The unity of mathematics as a method of discovery: Wiles' example

(A 1993 talk reworked for the 7th French Philosophy of Mathematics Workshop 5-7 November 2015, University of Paris-Diderot)

Jean Petitot
CAMS (EHESS), Paris

Introduction

At the end of August 1993 at the *XIXth International Congress of History of Science* organized in Zaragoza by my colleague Jean Dhombres, I gave a talk "Théorème de Fermat et courbes elliptiques modulaires"[45] in a workshop organized by Marco Panza. It was about the recent (quasi)-proof of the Taniyama-Shimura-Weil conjecture (*TSW*) presented by Andrew Wiles in three lectures "Modular forms, elliptic curves, and Galois representations" at the Conference "*p*-adic Galois representations, Iwasawa theory, and the Tamagawa numbers of motives" organized by John Coates. at the Isaac Newton Institute of Cambridge on June 21-23, 1993.

But the proof, which, as you know, implies Fermat Last Theorem (FLT), was not complete as it stood and contained a gap pointed out by Nicholas Katz (who, by the way, was one of the unique colleagues of Wiles at Princeton brougth into confidence).

It has been completed in a joined work with Richard Taylor (September 19, 1994: "I've got it!"), sent to some colleagues (including Faltings) on October 6, 1994, submitted on October 25, 1994, and published in 1995 [73] and [74].

Until 1997, I attended the Bourbaki seminars of Serre [55] and Oesterlé [44] at the Institut Henri Poincaré, worked a lot to understand the proof, and gave some lectures on it.

Today, I will use this old technical stuff on TSW but with a new focus.

In a presentation of the proof, Ram Murty ([41], p.1) speaks of "Himalayan peaks" that hold the "secrets" of such results. I will carry this excellent metaphor further.

The mathematical universe is like an Hymalayan mountain chain surrounded by the plain of elementary mathematics. A proof is like a path and a conjecture is like a peak or the top of a ridge to be reached. But not all paths are "conceptualizable" i.e. conceptually describable. Valleys are "natural" *mono*-theoretical conceptualizable paths.

But, if the conjecture is "hard", its peak cannot be reached along a valley starting from scrach in the plain. One has to reach internal "hanging valleys" suspended over lower valleys. This corresponds to the abstraction of relevant abstract structures. One has also to change valley using saddles, tunnels, passes, canyons, and also conceptual crossroads. One can also follow ridges between two valleys (two theories).

What is essential is that all these routes are internal to the whole Himalayan chain, and it is here that Lautman's concept of *unity* of mathematics enters the stage (Lautman is my hero in philosophy of mathematics).

A conceptually complex proof is a very uneven, rough, rugged *multi*-theoretical conceptualizable route.

It is this *holistic* nature of a complex proof which will be my main purpose. It corresponds to the fact that, even if FLT is very simple in its formulation, the deductive parts of its proof are widely *scattered in the global unity* of the mathematical universe. As was emphasized by Israel Kleiner ([31], p.33):

> *"Behold the simplicity of the question and the complexity of the answer! The problem belongs to number theory – a question about positive integers. But what area does the proof come from? It is unlikely one could give a satisfactory answer, for the proof brings together many important areas – a characteristic of recent mathematics."*

Wiles proof makes an extremely long detour to connect FLT with a great conjecture on elliptic curves, the Taniyama-Shimura-Weil conjecture (TSW). As was emphasized by Barry Mazur ([37], p. 594):

> *"The conjecture of Shimura-Taniyama-Weil is a profoundly unifying conjecture — its very statement hints that we may have to look to diverse mathematical fields for insights or tools that might leads to its resolution."*.

In the same paper, Mazur adds:

> *"One of the mysteries of the Shimura-Taniyama-Weil conjecture, and its constellation of equivalent paraphrases, is that although it is indeniably a conjecture "about arithmetic", it can be phrased variously, so that: in one of its guises, one thinks of it as being also deeply "about" integral transforms in the theory of one complex variable; in another as being also "about" geometry."*

All these quotations point out that the proof unfolds in the labyrinth of many different theories.

In many cases, it is possible to formulate "translations" as functors from one category to another (as in algebraic topology). One can say that a "direct and simple" proof is a sequence of deductive steps inside a single category, while an "indirect and complex" proof is a proof using many functorial changes of category.

But we need a lot of other conceptual operations to reach a correct comprehension of what is traveling inside the unity of mathematics. Albert Lautman was the first to investigate this problem.

Kummer's cyclotomic route

I will not go into the classical history of FLT, which is a true Odyssea. As you know, the first great general result ("general" means here for an infinite number of primes) is due to Kummer and results from the deep arithmetic of cyclotomic fields.

The case $n = 4$ was proved by Fermat himself using a "descent argument" based on the fact that if (a, b, c) is a Pythagorean triple (that is a triple of positive integers such that $a^2 + b^2 = c^2$), then the area $ab/2$ of the right triangle of sides a, b, c cannot be a square.

Then, during what could be called an "Eulerian" period, many particular cases where successively proved by Sophie Germain, Dirichlet, Legendre, Lamé, etc. using a fundamental property of *unique factorization of integers in prime factors* in algebraic extensions of $\mathbb{Q}$. But this property is *not* always satisfied.

FLT for regular primes

In 1844 Ernst Kummer was able to abstract the property for a prime l to be *regular*, proved FLT for all regular primes and explained that the *irregularity* of primes was the main obstruction to a natural algebraic proof. As you know, it is for this proof that Kummer invented the concept of "ideal" number and proved his outstanding result that unique factorization in prime factors remains valid for "ideal" numbers.

After this breakthrough, a lot of particular cases of irregular primes were proved which enabled to prove FLT up to astronomical l; and a lot of computational verifications were made. But no *general* proof was found.

As reminded by Henri Darmon, Fred Diamond and Richard Taylor in their 1996 survey of Wiles ([12], p.4):

> *"The work of Ernst Eduard Kummer marked the beginning of a new era in the study of Fermat's Last Theorem. For the first time, sophisticated concepts of algebraic number theory and the theory of L-functions were brought to bear on a question that had until then been addressed only with elementary methods. While he fell short of providing a complete solution, Kummer made substantial progress. He showed how Fermat's Last Theorem is intimately tied to deep questions on class numbers of cyclotomic fields."*

For l a prime number > 2, Kummer's basic idea was to factorize Fermat equation in the ring $\mathbb{Z}[\zeta]$ where ζ is a primitive l-th root of unity and to work in the *cyclotomic extension* $\mathbb{Z}[\zeta] \subset \mathbb{Q}(\zeta)$. This route was opened by Gauss for $l = 3$ ($\zeta = j$).

In $\mathbb{Z}[\zeta]$ we have the factorization into linear factors

$$x^l - 1 = \prod_{j=0}^{j=l-1} \left(x - \zeta^j\right).$$

The polynomial

$$\Phi(x) = x^{l-1} + \ldots + x + 1 = \prod_{j=1}^{j=l-1} \left(x - \zeta^j\right) \text{ (beware: } j = 1)$$

is irreducible over $\mathbb{Q}$ and is the minimal polynomial defining ζ ($\Phi(\zeta) = 0$). We note that $\Phi(1) = l$.

- The conjugates of ζ are $\zeta^2, \ldots, \zeta^{l-1}$,
- $\mathbb{Q}(\zeta)$ is the splitting field of $\Phi(x)$ over $\mathbb{Q}$ and $\mathbb{Q}(\zeta)/\mathbb{Q}$ is a Galois extension of degree $[\mathbb{Q}(\zeta) : \mathbb{Q}] = l - 1$.
- $\mathbb{Z}[\zeta]$ has for $\mathbb{Z}$-base $1, \zeta, \ldots, \zeta^{l-2}$.
- The prime l is totally ramified in $\mathbb{Z}[\zeta]$.

More precisely, $(1 - \zeta)$ is a prime ideal of $\mathbb{Z}[\zeta]$, the quotient $\mathbb{Z}[\zeta]/(1 - \zeta)$ is the finite field $\mathbb{F}_l$ and there exists some unit u s.t.

$$\begin{aligned} l &= u(1-\zeta)^{l-1} \text{ (product of elements)} \\ (l) &= (1-\zeta)^{l-1} \text{ (product of ideals)} \end{aligned}$$

since the $u_j = (1 - \zeta^j)/(1 - \zeta) = 1 + \zeta + \ldots + \zeta^{j-1}$ are units.

- $\mathbb{Z}[\zeta]$ is a unique factorization domain for $l \leq 19$ but not for $l = 23$ (it was a great discovery of Kummer).

Let us remind here some general conceptual properties of finite algebraic extensions which are at the origin of *abstract* algebra. They provide a *structural* picture enabling to manage intractable concrete computations.

Let $K/\mathbb{Q}$ be a finite algebraic extension of degree d. There are prime ideals $\mathfrak{p}$ of $\mathcal{O}_K$ (the ring of integers of K) over p (i.e. $\mathfrak{p} \cap \mathbb{Z} = (p)$, notation $\mathfrak{p} \mid (p)$).

Polynomials irreducible over $\mathbb{Q}$ can become reducible and factorize over K. So, (p) splits in $\mathcal{O}_K$ as a product of primes

$$p\mathcal{O}_K = \prod_{j=1}^{j=r} \mathfrak{p}_j^{e_j} \text{ with } \mathfrak{p}_j \mid (p)$$

and we have

$$\mathcal{O}_K/p\mathcal{O}_K = \bigoplus_{j=1}^{j=r} \mathcal{O}_K/\mathfrak{p}_j^{e_j} \ .$$

As you know, three types of numbers are essential to understand the behavior of the primes p in K.

1. The number r of factors: it as to do with the *decomposition* of (p).
2. The exponents e_j are called the *degrees of ramification* of the $\mathfrak{p}_j$ in $K/\mathbb{Q}$. The extension $K/\mathbb{Q}$ is said *unramified* at $\mathfrak{p}_j$ if $e_j = 1$, and $K/\mathbb{Q}$ is said *unramified* at p if it is unramified at any $\mathfrak{p}_j \mid (p)$, i.e. if all $e_j = 1$.
3. The residue field $\mathcal{O}_K/\mathfrak{p}_j$ is an algebraic extension of $\mathbb{F}_p$ of degree f_j called the *residue or inertia degree*. Therefore $\mathcal{O}_K/\mathfrak{p}_j = \mathbb{F}_{p^{f_j}}$

These numbers are linked by a fundamental relation:

$$\sum_{j=1}^{j=r} e_j f_j = d \ .$$

If $K/\mathbb{Q}$ is *Galois* (i.e. $\mathbb{Q}$ is the subfield fixed by the automorphism group of $K/\mathbb{Q}$), then the Galois group $\mathrm{Gal}(K/\mathbb{Q})$ acts transitively upon the $\mathfrak{p}_j$ and conjugate the r factors. All the ramification degrees are the same, $e_j = e$, and the same applies for the inertia degrees $f_j = f$. The fundamental relation becomes:

$$ref = d \ .$$

Three cases are particularly interesting:

1. $e = d$, and $f = r = 1$. (p) is not decomposed, $p\mathcal{O}_K = \mathfrak{p}^d$ and $\mathcal{O}_K/\mathfrak{p} =\mathbb{F}_p$. The prime p is said *totally ramified* in K.
2. $r = d$, and $e = f = 1$. In that case, (p) is said *totally decomposed*, $p\mathcal{O}_K = \prod\limits_{j=1}^{j=d} \mathfrak{p}_j$, $\mathcal{O}_K/\mathfrak{p}_j{=}\mathbb{F}_p$, and $\mathcal{O}_K/p\mathcal{O}_K = \bigoplus_{j=1}^{j=n} \mathbb{F}_p$.
3. $f = d$, and $e = r = 1$. In that case, p is said *inert* in K, which means that p remains prime in $\mathcal{O}_K$. But now, the residue field is $\mathcal{O}_K/p\mathcal{O}_K = \mathbb{F}_p^d$.

The three properties: decomposition, ramification, and inertia can be easily red on *subgroups* of the Galois group $\mathrm{Gal}\,(K/\mathbb{Q})$, and therefore, via the Galois correspondance, on *intermediary extensions* $K/L/\mathbb{Q}$.

If $g \in \mathrm{Gal}\,(K/\mathbb{Q})$, then g acts on the residue fields as

$$\overline{g} : \mathcal{O}_K/\mathfrak{p}_j \to \mathcal{O}_K/g\left(\mathfrak{p}_j\right).$$

So, if g *fixes* $\mathfrak{p}_j$, then $\overline{g} \in \mathrm{Gal}\left(\left(\mathcal{O}_K/\mathfrak{p}_j\right)/\mathbb{F}_p\right)$. These g stabilizing $\mathfrak{p}_j$ consitute the *decomposition group* $D = D_{\mathfrak{p}_j}$ of $\mathfrak{p}_j$ and the kernel $I = I_{\mathfrak{p}_j}$ of $g \mapsto \overline{g}$ (i.e. g acts trivially on the residue field) is called the *inertia group* of $\mathfrak{p}_j$.

D fixes a sub-extension $K^D/\mathbb{Q}$ of $K/\mathbb{Q}$ and defines an extension K/K^D which is the smallest extension $K/K^D/\mathbb{Q}$ where the prime $\mathfrak{q}_j = \mathfrak{p}_j \cap \mathcal{O}_{K^D}$ does not split because the complete possible decomposition of $\mathfrak{p}_j$ in $\mathcal{O}_K$ is already done in $\mathcal{O}_{K^D}$.

K^D is called the *decomposition field* of p in K.

The prime p is totally decomposed in K^D and therefore $\left[K^D : \mathbb{Q}\right] = r$.

The inertia subgroup I corresponds to a sub-extension K^I of K which is also an extension of K^D.

So we have the tower of extensions $K/K^I/K^D/\mathbb{Q}$.

1. $K^D/\mathbb{Q}$ explains decomposition.
2. K^I/K^D explains inertia, that is $\mathfrak{q}_j$ (above p and under $\mathfrak{p}_j$) remains inert between K^D and K^I.
3. And finally, K/K^I explains ramification: all the $\mathfrak{q}_j$ become totally ramified as $\mathfrak{p}_j^{e_j}$.

$$\begin{tikzcd}
 & \mathfrak{p}_j \arrow[dd, "e^j"'] & K & \\
\text{ramification} & & & K/K^I \text{ is the source of ramification} \\
 & \mathfrak{q}_j \arrow[dd, "f^j"'] & K^I & \\
\text{inertia} & & & K^I/K^D \text{ is the source of inertia} \\
 & \mathfrak{q}_j \arrow[dd, "r"'] & K^D & \\
\text{decomposition} & & & K^D/K \text{ is the source of decomposition} \\
 & p & \mathbb{Q} &
\end{tikzcd}$$

For the cyclotomic field $\mathbb{Q}(\zeta)$, there exist three simple behaviors for natural primes p in $\mathbb{Z}[\zeta]$ (there exists a more complicated 4-th case).

1. If $p = l = u.(1-\zeta)^{l-1}$, then p is totally ramified.
2. If $p \equiv 1 \mod l$, then p is totally decomposed.
3. If $f = p - 1$, $e = 1$, then p is inert.

In $\mathbb{Z}[\zeta]$ we get the decomposition

$$z^l = x^l + y^l = \prod_{j=0}^{j=l-1} \left(x + \zeta^j y \right) .$$

If, in $\mathbb{Z}[\zeta]$, the unique factorization of an integer in prime factors (UF) were valid, *then* we would use the fact that all the factors $\left(x + \zeta^j y \right)$ are l powers and we would conclude. *But*, in $\mathbb{Z}[\zeta]$, UF is *not* necessarily true. However, Kummer proved it remains valid for ideals.

To prove FLT in this context, we suppose that a non trivial solution (a, b, c) exists and we look at its relations with the prime power l. In the computations the property of "*regularity*" enters the stage to derive *impossible* congruences.

Case I

Suppose first that x and y are *prime to l*. This implies that the ideals $\left(x+\zeta^j y\right)$ are *relatively prime*.

As the product of the $\left(x+\zeta^j y\right)$ is the l-th power $(z)^l$, each $\left(x+\zeta^j y\right)$ is therefore a l-th power and we have in particular

$$(x+\zeta y)=\mathfrak{a}^l$$

which shows that $\mathfrak{a}^l$ is a *principal* ideal.

It is here that the property of regularity enters the stage.

- *Intuitive definition*. l is a regular prime if when a l-th power $\mathfrak{a}^l$ of an ideal $\mathfrak{a}$ is principal $\mathfrak{a}$ is already itself a principal ideal.
- *Technical definition*. l is a regular prime if it doesn't divide the class number h_l of the cyclotomic field $\mathbb{Q}(\zeta)$, the class number h_l "measuring" the failure of UF in $\mathbb{Z}[\zeta]$.

As $\mathfrak{a}^l$ is principal, if l is a regular prime, $\mathfrak{a}$ is principal: $\mathfrak{a} = (t)$, $(x + \zeta y) = (t)^l$ and there exists therefore some unit u in $\mathbb{Z}[\zeta]$ s.t.

$$x + \zeta y = ut^l.$$

The idea is then to compare $x + \zeta y$ with its complex conjugate $x + \overline{\zeta} y$ using congruences $\mathrm{mod}\, l$ in $\mathbb{Z}[\zeta]$.

Using the fact that $\left\{1, \zeta, \ldots, \zeta^{l-2}\right\}$ is an integral basis of $\mathbb{Z}[\zeta]$ over $\mathbb{Z}$ and developing t as $t = \sum_{i=0}^{i=l-2} \tau_i \zeta^i$, one shows first that $t^l \equiv \overline{t}^l \bmod l\mathbb{Z}[\zeta]$. Secondly, using a lemma of Kronecker, one shows that u being a unit, there exists j s.t. $\frac{u}{\overline{u}} = \zeta^j$. One concludes that

$$x + \zeta y = ut^l = \zeta^j \overline{u} t^l \equiv \zeta^j \overline{u} \overline{t}^l \bmod l\mathbb{Z}[\zeta] \equiv \zeta^j \left(x + \overline{\zeta} y\right) \bmod l\mathbb{Z}[\zeta] \tag{C}$$

We get therefore $\bmod l\mathbb{Z}[\zeta]$ *a linear relation* between $1, \zeta, \zeta^j, \zeta^{j-1}$ (we use $\zeta^j \overline{\zeta} = \zeta^{j-1}$) with integral coefficients x, y coming from the supposed solution (x, y, z) of Fermat equation.

But the congruence (C) is impossible. Indeed if $1, \zeta, \zeta^j, \zeta^{j-1}$ are different powers then then they are independent in $\mathbb{Z}[\zeta]$ over $\mathbb{Z}$. When it is not the case $(j = 0, j = 1, j = 2, j = l - 1)$, one proves the particular cases.

Case II

The real difficulty is the case II when one of x, y, z is divided by l. I will skip it here.

Kummer's proof is marvelous and played a fundamental role in the elaboration of modern arithmetical tools. Its essential achievement is to do arithmetic no longer in $\mathbb{Z}$ but in the ring of integers $\mathbb{Z}[\zeta]$ of the cyclotomic field $\mathbb{Q}(\zeta)$. But it remains a proof developed inside a *single* theory, namely algebraic number theory.

In Summer 1847, Kummer not only proved FLT for l regular but, reinterpreting a formula of Dirichlet, gave a deep criterion for a prime l to be regular. As Edwards emphasizes [20], this

> *"must be regarded as an extraordinary tour de force."*

Characterization of regular primes. A prime l is regular iff it doesn't divide the numerators of any of the *Bernouilli numbers* $B_2, B_4, \ldots, B_{l-3}$.

For instance 37 is an irregular prime since 37 divides the numerator 7709321041217 of B_{32} and $32 < 37 - 3 = 34$.

Bernouilli numbers are defined by the series

$$\frac{x}{e^x - 1} = \sum_{n=0}^{n=\infty} B_n \frac{x^n}{n!}$$

They are also defined by the recurrence relations $B_0 = 1$, $1 + 2B_1 = 0$, $1 + 3B_1 + 3B_2 = 0$, $1 + 4B_1 + 6B_2 + 4B_3 = 0$, $1 + 5B_1 + 10B_2 + 10B_3 + 5B_4 = 0$

$$(n+1)\,B_n = -\sum_{k=0}^{k=n-1} \binom{n+1}{k} B_k$$

where the binomial coefficients $\binom{n}{k} = \frac{n!}{(n-k)!k!}$.

We have $B_1 = -\frac{1}{2}$, $B_2 = \frac{1}{6}$, $B_3 = 0$, $B_4 = -\frac{1}{30}$, $B_5 = 0$, $B_6 = \frac{1}{42}$, $B_7 = 0$, etc. All the B_n for $n > 1$ odd vanish.

A theorem due to Von Staudt and Claussen asserts that the denominator D_n of the B_n are the product of the primes such that $(p-1) \mid n$. In fact, $B_{2k} + \sum_{p \text{ s.t. } p-1 \mid 2k} \frac{1}{p}$ is a rational integer and $p \mid D_{2k}$ iff $(p-1) \mid 2k$ and then $pB_{2k} \equiv -1 \mod p$.

Bernouilli numbers are ubiquitous in arithmetics and closely related to the values of Riemann Zeta function (see below) at even integers $2k$ and negative odd integers $1-2k$ $(k>0)$:

$$\zeta(2k) = (-1)^{k-1}\frac{(2\pi)^{2k} B_{2k}}{2(2k)!}, \zeta(1-2k) = -\frac{B_{2k}}{2k}$$

For instance, in the case $k=1$, we find $\zeta(2) = \sum_{n\geq 1}\frac{1}{n^2} = \frac{4\pi^2}{2.2}B_2 = \frac{\pi^2}{6}$ and in the case $k=2$, we find $\zeta(4) = \sum_{n\geq 1}\frac{1}{n^4} = -\frac{16\pi^4}{2.24}B_4 = \frac{\pi^4}{90}$, values Euler already knew.

Kummer theorem, which in that sense is deeply linked with Riemann ζ function, follows from the fact that if K^+ is the maximal real subfield $\mathbb{Q}\left(\zeta+\overline{\zeta}\right)$ $(\overline{\zeta}=\zeta^{-1})$ of $\mathbb{Q}(\zeta)$ and h^+ the class number of K^+ then $h=h^+h^-$, h^+ being computable in terms of special units (it is a difficult computation) and h^-, called the relative class number, in terms of Bernouilli numbers: $l \mid h^-$ iff l divides the numerators of the Bernouilli numbers $B_2, B_4, \ldots, B_{l-3}$.

If l is regular $l \nmid h$ and therefore $l \nmid h^-$. Kummer proved also that $l \mid h^+$ implies $l \mid h^-$ and therefore $l^2 \mid h$ and also $l \mid h \Leftrightarrow l \mid h^-$. The Kummer-Vandiver conjecture claims that in every case $l \nmid h^+$ and that l is irregular iff $l \mid h^-$. It has been verified up to $l < 2^{27} = 134\ 217\ 728$ by David Harvey.

Further advances along the cyclotomic route

After Kummer's intensive and extensive computations and theoretical breakthrough, many people devoted a lot of works to the incredibily more complex irregular case, trying to deepen the knowledge of the structure of cyclotomic fields (see Washington's book [70] and Rosen's survey [49]).

Kummer himself weakened his regularity condition and succeeded in proving FLT for $l < 100$ because the irregular primes < 100, namely 37, 59, and 67 satisfy these weaker criteria. But such criteria are extremely computation consuming.

This point is particularly interesting at the epistemological level. Kummer's systematic computations for l regular opened the way to abstract structural algebra *à la* Dedekind-Hilbert.

A particularly important work on the cyclotomic route were that of Harry Schultz Vandiver (1882-1973) who proved in the late 1920s that if the Bernouilli numbers B_i for $i = 2, 4, \ldots, l - 3$ are not divisible by l^3 and if $l \nmid h_l^+$ then the second case of FLT is true for l.

Vandiver proposed also a key conjecture:

Vandiver conjecture: $l \nmid h_l^+$.

Vandiver began to use such criteria "to test FLT computationnally" (Rosen [49]) and, with the help of Emma and Dick Lehmer for computations, proved FLT up to $l \sim 4.000$ and, in Case I, for $l < 253.747.889$.

In beautiful papers, Leo Corry [9] and [10] analyzed the computational aspects of FLT after the introduction of computers.

- In 1949 John von Neumann constructed the first modern computer ENIAC. As soon as 1952 E. and D. Lehmer used softwares implementing the largest criteria for proving FLT, first with ENIAC, then at the NBS (National Bureau of Standards) with SWAC (Standards Western Automatic Computer, 1.600 additions and 2.600 multiplications per second).
- They discovered new irregular primes such as 389, 491, 613, and 619. To prove that 1693 is irregular took 25mn.
- In 1955, to prove FLT for $l < 4,000$ took hundred hours of SWAC.
- In 1978, Samuel Wagstaff succeeded up to $l < 125,000$.
- In 1993, just before Wiles' proof, FLT was proved up to $l \sim 4\ 000\ 000$ (Buhler) and, in Case I, for $l < 714\ 591\ 416\ 091\ 389$ (Grandville).

But in spite of deep results of Stickelberger, Herbrand, etc. there remain apparently *intractable obstructions* on the cyclotomic route for irregular primes. It seemed that such a *purely algebraic* strategy didn't succeed to break the problem.

As was emphasized by Charles Daney [11]

> *"Despite the great power and importance of Kummer's ideal theory, and the subtlety and sophistication of subsequent developments such as class field theory, attempts to prove Fermat's last theorem by purely algebraic methods have always fallen short."*

We will see that Wiles' proof uses a very strong "non abelian" generalization of the classical "abelian" class field theory.

Faltings theorem and the Mordell-Weil conjecture

The natural context of a proof of FLT seems to be algebraic geometry since Fermat equation

$$x^l + y^l = z^l$$

is the homegeneous equation of a projective plane curve F. The equation has rational coefficients and FLT says that, for $l \geq 3$, the curve F has no rational points.

So FLT is a particular case of computing the cardinal $|F(\mathbb{Q})|$ of the set of rational points of a projective plane curve F defined over $\mathbb{Q}$. To solve the problem, one needs a deep knowledge of the arithmetic properties of *infinetely many* types of projective plane curves since the genus g of F is

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

This genus increases quadratically with the degree l. We note that for $l \geq 4$ we have $g \geq 3$. But of course it is extremely difficult to prove *general* arithmetic theorems valid for infinitely many sorts of classes of curves.

A great achievement in this direction was the demonstration by Gerd Faltings of the celebrated *Mordell-Weil conjecture*.

Theorem (Faltings). Let C be a smooth connected projective curve defined over a number field K and let $K \subset K'$ be an algebraic extension of the base field K. Let g be the genus C.

1. If $g = 0$ (sphere) and $C(K') \neq \emptyset$, then C is isomorphic over K' to the projective line $\mathbb{P}^1$ and there exist *infinitely many* rational points over K'.
2. If $g = 1$ (elliptic curve), either $C(K') = \emptyset$ (no rational points over K') or $C(K')$ is a finitely generated $\mathbb{Z}-$module (Mordell-Weil theorem, a deep generalization of Fermat descent method).
3. If $g \geq 2$, $C(K')$ is *finite* (Mordell-Weil conjecture, Faltings theorem).

Faltings theorem is an extremely difficult one which won him the Fields medal in 1986. But for FLT we need to go from "$C(K')$ finite" to "$C(K') = \emptyset$". The gap is too large. We need to find *another route*.

Hellegouarch and Frey: opening the elliptic route

In 1969 Yves Hellegouarch introduced an "elliptic trick". His idea was to use an hypothetical solution $a^l + b^l + c^l = 0$ of Fermat equation (l prime ≥ 5, $a, b, c \neq 0$ pairwise relatively prime) *as parameters for an elliptic curve* (EC) defined over $\mathbb{Q}$, namely the curve E:

$$y^2 = x\left(x - a^l\right)\left(x + b^l\right) = x^3 + \left(b^l - a^l\right)x^2 - (ab)^l x$$

Hellegouarch analyzed the l-torsion points of E (see below) and found that the extension of $\mathbb{Q}$ by their coordinates had *very strange ramification properties* (it is unramified outside 2 and l) (see below).

Seventeen years later, in 1986, Gerhard Frey refined this key idea which led to Wiles-Taylor proof in 1995.

The EC E is *regular*. Indeed its equation is of the form

$$F(x, y) = y^2 - f(x) = y^2 - x\left(x - a^l\right)\left(x + b^l\right) = 0$$

and a singular point must satisfy $\frac{\partial F}{\partial x} = \frac{\partial F}{\partial y} = 0$. The condition $\frac{\partial F}{\partial y} = 0$ implies $y = 0$ and therefore $f(x) = 0$, while the condition $\frac{\partial F}{\partial x} = 0$ implies $f'(x) = 0$. So the x coordinate of a singular point must be a multiple root of the cubic equation $f(x) = 0$, but this is impossible for $f(x) = x\left(x - a^l\right)\left(x + b^l\right)$.

A Frey curve E is given in the Weierstrass form:

$$y^2 = x^3 + \frac{b_2}{4}x^2 + \frac{b_4}{2}x + \frac{b_6}{4}$$

Its *discriminant* is given by the general formula:

$$\Delta = -\left(b_2\right)^2 b_8 - 8\left(b_4\right)^3 - 27\left(b_6\right)^2 + 9b_2 b_4 b_6$$

with $4b_8 = b_2 b_6 - \left(b_4\right)^2$. We have $b_2 = 4\left(b^l - a^l\right)$, $b_4 = -2a^l b^l$, $b_6 = 0$, $b_8 = -a^{2l}b^{2l}$ and therefore

$$\Delta = 16\left(a^l b^l c^l\right)^2$$

E is regular, iff $\Delta \neq 0$ and it the case here.

But, if we *reduce* $E \mod p$ (which is possible since the coefficients of E are in $\mathbb{Z}$), the reduction E_p will be singular if $p \mid \Delta$. But since a and b are relatively prime, we cannot have at the same time $a^l \equiv 0 \mod p$ and $b^l \equiv 0 \mod p$, and so we cannot have a triple root.

The singularity of E_p can only be a normal crossing of two branches (a node). ECs sharing this property are called *semi-simple*.

Another extremely important invariant of an EC is its *conductor* N which, according to Henri Darmon ([13], p.1398), is

> *"an arithmetically defined quantity that measures the Diophantine complexity of the associated cubic equation."*

In the semi-simple case (where all singular reductions E_p are nodes) N is rather simple: it is the *square free* the product

$$N = \prod_{p|\Delta} p = \prod_{p|abc} p \ .$$

(2 which divides Δ divides also *abc* since one of the a, b, c is even).

As Δ is proportional to $(abc)^{2l}$ while $N \leq abc$, we see that $\Delta \geq CN^{2l}$ for a constant C.

