

7ème Rencontres Françaises de Philosophie des Mathématiques
5-7 novembre 2015, Université Paris-Diderot

The Unity of Mathematics as a Discovery Method

A 1993 talk reworked for the Workshop

Jean Petitot, CAMS (EHESS)

November 2015

Contents

1	Introduction	2
2	Kummer's cyclotomic route	5
2.1	FLT for regular primes	5
2.2	Further advances along the cyclotomic route	8
3	Mordell-Weil-Faltings' general algebraic route	10
4	Hellegouarch and Frey: opening the elliptic route	11
5	The L-function of an elliptic curve	14
6	Riemann's ζ-function and Dirichlet's L-functions	14
7	Modularity	16
7.1	General definition	16
7.2	EC as complex tori	16
7.3	$SL(2, \mathbb{Z})$ action	17
7.4	Modular functions for $SL(2, \mathbb{Z})$	18
7.5	Fourier expansion at infinity, modular forms, and cusp forms	19
7.6	L -functions of cusp forms	20
7.7	The modular curve $X_0(N)$	22

7.8	Modular elliptic curves	23
7.9	New eigenforms and Hecke's algebras	25
8	The core of the proof	25
8.1	“Un extraordinaire carrefour” (Dieudonné)	25
8.2	Two classes of L -functions (Knapp, Murty)	26
8.3	Encoding geometrico-arithmetic information into L -functions	27
9	TSW and Ribet theorem	27
10	Wiles' travel	30
10.1	Encoding information into Galois representations	30
10.2	The Frobenius	33
10.3	Galois representations and L_E -functions	34
10.4	Serre conjecture	35
10.5	A letter of Jean-Pierre Serre to Alexander Grothendieck	37
10.6	Deligne theorem	39
10.7	Langlands-Tunnell theorem	41
10.8	Wiles new route: SSML for $p = 3, 5$	41
10.9	Lifting to p -adic representations	45
10.10	Infinitesimal deformations and cohomology	47
10.11	Deformation data and Mazur conjectures	49
10.12	Complete intersections and Gorenstein property	54
10.13	Selmer groups	56
11	Some developements since 1994.	57
12	Conclusion: the conceptual complexity of a proof	59

1 Introduction

1. Kant used to claim that

“philosophical knowledge is rational knowledge from concepts, mathematical knowledge is rational knowledge from the construction of concepts” (A713/ B741).

As I am rather Kantian, I consider that philosophy of mathematics has to do with “rational knowledge from concepts” in mathematics, that is with the constitutive role of concepts in mathematics.

2. But “concept” in what sense? Well, in the sense introduced by Galois and deeply developped through Hilbert to Bourbaki. Galois said:

“Il existe pour ces sortes d’équations un certain ordre de considérations métaphysiques qui planent sur les calculs et qui souvent les rendent inutiles.”

“Sauter à pieds joints sur les calculs, grouper les opérations, les classer suivant leur difficulté et non suivant leur forme, telle est selon moi la mission des géomètres futurs.”

So, I use “concept” in the *structural* sense. In this perspective, philosophy of mathematics has to do with the dialectic between, on the one hand, logic and computations, and, on the other hand, structural concepts.

3. In mathematics, the context of justification is proof. It has been tremendously investigated. But the context of discovery remains mysterious and is very poorly understood. I think that structural concepts play a crucial role in it.

4. In this general perspective, my purpose is to investigate what could mean “complex” in a *conceptually complex proof*. The only way is to look at a relevant example.

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” 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 Isaac Newton Institute of Cambridge on June 21-23, 1993 at the Conference “*p*-adic Galois representations, Iwasawa theory, and the Tamagawa numbers of motives” organized by John Coates.

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 brought into confidence), but was completed in a joined work with Richard Taylor (September 19, 1994: “I’ve got it!” the English version of Eureka!), sent to some colleagues (including Faltings) on October 6, 1994 and submitted on October 25, 1994. Until 1997, I attended the seminars of Serre and Oesterlé at the IHP and worked a lot to understand this proof and I gave some seminar on it. I will use this stuff today.

In a presentation of the proof, Ram Murty speaks of “Himalayan peaks” that hold the “secrets” of such results. I will use this excellent metaphor. 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. Valleys are “natural” mono-theoretical paths. But, if the conjec-

ture is “hard”, the peak cannot be reached along a valley starting from scratch in the plain. One has to reach first internal “hanging valleys” suspended over lower valleys. This corresponds to the abstraction of an abstract structure. One has also to change valley using saddles, tunnels, passes, canyons. 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* route.

It is this *holistic* nature of the 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 ([102], 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.*” (Our emphasis)

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 ([28], 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 lead to its resolution. (...) It does not seem unnatural to look to differential geometry for progress with this conjecture, or to partial differential equations and the study of the eigenvalue problem for elliptic operators, or to the representation theory of reductive groups... . It would be no surprise if ideas from the classical theory of one complex variable and the Mellin transform were relevant, or of Algebraic Geometry... . But perhaps one should also look in the direction of Kac-Moody algebras, loop groups, or \mathcal{D} -modules, perhaps to ideas that have been, or will be, imported from Physics... .” (Our emphasis).

In the same paper (p. 596), 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 travelling inside the unity of mathematics. Albert Lautman was the first to investigate this problem.

2 Kummer’s cyclotomic route

I will be extremely brief concerning the classical history of FLT. 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.

2.1 FLT for regular primes

In a nutshell we can say that, during what could be called an “Eulerian” period, many particular cases were 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. 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 reminded by Henri Darmon, Fred Diamond and Richard Taylor in their 1996 survey of Wiles [9]:

“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." (p. 4)

It is for this proof that Kummer invented the concept of "ideal" number which will become with Dedekind the founding concept of *ideal* of a ring (the basis of commutative algebra) 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.

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

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

the polynomial

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

is irreducible over \mathbb{Q} and is the minimal polynomial defining ζ ($\Phi(\zeta) = 0$). The conjugates of ζ are $\zeta^2, \dots, \zeta^{l-1}$, $\mathbb{Q}(\zeta)$ is the splitting field of $\Phi(x)$ over \mathbb{Q} and $\mathbb{Q}(\zeta)/\mathbb{Q}$ is a Galois extension. $\mathbb{Z}[\zeta]$ has for \mathbb{Z} -base $1, \zeta, \dots, \zeta^{l-2}$. We note that $\Phi(1) = l$. The prime l is totally ramified in $\mathbb{Z}[\zeta]$ (and in fact is the only ramified prime). More precisely, $(1 - \zeta)$ is a prime ideal of $\mathbb{Z}[\zeta]$, the quotient $\mathbb{Z}[\zeta]$ is the finite field \mathbb{F}_l and there exists some unit u s.t.

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

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

$\mathbb{Z}[\zeta]$ is a unique factorization domain for $l \leq 19$ but not for $l = 23$.

More generally, let K/\mathbb{Q} be an algebraic, finite, Galois extension of degree d and \mathfrak{L} a prime ideal of \mathcal{O}_K , the ring of integers of K , over l (i.e.

$\mathfrak{L} \cap \mathbb{Z} = (l)$, notation $\mathfrak{L} \mid l$). Polynomials irreducible over \mathbb{Q} can become reducible and factorize over K . If (l) splits in \mathcal{O}_K as a product of primes

$$l = \prod_{j=1}^{j=r} \mathfrak{L}_j^{e_j} \text{ with } \mathfrak{L}_j \mid l$$

the exponents e_j are called the *degrees of ramification* of the \mathfrak{L}_j in K/\mathbb{Q} . The extension K/\mathbb{Q} is said *unramified* at \mathfrak{L}_j if $e_j = 1$, and K/\mathbb{Q} is said *unramified* at l if it is unramified at any $\mathfrak{L}_j \mid l$, i.e. if all $e_j = 1$. The residue field $\mathcal{O}_K/\mathfrak{L}_j$ is an algebraic extension of \mathbb{F}_l of degree f_j called the residue degree and we have $\sum_{j=1}^{j=r} e_j f_j = d$.

A very useful geometric intuition is to think of the extension \mathcal{O}_K/\mathbb{Z} as a *geometric projection* of schemes $\text{Spec}(\mathcal{O}_K) \rightarrow \text{Spec}(\mathbb{Z})$ with fibers over the primes l .

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

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

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 $(x + \zeta^j y)$ 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.

The property of regularity is the following:

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 $\mathbb{Z}[\zeta]$.

Characterization. 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 emphasizes Edwards [16], this

“must be regarded as an extraordinary *tour de force*.”

Kummer theorem (1847, paper sent to Dirichlet). A prime l is regular iff it doesn't divide the numerators of any of the *Bernoulli numbers* B_2, B_4, \dots, B_{l-3} . For instance 37 is an irregular prime since 37 divides the numerator of B_{32} .

Introduced in 1713 in *Ars Conjectandi*, Bernoulli numbers are defined by the series

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

or 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 $B_1 = -\frac{1}{2}$, $B_2 = \frac{1}{6}$, $B_3 = 0$, $B_4 = -\frac{1}{30}$, $B_5 = 0$, $B_6 = \frac{1}{42}$, $B_7 = 0$, $B_8 = -\frac{1}{30}$, VERIFIER B8 etc. All the B_n for $n > 1$ odd vanish. A theorem due to Von Staudt and Clausen 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 $pB_{2k} \equiv -1 \pmod{p}$.

