, H^1_{\text{ét}}(U, \mathcal{E}))$$

which from the hypothesis on $\mathcal{E}_0$ is a polynomial in $\mathbf{Q}[T]$, so the eigenvalues are algebraic.

We now look at the Euler product (21) (putting maybe $T = q^{-s}$ this time to get a more familiar picture), the assumption on the weight implies that the product converges absolutely for $\text{Re}(s - \beta/2) > 1$, namely $\text{Re}(s) > \beta/2 + 1$, which means for

$$|T| < q^{-\frac{\beta}{2}-1}$$

and so $L(\mathcal{E}_0)$ has no zeros in this region, which gives the result.

◇

We abstract the situation where we were left at the end of the previous section. We have a smooth affine curve $U_0 = \mathbf{P}^1_0 \setminus S$ over $\mathbf{F}_q$, a sheaf $\mathcal{E}_0$ (equivalently, a $\mathbf{Q}_\ell$-vector space E_0 of finite dimension on which $\pi_1(U_0, u)$ acts), and moreover there is an alternating non-degenerate bilinear form

$$\mathcal{E}_0 \times \mathcal{E}_0 \to \mathbf{Q}_\ell$$

(here we cheat: $< \cdot, \cdot >$ might not be non-degenerate on E, namely we could have $E \cap E^\perp \neq 0$; that this is so is known as the hard Lefschetz theorem, and was actually only proved by Deligne from the Riemann Hypothesis; however, one can proceed – as he did –, in the same way, with some more minor headaches, by replacing E by $E/E \cap E^\perp$.)

On the E_0 side this means that the corresponding bilinear form

$$E_0 \times E_0 \to \mathbf{Q}_\ell$$

is invariant by the action of $\pi_1(U_0, u)$. Moreover the Frobenius at x acts by a transformation of determinant q^{d-1}, which means precisely that the determinant of the representation on E_0 (a character), is associated to a sheaf of weight $n(d-1)$, n being the dimension of E_0 as $\mathbf{Q}_\ell$-vector space (this is because $\mathcal{E}_0$ is a subsheaf of $R^{d-1}f_{0*}\mathbf{Q}_\ell$.)

So the image G^a of $\pi_1(U_0, u)$ in the group of linear transformations of E is contained in the symplectic group of this bilinear form; its closure in the Zariski topology of the symplectic group is called the arithmetic monodromy group and has considerable importance. It contains as a subgroup the geometric monodromy group G^g which is the closure of the image of $\pi_1(U, u)$ in the symplectic group.

To these facts, Deligne adds the following ingredients:

Proposition 3 *(i) The monodromy group G^g is equal to the symplectic group (i.e., it is as big as it possibly can.)*

(ii) For any closed point x of U_0, the local factor at x of the L-function of $\mathcal{E}_0$, $\det(1 - F_x T^{\deg(x)} \,|\, E_0)$, is a polynomial with rational coefficients.

The first part (due to Kazhdan-Margulis) is deduced from the Lefschetz theory, and in particular from the Picard-Lefschetz formula (used to show that the representation of $\pi_1(U, u)$ on E is absolutely irreducible), and a result concerning the representations of the symplectic group.

The second part comes from rather intricate (but not very difficult) manipulations with the zeta function of the fibers of the Lefschetz pencil (which as a zeta function has coefficients in $\mathbf{Q}$) and its expression given by Grothendieck's theory.

We now start from these data, that is a curve U_0, a sheaf $\mathcal{E}_0$, with a non-degenerate alternating form, whose determinant has weight $n(\beta)$, which all together satisfy the conditions (i) and (ii) of the proposition and deduce…

Theorem 4 *(Deligne).*

In this situation, all eigenvalues of $F^\star$ acting on $H^1_c(U,\mathcal{E})$ are algebraic numbers and all their complex conjugates α satisfy

$$|\alpha| \leqslant q^{\frac{\beta+1}{2}+\frac{1}{2}}.$$

Since $\beta = d-1$ in the case under consideration, this will, at long last, conclude the proof of the Riemann Hypothesis for smooth projective varieties over finite fields.

Proof.

If we can deduce from the assumptions that $\mathcal{E}_0$ itself has weight β, we will be able to apply the lemma (the vanishing of $H^0_{\text{ét}}$ and $H^2_{\text{ét}}$ here is easy from the general theory; remember – well, I didn't mention it – that those $H^i_{\text{ét}}$ for the non-projective U_0 are cohomology groups with compact support, which accounts for the different behavior; H^0_c being 0, for instance, means that there are no sections supported on a finite set of points, which is pretty obvious).