This property is in fact quite "extraordinary" since it violates the very plausible following Szpiro conjecture saying that the discriminant is bounded by a *fixed* power of the conductor:

Szpiro Conjecture. If E is any elliptic curve defined over $\mathbb{Q}$, for every $\varepsilon > 0$ there exists a constant D s.t. $|\Delta| < DN^{6+\varepsilon}$.

Another fondamental invariant of E is the *modular invariant* j defined by

$$j = \frac{\left((b_2)^2 - 24b_4\right)^3}{\Delta}$$

Hellegouarch and Frey idea is that, as far as (a, b, c) is a solution of Fermat equation and is supposed to be too exceptional to exist, the associated curve E must also be in some sense "too exceptional" to exist: exceptional numbers must parametrize exceptional objects.

We meet here a spectacular example of a *translation strategy* which consists in coding solutions of a first equation into *parameters* of a second object of a completely different nature and using the properties of the second object for gathering informations on the solutions of the first equation.

In the Himalayan metaphor, this type of methodological move consists in finding a sort of "tunnel" or "canyon" between two valleys.

G. Frey was perfectly aware of the originality of his method. In his paper [25] he explains:

> *"In the following paper we want to relate conjectures about solutions of the equation $A - B = C$ in global fields with conjectures about elliptic curves."*
> *"An overview over various conjectures and implications discussed in this paper (...) should show how ideas of many mathematicians come together to find relations which could give a new approach towards Fermat's conjecture."*

Frey's "come together" is like Kleiner's "bring together" and emphasizes the holistic nature of the proof.

The advantages of Frey's strategic "elliptic turn" are multifarious:

1. Whatever the degree l could be, we work always on an elliptic curve and we shift therefore from the full universe of algebraic plane curves $x^l + y^l = z^l$ to a *single* class of curves. It is a fantastic reduction of the diversity of objects.
2. Elliptic curves are by far the best known of all curves and their fine Diophantine and arithmetic structures can be investigated using *non elementary* techniques from analytic number theory.
3. For elliptic curves a strong criterion of "normality" is available: "good" elliptic curves are *modular* in the sense they can be parametrized by modular curves.
4. A well known conjecture, the *Taniyama-Shimura-Weil conjecture*, says in fact that *every* elliptic curve is modular.

From Frey's idea one can derive a natural schema of proof for FLT:

(a) Prove that Frey ECs are not modular.
(b) Prove the Taniyama-Shimura-Weil conjecture.

Step 1 was achieved by Kenneth Ribet who proved that Taniyama-Shimura-Weil implies FLT and triggered a revolutionary challenge, and

step 2 by Andrew Wiles and Richard Taylor for the so called "semi-stable" case, which is sufficient for FLT since Frey ECs are semi-simple.

In such a perspective, FLT is no longer an isolated curiosity, as Gauss claimed, but a consequence of general deep arithmetic constraints.

The L-function of an elliptic curve

To define what is a modular elliptic curve E defined over $\mathbb{Q}$, we have to associate to E a L-function L_E which counts in some sense the number of integral points of E.

E has an infinity of points over $\mathbb{C}$ (but can have no points on $\mathbb{Q}$). However, if we reduce $E \mod p$ (p a prime number), its reduction E_p will necessarily have a finite number of points $N_p = \#E_p(\mathbb{F}_p)$ over the finite field $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$.

The most evident arithmetic data on E consists therefore in combining these local data N_p relative to the different primes p.

This is a general idea. Any EC (more generally any algebraic variety) defined over $\mathbb{Q}$ can be interpreted as an EC with points in $\mathbb{Q}$, in algebraic number fields K, in $\overline{\mathbb{Q}}$, $\mathbb{R}$, $\mathbb{C}$, $\mathbb{F}_p$, $\mathbb{F}_{p^n}$, $\overline{\mathbb{F}_p}$, etc.

The *L*-function L_E of E is defined as an *Euler product*, that is a product of one factor for each p. We must be cautious since for p dividing the discriminant Δ of E, the reduction is "bad", i.e. E_p is singular (it is a node: semi-simplicity).

For technical reasons (see below), it is better to use the difference $a_p = p + 1 - N_p$. In the good reduction case (where E_p is itself an EC) we can generalize the counting to the finite fields $\mathbb{F}_{p^n}$ and show that the a_{p^n} are determined by the a_p via the formula

$$\frac{1}{1 - \frac{a_p}{p^s} + \frac{1}{p^{2s-1}}} = 1 + \frac{a_p}{p^s} + \frac{a_{p^2}}{p^{2s}} + \cdots$$

In the bad reduction case, we must use $\left(1 - a_p p-^{s}\right)^{-1}$.

So, the good choice of an Euler product is the following, which defines the L-function $L_E(s)$ of the elliptic curve E:

$$L_E(s) = \prod_{p|\Delta} \frac{1}{1 - \frac{a_p}{p^s}} \prod_{p\nmid\Delta} \frac{1}{1 - \frac{a_p}{p^s} + \frac{1}{p^{2s-1}}}$$

As $1 \leq N_p \leq 2p + 1$ (we count the point at infinity), then $|a_p| \leq p$, and therefore $L_E(s)$ converges for $\Re(s) > 2$. In fact, a theorem due to Hasse asserts that $|a_p| \leq 2\sqrt{p}$ and therefore $L_E(s)$ converges for $\Re(s) > 3/2$.

We will see below with Hecke's theory how the L-functions are constructed.

As explained Anthony Knapp [32], the L-function L_E

> *"encode geometric information, and deep properties of the elliptic curve come out (partly conjecturally) as a consequence of properties of these functions."*

And as for Riemann's Zeta function:

> *"It is expected that deep arithmetic information is encoded in the behavior of $L_E(s)$ beyond the region of convergence".*

Riemann's ζ-function

To understand the relevance of the L-functions L_E, we have to come back to Riemann's ζ-function, which is the great inspirer.

The zeta function $\zeta(s)$ encodes deep arithmetic properties in analytic structures.

Its initial definition is extremely simple and led to a lot of computations since Euler time:

$$\zeta(s) = \sum_{n \geq 1} \frac{1}{n^s}$$

which is a series absolutely convergent for integral exponents $s > 1$.

Euler already proved $\zeta(2) = \pi^2/6$ and $\zeta(4) = \pi^4/90$.

A trivial expansion shows that, in the convergence domain, the sum is equal to an infinite Euler product containing a factor for each prime p (we denote $\mathcal{P}$ the set of primes):

$$\zeta(s) = \prod_{p \in \mathcal{P}} \left(1 + \frac{1}{p^s} + \dots \frac{1}{p^{ms}} + \dots \right) = \prod_{p \in \mathcal{P}} \frac{1}{1 - \frac{1}{p^s}}.$$

The fantastic strength of the zeta function as a tool comes from the fact that *it can be extended by analytic continuation to the complex plane*.

- First s can be extended to complex numbers s of real part $\Re(s) > 1$, and moreover
- $\zeta(s)$ can be extended by analytic continuation to a meromorphic function on the entire complex plane $\mathbb{C}$ with a pole at $s = 1$.

Theta function and Mellin transform

The zeta function encodes very deep arithmetic properties.

Riemann proved in his celebrated 1859 paper "Über die Anzahl der Primzahlen unter einer gegeben Grösse" ("On the number of prime numbers less than a given quantity") [48] that it manifests beautiful properties of symmetry.

This can be made explicit noting that $\zeta(s)$ is related by a transformation called the *Mellin transform* to the *theta function* which possesses beautiful properties of automorphy, where "automorphy" means invariance of a function $f(\tau)$ defined on the Poincaré plane $\mathcal{H}$ (complex numbers τ of positive imaginary part $\Im(\tau)$) relatively to a countable subgroup of the group acting on $\mathcal{H}$ by homographies (also called Möbius transformations) $\gamma(\tau) = \frac{a\tau+b}{c\tau+d}$. (See below)

The theta function $\Theta(\tau)$ is defined on the half plane $\mathcal{H}$ as the series

$$\Theta(\tau) = \sum_{n\in\mathbb{Z}} e^{in^2\pi\tau} = 1 + 2\sum_{n\geq 1} e^{in^2\pi\tau}$$

$\Im(\tau) > 0$ is necessary to warrant the convergence of $e^{-n^2\pi\Im(\tau)}$.

We will see later that $\Theta(\tau)$ is what is called a *modular form* of level 2 and weight $\frac{1}{2}$. Its automorphic symmetries are

1. Symmetry under translation: $\Theta(\tau+2) = \Theta(\tau)$ (level 2, trivial since $e^{2i\pi} = 1$ implies $e^{in^2\pi(\tau+2)} = e^{in^2\pi\tau}$).
2. Symmetry under inversion: $\Theta(\frac{-1}{\tau}) = \left(\frac{\tau}{i}\right)^{\frac{1}{2}}\Theta(\tau)$ (weight $\frac{1}{2}$, proof from Poisson formula).

If $f : \mathbb{R}^+ \to \mathbb{C}$ is a complex valued function defined on the positive reals, its *Mellin transform* $g(s)$ is defined by the formula:

$$g(s) = \int_{\mathbb{R}^+} f(t) t^s \frac{dt}{t}$$

Let us compute the Mellin transform of $\Theta(it)$ or more precisely, using the formula $\Theta(\tau) = 1 + 2\tilde{\Theta}(\tau)$, of $\tilde{\Theta}(it) = \frac{1}{2}\left(\Theta(it) - 1\right)$:

$$\Lambda(s) = \frac{1}{2} g\left(\frac{s}{2}\right) = \frac{1}{2}\int_0^\infty \left(\Theta(it) - 1\right) t^{\frac{s}{2}} \frac{dt}{t} = \sum_{n \geq 1} \int_0^\infty e^{-n^2 \pi t} t^{\frac{s}{2}} \frac{dt}{t}$$

In each integral we make the change of variable $x = n^2 \pi t$. The integral becomes:

$$n^{-s} \pi^{-\frac{s}{2}} \int_0^\infty e^{-x} x^{\frac{s}{2}-1} dx$$

But $\int_0^\infty e^{-x} x^{\frac{s}{2}-1} dx = \Gamma\left(\frac{s}{2}\right)$ where $\Gamma(s) = \int_0^\infty e^{-x} x^{s-1} dx$ is the *gamma function*.

Therefore

$$\Lambda(s) = \pi^{-\frac{s}{2}} \Gamma\left(\frac{s}{2}\right) \left(\sum_{n \geq 1} \frac{1}{n^s}\right) = \zeta(s) \Gamma\left(\frac{s}{2}\right) \pi^{-\frac{s}{2}}$$

This remarkable expression enables to use the automorphic symmetries of the theta function to derive a *functional equation* satisfied by the lambda function, and therefore by the zeta function.

Functional equation

Indeed, let us write $\Lambda(s) = \int_0^\infty = \int_0^1 + \int_1^\infty$ and use the change of variable $t = \frac{1}{u}$ in the first integral. Since $\frac{i}{u} = -\frac{1}{iu}$ and

$$\Theta\left(\frac{i}{u}\right) = \Theta\left(-\frac{1}{iu}\right) = \left(\frac{iu}{i}\right)^{\frac{1}{2}} \Theta\left(iu\right) = u^{\frac{1}{2}} \Theta\left(iu\right)$$

due to the symmetry of Θ under inversion, we verify that the $\int_0^1$ part of $\Lambda(s)$ is equal to the $\int_1^\infty$ part of $\Lambda(1-s)$ and vice-versa and therefore the lambda function satisfies the functional equation

$$\Lambda(s) = \Lambda(1-s)$$

As $\zeta(s)$ is well defined for $\Re(s) > 1$, it also well defined, via the functional equation of Λ, for $\Re(s) < 0$, the difference between the two domains coming from the difference of behavior of the gamma function Γ.

We can easily extend $\zeta(s)$ to the domain $\Re(s) > 0$ using the fact that $\zeta(s)$ has a pole of order 1 at $s = 1$ and computing $\zeta(s)$ as

$$\zeta(s) = \frac{1}{s-1} + \cdots$$

$\Lambda(s)$ being now defined on the half plane $\Re(s) > 0$, the functional equation can be interpreted as a symmetry relative to the line $\Re(s) = \frac{1}{2}$, hence the major role of this line which is called the *critical line* of $\zeta(s)$.

The Γ function has no zeroes but has poles exactly on negative integers $-k$ ($k \geq 0$) where it has residue $\frac{(-1)^k}{k!}$.

For $s = -2k$ with $k > 1$, the functional equation reads

$$\zeta(-2k)\Gamma\left(-k\right)\pi^k = \zeta(1+2k)\Gamma\left(\frac{1+2k}{2}\right)\pi^{-\frac{1+2k}{2}}$$

and as the rhs is finite (the only pole of $\zeta(s)$ is $s = 1$) while $\Gamma\left(-k\right)$ is a pole, we must have $\zeta(-2k) = 0$.

These are called the *trivial zeroes* of the zeta function.

One of the main interests of $\zeta(s)$ is to have *non trivial zeroes which encode the distribution of primes* in the following sense.

For x a positive real, let $\pi(x)$ be the number of primes $p \leq x$.

From Gauss (1792, 15 years old) and Legendre (1808) to Hadamard (1896) and De La Vallée Poussin (1896) an asymptotic formula, called the *prime number theorem*, was proved and deeply investigated:

$$\pi(x) \sim \frac{x}{\log(x)}$$

A better approximation, due to Gauss (1849), is $\pi(x) \sim \mathrm{Li}(x)$ where the logarithmic integral is $\mathrm{Li}(x) = \int_2^x \frac{dx}{\log(x)}$.

For small n, $\pi(x) < \mathrm{Li}(x)$, but Littelwood proved in 1914 that the inequality reverses an infinite number of times.

The prime number theorem is a consequence of the fact that $\zeta(s)$ has no zeroes on the line $1 + it$ (recall that 1 is the pole of $\zeta(s)$). It as been improved with better approximations by many great arithmeticians.

In his 1859 paper, Riemann proved the fantastic result that $\pi(x)$ can be computed as the sum of a series whose terms are indexed by the non trivial zeroes of $\zeta(s)$.

It can be proved easily that all the non trivial zeroes of $\zeta(s)$ must lie inside the critical strip $0 < \Re(s) < 1$. Due to the functional equation they are symmetric relatively to the critical line and it is known that there exist an infinity of zeroes on the critical line and that the zeroes "concentrate" in a precise sense on the critical line.

An enormous amount of computations from Riemann time to actual supercomputers (ZetaGrid: more than 10^{12} zeroes in 2005) via Gram, Backlund, Titchmarsh, Turing, Lehmer, Lehman, Brent, van de Lune, Wedeniwski, Odlyzko, Gourdon, and others, shows that all computed zeroes lie on the critical line $\Re(s) = \frac{1}{2}$.

The celebrated *Riemann hypothesis*, one of the deepest unsolved problem (8th Hilbert problem), claims that in fact they all lie on the critical line.

Dirichlet's *L*-functions

Dirichlet's L-functions generalize $\zeta(s)$. They have the general form

$$\sum_{n\geq 1} \frac{a_n}{n^s}$$

and under some "multiplicative" conditions on the a_n can be factorized into Euler products

$$\prod_{p\in\mathcal{P}} \left(1 + \frac{a_p}{p^s} + \ldots \frac{a_{p^m}}{p^{ms}} + \ldots\right)$$

1. The condition is of course that the coefficients a_n are *multiplicative* in the sense that $a_1 = 1$ and, if $n = \prod p_i^{r_i}$, $a_n = \prod a_{p_i^{r_i}}$.
2. Moreover if the a_n are *strictly multiplicative* in the sense that $a_{p^m} = (a_p)^m$ then the series can be factorized in a *first degree* (or linear) Euler product

$$\prod_{p \in \mathcal{P}} \frac{1}{1 - \frac{a_p}{p^s}}.$$

3. If $a_1 = 1$ and if for every prime p there exists an integer d_p s.t.

$$a_{p^m} = a_p a_{p^{m-1}} + d_p a_{p^{m-2}}$$

then the series can be factorized in a *second degree* (or quadratic) Euler product

$$\prod_{p \in \mathcal{P}} \frac{1}{1 - \frac{a_p}{p^s} - \frac{d_p}{p^{2s}}}$$

The most important examples of Dirichlet series are given by Dirichlet L-functions where the a_n are the values $\chi(n)$ of a *character* $\bmod m$, that is of a multiplicative morphism

$$\chi : (\mathbb{Z}/m\mathbb{Z})^* \to \mathbb{C}$$

$$L_\chi = \sum_{n \geq 1} \frac{\chi(n)}{n^s}$$

As χ is multiplicative, the a_n are strictly multiplicative and the series can be factorized in a *first degree* Euler product.

The theory of the zeta function can be straightforwardly generalized (theta function, automorphy symmetries, lambda function, functional equation) to these Dirichlet L-functions.

Modularity

We have defined L-functions L_E of EC. We will now define a *completely different* class of L-functions L_f associated to what are called *modular forms*. By construction, the L_f have extremely deep arithmetic properties. An EC curve is said *modular* if there exists a "good" f s.t. $L_E = L_f$.

By definition, *modular* EC have strong arithmetic properties and therefore to say that *all* EC are modular is to say that there exist highly non trivial constraints and that such constraints imply FLT.

We have to define f and L_f.

ECs as complex tori

As cubic plane projective curves, EC are commutative algebraic groups. Let P and Q be two points of E. As the equation is cubic, the line PQ intersects E in a third point R. The group law is then defined by setting $P + Q + R = 0$.

A great discovery (Abel, Jacobi, up to Riemann) is that they are isomorphic to their Jacobian, which is a complex torus.

Let $E = E_{\text{cub}}$ be a regular cubic. Topologically it is a torus and it is endowed with a complex structure making it a compact Riemann surface.

Let γ_1 and γ_2 be two loops corresponding to a parallel and a meridian of E (they constitute a $\mathbb{Z}$-basis of the first integral homology group $H_1(E, \mathbb{Z})$).

Up to a factor, there exists a single *holomorphic* 1-form ω on E. Its periods $\omega_i = \int_{\gamma_i} \omega$ generate a lattice Λ in $\mathbb{C}$ and we can consider the torus $E_{\mathrm{tor}} = \mathbb{C}/\Lambda$ which is called the *Jacobian* of E.

If a_0 is a base point in E, the integration of the 1-form ω defines a map

$$\begin{array}{rcl} \Phi : E_{\mathrm{cub}} & \longrightarrow & E_{\mathrm{tor}} \\ a & \longmapsto & \int_{a_0}^{a} \omega \end{array}$$

(the map is well defined since two pathes from a_0 to a differ by a $\mathbb{Z}$-linear combination of the γ_i and the values of ω differ by a point of the lattice Λ).

Theorem. Φ is an *isomorphism* between E_{cub} and E_{tor}.

Elliptic functions

We consider now the representation of elliptic curves as complex tori $E = \mathbb{C}/\Lambda$ with Λ a lattice $\{m\omega_1 + n\omega_2\}_{m,n\in\mathbb{Z}}$ in $\mathbb{C}$ with $\mathbb{Z}$-basis $\{\omega_1, \omega_2\}$. If $\tau = \omega_2/\omega_1$, we can suppose $\mathrm{Im}\,(\tau) > 0$, that is $\tau \in \mathcal{H}$ where $\mathcal{H}$ is the Poincaré upper half complex plane.

The EC defined by $\{1, \tau\}$ is denoted Λ_τ.

The complex valued functions f on $E = \mathbb{C}/\Lambda$ are *doubly periodic* functions on $\mathbb{C}$. They are called *elliptic functions*. E being compact, such an f cannot be holomorphic without being constant according to Liouville theorem; f can only be a *meromorphic* function if it is not constant.

Applying the residue theorem successively to f, f'/f, and zf'/f we can show:

1. f possesses at least 2 poles.
2. If the m_i are the order of the singular points a_i (poles and zeroes) of f, $\sum m_i = 0$ (this says that the *divisor* $\mathrm{div}(f)$ is of degree 0).
3. $\sum m_i a_i \equiv 0 \mod \Lambda$.

One elliptic function is of particular interest since it generates with its derivative the field of all elliptic functions. It is the Weierstrass function $\wp(z)$ which is the most evident even function having a double pole at the points of the lattice Λ.

Let $\Lambda' = \Lambda - \{0\}$, the definition is:

$$\wp(z) = \frac{1}{z^2} + \sum_{\omega \in \Lambda'} \left(\frac{1}{(z-\omega)^2} - \frac{1}{\omega^2} \right)$$

The derivative $\wp'(z)$ is an odd function possessing triple poles at the points of Λ:

$$\wp'(z) = -2 \sum_{\omega \in \Lambda} \frac{1}{(z-\omega)^3}$$

Theorem. $\wp(z)$ and $\wp'(z)$ generate the field of elliptic functions on the elliptic curve $E = \mathbb{C}/\Lambda$.

What are the relations between these two definitions of elliptic curves, one algebraic and the other analytic?

In one sense, from complex tori to cubics, the relation is quite simple. Indeed $\wp(z)^3$ and $\wp'(z)^2$ have both a pole of order 6 at 0 and must be related. Some (tedious) computations on their Laurent expansions show that there exists effectively an algebraic relation between $\wp(z)$ and $\wp'(z)$, namely

$$\wp'(z)^2 = 4\wp(z)^3 - g_2\wp(z) - g_3$$

with $g_2 = 60G_4$ and $g_3 = 140G_6$, G_m being the *Eisenstein series*

$$G_m = \sum_{\omega \in \Lambda'} \frac{1}{\omega^m}$$

This means that $(\wp(z), \wp'(z))$ is on the elliptic curve E_{cub} of equation

$$y^2 = 4x^3 - g_2 x - g_3$$

which discriminant is:

$$\Delta = (g_2)^3 - 27\,(g_3)^2$$

the lattice Λ corresponding to the point at infinity in the y direction.

One can verify that $\Delta \neq 0$ and that E is therefore *regular*.

One of the great advantage of the torus representation is that the group structure become evident. Indeed $E_{\mathrm{tor}} = \mathbb{C}/\Lambda$ inherits the additive group structure of $\mathbb{C}$ and through the parametrization by $\wp(z)$ and $\wp'(z)$ this group structure is transfered to E_{cub}.

$SL(2,\mathbb{Z})$ action

The isomorphism between an EC and its Jacobian is the beginning of the great story of *Abelian varieties*.

In this context, *where algebraic structures are translated and coded into analytic ones*, one can develop an extremely deep theory of *arithmetic* properties of elliptic curves. Its "deepness" comes from *the analytic coding of arithmetics*.

Let $E = \mathbb{C}/\Lambda$ be an EC considered as a complex torus. To correlate *univocally* E and its "module" τ we must look at the transformation of τ when we change the $\mathbb{Z}$-basis of Λ. Let $\{\omega_1', \omega_2'\}$ another $\mathbb{Z}$-basis.

We have $\begin{pmatrix} \omega_2' \\ \omega_1' \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \omega_2 \\ \omega_1 \end{pmatrix}$ with $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ an integral matrix.

But γ must be inversible and its inverse must therefore be also an integral matrix, so $\mathrm{Det}\,(\gamma) = ad - bc = 1$ and $\gamma \in SL(2, \mathbb{Z})$.

γ acts on τ via fractional linear Möbius transformations:

$$\gamma\,(\tau) = \frac{a\tau + b}{c\tau + d}\ .$$

Modular functions for $SL(2,\mathbb{Z})$

The concept of modular form arises naturally when we consider *holomorphic* $SL(2,\mathbb{Z})$*-invariant differentials* on the Poincaré half-plane $\mathcal{H}$. Let $f(\tau)d\tau$ be a 1-form on $\mathcal{H}$ with f an holomorphic function on $\mathcal{H}$ and consider $f(\tau')d\tau'$ with $\tau' = \gamma(\tau)$. We have

$$\begin{aligned} f(\tau')d\tau' &= f\left(\frac{a\tau+b}{c\tau+d}\right)\frac{(c\tau+d)\,a-(a\tau+b)\,c}{(c\tau+d)^2}d\tau \\ &= f\left(\frac{a\tau+b}{c\tau+d}\right)\frac{1}{(c\tau+d)^2}d\tau \text{ since } ad-bc=1 \end{aligned}$$

We see that in order to get the invariance $f(\tau)d\tau = f(\tau')d\tau'$ we need $f\left(\frac{a\tau+b}{c\tau+d}\right)\frac{1}{(c\tau+d)^2} = f(\tau)$, i.e.

$$f(\gamma(\tau)) = (c\tau+d)^2 f(\tau)\ .$$

Hence the general definition:

Definition. An holomorphic function on $\mathcal{H}$ is a *modular function of weight* k if $f(\gamma(\tau)) = (c\tau + d)^k f(\tau)$ for every $\gamma \in SL(2, \mathbb{Z})$.

We note that the definition implies $f = 0$ for *odd* weights since $-I \in SL(2, \mathbb{Z})$ and if k is odd

$$f(-I\tau) = f\left(\frac{-\tau}{-1}\right) = f(\tau) = (-1)^k f(\tau) = -f(\tau)$$

The weight 0 means that the *function* f is $SL(2, \mathbb{Z})$-invariant. The weight 2 means that the 1-*form* $fd\tau$ is $SL(2, \mathbb{Z})$-invariant.

A modular function of weight k can also be interpreted as an homogeneous holomorphic function of degree $-k$ defined on the lattices Λ. If we define $\tilde{f}(\Lambda)$ by $\tilde{f}(\Lambda) = \omega_1^{-k} f(\tau)$ we see that for f to be modular of weight k is equivalent to $\tilde{f}(\alpha\Lambda) = \alpha^{-k} f(\Lambda)$.

To be modular, f has only to be modular on generators of $SL(2, \mathbb{Z})$, two generators being the translation $\tau \rightarrow \tau + 1$ and the inversion $\tau \rightarrow -1/\tau$. Therefore f is modular of weight k iff

$$\begin{cases} f(\tau + 1) = f(\tau) \\ f\left(-\frac{1}{\tau}\right) = (-\tau)^k f(\tau) \end{cases}$$

These are properties of *automorphy*, where “automorphy” means some sort of invariance of entities defined on the Poincaré plane $\mathcal{H}$ with respect to a countable subgroup of the group $SL(2, \mathbb{Z})$ acting naturally on $\mathcal{H}$.

We already met modular functions in the theory of elliptic curves:

1. the Eisenstein series G_{2k} of weight $2k$,
2. the *elliptic invariants* which are the coefficients g_2 of weight 4 and g_3 of weight 6 of the Weierstrass equation associated to a complex torus,
3. the discriminant $\Delta = (g_2)^3 - 27\,(g_3)^2$ of weight 12,
4. the modular invariant j of weight 0.

Fourier expansion, modular forms, and cusp forms

The fact that a modular function f is invariant by the translation $\tau \rightarrow \tau + 1$ means that it is *periodic* of period 1 and therefore can be expanded into a *Fourier series*

$$f(\tau) = \sum_{n \in \mathbb{Z}} c_n e^{2i\pi n\tau} = \sum_{n \in \mathbb{Z}} c_n \kappa^n \text{ with } \kappa = e^{2i\pi\tau}$$

The variable $\kappa = e^{2i\pi\tau}$ is called the *nome* (and is traditionally denoted by q). It is a mapping $\mathcal{H} \rightarrow \mathbb{D} - \{0\}$, $\tau \mapsto \kappa = e^{2i\pi\tau}$ which uniformizes $\mathcal{H}$ at infinity in the sense that, if $\tau = x + iy$, $\kappa = e^{2i\pi x} e^{-2\pi y} \underset{y \rightarrow \infty}{\longrightarrow} 0$. The boundary $y = 0$ of $\mathcal{H}$ maps cyclically on the boundary $\mathbb{S}^1 = \partial\mathbb{D}$ of $\mathbb{D}$.

If we use this representation, the second property of modularity

$$f\left(-\frac{1}{\tau}\right) = (-\tau)^k f(\tau)$$

imposes very strict *constraints* on the Fourier coefficients c_n and therefore modular functions generate *very special series* $\{c_n\}_{n\in\mathbb{Z}}$.

For controlling the holomorphy of f *at infinity* one introduces two restrictions on the general concept of a modular *function* of weight k.

- *Definition*. f is called a modular *form* of weight k if f is *holomorphic* at infinity, that is if its Fourier coefficients $c_n = 0$ for $n < 0$.
- *Definition*. Moreover, f is called a *cusp form* if f *vanishes* at infinity, that is if $c_0 = 0$ (then $c_n = 0$ for $n \leq 0$).

It is traditional to note M_k the space of modular forms of weight k, and $S_k \subset M_k$ the space of cusp forms of weight k.

Eisenstein series

$$G_k(\tau) = \sum_{(m,n)\in\mathbb{Z}\times\mathbb{Z}-\{0,0\}} \frac{1}{(m\tau+n)^k}$$

are modular forms. The power k must be even $(k = 2r)$ for if k is odd the $(-m,-n)$ and (m,n) terms cancel.

The discriminant Δ of elliptic curves,

$$\Delta(\tau) = (g_2(\tau))^3 - 27(g_3(\tau))^2$$

with $g_2(\tau) = 60G_4(\tau)$ and $g_3(\tau) = 140G_6(\tau)$, is a modular function of weight 12.

It expands into

$$\Delta(\tau) = q - 24q^2 + 252q^3 - 1472q^4 + \dots$$

One can show that it is given by the infinite product

$$\Delta(\tau) = q \prod_{r=1}^{r=\infty} (1 - q^r)^{24}$$

It is therefore a *cusp form* $\Delta \in S_{12}$. We note that $\Delta(\tau) = 0$

- exactly for $q^r = 1$,
- that is $e^{2i\pi r\tau} = 1$,
- that is $r\tau \in \mathbb{Z}$,
- that is $\tau \in \mathbb{Q}$,
- that is for the rational points on the boundary of $\mathcal{H}$, which are called *cusp points*.

$\Delta(\tau)$ vanishes nowhere on $\mathcal{H}$.

On the contrary, the modular invariant j of weight 0 expands into

$$j(\tau) = \frac{1}{q} + 744 + 196\,884q + 21\,493\,760q^2 + \dots$$

It has a pole at infinity and fails to be a modular form.

The fundamental importance of the Eisenstein series and the discriminant is that they enables to determine the spaces M_k and S_k.

We will see later that they are eigenvectors of the Hecke operators defined on the spaces M_k and S_k.