Bernoulli numbers are ubiquitous in arithmetics and closely related to the values of Riemann Zeta function 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 $\zeta(2) = \sum_{n \geq 1} \frac{1}{n^2} = \frac{4\pi^2}{2 \cdot 2} B_2 = \frac{\pi^2}{6}$ and in the case $k = 2$, we find $\zeta(4) = \sum_{n \geq 1} \frac{1}{n^4} = -\frac{16\pi^4}{2 \cdot 24} B_4 = \frac{\pi^4}{90}$, 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 $\mathbb{Q}(\zeta + \bar{\zeta})$ ($\bar{\zeta} = \zeta^{-1}$) 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 Bernoulli numbers: $l \mid h^-$ iff l divides the numerators of the Bernoulli numbers B_2, B_4, \dots, 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.

2.2 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 incredibly more

complex irregular case, trying to deepen the knowledge of the structure of cyclotomic fields (see Washington’s book [45] and Rosen’s survey [37]).

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. Tackling the irregular case has been computationally quite inaccessible for a long time and one had to wait until the construction of the first computers before resuming Kummer and Dirichlet’s style.

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 Bernoulli numbers B_i for $i = 2, 4, \dots, l - 3$ are not divisible by l^3 and if $l \nmid h_l^+$ then the second case of FLT is true for l .

The two “cases” of FLT are:

1. the first case is when l is supposed relatively prime to x, y, z .
2. the second case is when l is supposed to divide exactly one of x, y, z (l cannot divide the three since they are relatively prime and cannot divide two of them since it would then divide the three).

Vandiver proposed also a key conjecture:

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

Vandiver began to use such criteria “to test FLT computationally” (Rosen [37], p. 516) 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 his beautiful papers [6], [7], Leo Corry analyzed the computational aspects of FLT after the introduction of computers. In 1949 John von Neumann (1904–1957) 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), the fastest computer of the time (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 ([8])

“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.

3 Mordell-Weil-Faltings’ general algebraic route

One can consider that the natural context of a proof of FLT is general algebraic geometry since Fermat equation

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

is the homogeneous equation of a projective plane curve F . The equation has rational coefficients and FLT says that, for $l \geq 3$, 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 *infinitely many* types of projective plane curves since the genus g of F is

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

and 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 of genus g defined over a number field K and let K'/K be an algebraic extension of the base field K .

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 difference is too huge. We need to take *another route*.

4 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 :

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

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 1994.

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

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

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 $f(x) = x(x - a^l)(x + b^l)$.

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

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

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

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

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

But, if we *reduce* $E \bmod p$ (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 $a^l \equiv 0 \pmod p$ and $b^l \equiv 0 \pmod p$, 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, is

“an arithmetically defined quantity that measures the Diophantine complexity of the associated cubic equation.”¹

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

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

As Δ is proportional to $(a^l b^l c^l)^2$ 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 fundamental invariant of E is the *modular invariant* j defined by

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

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.

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.

¹Darmon [9], p. 1398.

G. Frey was perfectly aware of the originality of his method. In his paper 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.” (Our emphasis.)

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 diversity.
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:

1. Prove that Frey elliptic curves are not modular.
2. 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.

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

5 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 $E \bmod p$ (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 depending on 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 $\mathbb{Q} \subset K \subset \overline{\mathbb{Q}}, \mathbb{R}, \mathbb{C}, \mathbb{F}_p, \mathbb{F}_{p^n}, \overline{\mathbb{F}_p}$.

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.

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

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

In the bad reduction case, we must use $(1 - a_p p^{-s})^{-1}$. 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 $\Re(s) > 3/2$.

6 Riemann’s ζ -function and Dirichlet’s L -functions

To understand the relevance of the L -functions L_E , we have to come back to Riemann’s ζ -function.

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 at 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 note \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 zeta function is a symbolic expression associated to the distribution of primes, which is well known to have a very mysterious structure. Its fantastic strength 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 *real* $s > 1$, secondly s can be extended to *complex* numbers s of real part $\Re(s) > 1$, and thirdly $\zeta(s)$ can be extended by analytic continuation to a meromorphic function on the entire complex plane \mathbb{C} with a pole at $s = 1$.

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

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

and under some conditions on the a_n can be factorized in Euler products

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

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* mod m , that is of a multiplicative morphism

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

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

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).

7 Modularity

7.1 General definition

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. $L_E = L_f$. This implies that there exists, associated to f , an analytic map

$$F : X_0(N) \rightarrow E$$

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

By definition, modular EC have strong arithmetic properties and therefore to say that *all* EC are modular is to say that in spite of the strong irregularity of the distribution of primes, there are highly non trivial constraints and that such constraints imply FLT.

We have to define f , L_f and $X_0(N)$.

7.2 EC 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 a complex torus. There is a completely different way of looking at elliptic curves, the equivalence of the two perspectives being one of the greatest achievements of mathematics in the first half of the XIXth century (Abel, Jacobi, etc.). It belongs to another theory, namely the theory of analytic complex functions. The problem is to study *doubly periodic* analytic functions on the complex plane \mathbb{C} . Let (ω_1, ω_2) be the two periods. We look for analytic functions $f(z)$ such that $f(z + m\omega_1 + n\omega_2) = f(z)$ for all $m, n \in \mathbb{Z}$. As ω_1 and ω_2 cannot be colinear, $\text{Im}(\omega_1/\omega_2) \neq 0$ and changing eventually a sign we can suppose $(\omega_1/\omega_2) > 0$. Let Λ be the lattice $\{m\omega_1 + n\omega_2\}_{m,n \in \mathbb{Z}}$ in \mathbb{C} and E the quotient space $E = \mathbb{C}/\Lambda$; E is a complex torus and f is defined on E . f is called an *elliptic function*. E being compact, f cannot be holomorphic without being constant according to Liouville theorem; f can only be a *meromorphic* function if it is not constant.

Let $E = E_{\text{cub}}$ 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 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 $E_{\text{tor}} = \mathbb{C}/\Lambda$ 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

$$\begin{aligned} \Phi : E_{\text{cub}} &\rightarrow E_{\text{tor}} \\ a &\mapsto \int_{a_0}^a \omega \end{aligned}$$

(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} .

This is the beginning of the great story of *Abelian varieties*.

It is in this context, *where algebraic structures are translated and coded in analytic ones*, that one can develop an extremely deep theory of *arithmetic* properties of elliptic curves. Its “deepness” comes from *the analytic coding of arithmetics*.

7.3 $SL(2, \mathbb{Z})$ action

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 $\text{Im}(\tau) > 0$, that is $\tau \in \mathcal{H}$ where \mathcal{H}

is the Poincaré upper half complex plane. 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 $\{\omega'_2, \omega'_1\}$ 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 invertible and its inverse must therefore be also an integral matrix, so $\text{Det}(\gamma) = ad - bc = 1$ and $\gamma \in SL(2, \mathbb{Z})$. γ acts on τ via Möbius transformations:

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

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

7.4 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 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 $f dz$ is $SL(2, \mathbb{Z})$ -invariant.

To be modular, f has only to be modular on generators of $SL(2, \mathbb{Z})$, two generators being the translation $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ acting by $\tau \rightarrow \tau + 1$

and the inversion $S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$ acting by $\tau \rightarrow -1/\tau$. 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 *automorphy* properties, where. “automorphy” means invariance of a function $f(\tau)$ defined on the Poincaré plane \mathcal{H} relatively to a countable subgroup of the group acting on \mathcal{H} by homographies (also called Möbius transformations) $\gamma(\tau) = \frac{a\tau+b}{c\tau+d}$.

7.5 Fourier expansion at infinity, 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 $\tau = x + iy$, $\kappa = e^{2i\pi x} e^{-2\pi y} \xrightarrow{y \rightarrow \infty} 0$. 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 .

7.6 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.

The Mellin transform was introduced as a fundamental link between Riemann's ζ -function and the Θ -function having strong automorphy properties. It implies the fundamental functional equation satisfied by the ζ -function.

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

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

$\Im(\tau) > 0$ (i.e. $\tau \in \mathcal{H}$) is necessary to warrant the convergence of $e^{-n^2\pi\Im(\tau)}$. $\Theta(\tau)$ is a *modular form* of level 2 and weight $\frac{1}{2}$. Its automorphic symmetries are

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

Now, if $f : \mathbb{R}^+ \rightarrow \mathbb{C}$ 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 $\tilde{\Theta}(it) = \frac{1}{2}(\Theta(it) - 1)$:

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

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*, and 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. 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

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

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^∞ 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 is 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

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

$\Lambda(s)$ being now define 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 $\zeta(s)$.

Paralleling the case of Riemann ζ function for which the function

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

was the Mellin transform of the theta function $\frac{1}{2} (\Theta(it) - 1)$, 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

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

7.7 The modular curve $X_0(N)$

We need to introduce now the *modular curves* $X_0(N)$ of different *levels* N . For $N = 1$, $X_0(1)$ 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 fundamental domain R . It is well known that R is the domain of \mathcal{H} defined by $-2 \leq \Re(\tau) < 2$ 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 $\zeta_3 + 1 = -\zeta_3^2 = \rho = e^{i\frac{\pi}{3}}$.

It can be shown that the field of meromorphic functions $K(X_0(N))$ is generated by the modular invariant j .

The *modular curve* of level N , $X_0(N)$, classifies pairs (Λ, C) of a lattice Λ and a cyclic subgroup C of order N , that is 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 ($1x = x = 0$).

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, ..., 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

$X_0(N)$ is 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 :

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

We note that $\begin{pmatrix} 1 & N \\ 0 & 1 \end{pmatrix} \in \Gamma_0(N)$. Of course $\Gamma_0(1) = SL(2, \mathbb{Z})$. 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: $\beta_j^{-1}(\infty) = -\frac{d_j}{c_j} \in \mathbb{Q}$.

