The Generalized Fermat Equation

Michael Bennett, Preda Mihăilescu and Samir Siksek

Abstract We survey approaches to solving the generalized Fermat equation

$$x^p + y^q = z^r$$

in relatively prime integers x, y and z, and integers p, q and $r \geq 2$.

2010 Mathematics Subject Classification: 11D41, 11G05, 11R18

Key Words and Phrases: Fermat's Last Theorem, Cyclotomic theory, Elliptic Curves, Modularity

1 Le Roy est mort - Vive le Roy

Pythagoras' formula was purportedly kept secret within the closed circle of his initiates – but, as with any fact of nature, it became eventually widely known: the squares of the cathetes sum to the square of the hypotenuse. In the spirit of arithmetic, spread eight centuries later by Diophantus of Alexandria, one may instead phrase this statement in terms of integral solutions of the equation

$$x^2 + y^2 = z^2, \quad \text{with } x, y, z \in \mathbb{N} \text{ and } \gcd(x, y, z) = 1, \tag{1}$$

M.A. Bennett
Department of Mathematics, University of British Columbia, Canada
e-mail: bennett@math.ubc.ca

P. Mihăilescu
Mathematisches Institut der Universität Göttingen, Germany
e-mail: preda@uni-math.gwdg.de

S. Siksek
Mathematics Institute, University of Warwick, Coventry, CV4 7AL, United Kingdom
e-mail: s.siksek@warwick.ac.uk

or, equivalently, ask for the coordinates of all rational points on the unit circle. All these are variants of the problem appearing in the second book of Diophantus, in the chapter numbered VIII – often quoted accordingly as Diophantus II.VIII. We nowadays call the solutions to equation (1) *Pythagorean triples*. Already in Diophantus one finds parametrizations for these solutions, given by

$$x = 2uv; \quad y = u^2 - v^2; \quad z = u^2 + v^2, \tag{2}$$

where u and v are positive integers. This fact is easy to verify in one direction; the proof that all triples are parametrized in this form is one of the most popular of ancient mathematics and is widely taught to this day.

The work of Diophantus on Arithmetic, a collection of 12 volumes written in Greek, was lost for centuries. Only in the late sixteenth century were half of these books, namely I–III and VIII–X, rediscovered in the form of a later Byzantine transcription. These were translated into latin by Bombelli, and then subsequently published in Basel by Xylander. It was through the annotated translation of Gaspard Bachet from 1621 that the *Arithmetica* finally received a wide diffusion, capturing the interest of mathematicians of the epoch. Among them, Fermat, a lawyer from Toulouse, was particularly impressed by the beauty of Diophantus' solution to (1). It is on the margin of the text to Diophantus' Problem II.VIII that Fermat wrote in 1634 his historical note concerning a short proof of the fact that the equation

$$x^n + y^n = z^n, \text{ with } x,y,z \in \mathbb{N} \text{ and } \gcd(x,y,z) = 1, \tag{3}$$

has no solution for $n > 2$, a proof which the margin of Bachet's book was insufficiently large to contain.

The assertion, henceforth bearing the name Fermat's Last Theorem – hereafter denoted FLT for concision – remained an open problem for more than three centuries. Attempts to prove FLT led to some of the most significant developments in mathematics of the past three hundred years; it is fair to say that it is one of the problems that has generated the most mathematics in history. The first systematic approach, initiated by Kummer in the mid 19th century, was based on the theory of algebraic number fields and in particular *cyclotomic fields*. The conjecture was finally proved in 1994 by Wiles, with the help of Taylor, building on a series of ideas and results, due to Hellegouarch, Frey, Serre and Ribet, that connect the Fermat equation to elliptic curves, modular forms and Galois representations.