1. $M_0 \simeq \mathbb{C}$ since an f which is $SL(2,\mathbb{Z})$-invariant and holomorphic on $\mathcal{H}$ and at infinity is holomorphic on the quotient $(\mathcal{H}/SL(2,\mathbb{Z})) \cup \{\infty\}$ which is compact. f is therefore constant by Liouville theorem.
2. $M_k = 0$ for $k < 0$ since if $f \neq 0 \in M_k$, then f^{12} is of weight $12k$, Δ^{-k} is of weight $-12k$, and $f^{12}\Delta^{-k} \in M_0$ but is without constant term. Therefore $f = 0$.
3. $M_k = 0$ for k odd since, if we take $\gamma = -I$, $f(\gamma(\tau)) = f(\tau) = -f(\tau)$, and $f \equiv 0$.
4. $M_k = 0$ for $k = 2$.
5. For k even $k > 2$, $M_k = \mathbb{C}G_k \oplus S_k$ since S_k is of codimension 1 in M_k and G_k has a constant term.
6. $S_k \simeq M_{k-12}$. Indeed if $f \in S_k$, $f/\Delta \in M_{k-12}$. Since $\Delta \neq 0$, f/Δ (which is of weight $k-12$) is holomorphic on $\mathcal{H}$ and, as $c_n = 0$ for $n \leq 0$ for f and Δ, $c_n = 0$ for $n < 0$ for f/Δ and $f/\Delta \in M_{k-12}$. Reciprocally, if $g \in M_{k-12}$ then $g\Delta \in S_k$. $S_k \simeq M_{k-12}$ implies, via (2), $\dim(S_k) = 0$ for $k < 12$ and, via (5), $\dim(M_k) = 1$ for $k < 12$.

It is therefore easy to compute the dimension of M_k: e.g. for $k = 12$, via (6) and (1), $\dim(S_k) = \dim(M_0) = 1$ and, via (5), $\dim(M_k) = 2$.

k	0	1	2	3	4	5	6	7
$\dim(M_k)$	1	0	0	0	1	0	1	0

k	8	9	10	11	12	13	14	15
$\dim(M_k)$	1	0	1	0	2	0	1	0

Such dimensions imply a lot of deep arithmetical relations because every time we can associate to d situations d modular forms of M_k and we have $d > \dim(M_k)$, then, as was emphasized by Don Zagier ([75], (p.240)),

> *"We get "for free" information – often highly non trivial – relating these different situations."*

Moreover we will see that the M_k are spanned by modular forms whose Fourier series have *rational* coefficients c_n. As Don Zagier also explains:

> *"It is this phenomenon which is responsible for the richness of the arithmetic applications of the theory of modular forms."*

We have seen that

$$\Delta(\tau) = (60 G_4(\tau))^3 - 27 (140 G_6(\tau))^2 .$$

It is a general fondamental fact:

Theorem. Every modular form can be expressed in a unique way as a *polynomial* in G_4 and G_6.

L-functions of cusp forms

If f is a cusp form of weight k, i.e. $f \in S_k$, then

$$f(\tau) = \sum_{n \geq 1} c_n \kappa^n$$

with the nome $\kappa = e^{2i\pi\tau}$. We associate to f the L-function:

$$L_f(s) = \sum_{n \geq 1} \frac{c_n}{n^s}$$

having the same coefficients. These L-functions encode a lot of arithmetical information. They come essentially as *Mellin transform* of their generating cusp form.

Paralleling the case of Riemann ζ function, we introduce the Mellin transform

$$\Lambda_f(s) = \int_0^\infty f\,(it)\,t^s \frac{ds}{s}$$

of the cusp form f on the positive imaginary axis and we compute

$$\Lambda_f(s) = \frac{1}{(2\pi)^s}\Gamma(s)L_f(s)$$

The modular invariance of f and its good behavior at infinity imply that the c_n are bounded in norm by $n^{k/2}$ and therefore $L_f(s)$ is absolutely convergent in the half-plane $\Re(s) > \frac{k}{2} + 1$.

As the Riemann ζ function, the L-functions $L_f(s)$ satisfy a *functional equation*. It is the content of a deep theorem due to Hecke:

Hecke theorem. $L_f(s)$ and $\Lambda_f(s)$ are *entire* functions and $\Lambda_f(s)$ satisfies the functional equation

$$\Lambda_f(s) = (-1)^{k/2}\Lambda_f(k-s)$$

The modular curve $X_0\left(1\right)$

We need to introduce now the *modular curves* $X_0\left(N\right)$ of different *levels* N.

For $N = 1$, $X_0\left(1\right)$ is the compactification of the quotient $\mathcal{H}/SL(2,\mathbb{Z})$ of $\mathcal{H}$ by the modular group $SL(2,\mathbb{Z})$, i.e. of its standard fondamental domain R.

R is the domain of $\mathcal{H}$ defined by $-\frac{1}{2} \leq \Re\left(\tau\right) < \frac{1}{2}$ and $|\tau| > 1$. It contains on its boundary the 3 remarkable points $i = e^{i\frac{\pi}{2}}$, $\zeta_3 = e^{2i\frac{\pi}{3}} = \rho^2$, and $\zeta_3 + 1 = -\zeta_3^2 = \rho = e^{i\frac{\pi}{3}}$.

The modular invariant j maps R conformally onto $\mathbb{C} \cup \{\infty\}$ with cuts on the real axis along $\{-\infty, 0\}$ and $\{1, \infty\}$. As the discriminant Δ has only a simple zero at ∞, j has only a single simple pole at ∞.

It can be shown that the field of meromorphic fonctions $K\left(X_0\left(1\right)\right)$ is generated by the modular invariant j.

Theorem. $K\left(X_0(1)\right) = \mathbb{C}\left(j\right)$.

The inclusion $\mathbb{C}\left(j\right) \subseteq K\left(X_0(1)\right)$ is trivial. Conversely, let $f\left(\tau\right) \in K\left(X_0(1)\right)$ with poles π_i (counted with multiplicity). Consider the function $g\left(\tau\right) = f\left(\tau\right) \prod_i \left(j\left(\tau\right) - j\left(\pi_i\right)\right)$. It is a modular function of weight 0 and level 1 without poles in $\mathcal{H}$. If g has a pole of order n at ∞ there exists c s.t. $g - cj^n$ is without pole in $\overline{\mathcal{H}}$ and is therefore constant. This implies $f\left(\tau\right) \in \mathbb{C}\left(j\right)$.

Elliptic points

The modular curve $X_0(N)$ and its congruence group $\Gamma_0(N)$

The *modular curve* of level N, $X_0(N)$, classifies pairs (Λ, C) of a lattice Λ and a N-cyclic group C of torsion points.

The *modular curve* of level N, $X_1(N)$, classifies pairs (Λ, x) of a lattice Λ and a N-torsion point x $(Nx = 0)$.

For the lattice $\Lambda_\tau = \mathbb{Z} \oplus \tau\mathbb{Z}$ $(\tau \in \mathcal{H})$ of basis $\{1, \tau\}$, C_τ is simply the cyclic subgroup generated by $1/N$.

For $N = 1$, C is reduced to the origin 0 $(1x = x = 0)$.

The $X_0(N)$ are intimately associated to the congruence groups $\Gamma_0(N)$ which are *smaller* than $SL(2,\mathbb{Z})$. This corresponds to the introduction of the key concept of *level N* of a modular function, the classical ones being of level 1.

The congruence subgroup $\Gamma_0(N)$ of $SL(2,\mathbb{Z})$ is defined by a restriction on the term c:

$$\begin{aligned}\Gamma_0(N) &= \left\{\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2,\mathbb{Z}) : c \equiv 0 \mod N\right\} \\ &= \left\{\begin{pmatrix} a & b \\ kN & d \end{pmatrix} \in SL(2,\mathbb{Z})\right\}\end{aligned}$$

In $\Gamma_1(N)$ we have moreover $a, b \equiv 1 \mod N$.
We note that $\begin{pmatrix} 1 & N \\ 0 & 1 \end{pmatrix} \in \Gamma_0(N)$. Of course $\Gamma_0(1) = SL(2,\mathbb{Z})$.

A fundamental domain R_N of $\Gamma_0(N)$ can be generated from R and, if $N > 1$, has *cusps* which are rational points on the boundary of $\mathcal{H}$.

Indeed, let $\Gamma_0(1) = \bigcup_j \beta_j \Gamma_0(N)$, $\beta_j = \begin{pmatrix} a_j & b_j \\ c_j & d_j \end{pmatrix} \in SL(2, \mathbb{Z})$, be a decomposition of $\Gamma_0(1)$ in $\Gamma_0(N)$-orbits.

A fundamental domain R_N of $\Gamma_0(N)$ is $R_N = \bigcup_j \beta_j^{-1}(R)$ where R is a fundamental domain of $SL(2, \mathbb{Z})$, $\left(\beta_j^{-1} = \begin{pmatrix} d_j & -b_j \\ -c_j & a_j \end{pmatrix}\right)$, and the cusps of R_N are the rational points of the boundary of $\mathcal{H}$ image of the infinite point: $\beta_j^{-1}(\infty) = -\frac{d_j}{c_j} \in \mathbb{Q}$.

$X_0(N)$ is the compactification of the quotient of $\mathcal{H}$ by $\Gamma_0(N)$.

The genus of $X_0(N)$

Let $g(N)$ be the genus of $X_0(N)$. Barry Mazur proved a beautiful theorem on $g(N)$. For low genus he got:

genus g	level N
0	$1, \ldots, 10, 12, 13, 16, 18, 25$
1	$11, 14, 15, 17, 19, 20, 21, 24, 27, 32, 36, 49$
2	$22, 23, 26, 28, 29, 31, 37, 50$

We will use in particular the crucial fact that $g(2) = 0$.

The remarkable general formula is a sum of four terms:

$$g = 1 + \frac{\mu}{12} - \frac{\nu_2}{4} - \frac{\nu_3}{3} - \frac{\nu_\infty}{2}$$

with

$$\begin{cases} \mu = [SL(2,\mathbb{Z}) : \Gamma_0(N)] = N \prod\limits_{p \nmid N} \left(1 + \frac{1}{p}\right) \\ \nu_2 = \prod\limits_{p|N} \left(1 + \left(\frac{-1}{p}\right)\right) \text{ if } 4 \nmid N \text{ and } = 0 \text{ if } 4 \mid N \\ \nu_3 = \prod\limits_{p|N} \left(1 + \left(\frac{-3}{p}\right)\right) \text{ if } 9 \nmid N \text{ and } = 0 \text{ if } 9 \mid N \\ \nu_\infty = \sum\limits_{d \geq 0, d|N} \varphi\left(d, \frac{N}{d}\right) \text{ where } \varphi \text{ is the Euler function} \end{cases}$$

where in the second and third equations $\left(\frac{-1}{p}\right)$ and $\left(\frac{-3}{p}\right)$ are the Legendre symbols:

$$\left(\frac{-1}{p}\right) = \begin{cases} 0 \text{ if } p = 2 \\ 1 \text{ if } p \equiv 1 \pmod 4 \\ -1 \text{ if } p \equiv 3 \pmod 4 \end{cases}$$

$$\left(\frac{-3}{p}\right) = \begin{cases} 0 \text{ if } p = 3 \\ 1 \text{ if } p \equiv 1 \pmod 3 \\ -1 \text{ if } p \equiv 2 \pmod 3 \end{cases}$$

The field of rational functions of $X_0(N)$

A perspicuous way of defining the modular curve $X_0(N)$ is to do it from its *field K of rational functions*. This is the way adopted by David Rohrlich [50].

One starts with an EC $\mathcal{E}$ no longer defined over $\mathbb{Q}$ but over the field of *rational functions* $\mathbb{Q}(t)$. Moreover, one requires that its j-invariant should be $j(\mathcal{E}) = t$.

This means that we look in fact at a family $\mathcal{E} = (E_t)$ of EC over $\mathbb{Q}$ having t as j-invariant. If t is noted j we want the "tautology" $j = j$.

An example of such a curve $\mathcal{E}$ is given by the Weierstrass equation

$$y^2 = 4x^3 - \frac{27t}{t-1728}x - \frac{27t}{t-1728}$$

(using the formula for j it is trivial to verify that $j = t$).

One chooses then a point $\mathcal{P}$ of order N on $\mathcal{E}$ and looks at the cyclic group $\mathcal{C}$ of order N generated by $\mathcal{P}$. $\mathcal{C}$ is a family of cyclic groups C_t of the E_t parametrized by t. In some sense, $(\mathcal{E}, \mathcal{C})$ is a *generic* or universal elliptic curve endowed with the supplementary structure $\mathcal{C}$.

The subfield of $\overline{\mathbb{Q}}(t)$ fixed by the automorphisms $\sigma \in \mathrm{Gal}\left(\overline{\mathbb{Q}}(t)/\mathbb{Q}(t)\right)$ which fix $\mathcal{C}$ (i.e. such that $\sigma(\mathcal{C}) = \mathcal{C}$) defines a finite extension K of $\mathbb{Q}(t)$ whose field of constants is $\mathbb{Q}$ $(\overline{\mathbb{Q}} \cap K = \mathbb{Q})$.

K is the field of rational functions of a smooth projective curve over $\mathbb{Q}$ and this curve is nothing else than $X_0(N)$.

The link with the previous definition is done using the remark that K is in fact a subfield of $\mathbb{Q}(t, \mathcal{E}[N])$ and the theorem that

$$\mathrm{Gal}\left(\mathbb{Q}(t, \mathcal{E}[N])/\mathbb{Q}(t)\right) \simeq GL(2, \mathbb{Z}/N\mathbb{Z})$$

One associates now to the subgroup H of $GL(2, \mathbb{Z}/N\mathbb{Z})$ defining K a subgroup Γ of $SL(2, \mathbb{Z})$ which is the transpose of the inverse image of $H \cap SL(2, \mathbb{Z})$ by the quotient $SL(2, \mathbb{Z}) \rightarrow SL(2, \mathbb{Z}/N\mathbb{Z})$.

Using the fact that $-I \in H$ and that the "determinant" map $\det : H \rightarrow (\mathbb{Z}/N\mathbb{Z})^*$ is surjective, one shows that Γ is nothing else than $\Gamma_0(N)$ and that $X_0(N)(\mathbb{C}) \simeq \overline{\mathcal{H}}/\Gamma_0(N)$ (see Rohrlich [50]).

This description makes evident that $X_0(N)$ classifies the pairs (E, C).

In fact the field of rational functions on $X_0(N)$ is easy to compute. Let $j_N(\tau)$ be the function defined by $j_N(\tau) = j(N\tau)$.

Theorem. $K(X_0(N)) = \mathbb{C}(j, j_N)$.

Indeed, let α_i be representants of the orbits of $\Gamma_0(N)$ acting on the set $M(N)$ of integral matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with determinant $ad - bc = N$ and $c \equiv 0 \bmod N$. The α_i are chosen as $\begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$ with $ad = N$, $d \geq 1$, $0 \leq b < d$, $(a, b, d) = 1$. Then $j_N = j \circ \alpha$ with $\alpha = \begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix}$ is a root of the polynomial

$$\Phi_N(x) = \prod_{i=1}^{i=\mu(N)} (x - j \circ \alpha_i) \text{ with (see above)}$$

$$\mu(N) = [SL(2, \mathbb{Z}) : \Gamma_0(N)] = N \prod_{p|N} \left(1 + \frac{1}{p}\right)$$

But $\Phi_N(x)$ has its coefficients in $\mathbb{Z}[j]$, is irreducible over $\mathbb{C}(j)$ and is the minimal polynomial of j_N over $\mathbb{C}(j)$. We have therefore (see Boston)

$$K(X_0(N)) = K(X_0(1))(j_N) = \mathbb{C}(j, j_N)$$

One generalizes trivially the definition of modularity to this more general context.

1. A modular function of weight k and level N is an $f(\tau)$ satisfying the invariance condition $f(\gamma(\tau)) = (c\tau + d)^k f(\tau)$ $\forall \gamma \in \Gamma_0(N)$.
2. A modular function of weight k and level N is a modular form $f(\tau) \in M_k(N)$ if it is holomorphic not only at infinity but also at the cusps.
3. A modular form of weight k and level N is a cusp form $f(\tau) \in S_k(N)$ if moreover it vanishes at infinity and at the cusps. The dimension of $S_k(N)$ is the genus $g(N)$ of the modular curve $X_0(N)$.
4. If $f(\tau) \in M_k(N)$, $f(\tau)$ is N-periodic and can be developped at infinity in a Fourier series $f(\gamma(\tau)) = \sum\limits_{n \geq 0} c_n \kappa^n$ with nome $\kappa = e^{\frac{2i\pi\tau}{N}}$

A further generalization consists in introducing a *character*

$$\varepsilon : \left(\frac{\mathbb{Z}}{N\mathbb{Z}}\right)^{+} \to \mathbb{C}^{\times}$$

(what is called in German a *Nebentypus*) and defining the invariance condition no longer by $f\left(\gamma\left(\tau\right)\right) = \left(c\tau + d\right)^{k} f(\tau)$ but by

$$f\left(\gamma\left(\tau\right)\right) = \left(c\tau + d\right)^{k} \varepsilon(d) f(\tau) \ .$$

We get that way spaces $M_k\left(N, \varepsilon\right)$ and $S_k\left(N, \varepsilon\right)$.

The Jacobian $J_0(N)$

Let g be the genus of the modular curve $X_0(N)$ and let $(c_1, \ldots, c_{2g})$ be a $\mathbb{Z}$-basis of its integral homology $H_1(X_0(N), \mathbb{Z})$. Let $(\omega_1, \ldots, \omega_g)$ be the dual $\mathbb{C}$-basis of the cohomology group $H^1(X_0(N), \mathbb{Z})$ and $(f_1, \ldots, f_g)$ the associated basis of $S_2(N)$. One defines a map Θ — called the *Abel-Jacobi* morphism — from the modular curve $X_0(N)$ on $\mathbb{C}^g$ by

$$\Theta(\tau) = \left\{ \int_{\tau_0}^{\tau} f_j(z)\, dz \right\}_{j=1,\ldots,g}$$

where τ_0 is a base point on $X_0(N)$. $\Theta(\tau)$ is well defined modulo the lattice $\Lambda(X_0(N))$ generated over $\mathbb{Z}$ by the $2g$ points of $\mathbb{C}^g$

$$u_k = \left\{ \int_{c_k} f_j(z)\, dz \right\}_{j=1,\ldots,g}$$

The Jacobian $J_0(N)$ is the quotient $\mathbb{C}^g / \Lambda(X_0(N))$.

Modular elliptic curves

Eichler and Shimura investigated the possibility of expressing the L-function $L_E(s)$ of an EC as a *modular* L-function $L_f(s)$ for a certain modular form f (i.e. a $\Gamma_0(N)$-invariant holomorphic differential $f(z)dz$ on the modular curve $X_0(N)$).

For the construction of an E from an f to be possible, f must be a cusp form of level N and weight 2. Let therefore $f \in S_2(N)$.

We integrate the differential $f(z)dz$ and get the function on $\mathcal{H}$

$$F(\tau) = \int_{\tau_0}^{\tau} f(z)dz$$

where τ_0 is a base point in $\mathcal{H}$.

Let now $\gamma \in \Gamma_0(N)$. Since $f(z)dz$ is $\Gamma_0(N)$-invariant, we have:

$$\begin{aligned} F(\gamma(\tau)) &= \int_{\tau_0}^{\gamma(\tau)} f(z)dz = \int_{\tau_0}^{\gamma(\tau_0)} + \int_{\gamma(\tau_0)}^{\gamma(\tau)} = \int_{\tau_0}^{\gamma(\tau_0)} + \int_{\tau_0}^{\tau} \\ &= F(\tau) + \Phi_f(\gamma) \text{ with } \Phi_f(\gamma) = \int_{\tau_0}^{\gamma(\tau_0)} f(z)dz \end{aligned}$$

Φ_f is a map $\Phi_f : \Gamma_0(N) \to \mathbb{C}$ and we see that *if* its image $\Phi_f(\Gamma_0(N))$ is a lattice Λ in $\mathbb{C}$ then the primitive $F(\tau)$ becomes a map

$$F : X_0(N) \to E = \mathbb{C}/\Lambda$$

which yields *a parametrization of the elliptic curve E by the modular curve $X_0(N)$*.

In that case E is called a *modular elliptic curve*.

Following Barry Mazur [36] we make a remark on this definition. We have seen that, as far as it is isomorphic with its Jacobian, a general EC E admits an *Euclidean* covering by $\mathbb{C}$, $\pi : \mathbb{C} \rightarrow E = \mathbb{C}/\Lambda$.

If E is defined over $\mathbb{Q}$ (that is "arithmetic") and modular, it admits also an *hyperbolic* covering by a modular curve $F : X_0 (N) \rightarrow E$ defined over $\mathbb{Q}$.

But the two types of coverings are completely different (the text was written in 1989 when the STW conjecture was still a conjecture).

> *"It is the confluence of two uniformizations, the Euclidean one, and the (conjectural) hyperbolic one of arithmetic type, that puts an exceedingly rich geometric structure on an arithmetic elliptic curve, and that carries deep implications for arithmetic questions."*

The great result of Eichler-Shimura's very technical construction is that if f is a *newform* (in the sense of Atkin and Lehner, see next section) then

1. Λ is effectively a lattice in $\mathbb{C}$;
2. $X_0(N)$, E and $F : X_0(N) \to E$ are defined over $\mathbb{Q}$ in a *compatible* way;
3. and the L-functions of the elliptic curve E and the cusp form f are equal: $L_E(s) = L_f(s)$.

The construction is mediated by the Jacobian curve $J_0(N)$ of the modular curve $X_0(N)$, the elliptic curve E being a *quotient* of the Jacobian. It is an astonishing result. As Knapp [32] explains:

> *"Two miracles occur in this construction [modular EC]. The first miracle is that $X_0(N)$, E, and the mapping can be defined compatibly over $\mathbb{Q}$. (...) The second miracle is that the L function of E matches the L function of the cusp form f."*

Hecke algebras and new eigenforms

What are newforms? Up to now, the L_f was defined as series $L_f(s) = \sum_{n\geq 1} \frac{c_n}{n^s}$ with $f(\tau) = \sum_{n\geq 1} c_n \kappa^n$ a Fourier series.

But, by construction, the L_E are Euler products encoding information *prime by prime*.

We need therefore to know what modular forms can be also Euler products. It is the scope of Hecke operators. The problem is rather technical and difficult.

For $SL(2,\mathbb{Z})$, Hecke's very beautiful idea was to solve it in two steps:

1. find linear operators $T_k(m)$ on the vector spaces M_k of modular forms which satisfy the relations of an Euler product;
2. look at their *simultaneous* eigenfunctions, which exist since the algebra $\mathcal{T}_k$ of the $T_k(m)$ is commutative.

These very particular modular *eigen*forms inherit very particular properties from those of Hecke operators. Their coefficients c_n are *algebraic integers* and satisfy the multiplicative relation $c_{nm} = c_n c_m$ if $(m,n) = 1$.

The Dirichlet L-function $L_f(s) = \sum_{n\geq 1} \frac{c_n}{n^s}$ can then be expressed as an Euler product.

The simplest way of defining Hecke operator is to start with the free group $\mathcal{L}$ generated by the lattices Λ of $\mathbb{C}$ (recall that it is the origin of the $SL(2,\mathbb{Z})$ action on the Poincaré half-plane $\mathcal{H}$).

If we consider a lattice Λ and magnify it into the sublattice $n\Lambda$, there will exist *intermediary* lattices Λ' s.t. $n\Lambda \subseteq \Lambda' \subseteq \Lambda$. In that case the larger torus $\mathbb{C}/n\Lambda$ projects onto the smaller one $\mathbb{C}/\Lambda'$.

We write $[\Lambda : \Lambda'] = n$. If $\{\omega'_1, \omega'_2\}$ and $\{\omega_1, \omega_2\}$ are respective $\mathbb{Z}$-basis of Λ' and Λ we have

$$\begin{pmatrix} \omega'_2 \\ \omega'_1 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \omega_2 \\ \omega_1 \end{pmatrix}$$

with $\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M(n)$ the set of integral matrices with determinant n.

$SL(2,\mathbb{Z})$ acts on $M(n)$ and decomposes it in orbits. We can choose as representing elements the matrices $\alpha_i = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$ with $ad = n$, $d \geq 1$, $0 \leq b < d$. There are $\nu(n) = \sigma_1(n) = \sum_{d|n} d$ of them and we have

$$M(n) = \bigcup_{i=1}^{i=\nu(n)} SL(2,\mathbb{Z})\,\alpha_i$$

Hecke operators construct the sum of such Λ':

Definition. The Hecke operator $T(n) : \mathcal{L} \rightarrow \mathcal{L}$ is the additive operator associating to any lattice Λ the sum of the lattices Λ' s.t. $[\Lambda : \Lambda'] = n$:

$$\begin{array}{rcl} T(n) : \mathcal{L} & \rightarrow & \mathcal{L} \\ \Lambda & \mapsto & T(n)(\Lambda) = \sum_{[\Lambda:\Lambda']=n} \Lambda' \end{array}$$

We have of course

$$T(n) = \sum_{i=1}^{i=\nu(n)} \alpha_i(\Lambda)$$

It is easy to extend the definition of Hecke operators to modular forms. Let us first consider homogeneous functions $\tilde{f}$ of degree $-k$ on the Λ: $\tilde{f}(\alpha\Lambda) = \alpha^{-k}\tilde{f}(\Lambda)$. We define

$$T_k(n)\left(\tilde{f}(\Lambda)\right) = n^{k-1} \sum_{[\Lambda:\Lambda']=n} \tilde{f}(\Lambda')$$

the factor n^{k-1} coming from homogeneity.

Modular functions $f(\tau)$ are related to $\tilde{f}(\Lambda)$ by

$$f(\tau) = \tilde{f}(\Lambda_\tau) = (\omega_1)^k \tilde{f}(\Lambda).$$

Computations yield for the action of Hecke operators on modular *forms* $f(\tau) \in M_k$, the following explicit formulae:

Proposition. Let $f(\tau) \in M_k$, $f(\tau) = \sum_{n \geq 0} c_n q^n$, be a modular form of weight k. Then $T_k(m)(f(\tau)) \in M_k$, $T_k(m)(f(\tau)) = \sum_{n \geq 0} b_n q^n$ with

$$\begin{cases} b_0 = c_0 \sigma_{k-1}(m) \text{ where } \sigma_j(m) = \displaystyle\sum_{d|m} d^j \\ b_1 = c_m \\ b_n = \displaystyle\sum_{a|(n,m)} a^{k-1} c_{\frac{nm}{a^2}} \text{ for } n > 1 \end{cases}$$

This result shows first that $c_0 = 0 \Rightarrow b_0 = 0$ and therefore if $f(\tau) \in S_k$, $T_k(m)(f(\tau)) \in S_k$. On the other hand, if $m = p$ is prime,

$$\begin{cases} b_0 = c_0 \\ b_1 = c_m \\ b_n = \sum\limits_{a|(n,p)} a^{k-1} c_{\frac{np}{a^2}} = c_{np} \text{ (only the term } a = 1\text{) for } n > 1 \text{ if } p \nmid n \\ b_n = c_{np} + p^{k-1} c_{\frac{n}{p}} \text{ (the terms } a = 1,\ a = p\text{) for } n > 1 \text{ if } p \mid n \end{cases}$$

Hecke theorem. On M_k the $T_k(m)$ constitue a *commutative* algebra $\mathcal{T}_k$ generated by the $T_k(p)$ and we have the product formulae

$$\begin{cases} T_k(p^r)T_k(p) = T_k(p^{r+1}) + p^{k-1}T_k(p^{r-1}) \\ T_k(m)T_k(n) = \sum\limits_{a|(n,m)} a^{k-1} T_k\left(\frac{mn}{a^2}\right) \\ T_k(m)T_k(n) = T_k(mn) \text{ if } (m,n) = 1 \end{cases}$$

But these are precisely the equivalent for operators of the *multiplicative formulae for quadratic Euler products*:

$$a_{p^r}a_p = a_{p^{r+1}} - d_p a_{p^{r-1}}.$$

We can be even more precise when we restrict Hecke operators to the space of *cusp* forms S_k. Let $\tau = \rho + i\sigma$. The measure $\frac{d\rho d\sigma.}{\sigma^2}$ on $\mathcal{H}$ is $SL(2,\mathbb{Z})$-invariant and, if R is a fundamental domain of $SL(2,\mathbb{Z})$,

$$\langle f, g\rangle = \int_R f(\tau)\,\bar{h}(\tau)\,\sigma^k \frac{d\rho d\sigma.}{\sigma^2}$$

is a scalar product, called *Petersson product*, on S_k.

Petersson theorem. On S_k the Hecke operators $T_k(n)$ are self-adjoint for the Petersson scalar product $\langle f, g\rangle$.

Petersson theorem implies that S_k possesses an *orthogonal* basis of *simultaneous* eigenvectors of the Hecke operators $T_k(n)$.

Let $f(\tau) \in S_k$ be such a simultaneous eigenvector.

- For every n, $T_k(n)f(\tau) = \lambda(n)f(\tau)$.
- If $f(\tau) = \sum_{r\geq 1} c_r q^r$ and $T_k(n)f(\tau) = \sum_{r\geq 1} b_r q^r$, we have therefore $b_r = \lambda(n)c_r$ for $r \geq 1$.
- But we have seen that $b_1 = c_n$.
- So $c_n = \lambda(n)c_1$ and $b_r = \lambda(n)c_r = \lambda(n)\lambda(r)c_1$.
- If we normalize $f(\tau)$ by setting $c_1 = 1$, we get $c_n = \lambda(n)$

So, as the $\lambda(n)$ are eigenvalues of the $T_k(n)$, *the multiplicative properties of Hecke operators become shared by the coefficients of the eigen cusp form* $f(\tau)$:

$$\begin{cases} c_{p^r}c_p = c_{p^{r+1}} + p^{k-1}c_{p^{r-1}} \\ c_m c_n = \sum\limits_{a|(n,m)} a^{k-1} c_{\frac{mn}{a^2}} \\ c_m c_n = c_{mn} \text{ if } (m,n) = 1 \end{cases}$$

These multiplicative properties imply immediately that the Dirichlet L-function $L_f(s)$ of f can be expressed by *a second order Euler product*:

$$L_f(s) = \prod_{p \in \mathcal{P}} \frac{1}{1 - \frac{c_p}{p^s} + \frac{1}{p^{1-k+2s}}}$$

which is of the standard form

$$\prod_{p \in \mathcal{P}} \frac{1}{1 - \frac{a_p}{p^s} - \frac{d_p}{p^{2s}}}$$

with $a_p = c_p$ and $d_p = -p^{k-1}$.

$L_f(s)$ converges for $\Re(s) > \frac{k}{2} + 1$ and has a single simple pole at $s = k$.

Atkin and Lehner newforms

This can be generalized to $\Gamma_0(N)$ with some technicalities solved by Atkin and Lehner with the concept of *newform*.