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 (since $\gamma = \begin{pmatrix} 1 & N \\ 0 & 1 \end{pmatrix} \in \Gamma_0(N)$ and $f(\gamma(\tau)) = f(\tau + N) = f(\tau)$) and can be developed at infinity in a Fourier series $f(\gamma(\tau)) = \sum_{n \geq 0} c_n \kappa^n$ with nome $\kappa = e^{\frac{2i\pi\tau}{N}}$.

A further generalization consists in introducing a *character*

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

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

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

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

7.8 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 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:

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

Φ_f is a map $\Phi_f : \Gamma_0(N) \rightarrow \mathbb{C}$ 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

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

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 we make a remark on this definition. We have seen that, as far as it is isomorphic with its Jacobian, a general elliptic curve E defined over \mathbb{C} 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 action of the lattice Λ on \mathbb{C} being commutative while the action of $\Gamma_0(N)$ on \mathcal{H} is non commutative.

“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.”²

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

1. Λ is effectively a lattice in \mathbb{C} ;

²Mazur [28], p. 607. The text was written in 1989 when the *STW* conjecture was still a conjecture.

2. $X_0(N)$, E and F 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: $L_E(s) = L_f(s)$.

7.9 New eigenforms and Hecke's algebras

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 $f(\tau) = \sum_{n \geq 1} c_n \kappa^n$ and not as Euler products. But, by definition, the L_E are Euler product 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 *eigenforms* 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 be expressed as an Euler product.

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: $S_k(N) = S_k^{\text{old}}(N) \oplus S_k^{\text{new}}(N)$.

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

8 The core of the proof

8.1 "Un extraordinaire carrefour" (Dieudonné)

Modularity is the core of the proof because it is an "extraordinaire carrefour" between many theories. In his *Panorama des Mathématiques pures. Le choix bourbachique*, 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.

8.2 Two classes of L -functions (Knapp, Murty)

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} 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 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.”

“Automorphic L -functions have more manageable analytic properties, but they initially have little to do with algebraic number theory or algebraic geometry. The fundamental objective is to prove that motivic L -function are automorphic.”

Ram Murty 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 that every elliptic curve over \mathbb{Q} is ‘modular’ is certainly the most intriguing reciprocity law of our time. The ‘Himalayan peaks’ that hold the secrets of this non abelian reciprocity law challenge humanity.”

8.3 Encoding geometrico-arithmetic information into L -functions

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”.(Knapp)

And as for the zeta function:

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

The situation is well described by A. Knapp:

“We have two kinds of L functions, the kind from cusp forms that we understand very well and the kind from elliptic curves that contains a great deal of information.”

Of course it would be a “miracle” that two L functions belonging to these two completely different classes will be the same. But it is precisely the astonishing result proved by Eichler et Shimura. As Knapp 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 .”

9 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 elliptic curve is isogenous (that is a covering of finite degree) with a modular elliptic curve coming from an $X_0(N)$ and a $f \in S_2^{\text{new}}(N)$ by the Eichler-Shimura construction.

In a Notice of the AMS of November 1995 Serge Lang [25] recalls some elements of the rich story of the conjecture from the pioneering works of Artin, Hasse and Hecke, explains how Taniyama became interested in 1955 by ζ - and L -functions which are Mellin transforms of automorphic forms, how Shimura proved in the late 1950s that the modular elliptic curves have a ζ -function sharing all the good properties one can expect and formulated the conjecture that all elliptic curves defined on \mathbb{Q} are modular. 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 (it is a technical definition: $N_E = \prod_{p|\Delta} p^{n(p)}$ with $n(p) = 1$ if E_p is a node, $n(p) = 2$ if $p > 3$ and E_p is a cusp, $n(p)$ is given by Tate's algorithm for $p = 2, 3$).

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* \rightarrow *HW* is easy since if two elliptic curves defined over \mathbb{Q} are isogeneous over \mathbb{Q} then their L -functions are equal. So E is isogeneous to E' with $L_{E'}(s) = L_f(s)$ for a certain $f \in S_2^{\text{new}}(N)$ and $L_E(s) = L_{E'}(s) = L_f(s)$. The implication *HW* \rightarrow *TSW* is less evident. *HW* implies that $\exists f \in S_2^{\text{new}}(N)$ 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 isogeneous 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

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

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 [35] (p. 16), Ribet gives 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 fact that $S_2(2) = 0$ which 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 elliptic curve 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 : $c_n \equiv c'_n \pmod{l} \forall n \geq 1$.

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 $c_n \equiv c'_n \pmod{l} \forall n \geq 1$. We then apply the lemma:

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 , $S_2(2) = 0$. As $S_2(2) = 0$, we get for $n = 1$ the congruence $(c_1 = 1) \equiv (c'_1 = 0) \pmod{l}$ which is clearly impossible and $TSW \Rightarrow$ Fermat. So under an incredibly complicated travel inside the 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.

10 Wiles' travel

In his reference paper, Wiles summarizes the story of his proof.

“I began working on these problems in the late summer of 1986 immediately on learning of Ribet's result.”

10.1 Encoding information into Galois representations

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

This encoding is particularly interesting for the following reason. Until now we met two definitions of modular elliptic curves defined over \mathbb{Q} . A *geometric* definition: they are quotients of modular curves $X_0(N)$, and an *analytic* definition: they are associated to modular forms. But as was emphasized by Charles Daney ([8], 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 as a complex torus \mathbb{C}/Λ . 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, $\{x_1, x_2\}$ is a \mathbb{Z} -basis of $E[N]$.

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

As the N -torsion points are $\overline{\mathbb{Q}}$ -points, we can consider the extension $\mathbb{Q}(E[N])$ of the base field \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 $\sigma \in \text{Gal}(\mathbb{Q}(E[N])/\mathbb{Q})$ 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*,

$$\bar{\rho}_{E,N} : G = \text{Gal}(\mathbb{Q}(E[N])/\mathbb{Q}) \rightarrow GL_2\left(\frac{\mathbb{Z}}{N\mathbb{Z}}\right) ;$$

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

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

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

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

$$\bar{\rho}_{E,p} : G = \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(\mathbb{F}_p)$$

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

As you know, the “absolute” Galois group $G = \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ is one of the deepest objects of Arithmetics and a lot of work has been devoted to its comprehension.

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 $\text{Gal}(K/\mathbb{Q})$ 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, 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’”

in the sense of the following theorem:

Theorem (Serre). The index of the image $\bar{\rho}_{E,N}(G)$ 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 $E(\overline{\mathbb{Q}})$ and consider the automorphism group

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

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

Theorem. The index of the image $\bar{\rho}_{E,\infty}(G)$ of G in $GL_2(\widehat{\mathbb{Z}})$ 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 $GL_2(\widehat{\mathbb{Z}}) = \text{Aut}(E[\infty])$ is a product of factors corresponding to the different primes:

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

and the representation $\bar{\rho}_{E,\infty} : G = \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(\widehat{\mathbb{Z}})$ is a “product” of $\bar{\rho}_{E,p^\infty} : G = \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(\mathbb{Z}_p)$.

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

Serre proved the theorem:

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

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

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

10.2 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 $\bar{\mathbb{F}}_q$ of \mathbb{F}_q the Frobenius is defined as $\bar{\text{Frob}}_q : x \rightarrow x^q$. It is the generator of the Galois group $\text{Gal}(\mathbb{F}_{q^n}$ or $\bar{\mathbb{F}}_q/\mathbb{F}_q$). 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. $\text{Gal}(\mathbb{F}_{q^n}/\mathbb{F}_q)$ is a cyclic group of order n . It can be *lifted* to a Frobenius Frob_q in the absolute Galois group $\text{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$.

More precisely, 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. $\mathfrak{q} \mid (q)$) are conjugated by the Galois group $\text{Gal}(K/\mathbb{Q})$. If $\mathfrak{q} \mid (q)$, the *decomposition group* of \mathfrak{q} is

$$D_{\mathfrak{q}} = \{\sigma \in \text{Gal}(K/\mathbb{Q}) \mid \sigma \text{ fixes } \mathfrak{q} \text{ (i.e. } \sigma(\mathfrak{q}) = \mathfrak{q})\}$$

At the level of residual fields it is the Galois group of the extension $F_{\mathfrak{q}}/\mathbb{F}_q$ of the residual fields $\mathcal{O}_K/\mathfrak{q} = F_{\mathfrak{q}}$ and $\mathbb{Z}/q\mathbb{Z} = \mathbb{F}_q$. But $F_{\mathfrak{q}}/\mathbb{F}_q$ is a Galois extension whose Galois group is the cyclic group generated by the $\bar{\text{Frob}}_q : x \mapsto x^q$. Let φ be the map $\varphi : D_{\mathfrak{q}} \rightarrow \text{Gal}(F_{\mathfrak{q}}/\mathbb{F}_q)$ which associates to $\sigma \in \text{Gal}(K/\mathbb{Q})$ the induced automorphism $\bar{\sigma}$ of $\mathcal{O}_K/\mathfrak{q}$ (which is well defined since σ fixes \mathfrak{q}). The map φ is surjective and its kernel $\text{Ker}(\varphi) = I_{\mathfrak{q}} \subset D_{\mathfrak{q}}$ is called the *inertia subgroup* of \mathfrak{q} .

$$I_{\mathfrak{q}} = \{\sigma \in D_{\mathfrak{q}} \mid \sigma(x) = x \pmod{\mathfrak{q}}\}$$

The elements of $I_{\mathfrak{q}}$ fix the residue field $\mathcal{O}_K/\mathfrak{q} = F_{\mathfrak{q}}$ and, via the map φ , the quotient $D_{\mathfrak{q}}/I_{\mathfrak{q}}$ becomes isomorphic to the Galois group $\text{Gal}(F_{\mathfrak{q}}/\mathbb{F}_q)$. The cardinal $|I_{\mathfrak{q}}|$ of the inertia subgroup is the *ramification index* e , and the prime q is said *unramified* in K/\mathbb{Q} iff φ is injective, that is iff $I_{\mathfrak{q}} = \{1\}$.

