Simply Fermat's Last Theorem

Ayleen Roselie

April 2026

1 Introduction

Fermat's last theorem is renowned as one of, if not *the* hardest theorem in all of mathematics. The underlying complexity of the theorem lies underneath it's seemingly simple statement ;

$$x^n + y^n = z^n \text{ has no integer solutions for n} \geq 3$$

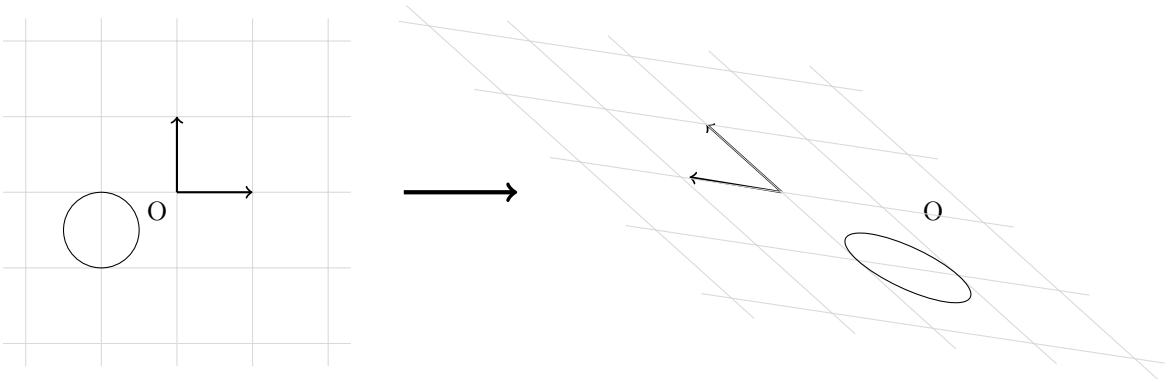
It was first proposed by Pierre de Fermat in 1637, and remained unsolved for 358 years until it was officially proven by Andrew Wiles in 1994.

Fermat's Last Theorem is extraordinarily difficult. Many mathematicians have attempted to work on it, but there has always been a subtle gap which flaws their proofs. It became clear that FLT is not the kind of problem one can solve only by the use of traditional algebra and number theory.

The final proof of Fermat's Last Theorem, though it was formally completed by Sir Andrew Wiles, is the product of years of work poured in by many mathematicians worldwide. What makes it so notable is how it connected many seperate and seemingly unrelated fields of math.

2 Modular Forms

Imagine an infinitely large sheet of stretchy paper. The paper can be stretched and squished like dough in general, but it must be kept flat - as it is "paper" after all. One point of the paper is pinned down. That point is *fixed in place*, and is called the origin point. Grid lines are then drawn on the paper.



It can be noticed that, the grid lines remain straight, parallel and evenly spaced after the stretchy paper was morphed. Mathematicians call 'morphs' like these "*linear transformations*". 2-dimensional linear transformations are written mathematically using matrices. The matrix $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ means that we map the vector $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ to $\begin{pmatrix} a \\ c \end{pmatrix}$, and map the vector $\begin{pmatrix} 0 \\ 1 \end{pmatrix}$ to $\begin{pmatrix} b \\ d \end{pmatrix}$.

The linear transformation that has just been executed has a special property : it preserves the area of shapes. The area of the circle is unchanged under the transformation, even if it has shapeshifted into an ellipse. Transformations like these are part of a group [1] called the *special linear group*.

Modular forms are closely related to special linear groups. Specifically, $SL_2(\mathbb{Z})$. Don't get confused by the weird symbols! "SL" stands for "special linear", the "2" means we are referring to 2-dimensional transformations. Lastly, "$\mathbb{Z}$" means we're focusing on transformations over the integers (a,b,c, and d must be integers).

Modular forms map

$\mathbb{H}$ (complex numbers with positive imaginary component) $\to \mathbb{C}$ (the complex numbers)

They satisfy a functional equation which must be true for *all* elements of the modular group. Specifically, let $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ be *any* element of $SL_2(\mathbb{Z})$. A complex function f is a modular form of weight k for an integer value of k

$$\iff f(\frac{az+b}{cz+d}) = (cz+d)^k f(z)$$

For a function to be modular, it needs to satisfy a truckload of symmetries. For some values of k, there do not exist any modular forms! Notice in the following diagram all the beautiful symmetries owned by a modular form.