Even before Wiles announced his proof, various generalizations of Fermat's Last Theorem had already been considered, to equations of the shape

$$Ax^p + By^q = Cz^r,$$

for fixed integers A, B and C. In the case where $A = B = C = 1$, for reasons we discuss later, we focus our attention on the equation

$$x^p + y^q = z^r, \text{ with } x,y,z \in \mathbb{N}, \ \gcd(x,y,z) = 1 \ \text{ and } \ \tfrac{1}{p} + \tfrac{1}{q} + \tfrac{1}{r} < 1. \tag{4}$$

Perhaps the only solutions to this equation are those currently known; i.e. those with (x,y,z,p,q,r) coming from the solution to Catalan's equation $1^p+2^3=3^2$, and from the following nine identities:

$$2^5+7^2=3^4,\ \ 7^3+13^2=2^9,\ \ 2^7+17^3=71^2,\ \ 3^5+11^4=122^2,$$
$$17^7+76271^3=21063928^2,\ 1414^3+2213459^2=65^7,\ 9262^3+15312283^2=113^7,$$
$$43^8+96222^3=30042907^2 \quad \text{and} \quad 33^8+1549034^2=15613^3.$$

In the mid 1990s, Andrew Beal, a graduate in mathematics with computational interests, in the course of carrying out calculations related to Fermat's equation and its variations, noted that the solutions listed here all have the property that $\min\{p,q,r\}=2$ and made what is now termed the *Beal conjecture*, that there are no non-trivial solutions to (4), once we assume that $\min\{p,q,r\}\geq 3$. A prize for the solution to this problem now amounts to one million U.S. dollars. Other names attached to conjectures about equation (4) include the *Fermat–Catalan conjecture* and the *Tijdeman–Zagier conjecture*.

As we shall see, the few particular cases of these conjectures that have been investigated may well support the hope that this new generalization of Pythagoras' initial equation will also stimulate fascinating new mathematics.

Our main focus for the rest of the paper will be to describe two approaches to proving results about equation (4). The first of these is essentially a generalization of the cyclotomic methods that proved successful for Catalan's equation. The second is the adaptation of Wiles' proof of Fermat's Last Theorem to handle many special cases of equation (4).

2 Cyclotomic approaches and their limitations

Let us start by noticing that in both equations (3) and (4) one may assume all exponents to be prime: indeed, if there is a solution with non-prime exponents, by raising the variables to some power, one obtains a solution with prime exponents. It thus suffices to consider this case and show that it leads to no non-trivial solutions. By an elementary observation, sometimes attributed to Euler, we have that

$$G:=\gcd\left(x\pm y,\ \frac{x^p\pm y^p}{x\pm y}\right)$$

is a divisor of p, provided that the integers x,y are coprime[1]. In this section, we will focus our attention on the special case of (4), given by the *Fermat-Catalan equation*

$$x^p+y^p=z^q,\ \ \text{with } x,y,z\in\mathbb{N} \text{ and } \gcd(x,y,z)=1, \tag{5}$$

[1] One verifies this by letting $t=x\pm y$ and $x=t\mp y$, as well as using the fact that $(t,y)=1$.

where one may hope to apply cyclotomic theory in a way somewhat analogous to that of the Fermat equation. Equation (5) also serves as a generalization of the binary Catalan equation

$$x^p - y^q = 1 \tag{6}$$

which also resisted solution for more than a century (and was finally solved by the second author, eight years after Wiles' remarkable proof of Fermat's Last Theorem). Here *binary* refers to the presence of only two unknowns in this equation, a fact which facilitates an analytic approach to bounding the possible solutions.

Assuming that either (3) or (5) has a non-trivial solution for the prime p – or the prime pair (p,q) – one distinguishes the cases when $G = 1$ and $G = p$. The case where $p \nmid xyz$ has traditionally been termed the *First Case* of FLT, or FLT1; we retain this designation for the Fermat–Catalan equation. If $G = p$ and thus $p \mid xyz$, one may assume, in equation (3) at least, that $p \mid z$. It is further easy to show that, in this case,

$$p^2 \nmid \frac{x^p \pm y^p}{x \pm y}\,.$$

Since the Fermat equation is homogeneous, the assumption that x, y and z are pairwise coprime is straightforward – otherwise one could divide by the common divisor in a solution and obtain one in coprime integers. In the case of (5), however, with a little work, one can always construct infinitely many "trivial" solutions for which x, y and z fail to be coprime.