If q is *unramified* in K/\mathbb{Q} then $I_{\mathfrak{q}} = \{1\}$, $D_{\mathfrak{q}} \simeq \text{Gal}(F_{\mathfrak{q}}/\mathbb{F}_q)$ and the unique $\sigma_{\mathfrak{q}} \in D_{\mathfrak{q}} \simeq \text{Gal}(F_{\mathfrak{q}}/\mathbb{F}_q) \subset \text{Gal}(K/\mathbb{Q})$ associated to $\bar{\text{Frob}}_q$ by the inverse isomorphism φ^{-1} is written $\text{Frob}_{\mathfrak{q}}$ and is also called the Frobenius. It generates $D_{\mathfrak{q}}$ and is of order $f_{\mathfrak{q}}$. In \mathcal{O}_K , $\text{Frob}_{\mathfrak{q}}(x) \equiv x^q \pmod{\mathfrak{q}}$. All the different $\text{Frob}_{\mathfrak{q}}$ for $\mathfrak{q} \mid (q)$ are conjugate and they define up to conjugacy a Frobenius $\text{Frob}_q \in \text{Gal}(K/\mathbb{Q})$.

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 ([35], 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 $\overline{\mathbb{Q}}$.”

In that case the decomposition subgroups $D_{\mathfrak{q}}$ and the (big) inertia subgroups $I_{\mathfrak{q}}$ are all subgroups of the *absolute* Galois group $\text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ and are all conjugate in their respective classes. One writes $\text{Frob}_{\mathfrak{q}}$ the $\text{Frob}_{\mathfrak{q}}$, Frob_q being defined up to conjugacy.

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 since N_p is the product of primes $q \neq p$ whose power in the discriminant Δ_E of E is not 0 modulo p .

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 then the p^n -th roots of unity. Their group is of size p^n while $E[p^n]$ 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 $\text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ act therefore trivially.

10.3 Galois representations and L_E -functions

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

remarkable following theorem. For $\sigma \in G$, the image $\bar{\rho}_{E,p}(\sigma)$ 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 $\bar{\rho}_{E,p}$ 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 elliptic curve defined over \mathbb{Q} . Its Galois representation $\bar{\rho}_{E,p}$ satisfies the following properties:

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

This theorem remains valid for the p -adic limit $\rho_{E,p} : G \rightarrow GL_2(\mathbb{Z}_p)$, that is when we lift the residual situation at p to the local situation at p .

The study of this type of representations is a large generalization of *class field theory* which corresponds to the abelianization of $\text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ and to representations of dimension 1.

10.4 Serre 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 $\bar{\rho} : G = \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(\mathbb{F}_p)$ coming from the torsion points of an elliptic curve is modular. The representations $\bar{\rho}_{E,p} : G \rightarrow GL_2(\mathbb{F}_p)$ are continuous and their character $\text{Det } \bar{\rho}_{E,p}$ is odd. It can be shown that they are *absolutely irreducible* in the sense that they are irreducible and $\bar{\rho}_{E,p} \otimes_{\mathbb{F}_p} \overline{\mathbb{F}_p}$ is also irreducible.

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

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

Theorem. Serre conjecture implies Fermat last theorem.

The proof is analog to the previous 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

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

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

1. $\bar{\rho}_{E,l}$ is continuous since it factorizes through $\text{Gal}(K/\mathbb{Q})$ where K is the field generated by (the coordinates of) the l -division points.
2. $\bar{\rho}_{E,l}$ is absolutely irreducible.
3. $\bar{\rho}_{E,l}$ is unramified outside 2 and l and its ramification at l and 2 is, as says Bas Edixhoven ([15], p. 222) “very well behaved”. Indeed for $\bar{\rho}_{E,l}$ 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. $\bar{\rho}_{E,l}$ can be modular.
5. If Serre conjecture is true $\bar{\rho}_{E,l}$ is modular.
6. One shows, it is the difficult part of the proof, that for any f s.t. $\bar{\rho}_{E,l} = \rho_f$ 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 $\bar{\rho}$ and that, for a $\bar{\rho}$ coming from a Frey curve, this Artin conductor is minimal and equal to 2. As was emphasized by Yves Hellegouarch ([23], 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 $\bar{\rho}$ 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 $\bar{\rho}$ to be unramified at the points where E has good reduction. But in our case, $\bar{\rho}$ is also unramified at $p = l$ and $p \mid N$ with $p > 2$ and this is quite extraordinary. As Gerd Faltings formulates it:³

“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 $L_E(s) = \sum_{n \geq 1} \frac{a_n}{n^s}$. One shows first that, for almost every prime p , the Galois representation modulo p , $\bar{\rho}_{E,p}$, can be modular. If Serre conjecture is valid, then they are modular and $\bar{\rho}_{E,p} = \rho_{f_p}$ 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_p can be lifted to characteristic 0 to a modular cusp form $F = \sum_{n \geq 1} A_n \kappa^n$ s.t. $\tilde{F} \equiv f_p \pmod{p}$. But 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. $\tilde{F} = f_p \pmod{p}$ for an *infinite* set P of primes p . Let $q \nmid N$, then E has good reduction. Let $a_q = \text{Trace}(\text{Frob}_q)$. We have $a_q \equiv a_{q,p} \pmod{p}$ for every $q \neq p$ and therefore $A_q = a_q$ in \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 isogeneous over \mathbb{Q} , and E is therefore modular.

10.5 A letter of Jean-Pierre Serre to Alexander Grothendieck

The 31 December 1986, Jean-Pierre Serre wrote a very interesting and touching letter to Alexander Grothendieck 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 $\text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ ” , un travail que j’ai rédigé ces derniers mois, mais qui était en fait en chantier depuis une douzaine d’années. (...)

³Faltings [18], p. 744.

“ 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 *analogie* (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

$$\text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(\mathbb{F}_p)$$

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” .

10.6 Deligne theorem

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

Let $S_k(N, \varepsilon)$ be the space of cusp forms of weight k , level N and character (*Nebentypus*) ε . 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 $f = \sum_{n \geq 1} a_n \kappa^n \in S_k^{\text{new}}(N, \varepsilon)$ 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 $\sim : \mathcal{O}_f \rightarrow \mathbb{F}_p$ 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 completion of \mathcal{O}_f at \mathfrak{p} .

*Deligne theorem.*⁴ Under these hypotheses there exists a (unique) representation $\rho_f : G = \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(\mathcal{O}_{f,\mathfrak{p}})$ 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, that is its quotient $\bar{\rho}_f : G = \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(\mathbb{F}_p)$ defined via the map $\sim : \mathcal{O}_f \rightarrow \mathbb{F}_p$ satisfies for every prime $q \neq p$ and $q \nmid N$

1. $\bar{\rho}_f$ is unramified at q ;
2. $\text{Trace } \bar{\rho}_f(\text{Frob}_q) = \tilde{a}_q$;
3. $\text{Det } \bar{\rho}_f(\text{Frob}_q) = q^{k-1} \widetilde{\varepsilon}(q)$,
4. $\text{Det } \bar{\rho}_f(c) = -1$ where c is the complex conjugation.

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

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,

⁴For $k = 2$, the theorem has been proved before by Eichler and Shimura.

“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.”⁵

As Wiles explains, 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.” (Wiles [48], p. 445).

For 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 $\bar{\rho}_{E,p}$ is modular;
3. for every p , $\bar{\rho}_{E,p}$ 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 isogeneous 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 the interview quoted above) concerning Langlands problem:

“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.”

⁵Adler [1], 2, p. 3.

10.7 Langlands-Tunnell theorem

With Deligne theorem, we can associate to any suitable modular form a Galois representation mod p . But for the converse, there are only *very few results* constructing an eigen cusp form from a Galois representation. The most important one is the fundamental theorem of Langlands and Tunnell concerning Galois representations ρ of $G = \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ 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 $\Gamma_1(N) = \left\{ \gamma \in SL_2(\mathbb{Z}) \mid \gamma \equiv \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix} \pmod{N} \right\}$. Remember that

$$\Gamma_0(N) = \left\{ \gamma \in SL_2(\mathbb{Z}) \mid \gamma \equiv \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \pmod{N} \right\}$$

Langlands-Tunnell theorem. Let $\rho : G = \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(\mathbb{C})$ be a continuous irreducible representation with odd determinant $\text{Det } \rho(c) = -1$ ($c =$ complex conjugation). Suppose that the image $\rho(G)$ is a subgroup of S_4 (fundamental hypothesis of *dihedrality*). 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 $b_q = \text{Trace } \rho(\text{Frob}_q)$.

In our “Himalayan” metaphor, this non trivial theorem could be considered as a sort of forced, narrow and very elevated “mountain pass”.

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

According to Wiles:

“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.” (Wiles [48], p. 444).

