Fermat's Last Theorem

Colin McLarty

Contents

Abstract

For 300 years, Fermat's Last Theorem seemed to be pure arithmetic little connected even to other problems in arithmetic. But the last decades of the twentieth century saw the discovery of very special cubic curves, and the rise

C. McLarty (✉)
Case Western Reserve University, Cleveland, OH, USA
e-mail: colin.mclarty@case.edu

B. Sriraman (ed.), *Handbook of the History and Philosophy of Mathematical Practice*,
https://doi.org/10.1007/978-3-031-40846-5_44

of the huge theoretical Langlands Program. The Langlands perspective showed those curves are so special they cannot exist, and thus proved Fermat's Last Theorem. With many great contributors, the proof ended in a deep and widely applicable geometric result relating nice curves in rational coordinates to very nice surfaces in complex coordinates. This geometry is the only proof strategy for FLT yet known. Outstanding but not atypical of current mathematics, the proof challenges common philosophic views on abstraction and structure and raises still-open logical questions in arithmetic.

Keywords

Structuralism · Fermat's Last Theorem · Geometry · Peano Arithmetic

1 Introductory Overview

For most of its 350-year history, Fermat's Last Theorem (FLT) was intriguing arithmetic of no special philosophic interest. Even in arithmetic, it had little relation to other theorems. A generalized "prime factorization" misused in a fallacious proof of FLT has been a driving force in number theory to this day. That has gotten philosophers' and historian's attention as an example of mathematical concept formation, see Sect. 3.2. But that strategy did not prove FLT (yet).

The proof of FLT in 1995 had two beginnings in the late 1960s. It involves much other mathematics. And it challenges philosophic views of modern mathematics.

One beginning was a surprising specific geometric observation on FLT. The other was the vast theoretical *Langlands Program* in arithmetic algebraic geometry, see Sect. 6. These two had no connection at first except that both descended from André Weil's lifelong project to geometrize number theory. Weil sketched in his dissertation (1928), and explained at length in later works (1987, 1991), how he built that project on his youthful reading of Diophantus, Fermat, and Poincaré.

The geometric observation came when Yves Hellegouarch and Vadim Andreevich Demyanenko independently noticed that if any of the curves defined by certain cubic polynomials had certain odd geometric properties, then FLT was false. This was not the most exciting direction of inference: No one knew how to find curves with those properties, and people rather expected there were none, since people expected FLT was true. And in the other direction, a proof that no such curves exist would only rule out one kind of counterexample to FLT. It would not prove FLT.

Still, it was a striking new connection for FLT.

Poincaré (1901) related arithmetic to the geometry of cubic curves following leads in Fermat among others. Poincaré's work was developed notably by Louis Mordell and André Weil and the subject was well established by the 1970s. Intriguingly, though, Hellegouarch and Demyanenko used a less known aspect of it, namely *rational torsion points* on *elliptic curves.* Jean-Pierre Serre at the famed Séminaire Bourbaki had to say one knows very little about them (1971, p. 282). Within a few

years, Barry Mazur (1977a, 1978) changed that with deep results on torsion and on *modularity:*

> Although it was not realized at the time, the chain of ideas that was to lead to a proof of Fermats Last theorem had already been set in motion by Barry Mazur in [these papers]. (Darmon et al. 1997, p. 8)

Mazur was not aiming at FLT per se. He was interested in geometrizing arithmetic in general.

Gerhard Frey (1986), with crucial input from Serre, gave a strategy for FLT using modularity: Show any counterexample to FLT would give an elliptic curve that is *not* modular. But also prove the *Modularity Theorem* (MT), saying every relevant elliptic curve *is* modular. Number theorists had known for years that MT would be powerfully useful if it was true. Where FLT was easy to state but had few consequences, MT was hard to state but had rich consequences.

Up to this point, all the arguments were virtually pure arithmetic. They made subtle calculations using sophisticated geometric concepts in arithmetic but used no higher order logic, except some real and complex analysis so concretely calculational that it should be conservative over (first order) Peano Arithmetic (PA). The arguments were basically expressed in PA except that the authors were not interested in formalization.

Experts found Frey's strategy persuasive and Ken Ribet (1990), drawing on ideas from Serre and Mazur, proved counterexamples to FLT would produce non-modular elliptic curves. Few believed the other step, MT, was within reach. Andrew Wiles pursued it in near total secrecy for 6 years, though, and shocked the world in 1993 by proving enough of MT to give FLT as corollary (with some corrections over the next 2 years). Ribet's and Wiles' proofs too are led by directly arithmetical ideas but both invoke higher level apparatus. Within 5 years, Wiles' method of *modularity lifting* plus a lot of skilled calculating led to a proof of the full Modularity Theorem and made modularity even more useful than had been known. While the proof of MT is extremely valuable, it was FLT that inspired Wiles since age 10 and FLT made headlines around the world and Fermat's own statement of it is quoted in Latin at the start of Wiles (1995a).

2 Basics

A *Pythagorean triple* is any triple $\langle a, b, c\rangle$ of nonzero integers with

$$a^2 + b^2 = c^2.$$

Some sources require a, b, c positive. This makes little difference since $x^2 = (-x)^2$.

Euclid's *Elements* Book 10 proposition 19 shows how to generate infinitely many of these. For any distinct, nonzero integers m, n take

$$a = 2mn, b = m^2 - n^2, c = m^2 + n^2. \quad (1)$$

Routine calculation shows this gives Pythagorean triples.

Conversely, every Pythagorean triple is a multiple of one formed by Eq. 1. There are several interestingly different proofs of this: by using ordinary prime factorization, or by factorization with complex Gaussian integers $a + b\sqrt{-1}$, or by the geometry of rational points on a circle. Each of these proofs is common in textbooks and they are all in the Wikipedia article *Pythagorean triples*.

Sometime around 1630 Fermat wrote in the margin of a book that he had a "marvelous proof" that there are no nonzero integer solutions for cubes or fourth powers or any higher order.

$$a^3 + b^3 = c^3 a^4 + b^4 = c^4 a^n + b^n = c^n \, n \geq 5.$$

Essentially Fermat's own proof for fourth powers is common in textbooks today. No proof by him of the cubic case is known. He may have believed for the rest of his life that he had a proof of the cubic case, but it seems he fairly soon came to believe he had no proof for higher powers. See Edwards (1977, p. 2) and Weil (1987, p. 104).