So we need to prove that the local factors

$$\det(1 - F_x T^{\deg(x)} \mid E_0) = \prod_j (1 - \alpha_j(x) T^{\deg(x)})$$

have algebraic eigenvalues $\alpha_j(x)$ of modulus $q^{\beta \deg(x)/2}$.

The analogy here is with modular forms: if those were the local factors of the L-function of a modular form f, and the corresponding problem that of proving the Ramanujan conjecture, we would get a non-trivial estimate by using the Rankin-Selberg method. And we can do exactly the same here: the Rankin-Selberg convolution of $\mathcal{E}$ with itself is simply the L-function whose local factors are

$$\prod_{i,j} (1 - \alpha_i(x)\alpha_j(x) T^{\deg(x)})$$

and they are just the local factors associated to the tensor product $\mathcal{E}_0 \otimes \mathcal{E}_0$ (or $E_0 \otimes E_0$ with the natural action of the fundamental group).

By hypothesis the local factors of $\mathcal{E}_0$ have rational coefficients and this implies quickly that those of the tensor product are power series in T with *positive* (in the French sense) rational coefficients.

On the other hand the expression as a rational function is

$$Z(\mathcal{E}_0 \otimes \mathcal{E}_0) = \frac{\det(1 - F^\star T \mid H^1_{\text{ét}}(U, \mathcal{E} \otimes \mathcal{E}))}{(1 - q^{\beta+1}T)^n}$$

(where n is the dimension of E_0 as $\mathbf{Q}_\ell$-vector space). The precise form of the denominator is obtained by applying the general formulas giving $H^0_{\text{ét}}$ and $H^2_{\text{ét}}$ in this situation:

$$H^0_{\text{ét}}(U,\mathcal{E}) = E^{\pi_1(U,u)}$$

and (by Poincaré duality from this)

$$H^2_{\text{ét}}(U, \mathcal{E}) = (E^*)_{\pi_1(U,u)}$$

(the coinvariants of the action, namely the largest quotient of E^* on which π_1 acts trivially). One gets $H^0_{\text{ét}} = 0$ because the action is irreducible, and $H^2_{\text{ét}}$ is determined using the theory of the invariants of the symplectic group, and the fact that the determinant has weight $n\beta$.

In particular, $Z(\mathcal{E}_0 \otimes \mathcal{E}_0)$ has poles only at $T = q^{-\beta-1}$. On the other hand, $\alpha_j(x)^{2/\deg(x)}$ is a pole of the local factor at x. Since we know that we are dealing with a power series with positive coefficients, the usual lemma implies

$$q^{-\beta-1} \leqslant |\alpha_j(x)|^{2/\deg(x)}$$

or equivalently

$$|\alpha_j(x)| \leqslant q^{\deg(x)(\frac{\beta}{2}+\frac{1}{2})}$$

which is a first result.

Must we stop here? Of course not, as with modular forms we want to study higher convolutions of $\mathcal{E}_0$. But the point here is that whereas we can't go very far with modular forms because the poles of those higher powers are not known, here Grothendieck's theory, by an argument similar to that for the poles of the first convolution, permits the complete determination of the poles of

$$L(\mathcal{E}_0^{\otimes 2k})$$

for any $k \geqslant 1$. The crucial point is that we know the monodromy group G^g to be the full symplectic group, which shows that the identification of those poles is simply a question in the theory of the symplectic group and its representations, quite independently from any arithmetical problem.

By the same reasoning it shows that

$$L(\mathcal{E}_0^{\otimes 2k}) = \frac{\det(1 - F^\star T \mid H^1_{\text{ét}}(U, \mathcal{E}^{\otimes 2k}))}{(1 - q^{k\beta+1}T)^n}$$

whence as before

$$q^{-k\beta-1} \leqslant |\alpha_j(x)|^{2k/\deg(x)}$$

or

$$|\alpha_j(x)| \leqslant q^{\deg(x)(\frac{\beta}{2}+\frac{1}{2k})}.$$

Letting k tend to infinity gives $|\alpha_j(x)| \leqslant q^{\deg(x)\beta/2}$, and then Poincaré duality again reverses the inequality, thereby concluding the proof.

