A talk written at the
Swiss Federal Institute of Technology In Zurich
on the topic of

Fermat's Last Theorem

by Clara Invernizzi and Ryan Rueger
for the seminar *Elliptic Curves* held by
Dr. M. Schwagenscheidt in
Zurich, November 2020

Introduction

In this talk we will discuss a famous theorem of number theory which was only a conjecture for over 300 years. We will first introduce where does the conjecture comes from and then we will understand the main ideas of the proof following Kramer's paper[4].

Pierre de Fermat was born the 26th of August in 1601 in France, he was a lawyer in Toulouse but in his free time he studied mathematics. In particular, he studied Diophantus' book *Arithmetica* and wrote a lot of marginal notes in his copy of Diophantus' book. Most of his notes were observations without rigorous proofs. Five years after Fermat's death in 1665, his son published a new version of Diophantus'*Arithmetica* containing Fermat's notes and in the next years many mathematicians contributed to rigorously prove all the observations he did.

The theorem we will talk about is the one that Fermat stated after reading the part of Diophantus's book where the *Pythagorean triples*, i.e. triples (a, b, c) of integer numbers satisfying $a^2 + b^2 = c^2$, were studied. Diophantus already knew a method to construct these triples and hence showed that there are infinitely many Pythagorean triples. In fact, one can find all *primitive Pythagorean triples*, which are positive Pythagorean triples (a, b, c) satisfying $\gcd(a, b, c) = 1$ and a is even, and then construct any other solution to $a^2 + b^2 = c^2$ by changing signs, permuting a, b and by multiplying with some non-zero integer. We can summarize the construction of the primitive Pythagorean triples in the following theorem[2, p.36]

Theorem 1. *A triple of positive integers (a, b, c) is a primitive Pythagorean triple if and only if there exist two integers x, y with $x > y > 0$, $\gcd(x, y) = 1$ and with different parity such that*

$$a = 2xy$$
$$b = x^2 - y^2$$
$$c = x^2 + y^2.$$

Fermat asked himself how many such triples of solutions exists if one substitutes the exponent 2 with an integer $n \geq 3$. After his attempts to find such triples, he stated the famous theorem:

Theorem 2 (Fermat's Last Theorem)**.** *For every integer $n \geq 3$, there exist no three integers a, b, c with $abc \neq 0$ satisfying:*

$$a^n + b^n = c^n.$$

Fermat also wrote in the margin that he had a proof for this claim, but there was not enough space to write it there. However, no one found a proof for this claim until 1995, so the name of the theorem became "Fermat's Last Theorem" because it was the only claim from Fermat's notes that were still not proven or contradicted.

From 1637 to 1980

First we should note that by prime factorization it is enough to prove Fermat's Last Theorem for every prime exponent $p > 2$ and for the exponent $n = 4$. Indeed, if $n > 2$ is an integer, then either $n = pl$ for a prime $p > 2$ and some $l \in \mathbb{N}$ or $n = 4l$ for some $l \in \mathbb{N}$. In both cases we can apply the following remark:

Remark. If Fermat's Last Theorem holds for $p \in \mathbb{N}$ then it holds also for $n = pl$, for every $l \in \mathbb{N}$.

Proof. We will prove it by contraposition. Suppose that Fermat's Last Theorem does not hold for $n = pl$. Then there exists non-zero integers a, b, c such that $a^n + b^n = c^n$. But this means that $(a^l)^p + (b^l)^p = (c^l)^p$. Since a^l, b^l, c^l are non-zero integers, it follows that the theorem does not hold for p too. □

In Fermat's notes one could find only a proof for the case in which the exponent is 4, in what follows we will explain the procedure of his proof[2, Lecture II].

Theorem 3. *There exist no integers a, b, c with $abc \neq 0$ such that*

$$a^4 + b^4 = c^2. \tag{1}$$

Note that if this theorem holds, then also Fermat's Last Theorem for $n = 4$ must hold: suppose that there exists $a, b, c \in \mathbb{N}$ such that $a^4 + b^4 = c^4$, then setting $x = c^2$, it follows that $a^4 + b^4 = x^2$ and it would be a contradiction to Theorem 3.