Among the cusp forms of level N, some come from a cusp form of sublevel N/r. They are called "old" forms.

$S_k(N)$ is the orthogonal sum of the subspaces of old and new (i.e. non old) forms: $S_k(N) = S_k^{\mathrm{old}}(N) \oplus S_k^{\mathrm{new}}(N)$.

If $f(\tau) \in S_k^{\mathrm{new}}(N)$ is a *new* form, everything is fine: $f(\tau)$ possesses at the same time an Euler product and a functional equation.

More precisely, if $f(\tau) \in S_k(N)$ we can associate to it by Mellin transform a Dirichlet L-function $L_f(s)$.

But we must be careful since for $N > 1$ the inversion $\tau \rightarrow -\frac{1}{\tau}$ of matrix $\alpha = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ is no longer in $\Gamma_0(N)$.

But we can use the transformation $\tau \rightarrow -\frac{1}{N\tau}$ and the operator

$$w_N(f(\tau)) = N^{-\frac{k}{2}} \tau^{-k} f\left(-\frac{1}{N\tau}\right)$$

wich leaves stable $M_k(N)$ and $S_k(N)$.

As w_N is an *involution*, called Fricke involution, the spaces $M_k(N)$ and $S_k(N)$ split into eigenspaces $M_k^{\pm}(N)$ and $S_k^{\pm}(N)$ of the eigenvalues ± 1.

Then, Hecke theorem can be generalized to eigenvectors of the Fricke involution:

Hecke theorem. If $f(\tau) \in S_k^{\pm}(N)$, its L-function is an entire function and $\Lambda_f(s) = N^{\frac{s}{2}} \frac{1}{(2\pi)^s} \Gamma(s) L_f(s)$ satisfies the functional equation

$$\Lambda_f(s) = \pm(-1)^{k/2} \Lambda_f(k-s)$$

So we have weakened the concept of cusp form in imposing less symmetries, but at the same time we have strengthen it in imposing its vanishing at its cusps and its "parity" relative to Fricke involution w_N.

We want now to generalize also Hecke operators and Euler products. The problem is rather subtle since N has prime factors $p \mid N$ and we cannot control easily the relation between w_N and the Hecke operators $T_k(p)$ for $p \mid N$.

For the expansions at infinity, we find essentially the same formulae as before.

We get also the same multiplicative recurrence formulae for $p \nmid N$, *but for $p \mid N$ we get another formula which is purely multiplicative*:

Proposition. If $p \mid N$, $T_k(p^r) = T_k(p)^r$.

Hence the generalization of Hecke theorem:

Generalized Hecke theorem. On $M_k(N)$ the $T_k(m)$ constitute a commutative algebra $\mathcal{T}_k$ generated by the $T_k(p)$ and

$$\begin{cases} T_k(p^r)T_k(p) = T_k(p^{r+1}) + p^{k-1}T_k(p^{r-1}) \text{ if } p \nmid N \\ T_k(p^r) = T_k(p)^r \text{ if } p \mid N \\ T_k(m)T_k(n) = \sum\limits_{a \mid (n,m)} a^{k-1}T_k\left(\frac{mn}{a^2}\right) \\ T_k(m)T_k(n) = T_k(mn) \text{ if } (m,n) = 1 \end{cases}$$

Petersson's theorem can also be generalized, the scalar product being defined now by integration on a fundamental domain R_N of $\Gamma_0(N)$.

Petersson theorem. The Hecke operator $T_k(n)$ is self-adjoint on $S_k\,(N)$ if $(n, N) = 1$.

Let $f\,(\tau) \in S_k\,(N)$ be a cusp form of weight k and level N which is a commun eigenvector of *all* the $T_k(n)$. Due to Hecke theorem, its Dirichlet L-function $L_f(s)$ can be expressed by a *second order* Euler product but with a *first order* part corresponding to the $p \mid N$:

$$L_f(s) = \prod_{\substack{p\in\mathcal{P}\\ p|N}} \frac{1}{1-\frac{c_p}{p^s}} \prod_{\substack{p\in\mathcal{P}\\ p\nmid N}} \frac{1}{1-\frac{c_p}{p^s}+\frac{1}{p^{1-k+2s}}}$$

$L_f(s)$ converges for $\Re(s) > \frac{k}{2}+1$ and has a single simple pole at $s = k$.

As we have already noticed, the main difficulty encountered with these generalizations of the case $SL(2,\mathbb{Z})$ to the case $\Gamma_0(N)$ concerns the control of the relation between w_N and $T_k(p)$ when $p \mid N$.

It is this problem which has been solved by Atkin and Lehner using the concept of *newform*.

The Eichler-Shimura construction

Let us come back to the Eichler-Shimura construction presented above.

One shows first that there exists a basis f_j of $S_2(N)$ in which the Hecke operators $T_2(n)$ are represented by *integral* matrices.

The pairs (Λ, C) of a lattice Λ and a cyclic subgroup C of order N being classified by the modular curve $X_0(N)$, Hecke operators, which act on the (Λ, C), act also on the divisor group $\mathrm{Div}(X_0(N))$ of $X_0(N)$, which is the group of the algebraic $\mathbb{Z}$-sums of points of $X_0(N)$.

Let $M(n, N)$ be the set of integral matrices with determinant n, $\alpha = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in M(n)$, satisfying not only the level condition $c \equiv 0 \bmod N$ but also the condition $(a, N) = 1$.

A simple set of α_i are the matrices $\alpha_i = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$ with $ad = n$, $d \geq 1$, $(a, N) = 1$, $0 \leq b < d$. If $M(n, N) = \bigcup_{i=1}^{i=r} \Gamma_0(N)\alpha_i$ where the α_i are representants of the orbits of $\Gamma_0(N)$ acting on $M(n, N)$, and if $[\tau]$ is the class of $\tau \in \mathcal{H}$ in $X_0(N)$, then

$$T_2(n)(\tau) = \sum_{i=1}^{i=r} [\alpha_i \tau] \in \mathrm{Div}(X_0(N))$$

More explicitely we have (see Rohrlich) the formulae for $T_2(p)$

$$T_2(p)(\tau) = \sum_{\nu=0}^{\nu=p-1} \frac{\tau+\nu}{p} + p\tau \text{ if } p \nmid N$$

$$T_2(p)(\tau) = \sum_{\nu=0}^{\nu=p-1} \frac{\tau+\nu}{p} \text{ if } p \mid N$$

Hecke operators $T_2(n)$ act also on the first integral homology group $H_1(X_0(N), \mathbb{Z})$ of the modular curve $X_0(N)$.

Indeed, let c be a loop on $X_0(N)$ and let $(\tau_0, \gamma(\tau_0))$ be a path lifting c in $\overline{\mathcal{H}}$ (τ_0 being a base point in $\mathcal{H}$ and $\gamma \in \Gamma_0(N)$).

If ω is the differential form $\omega = f(z)dz$ then $\int_{\tau_0}^{\gamma(\tau_0)} f(z)dz = \int_c \omega = \Phi_f(\gamma)$. It is then natural to define the action of $T_2(n)$ on the path $(\tau_0, \gamma(\tau_0))$ by

$$T_2(n)((\tau_0, \gamma(\tau_0))) = \sum_{i=1}^{i=r} (\tau_0, \gamma_i(\tau_0))$$

γ_i being defined by $\alpha_i \gamma = \gamma_i \alpha_{j(i)}$ in such a way that, $T_2(n)(\omega)$ being the action of $T_2(n)$ on $\omega = f(z)dz$ deduced from its action on $f \in S_2(N)$, we have

$$\int_{T_2(n)(c)} \omega = \int_c T_2(n)(\omega)$$

Using these different actions of Hecke operators, one proves the following theorem:

Theorem. In the bases $(c_1, \ldots, c_g)$, $(\omega_1, \ldots, \omega_g)$, and $(f_1, \ldots, f_g)$, the matrices of Hecke operators $T_2(n)$ are *integral* matrices and there eigenvalues are therefore *algebraic integers*.

Now we can define what is called a $\mathbb{Q}$-*structure* on the modular curve $X_0(N)$.

We have seen above that the field $K(X_0(N))$ of meromorphic functions on $X_0(N)$ is generated by the modular function j and its N-level transform $j_N = j \circ \begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix}$: $K(X_0(N)) = \mathbb{C}(j, j_N)$.

The rational $\mathbb{Q}$-structure is defined by the subfield $\mathbb{Q}(j, j_N)$ of $\mathbb{C}(j, j_N)$. A meromorphic function $f(\tau)$ belongs to $\mathbb{Q}(j, j_N)$ iff its Laurent series $\sum_{n \geq -M} c_n \kappa^n$ has all its coefficients c_n *rational*.

Let us now extend the action of Hecke operators to the Jacobian $J_0(N)$ of $X_0(N)$. We have the composed map

$$\begin{tikzcd}
\overline{\mathcal{H}} \arrow[r, "\pi"] \arrow[rr, bend right, "\tilde{\Theta}"'] & X_0(N) \arrow[r, "\Theta"] & J_0(N)
\end{tikzcd}$$

$$\tilde{\Theta}(\tau) = \left\{ \int_{\tau_0}^{\tau} f_j(z)\, dz \right\}_{j=1,\ldots,g}$$

If $\{z_j\}_{j=1,\ldots,g}$ are therefore local complex coordinates on the Jacobian $J_0(N)$, the inverse images by $\tilde{\Theta}$ of their differentials dz_j are the 1-forms on $\overline{\mathcal{H}}$ $\tilde{\Theta}^*(dz_j) = \omega_j = f_j(\tau)\, d\tau$.

But since $J_0(N)$ is an abelian group, the map Θ can be extended by linearity to the divisor group $\mathrm{Div}(X_0(N))$ and as the Hecke operators $T_2(n)$ act on this divisor group, they act also, by composition with Θ, on $J_0(N)$ via $T_2(n)^* = \Theta \circ T_2(n)$

$$T_2(n)^*(\tau) = \left\{ \sum_{i=1}^{i=r} \int_{\tau_0}^{\alpha_i \tau} f_j(z)\, dz \right\}_{j=1,\ldots,g}$$

This action enables to define an Hecke *endomorphism* $\mathbf{t}(n)$ of the Jacobian $J_0(N)$ through the following commutative diagram:

$$\begin{tikzcd}
X_0(N) \arrow[r, "T_2(n)"] \arrow[d, "\Theta"] \arrow[dr, "T_2(n)^*"] & \mathrm{Div}(X_0(N)) \arrow[d, "\Theta"] \\
J_0(N) \arrow[r, Rightarrow, "\mathbf{t}(n)"'] & J_0(N)
\end{tikzcd}$$

The Hecke endomorphism $\mathbf{t}(n)$ exists because the mapping $\Theta : X_0(N) \to J_0(N)$ of the compact Riemann surface $X_0(N)$ onto its Jacobian $J_0(N)$ is *universal*.

This means that if $F : X_0(N) \to V$ is a morphism of $X_0(N)$ on an abelian variety V then $F(\tau) = f \circ \Theta(\tau) + F(\tau_0)$ (where τ_0 is a base point of $X_0(N)$) with $f : J_0(N) \to V$ a morphism of abelian varieties. So

$$T_2(n)^*(\tau) = \mathbf{t}(n) \circ \Theta(\tau) + T_2(n)^*(\tau_0).$$

Proposition. All these constructions are simultaneously defined over $\mathbb{Q}$ in a compatible way.

Let then A be the abelian submanifold of $J_0(N)$ defined as the sum of the images of the $\mathbf{t}(p)$

$$A = \sum_{\text{almost all } p} \mathrm{Im}\left(\mathbf{t}(p) - c_p\right)$$

Theorem of Eichler-Shimura-Igusa. Let $f \in S_2^{\mathrm{new}}(N)$ be a new cusp eigenform of $S_2\left(N\right)$ and let $f = \sum_{n \geq 1} c_n \kappa^n$ be its Fourier expansion at infinity. There exists an elliptic curve E and a projection $F : X_0\left(N\right) \rightarrow E$ (an hyperbolic parametrization) s.t.

1. E is defined over $\mathbb{Q}$ and E is the quotient of the Jacobian $J_0\left(N\right)$ by the subgroup A.
2. The Hecke operators $\mathbf{t}(n) \in \mathrm{End}\left(J_0\left(N\right)\right)$ leave A stable and act on A by multiplication by the coefficient c_n.
3. The differential 1-form ω_f on $X_0\left(N\right)$ associated with f is a multiple of $F^*\left(\Omega\right)$ where Ω is the invariant differential on E.
4. The set $\Lambda_f = \left\{ \Phi_f\left(\gamma\right) = \int_{\tau_0}^{\gamma(\tau_0)} f(z)dz \mid \gamma \in \Gamma_0(N) \right\}$ is a lattice in $\mathbb{C}$ and $E \simeq \mathbb{C}/\Lambda_f$.
5. The coefficients c_p are equal to the $a_p = p + 1 - \#E_p\left(\mathbb{F}_p\right)$ and therefore $L_E(s) = L_f(s)$.

As E is an abelian variety, the fact that it is an elliptic curve comes from the fact that it is of dimension 1.

This is proved using the fact that, since E is by definition the quotient $J_0(N)/A$, holomorphic differential 1-forms ω on E are lifted into holomorphic differential 1-forms ω^* on $J_0(N)$ s.t. $\mathbf{t}(p)^*(\omega^*) = c_p\omega^*$.

These correspond to cusp forms $f' \in S_2(N)$ which, for almost all p, are eigenfunctions of the Hecke operators $T_2(p)$ with eigenvalue c_p.

But as f is a *new* form, the space of such f' *is of dimension* 1 and therefore there is essentially one and only one holomorphic differential 1 -forms ω on E and $\dim(E) = 1$.

If $f \in S_2(N)$, $f = \sum_{j=1}^{j=g} r_j f_j$ the f_j corresponding to the dz_j, we have with the above notations:

$$\begin{aligned}\omega_f(u_k) &= \sum_{j=1}^{j=g} r_j dz_j(u_k) = \sum_{j=1}^{j=g} r_j (u_k)_j \\ &= \sum_{j=1}^{j=g} r_j \left(\int_{c_k} f_j(z)\, dz \right) = \int_{c_k} \left(\sum_{j=1}^{j=g} r_j f_j(z) \right) dz = \int_{c_k} f(z) dz\end{aligned}$$

and therefore $\omega_f(\Lambda(X_0(N))) = \Lambda_f$ and Λ_f is a lattice in $\mathbb{C}$ and $E \simeq \mathbb{C}/\Lambda_f$.

For the proof of the matching of the L-functions $L_E(s) = L_f(s)$ one uses the *Frobenius morphism* φ. We will return on that below.

φ is the morphism of degree p which fixes the points of E modulo p, that is the points of $E_p(\mathbb{F}_p)$. We have the equivalence

$$x \in E_p(\mathbb{F}_p) \Leftrightarrow x \in \mathrm{Ker}(1-\varphi)$$

and, as $\# \, \mathrm{Ker}(1-\varphi) = \deg(1-\varphi)$,

$$N_p = \# E_p(\mathbb{F}_p) = \deg(1-\varphi) \ .$$

Now, one can show that there exists a *dual* morphism $\widehat{\varphi}$ s.t. $(1-\widehat{\varphi}) \circ (1-\varphi) = \deg(1-\varphi)\, \mathrm{Id}$ of E and $\widehat{\varphi} \circ \varphi = p\, \mathrm{Id}$.

So, if we compute into the ring $\mathrm{End}\,(E_p)$ of endomorphisms of E reduced modulo p, we find:

$$\begin{aligned}\#E_p\,(\mathbb{F}_p)\,\mathrm{Id} &= \deg\,(1-\varphi)\,\mathrm{Id} = (1-\widehat{\varphi})\circ(1-\varphi)\\ &= 1-(\varphi+\widehat{\varphi})+\widehat{\varphi}\circ\varphi = 1-(\varphi+\widehat{\varphi})+p\,\mathrm{Id}\end{aligned}$$

This shows that in $\mathrm{End}\,(E_p)$

$$\varphi+\widehat{\varphi} = (p+1-\#E_p\,(\mathbb{F}_p))\,\mathrm{Id} = a_p\,\mathrm{Id}$$

But it can be proved that

Proposition. Modulo p, the Hecke operator $T_2(p)$ acts on $X_0\,(N)$ as $\varphi+\widehat{\varphi}$.

And, as the action of $T_2(p)$ is also the multiplication by c_p, we get finally $a_p = c_p$ and the desired identity $L_E(s) = L_f(s)$.

The core of the proof

Modularity is the core of the proof because it is an "extraordinaire carrefour" between many theories. It is an "holistic" concept. In his *Panorama des Mathématiques pures. Le choix bourbachique* [18], Jean Dieudonné gives specifically the example of modular forms:

> *"La théories des formes automorphes et des formes modulaires est devenue un extraordinaire carrefour où viennent réagir les unes sur les autres les théories les plus variées : Géométrie analytique, Géométrie algébrique, Algèbre homologique, Analyse harmonique non commutative et Théorie des nombres."*

As all creative mathematicians, Jean Dieudonné was convinced that the mathematical interest of a proof depends upon its capacity of *circulating* between many *heterogeneous* theories and of *translating* some parts of theories into completely different other ones.

Two classes of L-functions

We met *two classes* of L-functions, those L_E associated to elliptic curves and those L_f associated to cusp modular forms. In the case of modular elliptic curves, the two L-functions are equal (Eichler-Shimura).

The Taniyama-Shimura-Weil conjecture that every elliptic curve over $\mathbb{Q}$ (i.e. "arithmetic") is modular says therefore that the two classes are identical. It is a conjecture on the equivalence between two completely different ways of constructing objects of a certain type (L-functions).

Its deepness has been very well formulated by Anthony Knapp [32] who explained that XXth century mathematics discovered

> *"a remarkable connection between automorphy and arithmetic algebraic geometry. "*

"This connection first shows up in the coincidence of L-functions that arise from some very special modular forms ('automorphic' L-functions) with L-functions that arise from number theory ('arithmetic' or 'geometric' L-functions, also called 'motivic')."

"The automorphic L-functions have manageable analytic properties, while the arithmetic L-functions encode subtle number-theoretic information. The fact that the arithmetic L-functions are automorphic enables one to bring a great deal of mathematics to bear on extracting the number-theoretic information from the L-function."

Ram Murty [41] also emphasized the point:

> *"In its comprehensive form, an identity between an automorphic L-function and a 'motivic' L-function is called a reciprocity law. (. . .) The conjecture of Shimura-Taniyama (...) is certainly the most intringuing reciprocity law of our time. The 'Himalayan peaks' that hold the secrets of this non abelian reciprocity law challenge humanity."*

TSW and Ribet theorem

The *Taniyama-Shimura-Weil conjecture (TSW)* (conjectured by Yutaka Taniyama in 1955 and formulated precisely by Goro Shimura in the early 1960s) says that every EC is isogenous (that is a covering of finite degree) with a modular EC coming from an $X_0(N)$ and a $f \in S_2^{\mathrm{new}}(N)$ by the Eichler-Shimura construction.

Shimura proved himself in 1971 that his conjecture is true for elliptic curves with *complex multiplication* (there exists a complex number $\alpha \notin \mathbb{Z}$ s.t. $\alpha\Lambda \subset \Lambda$).

A result due to Carayol says that the level N must be equal to the *conductor* N_E of E.

TSW conjecture is equivalent to another celebrated conjecture:

Hasse-Weil conjecture. The L-functions $L_E(s)$ of elliptic curves share the same automorphy properties as the L-functions $L_f(s)$.

Theorem. *TWS* conjecture and the Hasse-Weil conjecture are equivalent.

The implication $TWS \longrightarrow HW$ is easy since if two ECs defined over $\mathbb{Q}$ are isogenous over $\mathbb{Q}$ then there L -functions are equal. So E is isogenous to E' with $L_{E'}(s) = L_f(s)$ for a certain $f \in S_2^{\mathrm{new}}(N)$ and $L_E(s) = L_{E'}(s) = L_f(s)$.

The implication $HW \to TSW$ is less evident. HW implies that $\exists f \in S_2^{\mathrm{new}}(N)$ with $L_E(s) = L_f(s)$. The Eichler-Shimura construction associates to f a modular elliptic curve E' with $L_{E'}(s) = L_f(s)$. So, $L_E(s) = L_{E'}(s)$ and we can conclude using a theorem of Faltings:

Theorem (Faltings). If $L_E(s) = L_{E'}(s)$ then E and E' are isogenous over $\mathbb{Q}$.

Theorem. **TSW implies FLT.**

Let $a^l + b^l + c^l = 0$ be an hypothetic solution of Fermat theorem (prime $l \geq 5$ and a, b, c relatively prime). We consider the associated Frey elliptic curve E of equation

$$y^2 = x\left(x - a^l\right)\left(x + c^l\right)$$

We know that the discriminant is $\Delta = 16(abc)^{2l}$ and that the conductor is $N = \prod_{p|abc} p$ due to semi-simplicity. Ribet proved that these values *forbid E to be modular*.

In [46] (p.16), Ribet gave the following conceptual description of Frey's strategy:

> *"From Frey's point of view, the main "unexpected" property of E is that Δ [the minimal discriminant] is a product of a power of 2 and a perfect l th power, where l is a prime ≥ 5. Frey translated this property into a statement about the Néron model for E: if p is an odd prime at which E has bad reduction, the number of components in the mod p reduction of the Néron model is divisible by l. Frey's idea was to compare this number to the corresponding number for the Jacobian of the modular curve $X_0(N)$, where N is the conductor of E. Frey predicted that a discrepancy between the two numbers would preclude E from being modular. In other words, Frey concluded heuristically that the existence of E was incompatible with the Taniyama-Shimura conjecture, which asserts that all elliptic curves over $\mathbb{Q}$ are modular."*

Ribet theorem is a *descent* result. The idea is to show that the level N can be reduced to the case $N = 2$ and then to use the

Lemma. $S_2(2) = 0$.

Indeed, in the $N = 2$ case, the result of Barry Mazur on the genus $g(N)$ of $X_0(N)$ says that the modular curve $X_0(2)$ is of genus $g = 0$ (it is topologically a sphere) and there exist therefore *no* non trivial holomorphic differential ω on $X_0(2)$ (the differential dz has a pole at infinity). As an $f \in S_2(2)$ corresponds to an ω, $S_2(2) = 0$.

The fact that $S_2(2) = 0$ shows that a parametrization associated to a modular form f cannot exist. The *reduction to level* 2 is a consequent of a theorem of Ribet.

Ribet theorem (*Serre ε-conjecture*). Let E be an elliptic curve defined over $\mathbb{Q}$ having discriminant Δ with prime decomposition $\Delta = \prod_{p|\Delta} p^{\delta_p}$ and conductor $N = \prod_{p|\Delta} p^{f_p}$. If E is a modular EC of level N associated to a cusp form $f \in S_2(N)$, if l is a prime dividing the power δ_p of p in Δ and if $f_p = 1$ (that is if $p \parallel N$ in the sense $p \mid N$ but $p^2 \nmid N$) then *modulo l* the modular parametrization can be reduced to level $N' = N/p$ mod l in the sense that there exists a cusp form $f' \in S_2(N')$ s.t. the coefficients of f and f' are equal modulo l: $c_n \equiv c'_n \; l \; \forall n \geq 1$.

Let us apply Ribet theorem to the Frey curve.

We know that $\Delta = 16a^{2l}b^{2l}c^{2l}$. As a, b, c are relatively primes, for $p \neq 2$, if $p \mid \Delta$, we have $2l \mid \delta_p$, hence $l \mid \delta_p$, and $f_p = 1$ and we can apply the theorem.

For $p = 2$ the situation is different since $4 + 2l \mid \delta_2$ and therefore $l \nmid \delta_2$ (if $\delta_2 = lm$ and $4 + 2l = n\delta_2$, then $4 + 2l = nlm$ and $l \mid 4$, but l is odd) and the reduction of levels leads to $N' = 2$.

So there exists $f' \in S_2(2)$ such that $c_n \equiv c'_n \mod l \ \forall n \geq 1$. We then apply the lemma $S_2(2) = 0$.

So under an incredibly complicated travel inside the Himalayan unity of mathematics and the *TSW* conjecture, the proof of Fermat theorem boils down to the *topological obstruction* that a torus of genus 1 cannot be parametrized by a sphere of genus 0.

An incredibly complex *arithmetic* impossibility is translated into a trivial *topological* impossibilty.

Wiles' travel

In his reference paper [73] summarizing the story of his proof, Wiles says

> *"I began working on these problems in the late summer of 1986 immediately on learning of Ribet's result."*

To prove *TSW*, he used deep works of Jean-Pierre Serre and Barry Mazur on a specific class of objects called *Galois representations* naturally associated to ECs and introduced in the 1940–1950's by André Weil and John Tate.

We meet here another extraordinary example of encoding informations of a theory into another theory. The arithmetic informations we will focus on are associated to *torsion points* (also called "division" points) of ECs.

Encoding information into Galois representations

This encoding is particularly interesting for the following reason. Until now we met two definitions of modular ECs defined over $\mathbb{Q}$:

- a *geometric* definition: they are quotients of modular curves $X_0(N)$,
- an *analytic* definition: they are associated to modular forms f.

But as was emphasized by Charles Daney ([11], p.24), these two definitions respectively geometric and analytic are *difficult* to use.

> *"The difficulty, perhaps, lies in the disparity between the essentially analytic nature of the properties and the algebraic nature of an elliptic curve and the kind of problems to which we want to apply the theory. (...) We seem to need some more algebraic formulation of what it means for an elliptic curve to be modular."*

It is here that Galois representations enter the stage.

Let E be an elliptic curve identified with its Jacobian J, which is a complex torus $\mathbb{C}/\Lambda$. The torsion points of order N of $E(\mathbb{C})$ correspond to those of the smaller lattice $\frac{1}{N}\Lambda$, that is those satisfying $Nx = 0$. Their set T_N is trivially isomorphic to $\frac{\mathbb{Z}}{N\mathbb{Z}} \times \frac{\mathbb{Z}}{N\mathbb{Z}}$.

So, the torsion points of $E(\mathbb{C})$ constitute a group $E[N](\mathbb{C}) \simeq \frac{\mathbb{Z}}{N\mathbb{Z}} \times \frac{\mathbb{Z}}{N\mathbb{Z}}$. If $\{\omega_1, \omega_2\}$ is a $\mathbb{Z}$-basis of Λ, $\{\omega_1/N, \omega_2/N\}$ is a $\mathbb{Z}$-basis of Λ/N and if x_i corresponds to ω_i/N by the isomorphism of E with its Jacobian, $\{x_1, x_2\}$ is a $\mathbb{Z}$-basis of $E[N](\mathbb{C})$.

If x is a p^m-division point of E and if $n > m$ then, a fortiori x is a p^n-division point. The $E[p^n]$ constitute a projective system whose projective limit $E[p^\infty]$ is called the p-adic *Tate module* of E.

Suppose now that E is “arithmetic” (i.e. defined over $\mathbb{Q}$). Then the N-torsion points are *algebraic* over $\mathbb{Q}$ (look at the formulae of division on E) and $E[N](\mathbb{C}) = E[N](\overline{\mathbb{Q}})$.

It is natural to look at *rational* N-torsion points, that is at $E[N](\mathbb{Q})$. The structure of their subgroup has been clarified by Lutz and Nagel in the 1930s. Long after, Barry Mazur proved a beautiful theorem giving their exhaustive list.

Mazur theorem. The only groups which appear as rational torsion groups of ECs defined over $\mathbb{Q}$ are:

1. $\mathbb{Z}/N\mathbb{Z}$ for $N = 1, 2, \ldots, 10, 12$.
2. $\mathbb{Z}/2N\mathbb{Z} \oplus \mathbb{Z}/2\mathbb{Z}$ for $N = 1, \ldots, 4$.

But between $\mathbb{Q}$ and $\overline{\mathbb{Q}}$ there is the whole universe of algebraic number fields!

As the N-torsion points are $\overline{\mathbb{Q}}$-points, we can consider the extension $\mathbb{Q}(E[N])$ of $\mathbb{Q}$ defined by the adjunction of their coordinates.

It can be shown that $\mathbb{Q}(E[N])$ is an algebraic Galois extension of $\mathbb{Q}$ and we can consider the way the elements $\sigma \in \mathrm{Gal}(\mathbb{Q}(E[N])/\mathbb{Q})$ act on $\mathbb{Q}(E[N])$.

In the $\mathbb{Z}$-basis $\{x_1, x_2\}$ of $E[N]$, any such automorphism σ of $\mathbb{Q}(E[N])$ over $\mathbb{Q}$ is represented by a 2×2 matrix and we get therefore a representation, called a *Galois representation*,

$$\overline{\rho}_{E,N} : G = \mathrm{Gal}(\mathbb{Q}(E[N])/\mathbb{Q}) \to GL_2\left(\frac{\mathbb{Z}}{N\mathbb{Z}}\right) .$$

This representation is *injective* (one-to-one) and makes $\mathrm{Gal}\left(\mathbb{Q}\left(E\left[N\right]\right)/\mathbb{Q}\right)$ a *subgroup* of $GL_2\left(\frac{\mathbb{Z}}{N\mathbb{Z}}\right)$. Indeed let g s.t. $\overline{\rho}_{E,N}\left(g\right)=\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, then g leaves invariant the $\mathbb{Z}$-basis $\{x_1, x_2\}$ of $E\left[N\right]$ and therefore $g=\mathrm{Id}$.

More generally, if K is an extension of $\mathbb{Q}$ containing $\mathbb{Q}(E[N])$, we get a representation $\overline{\rho}_{E,N} : \mathrm{Gal}(K/\mathbb{Q}) \to GL_2\left(\frac{\mathbb{Z}}{N\mathbb{Z}}\right)$. In particular,

- for $K = \overline{\mathbb{Q}}$ we get a Galois representation

$$\overline{\rho}_{E,N} : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2\left(\frac{\mathbb{Z}}{N\mathbb{Z}}\right)$$

- and in the case where $N = p$ is a prime, we get a Galois representation

$$\overline{\rho}_{E,p} : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2\left(\mathbb{F}_p\right)$$

of the "absolute" Galois group $\mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right)$.

This representation is called "*continuous*" in the sense it factorizes through the Galois group $\mathrm{Gal}(K/\mathbb{Q})$ of a *finite* algebraic Galois extension $K/\mathbb{Q}$.

$\overline{\rho}_{E,p}$ can be reducible. E.g., it is trivial if the p-torsion points are rational.

To make things more concrete, let us take the simplest case $p = 2$.

We have $E[2] = \{(0,\infty),(\alpha_1,0),(\alpha_2,0),(\alpha_3,0)\}$ where the α_i are the 3 roots of the cubic polynomial $f(x)$ in the equation $y^2 = f(x)$ of E and G permutes these roots.