For ease of reference, define F_n to be the Fermat equation with exponent n including the requirement that none of *a, b, c* are 0:

$$a^n + b^n = c^n \text{ with integers } abc \neq 0.$$

Notice F_n has rational solutions if and only if it has integer solutions. An integer solution is rational. And conversely any rational solution can be expressed by fractions with a common denominator, and then

$$(a/d)^n + (b/d)^n = (c/d)^n \text{ is equivalent to } a^n + b^n = c^n.$$

So it does not matter whether FLT is stated for integers or for rational numbers.

Also, if F_n has a solution, and n factors as $n = k \cdot m$, then F_m has a solution:

$$a^{k \cdot m} + b^{k \cdot m} = c^{k \cdot m} \text{ is equivalent to } \left(a^k\right)^m + \left(b^k\right)^m = \left(c^k\right)^m. \quad (2)$$

Every integer $n > 1$ has prime factors, and Fermat proved F_4 has no solutions. So if FLT is not true, there must be some odd prime p such that F_p has solutions.

3 A Proof for $n = 3$ and a Valuable Mistake

In 1770, Leonhard Euler published a brilliant but unclear proof of the cubic case, saying F_3 has no integer solutions. It centered on factoring numbers of the form $a^2 + 3b^2$ where *a, b* are integers. Some steps were not justified clearly.

Edwards (1977, pp. 43ff.) relates Euler's proof to using complex numbers of the form $a + b\sqrt{-3}$ with ordinary integers *a, b.* These numbers add and multiply, keeping this form:

$$\left(a + b\sqrt{-3}\right) + \left(c + d\sqrt{-3}\right) = (a + c) + (b + d)\sqrt{-3} \tag{3}$$

$$\left(a + b\sqrt{-3}\right) \cdot \left(c + d\sqrt{-3}\right) = (ac - 3bd) + (ad + bc)\sqrt{-3}. \tag{4}$$

This connects to Euler's argument by a special case of Eq. 4:

$$\left(a + b\sqrt{-3}\right) \cdot \left(a - b\sqrt{-3}\right) = a^2 + 3b^2.$$

Euler's unwarranted assumption could be justified in these terms if unique prime factorization generalizes from the ordinary integers $\mathbb{Z}$ to these numbers $a + b\sqrt{-3}$, with the primes existing in this new context. And it sort of does. Only it requires using terms $\left(1 + \sqrt{-3}\right)/2$ with denominator 2 as well as integer multiples of $\sqrt{-3}$. This indeed is one way to prove F_3 has no solutions.

In modern terms, the ring called $\mathbb{Z}\left[\left(1 + \sqrt{-3}\right)/2\right]$ has unique prime factorization:

$$\mathbb{Z}\left[\frac{1 + \sqrt{-3}}{2}\right] = \left\{a + b \cdot \frac{1 + \sqrt{-3}}{2} \middle| a, b \in \mathbb{Z}\right\}.$$

In this context, for example, 7 is no longer a prime since it has smaller factors

$$7 = 2^2 + 3 \cdot 1^2 = \left(2 + \sqrt{-3}\right)\left(2 - \sqrt{-3}\right).$$

In the ring $\mathbb{Z}\left[\left(1 + \sqrt{-3}\right)/2\right]$, it turns out $2 + \sqrt{-3}$ and $2 - \sqrt{-3}$ are prime.

Rings like $\mathbb{Z}\left[\left(1 + \sqrt{-3}\right)/2\right]$ are today called *rings of integers of algebraic number fields.* They do not always have unique prime factorization and that has been the topic of a great deal of number theory. See many algebraic number theory textbooks, including, for example, Stewart and Tall (2001, Chap. 3), and see Sect. 3.2 below.

3.1 Proofs by Factorization

Over the next century, Euler's proof for F_3 was perfected, others used similar arguments to prove F_5 has no solutions, and Gabriel Lamé did it for F_7. Then in 1847, Lamé unified those proofs into a purported proof of all of FLT. He expressed the idea quite clearly and so he was clearly wrong. But he was not wholly wrong. The historic events are well known and vividly described by Edwards (1977, pp. 76ff.) and Stewart and Tall (2001, pp. 185ff.).

Because people knew our Eq. 2, Lamé just had to show F_p has no solution with odd prime p. For that case, Lamé would take a *primitive p-th root of unity.* That is a complex number ζ_p such that $\zeta_p^p = 1$ while $\zeta_p \neq 1$. These always exist. They are those roots of $X^p - 1$ that are not roots of $X - 1$, so long division of polynomials shows they are exactly the solutions to

$$X^{p-1} + X^{p-2} + \cdots + X + 1 = 0. \tag{5}$$

To verify this, just multiply that polynomial by $X - 1$.

Using Eq. 5, simple calculation (try it by hand for $p = 3$) shows for any integers *a, b* there is a factorization

$$a^p + b^p = (a + b)(a + b\zeta_p)\left(a + b\zeta_p^2\right)\cdots\left(a + b\zeta_p^{p-1}\right).$$

Thus any solution $\langle a, b, c\rangle$ to F_p has a factorization

$$(a + b)(a + b\zeta_p)\left(a + b\zeta_p^2\right)\cdots\left(a + b\zeta_p^{p-1}\right) = c^p. \tag{6}$$

This is all correct. Then Lamé proved Eq. 6 is impossible by using further steps which, as his colleague Joseph Liouville quickly pointed out, amounted to an unjustified assumption of unique prime factorization for this sort of number.

Liouville knew this was a major gap. Then Liouville got a letter from Ernst Kummer showing unique factorization actually fails for $p = 23$. In fact, it fails for all primes $p \geq 23$ (see the Wikipedia article *Cyclotomic fields* and references there).

Kummer extended this kind of argument to give some results even in contexts where unique factorization failed. Further work on FLT in this line became extremely sophisticated and unavoidably technical. It is well covered in Edwards (1977). This strategy remained the main engine for proving new cases of FLT right up until the 1980s. It did not lead to the current proof of FLT.