In the proof of Theorem 3 we will use the infinite descent method which consists in assuming to have a integer solution to the problem and then showing that there exists a smaller solution. Since the same argument can be repeated infinitely many times, this leads to an infinite decreasing sequence of positive integers $a_0 > a_1 > ... > 0$ which is absurd.

Proof of Theorem 3. Assume by contradiction that there exists $a, b, c \in \mathbb{N}$ satisfying equation (1) and without loss of generality we can assume that $\gcd(a, b, c) = 1$. One of the three integers must be even, so we can assume that a is even. We then know that (a^2, b^2, c) is a Pythagorean triple and by Theorem 1 there exists integers x, y of different parity with $x > y > 0$ and $\gcd(x, y) = 1$ such that

$$a^2 = 2xy \tag{2}$$
$$b^2 = x^2 - y^2 \tag{3}$$
$$c = x^2 + y^2. \tag{4}$$

By (3) it follows $x^2 = b^2 + y^2$ and $\gcd(x, b, y) = 1$. Hence using again Theorem 1, there exists integers $w, z \in \mathbb{N}$ of different parity with $w > z > 0$ and $\gcd(w, z) = 1$ such that

$$y = 2wz \tag{5}$$
$$b = w^2 - z^2 \tag{6}$$
$$x = w^2 + z^2. \tag{7}$$

By (2), (5) and (7), we get that $a^2 = 2xy = 4wz(w^2 + z^2)$. Since $w, z, (w^2 + z^2)$ are pairwise coprime and by uniqueness of the factorization into primes, it follows that $w, z, w^2 + z^2$ are squares of positive integers $e, f, g \in \mathbb{N}$:

$$w = e^2$$
$$z = f^2$$
$$w^2 + z^2 = g^2$$

with the property $e^4 + f^4 = g^2$. Since $c = x^2 + y^2 = (w^2 + z^2)^2 + 4w^2z^2 > g^4 > g > 0$, we found a new triple (e, f, g) satisfying (1) with g strictly smaller than c. Hence we can construct infinitely many triples of positive integers (e_i, f_i, g_i) with $g_i > g_{i+1}$ satisfying (1), but this leads to a contradiction. □

Using the same method of infinite descent, Leonhard Euler (1707-1783) proved the case for $n = 3$ and Peter Gustav Lejeune Dirichlet (1805-1859) proved the case for $n = 5$ in 1825. Independently from Dirichlet, Adrien-Marie Legendre (1752-1833) found another proof for the exponent $n = 5$ in the same year. Another relevant proof came by Ernst Eduard Kummer (1810-1893) who proved the theorem for all prime exponents smaller than 100, except for $37, 59, 67$. In 1976 the theorem was proven for all prime exponents smaller than 125.000, hence the proof was still not complete.

A more rigorous approach is required: A link to Modular Forms

Until now, much of the work on Fermat's Last Theorem was very manual. As evidenced by the 300 years of fruitless work by an array of established heavyweights in the mathematical community, more theory needed to be developed before any meaningful progress could be made on the proof as a whole.

To do this, we will revisit cusp forms in a slightly more generalised form. These generalised cusp forms will lead us to the definition of modular elliptic curves which are the object of the Taniyama-Shimura conjecture; the final theorem that remained to be proven to show Fermat's Last Theorem.

In short, a modular elliptic curve is a curve that has a certain special associated (generalised) cusp form. The conditions that this (generalised) cusp form must fulfill can be read off from its Fourier coefficients.

As a quick reminder, we will give the definition of a

Definition 4 (Cusp form)**.** A *cusp form* ζ of integral *weight* k is a holomorphic function defined on the upper complex half plane $\mathbb{H}$ with a Fourier expansion

$$\zeta(\tau) = \sum_{n=1}^{\infty} c_n e^{2\pi \mathrm{i}\, \tau}$$

that satisfies the relation

$$\zeta(G\tau)(G_{21}\tau + G_{22})^{-k} = \zeta(\tau)$$

for all fractional linear transformations[1] $G = (G_{11}, G_{12}; G_{21}, G_{22})$ in $\Gamma = \mathrm{SL}(2;\mathbb{Z})$. In the previous lecture, we introduced the handy notation $\zeta|_k G = \zeta(G\tau)(G_{21}\tau + G_{22})^{-k}$ to denote this behaviour. We say that such a cusp form is *normed* if $c_1 = 1$.