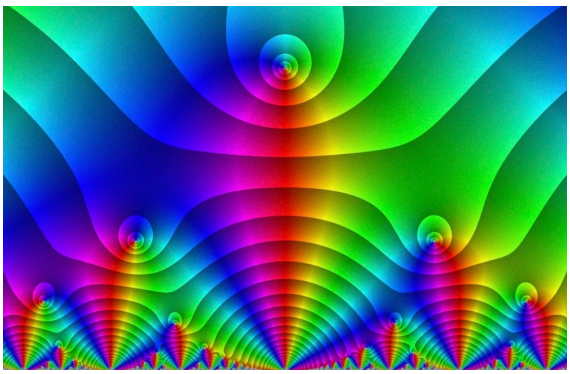


Figure 1: An example of a modular form (image from Quanta Magazine)

[1] A group is a set G combined with a binary operator $\circ$ (an operator which takes in 2 inputs and produces 1 output. ex : addition, multiplication) which satisfies the following properties :

- **Associativity** for elements a, b, and c in G, $a \circ (b \circ c) = (a \circ b) \circ c$.
- **Closure** for elements a and b in G, $a \circ b$ is also in G.
- **Identity** there exists an element e in G such that for all elements a in G, $a \circ e = e \circ a = a$.
- **Inverse** for all elements a in G, there exists an element a^{-1} in G such that, $a \circ a^{-1} = a^{-1} \circ a = e$.

More properties of modular forms :

- Note that, $\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in SL_2(\mathbb{Z})$. This by definition implies that

$$f(\frac{z+1}{1}) = (1)^k f(z)$$
$$f(z+1) = f(z)$$

  So, modular forms must have a period of exactly 1.

- Modular forms are holomorphic. This means that the derivative[2],

$$\lim_{h \to 0} \frac{f(z+h) - f(z)}{h}$$

  is always the same (and exists) for all values of z in $\mathbb{H}$, regardless of how h approaches 0 in the complex plane.

Another term we will use later on in this essay is the *cusp form*. A cusp form is a specific type of modular form, which has a value of exactly 0 at its cusps[3].

Also note that the modular forms we have been discussing so far are modular forms of *level 1*. If a modular form is of level N, then it satisfies

$$f(\frac{az+b}{cz+d}) = (cz+d)^k f(z)$$

only for matrices in the congruence subgroup $\Gamma_0(N) \subseteq SL_2(\mathbb{Z})$. $\Gamma_0(N)$ contains the matrices $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $c \equiv 0 \pmod{N}$. This means that, modular forms of higher levels are restricted to less symmetries than those of lower levels. Also, with the modular forms of level 1 being the most symmetric out of all.

Because of this, if a modular form is of level 1, it will also be of levels 2,3,4, and so on. A modular form of level 2 will also be one of all the higher (even) levels. So, a useful term we use to describe the minimal value of N for which a function is modular is the *level of a newform*.

A modular form being a newform of level N means that it is *new*. More precisely, that it isn't a modular form of any levels lower than N. So, "the level of a newform" is the same as "the minimal level of a modular form".

[2]The derivative of a function is its rate of change. h is the change in input, and $f(z+h) - f(z)$ is the change in output caused.

[3]A *cusp* is a complex number $a + i\infty$, i.e its imaginary component is ∞ (this may be a bit unusual, but we will treat ∞ as a number).