An excellent introduction to the first Wiles proof is the text of Karl Rubin and Alice Silverberg [38]. “A report on Wiles’ Cambridge lectures”, *Bulletin of the AMS* (1994).

As emphasized by Charles Daney ([8], 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 2003 [2], what he called “the big picture”:

“An outline to the strategy of the proof has been given. 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 (and so in particular the elliptic curve ones do, implying that Frey’s curves are modular, enough for a contradiction). 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 first key idea of Wiles is to weaken *TSW* by considering it *modulo* p and to relativize it to a *single* prime p . So the strategy is to work modulo p (i.e. with characteristic p) and then to *lift* the results to characteristic 0. The transformed conjecture is called the “semi-stable modular lifting conjecture”.

As pointed out by Kenneth Ribet ([35], p. 18)

“Wiles’s approach to the Taniyama-Shimura conjecture is ‘orthogonal’ to one based on consideration of the varying $\bar{\rho}_{E,p}$.”

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 (p. 446). Instead of looking at all representations $\bar{\rho}_{E,p}$ and try to prove that an infinity of them are modular, Wiles chose to focus on a *single* prime p and to prove that the p -adic *lifting* $\rho_{E,p^\infty} : G = \text{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(\mathbb{Z}_p)$ is modular. This would be sufficient since

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

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

- (a) $\bar{\rho}_{E,p}$ is irreducible,
- (b) E is modular *but only* mod \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), i.e. there exists an eigenform $f \in S_2(N)$, $f = \sum_{n \geq 1} a_n \kappa^n$, satisfying $a_q \equiv q + 1 - \#E(\mathbb{F}_q) \pmod{\mathfrak{p}}$ (*very approximative* equality) for almost every prime q ,

then E is *really modular*, i.e. there exists an 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 conjecture remains highly non trivial since, as was emphasized by Karl Rubin and Alice Silverberg ([38], p. 21)

“There is no known way to produce such a form in general.”

It is why, as explained by Richard Taylor (in his Harvard interview)

“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 semi-stable modular lifting 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:

“Our approach to the study of elliptic curves is via their associated Galois representation. Suppose that $\bar{\rho}_p$ is the representation of $\text{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ on the p -division points of an elliptic curve over \mathbb{Q} , and suppose for the moment that $\bar{\rho}_3$ is irreducible. The choice of 3 is critical because a crucial theorem of Langlands and Tunnell shows that if $\bar{\rho}_3$ is irreducible then it is also modular. We then proceed by showing that under the hypothesis that $\bar{\rho}_3$ is semi-stable at 3, together with some milder restrictions on the ramification of $\bar{\rho}_3$ at the other primes, every suitable lifting of $\bar{\rho}_3$ is modular.” (Wiles [48], p. 444).

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

Sketch of the proof. Let E be defined over \mathbb{Q} and semi-stable and suppose that the semi-stable modular lifting conjecture is true for $p = 3$. Suppose first that the Galois representation $\bar{\rho}_{E,3}$ 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 Langlands-Tunnell.

To construct ρ in our case, we consider $\bar{\rho}_{E,3} : G = \text{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(\mathbb{F}_3)$. It is irreducible by hypothesis. We use the key fact that $GL_2(\mathbb{F}_3)$ can be embedded in $GL_2(\mathbb{C})$ through the well suited morphism ψ which factorizes through $GL_2(\mathbb{Z}[i\sqrt{2}])$ and satisfies

$$\begin{cases} \text{Trace}(\psi(g)) \equiv \text{Trace}(g) \pmod{(1+i\sqrt{2})} \\ \text{Det}(\psi(g)) \equiv \text{Det}(g) \pmod{3} \end{cases}$$

We define explicitly ψ on the generators $\begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix}$ and $\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$ of $GL_2(\mathbb{F}_3)$ 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 $\rho = \psi \circ \bar{\rho}_{E,3} : G = (\bar{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(\mathbb{C})$ is irreducible with odd determinant $\text{Det} \rho(c) = -1$ and that $\text{Im}(\rho) \subseteq PGL_2(\mathbb{F}_3) \simeq S_4$. One can therefore apply Langlands-Tunnel. 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 $b_q = \text{Trace} \rho(\text{Frob}_q)$. From g , one constructs then an eigenform $f \in S_2(N) = S_2(\Gamma_0(N))$ s.t. $\forall n \ a_n \equiv b_n \pmod{\mathfrak{p}}$, where \mathfrak{p} is the prime ideal of $\bar{\mathbb{Q}}$ containing $1 + i\sqrt{2}$. The congruences show that the eigenform f satisfies (b) for the ideal $\mathfrak{p}' = \mathfrak{p} \cap \mathcal{O}_f$ and therefore E is modular.

Suppose now that the representation $\bar{\rho}_{E,3}$ is *reducible*. If the representation $\bar{\rho}_{E,5}$ is also reducible then E is modular. Indeed, the group of points of E over $\bar{\mathbb{Q}}$ contains a cyclic subgroup of order $15 = 3 \cdot 5$ which is G -stable. But the pairs (E, C) are classified by the *rational* points of the modular curve $X_0(15)$. But $X_0(15)$ has only 4 rational points and it can be shown that they all correspond to modular curves.

We can therefore suppose that $\bar{\rho}_{E,5}$ is *irreducible*. In that case, Wiles method is to construct *another* auxiliary elliptic curve E' defined over \mathbb{Q} and semi-stable s.t.

1. $\bar{\rho}_{E',5} = \bar{\rho}_{E,5}$, and
2. $\bar{\rho}_{E',3}$ 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 $q + 1 - \#E'(\mathbb{F}_q) \equiv \text{Trace } \bar{\rho}_{E',5}(\text{Frob}_q) \pmod{5}$. And, as $\bar{\rho}_{E',5} = \bar{\rho}_{E,5}$, we have the congruence

$$\text{Trace } \bar{\rho}_{E',5}(\text{Frob}_q) = \text{Trace } \bar{\rho}_{E,5}(\text{Frob}_q) \equiv q + 1 - \#E(\mathbb{F}_q) \pmod{5}$$

and f satisfies therefore the condition (b) of the semi-stable modular lifting conjecture for $p = 5$. We conclude that E is *modular*.

At this point, the main difficulty is to construct the auxiliary elliptic curve E' . The starting point is that elliptic curves E' satisfying $\bar{\rho}_{E',p} = \bar{\rho}_{E,p}$ are classified by the rational points of the Riemann surface $X(p)$ (defined over \mathbb{Q}) $X(p) = \mathcal{H}/\Gamma(p)$ where

$$\Gamma(p) = \{\gamma \in SL_2(\mathbb{Z}) \mid \gamma \equiv \text{Id} \pmod{p}\}$$

is the subgroup of integral matrices of $SL_2(\mathbb{Z})$ which are congruent to the identity matrix modulo 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 $X(5)$ $\bar{\rho}_{E',3}$ is *irreducible*.

One uses the fact 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 $G = \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ in $GL_2(\mathbb{F}_p)$ is maximal (that is equal to $GL_2(\mathbb{F}_p)$). 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 choose an E' which is “close” to E for the p -adic metric and use 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.

10.9 Lifting to p -adic representations

Up to now, we have considered only representations of $G = \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ into $GL_2(\mathbb{Z}/N\mathbb{Z})$ induced by the N -torsion (N -division) $\overline{\mathbb{Q}}$ -points $E[N] \simeq \frac{\mathbb{Z}}{N\mathbb{Z}} \times \frac{\mathbb{Z}}{N\mathbb{Z}}$ 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*

$$\rho_{E,p} : G = \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(\mathbb{Z}_p)$$

which satisfies the properties:

1. $\text{Det } \rho_{E,p} = \varepsilon_p$ (where ε_p is the cyclotomic character $\varepsilon_p : G \rightarrow \mathbb{Z}_p^\times$),
2. for almost every prime q , $\text{Trace } \rho_{E,p}(\text{Frob}_q) = q + 1 - \#E(\mathbb{F}_q)$ (exact equality).

(of course, through the quotient $\mathbb{Z}_p \rightarrow \mathbb{F}_p$, $\rho_{E,p}$ returns $\bar{\rho}_{E,p}$).

Once again, we will say that a *p-adic representation*

$$\rho : G = \text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(\mathbb{Z}_p)$$

is *modular* if there exists an eigenform $f \in S_2(N)$, $f = \sum_{n \geq 1} a_n \kappa^n$, s.t.

$\text{Trace } \rho(\text{Frob}_q) = a_q$ 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 $\bar{\rho}_{E,p}$ 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* $\rho_1, \rho_2 : G \rightarrow GL_2(\mathbb{Z}_p)$, ρ_1 coming from E/\mathbb{Q} and ρ_2 from a cusp form. We know that their residual representations mod p , $\bar{\rho}_1, \bar{\rho}_2 : G \rightarrow GL_2(\mathbb{F}_p)$, are equal and we want to gather some informations on the spaces of $\rho_3 : G \rightarrow GL_2(\mathbb{Z}_p)$ s.t. $\bar{\rho}_3 = \bar{\rho}_1 = \bar{\rho}_2$. In fact ρ_1 and ρ_2 share more properties than $\bar{\rho}_1 = \bar{\rho}_2$: they are unramified for almost every q (i.e. outside a finite set of “bad” primes). We consider only such representations $\rho : G \rightarrow GL_2(\mathbb{Z}_p)$.