We can now generalise this definition by replacing Γ with an arbitrary subgroup Π of Γ, to define a

Definition 5 (Generalised cusp form)**.** Let Π be a subgroup of the modular group Γ. A *generalised cusp form* ζ *with respect to* Π of integral *weight* k is a holomorphic function defined on the upper complex half plane $\mathbb{H}$ with a Fourier expansion

$$\zeta(\tau) = \sum_{n=1}^{\infty} c_n e^{2\pi \mathrm{i}\, \tau}$$

[1] That is, the action defined as $\tau \mapsto (G_{11}\tau + G_{12})/(G_{21}\tau + G_{22})$ whereby G_{ij} is the ij-th entry of the integral 2×2 matrix G.

that satisfies the relation

$$\zeta|_k P = \zeta$$

for all fractional linear transformations P in Π. Again, we say that such a generalised cusp form is *normed* if $c_1 = 1$.

Remark. It should be noted that some authors simply call our "generalised cusp forms" "cusp forms" making no distinction between the two. These authors generally rely on the convention, that not mentioning which subgroup the (generalised) cusp forms are respective to implies that they are talking about (what we initially called) a cusp form.

For a fixed weight k, the cusp forms of weight k form a complex vector space which we denote $\mathbb{S}_k$ (the "$\mathbb{S}$" originates from the German word for cusp form: "*Spitzenform*"). In the last talk, the dimension of $\mathbb{S}_k$ was calculated to be

$$\dim_{\mathbb{C}}(\mathbb{S}_k) = \begin{cases} \lfloor \frac{k}{12} \rfloor - 1 & \text{if } k \equiv 2 \pmod{12} \\ \lfloor \frac{k}{12} \rfloor & \text{if } k \not\equiv 2 \pmod{12}. \end{cases}$$

Similarly, for a fixed weight k, the generalised cusp forms of weight k with respect to the subgroup $\Pi \leq \Gamma$ form a complex vector space which we will denote $\mathbb{GS}_k(\Pi)$.

Unfortunately for us though, we cannot say anything about the dimension of $\mathbb{GS}(\Pi)$ in general since the smaller the subgroup Π is, the less conditions we impose on the holomorphic functions that constitute $\mathbb{GS}(\Pi)$. For example, setting $\Pi = \{\mathrm{id}_{2\times 2}\}$, we see that the requirement $\zeta|_k \mathrm{id}_{2\times 2} = \zeta$ is trivially satisfied by all functions ζ and as such doesn't impose any restrictions.

To better our prospects of finding any interesting results we will restrict our focus to a specific subgroup of the modular group

$$\Gamma_0(N) = \{M = (M_{11}, M_{12}; M_{21}, M_{22}) \in \Gamma \mid M_{21} \equiv 0 \pmod{N}\}.$$

This action, when applied to the upper half plane gives rise to an open Riemann surface, a special kind of complex manifold. Before we talk more about this, we will finally define a

Definition 6 (Modular elliptic curve)**.** An elliptic curve $E = V_{\mathbb{P}^2}(Y^2 - f(X))$ is said to be *modular of the level* N, if every prime p coprime to N results in a good reduction $E(\mathbb{F}_p)$ of $E = E(\mathbb{Q})$ and there exists a non-zero normed generalised cusp form ζ of weight

2 with respect to $\Gamma_0(N)$, such that the p-th Fourier coefficient satisfies

$$c_p = p - |E(\mathbb{F}_p)| = p - \left| \left\{ (x,y) \in {\mathbb{F}_p}^2 \mid y^2 - f(x) = 0 \right\} \right|.$$

Somewhat more succinctly: an elliptic curve E is said to be *modular of the level* N if every prime p coprime to N results in a good reduction of E and there exists a generalised cusp form in $\mathbb{GS}_2(\Gamma_0(N))$ whose p-th Fourier coefficient c_p is equal to $p - |E(\mathbb{F}_p)|$.

At first, this definition seems like it excludes a lot of elliptic curves. However, in 1957 it was conjectured by Yutaka Taniyama and Goro Shimura that in fact *all* elliptic curves are modular. Later work by André Weil[2] shed light on what the level of a given elliptic curve should be, and that it could be determined quite easily from the curve. Before we talk about the level in more detail, let us summarise a slightly vague version of the

Theorem 7 (Modularity Theorem)**.** *Every (rational) elliptic curve E is modular to the level $N = N(E)$.*

Initially, this theorem was called the "Taniyama-Shimura conjecture", and after Andrew Wiles proved the theorem it became the 'Taniyama-Shimura-Wiles theorem". Perhaps as an homage to the fact that many mathematicians worked on this theorem: providing motivation, supplementary proofs and insights...