2.1 History of "Fermat's Last Theorem"

As is well known, Fermat left no published proof of his conjecture. He did, however, provide a beautiful argument in the biquadratic or quartic case. To be precise, he considered the more general equation

$$x^4 + y^4 = z^2 \tag{7}$$

and, using some astute manipulation of Pythagorean triples, proved that if (x,y,z) is a non-trivial solution of (7) in, say, positive integers, then one can construct a further positive solution (x',y',z') of the same equation, in which $z' < z$. By repeating the procedure one eventually finds a solution with $z' = 1$, which implies that $x' \cdot y' = 0$, a contradiction. This was the first instance of the method of *infinite descent* in number theory. Euler gave a proof of the cubic case using such an argument; Gauss gave later an alternative proof using congruences. Although elementary, both methods require quite intricate computations and are not easy to memorize. We provide here a short

elementary proof, which uses some more recent ideas, that date back to Wieferich and Furtwängler[2]:

Lemma 1. *The equation $x^3+y^3+z^3=0$ has no non-trivial integer solutions.*

Proof. Assume that (x,y,z) is a non-trivial solution in coprime integers. Let

$$\rho = \frac{-1+\sqrt{-3}}{2}$$

be a third root of unity and $\mathscr{E} = \mathbb{Z}[\rho]$ be the ring of Eisenstein integers, which is a Euclidean ring. We assume first that $3 \nmid xyz$, so that both $x+y$, and x^2-xy+y^2 are cubes, say $x+y=s^3$, and

$$(x+\rho y) = (a+b\rho)^3$$

is the cube of a principal ideal. Exactly one of x,y,z must be even – we shall thus assume that $y=2v$ is even. Since the units of $\mathscr{E}$ are the sixth roots of unity, there is some $\delta \in <-\rho>$ such that

$$\begin{aligned}\delta(x+y\rho) = (a+b\rho)^3 &= a^3+b^3+3ab\rho(a+b\rho)\\ &= a^3+b^3-3ab^2+3ab(a-b)\rho.\end{aligned}$$

If $\delta = \pm 1$, then comparing coefficients implies that $y \equiv 0 \bmod 3$, contradicting our assumption. We thus have that

$$x+y\rho = \pm\rho^c(a+b\rho)^3, \text{ with } c = \pm 1\,.$$

We have chosen $y \equiv 0 \bmod 2$, whereby $x\rho^{-c} \equiv w^3 \bmod 2\mathscr{E}$ for some $w = \pm(a+b\rho) \in \mathscr{E}$. But $x \equiv x+y = s^3 \bmod 2$, whence we may conclude that

$$\rho^{-c} \equiv (w/s)^3 \bmod 2\,.$$

The ideal $\mathfrak{p} := (2) \subset \mathscr{E}$ is prime; let $\pi : \mathscr{E} \to \mathbb{F}_{2^2}$ be the natural projection. Since $c \not\equiv 0 \bmod 3$, it follows that $\pi(w/s) \in \mathbb{F}_{2^2}$ is a primitive 9-th root of unity, an impossibility. It remains, then, to consider the case where $3 \mid xyz$; we may suppose, without loss of generality, that $3 \mid z$. Appealing to (9), we find that there is a root of unity δ such that $(\alpha) = \delta\left(\frac{x+y\rho}{1-\rho}\right) = (a+b\rho)^3$ and $x+y = 9s^3$. Since $\frac{1-\rho}{1-\overline{\rho}} = -\rho$, we obtain after dividing the previous identity by its complex conjugate, that there is another root of unity, say δ', such that

$$\frac{x+y\rho}{x+y\overline{\rho}} = \frac{2x-y+y\sqrt{-3}}{2x-y-y\sqrt{-3}} = \delta' \cdot \left(\frac{a+b\rho}{a+b\overline{\rho}}\right)^3 \equiv \delta' \bmod 3\sqrt{-3}\cdot\mathscr{E}. \tag{8}$$