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 a “*point*” and the *local algebra* \mathbb{Z}_p is like a “*germ of deformation*” and therefore a lifting $\bar{\rho} \rightarrow \rho$ 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 , the “*valuation*” of n at the “*point*” p being the power of p in the decomposition of n . To localize the n in the neighborhood 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 residual 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 $p\mathbb{Z}_p$, residual 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 $|x|_p = p^{-v_p(x)}$. 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 $n \geq 1$.

Let us return to the lifting problem. We apply it 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 residual field k . These properties are the “good” properties for a \mathbb{Z}_p -algebra A in this context. We start from a representations $\bar{\rho} : G = \text{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(k)$ and we look for liftings $\rho : G \rightarrow GL_2(A)$ making the following diagram commutative ($i \circ \rho = \bar{\rho} \otimes_k \bar{k}$):

$$\begin{array}{ccc} & & GL_2(A) \\ & \nearrow \rho & \downarrow i \\ G & \xrightarrow{\bar{\rho} \otimes_k \bar{k}} & GL_2(\bar{k}) \end{array}$$

where $i : A \rightarrow \bar{k}$ is a morphism and $\bar{\rho} \otimes_k \bar{k}$ extends the field of scalars from k to \bar{k} .

This is what Wiles called the “*ring theoretic version*” of the problem.

10.10 Infinitesimal deformations and cohomology

Following the geometric analogy, it is natural to ask what can be an “*infinitesimal*” deformation in this algebraic context.

The key idea, introduced a long time ago by Alexander 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. $\varepsilon^2 = 0$.

Let us now proceed naively. Let $\rho : G \rightarrow GL_2(\mathbb{Z}_p)$. Its residual representation $\bar{\rho} : G \rightarrow GL_2(\mathbb{F}_p)$ 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 $\bar{\rho}$ as associating to every $\gamma \in G$ a 2×2 matrix

$$\begin{pmatrix} \sum_{n \geq 0} a_n(\gamma) p^n & \sum_{n \geq 0} b_n(\gamma) p^n \\ \sum_{n \geq 0} c_n(\gamma) p^n & \sum_{n \geq 0} d_n(\gamma) p^n \end{pmatrix} \in GL_2(\mathbb{Z}_p)$$

with the $a_0(\gamma)$, $b_0(\gamma)$, $c_0(\gamma)$, $d_0(\gamma)$ returning $\bar{\rho}$. The Taylor approximations consist in truncating the series at a certain order and in particular the first order linear approximation consists in a representation $\rho^1 : G \rightarrow GL_2(\mathbb{Z}/p^2\mathbb{Z})$ with matrices

$$\begin{pmatrix} a_0(\gamma) + a_1(\gamma)p & b_0(\gamma) + b_1(\gamma)p \\ c_0(\gamma) + c_1(\gamma)p & d_0(\gamma) + d_1(\gamma)p \end{pmatrix} \in GL_2(\mathbb{Z}/p^2\mathbb{Z})$$

where $p^2 = 0$ that is where p is treated as an *infinitesimal* ε .

Compute formally in $\mathbb{F}_p[\varepsilon]$ with $\rho^1 : G \rightarrow GL_2(\mathbb{F}_p[\varepsilon])$. ρ^1 is close to $\bar{\rho}$ and to compare them we write $\rho^1(g)\bar{\rho}(g)^{-1} = 1 + \varepsilon a(g)$ with $a(g) \in M_2(\mathbb{F}_p)$. We consider now the structure of G -module defined by $\bar{\rho}$ 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 $\bar{g} \in GL_2(\mathbb{F}_p)$, the action $\bar{g} * \alpha$ of \bar{g} on α is given by $\bar{g} * \alpha = \bar{g}\alpha\bar{g}^{-1}$ (what is called the adjoint representation). We write then that $\rho^1(g) = (1 + \varepsilon a(g))\bar{\rho}(g)$ is a representation that is $\rho^1(gh) = \rho^1(g)\rho^1(h)$. This imposes drastic conditions on the map $a : G \rightarrow M_2(\mathbb{F}_p)$, namely

$$a(gh) = a(g) + \bar{\rho}(g)^{-1} a(h) \bar{\rho}(g) \quad (**)$$

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 $\bar{\rho} : G \rightarrow GL_2(\mathbb{F}_p)$ and the cohomology group $H^1(G, M_2(\mathbb{F}_p))$. As Barry Mazur ([30], p. 245) explains:

‘‘First-order infinitesimal’’ information concerning the universal deformation ring [see below] attached to a representation $\bar{\rho}$

can be expressed in terms of group cohomology (of the adjoint representation of $\bar{\rho}$). 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 ([47], 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.”

10.11 Deformation data and Mazur conjectures

We have seen how to handle first order infinitesimal deformations of representations $\bar{\rho}_{E,p} : G = \text{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(\mathbb{F}_p)$ and how Galois cohomology enters the stage. Wiles wanted to show *that modularity is a liftable property*: if $\bar{\rho}_{E,p}$ is modular then its p -adic lifting $\rho_{E,p} : G = \text{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \rightarrow GL_2(\mathbb{Z}_p)$ is also modular. His strategy was to make an induction on “Taylor expansions”, that is to lift modularity to the successive n th-order infinitesimal deformations $\rho_{E,p} \bmod n$ and to pass to the limit. As comments Brian Conrad ([5], 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 Wiles 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.”

A deformation data is a pair $\mathfrak{D} = (\Sigma, t)$ where

1. Σ is a *finite* set of “bad” primes q outside of which representations are *unramified*, and
2. t is a set of relevant properties of representations ρ at p (to be “ordinary”, to be “flat”, etc.).

Once again, a representation $\bar{\rho} : G \rightarrow GL_2(k)$ is called \mathfrak{D} -*modular* if there exists an 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 $\bar{\rho}$.

For technical reasons, Wiles need to introduce local conditions which are essentially constraints on the p -adic representations $\rho_{E,p} = \bar{\rho}_{E,p^\infty}$ which lift local constraints on the residual representations $\bar{\rho}_{E,p}$. They essentially mean that ρ is unramified outside Σ and has the same behavior as its residual representation $\bar{\rho}$.

They are remarkably commented by Ken Ribet ([35], 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 $\bar{\rho}_{E,p}$; 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 $\bar{\rho} : G \rightarrow GL_2(k)$ (k an extension of \mathbb{F}_p) be *absolutely* irreducible (that is $\bar{\rho} \otimes_k \bar{k}$ is irreducible) and \mathfrak{D} -modular, then every \mathfrak{D} -lifting of $\bar{\rho}$ to the integer ring \mathcal{O} of a finite extension of \mathbb{Q}_p with residual field k is modular.

Wiles theorem. Mazur 1 \Rightarrow Semi-stable modular lifting conjecture.

Indeed let E be an elliptic curve defined over \mathbb{Q} and semi-stable which satisfies the conditions (a) and (b) of the semi-stable modular lifting conjecture for p and let $\bar{\rho}$ be the representation $\bar{\rho} = \bar{\rho}_{E,p}$. According to hypothesis (a) $\bar{\rho}$ is irreducible. One shows that it is also *absolutely* irreducible. The hypothesis (b) means that $\bar{\rho}$ is modular. It is then possible to find a deformation data \mathfrak{D} with $\Sigma = \{p\} \cup \{q \mid E \text{ has bad reduction at } q\}$ and to show that $\rho_{E,p}$ is a \mathfrak{D} -lifting of $\bar{\rho}$ and that $\bar{\rho}$ 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.” (our emphasis, Rubin-Silverberg [38], p. 26)

The reformulation is done in terms of *universal deformations* for $(\mathfrak{D}, \mathcal{O})$ -deformations, \mathcal{O} being the ring of integers of a finite extension of \mathbb{Q}_p

$$\begin{array}{ccc}
 & & GL_2(A) \\
 & \nearrow \rho & \downarrow i \\
 G & \xrightarrow{\quad} & GL_2(k) \\
 & \searrow \bar{\rho} &
 \end{array}$$

where A is a *local*, Noetherian, complete \mathcal{O} -algebra of residual field k . The concept of universal deformation is then associated to the existence of a very special algebra \mathfrak{R} . The concept of deformation comes from *differential geometry* and extends the analogy between algebra and geometry.

Mazur-Ramakrishna theorem. There exists a *universal* $(\mathfrak{D}, \mathcal{O})$ -lifting $\rho_{\mathfrak{R}} : G \rightarrow GL_2(\mathfrak{R})$ of $\bar{\rho}$, that is for every $(\mathfrak{D}, \mathcal{O})$ -lifting $\rho : G \rightarrow GL_2(A)$ there exists one and only one morphism of algebras $\varphi_{\rho} : \mathfrak{R} \rightarrow A$ s.t. the following diagram is commutative:

$$\begin{array}{ccc}
 & & GL_2(A) \\
 & \nearrow \rho & \downarrow \varphi_{\rho}^* \\
 G & \longrightarrow & GL_2(\mathfrak{R}) \\
 \parallel & \nearrow \rho_{\mathfrak{R}} & \downarrow i \\
 G & \xrightarrow{\bar{\rho}} & GL_2(k)
 \end{array}$$

This fundamental theorem means that the *functor* $A \rightsquigarrow \mathcal{L}(A)$ which associates to every \mathcal{O} -algebra A (they constitute a category) as above the set of liftings of $\bar{\rho} : G \rightarrow GL_2(k)$ to A (they constitute a category) is *representable* by \mathfrak{R} (in the sense of the interpretation of universal problems as representation of functors), and therefore that there exists an isomorphism

$$\mathcal{L}(A) = \text{Hom}_{\text{cont}}(\mathfrak{R}, A)$$

(where $\text{Hom}_{\text{cont}}(\mathfrak{R}, A)$ is the set of *continuous* homomorphisms from \mathfrak{R} to A). It 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 $\bar{\rho}$ 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 $\bar{\rho}$ and $\rho_{f, \mathfrak{p}} \otimes_{\mathcal{O}_f} \mathcal{O}_f$ is a $(\mathfrak{D}, \mathcal{O})$ -lifting of $\bar{\rho}$ 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 $q \in \Sigma$ and a morphism $j : T(N) \rightarrow \mathfrak{T}$ from the standard Hecke algebra $T(N)$ acting on $S_2(N)$ to \mathfrak{T} s.t. \mathfrak{T} is generated over \mathcal{O} by the images $j(T_q)$ of the Hecke operators T_q of $T(N)$ for $q \notin \Sigma$ (i.e. q “good” prime).