3 Fourier Series

Let $f(x)$ be a real-valued function with period T. A Fourier series breaks down $f(x)$ as an infinite sum of sine and cosine waves. Specifically (though we will not dive into the details here),

$$\begin{aligned} f(x) &= \sum_{n=0}^{\infty}[a_n cos(\frac{2n\pi x}{T}) + b_n sin(\frac{2n\pi x}{T})] \\ &= [a_0 cos(\frac{2(0)\pi x}{T}) + b_0 sin(\frac{2(0)\pi x}{T})] + \sum_{n=1}^{\infty}[a_n cos(\frac{2n\pi x}{T}) + b_n sin(\frac{2n\pi x}{T}) \\ &= [a_0 cos(0) + b_0 sin(0)] + \sum_{n=1}^{\infty}[a_n cos(\frac{2n\pi x}{T}) + b_n sin(\frac{2n\pi x}{T}) \\ &= a_0 + \sum_{n=1}^{\infty}[a_n cos(\frac{2n\pi x}{T}) + b_n sin(\frac{2n\pi x}{T})] \end{aligned}$$

A fun thing to notice is that, the period of $a_n cos(\frac{2n\pi x}{T}) + b_0 sin(\frac{2n\pi x}{T})$ is $\frac{T}{n}$. The larger that n gets, the smaller the period and thus the larger the frequency. This is not really rigorous, but it can be thought of as an increase in precision and accuracy too. Meanwhile, the constants a_n and b_n are the amplitudes of the waves.

For a complex function $f(t) : \mathbb{R} \to \mathbb{C}$ with period T,

$$f(t) = \sum_{n=-\infty}^{\infty} c_n e^{i(\frac{2\pi n}{T}t)} \text{ with } c_n \in \mathbb{C}$$

Chillax, it's not as complicated as it seems! First, recall that $re^{i\theta}$ represents a complex number, with r as the modulus[4], and θ is the argument[5]. We can think of this Fourier series geometrically, as *the sum of rotating vectors.*

The term $\frac{2\pi n}{T}$ is the frequency of the nth epicycle (a fancy word for *rotating vector*). The higher the frequency, the faster the rotation. If n is positive, the epicycle spins counter-clockwise. Otherwise, it spins clockwise. To decompose complex periodic functions into epicycles, we need both clockwise and counter-clockwise directions of rotation. This is why the sum goes from $-\infty$ to ∞ and not 0 to ∞ instead.

The constants c_n determine the phase shift[6] and modulus of the vectors. How so? c_n is also a complex number, so we can write it as

$$r_n e^{i\theta_n}$$

we can then re-write

$$f(t) = \sum_{n=-\infty}^{\infty} r_n e^{i(\frac{2\pi n}{T}t + \theta_n)}$$

Notice that at $t = 0$, the argument of the nth rotating vector is θ_n (its starting angle is θ_n). Also, the modulus of the rotating vector is r_n. This is precisely how c_n defines the phase shift and modulus of our epicycles.

[4]The modulus of a complex number is it's distance from the origin.
[5]The argument of a complex number is it's angle from the positive x-axis, measured counter clockwise.
[6]The phase shift of a rotating vector is it's starting angle.

Why did I write a whole section dedicated to Fourier series? It's because, we can identify modular forms by their Fourier series! This will be a crucial point later on.

But wait, a modular form is $\mathbb{H} \to \mathbb{C}$, not $\mathbb{R} \to \mathbb{C}$? So, we use the following very smart trick : Define

$$f_y(x) = f(x + iy) \text{ with } y \in \mathbb{R}^+$$

This way, $f_y(x)$ is a (periodic) function $\mathbb{R} \to \mathbb{C}$. We can then find the Fourier coefficients of the functions f_y (the c_n's as a function of y). Some math (which is outside the scope of this essay to be elaborated in detail) leads us to the following :

$$c_n(y) = a_n e^{-2\pi ny} \text{ with } a_n \in \mathbb{C}$$

Plugging this into the Fourier series formula, (note that $T = 1$, and with $z = x + iy$)

$$\begin{aligned} f_y(x) = f(z) &= \sum_{n=-\infty}^{\infty} c_n(y) e^{i(2\pi nx)} \\ &= \sum_{n=-\infty}^{\infty} a_n e^{-2\pi ny} e^{2\pi inx} \\ &= \sum_{n=-\infty}^{\infty} a_n e^{2\pi n(ix-y)} \\ &= \sum_{n=-\infty}^{\infty} a_n e^{2\pi n(iz)} \\ &= \sum_{n=-\infty}^{\infty} a_n e^{2\pi inz} \end{aligned}$$

Also, the holomorphic property of modular forms leads to the fact that $c_n = 0$ for negative values of n. We can also define $q = e^{2\pi iz}$. We end up with the following beautiful equation :

$$f(z) = \sum_{n=0}^{\infty} a_n q^n$$

where the constants a_n are called the Fourier coefficients of the modular form.