The group $GL_2\left(\frac{\mathbb{Z}}{2\mathbb{Z}}\right)$ is isomorphic with the group S_3 of permutations on 3 elements a, b, c and the image of G by $\rho_{E,2}$ in $GL_2\left(\frac{\mathbb{Z}}{2\mathbb{Z}}\right) \simeq S_3$ is isomorphic to $\mathrm{Gal}(K/\mathbb{Q})$ where K is the splitting field of the polynomial $f(x)$.

In his 1972 paper "Propriétés galoisiennes des points d'ordre fini des courbes elliptiques" dedicated to André Weil [53] , Jean-Pierre Serre explains that

> *"Il s'agit de prouver que les groupes de Galois associés aux points d'ordre fini des courbes elliptiques sont 'aussi gros que possible.'"*

Theorem (Serre). The index of the image $\overline{\rho}_{E,N}(G)$ of G is bounded in $GL_2\left(\frac{\mathbb{Z}}{N\mathbb{Z}}\right)$ by a constant depending only on E.

Let $E[\infty] = \bigcup_{N \in \mathbb{N}} E[N]$ be the subgroup of *all* torsion points in $E\left(\overline{\mathbb{Q}}\right)$ and consider the automorphism group

$$\varprojlim GL_2\left(\frac{\mathbb{Z}}{N\mathbb{Z}}\right) = GL_2\left(\varprojlim \frac{\mathbb{Z}}{N\mathbb{Z}}\right) = GL_2\left(\widehat{\mathbb{Z}}\right)$$

Let $\overline{\rho}_{E,\infty} : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2\left(\widehat{\mathbb{Z}}\right)$ be the limit of the $\overline{\rho}_{E,N}$ then

Theorem. The index of the image $\overline{\rho}_{E,\infty}(G)$ of G in $GL_2\left(\widehat{\mathbb{Z}}\right)$ is *finite.*

These results can be formulated in the p-adic framework.

Indeed $E[\infty] = \bigcup_{N \in \mathbb{N}} E[N] = \bigoplus_{p \text{ prime}} E[p^\infty]$ with $E[p^\infty]$ the p-adic Tate module, and $GL_2\left(\widehat{\mathbb{Z}}\right) = \mathrm{Aut}(E[\infty])$ is a product of factors corresponding to the different primes:

$$GL_2\left(\widehat{\mathbb{Z}}\right) = \mathrm{Aut}(E[\infty]) = \prod_{p \text{ prime}} \mathrm{Aut}(E[p^\infty]) \simeq \prod_{p \text{ prime}} GL_2(\mathbb{Z}_p)$$

and the representation $\overline{\rho}_{E,\infty} : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2\left(\widehat{\mathbb{Z}}\right)$ is a "product" of $\overline{\rho}_{E,p^\infty} : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2(\mathbb{Z}_p)$.

The representations $\overline{\rho}_{E,\infty}$ encode a lot of informations on the elliptic curve E. For instance, $\overline{\rho}_{E,\infty}$ and $\overline{\rho}_{E',\infty}$ are isomorphic iff E and E' are *isogenous*.

Serre proved the theorem:

Theorem (Serre). For almost every prime p, $\overline{\rho}_{E,p^\infty}$ is *surjective*: $\overline{\rho}_{E,p^\infty}(G) = GL_2(\mathbb{Z}_p)$.

The main obstruction to the surjectivity of $\overline{\rho}_{E,p^\infty}$ is the existence of a $\mathbb{Q}$-rational point of order p.

So, we can say that the $\overline{\rho}_{E,p} : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \rightarrow GL_2(\mathbb{F}_p)$ are generically (i.e. for almost all p) surjective, and therefore *isomorphisms* $\mathrm{Gal}(K/\mathbb{Q}) \rightarrow GL_2(\mathbb{F}_p)$.

The Frobenius

In order to go further, we need *Frobenius* morphisms.

For the algebraic extensions $\mathbb{F}_{q^n}$ of $\mathbb{F}_q$ and the algebraic closure $\overline{\mathbb{F}}_q$ of $\mathbb{F}_q$ the Frobenius is defined as $\overline{\mathrm{Frob}}_q : x \to x^q$.

- It is the generator of the Galois group $\mathrm{Gal}\left(\mathbb{F}_{q^n} \text{ or } \overline{\mathbb{F}}_q/\mathbb{F}_q\right)$.
- As, due to Fermat little theorem, $x^q = x$ for every $x \in \mathbb{F}_q$, it is the identity on $\mathbb{F}_q$.
- On $\mathbb{F}_{q^n}$ it is a $\mathbb{F}_q$-automorphism of order n, i.e. $\mathrm{Gal}\left(\mathbb{F}_{q^n}/\mathbb{F}_q\right)$ is a cyclic group of order n.

A key fact is that it can be *lifted* to a Frobenius Frob_q in the Galois group $\mathrm{Gal}\left(K/\mathbb{Q}\right)$ of any Galois extension $K/\mathbb{Q}$ where q is unramified.

Indeed, let $K/\mathbb{Q}$ be a Galois extension and $\mathcal{O}_K$ the ring of integers of K. The prime ideals $\mathfrak{q}$ of $\mathcal{O}_K$ s.t. $q \in \mathfrak{q}$ (i.e. $(q) \subset \mathfrak{q}$, i.e. $\mathfrak{q} \mid (q)$) are conjugated by the Galois group $\mathrm{Gal}\,(K/\mathbb{Q})$.

If q is *unramified* in $K/\mathbb{Q}$ then its inertia group $I_{\mathfrak{q}} = \{1\}$ is trivial, its decomposition group $D_{\mathfrak{q}}$ is isomorphic to $\mathrm{Gal}\,(F_{\mathfrak{q}}/\mathbb{F}_q)$ and the unique $\sigma_{\mathfrak{q}} \in D_{\mathfrak{q}} \simeq \mathrm{Gal}\,(F_{\mathfrak{q}}/\mathbb{F}_q)$ associated to $\overline{\mathrm{Frob}}_q$ is written $\mathrm{Frob}_{\mathfrak{q}}$ and is also called the Frobenius.

All the different $\mathrm{Frob}_{\mathfrak{q}}$ for $\mathfrak{q} \mid (q)$ are conjugate and they define up to conjugacy a Frobenius $\mathrm{Frob}_q \in \mathrm{Gal}\,(K/\mathbb{Q})$.

One can generalize the finite degree case to the *infinite* degree case of the algebraic closure $K = \overline{\mathbb{Q}}$, and $F_{\mathfrak{q}} = \overline{\mathbb{F}}_q$. As explained by Kenneth Ribet ([46], p. 12), in that special but very fundamental case,

> *"One can think of $\mathfrak{q}$ as a coherent set of choices of primes lying over q in the rings of integers of all finite extensions of $\mathbb{Q}$ in $\overline{\mathbb{Q}}$."*

One says that $\overline{\rho}_{E,p}$ is *unramified* at q if $\overline{\rho}_{E,p}$ is trivial on the inertia group I_q.

The *conductor* N_p of the representation $\overline{\rho}_{E,p}$ is defined as

$$N_p = \prod_{\substack{q \neq p \\ q \text{ ramified}}} q^{n(\rho,q)}$$

where $n(\rho, q)$ is the *degree of ramification* of $\overline{\rho}_{E,p}$ at the prime $q \neq p$. N_p divides the conductor N_E of E.

An important theorem relates the properties of *ramification* of $\overline{\rho}_{E,p}$ to the properties of *reduction* of E: if $q \neq p$ and $q \nmid N_E$ (good reduction) then $\overline{\rho}_{E,p}$ is unramified at q. Further:

Theorem of Néron, Ogg, Shafarevich. Let $q \neq p$. Then E has good reduction at q iff the representation $\overline{\rho}_{E,p^\infty}$ on the p-adic Tate module is unramified at q. In particular, if $E/\mathbb{Q}$ and $E'/\mathbb{Q}$ are isogenous they have the same primes of good and bad reduction.

Suppose, e.g., that E is *semi-stable* at q, its reduction mod q being a node.

The group of regular points is then the multiplicative group $\mathbb{G}_m$ of $\mathbb{C}$ and the p^n-torsion points are the p^n-th roots of unity.

Their group is of size p^n while $E\left[p^n\right]$ is of size p^{2n}. So a lot of p^n-torsion points are killed by the reduction mod q. Hence the ramification.

Another related result concerns the links between the *reducibility* of $\overline{\rho}_{E,p}$ and the *rationality* of the corresponding point of $X_0(p)$:

Theorem. $\overline{\rho}_{E,p}$ is reducible iff the corresponding point of $X_0(p)$ is rational.

This is due to the fact that rational points of $X_0(p)$ correspond to curves whose p-division points are *rational* and on which $\mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right)$ act therefore trivially.

Galois representations and L_E-functions

The consideration of the Galois representations $\overline{\rho}_{E,p}$ of $G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right)$ is relevant because they have deep links with L-functions.

It is due to the remarkable following theorem.

For $\sigma \in G$, the image $\overline{\rho}_{E,p}(\sigma)$ is a matrix $GL_2(\mathbb{F}_p)$ and this matrix has two invariants belonging to $\mathbb{F}_p$, its *trace* and its *determinant*.

The theorem shows in particular that the different $\overline{\rho}_{E,p}$ encode the *counting* of points of E over the different prime fields $\mathbb{F}_q$ with q another prime (beware: we are considering *two* primes p and q).

Theorem. Let E be an EC defined over $\mathbb{Q}$. Its Galois representation $\overline{\rho}_{E,p}$ satisfies the following properties:

1. $\mathrm{Trace}\,\overline{\rho}_{E,p}(\mathrm{Frob}_q) \equiv q + 1 - \#E_q(\mathbb{F}_q)\,(= a_q)$ mod p for almost every prime q (essentially $q \neq p$ and $q \nmid N$). This is the reason why we used a_q instead of $\#E_q(\mathbb{F}_q)$.
2. $\mathrm{Det}\,\overline{\rho}_{E,p} = \overline{\varepsilon}_p$ where $\overline{\varepsilon}_p : G \to \mathbb{F}_p^\times$ is the cyclotomic character giving the action of G on the p-th roots of unity, and in particular $\mathrm{Det}\,\overline{\rho}_{E,p}(\mathrm{Frob}_q) \equiv q \pmod p$.
3. $\mathrm{Det}\,\overline{\rho}_{E,p}(c) = \overline{\varepsilon}_p(c) = -1$ (i.e. $\overline{\rho}_{E,p}$ is odd) since the complex conjugation c acts on a pth root of unity ζ by $\zeta \mapsto \zeta^{-1}$. (Complex conjugation can be interpreted as Frob_∞, the Frobenius of the "infinite" prime corresponding to $\mathbb{R}$.)

This theorem remains valid (with $=$ and no longer $\equiv$) for the p-adic limit $\rho_{E,p} : G \to GL_2(\mathbb{Z}_p)$, that is when we lift the *residual* situation at p to the p-adic *local* situation at p.

Serre's conjecture

The conjecture which is the equivalent to the *TSW* conjecture for Galois representations is due to Jean-Pierre Serre and says essentially that every Galois representation

$$\overline{\rho} : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2\left(\mathbb{F}_p\right)$$

coming from the torsion points of an elliptic curve is modular.

The representations $\overline{\rho}_{E,p} : G \to GL_2\left(\mathbb{F}_p\right)$ are continuous and their character $\mathrm{Det}\,\overline{\rho}_{E,p}$ is odd. It can be shown that they are *absolutely irreducible* in the sense that they are irreducible and $\overline{\rho}_{E,p} \otimes_{\mathbb{F}_p} \overline{\mathbb{F}}_p$ is also irreducible.

Serre conjecture. Let $\overline{\rho} : G \to GL_2\left(\mathbb{F}_p\right)$ be a continuous, and absolutely irreducible Galois representation with $\mathrm{Det}\,\overline{\rho}\,(c) = -1$ (that is $\overline{\rho}$ *can be* modular). Then $\overline{\rho}$ *is* effectively modular: there exist a level $N \geq 1$, a weight $k \geq 2$, a character $\chi : \left(\frac{\mathbb{Z}}{N\mathbb{Z}}\right)^{+} \to \mathbb{C}^{\times}$, and a new cusp form $f \in S_k^{\mathrm{new}}(N, \chi)$ s.t. $\overline{\rho} = \rho_f$.

Serre made precise propositions for the weight k, the conductor N, the *Nebentypus* χ and the newform f.

Theorem. Serre conjecture implies FLT.

The proof is similar to the proof that STW implies Fermat.

Let $a^l + b^l + c^l = 0$ be an hypothetical solution of Fermat theorem for a prime $l \geq 5$ and a, b, c relatively prime non vanishing integers. We consider once again the associated Frey elliptic curve E

$$y^2 = x\left(x - a^l\right)\left(x + b^l\right)$$

and this time we consider the *very particular* Galois representation $\overline{\rho}_{E,l} : G \rightarrow GL_2\left(\mathbb{F}_l\right)$ defined by the points of l-torsion, where l is now *the power in Fermat equation*.

1. $\overline{\rho}_{E,l}$ is continuous since it factorizes through $\mathrm{Gal}\,(K/\mathbb{Q})$ where K is the field generated by (the coordinates of) the l-division points.
2. $\overline{\rho}_{E,l}$ is absolutely irreducible.
3. $\overline{\rho}_{E,l}$ is unramified outside 2 and l and its ramification at l and 2 is, as says Bas Edixhoven, “very well behaved”. Indeed for $\overline{\rho}_{E,l}$ to be ramified at $q \neq 2, l$ we must have (since $q \mid \Delta$) $q \mid abc$. But in that case we get a node (semi-simplicity) with l dividing the exponent $2l$ of q in Δ, and this implies the non ramification.
4. $\overline{\rho}_{E,l}$ can be modular.
5. If Serre conjecture is true, then $\overline{\rho}_{E,l}$ is modular.
6. One shows, it is the difficult part of the proof, that for any f s.t. $\overline{\rho}_{E,l} = \rho_f$ we must have $(N, k, \chi) = (2, 2, 1)$.
7. One concludes with the same argument as before: $S_2(2, 1) = 0$ since $X_0(2)$ is of genus $g = 0$.

Step 6 uses a theorem due to Barry Mazur and an adaptation of Ribet theorem which say essentially that

- we can choose as conductor N the Artin conductor of $\overline{\rho}$, and
- for a $\overline{\rho}$ coming from a Frey curve, this Artin conductor is minimal and equal to 2.

As was emphasized by Yves Hellegouarch ([29], p.329):

> *"La 'philosophie' qui rend ces conjectures si précieuses tient au fait que la représentation ρ_f liée à une nouvelle forme f de niveau N peut être beaucoup plus simple que ce que l'on pouvait attendre : en particulier son conducteur d'Artin N_ρ peut être beaucoup plus petit que N. La forme f est alors congrue modulo p à une forme dont le niveau est un très petit diviseur de N, ce qui conduit à des conséquences merveilleuses."*

These arguments (implying that $\overline{\rho}$ is absolutely irreducible, unramified at p and flat at l) give a proof of the implication $STW \Rightarrow$ Fermat.

Of course it is normal for $\overline{\rho}$ to be unramified at the points where E has good reduction.

But in our case, $\overline{\rho}$ is also unramified at $p = l$ and $p \mid N$ with $p > 2$ and this is quite extraordinary. As Gerd Faltings formulates it ([22], p.744):

> *"The l-division points behave as if E had good reduction at all $p > 2$."*

But this is impossible.

In fact Serre conjecture is *stronger* than the *TSW* conjecture. Indeed:

Theorem. Serre conjecture implies *TSW* conjecture.

Sketch of the proof. Let E be of conductor N with Hasse-Weil L-function $L_E(s) = \sum\limits_{n\geq 1} \frac{a_n}{n^s}$.

One shows first that, for almost every prime p, the Galois representations modulo p, $\overline{\rho}_{E,p}$, can be modular.

If Serre conjecture is valid, then they are modular and $\overline{\rho}_{E,p} = \rho_{f_p}$ for a cusp form $f_p \in S_2(N, 1)$ whose coefficients $a_{q,p}$ for $q \nmid N$ are eigenvalues of Hecke operators.

But f_q can be lifted to characteristic 0 to a modular cusp form $F = \sum_{n \geq 1} A_n \kappa^n$ s.t. $\widetilde{F} \equiv f_p \mod p$.

As the weight k and the level N are fixed, there exist only a *finite* number of possible F. There exists therefore an F s.t. $\widetilde{F} = f_p \mod p$ for an *infinite* set P of primes p.

Let $q \nmid N$. Then E has good reduction. Let $a_q = \mathrm{Trace}\left(\mathrm{Frob}_q\right)$. We have $a_q \equiv a_{q,p} \mod p$ for every $q \neq p$ and therefore $A_q = a_q$ in $\overline{\mathbb{F}_p}$ for every $p \in P - \{q\}$ and, as P is infinite, $A_q = a_q$ for every $q \nmid N$.

This shows that $A_q \in \mathbb{Z}$ and that the A_q define a modular curve E_F of level N'/N, E and E_F sharing equivalent q-adic representations.

But, due to Faltings theorem, this implies that E and E_F are isogenous over $\mathbb{Q}$, and E is therefore modular.

A letter of Jean-Pierre Serre to Alexandre Grothendieck

The 31 December 1986, Jean-Pierre Serre wrote a very interesting and touching letter to Alexandre Grothendieck [27] announcing his conjecture. Let us quote it.

" Cher Grothendieck,
"Tu vas recevoir un de ces jours une copie de "Sur les représentations modulaires de degré 2 de $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$", un travail que j'ai rédigé ces derniers mois, mais qui était en fait en chantier depuis une douzaine d'années. (...)

"Tu te souviens sans doute de la conjecture avancée par Weil en 1966 : toute courbe elliptique sur $\mathbb{Q}$ est "modulaire". (...)

Le grand intérêt de cette conjecture est qu'elle décrit comment on peut obtenir les motifs les plus simples qui soient : ceux de dimension 2, de hauteur 1 et de corps de base $\mathbb{Q}$.

En particulier, si la conjecture est vraie (et elle a été vérifiée numériquement dans de très nombreux cas), la fonction zêta du motif a les propriétés analytiques (prolongement et équation fonctionnelle) que l'on pense.

"Plus généralement, toutes les fonctions zêta attachées aux motifs devraient (conjecturalement) provenir de "représentations modulaires" convenables; il y a là-dessus des conjectures assez précises de Langlands et Deligne.

“Ce que j’ai essayé de faire dans le texte que je t’envoie, c’est un *analogue* (modulo p) de la conjecture de Weil en question. On veut décrire en termes de formes modulaires (modulo p) certaines représentations galoisiennes. Ces représentations sont en apparence très spéciales; ce sont des représentations

$$\mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \rightarrow GL_2\left(\mathbb{F}_p\right)$$

irréductibles (sinon ce n’est pas très intéressant) et de déterminant impair (la conjugaison complexe doit avoir un déterminant égal à -1).

La conjecture que je fais est que toutes ces représentations sont “modulaires” , i.e. proviennent de formes modulaires modulo p dont je prédis en outre le niveau et le poids (la recette prédisant le niveau est très naturelle — celle du poids ne l’est pas).

Bien entendu, je ne suis pas du tout sûr que cette conjecture soit vraie! Mais elle est étayée par quantité d'exemples mi-théoriques, mi-numériques, et j'ai fini par me décider à la publier.

D'autant plus que ses applications sont nombreuses :

a) elle entraîne la conjecture de Weil citée au début, ainsi que des conjectures analogues sur des motifs de hauteur > 1 (...); a priori, cela peut paraître surprenant : comment déduire un énoncé de caractéristique 0 d'un énoncé de caractéristique p? C'est beaucoup moins surprenant lorsqu'on se rend compte qu'on a une infinité de p à sa disposition.

b) elle entraîne le (grand) théorème de Fermat, ainsi que des variantes assez surprenantes: non-existence de solutions non triviales de $x^p + y^p + \ell z^p = 0$, $p \geq 11$, pour ℓ premier égal à 3, 5, 7, 11, 17, 19,. . . (mais la méthode ne s'applique pas à $\ell = 31$).

c) elle entraîne que tout schéma en groupes sur $\mathbb{Z}$, plat, fini, de type (p, p) est somme directe (pour $p \geq 3$) de copies de $\mathbb{Z}/p\mathbb{Z}$ et de μ_p. (Attention: il ne s'agit que de schémas de rang 2. Je ne sais rien faire pour un rang plus grand.).

"Bien sûr, on serait un peu plus rassuré si on savait faire une conjecture générale (sur un corps global quelconque, pour des représentations de dimension quelconque). J'y ai souvent réfléchi, mais je ne vois pas comment faire (et cependant je suis sûr que c'est possible, au moins dans certains cas). On verra bien. . .

"Bien à toi — et meilleurs vœux pour 1987.
"J-P. Serre" .

Deligne's theorem

We must define now *modular* Galois representations.

We have encoded a lot of arithmetic informations on ECs in Galois representations $\overline{\rho}_{E,p} : \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \rightarrow GL_2\left(\mathbb{F}_p\right)$.

Now, due to a fundamental work of Pierre Deligne in 1969, one can also associate such Galois representations to *modular forms*.

Hence the strategic idea of proving *TSW* conjecture by proving that the "arithmetic" $\overline{\rho}_{E,p}$ are modular.

Let $S_k(N, \varepsilon)$ be the space of cusp forms of weight k, level N and character (*Nebentypus*) ε.

The Hecke operators $T_k(\ell)$ for ℓ prime (they generate all the $T_k(n)$) act on $S_k(N, \varepsilon)$ and commute between them.

Let $\lambda(n)$ be the eigenvalues of a common *new* eigenform $f = \sum_{n \geq 1} a_n \kappa^n \in S_k^{\mathrm{new}}(N, \varepsilon)$ of the $T_k(n)$, let $\mathcal{O}_f$ be the ring generated by the $\lambda(\ell)$ and the $\varepsilon(\ell)$ and K_f the quotient field.

Let $\sim: \mathcal{O}_f \to \mathbb{F}_p$ be a morphism of $\mathcal{O}_f$ into the finite field $\mathbb{F}_p$.

For p a prime non dividing N, let $\mathfrak{p}$ be a prime ideal of $\mathcal{O}_f$ above p, and $\mathcal{O}_{f,\mathfrak{p}}$ the $\mathfrak{p}$-adic completion of $\mathcal{O}_f$ at $\mathfrak{p}$ (local $\mathfrak{p}$-adic ring).

Deligne theorem. (For $k = 2$, the theorem has been proved before by Eichler and Shimura.) Under these hypotheses there exists a (unique) Galois representation $\rho_f : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2\left(\mathcal{O}_{f,\mathfrak{p}}\right)$ associated with f, which is continuous, semi-simple and unramified for $q \neq p$ and $q \nmid N$ and has the "good" properties for the trace, the determinant and the character.

Namely its quotient $\overline{\rho}_f : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2\left(\mathbb{F}_p\right)$ defined via the map $\sim: \mathcal{O}_f \to \mathbb{F}_p$ satisfies for every prime $q \neq p$ and $q \nmid N$

1. $\overline{\rho_f}$ is unramified at q;
2. $\mathrm{Trace}\,\overline{\rho}_f\left(\mathrm{Frob}_q\right) = \widetilde{a_q}$;
3. $\mathrm{Det}\,\overline{\rho}_f\left(\mathrm{Frob}_q\right) = q^{k-1}\widetilde{\varepsilon\left(q\right)}$,
4. $\mathrm{Det}\,\overline{\rho}_f\left(c\right) = -1$ where c is the complex conjugation.

Moreover, the same properties are true for ρ_f:
$\mathrm{Trace}\,\rho_f\left(\mathrm{Frob}_q\right) = a_q$, $\mathrm{Det}\,\rho_f\left(\mathrm{Frob}_q\right) = q^{k-1}\varepsilon\left(q\right)$, and
$\mathrm{Det}\,\rho_f\left(c\right) = -1$.

For such modular Galois representations everything is fine.

Theorem. For elliptic curves E defined over $\mathbb{Q}$ the following properties are equivalent:

1. E is modular and associated to a newform f;
2. there exists a prime p s.t. the Galois representation $\overline{\rho}_{E,p}$ is modular;
3. for every p, $\overline{\rho}_{E,p}$ is modular;
4. there exists a covering $\pi : X_0 (N_E) \rightarrow E$ of E by the modular curve $X_0 (N_E)$;
5. E is isogenous to the modular abelian variety defined by f.

We have therefore a two completely different ways to Galois representations: elliptic curves and modular forms, and the *unity* of this double way inside the whole unity of mathematics is particularly deep and striking.

As said Richard Taylor in an interview at Harvard [64] concerning Langlands program:

> *"The answer is to my mind extremely surprising; it invokes extremely different objects. You start with this algebraic structure and end up using what are called modular forms, which relate to complex analysis."*

This provides a method of proof. The idea is now to *translate* all the problematic of the *TSW* conjecture and *FLT* into this new context of Galois representations. As Allan Adler explains very well in [1],

> *"The point is that to every elliptic curve one can associate a Galois representation, while in some cases one knows how to associate a Galois representation to a modular form. The idea then is to show that the Galois representation associated to the semi-stable elliptic curve E is of the type one gets from modular forms."*

As Wiles explains for his part ([73], p.445), his aim was to prove a sort of *converse* of Deligne's theorem:

> *"We will be concerned with trying to prove results in the opposite direction, that is to say, with establishing criteria under which a $\mathfrak{p}$-adic representation arises in this way from a modular form."*

Langlands-Tunnell theorem

With Deligne theorem, we can associate to any suitable modular form a Galois representation mod p. But for the converse, which is the objective sought by Wiles, we meet a key difficulty:

there are only *very few results* constructing a cusp eigenform from a Galois representation.

The most important one is the fundamental theorem of Langlands and Tunnell concerning Galois representations ρ of $G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right)$ in $GL_2(\mathbb{C})$ (and not in $GL_2(\mathbb{F}_p)$: representations in $GL_2(\mathbb{C})$ are *Artin* representations).

To formulate it we need to define the smaller congruence group $\Gamma_1(N) = \left\{ \gamma \in SL_2(Z) \mid \gamma \equiv \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \bmod N \right\}$. Remember that $\Gamma_0(N) = \left\{ \gamma \in SL_2(Z) \mid \gamma \equiv \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \bmod N \right\}$

Langlands-Tunnell theorem. Let $\rho : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2(\mathbb{C})$ be a continuous irreducible representation with odd determinant $\mathrm{Det}\,\rho(c) = -1$ ($c =$ complex conjugation). Suppose that the image $\rho(G)$ is a subgroup of S_4 (fondamental hypothesis of *diehedrality*). Then there exist a level N and an eigenform $g \in S_1(\Gamma_1(N))$, $g = \sum_{n \geq 1} b_n \kappa^n$, s.t., for almost every prime q, one has $b_q = \mathrm{Trace}\,\rho(\mathrm{Frob}_q)$.

In our "Hymalayan" metaphor, this highly non trivial theorem on Artin representations could be considered as a sort of forced, narrow and very elevated "mountain pass".

In fact Langlands-Tunnell theorem is valid under the more general condition that $\rho(G)$ is a *solvable* subgroup of $GL_2(\mathbb{C})$ (a group G is solvable if there exists a finite chain of normal subgroups from $\{1\}$ to G whose successive quotients are *abelian*).

As every continuous representation $\rho : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2(\mathbb{C})$ factors through the finite Galois group $\mathrm{Gal}(K/\mathbb{Q})$ of a finite algebraic extension, its image $\rho(G)$ in $GL_2(\mathbb{C})$ (in fact $PGL_2(\mathbb{C})$) is *finite* and is the group of symmetry of a regular polyhedron in $\mathbb{R}^3$.
As ρ is irreducible, the degenerate case of a regular polygon in a plane is excluded and $\rho(G)$ is therefore either A_5 (icosahedral and dodecahedral cases), S_4 (octahedral and cubic cases), A_4 (tetrahedral) or D_{2n} (dihedral case).

- The dihedral case was solved by Hecke;
- A_5 is not solvable;
- the cases of interest are therefore S_4 and A_4.

Langlands-Tunnell theorem doesn't afford directly an eigenform g but an odd *automorphic* cuspidal representation $\pi(\rho)$ of the adelic group $GL_2(\mathbb{A}_{\mathbb{Q}})$ (see Gelbart), where the group of *adeles* of $\mathbb{Q}$ is

$$\mathbb{A}_{\mathbb{Q}} = \left\{ (\alpha_p) \in \mathbb{R} \times \prod_{p \in \mathcal{P}} \mathbb{Q}_p \mid \alpha_p \in \mathbb{Z}_p \text{ for almost every } p \right\}$$

This means that $\pi(\rho) = \otimes_p \pi_p (GL_2(\mathbb{A}_{\mathbb{Q}}))$ is a family $\{\pi_p\}$ (including π_∞) of irreducible unramified representations of the local groups $GL_2(\mathbb{Q}_p)$ (with $\mathbb{Q}_\infty = \mathbb{R}$).

Such a representation $\pi(\rho)$ has the property that, for almost every q, $\text{Trace}\, \rho(\text{Frob}_q) = \text{Trace}\left(t_{\pi_q}\right)$ where $t_{\pi_q} = \begin{pmatrix} \mu_1(p) & 0 \\ 0 & \mu_2(p) \end{pmatrix}$, with μ_1 and μ_2 being the two unramified characters of $\mathbb{Q}_p^\times$ inducing π_p (t_{π_q} is called the Langlands class of π_q in $GL_2(\mathbb{C})$). The traces $\mu_1(p) + \mu_2(p)$ give the coefficients b_p of g.

Wiles new route: SSML for $p = 3, 5$

According to Wiles ([73], p.444):

> *"The key development in the proof is a new and surprising link between two strong but distinct traditions in number theory, the relationship between Galois representations and modular forms on the one hand and the interpretation of special values of L-functions on the other."*

An excellent introduction to the first Wiles proof is the text of Karl Rubin and Alice Silverberg [51] "A report on Wiles' Cambridge lectures", *Bulletin of the AMS* (1994).

As emphasized by Charles Daney ([11], p.2), Wiles theorem

> *"can be seen to be both surprising and beautiful. The reason is that it concerns two apparently quite different sorts of mathematical objects — elliptic curves and modular forms. Each of these is relatively simple and has been studied intensively for ever* 100 *years. Along the way some very surprising parallels have been observed in the theory of each. And the theorem states that the parallels are in fact the results of a fundamental underlying connection between the two."*

Wiles strategy was defined in the following way by Nigel Boston in 2003 [4], in what he called “the big picture”:

> *“A counterexample to Fermat’s Last Theorem would yield an elliptic curve (Frey’s curve) with remarkable properties. This curve is shown as follows not to exist. Associated to elliptic curves and to certain modular forms are Galois representations. These representations share some features, which might be used to define admissible representations. The aim is to show that all such admissible representations come from modular forms (...). We shall parametrize special subsets of Galois representations by complete Noetherian local rings and our aim will amount to showing that a given map between such rings is an isomorphism. This is achieved by some commutative algebra, which reduces the problem to computing some invariants, accomplished via Galois cohomology.”*

A key idea of Wiles is to *weaken TSW* by considering it *modulo p* and then to try to *lift* it p-adically to characteristic 0.