T3 There exists a $(\mathfrak{D}, \mathcal{O})$ -lifting of $\bar{\rho}$, $\rho_{\mathfrak{T}} : G \rightarrow GL_2(\mathfrak{T})$, s.t.

$$\text{Trace } \rho_{\mathfrak{T}}(\text{Frob}_q) = j(T_q) \text{ for every “good” prime } q .$$

T4 If ρ is a modular $(\mathfrak{D}, \mathcal{O})$ -lifting of $\bar{\rho}$ 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:

$$\begin{array}{ccc} & & GL_2(\mathfrak{T}) \\ & \nearrow^{\rho_{\mathfrak{T}}} & \downarrow \psi_{\rho}^* \\ G & \longrightarrow_{\rho} & GL_2(\mathfrak{K}) \end{array}$$

As $\rho_{\mathfrak{T}}$ is a $(\mathfrak{D}, \mathcal{O})$ -lifting of $\bar{\rho}$, Mazur-Ramakrishna theorem implies that there exists one and only one morphism of algebras $\varphi : \mathfrak{K} \rightarrow \mathfrak{T}$ s.t. $\rho_{\mathfrak{T}} = \varphi \circ \rho_{\mathfrak{K}}$. The map φ is *surjective* since

$$\forall q \notin \Sigma, \varphi(\text{Trace } \rho_{\mathfrak{K}}(\text{Frob}_q)) = \text{Trace } \rho_{\mathfrak{T}}(\text{Frob}_q) = j(T_q)$$

and the $j(T_q)$ generate \mathfrak{T} by (2).

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. $\varphi : \mathfrak{K} \rightarrow \mathfrak{T}$ is an *isomorphism*.

Theorem. Mazur conjecture 2 implies Mazur conjecture 1.

Sketch of the proof. Let $\bar{\rho} : G \rightarrow GL_2(k)$ be absolutely irreducible and \mathfrak{D} -modular. If ρ is a \mathfrak{D} -lifting of $\bar{\rho}$ to A , we want to show that ρ is modular. We first extend ρ and $\bar{\rho}$ to \mathcal{O} and ρ becomes a $(\mathfrak{D}, \mathcal{O})$ -lifting. Let $\psi_{\rho} : \mathfrak{K} \rightarrow A$ be the morphism of algebras asserted by Mazur-Ramakrishna theorem. If $\varphi : \mathfrak{K} \rightarrow \mathfrak{T}$ is an *isomorphism* we can consider the *inverse* map $\varphi^{-1} : \mathfrak{T} \rightarrow \mathfrak{K}$ and the composed map $\psi = \psi_{\rho} \circ \varphi^{-1} : \mathfrak{T} \rightarrow A$

$$\psi : \mathfrak{T} \xrightarrow{\varphi^{-1}} \mathfrak{K} \xrightarrow{\psi_{\rho}} A$$

We deduce from (T3) that $\psi(T_q) = \text{Trace } \rho(\text{Frob}_q)$ for almost every prime q . Shimura results imply then the existence of an eigenform $f \in S_2(N)$, $f = \sum_{n \geq 1} a_n \kappa^n$, s.t. $a_q = \text{Trace } \rho(\text{Frob}_q) = \psi(T_q)$ for almost every prime q . But this implies that the representation ρ is modular.

10.12 Complete intersections and Gorenstein property

We get universal local algebras \mathfrak{R} associated to deformation data $\mathfrak{D} = (\Sigma, t)$. 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 $\mathcal{L}(A) = \{\text{lifting of } \bar{\rho} \text{ to } A\}$. 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 “tangent space”* at its maximal ideal $\mathfrak{m}_{\mathfrak{R}}$.

The “tangent space” to \mathfrak{R} is $\mathcal{T}_{\mathfrak{R}} = \text{Hom}_k(\mathfrak{R}, k[\varepsilon])$ the set of $k[\varepsilon]$ -points of \mathfrak{R} . It corresponds to first order infinitesimal deformations and its dimension can be computed using Galois cohomology. For every morphism $t : \mathfrak{R} \rightarrow k[\varepsilon]$ the maximal ideal $\mathfrak{m}_{\mathfrak{R}}$ of the local ring \mathfrak{R} maps onto $k[\varepsilon]$ with a kernel containing $\mathfrak{m}_{\mathfrak{R}}^2$ and $\mathcal{T}_{\mathfrak{R}} = \text{Hom}_k(\mathfrak{R}, k[\varepsilon])$ is the dual space of the “cotangent space”

$$\mathcal{T}_{\mathfrak{R}}^* = \frac{\mathfrak{m}_{\mathfrak{R}}}{\mathfrak{m}_{\mathfrak{R}}^2}$$

The problem is now to prove that $\varphi : \mathfrak{R} \rightarrow \mathfrak{T}$ is an *isomorphism*. Surjectivity is “easy”. For injectivity, Wiles introduced a fundamental numerical criterion. The idea is to *bound* the order of “tangent spaces” at prime ideals of \mathfrak{R} .

If $\bar{\rho}$ is \mathfrak{D} -modular, there exists an 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 $\bar{\rho}$. If $\mathcal{O}_f \subset \mathcal{O}$ (with K the field of fractions of \mathcal{O}), then $\rho_{f,\mathfrak{p}} \otimes_{\mathcal{O}_f} \mathcal{O}$ is a $(\mathfrak{D}, \mathcal{O})$ -lifting of $\bar{\rho}$. As the Galois representation $\rho_{f,\mathfrak{p}} \otimes_{\mathcal{O}_f} \mathcal{O}$ is modular by construction, due to the universality property T4, there exists one and only one morphism $\pi_{\mathfrak{T}} : \mathfrak{T} \rightarrow \mathcal{O}$ s.t. the composed map $G \xrightarrow{\rho_{\mathfrak{T}}} GL_2(\mathfrak{T}) \xrightarrow{\pi_{\mathfrak{T}}} GL_2(\mathcal{O})$ satisfies

$$\pi_{\mathfrak{T}} \circ \rho_{\mathfrak{T}} = \rho_{f,\mathfrak{p}} \otimes_{\mathcal{O}_f} \mathcal{O}.$$

Let $\mathfrak{p}_{\mathfrak{T}} = \text{Ker}(\pi_{\mathfrak{T}})$ and

$$\mathfrak{p}_{\mathfrak{R}} = \text{Ker}(\pi_{\mathfrak{R}} \circ \varphi) = \varphi^{-1}(\mathfrak{p}_{\mathfrak{T}}) = \text{Ker}(\pi_{\mathfrak{R}})$$

where $\pi_{\mathfrak{R}}$ is the (unique) map $\pi_{\mathfrak{R}} : \mathfrak{R} \rightarrow \mathcal{O}$ by the universal property of \mathfrak{R} . We have therefore:

$$\begin{array}{ccc} \mathfrak{R} & \begin{array}{c} \xrightarrow{\varphi^{-1}} \\ \xleftarrow{\varphi} \end{array} & \mathfrak{T} \\ \pi_{\mathfrak{R}} \searrow & & \swarrow \pi_{\mathfrak{T}} \\ & \mathcal{O} & \end{array}$$

The property (T2) of \mathfrak{T} (to be generated over \mathcal{O} by the Hecke operators $j(T_q)$ for “good” q) and the fact that, for almost every prime q , $\text{Trace } \rho_{f,p}(\text{Frob}_q) = a_q$ imply that, for almost every prime q , $\pi_{\mathfrak{T}}(T_q) = a_q$.

Now, we use the fact that the cotangent spaces of the schemes $\text{Spec}(\mathfrak{R})$ and $\text{Spec}(\mathfrak{T})$ at the points $\mathfrak{p}_{\mathfrak{R}} = \text{Ker}(\pi_{\mathfrak{R}})$ and $\mathfrak{p}_{\mathfrak{T}} = \text{Ker}(\pi_{\mathfrak{T}})$ are respectively $\frac{\mathfrak{p}_{\mathfrak{R}}}{\mathfrak{p}_{\mathfrak{R}}^2}$ and $\frac{\mathfrak{p}_{\mathfrak{T}}}{\mathfrak{p}_{\mathfrak{T}}^2}$. As explains Barry Mazur ([30], 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.”

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 $\text{Hom}_{\mathcal{O}}(\mathfrak{T}, \mathcal{O})$.