3.2 Further on Factorization in Rings

Fermat's Last Theorem was not the first reason mathematicians studied unique prime factorization in these algebraic contexts. But it was one. And the tools created to study factorization pervade all of algebraic geometry and algebraic number theory today, naturally including the proof of FLT. So a few more words on the topic are appropriate though it is not a specific theme leading to Wiles' proof of FLT.

Mathematicians made two related responses to failure of unique prime factorization in a ring R. First was to define new *ideal divisors* for R, which are not elements of R but serve some of the same purposes as ordinary prime factorization in the integers $\mathbb{Z}$. Second was to measure "how close" R is to having unique prime factorization, which became one source of the far-reaching subject of *Class Field Theory* (Cox 1989; Kato et al. 2011). Kummer himself took early steps on both of these Mazur (1977b).

Today *ideals of rings* are bread and butter for algebraists, and often get into sophomore algebra courses. Philosophers and historians have taken this as a major example of mathematical concept formation. See, for example, Avigad (2006) and Boniface (2016). Class Field Theory is so imposing philosophers have barely studied it, and then indirectly through another of its sources the Quadratic Reciprocity Theorem. See, for example, Tappenden (2008), Yap (2011, 2020), and D'Alessandro (2020).

4 Poincaré, Mordell, and Weil

> What we possess of [Fermat's] methods for dealing with curves of genus 1 [i.e. elliptic curves] is remarkably coherent; it is still the foundation for the modern theory of such curves. It naturally falls into two parts; the first one, directly inspired by Diophantus, may conveniently be termed a method of ascent, in contrast with the descent which is rightly regarded as Fermat's own. (Weil 1987, p. 104)

Weil (1987, pp. 104–18) features the rich history of connections between arithmetic and cubic curves. See also Stillwell (1995). Whether Diophantus or Fermat themselves thought of their calculations in geometric terms is not clear. Poincaré (1901) did explicitly when he systematized Fermat's methods, related them to Kronecker's and Weierstrass's complex analysis, and expanded them to new problems.

4.1 Rational Roots of Cubic Polynomials

This computation was obvious to Fermat:

Theorem 1 *If a cubic polynomial P(X) in one variable with rational coefficients has two rational roots, or one rational root of multiplicity 2, then all its roots are rational.*

Proof. If α_1 is a rational root of $P(X)$, then $X - \alpha_1$ divides $P(X)$ and the quotient $Q(X)$ is quadratic with rational coefficients. If α_2 is a distinct rational root of $P(X)$, then $X - \alpha_2$ divides $Q(X)$. If α_1 is a double root of $P(X)$, then $X - \alpha_1$ divides $Q(X)$. Either way leaves a rational linear polynomial which necessarily has a rational root, and that is all the roots of $P(X)$. QED.

Notice this is pure arithmetic calculation, requiring no general notion of real or complex numbers let alone, say, the fundamental theorem of algebra.

Moving to two variables, and thus to the geometry of curves in the plane, does not change the idea.

Definition 1 A *rational point* in the coordinate plane $\mathbb{R}^2$ is a point with both coordinates rational numbers.

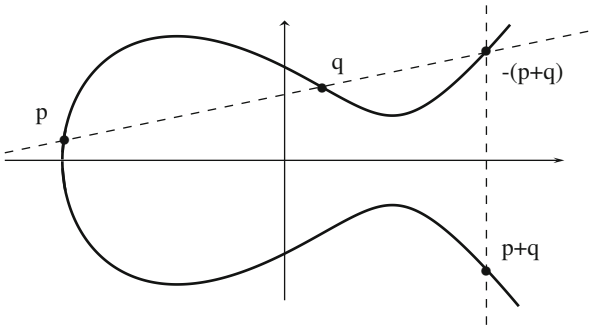

Fig. 1 The addition law on the elliptic curve defined by $Y^2 = X^3 - X + 1$

The curve in Fig. 1 is defined by a cubic equation in two variables

$$Y^2 = X^3 - X + 1. \tag{7}$$

Theorem 1 easily implies: If a straight line meets that curve in three points, and two of them are rational, then so is the third. And a line tangent to that curve at a rational point meets the curve in another rational point. To make those true even for vertical lines, we speak of a rational point on the curve at vertical infinity.

Equation 7 has six obvious rational (indeed integral) solutions: namely $\langle 0, \pm 1\rangle$, and $\langle -1, \pm 1\rangle$, and $\langle 1, \pm 1\rangle$. Take one of them, say $r = \langle 0, 1\rangle$. Routine calculation gives the tangent line to the curve at r, and by Theorem 1, it meets the curve in a new rational point, say s. Then take the tangent to the curve at s to find another rational point, and so on. Diophantus used this kind of algebra to find nontrivial rational solutions to suitable problems by starting from trivial ones.

Two questions naturally arise:

1. Is every rational solution to Eq. 7 generated this way starting from those six points? If not, does some other finite set of points generate them all?
2. Starting from a given one, say r, will this method continue generating new rational points? Or will it eventually return to r and start repeating?

The pursuit of those questions, and several other questions from number theory and from complex analysis, led to the idea of addition laws on elliptic curves.

4.2 The Addition Law on an Elliptic Curve

A wondrous geometric fact is that intersecting lines with a cubic curve produces an *addition law* so that the points of the curve form an Abelian group.

Definition 2 A *real elliptic curve* is a curve defined by an equation

$$Y^2 = X^3 + AX + B \tag{8}$$

where *A, B* are real numbers and the roots of $X^3 + AX + B$ are all distinct.

The addition law is uniquely determined by two rules:

1. The point at vertical infinity will be the 0 element.
2. For any three collinear points *p, q, r* on the curve, $p + q + r = 0$. (If the tangent to the curve at *p* also meets the curve at q, then $2p + q = 0$. And if *p* is an inflexion point then $3p = 0$.)

By these rules, if the line between points *r, s* on the curve is vertical, then

$$r + s + 0 = 0.$$

So define $-r = s$. Then look at the example in Fig. 1. The line connecting points *p, q* meets the curve in a third point and rule 2 says these three points add up to 0, so call that third point $-(p + q)$. Now the vertical line through $-(p + q)$ meets the curve in the negative of $-(p + q)$, thus $p + q$.