The transformed conjecture is called the "semi-stable modular lifting conjecture".

As pointed out by Kenneth Ribet ([46], p.18) :

> *"Wiles's approach to the Taniyama-Shimura conjecture is 'orthogonal' to one based on consideration of the varying* $\overline{\rho}_{E,p}$*."*

He didn't look, as Serre and Drinfeld suggested, for "a compatible system of *p*-adic representations" but followed rather the suggestion by Mazur and Fontaine to use restrictions on the decomposition and inertia groups.

Instead of looking at all representations $\overline{\rho}_{E,p}$ and try to prove that an infinity of them are modular, he chose to focus on a *single* prime p and to prove that the p-adic *lifting*

$$\rho_{E,p^\infty} : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \rightarrow GL_2\left(\mathbb{Z}_p\right)$$

is modular.

This would be sufficient since

$$\mathrm{Trace}\, \rho_{E,p^\infty}\left(\mathrm{Frob}_q\right) \equiv a_q \text{ for every } q \neq p, q \nmid N$$

Semi-stable modular lifting conjecture (SSML). Suppose that E is *semi-stable* and that there exists a single prime $p \geq 3$ s.t.

(a) $\overline{\rho}_{E,p}$ is irreducible,

(b) E is modular *but only* $\bmod \mathfrak{p}$ (where the ideal $\mathfrak{p}$ lifts p in the ring of integers $\mathcal{O}_f$ of the extension $\mathbb{Q}(a_n)$ of $\mathbb{Q}$ by the algebraic integers a_n with $a_q \equiv q+1-\#E(\mathbb{F}_q)$), i.e. there exists a cusp eigenform $g \in S_2(N)$, $g = \sum_{n \geq 1} b_n \kappa^n$, satisfying
$$b_q \equiv q+1-\#E_q(\mathbb{F}_q) \mod \mathfrak{p}$$ (*very approximative* equality) for almost every prime q,

then E is *really modular*, i.e. there exists a cusp eigenform $f \in S_2(N)$, $f = \sum_{n \geq 1} a_n \kappa^n$, satisfying $a_q = q+1-\#E(\mathbb{F}_q)$ (*exact* equality) for almost every prime q.

Even if weaker than *TSW* the *SSML* conjecture remains highly non trivial since, as was emphasized by Rubin and Silverberg ([51], p.21):

> *"There is no known way to produce such a form in general."*

It is why, as explained by Taylor in his Harvard interview [64]:

> *"The big problem has been to start with a representation of the Galois group and try to produce a modular form."*

Wiles strategy is based on the fact that the SSLM conjecture *for the first two primes* $p = 3, 5$ is *sufficient* to prove the *semi-stable* *TSW* conjecture, which is itself sufficient for *FLT*.

The key reason is that the group $PGL_2(\mathbb{F}_3)$ is isomorphic to the symmetric group S_4 of permutations of 4 elements and that for this *extremely special dihedral case* there exists the Langlands-Tunnell result of modularity.

As Wiles explains in his paper regarding his "first real breakthrough" ([73], p.444):

> *"Suppose that $\overline{\rho}_p$ is the representation of $\mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right)$ on the p-division points of an elliptic curve over $\mathbb{Q}$, and suppose for the moment that $\overline{\rho}_3$ is irreducible. The choice of 3 is critical because a crucial theorem of Langlands and Tunnell shows that if $\overline{\rho}_3$ is irreducible then it is also modular. We then proceed by showing that under the hypothesis that $\overline{\rho}_3$ is semi-stable at 3, together with some milder restrictions on the ramification of $\overline{\rho}_3$ at the other primes, every suitable lifting of $\overline{\rho}_3$ is modular."*

Theorem. Semi-stable modular lifting conjecture for $p = 3, 5 \Rightarrow$ semi-stable $TSW \Rightarrow FLT$. (The case $p = 5$ is needed when $\overline{\rho}_{E,3}$ is reducible.)

Sketch of the proof (see Rubin-Silverberg [51]). Let E be defined over $\mathbb{Q}$ and semi-stable and suppose that the semi-stable modular lifting conjecture is true for $p = 3$.

Semi-stability is here a key property. So Shimura's theorem proving STW in the complex multiplication case is of no help, since if E has complex multiplication its modular j-invariant $j(E)$ is an integer and semi-stability would imply good reduction everywhere which is impossible for an $E/\mathbb{Q}$.

Suppose first that the Galois representation $\overline{\rho}_{E,3}$ is *irreducible* (hypothesis (a)). Then E will be modular via the semi-stable modular lifting conjecture if hypothesis (b) is verified. For proving (b) one relies upon a the Langlands-Tunnell theorem.

To construct ρ in our case, we consider

$$\overline{\rho}_{E,3} : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2\left(\mathbb{F}_3\right).$$

It is irreducible by hypothesis. We use the key fact that $GL_2\left(\mathbb{F}_3\right)$ can be embedded in $GL_2\left(\mathbb{C}\right)$ through a well suited morphism ψ *which factorizes through* $GL_2\left(i\mathbb{Z}\sqrt{2}\right)$ and satisfies

$$\begin{cases} \mathrm{Trace}\left(\psi(g)\right) = \mathrm{Trace}\left(g\right) \mod \left(1 + i\sqrt{2}\right) \\ \mathrm{Det}\left(\psi(g)\right) = \mathrm{Det}\left(g\right) \mod (3) \end{cases}$$

If $\begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$ are generators of $GL_2(\mathbb{F}_3)$, we define explicitely ψ by $\psi\left(\begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix}\right) = \begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix}$ and $\psi\left(\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}\right) = \begin{pmatrix} i\sqrt{2} & 1 \\ 1 & 0 \end{pmatrix}$.

One shows that $\rho = \psi \circ \overline{\rho}_{E,3} : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2(\mathbb{C})$ is irreducible with odd determinant $\mathrm{Det}\,\rho(c) = -1$ and that $\mathrm{Im}(\rho) \subseteq PGL_2(\mathbb{F}_3) \simeq S_4$.

- $\rho = \psi \circ \overline{\rho}_{E,3} : G = \left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2\left(\mathbb{C}\right)$ has odd determinant $\operatorname{Det} \rho(c) = -1$ since $\overline{\rho}_{E,3}$ is odd, $\operatorname{Det} \rho(c) \equiv \operatorname{Det} \overline{\rho}_{E,3}\left(c\right) \pmod 3$ and $-1 \not\equiv 1 \pmod 3$;
- $\operatorname{Im}\left(\rho\right) \subseteq PGL_2\left(\mathbb{F}_3\right) \simeq S_4$ is solvable since S_4 is solvable;
- ρ is irreducible, since $\overline{\rho}_{E,3}$ is absolutely irreducible (see Gelbart). Indeed as $\operatorname{Det} \overline{\rho}_{E,3}\left(c\right) = -1$, $\overline{\rho}_{E,3}\left(c\right)$ has two different eigenvalues (1 and -1) in $\mathbb{F}_3$ and, as far as the only matrices of $M_2\left(\overline{\mathbb{F}}_3\right)$ commuting with $\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$ and non diagonal matrices $\overline{\rho}_{E,3}\left(g\right)$ are the λI, Schur's Lemma implies the irreducibility over $\overline{\mathbb{F}}_3$. If ρ was reducible, then its image in $GL_2\left(\mathbb{C}\right)$ would be abelian since, complex representations of a compact group being completely reducible, ρ would be the direct sum of two characters. But this would imply that $\overline{\rho}_{E,3}$, which is absolutely irreducible, would have an image in $GL_2\left(\overline{\mathbb{F}}_3\right)$ which would be at the same time abelian and irreducible, which is impossible.

One can therefore apply Langlands-Tunnel.

There exist a level N and an eigenform $g \in S_1\left(\Gamma_1\left(N\right), \varepsilon\right)$, $g = \sum_{n \geq 1} b_n \kappa^n$ (with coefficients $b_n \in \mathbb{Z}\left[i\sqrt{2}\right]$), s.t. for almost every prime q one has $b_q = \mathrm{Trace}\,\rho\left(\mathrm{Frob}_q\right)$.

From g, one constructs then a cusp eigenform $f \in S_2\left(N\right) = S_2\left(\Gamma_0\left(N\right)\right)$ s.t. $\forall n\ a_n \equiv b_n \mod \mathfrak{p}$, where $\mathfrak{p}$ is the prime ideal of $\overline{\mathbb{Q}}$ containing $1 + i\sqrt{2}$ (and hence 3).

The congruences show that the eigenform f satisfies (b) (that is modularity modulo a prime over p) for the ideal $\mathfrak{p}' = \mathfrak{p} \cap \mathcal{O}_f$ and therefore E is modular if $SSML$ is true.

The idea for passing from g to f is to multiply g by a non cuspidal form of weight 1, h, which is $\equiv 1 \pmod 3$. Then $gh = \sum\limits_{n \geq 1} c_n \kappa^n$ will be of weight 2 with $c_n \equiv b_n \pmod{\mathfrak{p}}$.

A possibility is to take the Eisenstein series

$$h(\tau) = 1 + 6 \sum_{n \geq 1} \sum_{d|n} \chi(d) \kappa^n$$

where χ is the odd Dirichlet character $(\mathrm{mod}\, 3)$ associated to the imaginary quadratic extension $\mathbb{Q}\left(i\sqrt{3}\right)/\mathbb{Q}$ (we have $\chi(d) = 0, 1, -1$ according to the cases $d \equiv 0, 1, -1 \pmod 3$).
$h \in M_1(3, \chi)$ and $gh \in S_2(3N, \varepsilon\chi)$, the level N and the Nebentypus ε being that of g.

One cannot conclude directly the modularity of $\overline{\rho}_{E,3}$ because gh is not an eigenform but only an eigenform $\pmod{\mathfrak{p}}$. But, due to the theorem of Deligne and Serre, there exists an eigenform $f \in S_2(3N, \varepsilon\chi)$, $f = \sum_{n\geq 1} a_n \kappa^n$ s.t. $a_n \equiv c_n \pmod{\mathfrak{p}'}$ for a $\mathfrak{p}' \mid \mathfrak{p}$ and we are done.

At the end of the proof we used the fact that the restriction of $\overline{\rho}_{E,3}$ to the subgroup $\mathrm{Gal}\left(\mathbb{Q}\left(i\sqrt{3}\right)/\mathbb{Q}\right)$ of $\mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right)$ is *absolutely* irreducible, that is irreducible as a representation in $GL_2\left(\overline{\mathbb{F}}_3\right)$ (see Wiles and Boston).

If it was not the case, the image H would be different from the permutation group S_4 otherwise $\overline{\rho}_{E,3}$ would be surjective and the image of $\mathrm{Gal}\left(\mathbb{Q}\left(i\sqrt{3}\right)/\mathbb{Q}\right)$ would be $SL_2(\mathbb{F}_3)$ and $\overline{\rho}_{E,3}\Big|_{\mathrm{Gal}\left(\mathbb{Q}\left(i\sqrt{3}\right)/\mathbb{Q}\right)}$ would be absolutely irreducible.

We consider therefore the subgroups of S_4.

Now,

- H cannot be included into the subgroup A_4 otherwise $\mathrm{Det}\left(\overline{\rho}_{E,3}\right)$ would be trivial (and not odd).
- H cannot be included into a subgroup S_3 of S_4 otherwise $\overline{\rho}_{E,3}$ would be reducible.
- So H is necessarily a dihedral subgroup of order 8 or a subgroup of index 2 of such a group.

As E is semi-stable, $\overline{\rho}_{E,3}$ can be ramified only at 3 (since the inertia groups I_q for $q \neq 3$ in the splitting field of $\overline{\rho}_{E,3}$ have an order dividing 3).

The abelianization H^{ab} of H would then be an abelian Galois extension of $\mathbb{Q}$ of degree 4 ramified only at 3. But this is impossible.

Suppose now that the representation $\overline{\rho}_{E,3}$ is *reducible*.

If the representation $\overline{\rho}_{E,5}$ is also reducible then E is modular.

Indeed, the group of points of E over $\overline{\mathbb{Q}}$ contains a cyclic subgroup of order $15 = 3.5$ which is G-stable. But the pairs (E, C) are classified by the *rational* points of the modular curve $X_0\,(15)$.

Now, $X_0\,(15)$ has only 4 rational points and it can be shown that they all correspond to modular curves.

To show that, we use Mazur's formula for the genus $g\,(N)$ of $X_0\,(N)$, and we find $g\,(15) = 1$, which means that $X_0\,(15)$ is an elliptic curve.

Indeed (see Rubin [52]),

$$g(15) = 1 + \frac{\mu(\Gamma(15))}{12} - \frac{\nu_2(\Gamma(15))}{4} - \frac{\nu_3(\Gamma(15))}{3} - \frac{\nu_\infty(\Gamma(15))}{2}$$

where $\mu(\Gamma(15)) = [SL(2,\mathbb{Z}) : \{\pm 1\}\Gamma(15)]$, $\nu_i(\Gamma(15))$ for $i = 2,3$ is the number of elliptic points of order i, and $\nu_\infty(\Gamma(15))$ is the number of cusps.

Computations give $\mu(\Gamma(15)) = 24$, $\nu_i(\Gamma(15)) = 0$ for $i = 2,3$, and $\nu_\infty(\Gamma(15)) = 4$. Hence $g(15) = 1 + \frac{24}{12} - \frac{4}{2} = 1$.

A Weierstrass equation of $X_0\left(15\right)$ over $\mathbb{Q}$ is

$$y^2 = x\left(x+3^2\right)\left(x-4^2\right)$$

$X_0\left(15\right)$ has only 8 rational points including its 4 cusps which don't correspond to any elliptic curve. The other 4 points correspond to elliptic curves of modular invariant

$$j \in \left\{-\frac{25}{2}, -\frac{5^2 \times 241^3}{2^3}, -\frac{5 \times 29^3}{2^5}, \frac{5 \times 211^3}{2^{15}}\right\}$$

which possess a *rational* cyclic subgroup C_{15}, are isogenous to

$$E_0 : y^2 + xy + y = x^3 - x - 2$$

and are all modular. As their conductor is $N = 50$, which is not square free, these curves are not semi-stable.

We can therefore suppose that $\overline{\rho}_{E,5}$ is *irreducible* , which implies as before that the restriction of $\overline{\rho}_{E,5}$ to the subgroup $\mathrm{Gal}\left(\mathbb{Q}\left(i\sqrt{5}\right)/\mathbb{Q}\right)$ of $\mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right)$ is absolutely irreducible since the only non trivial extension of $\mathbb{Q}\left(i\sqrt{5}\right)$ unramified outside 5 is $\mathbb{Q}\left(\zeta^5\right)$ which is abelian over $\mathbb{Q}$.

In that case, Wiles method is to use a "3-5 switch" trick due to Mazur and to construct *another* auxiliary elliptic curve E' defined over $\mathbb{Q}$ and semi-stable s.t.

1. $\overline{\rho}_{E',5} = \overline{\rho}_{E,5}$, and
2. $\overline{\rho}_{E',3}$ is *irreducible*.

Let us suppose that E' is constructed. According to the case explained before, E' is modular.

Let $f \in S_2(N)$, $f = \sum_{n \geq 1} a_n \kappa^n$, be the associated eigenform. For almost every prime q we have $a_q = q + 1 - \#E'(\mathbb{F}_q)$. But

$$q + 1 - \#E'(\mathbb{F}_q) \equiv \mathrm{Trace}\, \overline{\rho}_{E',5}(\mathsf{Frob}_q) \bmod 5.$$

And, as $\overline{\rho}_{E',5} = \overline{\rho}_{E,5}$, we have the congruence

$$\mathrm{Trace}\, \overline{\rho}_{E',5}(\mathsf{Frob}_q) = \mathrm{Trace}\, \overline{\rho}_{E,5}(\mathsf{Frob}_q) \equiv q + 1 - \#E(\mathbb{F}_q) \bmod 5$$

and f satisfies therefore the condition (b) of the semi-stable modular lifting conjecture for $p = 5$.

We conclude that E is *modular* if the *SSML* conjecture is true.

At this point, the main difficulty is to construct the auxiliary EC E'.

The sarting point is that ECs E' satisfying $\overline{\rho}_{E',p} = \overline{\rho}_{E,p}$ are classified by the rational points of the Riemann surface $X(p)$ (defined over $\mathbb{Q}$) associated to $\Gamma(p)$ the subgroup of integral matrices of $SL_2(\mathbb{Z})$ which are congruent to the identity matrix modulo p:

$$\Gamma(p) = \{\gamma \in SL_2(\mathbb{Z}) \mid \gamma \equiv \mathrm{Id} \bmod p\}$$

We will use again a *topological* argument, namely that $X(p)$ is of genus $g = 0$ for $p \leq 5$. But when $g = 0$, if there exists a rational point (which is the case here with $E' = E$) then there exist an *infinite number* of rational points. One then shows:

Proposition. For an *infinite* number of rational points of $X(5)$ $\overline{\rho}_{E',3}$ is *irreducible*.

One uses Serre's result that if E' is a *generic* point (and therefore not rational) of $X(5)$ then its Galois group given by its p-torsion points is "big" in the sense that the image of $G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right)$ in $GL_2\left(\mathbb{F}_p\right)$ is maximal (that is equal to $GL_2\left(\mathbb{F}_p\right)$).

But a theorem due to Hilbert, called the *irreducibility theorem*, says that "many" specializations of a generic point have the same Galois group and we can conclude.

One shows next that E' can be chosen semi-stable. If the prime $q \neq 5$ semi-stability reads on $E'[5]$ and as $E'[5] = E[5]$ and E is semi-stable at q by hypothesis, E' is also semi-stable at q.

For $q = 5$ one chooses an E' which is "close" to E for the p-adic metric and uses the fact that semi-stability is an *open* property. As E is semi-stable at 5 by hypothesis, E' is also semi-stable at 5.

Lifting to p-adic representations

Up to now, we have considered only representations of $G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right)$ into $GL_2\left(\mathbb{Z}/N\mathbb{Z}\right)$ induced by the N-torsion (N-division) $\overline{\mathbb{Q}}$-points $E\left[N\right] \simeq \frac{\mathbb{Z}}{N\mathbb{Z}} \times \frac{\mathbb{Z}}{N\mathbb{Z}}$ of ECs.

We will now look at *all* the representations associated to the successive powers p^k of a prime p. Taking their projective limit, we get a continuous representation in the algebra $\mathbb{Z}_p$ of *p-adic integers*

$$\rho_{E,p} : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2\left(\mathbb{Z}_p\right)$$

which satisfies the properties (see Stevens [63], p. 6):

1. Det $\rho_{E,p} = \varepsilon_p$ (where ε_p is the cyclotomic character $\varepsilon_p : G \to \mathbb{Z}_p^\times$) and $\rho_{E,p}$ is odd;
2. $\rho_{E,p}$ is unramified outside pN_E;
3. for almost every prime q,
   Trace $\rho_{E,p}(\mathrm{Frob}_q) = q + 1 - \#E(\mathbb{F}_q)$ (exact equality);
4. If $q \neq p$, $\overline{\rho}_{E,p}$ is unramified at q iff p divides the order of the discriminant Δ_E at q;
5. $\overline{\rho}_{E,p}$ is "flat" at p (see the definition below) iff p divides the order of Δ_E at p.

Of course, through the quotient $\mathbb{Z}_p \to \mathbb{F}_p$, $\rho_{E,p}$ returns $\overline{\rho}_{E,p}$.

Once again, we will say that a p-adic representation

$$\rho : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2\left(\mathbb{Z}_p\right)$$

is *modular* if there exists a cusp eigenform $f \in S_2\left(N\right)$, $f = \sum_{n \geq 1} a_n \kappa^n$, s.t. $\mathrm{Trace}\,\rho\left(\mathrm{Frob}_q\right) = a_q$ for almost every prime q in a well suited extension of $\mathbb{Z}_p$ (for instance a completion $\mathcal{O}_{f,\mathfrak{p}}$ for $\mathfrak{p} \cap \mathbb{Z} = p\mathbb{Z}$).

The semi-stable modular lifting conjecture says essentially that, given E defined over $\mathbb{Q}$ and semi-stable and $p \geq 3$, if $\overline{\rho}_{E,p}$ is irreducible and modular then $\rho_{E,p}$ is modular.

We see that this is a problem of *lifting the modularity property* from the *prime* field $\mathbb{F}_p$ of characteristic p to the *p-adic* algebra $\mathbb{Z}_p$ which is the ring of integers of the *local* field $\mathbb{Q}_p$ of characteristic 0.

In this context the strategy has been pedagogically very well explained by Allan Adler [1].

We have two p-adic Galois representation $\rho_1, \rho_2 : G \to GL_2(\mathbb{Z}_p)$, ρ_1 coming from $E/\mathbb{Q}$ and ρ_2 from a cusp form.

We know that their residual representations $\mathrm{mod}\, p$, $\overline{\rho}_1, \overline{\rho}_2 : G \to GL_2(\mathbb{F}_p)$, are equal and we want to gather some informations on the spaces of $\rho_3 : G \to GL_2(\mathbb{Z}_p)$ s.t. $\overline{\rho}_3 = \overline{\rho}_1 = \overline{\rho}_2$.

In fact ρ_1 and ρ_2 share more properties than $\overline{\rho}_1 = \overline{\rho}_2$: they are unramified for almost every q (i.e. outside a finite set of “bad” primes). We consider only such representations $\rho : G \to GL_2(\mathbb{Z}_p)$.

At this point, we use the deep *analogy between arithmetics and geometry* linking finite fields $\mathbb{F}_p$ and p-adic fields $\mathbb{Q}_p$: $\mathbb{F}_p$ is like values of "functions" at the "*point*" p and the *local* algebra $\mathbb{Z}_p$ is like a "*germs of functions*" in the neighborhood of p.

Therefore a lifting $\overline{\rho} \to \rho$ is like to lift the value of a fonction at a point to a germ of function near the point.

As you know, this deep longstanding analogy dates back to Dedekind, Weber and Hensel who considered the integers n as "functions" over the primes p, using the "valuation" $v_p(n)$ of n at the "points" p, i.e. the power of p in the decomposition of n into prime factors.

To localize the "functions" n in the neiborhood of p we consider first $S = \mathbb{Z} - (p)$ and make the elements of S invertible.

We get the local ring $\mathbb{Z}_{(p)}$ with maximal ideal $\mathfrak{m}_{(p)} = p\mathbb{Z}_{(p)}$ and residue field $\mathbb{Z}_{(p)}/p\mathbb{Z}_{(p)} = \mathbb{F}_p$.

If $n \in \mathbb{Z}$, to look at n "locally" at p is to look at n in $\mathbb{Z}_{(p)}$. The "value" of n at p is its class in $\mathbb{F}_p$, i.e. $n \bmod p$ and the local structure of n at p can be read in $\mathbb{Z}_{(p)}$.

In the local ring $\mathbb{Z}_{(p)}$ every ideal is equal to some power p^k of p.

The successive quotients $\mathbb{Z}_{(p)}/p^{k+1}\mathbb{Z}_{(p)}$ are like successive approximations of order k of the elements $\mathbb{Z}_{(p)}$ (expansion of natural integers n in base p).

Indeed, to make $p^{k+1} = 0$ is to approximate n by a sum $\sum_{i=0}^{i=k} n_i p^i$ with all $n_i \in \mathbb{F}_p$.

It is well known that $|x|_p = p^{-v_p(x)}$ is an *ultrametric norm* on $\mathbb{Q}$. The *projective limit*

$$\mathbb{Z}_p = \varprojlim \frac{\mathbb{Z}}{p^k\mathbb{Z}}$$

is a "profinite" *local* ring with maximal ideal $p\mathbb{Z}_p$, residue field $\frac{\mathbb{Z}_p}{p\mathbb{Z}_p} = \frac{\mathbb{Z}}{p\mathbb{Z}} = \mathbb{F}_p$ and fraction field $\mathbb{Q}_p = \ \mathbb{Q} \otimes_{\mathbb{Z}} \mathbb{Z}_p = \mathbb{Z}_p\left(\frac{1}{p}\right)$.

$\mathbb{Z}_p$ is compact (due to Tychonoff theorem), totally discontinuous (it is a Cantor set) as limit of discrete structures, and is the *completion* of $\mathbb{Z}$ for the p-adic absolute value $|x|_p = p^{-v_p(x)}$.

$\mathbb{Z}$ is a dense subring of $\mathbb{Z}_p$ and $\mathbb{Q}$ is a dense subfield of $\mathbb{Q}_p$.

For a polynomial $P(x) \in \mathbb{Z}[x]$, to have a root in $\mathbb{Z}_p$ is to have a root $\mod p^n$ for every $n \geq 1$.

In the p-adic topology of $\mathbb{Z}_p$, the ideal $p^n\mathbb{Z}_p$ is analogous to the closed ball of radius $\frac{1}{p^n}$.

An intuitive way of looking at p-adic integers $a \in \mathbb{Z}_p$ is to consider them as *series* in p

$$a = \sum_{k=0}^{k=\infty} a_k p^k$$

with $a_k \in \mathbb{F}_p$, the valuation $v_p(a)$ being the least power k such that $a_k \neq 0$.

This representation leads to an analogy with the *Taylor expansion* of a smooth function.

The geometric analogy can be rigourously justified using the concept of *scheme*:

- primes p are the points of the spectrum $\mathrm{Spec}(\mathbb{Z})$ of $\mathbb{Z}$,
- the finite prime fields $\mathbb{F}_p$ are the fibers of the structural sheaf $\mathcal{O}$ of $\mathbb{Z}$,
- integers n are global sections of $\mathcal{O}$,
- $\mathbb{Q}$ is the field of fractions of $\mathcal{O}$.

In this context, $\mathbb{Z}_p$ and $\mathbb{Q}_p$ correspond to the localization of global sections, analog to what are called *germs* of sections in classical geometry.

We apply the analogy to the case where we have a finite algebraic extension k of $\mathbb{F}_p$ and a $\mathbb{Z}_p$-algebra A which is

- *Noetherian* (every prime ideal is finitely generated),
- *local* (there is only one maximal ideal $\mathfrak{p}$),
- *complete* (complete for the Krull topology defined by the successive powers of $\mathfrak{p}$),
- with residue field k.

These properties are the “good” ones for a $\mathbb{Z}_p$-algebra A in this context.

We start with a representation $\overline{\rho} : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2(k)$ and we look for liftings $\rho : G \to GL_2(A)$ making the following diagram commutative ($i \circ \rho = \overline{\rho} \otimes_k \overline{k}$):

$$\begin{tikzcd}
 & GL_2(A) \arrow[d, "i"] \\
G \arrow[ur, "\rho"] \arrow[r, "\overline{\rho}\otimes_k \overline{k}"'] & GL_2\left(\overline{k}\right)
\end{tikzcd}$$

where $i : A \to \overline{k}$ is a morphism and $\overline{\rho} \otimes_k \overline{k}$ extends the field of scalars from k to $\overline{k}$.

This is what Wiles called the "*ring theoretic version*" of the problem.

Infinitesimal deformations

Following the geometric analogy, it is natural to ask what can be a finite order "*infinitesimal*" deformation and in particular a "*tangent*" deformation in this algebraic context.

The key idea, introduced a long time ago by Alexandre Grothendieck, is to define a "tangent vector" of a K-algebra R as a morphism $t : R \rightarrow K[\varepsilon]$ of R into the algebra of *dual numbers* $K[\varepsilon] = \frac{K[T]}{(T^2)}$.

The idea (very old, as old as Leibniz Calculus and introduced by Nieuwentijt) is that a tangent vector is a linear approximation of a Taylor expansion and can be defined and computed using first order *nilpotent* infinitesimal ε s.t. $\varepsilon^2 = 0$.

From infinitesimals to cohomology

Let us now proceed naively. Let $\rho : G \rightarrow GL_2(\mathbb{Z}_p)$. Its residual representation $\overline{\rho} : G \rightarrow GL_2(\mathbb{F}_p)$ associates to every $\gamma \in G$ a 2 × 2 matrix

$$\begin{pmatrix} a_0(\gamma) & b_0(\gamma) \\ c_0(\gamma) & d_0(\gamma) \end{pmatrix} \in GL_2(\mathbb{F}_p).$$

Using the representation of p-adic integers as "Taylor series" we can consider the lifting ρ of $\overline{\rho}$ as associating to every $\gamma \in G$ a 2 × 2 matrix

$$\begin{pmatrix} \sum\limits_{n\geq 0} a_n(\gamma)\, p^n & \sum\limits_{n\geq 0} b_n(\gamma)\, p^n \\ \sum\limits_{n\geq 0} c_n(\gamma)\, p^n & \sum\limits_{n\geq 0} d_n(\gamma)\, p^n \end{pmatrix} \in GL_2(\mathbb{Z}_p)$$

with the $a_0(\gamma)$, $b_0(\gamma)$, $c_0(\gamma)$, $d_0(\gamma)$ returning $\overline{\rho}$.

The Taylor approximations consist in truncating the series at a certain order and in particular the first order linear approximation consists in a representation $\rho^1 : G \to GL_2\left(\mathbb{Z}/p^2\mathbb{Z}\right)$ with matrices

$$\begin{pmatrix} a_0(\gamma) + a_1(\gamma)\,p & b_0(\gamma) + b_1(\gamma)\,p \\ c_0(\gamma) + c_1(\gamma)\,p & d_0(\gamma) + d_1(\gamma)\,p \end{pmatrix} \in GL_2\left(\mathbb{Z}/p^2\mathbb{Z}\right)$$

where $p^2 = 0$, that is where p is treated as an *infinitesimal* ε.

Compute formally in $\mathbb{F}_p[\varepsilon]$ with $\rho^1 : G \to GL_2(\mathbb{F}_p[\varepsilon])$.

ρ^1 is close to $\overline{\rho}$ and to compare them we write $\rho^1(g)\overline{\rho}(g)^{-1} = 1 + \varepsilon a(g)$ with $a(g) \in M_2(\mathbb{F}_p)$.