Some additional information: notice that, if we plug in $z = a + i\infty$ we will get

$$q = e^{2\pi iz} = e^{2\pi i(a+i\infty)} = e^{2\pi ia} e^{2\pi i^2 \infty} = e^{2\pi ia} e^{-\infty} = 0$$

$$a_0 = 0 \Rightarrow f(a + i\infty) = \sum_{n=0}^{\infty} a_n q^n = 0$$

So, the 0th Fourier coefficient of a cusp form must be 0.

4 Elliptic Curves

Elliptic curves are non-singular, meaning that they have no cusps (sharp, pointy bits) nor self intersections. This means that they are *smooth* (infinitely differentiable). Another property is

that they are symmetric over the x-axis.

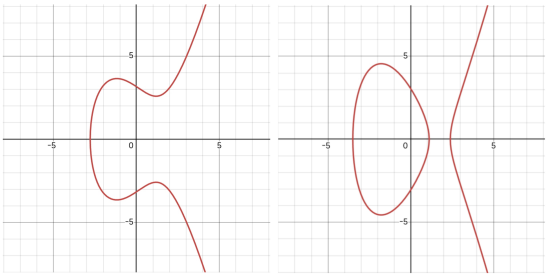


Figure 2: the two possible geometric shapes of an elliptic curve (graphs made on Desmos)

A Weierstrass equation generates an elliptic curve. It is often written in short form as $y^2 = x^3 + ax + b$. In the context of Fermat's Last Theorem, we focus on elliptic curves over the rationals. This means that the constants $a, b \in \mathbb{Q}$.

The discriminant of an elliptic curve determines whether it is non-singular. Commonly denoted Δ, it is equal to

$$-16(4a^3 + 27b^2)$$

where a,b are the coefficients of the Weierstrass equation that generates the elliptic curve. An elliptic curve is non-singular if and only if $\Delta \neq 0$

We can identify elliptic curves by their solutions in the field [7] $\mathbb{F}_p$ [8](with p being a prime number). Worded more simply,

$$x, y \in \{0, 1, 2, ..., p-1\} \text{ such that } y^2 \equiv x^3 + ax + b \text{ (mod p)}$$

Define $\#E(\mathbb{F}_p)$ as the number of solution pairs (x, y). It turns out to be the case that $\#E(\mathbb{F}_p)$ is always $\leq p + 1$. In other words,

$$\#E(\mathbb{F}_p) = p + 1 - a_p(E)$$

with $a_p(E)$ being an *error term*. Also, the letter 'E' represents the elliptic curve itself.

[7] Fields are quite similar to groups. A field is a set combined with 2 binary operators (denoted + and $\cdot$, called 'addition' and 'multiplication', and indeed behave like addition and multiplication. But, these two symbols don't literally represent actual addition and multiplication!), which satisfies the following properties :

- **Associativity** both + and $\cdot$ must be associative.
- **Commutativity** both + and $\cdot$ must be commutative. i.e, for elements a, b in the field, $a + b = b + a$ and $a \cdot b = b \cdot a$.
- **Closure** the set is closed under both + and $\cdot$.
- **Identity** there exists both an additive identity (0) and a multiplicative identity (1).
- **Inverse** all elements a in the field have an additive inverse. All nonzero elements a in the field have a multiplicative inverse.
- **Distributivity** multiplication distributes over addition. For elements a, b, c in the field, $a \cdot (b+c) = a \cdot b + a \cdot c$.

[8] $\mathbb{F}_p$ can be thought of as 'integers modulo p'.

5 The Frey Curve

The Taniyama-Shimura-Weil conjecture states that[9],

$$\forall E \ , \ \exists f(z) \ \text{ such that } \ \forall \text{primes } p \nmid N \ , \ a_p(E) = a_p(f)$$

with E being an elliptic curve (over the rationals $\mathbb{Q}$), $f(z)$ a weight-2 cusp form on $\Gamma_0(N)$, $a_p(E)$ being the error term of E in the field $\mathbb{F}_p$, and $a_p(F)$ being the p^{th} Fourier coefficient of f.