If $2x \not\equiv y \bmod 3$, letting $y' \equiv y/(2x-y) \bmod 3$, the last identities imply that

[2] The second author found this proof, confronted with the own incapacity to recall the details of the classical proofs, for a seminar. It is possible that the proof may have been known, but we found no reference to it in the literature

$$\frac{1+y'\sqrt{-3}}{1-y'\sqrt{-3}} \equiv \delta' \bmod 3\sqrt{-3}\cdot\mathscr{E},$$

whence

$$y' \equiv y \equiv 0 \bmod 3,$$

contradicting the assumption that $3 \mid z$ and $\gcd(x,y,z) = 1$. Finally, suppose that $2x - y = -3d$. Inserting this value into (8) yields

$$\frac{d\sqrt{-3}+y}{d\sqrt{-3}-y} \equiv \delta' \bmod 3\sqrt{-3}\cdot\mathscr{E},$$

whereby $\delta' = -1$ and $d \equiv 0 \bmod 3$. Then $2x - y \equiv x + y \equiv 0 \bmod 9$. Summing these two congruences, we find $3x \equiv 0 \bmod 9$, and thus $3 \mid x$, a contradiction which completes the proof.

□

After Euler and Gauss, the quintic and septic cases of FLT were solved with contributions from Dirichlet, Lamé, Legendre and Cauchy. In his proof of the quintic case, Dirichlet distinguished the cases $p \nmid xyz$ and $p \mid xyz$, using descent for the proof of the second case. Lamé announced in 1841 a full proof of the general case of FLT. Unfortunately, his proof relied implicitly upon the (incorrect) assumption that the integers of the form

$$\alpha = \sum_{k=0}^{p-2} a_k \zeta^k,$$

with $a_k \in \mathbb{Z}$ and ζ a p-th root of unity, form a ring with unique factorization. Kummer demonstrated that this assumption is, in general, false and that the smallest prime for which it fails is $p = 23$. In order to circumvent this difficulty, Kummer proceeded with an investigation of divisibility in the rings of algebraic integers of the p-th cyclotomic field, and introduced the notion of *ideal numbers*, a larger group in which unique factorization was recovered. This work, stemming from a desire to attack the Fermat problem, led, in the following decades, to the theory of ideals and the work of Dedekind, giving rise to the fundamental results underlying what we presently know as algebraic number theory.

If p is an arbitrary odd prime, we let $\mathbb{K} = \mathbb{Q}[\zeta]$ denote the p-th cyclotomic extension. The algebraic integers of this field are $\mathscr{O}(\mathbb{K}) = \mathbb{Z}[\zeta]$ and the ideals of this ring factorize uniquely as products of prime ideals. If I is the semigroup of ideals and P the one of principal ideals, i.e. ideals generated by single elements, the quotient $\mathscr{C}(\mathbb{K}) = I/P$ is a finite abelian multiplicative group, the class group. The prime p is called *regular*, if p does not divide the size $h(\mathbb{K}) = |\mathscr{C}(\mathbb{K})|$ of the class group, and *irregular* otherwise. With respect to the Fermat equation, we have in $\mathbb{Z}[\zeta]$ one of the following factorizations :

$$x^p = (x+y)\cdot\left(\frac{x^p+y^p}{x+y}\right) = (x+y)\cdot\prod_{c=1}^{p-1}(x+y\zeta^c), \quad \text{or}$$
$$z^p = p(x+y)\cdot\left(\frac{x^p+y^p}{p(x+y)}\right) = p(x+y)\cdot\prod_{c=1}^{p-1}\left(\frac{x+y\zeta^c}{1-\zeta}\right),$$

