# The unity of mathematics as a method of discovery

*(long version of a talk given at the 7th French Philosophy of Mathematics Workshop 5-7 November 2015, University of Paris-Diderot)*

Jean Petitot
CAMS (EHESS), Paris

# 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 will consider here that philosophy of mathematics has to do with "rational knowledge from 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 best 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"[45] in a workshop organized by Marco Panza. It was about the recent (quasi)-proof of the Taniyama-Shimura-Weil conjecture (*TSW*) presented by Andrew Wiles in three lectures "Modular forms, elliptic curves, and Galois representations" at the Conference "$p$-adic Galois representations, Iwasawa theory, and the Tamagawa numbers of motives" organized by John Coates. at the Isaac Newton Institute of Cambridge on June 21-23, 1993.

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

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

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

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

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

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

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

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

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

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

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

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

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

In the same paper, Mazur adds:

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

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

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

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

# Kummer's cyclotomic route

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

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

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

# FLT for regular primes

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

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

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

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

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

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

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

The polynomial

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

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

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

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

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

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

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

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

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

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

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

and we have

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

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

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

These numbers are linked by a fundamental relation:

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

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

$$ref = d \ .$$

Three cases are particularly interesting:

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

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

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

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

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

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

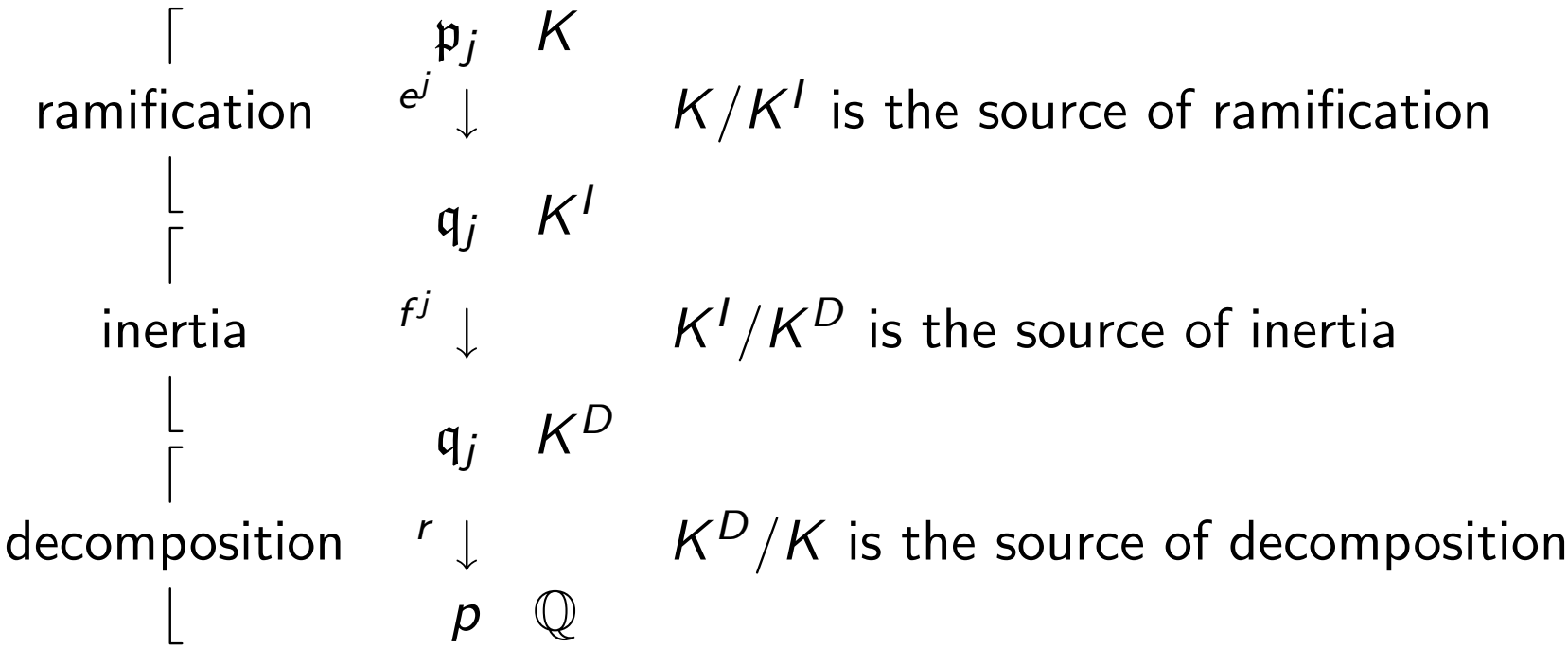
$K^D$ is called the *decomposition field* of $p$ in $K$.