So, let us try to gather some results about the generalised cusp forms of weight two with respect to $\Gamma_0(N)$. However, to do this, we must first garner a better understanding of the group $\Gamma_0(N)$ and its action on $\mathbb{H}$.

The quotient $\Gamma_0(N)\backslash\mathbb{H}$ is an open Riemann surface, that is, an open one-dimensional complex manifold. We can make this space compact (since it is Tychonoff) and denote it $\overline{\Gamma_0(N)\backslash\mathbb{H}}$. This is still a Riemann surface and we can introduce a further geometric constant, namely the

Definition 8 (Geometric Genus)**.** The *geometric genus* of a complex manifold M of dimension k is the dimension of the space of holomorphic k-forms on M.

Moreover, we know that the (geometric) genus of $\overline{\Gamma_0(N)\backslash\mathbb{H}}$ is given by

$$g(N) = g_N = \frac{N+1}{12} - \frac{1}{4}\left(1 + \left(\!\!\left(\frac{-1}{N}\right)\!\!\right)\right) - \frac{1}{3}\left(1 + \left(\!\!\left(\frac{-3}{N}\right)\!\!\right)\right)$$

where $(\!(a/p)\!)$ denotes the Legendre symbol of a with the prime p.

Theorem 9. *There is a natural isomorphism of vector spaces between the cusp forms of weight 2 with respect to $\Gamma_0(N)$ and the regular differential 1-forms on the complex*

[2] This is not a typo, it was indeed André Weil — not Andrew Wiles — that made this discovery.

manifold $\overline{\Gamma_0(N)\backslash\mathbb{H}}$. *The isomorphism* $\mathbb{S}_2(\Gamma_0(N)) \to \Omega^1\left(\overline{\Gamma_0(N)\backslash\mathbb{H}}\right)$ *is characterised by* $\zeta(\tau) \mapsto \zeta(\tau)\,\mathrm{d}\,\tau$. *Thus*

$$\dim_{\mathbb{C}} \mathbb{S}_2(\Gamma_0(N)) = g_N.$$

We can now immediately conclude that there are no modular elliptic curves of the level 2, since

$$\begin{aligned}
\dim_{\mathbb{C}}(\mathbb{S}_2(\Gamma_0(2))) &= g_2 \\
&= \frac{2+1}{12} - \frac{1}{4}\left(1 + \left(\!\!\left(\frac{-1}{2}\right)\!\!\right)\right) - \frac{1}{3}\left(1 + \left(\!\!\left(\frac{-3}{2}\right)\!\!\right)\right) \\
&= \frac{1}{4} - \frac{1}{4}\left(1 + \left(\!\!\left(\frac{-1}{2}\right)\!\!\right)\right) - \frac{1}{3}\left(1 + \left(\!\!\left(\frac{-3}{2}\right)\!\!\right)\right) \\
&= 0.
\end{aligned}$$

Frey curves

We now want to find a contradiction to the existence of solutions to $a^l + b^l = c^l$. Since we have seen that there are no modular elliptic curves of level 2, we should study what this "level" of an elliptic curve is. We will discover that it is called the *conductor*, and we will define it only for some special elliptic curves, the so-called *semistable* elliptic curve [5].

Definition 10 (Good/bad reductions)**.** Let

$$\bar{E} : Y^2 = X^3 + \bar{a}_1 X^2 + \bar{a}_2 X + \bar{a}_3 = f(X)$$

the reduction modulo p of an elliptic curve with integer coefficients. Then the reduction is called *good* if $\Delta \not\equiv 0 \mod p$. Otherwise the reduction is *bad* and two cases can happen:

1. If two zeros of $f(X)$ are equal, then the reduction is called a multiplicative reduction and the reduced curve $\bar{E}$ has a *node*.
2. If all the three zeros of $f(X)$ are equal, then it is an additive reduction and $\bar{E}$ has a *cusp*.