in case of FLT1 or FLT2, respectively. In either situation, writing $\alpha = \frac{x+\zeta y}{(1-\zeta)^e}$, with $e=0$ in the first case and $e=1$ in the second, one finds that α is coprime to $p(x+y)$ and is, in fact, a p-th power, albeit not one of an algebraic integer of $\mathbb{K}$, but rather of the ideal $\mathfrak{A} = (\alpha, z)$. We thus have

$$(\alpha) = \mathfrak{A}^p \quad \text{and} \quad N_{\mathbb{K}/\mathbb{Q}}(\alpha) = \frac{z^p}{p^e(x+y)}. \tag{9}$$

We note at this point that, in the case of the Fermat-Catalan equation, the same construction leads again to $(\alpha) = \mathfrak{A}^q$. Using the connection between class groups and factorization of ideals, Kummer proved his fundamental result on Fermat's Conjecture:

Theorem 1 (Kummer). *Equation (3) has no solutions for regular primes $p > 2$.*

For regular primes, it follows from (9) that there exists a $\rho \in \mathbb{Z}[\zeta]$ such that $(\alpha) = (\rho)^p$ is the p-th power of a principal ideal. Starting from this, the proof of FLT1 is relatively simple. For the second case, however, Kummer appealed to a sophisticated variant of infinite descent in the real field $\mathbb{K}^+ \subset \mathbb{K}$ – the method bears currently his name, Kummer descent. A modern, complete proof of this result can be found in the book of Washington [36], Chapter IX. One finds in Rassias' lovely introductory work for undergraduates [26], on page 147, more biographical details of Kummer's life.

Kummer's work was followed by a century of active research on the Fermat equation, which led to the establishment of a large number of conditions known to imply the truth of the Fermat Conjecture – see e.g. Ribenboim's famous survey [27]. However, before Wiles's breakthrough, there were only two known results known to hold for infinitely many exponents, namely the "elementary" proof [34] given by Terjanian in 1977 for the fact that (3) has no solution for *even* exponents $n > 2$ with $n \nmid xy$, as well as the deep analytic proof of Adelman, Fouvry and Heath-Brown, which showed that FLT1 holds for infinitely many primes[3]

There are, however, a number of results of interest on FLT that were established via cyclotomic methods before Wiles' proof. We review here a few of the most important of them:

(i) Wieferich and then Mirimanoff and Furtwängler proved that if FLT1 has a non-trivial solution, then

[3] One would expect, for various reasons, that regular primes occur more frequently than irregular ones. If we knew this, since it has been proved that there are infinitely many irregular primes, Kummer's result would already imply that there are infinitely many primes p for which FLT holds. However no proof of the fact that the set of regular primes is infinite is known, even now.

$$a^{p-1} \equiv 1 \bmod p^2, \quad \text{for } a \in \{2,3\}. \tag{10}$$

Variations on this theme was treated during the following decades by, among others, Morichima, Lehmer, Skula, Granville and Monagan, the last two of these eventually proving that if FLT1 had non-trivial solutions, then (10) holds for all primes $a \leq 89$. With this Granville and Monagan were able to prove that FLT1 has no solutions for

$$p < 714,591,416,091,389.$$

(ii) Eichler proved that if FLT1 has non-trivial solutions, then the p-rank of the p-part of the class group $A := \mathscr{C}(\mathbb{Q}[\zeta])_p$ of the p-th cyclotomic field is necessarily *large*, namely $p\text{-rk}(A_p) \geq \sqrt{p} - 1$. He thus improved upon an earlier result of Krasner, who had proved that if FLT1 had solutions and $p > n_0 = (45!)^{88}$, then the Bernoulli numbers $B_{p-1-2i}, i = 1,2,\ldots,\lfloor (\log p)^{1/3} \rfloor$ had numerators divisible by p; this implies in particular that $p\text{-rk}(A_p) \geq \lfloor (\log p)^{1/3} \rfloor$.