“The turning point in this and indeed in the whole proof came in the spring of 1991. In searching for a clue from commutative algebra I had been particularly struck some years earlier by a paper of Kunz [Ku2]. 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 $\pi_{\mathfrak{T}} : \mathfrak{T} \rightarrow \mathcal{O}$ corresponds therefore 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} :

$$\begin{array}{ccccc} \text{Hom}_{\mathcal{O}}(\mathfrak{T}, \mathcal{O}) & \xrightarrow{\sim} & \mathfrak{T} & \xrightarrow{\pi_{\mathfrak{T}}} & \mathcal{O} \\ & & \pi_{\mathfrak{T}} \mapsto & \xi & \mapsto \pi_{\mathfrak{T}}(\xi) \end{array}$$

Let η be the ideal $(\pi_{\mathfrak{T}}(\xi))$ of \mathcal{O} (η is well defined independently of the isomorphism $\text{Hom}_{\mathcal{O}}(\mathfrak{T}, \mathcal{O}) \simeq \mathfrak{T}$). Wiles gave a sufficient condition for $\varphi : \mathfrak{R} \rightarrow \mathfrak{T}$ 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 $\varphi : \mathfrak{R} \rightarrow \mathfrak{T}$ is an

isomorphism of complete intersection rings. An \mathcal{O} -algebra A is a complete intersection if $A \simeq \mathcal{O}[[T_1, \dots, T_r]] / (f_1, \dots, f_r)$ (as many relations as variables).

Wiles shows $\# \left(\frac{\mathfrak{p}_{\mathfrak{T}}}{\mathfrak{p}_{\mathfrak{R}}^2} \right) \geq \# \left(\frac{\mathcal{O}}{\eta} \right)$ and therefore, if $\# \left(\frac{\mathfrak{p}_{\mathfrak{R}}}{\mathfrak{p}_{\mathfrak{T}}^2} \right) \leq \# \left(\frac{\mathcal{O}}{\eta} \right)$ 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 “tangent 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. Indeed, as Darmon, Diamond and Taylor explain in [9],

“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!

10.13 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 “global” solution over \mathbb{Z} . The best known is *Minkowski’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 fundamental concept is commented in the following way by Barry Mazur ([29], p. 21) for any algebraic variety V :

“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*.

11 Some developements since 1994.

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 (Breuil et al. [3], 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.”

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 $\text{ord}_{s=1} L = r$, 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:

“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.”⁶

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 $\mathbb{Q}(\sqrt{2})$ (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 ([17]), 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.

Another conjecture asserts that if A is an abelian variety over \mathbb{Q} for which $\text{End}(A) \otimes_{\mathbb{Z}} \mathbb{Q}$ is a number field of degree equal to $\dim A$, then there

⁶Darmon [9], p. 1399.

exists an hyperbolic uniformization of A defined over \mathbb{Q} . As explains Ribet ([35], 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.”

Another line of generalization consists in studying higher dimensional representations $\rho : G \rightarrow GL_n(\overline{\mathbb{F}}_p)$ 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).

In what concerns *Serre's modularity conjecture* (which is stronger than the *TSW* conjecture), it has been proved in 2005 by Chandrashekar Khare in the level 1 case, and later in 2008 by Khare and Jean-Pierre Wintenberger.

12 Conclusion: the conceptual complexity of a proof

We have tried to present conceptually the elements of Wiles' proof of the Taniyama-Shimura-Weil conjecture and we emphasized the fact that its main steps consist in translating parts of theories into another theories in order to make explicit and tractable some pieces of information. To say that the proof is not “direct and simple” but “indirect and complex” is to say that the amount of such translational steps is very high.

References

- [1] Adler, A., *Lecture notes on Fermat's Last Theorem*, University of Rhode Island, URI, 1993.
- [2] Boston, N., *The Proof of Fermat Last Theorem*, University of Wisconsin-Madison, 2003.

- [3] 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.
- [4] Carayol, H., “Sur les représentations galoisiennes modulo l attachées aux formes modulaires”, *Duke Math. J.* 59 (1989), 785-801.
- [5] 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.
- [6] Corry, L., Hunting Prime Numbers—from Human to Electronic Computers, The Rutherford Journal, article030105.html
- [7] Corry, L., “Number crunching vs. number theory: computers and FLT, from Kummer to SWAC (1850-1960), *Arch. Hist. Exact Sci.*, 2007.
- [8] Daney, C., *The Mathematics of Fermat’s Last Theorem*, 1996. <http://cgd.best.vwh.net/home/flt/flt01.htm>
- [9] Darmon, H., Diamond, F., Taylor, R.L., “Fermat’s Last Theorem”, in *Current Developments in Mathematics*, International Press, Cambridge, MA, 1996, 1-154.
- [10] Deligne, P., “Formes modulaires et représentations l -adiques”, *Séminaire Bourbaki*, 1968-1969, Exposé 355, *Lect. Notes in Maths.* 179 (1971), 139-172.
- [11] Deligne, P., Serre, J-P., “Formes modulaires de poids 1”, *Ann. Sci. ENS*, 7 (1974), 507-530.
- [12] Diamond, F., “On deformations rings and Hecke rings”, *Ann. of Maths.*, 144, 1 (1996) 137-166.
- [13] Diamond, F., “The Taylor-Wiles construction and multiplicity one”, *Invent. Math.*, 128 (1997) 379-391.
- [14] Dieudonné, J., *Panorama des Mathématiques pures. Le choix bourbachique*, Paris, Gauthier-Villars, 1977.
- [15] 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.
- [16] Edwards, H. M., *Fermat’s Last Theorem: A Genetic Introduction to Algebraic Number Theory*, New-York, Springer, 1977.

- [17] Ellenberg, J., “ \mathbb{Q} -curves and Galois representations”, 2003. *Modular Curves and Abelian Varieties*, Progress in Mathematics, 224, 93-103, Birkhäuser Verlag, Basel, 2004.
- [18] Faltings, G., “The Proof of Fermat’s Last Theorem by R. Taylor and A. Wiles”, *Notices of the AMS*, **42**, 7, (1995), 743-746.
- [19] Frey, G., “Links between stable elliptic curves and certain Diophantine equations”, *Ann. Univ. Saraviensis, Ser. Math.*, 1 (1986), 1-40.
- [20] 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.
- [21] Grothendieck, A., Serre, J-P., *Correspondance Grothendieck-Serre*, (P. Colmez, J-P. Serre eds), Société Mathématique de France, Paris, 2001.
- [22] Hellegouarch, Y., “Points d’ordre $2p^h$ sur les courbes elliptiques”, *Acta. Arith.*, 26 (1974/75), 253-263.
- [23] Hellegouarch, Y., *Invitation to the Mathematics of Fermat-Wiles*. San Diego, Academic Press, 2002.
- [24] Knapp, A., *Elliptic curves*, Princeton University Press, 1992.
- [25] Lang, S., “Some History of the Shimura-Taniyama conjecture”, *Notices of the AMS*, **42**, 11, (1995), 1301-1307.
- [26] Mazur, B., “Modular curves and the Eisenstein ideal”, *Publ. Math. IHES*, 47 (1977), 33-186.
- [27] 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.
- [28] Mazur, B., “Number Theory as Gadfly”, *The American Mathematical Monthly*, **98**, 7, (1991), 593-610.
- [29] Mazur, B., “On the passage from local to global in number theory”, *Bulletin of the American Mathematical Society*, 29, 1, (1993), 14-50.
- [30] 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.

- [31] Oesterlé, J., “Nouvelles approches du théorème de Fermat”, *Séminaire Bourbaki* 694 (1987-1988), *Astérisque* 161/162, 165-186, 1988.
- [32] Oesterlé, J., “Travaux de Wiles (et Taylor)”, II, *Séminaire Bourbaki*, *Astérisque*, 237 (1996), 333-355.
- [33] Murty, R. M., “Selberg’s conjectures and Artin L -functions”, *Bulletin of the AMS*, 31, 1, 1-14, 1994.
- [34] Murty, V.K., “Modular elliptic curves”, *Seminar on Fermat’s Last Theorem*, Canadian Math. Soc. Conf. Proc., 17, 1995.
- [35] Ribet, K., “Galois Representations and Modular Forms”, *Bulletin of the AMS*, Oct. 1995, 375-402.
- [36] Ribet, K., “On modular representations of $\text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ arising from modular forms”, *Invent. Math.*, 100, 431-476, 1990.
- [37] 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.
- [38] Rubin, K., Silverberg, A., “A report on Wiles’ Cambridge lectures”, *Bulletin of the AMS*, 31, 1, 15-38, 1994.
- [39] Serre, J-P., “Propriétés galoisiennes des points d’ordre fini des courbes elliptiques”, *Invent. Math.*, 15 (1972), 259-331.
- [40] Serre, J-P., “Sur les représentations modulaires de degré 2 de $\text{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ ”, *Duke Math. J.*, 54, 179-230, 1987.
- [41] Serre, J-P., “Travaux de Wiles (et Taylor)”, I, *Séminaire Bourbaki*, *Astérisque* 237 (1996), 319-332).
- [42] Silverman, J., Tate, J., *Rational Points on Elliptic Curves*, New York, Springer-Verlag, 1992.
- [43] Vojta, P., “Diophantine Approximations and Value Distribution Theory”, *Lec. Notes in Math.*, 1239, Springer, 1989.
- [44] Taylor, R., Wiles, A., “Ring-theoretic properties of certain Hecke algebras”, *Ann. of Math. (2)* 141, 553–572, 1995.
- [45] Washington, L. C., *Introduction to Cyclotomic Fields* , New-York, Springer, 1997.

- [46] Washington, L. C., “Kummer’s lemma for prime power cyclotomic fields”, *J. Number Theory*, 40, (1992), 165-173.
- [47] 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.
- [48] Wiles, A., “Modular elliptic curves and Fermat’s Last Theorem”, *Ann. of Math. (2)* 141, 443–551, 1995.