Definition 11 (Semistable elliptic curves)**.** An elliptic curve E over $\mathbb{Q}$ is called *semistable* if for all primes p, the reduction modulo p of E is either good or has a node.

At this point we can give a definition of the conductor that will work for our case.

Definition 12 (Conductor for semistable elliptic curves)**.** For a semistable elliptic curve E over $\mathbb{Q}$, the *conductor* N_E is the product of the primes p for which the reduction modulo p of E has a node, in other words, it is the product of all primes p that divides the discriminant.

Note that there are only finitely many primes for which $\bar{E}$ has a bad reduction: by definition, the reduction modulo p is bad if p divides the discriminant, hence for all but finitely many primes the reduction is good. This implies that our definition of the conductor is well-defined.

Now, we can finally define the elliptic curves we are interested in. The name of this curves comes from Gerhard Frey, who studies this type of curves and linked them to Fermat's Last Theorem.

Definition 13 (Frey curves)**.** Assume that there exists $a, b, c \in \mathbb{N}$ with $\gcd(a, b, c) = 1$, and $l \in \mathbb{N}$ prime with $l > 5$, that satisfy:

$$a^l + b^l = c^l.$$

Then we define the *Frey curve* as:

$$E_{a,b,c} : Y^2 = X(X - a^l)(X + b^l) = X^3 + (b^l - a^l)X^2 - (ab)^l X$$

The discriminant of $E_{a,b,c}$ is given by:

$$\Delta_{E_{a,b,c}} = 16(abc)^{2l}.$$

Therefore, can note that the Frey curves are semistable: the roots of $E_{a,b,c}$ are $\{0, a^l, -b^l\}$, if p is a prime that divides a, then it cannot divide b too, otherwise p divides $a^l + b^l = c^l$ and this contradicts $\gcd(a, b, c) = 1$. Hence either $p|a^l$ or $p|b^l$ and in both cases the polynomial $X(X - \bar{a}^l)(X + \bar{b}^l)$ has two distinct roots. If p divides c, then $a^l + b^l \equiv 0 \mod p$, this means that $a^l \equiv b^l \mod p$ which cannot be 0 for otherwise it would again contradict $\gcd(a, b, c) = 1$. Hence, also in this case, the polynomial has two distinct roots.

Since Frey curves are semistable, we can use Definition 12 to conclude that the conductor of $E_{a,b,c}$ is:

$$N_{a,b,c} = 2 \cdot \prod_{\substack{p \text{ prime} \\ p|abc \\ p \neq 2}} p.$$

Ribet's Reduction

In 1975, Ken Ribet proved what was called the ε-conjecture, and with that, that the Taniyama-Shimura conjecture implied Fermat's Last theorem. After successfully proving the ε-conjecture, it became known as Ribet's theorem. We will give a simplified version of the theorem applied to the Frey-Curve.

Theorem 14 (Ribet)**.** *Suppose there is a non-trivial integral triple a, b, c together with an exponent $l > 5$ such that $a^l + b^l = c^l$. Now, given that the associated Frey-Curve*

$$E_{a,b,c} = V_{\mathbb{P}^2}(X(X - a^l)(X + b^l))$$

is modular to the level $N = N_E$, then it is also modular to the level N/p for any odd prime p dividing N.

We may conclude that Ribet's theorem shows us that *if* the Frey-Curve $E_{a,b,c}$ is modular to the level N_E, then it is modular to the level 2. This is a contradiction to the equality $a^l + b^l = c^l$, since we showed that there are no modular curves to the level 2 and that all elliptic curves are modular to the level of their conductor. The conductor having a factor of 2, was directly given by the equality $a^l + b^l = c^l$.

Bibliography

[1] J.H. Silverman and J.T. Tate, *Rational Points on Elliptic Curves*, Springer (1992).

[2] P. Ribenboim, *13 Lectures on Fermat's Last Theorem*, Springer-Verlag, New York-Heidelberg-Berlin, (1979).

[3] J.S. Milne, *Elliptic curves*, BookSurge Publishers, (2006).

[4] J. Kramer, *Der grosse Satz von Fermat - die Lösung eines 300 Jahre alten Problems*, In: Aigner M., Behrends E. (eds) Alles Mathematik, (2000).

[5] J. Kramer, *Über die Fermat-Vermutung*, El. Math. 50 (1995), pp.11-25.