We consider now the structure of G -module defined by $\overline{\rho}$ on $M_2(\mathbb{F}_p)$. $GL_2(\mathbb{F}_p)$ acts on $M_2(\mathbb{F}_p)$ by conjugation: if $\alpha \in M_2(\mathbb{F}_p)$ and $\overline{g} \in GL_2(\mathbb{F}_p)$, the action $\overline{g} * \alpha$ of $\overline{g}$ on α is given by $\overline{g} * \alpha = \overline{g}\alpha\overline{g}^{-1}$ (what is called the adjoint representation).

We write then that $\rho^1(g) = (1 + \varepsilon a(g))\overline{\rho}(g)$ is a representation, that is $\rho^1(gh) = \rho^1(g)\rho^1(h)$. This imposes drastic conditions on the map $a : G \to M_2(\mathbb{F}_p)$, namely

$$a(gh) = a(g) + \overline{\rho}(g)^{-1} a(h)\overline{\rho}(g) \qquad (*)$$

A key point is that this formula $(*)$ says that the map a is a 1-*cocycle* for the action of G on $M_2(\mathbb{F}_p)$ in the sense of *group cohomology*.

There exists therefore a fundamental link between the first order lifting of $\overline{\rho} : G \to GL_2(\mathbb{F}_p)$ and the cohomology group $H^1(G, M_2(\mathbb{F}_p))$. As Barry Mazur explains ([39], p.245):

> *"First-order infinitesimal" information concerning the universal deformation ring [see below] attached to a representation $\overline{\rho}$ can be expressed in terms of group cohomology (of the adjoint representation of $\overline{\rho}$). This is quite a general phenomenon, does not even depend upon the representability of the deformation problem, and has an appropriate variant for deformation problems subject to conditions."*

It is this idea which has been generalized at higher orders with an extraordinary virtuosity by Barry Mazur, Andrew Wiles and Richard Taylor.

As was emphasized by Lawrence Washington ([72], p.108)

> *"The main reason that Galois cohomology arises in Wiles' work is that certain cohomology groups can be used to classify deformations of Galois representations."*

We will follow Washington's presentation.

We look at representations $\overline{\rho} : G \to GL_2(\mathbb{F}_p) = \mathrm{Aut}\left(\mathbb{F}_p^2\right)$ transforming $\mathbb{F}_p^2$ into a G-module.

The action of G on $\mathbb{F}_p^2$ is naturally extended to $M_2(\mathbb{F}_p) = \mathrm{End}_{\mathbb{F}_p}(\mathbb{F}_p^2)$ via conjugation, $g \in G$ acting on a matrix $A \in M_2(\mathbb{F}_p)$ by

$$A \mapsto g * A = \overline{\rho}(g) A \overline{\rho}(g)^{-1}$$

(change of basis in $\mathbb{F}_p^2$).

When endowed with this G-action, $M_2(\mathbb{F}_p)$ is called *the adjoint representation* and is noted $Ad(\overline{\rho})$, its restriction to matrices of trace 0 being noted $Ad^0(\overline{\rho})$.

Now we want to extend $\overline{\rho}$ to infinitesimal deformations $\rho : G \rightarrow GL_2(\mathbb{F}_p[\varepsilon])$ (ρ is the ρ^1 above), two infinitesimal deformations ρ and ρ' being equivalent when they are conjugated via a change of basis B in $\mathbb{F}_p^2$ congruent mod ε to the identity of $\mathbb{F}_p^2$.

So $\rho' = B\rho B^{-1}$ with $B \in GL_2(\mathbb{F}_p[\varepsilon])$, $B \equiv I_{\mathbb{F}_p^2} \pmod{\varepsilon}$.

Proposition (see Washington [72]). The three following sets are isomorphic:

(a) The set $H^1(G, Ad(\overline{\rho}))$ of classes of cohomology of G in the adjoint representation $Ad(\overline{\rho})$,

(b) The set $\mathrm{Ext}^1(\mathbb{F}_p^2, \mathbb{F}_p^2)$ of extensions of the G-module $\mathbb{F}_p^2$ by itself, that is the set of exact sequences $0 \longrightarrow \mathbb{F}_p^2 \xrightarrow{\alpha} \mathfrak{F} \xrightarrow{\beta} \mathbb{F}_p^2 \longrightarrow 0$,

(c) The set of isomorphism classes of infinitesimal deformations of $\overline{\rho}$.

Let

$$0 \longrightarrow \mathbb{F}_p^2 \xrightarrow{\alpha} \mathfrak{F} \xrightarrow{\beta} \mathbb{F}_p^2 \longrightarrow 0$$

be an extension of $\mathbb{F}_p^2$ by itself. The surjection β admits a section $\varphi : \mathbb{F}_p^2 \longrightarrow \mathfrak{F}$ s.t.

$$\beta \circ \varphi = I_{\mathbb{F}_p^2} : 0 \longrightarrow \mathbb{F}_p^2 \xrightarrow{\alpha} \mathfrak{F} \underset{\varphi}{\overset{\beta}{\rightleftarrows}} \mathbb{F}_p^2 \longrightarrow 0.$$

If $m = (x, y)$ is a point of the $\mathbb{F}_p$-plane $\mathbb{F}_p^2$, we consider in the extension $\mathfrak{F}$ the difference between $\varphi(m)$ and its conjugation by g: $g * \varphi\left(g^{-1} * m\right) - \varphi(m)$. As the surjection β is a morphism of $\mathbb{F}_p[G]$-modules,

$$g * \varphi\left(g^{-1} * m\right) - \varphi(m) \in \mathrm{Ker}(\beta) = \alpha\left(\mathbb{F}_p^2\right).$$

If we apply α^{-1} (which is possible since α is injective) we get $a(g) : \mathbb{F}_p^2 \longrightarrow \mathbb{F}_p^2$ defined by

$$\begin{aligned} a(g)(m) &= \alpha^{-1}\left(g * \varphi\left(g^{-1} * m\right) - \varphi(m)\right) \\ &= \alpha^{-1}\left(\left(\rho(g) \circ \varphi \circ \overline{\rho}(g)^{-1} - \varphi\right)(m)\right) \end{aligned}$$

where $\rho(g)$ corresponds to the structure of G-module of $\mathfrak{F}$ ($\varphi\left(\mathbb{F}_p^2\right)$ is embedded in $\mathfrak{F}$).

The point is to verify that a satisfy the relation

$$a(gh) = a(g) + g * a(h)$$

and is therefore a 1-cocycle of G in the adjoint representation $Ad(\overline{\rho})$.

But

$$a\left(gh\right) = m \mapsto \alpha^{-1}\left(\left(\rho\left(g\right)\rho\left(h\right)\circ\varphi\circ\overline{\rho}\left(h\right)^{-1}\overline{\rho}\left(g\right)^{-1} - \varphi\right)\left(m\right)\right)$$

while

$$\begin{aligned}
a\left(g\right) + g * a\left(h\right) &= a\left(g\right) + \rho\left(g\right)a\left(h\right)\overline{\rho}\left(g\right)^{-1} \\
&= m \mapsto \\
&\alpha^{-1}\left(\left(\rho\left(g\right)\circ\varphi\circ\overline{\rho}\left(g\right)^{-1} - \varphi\right)\left(m\right)\right) + \\
&\rho\left(g\right)\left(\alpha^{-1}\left(\left(\rho\left(h\right)\circ\varphi\circ\overline{\rho}\left(h\right)^{-1} - \varphi\right)\left(m\right)\right)\right)\overline{\rho}\left(g\right)^{-1} \\
&= m \mapsto \\
&\rho\left(g\right)\left(\alpha^{-1}\left(\left(\rho\left(h\right)\circ\varphi\circ\overline{\rho}\left(h\right)^{-1}\right)\left(m\right)\right)\right)\overline{\rho}\left(g\right)^{-1} - \\
&\alpha^{-1}\left(\varphi\left(m\right)\right)
\end{aligned}$$

The two expressions are trivially the same.

Further, if we have two equivalent extensions defined by an isomorphism $\gamma : \mathfrak{F} \to \mathfrak{F}'$ (making the diagram of exact sequences commutative) with sections φ and φ', then the difference between the 1-cocycles a and a' is given by $g * f - f$ for the element

$$f = \alpha^{-1}\left(\gamma^{-1}\left(\varphi' - \gamma \circ \varphi\right)\right)$$

of $Ad\left(\overline{\rho}\right)$ and is therefore a coboundary.

$$\begin{tikzcd}
0 \arrow[r] & \mathbb{F}_p^2 \ni f(m) \arrow[r, shift left, "\alpha"] & \mathfrak{F} \arrow[l, shift left, "\alpha^{-1}"] \arrow[r, shift left, "\beta"] \arrow[d, shift left, "\gamma"] & \mathbb{F}_p^2 \arrow[l, shift left, "\varphi"] \arrow[r] & 0 \\
0 \arrow[r] & \mathbb{F}_p^2 \arrow[r, "\alpha'"'] & \mathfrak{F}' \arrow[u, shift left, "\gamma^{-1}"] \arrow[r, shift right, "\beta'"'] & \mathbb{F}_p^2 \ni m \arrow[l, shift right, "\varphi'"'] \arrow[r] & 0
\end{tikzcd}$$

We have therefore associated to the extension

$$0 \to \mathbb{F}_p^2 \xrightarrow{\alpha} \mathfrak{F} \xrightarrow{\beta} \mathbb{F}_p^2 \to 0$$

the class of cohomology in the adjoint representation $a \in H^1\left(G, Ad\left(\overline{\rho}\right)\right)$.

The link with infinitesimal deformations is done by interpreting the extensions $\mathfrak{F}$ as $\mathbb{F}_p^2 \underset{\mathbb{F}_p}{\otimes} \mathbb{F}_p[\varepsilon] = \mathbb{F}_p^2 \oplus \varepsilon\mathbb{F}_p^2$,

- the injection α being the inclusion of $\mathbb{F}_p^2$ as second component $\varepsilon\mathbb{F}_p^2$,
- the surjection β being the projection on the first component $\mathbb{F}_p^2$,
- and the section φ being simply the identity on this first component.

To define the action of G on $\mathfrak{F} = \mathbb{F}_p^2 \oplus \varepsilon\mathbb{F}_p^2$, we use a cocycle $a : g \mapsto a(g) \in Ad(\overline{\rho})$ and define an infinitesimal deformation of $\overline{\rho}$ by

$$\rho(g) = (I + \varepsilon a(g))\,\overline{\rho}(g)$$

The fact that a is a cocycle warrants the fact that ρ is a representation. Indeed, on the one hand we have

$$\begin{aligned}\rho(gh) &= (I+\varepsilon a(gh))\overline{\rho}(gh)\\ &= \left(I+\varepsilon\left(a(g)+\overline{\rho}(g)a(h)\overline{\rho}(g)^{-1}\right)\right)\overline{\rho}(g)\overline{\rho}(h)\\ &= \overline{\rho}(g)\overline{\rho}(h)+\varepsilon a(g)\overline{\rho}(g)\overline{\rho}(h)+\varepsilon\overline{\rho}(g)a(h)\overline{\rho}(g)^{-1}\overline{\rho}(g)\overline{\rho}(h)\\ &= \overline{\rho}(g)\overline{\rho}(h)+\varepsilon a(g)\overline{\rho}(g)\overline{\rho}(h)+\varepsilon\overline{\rho}(g)a(h)\overline{\rho}(h)\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\rho(g)\rho(h) &= (I+\varepsilon a(g))\overline{\rho}(g)(I+\varepsilon a(h))\overline{\rho}(h)\\ &= \overline{\rho}(g)\overline{\rho}(h)+\overline{\rho}(g)\varepsilon a(h)\overline{\rho}(h)+\varepsilon a(g)\overline{\rho}(g)\overline{\rho}(h)\\ &\text{since } \varepsilon^2=0\end{aligned}$$

and the two expressions are trivially the same.

We see conversely, that if $\rho(g)$ is a representation, then $a(g)$ is a 1-cocycle of $Ad(\overline{\rho})$ given by

$$(I+\varepsilon a(g)) = \overline{\rho}(g)\,\rho(g)^{-1} \ .$$

The equivalence of extensions and infinitesimal deformations can be shown in the following way. The 1-cocycle $\widetilde{a}$ determined by the extension $\mathfrak{F}$ is given here by

$$\begin{aligned}
\widetilde{a}(g)(m) &= \varepsilon^{-1}\left(\left(\rho(g)\circ\varphi\circ\overline{\rho}(g)^{-1}-\varphi\right)(m)\right) \\
&= \varepsilon^{-1}\left(\left((I+\varepsilon a(g))\,\overline{\rho}(g)\circ\varphi\circ\overline{\rho}(g)^{-1}-\varphi\right)(m)\right) \\
&= \varepsilon^{-1}(I+\varepsilon a(g)-I)(m) \ \text{ since } \varphi \text{ is the identity on } \mathbb{F}_p^2 \\
&= a(g)(m)
\end{aligned}$$

and $\widetilde{a} = a$.

It is easy to verify the remaining details of the proposition.

The previous elementary computations can be best expressed in terms of the linear groups $GL_2(\mathbb{F}_p)$ and $GL_2(\mathbb{F}_p[\varepsilon])$ (see Mazur).

If Γ is the kernel $\mathrm{Ker}\,(GL_2(\mathbb{F}_p[\varepsilon]) \to GL_2(\mathbb{F}_p))$ of the projection $\varepsilon \to 0$, $\Gamma = I + \varepsilon M_2(\mathbb{F}_p) \simeq \mathrm{End}_{\mathbb{F}_p}(\mathbb{F}_p^2)$ and we have the short exact sequence

$$1 \to \Gamma \to GL_2(\mathbb{F}_p[\varepsilon]) \to GL_2(\mathbb{F}_p) \to 1$$

splitted by the natural embedding $GL_2(\mathbb{F}_p) \hookrightarrow GL_2(\mathbb{F}_p[\varepsilon])$.

This exact sequence transforms $GL_2(\mathbb{F}_p[\varepsilon])$ in the semi-direct product $GL_2(\mathbb{F}_p) \times \Gamma \simeq GL_2(\mathbb{F}_p) \times M_2(\mathbb{F}_p)$ coming from the adjoint representation.

As $\det((I + \varepsilon a(g))\overline{\rho}(g)) = (1 + \varepsilon\,\mathrm{trace}\,(a(g)))\det(\overline{\rho}(g))$, we see that if we want infinitesimal deformations with constant determinant we must use cocyles of trace $= 0$ and work in the representation $Ad^0(\overline{\rho})$.

We remark also the following "good" property of infinitesimal deformations concerning *ramification*.

Ramification can be read on restricting $G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right)$ to the decomposition subgroups $G_p = \mathrm{Gal}\left(\overline{\mathbb{Q}}_p/\mathbb{Q}_p\right)$ and their inertia subgroups I_p.

If a is a 1-cocycle belonging to the kernel $H^1\left(G_p/I_p, Ad\left(\overline{\rho}\right)^{I_p}\right)$ of the restriction map

$$H^1\left(G, Ad\left(\overline{\rho}\right)\right) \to H^1\left(G_p, Ad\left(\overline{\rho}\right)\right) \to H^1\left(I_p, Ad\left(\overline{\rho}\right)\right)$$

then $\rho\mid_{I_p} = \overline{\rho}\mid_{I_p}$ and the ramification of ρ at p comes entirely from that of $\overline{\rho}$.

In particular, if $\overline{\rho}$ is unramified at p ($\overline{\rho}\mid_{I_p}$ is trivial) then ρ is also unramified at p.

Deformation data and Mazur's conjectures

We have seen how to handle first order infinitesimal deformations of representations $\overline{\rho}_{E,p} : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2\left(\mathbb{F}_p\right)$ and how Galois cohomology enters the stage.

Wiles wanted to show *that modularity is a liftable property*: if $\overline{\rho}_{E,p}$ is modular then its p-adic lifting

$$\rho_{E,p} : G = \mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right) \to GL_2\left(\mathbb{Z}_p\right)$$

is also modular.

His strategy was to make an induction on "Taylor expansions", that is to lift modularity to the successive nth-order infinitesimal deformations and to pass to the limit.

As comments Brian Conrad ([8], p.375):

> *"In order to carry out this procedure, there is an extremely delicate balancing act to handle, with (abstract) deformation rings on one side and (concrete) Hecke rings on the other side. The latter provides a link to modular forms and representations 'coming from modular forms', whereas the former provides a link to the particular representation of interest, $\rho_{E,p}$, which we want to prove 'comes from a modular form'. The relation between the two different types of rings – leading to the proof that they're isomorphic – is supplied by a numerical criterion from commutative algebra. The hard part is to check that this numerical criterion actually can be applied! In order to do this, one has to prove highly non-obvious theorems about the commutative algebra properties of the rings in question. This requires a very detailed understanding of both the deformation rings and the Hecke rings."*

Wiles consider liftings satisfying constraints called "deformation data" $\mathfrak{D}$ by Barry Mazur (this stuff is extremely technical). As he said:

> *"Mazur had been developing the language of deformations of Galois representations. Moreover, Mazur realized that the universal deformation rings he found should be given by Hecke rings, at least in certain special cases. This critical conjecture refined the expectation that all ordinary liftings of modular representations should be modular."*

The concept of deformation comes from *differential geometry* and extends the analogy between algebra and geometry.

For thechnical reasons, he needed to generalize the situation $\mathbb{F}_p$, $\mathbb{Z}_p$, $\mathbb{Q}_p$ to finite extensions k of $\mathbb{F}_p$ and to integer rings $\mathcal{O}$ of finite extensions K of $\mathbb{Q}_p$ with residue field k.

A deformation data is a pair $\mathfrak{D} = (\Sigma, \mathfrak{S})$ where

1. Σ is a *finite* set of “bad” primes q outside of which representations are *unramified*, and
2. $\mathfrak{S}$ is a set of relevant properties of representations ρ at p (to be “ordinary”, to be “flat”, etc.).

Once again, a representation $\overline{\rho} : G \longrightarrow GL_2(k)$ is called $\mathfrak{D}$-*modular* if there exists a cusp eigenform $f \in S_2(N)$ and a prime ideal $\mathfrak{p}$ over p $(\mathfrak{p} \mid p)$ in $\mathcal{O}_f$ s.t. the representation $\rho_{f,\mathfrak{p}}$ associated to f by the Eichler-Shimura construction is a $\mathfrak{D}$-lifting of $\overline{\rho}$.

For technical reasons, Wiles needed to introduce local conditions which are essentially constraints on the p-adic representations $\rho_{E,p} = \overline{\rho}_{E,p^\infty}$ which lift local constraints defined on the residual representations $\overline{\rho}_{E,p}$. They mean that ρ is unramified outside Σ and has the same behavior as its residual representation $\overline{\rho}$.

They are remarkably commented by Ken Ribet ([46], p.20):

> *"There is flexibility and tension implicit in the choice of these conditions. They should be broad enough to be satisfied by $\rho_{E,p}$ and tight enough to be satisfied only by lifts that can be related to modular forms. Roughly speaking, in order to prove the modularity of all lifts satisfying a fixed set of conditions, one needs to specify in advance a space of modular forms S so that the normalized eigenforms in S satisfy the conditions and such that, conversely, all lifts satisfying the conditions are plausibly related to forms in S. It is intuitively clear that this program will be simplest to carry out when the conditions are the most stringent and progressively harder to carry out as the conditions are relaxed."*

"A theme which emerges rapidly is that there are at least two sets of conditions of special interest. Firstly, one is especially at ease when dealing with the most stringent possible set of conditions which are satisfied by $\overline{\rho}_{E,p}$; this leads to what Wiles calls the "minimal" problem. Secondly, one needs at some point to consider some set of conditions which allows treatment of the lift $\rho_{E,p}$ — this lift is, after all, our main target. It would be natural to consider the most stringent such set. The two sets of conditions may coincide, but there is no guarantee that they do; in general, the second set of conditions is more generous than the first."

"Wiles provides a beautiful "induction" argument which enables him to pass from the minimal set of conditions to a non-minimal set. Heuristically, this argument requires keeping tabs on the set of those normalized eigenforms whose Galois representations are compatible with an incrementally relaxing set of conditions. As the conditions loosen, the set of forms must grow to keep pace with the increasing number of lifts. The increase in the number of lifts can be estimated from above by a local cohomological calculation. A sufficient supply of modular forms is then furnished by the theory of congruences between normalized eigenforms of differing level."

Mazur conjecture 1. Let $\overline{\rho} : G \longrightarrow GL_2(k)$ (k an extension of $\mathbb{F}_p$) be *absolutely* irreducible (that is $\overline{\rho} \otimes_k \overline{k}$ is irreducible) and $\mathfrak{D}$-modular, then every $\mathfrak{D}$-lifting of $\overline{\rho}$ to the integer ring $\mathcal{O}$ of a finite extension of $\mathbb{Q}_p$ with residue field k is modular.

Wiles theorem. Mazur 1 $\Rightarrow$ Semi-stable modular lifting conjecture.

Indeed let E be an EC defined over $\mathbb{Q}$, which is semi-stable and satisfies the conditions (a) and (b) of the SSML conjecture for p and let $\overline{\rho}$ be the representation $\overline{\rho} = \overline{\rho}_{E,p}$.

According to hypothesis (a) $\overline{\rho}$ is irreducible. One shows that it is also *absolutely* irreducible. The hypothesis (b) means that $\overline{\rho}$ is modular. Let $\mathfrak{D} = (\Sigma, \mathfrak{S})$ be the deformation data defined by

$$\Sigma = \{p\} \cup \{q \mid E \text{ has bad reduction at } q\}$$

and $\mathfrak{S}$ means “ordinary” if E has ordinary or bad reduction at q (ordinary reduction means good reduction and $E[q]$ has a subgroup of order q which is I_q-stable) and “flat” if E has supersingular reduction at q (supersingular reduction means good reduction and $E[q]$ has no subgroup of order q which is I_q-stable).

One shows that $\rho_{E,p}$ is a $\mathfrak{D}$-lifting of $\overline{\rho}$ and that $\overline{\rho}$ is $\mathfrak{D}$-modular. Mazur 1 implies that $\rho_{E,p}$ is modular and therefore E is modular.

In a second step, following once again the geometric analogy, one reformulates the first Mazur conjecture

> *"as a conjecture that the algebras which parametrize liftings and modular liftings of a given representation are isomorphic. It is this form of Mazur's conjecture that Wiles attacks directly." (Rubin-Silverberg)*

The reformulation is done in terms of *universal deformations* for $(\mathfrak{D}, \mathcal{O})$-deformations, $\mathcal{O}$ being the p-adic ring of integers of a finite extension of $\mathbb{Q}_p$

$$\begin{tikzcd}
 & GL_2\left(A\right) \arrow[d, "i"] \\
G \arrow[ur, "\rho"] \arrow[r, "\overline{\rho}"'] & GL_2\left(k\right)
\end{tikzcd}$$

where A is a *local*, Noetherian, complete $\mathcal{O}$-algebra of residue field k. The concept of universal deformation is then associated to the existence of a very special algebra $\mathfrak{R}$.

Mazur-Ramakrishna theorem. There exists a *universal* $(\mathfrak{D}, \mathcal{O})$-lifting $\rho_{\mathfrak{R}} : G \longrightarrow GL_2(\mathfrak{R})$ of $\overline{\rho}$, that is for every $(\mathfrak{D}, \mathcal{O})$-lifting $\rho : G \longrightarrow GL_2(A)$ there exists one and only one morphism of algebras $\varphi_\rho : \mathfrak{R} \longrightarrow A$ s.t. the following diagram is commutative:

$$\begin{tikzcd}
 & GL_2(A) \arrow[d, "\varphi_\rho^*"] \\
G \arrow[ur, "\rho"] \arrow[r, "\rho_{\mathfrak{R}}"'] \arrow[d, equal] & GL_2(\mathfrak{R}) \arrow[d, "i"] \\
G \arrow[ur] \arrow[r, "\overline{\rho}"'] & GL_2(k)
\end{tikzcd}$$

This fundamental theorem means that the *functor* $A \rightsquigarrow \mathcal{L}(A)$ which associates to every $\mathcal{O}$-algebra A as above the set of liftings of $\overline{\rho} : G \rightarrow GL_2(k)$ to A is *representable* by $\mathfrak{R}$.

There exists therefore an isomorphism

$$\mathcal{L}(A) \simeq \mathsf{Hom}_{\mathsf{cont}}(\mathfrak{R}, A)$$

where $\mathrm{Hom}_{\mathrm{cont}}(\mathfrak{R}, A)$ is the set of continuous homomorphisms from $\mathfrak{R}$ to A.

The theorem can be proved first without conditions and then relativized to representations ρ “with particularly desirable properties”. As Barry Mazur explains:

> *“The recipe for cutting down the “universal deformation” to these more specifically desirable Galois representations is (surprisingly enough!) at last conceptually nothing more than the “imposition” of local conditions at the ramified primes, and sometimes with the additional prescription of the appropriate global determinant.”*

But if $\overline{\rho}$ is $\mathfrak{D}$-*modular* with an eigenform f and a prime ideal $\mathfrak{p}$ of $\mathcal{O}_f$ s.t. $\rho_{f,\mathfrak{p}}$ is a $\mathfrak{D}$-lifting of $\overline{\rho}$ and $\rho_{f,\mathfrak{p}} \otimes_{\mathcal{O}_f} \mathcal{O}$ is a $(\mathfrak{D}, \mathcal{O})$-lifting of $\overline{\rho}$ then there exists also a *modular universal deformation* in the following sense:

T1 The $\mathcal{O}$-algebra A is a generalized *Hecke algebra* $\mathfrak{T}$ of operators satisfying the expected properties.

T2 There exists a level N divisible only by "bad" primes $q \in \Sigma$ s.t. $\mathfrak{T}$ is generated over $\mathcal{O}$ by the images T'_q of the T_q of $T(N)$ for $q \notin \Sigma$.

T3 There exists a $(\mathfrak{D}, \mathcal{O})$-lifting of $\overline{\rho}$, $\rho_{\mathfrak{T}} : G \rightarrow GL_2(\mathfrak{T})$, s.t.

$$\mathrm{Trace}\, \rho_{\mathfrak{T}}(\mathsf{Frob}_q) = j(T_q) \text{ for every "good" prime } q \ .$$

T4 If ρ is a modular $(\mathfrak{D}, \mathcal{O})$ -lifting of $\overline{\rho}$ to an A, then there exists one and only one $\mathcal{O}$-morphism $\psi_\rho : \mathfrak{T} \rightarrow A$ s.t. the following diagram is commutative:

$$\begin{tikzcd}
 & GL_2(\mathfrak{T}) \arrow[d, "\psi_\rho^*"] \\
G \arrow[ur, "\rho_{\mathfrak{T}}"] \arrow[r, "\rho"'] & GL_2(\mathfrak{R})
\end{tikzcd}$$

As $\rho_{\mathfrak{T}}$ is a $(\mathfrak{D}, \mathcal{O})$-lifting of $\overline{\rho}$, Mazur-Ramakrishna theorem implies that there exists one and only one morphism of algebras $\varphi : \mathfrak{R} \rightarrow \mathfrak{T}$ s.t. $\rho_{\mathfrak{T}} = \varphi \circ \rho_{\mathfrak{R}}$. The map φ is *surjective* since

$$\forall q \notin \Sigma, \ \varphi \left(\mathrm{Trace}\, \rho_{\mathfrak{R}} \left(\mathsf{Frob}_q \right) \right) = \mathrm{Trace}\, \rho_{\mathfrak{T}} \left(\mathsf{Frob}_q \right) = T'_q$$

and the T'_q generate $\mathfrak{T}$ by (T2).

Following the key idea that the general case is always modular, Mazur introduced a second conjecture saying intuitively that parametrizations of *ordinary* liftings and *modular* liftings are equivalent or that "universal" is equivalent to "modular universal", which is clearly a *translation* of the *TSW* conjecture in the context of universal deformations.

Mazur conjecture 2. $\varphi : \mathfrak{R} \rightarrow \mathfrak{T}$ is an *isomorphism*.

Theorem. Mazur conjecture 2 implies Mazur conjecture 1.

Sketch of the proof. Let $\overline{\rho} : G \to GL_2(k)$ be absolutely irreducible and $\mathfrak{D}$-modular. If ρ is a $\mathfrak{D}$-lifting of $\overline{\rho}$ to A, we want to show that ρ is modular.

We first extend ρ and $\overline{\rho}$ to $\mathcal{O}$ and ρ becomes a $(\mathfrak{D}, \mathcal{O})$-lifting. Let $\psi_\rho : R \to A$ be the morphism of algebras asserted by Mazur-Ramakrishna theorem. If $\varphi : R \to T$ is an *isomorphism* we can consider the *inverse* map $\varphi^{-1} : T \to R$ and the composed map $\psi = \psi_\rho \circ \varphi^{-1} : T \to A$

$$\psi : T \overset{\varphi^{-1}}{\to} R \overset{\psi_\rho}{\to} A$$

We deduce from (T3) that $\psi(T_q) = \mathrm{Trace}\,\rho(\mathrm{Frob}_q)$ for almost every prime q. Shimura results imply then the existence of an eigenform $f \in S_2(N)$, $f = \sum\limits_{n \geq 1} a_n q^n$, s.t. $a_q = \mathrm{Trace}\,\rho(\mathrm{Frob}_q) = \psi(T_q)$ for almost every prime q. But this implies that the representation ρ is modular.

Complete intersections and Gorenstein property

We get universal local algebras $\mathfrak{R}$ associated to deformation data $\mathfrak{D} = (\Sigma, \mathfrak{S})$. As we have noted, $\mathfrak{R}$ *represents* the functor $\mathcal{L}$ which associates functorially to every local algebra A as above the set of deformations $\mathcal{L}(A) = \{\text{lifting of } \overline{\rho} \text{ to } A\}$.

The problem is to measure in some sense the "*size*" of $\mathfrak{R}$. The simplest way to do that is to compute *the dimension of its "(co)tangent space"* at its maximal ideal $\mathfrak{m}_{\mathfrak{R}}$.

As explains Barry Mazur ([39], p.271):

> *"The intuition behind this definition is that if one thinks of $\mathfrak{R}$ as being "functions on some base-pointed space", then $\mathfrak{m}_{\mathfrak{R}}$ may be thought of as those functions vanishing at the base point, and $\mathcal{T}^*_{\mathfrak{R}}$ is the quotient of $\mathfrak{m}_{\mathfrak{R}}$ by the appropriate ideal (of "higher order terms" of these functions) so as to isolate the "linear parts" of these functions."*

Complete intersections and Gorenstein property

The problem is now to prove that $\varphi : \mathfrak{R} \to \mathfrak{T}$ is an *isomorphism*. There are two steps:

1. prove the isomorphism at the *(co)tangent* level;
2. prove that $\mathfrak{R}$ and $\mathfrak{T}$ have the property that *(co)tangent isomorphism implies isomorphism*.