Simple reasoning left to the reader will verify most of the Abelian group axioms:

$$p + q = q + p, p + 0 = p, \text{for every } p \text{ there is a } q \text{ with } p + q = 0.$$

The associative law $p + (q + r) = (p + q) + r$ is not easy but it does follow by a classical geometric property of algebraic curves. See, e.g., Reid (2013, Chap. 2) or Stewart and Tall (2001, p. 229).

So each elliptic curve *E* has a group of real points $E(\mathbb{R})$ which is Abelian and contains at least the point at vertical infinity as 0 element. The same algebra gives an Abelian group $E(\mathbb{C})$ of the complex points on *E*. These geometric facts were important in early nineteenth century complex analysis, long before they were cast in group-theoretic form (McKean and Moll 1999, Chap. 2).

4.3 Rational Points on Arithmetic Elliptic Curves

Another wondrous fact is that the rationality considerations of Sect. 4.1 and the geometry and analysis of 1.4.2 fit right together to put Fermat's ascent and descent into modern geometric form.

Definition 3 An *arithmetic elliptic curve* is an elliptic curve defined by $Y^2 = X^3 + AX + B$ with rational coefficients *A, B*.

Theorem 1 shows the addition law on an arithmetic elliptic curve takes rational points to rational points, and so gives an Abelian group $E(\mathbb{Q})$ of rational points.

In this notation, the Fermat method of ascent by repeated tangent lines, described just after Eq. 7, takes a point r, then forms $s = -2r$, and then $-2\ s = 4r$, then $-8r$, $16r$, ... multiplying by -2 each time. The modern approach gives more points by repeated addition. Starting from one rational point r form the series:

$$r, r + r = 2r, 2r + r = 3r, 3r + r = 4r, 4r + r = 5r, \ldots.$$

A crucial concept is *torsion*. Point r is a torsion point if $n \cdot r = 0$ for some positive integer n, so that $(n + 1)\ r$ is just r again. Notice 0 is a torsion point trivially. And the sum of torsion points is torsion because $n \cdot r = 0$ and $m \cdot s = 0$ implies $mn \cdot (r + s) = 0$.

So, for every arithmetic elliptic curve E, the Abelian group $E(\mathbb{Q})$ of rational points has a subgroup $E_{tor}(\mathbb{Q})$ of rational torsion points. The questions at the end of Sect. 4.1 take this form:

1. Is the Abelian group $E(\mathbb{Q})$ generated by some finite set of its points?
2. Does $E(\mathbb{Q})$ have nonzero torsion points? If so, what are they, and what is the structure of $E_{tor}(\mathbb{Q})$?

Mordell proved the answer to the first is yes. Then Weil used a more transparent and systematic method extending this to a far more general result (Cassels 1973, p. 500). The question of how many elements it takes to generate a given $E(\mathbb{Q})$ drew a lot of attention, produced great results, and remains open. See the Wikipedia article *Birch and Swinnerton-Dyer conjecture*. For a recent Fields Medalist's contribution stressing elementary arithmetic aspects, see Bhargava and Shankar (2015).

While there had been notable work on rational torsion points, for example, by Beppo Levi (Schappacher and Schoof 1996), the torsion groups $E_{tor}(\mathbb{Q})$ were not well understood until Mazur's papers in the 1970s. A key consequence of Mazur (1978) shows how specific the arithmetic gets, and it will reappear in the noted "3–5 trick" in Wiles' proof (Sect. 7.2). For any point r on any arithmetic elliptic curve E, the smallest nonzero $n \in \mathbb{N}$ with $nr = 0$ in $E(\mathbb{Q})$ has to be either in the range 1–10, or else 12. Wiles (1995a, p. 542) uses the fact that it cannot be 15.

However, Mordell also built on Poincaré (1901) to pose a conjecture for other shapes of algebraic curves, which Weil could not settle.

5 The First Deep Geometric Result on FLT

> I asked myself whether many problems of indeterminate Analysis [in arithmetic] could not be connected to each other by a new classification of higher degree polynomials. . .. Poincaré (1901, p. 21)

The most obvious classification of polynomials is by degree. These two equations both have degree 3:

$$Y^2 = X^3 - X + 1 \tag{9}$$

$$Y^2 = X^3 - 3X + 2 \tag{10}$$

But Hilbert and Hurwitz (1890) already knew that in terms of number theory, Eq. 10 does not act like a cubic, it acts like a quadratic. It acts like a degree 2 equation. Hilbert, Hurwitz, and Poincaré all knew the right explanation is the geometry of the curves defined by these equations. The difference shows in plots of their real points. Compare Figs. 1 and 2. The curve defined by Eq. 10 crosses itself. Any cubic curve that crosses itself like that will behave arithmetically like a quadratic curve.

Specifically, all those mathematicians knew the important classification of polynomials for number theory is not by degree but by the *genus* of their curves. Here "curves" means spaces of complex number solutions, including points at infinity. Hilbert, Hurwitz, and Poincaré drew on Riemann's classification of what are now called Riemann surfaces (Riemann 1851; Pont 1974).

5.1 Curves as Surfaces

This idea is needed to state Mordell's conjecture, and again to describe modularity in Sect. 6.2. The conceptual point is to show these theorems involve perfectly concrete visual geometry.

Analysts talk about the *complex plane* while algebraic geometers call that same thing the *complex line.* More generally, the analyst's *Riemann surfaces* can be the geometer's *algebraic curves.* Of course, analysts and geometers are often the same people and the dual terminology has even been used as a textbook title: *Algebraic Curves and Riemann Surfaces* (Miranda 1995).