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

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

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

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


$\mathfrak{p}_j$ $K$
ramification
$e^j \downarrow$
$K/K^I$ is the source of ramification
$\mathfrak{q}_j$ $K^I$
inertia
$f^j \downarrow$
$K^I/K^D$ is the source of inertia
$\mathfrak{q}_j$ $K^D$
decomposition
$r \downarrow$
$K^D/K$ is the source of decomposition
$p$ $\mathbb{Q}$

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

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

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

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

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

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

*Case I*

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

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

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

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

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

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

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

$$x + \zeta y = u t^l.$$

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

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

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

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

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

*Case II*

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

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

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

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

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

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

Bernouilli numbers are defined by the series

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

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

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

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

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

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

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

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

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

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

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

# Further advances along the cyclotomic route

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

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

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

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

Vandiver proposed also a key conjecture:

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

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

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

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

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

As was emphasized by Charles Daney [11]

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

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

# Faltings theorem and the Mordell-Weil conjecture

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

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

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

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

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

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

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

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

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

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

# Hellegouarch and Frey: opening the elliptic route

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

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

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

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

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

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

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

A Frey curve $E$ is given in the Weierstrass form:

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

Its *discriminant* is given by the general formula:

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

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

$$\Delta = 16\left(a^{\prime}b^{\prime}c^{\prime}\right)^2$$

$E$ 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 $a^l \equiv 0 \bmod p$ and $b^l \equiv 0 \bmod p$, and so we cannot have a triple root.

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

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

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

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

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

(2 which divides $\Delta$ divides also $abc$ since one of the $a, b, c$ is even).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(a) Prove that Frey ECs are not modular.

(b) Prove the Taniyama-Shimura-Weil conjecture.

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

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

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

# The $L$-function of an elliptic curve

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

$E$ has an infinity of points over $\mathbb{C}$ (but can have no points on $\mathbb{Q}$). However, if we reduce $E \bmod p$ ($p$ 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 $N_p$ relative to the different primes $p$.

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

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

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

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

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

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 $\Re(s) > 3/2$.

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

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

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

And as for Riemann's Zeta function:

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

# Riemann's $\zeta$-function

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

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

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

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

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

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

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

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

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

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

# Theta function and Mellin transform

The zeta function encodes very deep arithmetic properties.

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

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

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

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

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

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

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

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

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

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

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

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*.

Therefore

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

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

# Functional equation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

# Dirichlet's $L$-functions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

# Modularity

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

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

We have to define $f$ and $L_f$.

# ECs as complex tori

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

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

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

Let $\gamma_1$ and $\gamma_2$ be two loops corresponding to a parallel and a meridian of $E$ (they constitute a $\mathbb{Z}$-basis of the first integral homology group $H_1(E, \mathbb{Z})$).

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

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

$$\begin{array}{rccc} \Phi : & E_{\mathrm{cub}} & \to & E_{\mathrm{tor}} \\ & a & \mapsto & \int_{a_0}^{a} \omega \end{array}$$

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

*Theorem*. $\Phi$ is an *isomorphism* between $E_{\mathrm{cub}}$ and $E_{\mathrm{tor}}$.

# Elliptic functions

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

which discriminant is:

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

the lattice $\Lambda$ corresponding to the point at infinity in the $y$ direction.

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

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

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

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

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

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

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

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

$\gamma$ acts on $\tau$ via fractional linear Möbius transformations:

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

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

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

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

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

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

Hence the general definition:

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

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

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

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

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

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

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

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

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

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

# Fourier expansion, modular forms, and cusp forms

The fact that a modular function $f$ is invariant by the translation $\tau \longrightarrow \tau + 1$ 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 $\mathcal{H} \longrightarrow \mathbb{D} - \{0\}$, $\tau \mapsto \kappa = e^{2i\pi\tau}$ which uniformizes $\mathcal{H}$ at infinity in the sense that, if $\tau = x + iy$, $\kappa = e^{2i\pi x} e^{-2\pi y} \underset{y \to \infty}{\longrightarrow} 0$. The boundary $y = 0$ of $\mathcal{H}$ maps cyclically on the boundary $\mathbb{S}^1 = \partial\mathbb{D}$ of $\mathbb{D}$.

If we use this representation, the second property of modularity

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

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

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

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

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

Eisenstein series

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

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

The discriminant $\Delta$ of elliptic curves,

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

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

It expands into

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

One can show that it is given by the infinite product

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

We have seen that

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

It is a general fondamental fact:

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

# $L$-functions of cusp forms

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

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

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

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

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

Paralleling the case of Riemann $\zeta$ function, we introduce the Mellin transform

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

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

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

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

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

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

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

# The modular curve $X_0(1)$

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 fondamental domain $R$.

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

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

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

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

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

# Elliptic points

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

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

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

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

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

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

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

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

In $\Gamma_1(N)$ we have moreover $a, b \equiv 1 \mod N$.

We note that $\begin{pmatrix} 1 & N \\ 0 & 1 \end{pmatrix} \in \Gamma_0(N)$. Of course $\Gamma_0(1) = SL(2,\mathbb{Z})$.

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

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

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

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

# The genus of $X_0(N)$

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

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

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

The remarkable general formula is a sum of four terms:

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

with

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A further generalization consists in introducing a *character*

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

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

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

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

# The Jacobian $J_0(N)$

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

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

where $\tau_0$ is a base point on $X_0(N)$. $\Theta(\tau)$ is well defined modulo the lattice $\Lambda(X_0(N))$ generated over $\mathbb{Z}$ by the $2g$ points of $\mathbb{C}^g$

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

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

# Modular elliptic curves

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

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

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

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

where $\tau_0$ is a base point in $\mathcal{H}$.

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

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

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

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

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

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

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

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

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

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

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

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

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

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