(iii) In the first half of the 20th century, Harry Schulz Vandiver wrote extensively on Fermat's equation, partially fixing some gaps in earlier proofs of Kummer (and leaving a number of gaps himself, which were fixed only at the end of the century). We present below his main result as part of Theorem 2.

Bearing in mind the fact that the Fermat Conjecture has been proved, it is still of interest to analyze other approaches which may provide alternative proofs of this Theorem. There are currently two primes known for which the Wieferich condition (10) is satisfied with $a = 2$; none are known with $a = 3$. If one admits the heuristic assumption that the vanishing of a Fermat quotient

$$\varphi_p(a) \equiv \frac{a^{p-1} - 1}{p} \bmod p$$

has probability $1/p$, one may expect on *average* $O(\log\log(X))$ primes $p < X$ for which the quotient $\varphi_p(a) = 0$ for some fixed $a < p$. However, the same heuristic argument suggests that one can find at most one prime for which two or more Fermat quotients vanish simultaneously. One may formulate the following:

Conjecture 1. There exists a constant $c \geq 2$ such that for every prime $p \in \mathbb{N}$ there are less than c values $a \in \{2,3,\ldots,p-2\}$ with $\varphi_p(a) = 0$.

If $c < 87$, this conjecture implies FLT1.

Concerning the criteria in group (ii), Washington provides in [36] heuristic arguments suggesting that

$$p\text{-rk}(A_p) \ll O(\log(p))$$

for all primes. The Theorem of Eichler would imply FLT1 even if a rather weaker conjecture holds:

Conjecture 2. There is an integer $a > 2$ such that for all primes p, the p part of the class group of the p-th cyclotomic field has rank $p\text{-rk}(A_p) < p^{1/a}$.

We have mentioned above that there are infinitely many irregular primes. The first of these were discovered by Kummer, the smallest one being $p = 37$. However, if one instead considers the largest *real* subfield contained in $\mathbb{K}$, which is $\mathbb{K}^+ = \mathbb{Q}[\zeta + \overline{\zeta}]$, the class number h^+ of this field is apparently much smaller and seems to never be divisible by p. Kummer was the first to suggest, in a letter to Kronecker from 1852 (see [23]), that this fact might hold for all p. The fact played an important role in many of the papers of Vandiver, who was seemingly unaware of Kummer's letter. The assumption that $p \nmid h_p^+ = |\mathscr{C}(\mathbb{K}^*)|_p$ is therefore called the *Kummer-Vandiver* Conjecture, or simply the Vandiver Conjecture. The conjecture has also deep implications in K-theory; it has been verified numerically for all primes $p < 2^{27}$. We should mention, however, that there are specialists who accept some heuristic arguments which suggest that the Conjecture might have counterexamples that are as scarce as the Wieferich primes. If this were true, there might be as many as $\log\log(X)$ primes $p < X$ for which the conjecture is false. Those who believe the Vandiver Conjecture are guided by the fact that there are numerous striking and rather improbable consequences to $p \mid h_p^+$, and therefore the heuristic assumption that the value of the residue $h_p^+ \mod p$ is uniformly distributed may be false.

In the context of Fermat's Last Theorem, two additional conditions of a rather specialized nature play a role; we formulate them also as assumptions: they have been computationally verified to hold within the same range as the Kummer–Vandiver Conjecture.

Assumption C Assume that the exponent of A_p is p, whereby the p-part of the class group of the p-th cyclotomic field is annihilated by p.

Assumption D Assume that all the units $\delta \in \mathbb{Z}[\zeta_p + \overline{\zeta}_p]^\times$ for which there exists an algebraic integer $\rho \in \mathbb{Z}[\zeta_p + \overline{\zeta}_p]$ such that $\delta \equiv \rho^p \mod p^2\mathbb{Z}[\zeta_p + \overline{\zeta}_p]$ are global p-th powers.