◇

It is strangely comforting to think that the decisive steps in the proof of such a deep and important theorem can be understood quite easily.

7 Epilogue and further references

The proof of the Weil conjectures is not just a monument to itself. It is also one of the most powerful results in number theory and has many applications. I will mention some that I know.

Deligne himself had already proved that they implied the Ramanujan-Petersson conjecture about the Fourier coefficients of holomorphic modular forms of weight at least 2 on congruence subgroups of $SL(2, \mathbf{Z})$.

He also generalized in [De1] the estimate of Weil for Kloosterman sums, to treat some families of more general exponential sums in many variables: let $Q \in \mathbf{F}_q[X_1, \ldots, X_n]$ a polynomial of degree d prime to p such that the homogeneous component of degree d of Q defines a smooth hypersurface in $\mathbf{P}^{n-1}_{\mathbf{F}_q}$. Then for any non-trivial character ψ of $\mathbf{F}_q$

$$\Big| \sum_{x_1,\ldots,x_n \in \mathbf{F}_q} \psi(Q(x_1, \ldots, x_n)) \Big| \leqslant (d-1)^n q^{n/2}.$$

Deducing this from the Riemann hypothesis was actually rather difficult and involved some non-trivial geometric considerations. Such techniques were then superseded by Deligne's further results [De2] which are the basis of the works of Katz on exponential sums which have produced ever deeper results, with many applications to concrete problems of analytic number theory.

I will now list some references the Weil conjectures and their proof.

- [De1] is Deligne's original paper, which I've tried to explain. It is very readable, with very good background summaries of the various theories involved.
- [De2], Deligne's second article about the Weil conjecture, contains another proof, subsumed in a much more general statement about the weight of higher direct images of pure or mixed constructible ℓ-adic sheaves, which is apparently considerably more flexible and useful for applications, in particular for exponential sums (this is the basis of the works of Katz).
- [Har] is the classical introduction to schemes in algebraic geometry, and contains the background of all non-elementary work in this area, although it doesn't treat étale cohomology or the Weil conjectures, except for stating the theorems in an appendix.
- [Mil] is one of the rare books about étale cohomology outside the SGA bunker (hence the title), with probably enough details to qualify as complete proofs of the basic statements. It has the Lefschetz theory for surfaces, for instance. Recommended if only because of the epigraph.
- [De3], a.k.a SGA 4 1/2, has summaries with sketches of proof of most of the theory; it's very useful as a kind of road-map if, by a fine spring morning, you feel that you would like to know how the proper base change theorem is proved. Again, Deligne takes great care to relate the methods and results to their classical analogues.
- [Wei], of course, is the original paper proposing the conjectures; the commentaries in the collected works of Weil (if you can find them in the library) are very interesting, and you may also see there how the possible *interpretation* in terms of some mysterious cohomology theory was already highlighted by Weil.

I'm afraid there's too much French in all that; Katz [Kat] has a survey of the proof in English, and there is a book called 'Étale Cohomology and the Weil Conjectures' by Freitag and Kiehl, but I haven't really gone into it.

References

[De1] Deligne, P.: La conjecture de Weil, I, Pub. Math. I.H.E.S 43, 273-307 (1974).

[De2] Deligne, P.: La conjecture de Weil, II, Pub. Math. I.H.E.S 52, 137-252 (1980).

[De3] Deligne, P.: Cohomologie étale, Lecture Notes in Mathematics 569, Springer Verlag 1977.

[Har] Hartshorne, R.: Algebraic Geometry, Graduate Texts in Mathematics 52, Springer 1990.

[Kat] Katz, N.: An overview of Deligne's proof of the Riemann hypothesis for varieties over finite fields, Proc. of Symposia in Pure Math. 28, 275-305, Am. Math. Soc. 1976.

[Mil] Milne, J.: Étale cohomology, Princeton University Press, 1980.

[Ser] Serre, J.-P.: Valeurs propres des endomorphismes de Frobenius (d'après P. Deligne), Séminaire Bourbaki 446 (1973/74), or Oeuvres, Vol. III (Springer 1986).

[Sil] Silverman, J.: The arithmetic of elliptic curves, Grad. Texts in Math. 106, Springer Verlag (1986).

[Wei] Weil, A.: Number of solutions of equations in a finite field, Bull. Am. Math. Soc. 55, 497-508 (1952).