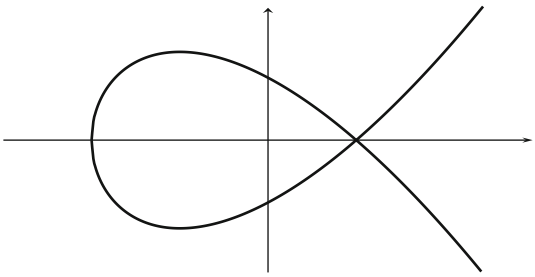

Fig. 2 The cubic curve defined by $Y^2 = X^3 - 3X + 2$

It is a choice between emphasizing topology or algebra. Topologically, the complex numbers form a plane with a real axis and imaginary axis, and small regions in a Riemann surface look like small regions in that plane. Algebraically, it is intuitive to say one equation in two variables defines a curve, and it is technically correct to say the complex numbers form a 1-dimensional complex vector space.

Visually, the curves in Figs. 1 and 2 formed by real solutions to Eqs. 9 and 10 are real cross-sections of complex curves – understanding that a complex curve is topologically a surface. Both curves go on to infinity so neither surface can be drawn whole to scale. But topologically, each curve expanding towards a point at infinity can be contracted down to put that point at a finite distance as in Fig. 3. Then the complex solutions to Eq. 9 form a torus. Those to Eq. 10 form a pinched torus where one part of the tube is pinched to a point. That pinch point is where the curve of real solutions in Fig. 2 crosses itself.

For more on genus, see numerous graphics in McKean and Moll (1999, Chap. 1) or the concise discussion in terms of algebra, topology, and differential geometry by Reid (2013, pp. 51ff.).

5.2 Mordell's Conjecture, Faltings' Theorem

In fact, quadratic curves (and self-crossing cubic curves) have genus 0, elliptic curves have genus 1, and when $n > 3$, then the curve of the Fermat equation F_n has genus greater than 1. Mordell conjectured that every algebraic curve with genus greater than 1 has at most finitely many rational points.

In 1983, Gerd Faltings proved this and much more (Mazur 1987). This was far-reaching arithmetic quite apart from FLT. It was beautiful classical arithmetic proved by insightful geometric means using the latest technology, including Grothendieck's *étale cohomology* (Faltings 1983). It was huge progress on FLT: No Fermat equation F_n with $n > 2$ has more than finitely many solutions! And several people gave explicit calculations based on this, concluding that F_n has no solutions, for the great majority of all $n > 2$ (Heath-Brown 1984). However, this proof did not identify which F_n have no solutions.

This result has not (yet) been the basis for any full proof of FLT.

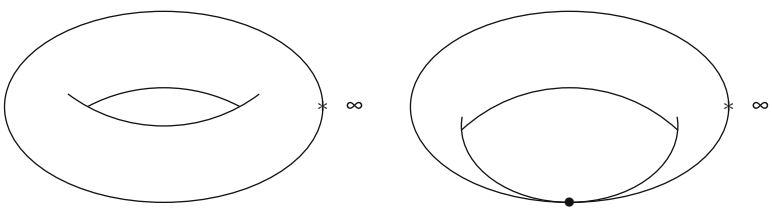

Fig. 3 The topology of a torus and a pinched torus, showing points at infinity

6 Elliptic Curves Meet FLT

Hellegouarch (1971) and Demyanenko (1971) independently used equations related to Eq. 11 below, to show that if certain kinds of torsion points can occur on elliptic curves, then FLT is false. It might naturally have occurred to one or both of them to turn this around and consider Eq. 11 for any hypothetical solution $\langle a, b, c\rangle$ to F_p for any odd prime p, but this never happened in print. Hellegouarch (2000) documents that he spoke on elliptic curves and FLT in the 1960s, but gives no details of what he said except that Serre corrected some errors.

At any rate, Hellegouarch and Demyanenko linked FLT to the arithmetic of elliptic curves. Mazur (1978) showed that Hellegouarch-Demyanenko torsion cannot occur on arithmetic elliptic curves; so these considerations would not refute FLT. But that did not prove FLT.

On the other hand, the beginnings of the Langlands Program are well documented, because in 1967, Robert Langlands wrote to Weil, to describe a huge unifying project (Mueller 2018). There has been great progress while a great deal remains to do.

> The Langlands Program is now a vast subject. There is a large community of people working on it in different fields: number theory, harmonic analysis, geometry, representation theory, mathematical physics. Although they work with very different objects, they are all observing similar phenomena. (Frenkel 2013, p. 77)

Wiles (1995a) is itself, technically, a step in the Langlands Program. But there is a persuasive argument that the overall concept of the Program mattered more in the process of proving FLT than the technical connection:

> Rather the crux of the Program is two pronged: its overall *vision* relating motives of all kinds to automorphic representations, and its *methods* which push representation theory to the forefront, and infuse the subject with a seemingly endless string of challenging problems. It is these aspects of the Langlands Program which (albeit indirectly) play a role in the proof of Fermat's Last Theorem. (Gelbart 1997, p. 191)

6.1 Frey-Hellegouarch Curves

Frey (1982) was apparently the first to associate with *every* integer solution to F_n

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

the elliptic curve defined by (Since Eq. 11 has nonzero term in x^2, it is not in the form of Eq. 8 used to define elliptic curves. But a linear change of variable $X' = X - ((a^n - b^n)/3)$ puts it in that form. Such issues often go without comment in the literature):

$$Y^2 = X(X - a^n)(X + b^n). \tag{11}$$

This is the Frey curve or Frey-Hellegouarch curve for the solution $\langle a, b, c\rangle$.

The goal would be to show this curve is impossible, and conclude there are no solutions to F_n. Frey suggested in a privately circulated note this curve could not be *modular,* as discussed in Sect. 6.2. A much fuller discussion was published with input from Serre (Frey 1986). See also Schappacher (1993/1994, p. 41).

This curve is better tied to F_n than appears at first. Obviously the roots of the righthand side are 0, and a^n and $-b^n$. But look at the differences between them:

$$a^n = a^n - 0b^n = 0 - (-b^n)c^n = a^n - (-b^n).$$

These differences are more intrinsic to the geometry of the curve than the roots are. A mere shift in the X-coordinate does not change the geometry. It changes the roots but does not change their differences. The most important classical invariant of a polynomial is its *discriminant,* defined by multiplying together the differences between its roots. The most obvious fact about the discriminant is that it equals 0 if and only if the polynomial has at least one multiple root.