N is the conductor of E. It is a positive integer which measures the complexity of the curve by its bad reductions[10] f_p.

$$f_p = \begin{cases} 0 & \text{if } E \text{ has no bad reduction mod } p \\ 1 & \text{if } E \text{ is semi-stable mod } p \\ \geq 2 & \text{if } E \text{ is unstable mod } p \end{cases} \quad , \qquad N = \prod p^{f_p}$$

We only look at primes $p \nmid N$, because $p \mid N \iff E$ has a bad reduction modulo p.

Overall in simpler terms, for all elliptic curves there exists a unique modular form such that: the error terms of the elliptic curve correspond to the Fourier coefficients of the modular form.

German mathematician Gerhard Frey suggested that if there exists (a, b, c) as a counterexample of Fermat's Last Theorem, then there exists an elliptic curve - called the Frey curve - which has "odd" properties. The Frey curve is[11]

$$y^2 = x(x - a^n)(x + b^n) \quad \text{with } a, b \text{ coprime}$$

A reasonable question to ask is, "but that isn't of the form $y^2 = x^3 + ax + b$?" Well, we can shape it into that form by doing a change of variables. Define

$$x = X - \frac{b^n - a^n}{3}$$

[9] If you are unfamiliar with logic notation, $\forall$ simply means 'for all', and $\exists$ means 'there exists'.

[10] A bad reduction is when the elliptic curve becomes singular when reduced modulo p. E is semi-stable if it has a node (self-intersection), meanwhile it is unstable if it has a cusp.

[11] If a, b are not coprime, let their GCD be equal to g. The triple $(\frac{a}{g}, \frac{b}{g}, \frac{c}{g})$ is a Fermat triple, with $\frac{a}{g}$ and $\frac{b}{g}$ coprime. So we can use those values as a, b instead.

Then, you will get

$$\begin{aligned}
y^2 &= x(x-a^n)(x+b^n) \\
&= x^3 + (b^n - a^n)x^2 - a^n b^n x \\
&= \left(X - \frac{b^n - a^n}{3}\right)^3 + (b^n - a^n)\left(X - \frac{b^n - a^n}{3}\right)^2 - a^n b^n \left(X - \frac{b^n - a^n}{3}\right) \\
&= \left[X^3 - (b^n - a^n)X^2 + \frac{(b^n - a^n)^2}{3}X - \frac{(b^n - a^n)^3}{27}\right] + \\
&\quad \left[(b^n - a^n)\left(X^2 - 2\frac{b^n - a^n}{3}X + \frac{(b^n - a^n)^2}{9}\right)\right] - \left[a^n b^n X - a^n b^n \frac{b^n - a^n}{3}\right] \\
&= \left[X^3 - (b^n - a^n)X^2 + \frac{(b^n - a^n)^2}{3}X - \frac{(b^n - a^n)^3}{27}\right] + \\
&\quad \left[(b^n - a^n)X^2 - 2\frac{(b^n - a^n)^2}{3}X + \frac{(b^n - a^n)^3}{9}\right] - \left[a^n b^n X - a^n b^n \frac{b^n - a^n}{3}\right] \\
&= X^3 + \left[-(b^n - a^n) + (b^n - a^n)\right]X^2 + \left[\frac{(b^n - a^n)^2}{3} - 2\frac{(b^n - a^n)^2}{3} - a^n b^n\right]X + \\
&\quad \left[-\frac{(b^n - a^n)^3}{27} + \frac{(b^n - a^n)^3}{9} - a^n b^n \frac{b^n - a^n}{3}\right] \\
&= X^3 + \left(-\frac{(b^n - a^n)^2}{3} - a^n b^n\right)X + \left(\frac{2(b^n - a^n)^3}{27} - a^n b^n \frac{b^n - a^n}{3}\right)
\end{aligned}$$

The above is just an unnecessary load of detail, though. It is definitely preferred and neater to just write the Frey curve as $y^2 = x(x - a^n)(x + b^n)$

But, in this form, we can find the discriminant of the Frey curve! With a bit of calculation, you will get that

$$\Delta = 16(abc)^{2n}$$

which is *extremely* large compared to the conductor of the curve. Also, it has a power of $2n$, which is a pretty high power. These are the "odd" properties of Frey Curves previously mentioned.