Surjectivity is "easy". For injectivity, Wiles introduced a fundamental numerical criterion. The idea is to find a *bound* for the order of "(co)tangent spaces" at prime ideals of $\mathfrak{R}$.

If $\overline{\rho}$ is $\mathfrak{D}$-modular, there exists a cusp eigenform $f \in S_2(N)$ and a prime ideal $\mathfrak{p} \mid p$ ($p \in \mathfrak{p}$) of $\mathcal{O}_f$ such that $\rho_{f,\mathfrak{p}}$ is a $\mathfrak{D}$-lifting of $\overline{\rho}$. If $\mathcal{O}_f \subset \mathcal{O}$ (with K the field of fractions of $\mathcal{O}$), then $\rho_{f,\mathfrak{p}} \bigotimes_{\mathcal{O}_f} \mathcal{O}$ is a $(\mathfrak{D}, \mathcal{O})$-lifting of $\overline{\rho}$.

As the Galois representation $\rho_{f,\mathfrak{p}} \bigotimes_{\mathcal{O}_f} \mathcal{O}$ is modular by construction, due to the universality property (T4), there exists one and only one morphism $\pi_{\mathfrak{T}} : \mathfrak{T} \to \mathcal{O}$ s.t. $\rho_{f,\mathfrak{p}} \bigotimes_{\mathcal{O}_f} \mathcal{O}$ factorizes through $\rho_{\mathfrak{T}}$, i.e. the composed map

$$G \xrightarrow{\rho_{\mathfrak{T}}} GL_2(\mathfrak{T}) \xrightarrow{\pi_{\mathfrak{T}}} GL_2(\mathcal{O})$$

satisfies

$$\pi_{\mathfrak{T}} \circ \rho_{\mathfrak{T}} = \rho_{f,\mathfrak{p}} \bigotimes_{\mathcal{O}_f} \mathcal{O} \ .$$

Let $\mathfrak{p}_{\mathfrak{T}} = \mathrm{Ker}(\pi_{\mathfrak{T}})$. Consider $\varphi : \mathfrak{R} \to \mathfrak{T}$ and

$$\mathfrak{p}_{\mathfrak{R}} = \mathrm{Ker}(\pi_{\mathfrak{T}} \circ \varphi) = \varphi^{-1}(\mathfrak{p}_{\mathfrak{T}}) = \mathrm{Ker}(\pi_{\mathfrak{R}})$$

where $\pi_{\mathfrak{R}}$ is the (unique) map $\pi_{\mathfrak{R}} : \mathfrak{R} \to \mathcal{O}$ given by the universal property of $\mathfrak{R}$.

We have therefore:

$$\begin{tikzcd}
\mathfrak{R} \arrow[rr, shift left, "\varphi^{-1}"] \arrow[dr, "\pi_{\mathfrak{R}}"'] & & \mathfrak{T} \arrow[ll, shift left, "\varphi"] \arrow[dl, "\pi_{\mathfrak{T}}"] \\
& \mathcal{O} &
\end{tikzcd}$$

The property (T2) of $\mathfrak{T}$ (to be generated over $\mathcal{O}$ by the Hecke operators T_q for "good" q) and the fact that, for almost every prime q, $\mathrm{Trace}\,\rho_{f,\mathfrak{p}}(\mathrm{Frob}_q) = a_q$ imply that, for almost every prime q, $\pi_{\mathfrak{T}}(T_q) = a_q$.

The cotangent spaces of the schemes $\mathrm{Spec}\,(\mathfrak{R})$ and $\mathrm{Spec}\,(\mathfrak{T})$ at the points $\mathfrak{p}_{\mathfrak{R}} = \mathrm{Ker}\,(\pi_{\mathfrak{R}})$ and $\mathfrak{p}_{\mathfrak{T}} = \mathrm{Ker}\,(\pi_{\mathfrak{T}})$ are respectively $\frac{\mathfrak{p}_{\mathfrak{R}}}{\mathfrak{p}_{\mathfrak{R}}^2}$ and $\frac{\mathfrak{p}_{\mathfrak{T}}}{\mathfrak{p}_{\mathfrak{T}}^2}$.

At this point Wiles uses a special property of the Hecke algebra $\mathfrak{T}$, namely to be a *Gorenstein ring*.

This result due to by Barry Mazur means that there exists a (non canonical) isomorphism of $\mathfrak{T}$-modules between $\mathfrak{T}$ and $\mathrm{Hom}_{\mathcal{O}}\,(\mathfrak{T}, \mathcal{O})$. As explained Wiles,

> *"The turning point in this and indeed in the whole proof came in the spring of* 1991*. (...) I had already needed to verify that the Hecke rings were Gorenstein in order to compute the congruences developed in Chapter 2. This property had first been proved by Mazur in the case of prime level and his argument had already been extended by other authors as the need arose."*

The morphism $\pi_{\mathfrak{T}} : \mathfrak{T} \to \mathcal{O}$ corresponds to an element ξ of $\mathfrak{T}$ and, via $\pi_{\mathfrak{T}}$ itself, to an element $\pi_{\mathfrak{T}}(\xi)$ of the ring $\mathcal{O}$:

$$\begin{tikzcd}
\mathrm{Hom}_{\mathcal{O}}(\mathfrak{T}, \mathcal{O}) \arrow[r, "\sim"] & \mathfrak{T} \arrow[r, "\pi_{\mathfrak{T}}"] & \mathcal{O} \\
\pi_{\mathfrak{T}} \arrow[r, mapsto] & \xi \arrow[r, mapsto] & \pi_{\mathfrak{T}}(\xi)
\end{tikzcd}$$

Let η be the ideal $(\pi_{\mathfrak{T}}(\xi))$ of $\mathcal{O}$ (η is well defined idependently of the isomorphism $\mathrm{Hom}_{\mathcal{O}}(\mathfrak{T}, \mathcal{O}) \simeq \mathfrak{T}$).

Wiles gave a sufficient condition for $\varphi : \mathfrak{R} \to \mathfrak{T}$ to be an isomorphism in terms of order of the “cotangent space” $\mathfrak{p}_{\mathfrak{R}}/\mathfrak{p}_{\mathfrak{R}}^2$.

As φ is onto, we already have

$$\#\left(\frac{\mathfrak{p}_{\mathfrak{R}}}{\mathfrak{p}_{\mathfrak{R}}^2}\right) \geq \#\left(\frac{\mathfrak{p}_{\mathfrak{T}}}{\mathfrak{p}_{\mathfrak{T}}^2}\right).$$

Theorem (Wiles).

- $\# \left(\frac{\mathcal{O}}{\eta} \right) \leq \# \left(\frac{\mathfrak{p}_{\mathfrak{T}}}{\mathfrak{p}_{\mathfrak{T}}^2} \right) \leq \# \left(\frac{\mathfrak{p}_{\mathfrak{R}}}{\mathfrak{p}_{\mathfrak{R}}^2} \right)$.
- If $\# \left(\frac{\mathfrak{p}_{\mathfrak{T}}}{\mathfrak{p}_{\mathfrak{T}}^2} \right)$ (and therefore $\# \left(\frac{\mathcal{O}}{\eta} \right)$) are *finite*, then $\mathfrak{T}$ and $\mathfrak{R}$ are *complete intersections* iff $\# \left(\frac{\mathfrak{p}_{\mathfrak{T}}}{\mathfrak{p}_{\mathfrak{T}}^2} \right) = \# \left(\frac{\mathcal{O}}{\eta} \right)$.
- Further, if $\# \left(\frac{\mathfrak{p}_{\mathfrak{R}}}{\mathfrak{p}_{\mathfrak{R}}^2} \right) = \# \left(\frac{\mathcal{O}}{\eta} \right)$, then $\varphi : \mathfrak{R} \to \mathfrak{T}$ is an *isomorphism* of complete intersection rings.

An $\mathcal{O}$-algebra A is a complete intersection if
$A \simeq \mathcal{O}\left[\left[T_1, \ldots, T_r \right]\right] / \left(f_1, \ldots, f_r \right)$ (as many relations as variables).

So, if $\#\left(\frac{\mathfrak{p}_{\mathfrak{R}}}{\mathfrak{p}_{\mathfrak{R}}^2}\right) \leq \#\left(\frac{\mathcal{O}}{\eta}\right)$, then we get the equalities

$$\#\left(\frac{\mathfrak{p}_{\mathfrak{R}}}{\mathfrak{p}_{\mathfrak{R}}^2}\right) = \#\left(\frac{\mathfrak{p}_{\mathfrak{T}}}{\mathfrak{p}_{\mathfrak{T}}^2}\right) = \#\left(\frac{\mathcal{O}}{\eta}\right)$$

and φ induces an isomorphism of the "cotangent spaces" of $\mathfrak{R}$ and $\mathfrak{T}$ at the corresponding "points" $\mathfrak{p}_{\mathfrak{R}}$ and $\mathfrak{p}_{\mathfrak{T}}$.

Due to the fact that $\mathfrak{T}$ is a *complete intersection* over $\mathcal{O}$, this "tangent isomorphism" implies that φ is an isomorphism. And we are done...

Indeed, as Darmon, Diamond and Taylor explain in [12]:

> *"The usefulness of the notion of complete intersections comes from the following two (vaguely stated) principles:*
>
> *1. Isomorphisms to complete intersections can often be recognized by looking at their effects on the tangent spaces.*
>
> *2. Isomorphisms from complete intersections can often be recognized by looking at their effects on the invariants η."*

The last difficulty in the proof of the *TSW* conjecture is then to *bound* the order $\#\left(\frac{\mathcal{O}}{\eta}\right)$. The new idea is to give a *cohomological* interpretation of "tangent spaces" in terms of *Selmer groups*. It is the most technical and difficult part of the proof!

Selmer groups

Selmer groups enter the stage because the classical local/global Hasse-Minkowski principle for solving Diophantine problems does not apply to ECs.

One considers solutions in the "local" fields $\mathbb{Q}_p$ and $\mathbb{R}$ as *local* solutions, and solutions in the "global" field $\mathbb{Q}$ as *global* solutions.

Of course, if global solutions on $\mathbb{Q}$ exist they can be localized and local solutions exist for every prime p, their *coherence* being ensured by the underlying global solution.

The main problem is to solve the *inverse* problem, that is when local solutions imply the existence of global solutions. It is highly non trivial.

Many classical theorems say that if a Diophantine equation $f = 0$ has “local” solutions at every “point” p (i.e. $\mod p$) and a solution over $\mathbb{R}$ (point at infinity) then it has a “gobal” solution over $\mathbb{Z}$. The best known is *Minkoswki's theorem* that proves this assertion for *quadratic forms* with *rational* coefficients.

But this Hasse principle is not verified by algebraic curves. In 1951, Selmer gave the famous counterexample of the projective cubic C over $\mathbb{Z}$ of equation

$$3x^3 + 4y^3 + 5z^3 = 0$$

which has solutions modulo every prime p and over $\mathbb{R}$ but has no rational point at all.

Let E be an elliptic curve defined over $\mathbb{Q}$. It is also trivially defined over $\mathbb{Q}_p$ and $\mathbb{R}$ since $\mathbb{Q}$ is a subfield of $\mathbb{Q}_p$ and $\mathbb{R}$.

The main question is to know in what exact sense the "local" elliptic curves $E/\mathbb{Q}_p$ and $E/\mathbb{R}$ determine the "global" one $E/\mathbb{Q}$.

An elliptic curve $E'/\mathbb{Q}$ such that $E'/\mathbb{Q}_p \simeq E/\mathbb{Q}_p$ and $E'/\mathbb{R} \simeq E/\mathbb{R}$ is called a *companion* of $E/\mathbb{Q}$ and the main problem is to compute what is called *the Selmer group* $\mathcal{S}(E)$ of the classes of isomorphisms of the companions of E.

This fondamental concept is commented in the following way by Barry Mazur for any algebraic variety V ([38], p.21):

> *"One can think of the cardinality of $\mathcal{S}(V)$ as roughly analogous to a class number, i.e., a measure of the extent to which local data (in this case, the isomorphism classes of $V/\mathbb{Q}_p$ for all p, and $V/\mathbb{R}$) determine or fail to determine global data (the isomorphism class of $V/\mathbb{Q}$). One might say that the local-to-global principle holds for a class of varieties $\mathcal{V}$ if $\mathcal{S}(V)$ consists of the single isomorphism class $\{V\}$ for each member V of $\mathcal{V}$."*

Using deep results of Rubin and Kolyvagin, Mazur proved the

Theorem (Mazur). The set $\mathcal{S}(C)$ of the non isomorphic companions of the Selmer curve C is constituted by the 5 curves defined over $\mathbb{Z}$: $3x^3 + 4y^3 + 5z^3 = 0$, $12x^3 + y^3 + 5z^3 = 0$, $15x^3 + 4y^3 + z^3 = 0$, $3x^3 + 20y^3 + z^3 = 0$, $60x^3 + y^3 + z^3 = 0$, and the last curve J is the common Jacobian of the four other curves and is the only one to have a $\mathbb{Q}$-rational point ($\{0, 1, -1\}$, it is unique).

In fact the natural interpretation of Selmer groups is *cohomological*.

Conclusion

Jean Dieudonné liked to claim in a rather provocative way:

> *"Vouloir faire des mathématiques une partie de la logique est une affirmation aussi absurde que celle qui consisterait à dire que les œuvres de Shakespeare ou de Goethe font partie de la grammaire!"*

Of course, literary works are made up of sentences. But besides their linguistic structure they have also a *narrative* structure.

It is the same thing in mathematics. Complex proofs have a conceptual "narrative" structure and, for me, philosophy of mathematics is prominently concerned with it.

Beside the logical context of justification, we need a "semiotic" investigation of the context of discovery to understand why "great" proofs are artworks.

Some developements since 1994

General TSW conjecture

Fred Diamond generalized Wiles result to the case of elliptic curves E defined over $\mathbb{Q}$ which are semi-stable only at $p = 3$ and $p = 5$. A corollary (Rubin-Silverberg) was that if the 2-division points of $E/\mathbb{Q}$ are rational then E is modular.

And finally in 1999 Christian Breuil, Brian Conrad, Diamond and Taylor generalized it to *all* elliptic curves $E/\mathbb{Q}$.

This final achievement required a lot of hard computations made possible by new techniques introduced by Breuil.

It is interesting to emphasize that these new translations show how the reinterpretation is a never-ended open process.

As the authors claim [5], p.848):

> *"In the key computation of the local deformation rings, we now make use of a new description (due to Breuil) of finite flat group schemes over the ring of integers of any p-adic field in terms of certain (semi)-linear algebra data. (...) It seems miraculous to us that these long computations with finite flat group schemes (...) give answers completely in accord with predictions made from much shorter computations with the local Langlands correspondence and the modular representation theory of* $GL_2(\mathbb{Q}_3)$*. We see no direct connection, but cannot help thinking that some such connection should exist."*

Serre's conjecture

Serre's modularity conjecture has been proved in 2005 by Chandrashekhar Khare in the level 1 case, and later on in 2008 by Khare and Jean-Pierre Wintenberger.

Swinnerton-Dyer conjecture

A lot of deep results "à la Fermat" on Diophantine equations proceed from these extraordinary achievements.

Other very important consequences concern the theory of elliptic curves, e.g. the celebrated Birch and Swinnerton-Dyer conjecture saying that $L_E(s)$ is analytic on the *whole* complex plane $\mathbb{C}$ (in particular at $s = 1$) with $\mathrm{ord}_{s=1} L = r$, where r is the *rank* of E.

Due to the Mordell-Weil theorem, the group of rational points $E(\mathbb{Q})$ is a finitely generated abelian group and is therefore of the form $E(\mathbb{Q}) = T + \mathbb{Z}^r$. We have Mazur's theorem for the torsion subgroup T and r is the rank.

As Henri Darmon explains ([12], p.1399):

> *"Knowing that E is modular also gives control on the arithmetic of E in other ways, by allowing the construction of certain global points on E defined over abelian extensions of quadratic imaginary fields via the theory of complex multiplication. Such analytic constructions of global points on E actually play an important role in studying the Birch and Swinnerton-Dyer conjecture through the work of Gross-Zagier and of Kolyvagin."*

$\mathbb{Q}$-curves

Other generalizations concern the situation of ECs defined not over $\mathbb{Q}$ but over an extension of $\mathbb{Q}$ such as $\mathbb{Q}(i)$ (imaginary quadratic field) or $\mathbb{Q}\left(\sqrt{2}\right)$ (totally real field). They are extremely difficult.

Of particular interest are what are called $\mathbb{Q}$-curves. They are elliptic curves E/K defined over a Galois extension K of $\mathbb{Q}$ having the property of being isogenous to all their Galois conjugates. As explains Jordan Ellenberg in ([21]), they constitute

> *"the 'mildest possible generalization' of the class of elliptic curves over $\mathbb{Q}$."*

Kenneth Ribet proposed the conjecture that an elliptic curve over $\mathbb{C}$ is modular iff it is a $\mathbb{Q}$-curve.

Abelian varieties

Another conjecture asserts that if A is an abelian variety over $\mathbb{Q}$ for which $\mathrm{End}\,(A) \underset{\mathbb{Z}}{\otimes} \mathbb{Q}$ is a number field of degree equal to $\dim A$, then there exists an hyperbolic uniformization of A defined over $\mathbb{Q}$. As explains Ribet ([46], p.383) :

> *"It is natural to regard Conjecture* 1 *and Conjecture* 2 *as generalizations of the Taniyama-Shimura conjecture. The first conjecture pertains to elliptic curves which are not necessarily defined over* $\mathbb{Q}$*, while the second pertains to abelian varieties over* $\mathbb{Q}$ *which are not necessarily elliptic curves. Neither of these conjectures is proved."*

Langlands program

Another line of generalization consists in studying higher dimensional representations $\rho : G \to GL_n\left(\overline{\mathbb{F}}_p\right)$ with $n > 2$. See e.g. the works of Avner Ash.

The *TSW* conjecture is part of the research program on the relations between Galois representations and automorphic forms known as *Langlands program*. Langlands conjectures have been proved in 1998 for local fields by Harris and Taylor and in 1999 for function fields by Louis Lafforgue (who won for that the Fields medal in 2002).

[1] Adler, A., *Lecture notes on Fermat's Last Theorem*, University of Rhode Island, URI, 1993.

[2] Artin, E., Tate, J., *Class Field Theory*, Benjamin, New York-Amsterdam, 1968.

[3] Ash, A., Gross, R., "Generalized non-abelian reciprocity laws: a context for Wiles' proof", *Bull. London. Math. Soc.*, **32**, (2000), 385-397. URL: http://fwww.bc.edu/MT/gross/avner.tex.txt.

[4] Boston, N., *The Proof of Fermat Last Theorem*, 2003 (COMPLETER)

[5] Breuil, C., Conrad, B., Diamond, F., Taylor, R., "On the modularity of elliptic curves over $\mathbb{Q}$: wild 3-adic exercises", *Journal of the American Mathematical Society*, **14**, 4, (2001), 843-939.

[6] Carayol, H., "Sur les représentations galoisiennes modulo *l* attachées aux formes modulaires", *Duke Math. J.* 59 (1989), 785-801.

[7] Cassels, J.W.S., *Lectures on Elliptic Curves*, New York, Cambridge University Press, 1991.

[8] Conrad, B., "The Flat Deformation Functor", *Modular Forms and Fermat's Last Theorem*, (G. Cornell, J. H. Silverman, G. Stevens eds), New-York, Springer, 1997, 373-420.

[9] Corry, L., Hunting Prime Numbers—from Human to Electronic Computers, The Rutherford Journal, article030105.html

[10] Corry, L., "Number crunching vs. number theory: computers and FLT, from Kummer to SWAC (1850-1960), *Arch. Hist. Exact Sci.*, 2007. COMPLETER

[11] Daney, C., *The Mathematics of Fermat's Last Theorem*, 1996. URL. COMPLETER.

[12] Darmon, H., Diamond, F., Taylor, R.L., "Fermat's Last Theorem", in *Current Developments in Mathemathics*, International Press, Cambridge, MA, 1996, 1-154.

[13] Darmon, H., "A Proof of the Full Shimura-Taniyama-Weil Conjecture is Announced", *Notices of the AMS*, December ,**46**, 11, (1999), 1397-1401.

[14] Deligne, P., "Formes modulaires et représentations *l*-adiques", *Séminaire Bourbaki*, 1968-1969, Exposé 355, *Lect. Notes in Maths.* 179 (1971), 139-172.

[15] Deligne, P., Serre, J-P., "Formes modulaires de poids 1", *Ann. Sci. ENS*, 7 (1974), 507-530.

[16] Diamond, F., "On deformations rings and Hecke rings", *Annals of Mathematics*, (2), **144**, 1, (1996), 137-166.

[17] Diamond, F., “The Taylor-Wiles construction and multiplicity one”, *Invent. Math.*, COMPLETER.

[18] Dieudonné, J., *Panorama des Mathématiques pures. Le choix bourbachique*, Paris, Gauthier-Villars, 1977.

[19] Edixhoven, B., “Serre's Conjectures”, *Modular Forms and Fermat's Last Theorem*, (G. Cornell, J. H. Silverman, G. Stevens eds), New-York, Springer, 1997, 209-242.

[20] Edwards, H. M., *Fermat's Last Theorem: A Genetic Introduction to Algebraic Number Theory*, New-York, Springer, 1977.

[21] Ellenberg, J., “$\mathbb{Q}$-curves and Galois representations”, 2003. COMPLETER

[22] Faltings, G., “The Proof of Fermat's Last Theorem by R. Taylor and A. Wiles”, *Notices of the AMS*, **42**, 7, (1995), 743-746.

[23] Fermat, P. de, *Oeuvres*, (4 vol.), Paris, Gauthier-Villars, 1891-1922.

[24] Frey, G., "Links between stable elliptic curves and certain Diophantine equations", *Ann. Univ. Saraviensis, Ser. Math.*, 1 (1986), 1-40.

[25] Frey, G., "Links between solutions of $A - B = C$ and elliptic curves", *Number Theory, Ulm1987, Proceedings* (H.P. Schlickewei, E. Wirsing, eds), *Lecture Notes in Mathematics*, 1380, Springer-Verlag, New York, 31-62, 1989.

[26] Gelbart, S., "Three Lectures on the Modularity of $\overline{\rho}_{E,3}$ and the Langlands Reciprocity Conjecture", *Modular Forms and Fermat's Last Theorem*, (G. Cornell, J. H. Silverman, G. Stevens eds), New-York, Springer, 1997, 155-207.

[27] Grothendieck, A., Serre, J-P., *Correspondance Grothendieck-Serre*, (P. Colmez, J-P. Serre eds), Société Mathématique de France, Paris, 2001.

[28] Hellegouarch, Y., “Points d’ordre $2p^h$ sur les courbes elliptiques”, *Acta. Arith.*, 26 (1974/75), 253-263.

[29] Hellegouarch, Y., *Invitation to the Mathematics of Fermat-Wiles*. San Diego, Academic Press, 2002.

[30] Igusa, J., I., *Theta functions*, Springer, 1972.

[31] Kleiner, I., “From Fermat to Wiles: Fermat’s Last Theorem becomes a Theorem”, Elem. Math., **55** (2000), 19-37.

[32] Knapp, A., *Elliptic curves*, Princeton University Press, 1992.

[33] Lafforgue, L., “Chtoucas de Drinfeld et correspondance de Langlands”, *Invent. Math.*, 147 (2002) 1-241.

[34] Lang, S., “Some History of the Shimura-Taniyama conjecture”, *Notices of the AMS*, **42**, 11, (1995), 1301-1307.

[35] Mazur, B., “Modular curves and the Eisenstein ideal”, *Publ. Math. IHES*, 47 (1977), 33-186.

[36] Mazur, B., "Deforming Galois representations", *Galois groups over* $\mathbb{Q}$ (Y. Ihara, K. Ribet, J.-P. Serre, eds.), *Math. Sci. Res. Inst. Publ.*, 16, Springer-Verlag, New York, 385-437, 1989.

[37] Mazur, B., "Number Theory as Gadfly", *The American Mathematical Monthly*, **98**, 7, (1991), 593-610.

[38] Mazur, B., "On the passage from local to global in number theory", *Bulletin of the American Mathematical Society*, 29, 1, (1993), 14-50.

[39] Mazur, B., "An Introduction to the Deformation Theory of Galois Representations", *Modular Forms and Fermat's Last Theorem*, (G. Cornell, J. H. Silverman, G. Stevens eds), New-York, Springer, 1997, 243-311.

[40] MFFLT, *Modular Forms and Fermat's Last Theorem*, (G. Cornell, J. H. Silverman, G. Stevens eds), New-York, Springer, 1997.

[41] Murty, R. M., “Selberg’s conjectures and Artin L-functions”, *Bulletin of the AMS*, 31, 1, 1-14, 1994.

[42] Murty, V.K., “Modular elliptic curves”, *Seminar on Fermat’s Last Theorem*, Canadian Math. Soc. Conf. Proc., 17, 1995. COMPLETER.

[43] Oesterlé, J., “Nouvelles approches du théorème de Fermat”, *Séminaire Bourbaki* 694(1987-1988), *Astérisque* 161/162, 165-186, 1988.

[44] Oesterlé, J., “Travaux de Wiles (et Taylor)”, II, *Séminaire Bourbaki*, *Astérisque*, 237 (1996), 333-355. COMPLETER

[45] Petitot, J., Théorème de Fermat et courbes elliptiques modulaires, *XIXth International Congress of History of Science*, Zaragoza, August 1993.

[46] Ribet, K.,. “Galois Representations and Modular Forms”, *Bulletin of the AMS*, Oct. 1995, 375-402.

[47] Ribet, K., “On modular representations of $\mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right)$ arising from modular forms”, *Invent. Math.*, 100, 431-476, 1990.

[48] Riemann, R., *Über die Anzahl der Primzahlen unter einer gegebenen Grösse*, Monatsberichte der Berliner Akademie, 1859. *Gesammelte Werke*, Teubner, Leipzig, 1892. (“On the number of prime numbers less than a given quantity”).

[49] Rosen, M., “Remarks on the History of Fermat's Last Theorem 1844 to 1984”, *Modular Forms and Fermat's Last Theorem*, (G. Cornell, J. H. Silverman, G. Stevens eds), New-York, Springer, 1997, 505-525.

[50] Rohrlich, D. F., “Modular Curves, Hecke Correspondences, *L*-Functions”, *Modular Forms and Fermat's Last Theorem*, (G. Cornell, J. H. Silverman, G. Stevens eds), New-York, Springer, 1997, 41-100.

[51] Rubin, K., Silverberg, A., “A report on Wiles' Cambridge lectures”, *Bulletin of the AMS*, 31, 1, 15-38, 1994.

[52] Rubin, K., "Modularity of Mod 5 Representations", *Modular Forms and Fermat's Last Theorem*, (G. Cornell, J. H. Silverman, G. Stevens eds), New-York, Springer, 1997, 463-474.

[53] Serre, J-P., "Propriétés galoisiennes des points d'ordre fini des courbes elliptiques", *Invent. Math.*, 15 (1972), 259-331.

[54] Serre, J-P., "Sur les représentations modulaires de degré 2 de $\mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right)$", *Duke Math. J.*, 54, 179-230, 1987.

[55] Serre, J-P., "Travaux de Wiles (et Taylor)", I, *Séminaire Bourbaki*, *Astérisque* 237 (1996), 319-332).

[56] Shimura, G., *Introduction to the arithmetic theory of automorphic functions*, Princeton Univ. Press, Princeton, NJ, 1971.

[57] Silverberg, A., "Explicit Families of Elliptic Curves with Prescribed Mod N Representations", *Modular Forms and Fermat's Last Theorem*, (G. Cornell, J. H. Silverman, G. Stevens eds), New-York, Springer, 1997, 447-461.

[58] Silverman, J. H., *The Arithmetic of Elliptic Curves*, New-York, Sringer-Verlag, 1986

[59] Silverman, J.,.Tate, J., *Rational Points on Elliptic Curves*, New York, Springer-Verlag, 1992.

[60] Silverman, J. H., *The Arithmetic of Elliptic Curves II*, New-York, Sringer-Verlag, 1994.

[61] Silverman, J.H., "A Survey of the Arithmetic Theory of Elliptic Curves", *Modular Forms and Fermat's Last Theorem*, (G. Cornell, J. H. Silverman, G. Stevens eds), New-York, Springer, 1997, 17-40.

[62] Smit, B. de, Rubin, K., Schoof, R., “Criteria for Complete Intersections”, *Modular Forms and Fermat’s Last Theorem*, (G. Cornell, J. H. Silverman, G. Stevens eds), New-York, Springer, 1997, 343-355.

[63] Stevens, G., “An Overview of the Proof of Fermat’s Last Theorem”, *Modular Forms and Fermat’s Last Theorem*, (G. Cornell, J. H. Silverman, G. Stevens eds), New-York, Springer, 1997, 1-16.

[64] Taylor, R., Interview with Scott Sheffield, Harvard Math Club.

[65] Taylor, R., “Galois representations”, *Proceedings of the International Congress of Mathematicians 2002*, **1**, 449-474.

[66] Taylor, R., “Galois representations”, *Annales de la Faculté des Sciences de Toulouse*, Série 6, **XIII**, 1, (2004), 73-119.

[67] Taylor, R., Wiles, A., “Ring-theoretic properties of certain Hecke algebras”, *Ann. of Math.* (2) 141, 553–572, 1995.

[68] Tilouine, J., "Hecke Algebras and the Gorenstein Property", *Modular Forms and Fermat's Last Theorem*, (G. Cornell, J. H. Silverman, G. Stevens eds), New-York, Springer, 1997, 327-342.

[69] Vandiver, H. S., *Fermat's Last Theorem and Related Topics in Number Theory*, Ann Arbor, MI, 1935.

[70] Washington, L. C., *Introduction to Cyclotomic Fields*, New-York, Springer, 1997.

[71] Washington, L. C., "Kummer's lemma for prime power cyclotomic fields", *J. Number Theory*, 40, (1992), 165-173.

[72] Washington, L. C., "Galois Cohomology", *Modular Forms and Fermat's Last Theorem*, (G. Cornell, J. H. Silverman, G. Stevens eds), New-York, Springer, 1997, 101-120.

[73] Wiles, A., "Modular elliptic curves and Fermat's Last Theorem", *Ann. of Math.* (2) 141, 443–551, 1995.

[74] Wiles, A., Taylor, R., “Ring-theoretic properties of certain Hecke algebras”, *Annals of Mathematics* (2) **141**, 3, (1995), 553-572.

[75] Zagier, D., “Introduction to Modular Forms”, *From Number Theory to Physics* (M. Waldschmidt, P. Moussa, J-M. Luck, C. Itzykson Eds), 1992, Springer, 238-291.