6.1.1 Some Details of Arithmetic and Geometry

First, given $a^n + b^n = c^n$ with $abc \neq 0$, the discriminant of $X(X - a^n)(X + b^n)$ *is* $-(abc)^{2n}$ and thus nonzero. So the Frey-Hellegouarch curve is elliptic. Its plot over the real numbers $\mathbb{R}$ has no self-crossings. It cannot look like Fig. 2.

Second, if a and b are divisible by a prime number ℓ, then $a^n + b^n = c^n$ implies c is too. And any common factor of a, b, c can be divided out to give a smaller solution to F_n.

$$(k \cdot h)^n + (k \cdot i)^n = (k \cdot j)^n \text{ implies } h^n + i^n = j^n \text{ when } k \neq 0.$$

So, dividing out common factors, we can assume no prime divides both a and b. Such a solution to F_n is called *primitive* since it is not a multiple of any smaller solution.

Work on FLT requires calculating modulo various primes ℓ. So it is helpful that, for a primitive solution $\langle a, b, c\rangle$, the roots a^n and $-b^n$ of the righthand polynomial in Eq. 11 cannot both equal the root 0 modulo one prime ℓ. Intuitively, doing algebraic geometry modulo ℓ, the Frey-Hellegouarch curve looks at worst like Fig. 2. It cannot degenerate beyond one self-crossing. In the standard terminology, when $n > 3$, the Frey-Hellegouarch curve for a primitive solution to F_n is *semistable.*

6.2 Modularity

> I want to spend a few minutes considering one example (a conjecture, in fact) which shows how Number Theory can sometimes contrive to be a helpful, and possibly inspirational, goad to the rest of the Mathematical Sciences. (Mazur 1991, p. 593).

There is an obvious kinship between the geometry of the complex plane $\mathbb{C}$ and of the Euclidean plane $\mathbb{R}^2$ The absolute value of a complex number $a + b\sqrt{-1}$ equals the Euclidean norm of the vector $\langle a, b\rangle \in \mathbb{R}^2$:

$$| a + b\sqrt{-1} | = \left(a^2 + b^2\right)^{\frac{1}{2}} = \left\| \langle a, b\rangle \right\|.$$

By mid-nineteenth century, a similar link was known between the upper half of the complex plane on one hand, and the hyperbolic plane of non-Euclidean geometry on the other (McKean and Moll 1999, Sect. 9).

$$\mathbb{H} = \left\{ a + b\sqrt{-1} \in \mathbb{C} | b > 0 \right\}.$$

Essentially every account of elliptic curves describes how they can be covered by the complex numbers so that the curve gets a locally Euclidean structure. This has been central to the subject since the nineteenth century.

Mazur (1991) explains a late twentieth century discovery by Gennadii Vladimirovich Belyi: An elliptic curve can also get a locally hyperbolic structure from the upper half complex plane $\mathbb{H}$, if and only if it is defined by Eq. 8 with algebraic numbers *A, B* as coefficients (i.e., *A, B* are themselves the roots of integer polynomials).

A local Euclidean structure plus a local hyperbolic structure on one curve can have interesting interactions. But Mazur (1991) describes how much more happens when the local hyperbolic structure has a particular arithmetic form. Elliptic curves meeting that condition are called *modular* elliptic curves. Frey had suggested that Frey-Hellegouarch curves cannot be modular. That suggestion was proved by Ribet (1990).

The conjecture described in the quote of Mazur was called the conjecture of Taniyama-Shimura-Weil. Now it is the *Modularity Thesis* or *Modularity Theorem* (MT). It says every arithmetic elliptic curve is modular! Given Ribet's proof, a proof of MT would prove FLT. Indeed, by the considerations described in Sect. 6.1.1, FLT would follow just from proving every semistable elliptic curve is modular. That seemed out of reach to most experts in 1991, but Wiles (1995a) would prove it.

7 The Ideas and Methods of Ribet's and Wiles's Proofs

The proofs discussed up to here used extremely sophisticated conceptions and calculations but – with the notable exception of Faltings (1983) – did not use machinery so advanced that a graduate student in number theory would not normally know it. The success of Faltings (1983) and the progress through Frey (1986), plus the whole tendency of the Langlands Program, brought such machinery to the fore.

7.1 Ribet

Ribet (1990) proved an arithmetic conjecture by Serre. It turned out the most useful calculational handle on modular covers of elliptic curves, for this purpose, is certain technical witnesses to such covers. These witnesses are called *newforms.* Every modular cover of a given elliptic curve will be witnessed by many different newforms. Ribet showed that if a Frey-Hellegouarch curve did admit a hyperbolic structure making it modular, then any newform witnessing it could be reduced to one on a lower *level,* and then a lower one, to the point of impossibility. So Frey-Hellegouarch curves cannot have such witnesses and cannot be modular.

The essence of the argument is geometrically motivated calculation. The steps of the proof are largely calculational, albeit the calculations are often organized in terms of uncountable structures like the real, complex, and p-adic numbers, as well as the absolute Galois group $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$ named in the paper's title.

As is typical of major proofs today, simply doing the calculations does not suffice to prove they work. Ribet uses considerable resources over several sections. He names just the chief ones in his summary:

> In §2, we recall material due to Raynaud [24] concerning Neron models of Jacobians. . .. In the next two §§, we recall the work of Deligne and Rapoport [4] and Cerednik-Drinfeld ([3], [7]) on the bad reduction of classical modular curves. . .. [To prove] the Main Theorem [w]e begin, in §6, with Mazur's result [20]. . .. (Ribet 1990, p. 433)

It should be obvious on its face that to bring all these resources into one comprehensible argument, the sources themselves must give concisely quotable theorems in relatively uniform terminology. This has to be done well to make proofs like Ribet's feasible to discover, and to make them reliable once given.

7.1.1 Structural Mathematics

Indeed the working methods are *structural* in that they define their objects just up to isomorphism, as philosophers have been remarking for some time. However, this has different goals and means in mathematics than philosophers usually discuss. Benacerraf (1965) depicts two young mathematicians using different "set theoretic" definitions of the natural numbers just because they were taught to. Benacerraf notes the two definitions work equally well because they produce isomorphic "structures," and proposes philosophers explore this as an issue for the ontology of mathematics. But it is also practical mathematics.