6 Ribet's Theorem

A Galois group, $\mathrm{Gal}(L/K)$, is the group of automorphisms[12] $\sigma : L \to L$, such that $\sigma(x) = x$ for all $x \in K$ (the points in the field K are fixed in place). Meanwhile, a Galois representation (often denoted ρ) is a homomorphism[13] from a Galois group to a general linear group[14] (denoted $GL(V)$ with V being a vector space).

Think about the field as a fabric, and the Galois group as the group of 'morphs' which preserve the underlying structure of the fabric. Now, we want to observe how these morphs interact with each other. Like, how two of them done one after another will lead to another morph which is still part of the Galois group.

Do these 'morphs' remind you of linear transformations?

[12] An automorphism is a bijective and structure-preserving map from a mathematical object (in this case a field) to itself, hence the $\sigma : L \to L$.

[13] A homomorphism is a structure-preserving, but not necessarily bijective map from one mathematical object to another (in this case, from one group to another)

[14] A general linear group is a group of non-degenerate linear transformations over a specific vector space, ex: the plane (see Section 2 on Modular Forms)

Linear transformations are clearly much easier to work with than these abstract morphs. So, if we want to study the structure of the Galois group, then we correspond it to a general linear group! This is the Galois representation.

An absolute Galois group of a field K is $\mathrm{Gal}(\overline{K}/K)$, with $\overline{K}$ being the closure[15] of K. Usually (including in the context of modular forms), Galois representations are defined on the absolute Galois group.

A Galois representation $\rho : \mathrm{Gal}(L/K) \to GL(V)$ is irreducible if there is no nonzero subspace $W \subset V$ such that

$$\forall w \in W \quad \text{and} \quad \forall g \in \mathrm{Gal}(L/K) \quad , \quad \rho(g) \cdot w \in W$$

This essentially means that, the only vector subspaces preserved after applying ρ is $\{0\}$ and V itself. In even simpler words, ρ cannot be broken down into smaller dimensional components. A Galois representation is *absolutely irreducible* if it is irreducible even after K is extended (usually to $\overline{K}$). As in, $\mathrm{Gal}(L/K)$ is irreducible and so is $\mathrm{Gal}(L/\overline{K})$.

Some notation : if f is a modular form, we use ρ_f to denote its usual (the associated) Galois representation. Meanwhile, $\overline{\rho_f}$ is a special kind of Galois representation, where you map the Galois group to $GL(\mathbb{F}_p)$ [16](for an arbitrary prime value of p).

Ribet's Theorem (Level-Lowering) states that :

Let $f(z)$ be a newform of level $N \cdot l$, with $l \nmid N$ (l is a prime), and $\overline{\rho_f}$ is absolutely irreducible. If at least one of the following is true :

- $\overline{\rho_f}$ is unramified at l.
- $l = p$ and $\overline{\rho_f}$ is flat at l.

Then there exists a newform $g(z)$ of level N such that $\overline{\rho_f} \simeq \overline{\rho_g}$.

Put simply, "unramified" and "flat" at l means that ρ behaves neatly around l. The symbol $\simeq$ means isomorphic, or that $\overline{\rho_f}$ has the same structure as $\overline{\rho_g}$.

The corresponding modular form of the Frey curve (if it exists) must satisfy all the properties above. So, we can apply the theorem! Because the Frey curve has some odd properties, it leads to the form being level-lowered down to level 2. But, there are no cusp forms of weight 2 and level 2.

This means that the Frey curve doesn't have a corresponding modular form! But then, this contradicts the Taniyama-Shimura-Weil conjecture, so overall, the Frey curve cannot even exist. Guess what that implies? That there does not exist a Fermat triple!

[15]The closure of a field K is the smallest (algebraically) closed field which contains it. This means that, any polynomial with coefficients in K has a root in $\overline{K}$.

[16]$GL(\mathbb{F}_p)$ is the general linear group over the field $\mathbb{F}_p$.