Generally, in practice, different constructions of one structure do not "work equally well." Different ones work best depending on context. And more:

1. A result proved in one context may be needed in another.
2. Efficiency may demand using a (reliably proven) result without lingering to specify any one construction or proof or context for its proof.

So mathematicians already have efficient, rigorous, routine ways to describe common structure behind different constructions.

The concept of structure that works for Ribet and his sources is the language of exact sequences and functors – in short categorical language. Some mathematicians will stress that Ribet and the others are not doing category theory. They are not lingering on categorical issues. But they use that language. That is how it is done.

7.2 Wiles

> To set the stage for the proof in (Wiles 1995a) one begins by replacing the problem on elliptic curves with a problem on Galois representations. (Wiles 1995b, p. 244)

After Ribet's proof, it remained to show all elliptic curves are modular – or at least all semistable ones – and that would prove FLT. Instead of working directly with hyperbolic covers of elliptic curves, Wiles used *Galois representations* which are algebraic constructs associated to elliptic curves. Wiles found modular elliptic curves can be recognized by their associated Galois representations, so these ones can be called *modular* representations. Then he used Mazur's theory of *deformations* of Galois representations (Mazur 1997).

Through all the technicalities, Wiles' proof has a clear conceptual outline often called an "$R = T$ strategy." In Wiles' case:

1. R captures a wide sense of "deformations" of modular representations of semistable elliptic curves, which reaches all Galois representations of those curves.
2. T captures a narrow sense of "deformations" of modular representations of semistable elliptic curves so the results of deformation remain modular.

The expression "$R = T$" is shorthand for proving the reach of T actually covers the reach of R (Saito 2013, Chap. 5). All the relevant Galois representations are reached from modular ones by deformations that preserve modularity. So all semistable elliptic curves are modular.

Of course Wiles' conceptual strategy did not erase concrete problems. Just the contrary. It pointed out certain concrete problems as decisive. The most famous example from arithmetic is the "3–5 trick," the topic of (Saito 2013, Chap. 4) and of many other publications and webpages. Wiles makes sophisticated use of a concrete fact of arithmetic on elliptic curves: the Abelian group $E(\mathbb{Q})$ of an arithmetic elliptic curve E cannot have both a point $r \neq 0$ with $3r = 0$ and a point $s \neq 0$ with $5\ s = 0$. It might have either one but cannot have both. This is not easy to prove. It was well known to Wiles since one proof of it is Mazur (1978). See Wiles (1995a, p. 542) or Saito's Proof of Proposition 4.3.

Wiles solved massive specific problems of many kinds to define the R and T he needed, and prove their relation. Wiles (1995a) gives 100 pages of detail, and that is dauntingly concise for all but leading experts. This has led to many more results on modularity and applications in number theory. And it finished the proof of FLT.

Here the numerous outside results needed, and the Grothendieckian functorial-categorical nature of the arguments, are even more prominent than in Ribet's case.

8 The Quondam Problem of Grothendieck Universes

> To avoid certain logical difficulties, we will accept the notion of a Universe, which is a set 'large enough' that the habitual operations of set theory do not go outside of it. (Grothendieck 1971, p. 146)

Whether this avoided difficulties or caused them has been a matter of controversy. Some mathematicians explicitly accept the large sets so as to use Grothendieck's large structure tools. Others reject the tools, or at least avoid using them in proofs, because of set theory. Everyone involved knew very well from the start how to give the specific number theoretic or geometric applications without these large structure tools. But the tools themselves were described, notably in Artin et al. (1972), using large cardinals which create a stronger set theory than ZFC.

The proofs bearing on FLT, including Ribet (1990) and Wiles (1995a), never mention Grothendieck universes. They have no reason to. But they refer to work in the Grothendieck school that does use them (McLarty 2010).

The set theoretic issue is moot now since even Grothendieck's largest structure tools turn out to be available in *finite order arithmetic,* far weaker than ZFC. Set theory at this level has been reinvented in many guises. It can take the form of Simple Type Theory (with arithmetic), or the Elementary Theory of the Category of Sets, or Zermelo set theory with Bounded Separation. Conservative extensions of any of these set theories can formalize Grothendieck's large structure tools by essentially the same definitions, and with the same theorems and proofs, as Grothendieck originally did (McLarty 2020).

Since this has been proved, no one who accepts anything like ZFC need feel any logical qualms about using these tools.

Intuitively, finite order arithmetic posits a set $\mathbb{N}$ of natural numbers, and a power set $P(\mathbb{N})$ containing all sets of natural numbers, and a power set $P^2(\mathbb{N})$ containing all sets of sets of natural numbers, and so on up through any finite level $P^n(\mathbb{N})$. This is a natural foundation for free and easy work in classical and basic abstract mathematics. It is too weak for some well-known purposes (Mathias 2001). But only very low levels of $P^n(\mathbb{N})$ are mentioned much in existing proofs of FLT. Likely no power sets are needed in principle, as discussed in the next section.

9 Peano Arithmetic

> $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Q}_p$, and $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, and the Tate modules are all completions of finitist arithmetical structures. A substantial amount of their first-order theories admits natural finite approximations that do not quantify over second-order entities. Putting proofs of MT into PA will

> involve finding finite approximations of MT and of all the other principles that go into its proof. (Macintyre 2011, p. 15)

Peano Arithmetic (PA) is an axiomatic theory in first order logic, widely described in logic texts and online. On its face, PA deals with the natural numbers

$$\mathbb{N} = \{0, 1, 2, 3 \ldots\}.$$

Methods already familiar to Leopold Kronecker (1882) allow PA to interpret the integers $\mathbb{Z}$, rational numbers $\mathbb{Q}$, and even algebraic numbers $\mathbb{A}$. The algebraic numbers A are all the real or complex roots of polynomials with integer coefficients. Finite lists of numbers can be expressed in PA. So, for example, PA can interpret polynomials with integer coefficients by using their lists of coefficients.

While PA cannot talk about arbitrary real, or complex, or p-adic numbers, it can talk about and prove things about rational approximations to all these. This is crucial to discussing proofs of FLT in PA. Of course PA cannot prove as much about approximations to those sorts of numbers as even Second Order Arithmetic can, let alone stronger set theories. As a rule of thumb, algebraic calculations with these kinds of numbers can be approximated in PA. Arguments that depend on the entirety of some infinite series (or an exact limit to it) are a problem for PA.

Also crucially, PA can define some sets of numbers though these sets will not exist in PA. For example, PA can define prime numbers.

Definition 4 A natural number $n > 1$ is *prime* if and only if no natural numbers *k, m* both >1 have $n = km$.

So the set of prime numbers is called *definable in* PA. Notably, PA can state and prove that every natural number is the product of some finite list of prime numbers (unique up to reordering the list).

For any arithmetic elliptic curve E, PA can define the groups $E(\mathbb{Q})$ and $E_{tor}(\mathbb{Q})$. But when E($\mathbb{Q}$) is infinite, it will not itself exist in PA.

All of the work described above through Sect. 6 fits easily into PA in these ways (except Faltings', which is not used in the proof of FLT). The same is likely true of Ribet (1990) described in Sect. 7. But Ribet deals with the absolute Galois group $\mathrm{Gal}\left(\overline{\mathbb{Q}}/\mathbb{Q}\right)$. Neither that group, nor even its elements, can be defined in PA. Its elements can only be approximated in PA. Ribet gives a sustained argument absorbing much previous work that should be better explored in logic than it is yet.

The Macintyre quote opening this section well indicates the real challenge for proving FLT in PA: finding approximations for the Modularity Theorem. He says

> Spelling out the approximations would be a lengthy enterprise, and here it would be useful to have some metatheorems to apply. (Macintyre 2011, pp. 15 f.)

He points to a wide array of results already known for some areas of the proof, or close to areas of the proof. Number theorists often seek such approximations for

practical use quite apart from logic. They can be hard to find, and informative when found. As to logic:

> [There are] a number of distinct regions of [Wiles'] proof, almost all hitherto unvisited by proof theorists. . .. Some but certainly not all this territory is known to some model theorists. . .. Macintyre (2011, p. 8)

The project has not been completed. And even just to show that FLT is provable in PA in principle, there is no means currently known except to show at least in principle that such approximations can be given.

Then there are other important methods in current number theory, notably étale cohomology, not used in proving FLT and not much explored by logicians. These may well turn out to be easier than MT to get into PA. The language of MT is analytic geometry over the complex numbers, which can go beyond PA. Étale cohomology is closer to arithmetic in the first place.

10 The Wide-Open Question

> I want to stress that it is by no means clear to me right now how much induction is needed for [. . .] a proof of Wiles' theorem in PA. In particular, I think there is little evidence that bounded arithmetic plus the totality of exponentiation would suffice. . .. [O]ne has so little experience of more modern mathematics in the system that one should be cautious. (Macintyre 2011, p. 9)

Bounds in arithmetic are akin to approximations. Proving approximations to MT in bounded arithmetic (with any given kind of bounds, such as exponential) is akin to approximating MT in PA to begin with. There is little evidence on what bounds might work for MT because little is known of approximations to it in PA at all.

As a general fact of logic any theorem provable in PA is provable in some strictly weaker fragment of PA. This is because a proof of a theorem uses only finitely many axioms, and PA proves the consistency of all its finitely axiomatized subtheories (Hájek and Pudlák 1993, p. 168).

But it is a striking fact of proof theoretic experience that many interesting theorems are provable in the specific agreeably weak and conceptually attractive fragment that Macintyre calls bounded arithmetic with exponentiation. It is also called EFA, abbreviating two of its common names *Elementary Function Arithmetic* and *Exponential Function Arithmetic.* See Avigad (2003).

Since Wiles' proof was literally published in the *Annals of Mathematics,* it is a natural test case for Harvey Friedman's Grand Conjecture:

> Every theorem published in the *Annals of Mathematics* whose statement involves only finitary mathematical objects (i.e., what logicians call an arithmetical statement) can be proved in EFA. (Avigad 2003, p. 258)

No clear example is yet known to refute this claim (Early issues of the *Annals* had light articles on superexponential functions in arithmetic but these are not seriously to the point of Friedman's conjecture.)

Whether MT might refute Friedman's Conjecture remains to be known.

The real value of pursuing provability of MT, or FLT, in PA is likely to be not just securing the answer "yes it is provable in PA" or "no it is not provable in PA," but learning what resources from PA those theorems actually use. What fragment of induction suffices? To lift words from the quote of Macintyre above, the value would be to add some experience of more modern mathematics to proof theory.

11 Conclusion

The proof of FLT raises conceptual points listed here with the most relevant sections. Section 6.2 on modular curves, and Sect. 7 on the ideas of Ribet's and Wiles' proofs, are pivotal to the whole argument:

1. It showcases the geometric unity of current mathematics. It refutes persistent rumors that mathematics today is fragmented into specialities and that geometry was neglected in the twentieth century. Sections 4, 5, and 6.
2. It typifies current major proofs as key steps are long and rely on many prior advanced theorems. To make this work, theorems must be stated concisely and precisely, by uniform means. This is the practical reason for *structuralist* mathematics. It is not like the versions of structuralism most philosophers of mathematics consider. Section 7.
3. Far from displacing concrete arithmetic (as critics of "modern abstraction" feared), the modern techniques mobilize it. No one could find such long intricate proofs as these *without* concrete arithmetic as a guide. (The entire article.)
4. Some structural tools originally relied on large sets, though practitioners knew how to cut those tools out of the proof if desired. Now it is known these tools themselves can be rigorously founded without large sets. Section 8.
5. It is overwhelmingly likely that FLT is provable in PA but Macintyre (2011) documents points where this has not been demonstrated. Section 9.
6. The wide-open logical issue is that the leading ideas of the proof are so explicit that FLT is probably provable in some truly finitary fragment of PA. We are far from verifying what fragment. Section 10.

References

Artin M, Grothendieck A, Verdier JL (1972) Theorie des Topos et Cohomologie Etale des Schémas. Séminaire de géométrie algébrique du Bois-Marie, 4. Springer, three volumes, cited as SGA 4, New York

Avigad J (2003) Number theory and elementary arithmetic. Philos Math 11:257–284