With these definitions, the following theorem holds:

Theorem 2. *Suppose that the Kummer–Vandiver Conjecture holds. If, additionally, Assumption C holds, then FLT1 has no solutions. If, instead, Assumption D holds, then FLT2 has no solutions.*

The first of these claims dates back to Vandiver, who, however supposed only that $p \nmid h_p^+$ and had not noticed the necessity of Assumption C. The correct statement was discovered by Sitaraman [32] in 1995. The second part of the theorem, together with its proof, are due to Kummer. We note that Theorem 2 provides the only known cyclotomic criterium which implies that there are no solutions to FLT2.

2.2 The Catalan equation

Due to results of Victor Lebesgue (1853) and Chao Ko (1962), which eliminated the cases of even exponents, the Catalan Conjecture was reduced to proving that

$x^p - y^q = 1$ has no solution with odd prime exponents (which are easily seen to be distinct). By considering cyclotomic factorizations, similar to the ones in (9), one obtains four cases. Cassels proved in 1962 that if (6) has a solution with odd exponents, then $p \mid y$ and $q \mid x$, while

$$\frac{x^p - 1}{p(x-1)} = v^q$$

for some rational integer v. One may define

$$\alpha = \frac{x-\zeta}{1-\zeta} \text{ and } \mathfrak{A} = (\alpha, y),$$

whence the analogue of equation (9) is $(\alpha) = \mathfrak{A}^q$.

Catalan's conjecture now follows from a combination of analytic methods with algebraic properties of cyclotomic fields. We present a brief exposition of some of the ideas that made this proof possible. Denoting by G the Galois group $\mathrm{Gal}(\mathbb{Q}[\zeta_p]/\mathbb{Q})$, one notices that the group ring $\mathbb{F}_q[G]$ acts on the class a of the ideal $\mathfrak{A}$; in other words, linear combinations of the type $\theta = \sum_{\sigma\in G} n_\sigma \cdot \sigma$, in which the integers n_σ are identified with their remainders modulo q, will act on the class a according to $a^\theta = \prod_{\sigma\in G} \sigma(a)^{n_\sigma}$. Since $a^q = 1$, we see that it suffices to consider $n_q \in \mathbb{F}_q$. We call an element $\theta \in \mathbb{F}_q[G]$ an *annihilator* of a if $a^\theta = 1$.

Suppose that we are able to find a non-trivial annihilator $(1+j)\theta \in \mathbb{F}_q[G]$ – here we denote the restricted action of complex conjugation to $\mathbb{K}$ by $j \in G$, so for instance $a^j = \overline{a}$. Then $\alpha^\theta = (\rho)^q$, for some $\rho \in \mathbb{K}^+$; the equality between principal ideals translates into an identity between algebraic numbers, involving an unknown unit ε: $\alpha^\theta = \varepsilon \cdot \rho^q$. Assume additionally, that there is a further $\theta' \in \mathbb{F}_q[G]$ such that $\varepsilon^{\theta'} \in (\mathbb{K}^\times)^q$. Given this, one is able to prove, using additional arguments about the structure of units in $\mathbb{K}$, that there is a number $\nu \in \mathbb{Z}[\zeta + \overline{\zeta}]$ such that $(x-\zeta)^{\theta\cdot\theta'} = \nu^q$: note that we eliminated the denominator of α! That these favourable assumptions situation can be shown to occur follows from an important theorem of Francisco Thaine [35] (which also leads to a cyclotomic proof of the Main Conjecture of one dimensional Iwasawa Theory).

Continuing, one now finds a multiple $\psi = \sum_{\sigma\in G} r_\sigma \sigma \in \mathbb{F}_q[G]$ of $\theta \cdot \theta'$ such that $\sum_{\sigma\in G} r_\sigma = hq$ for some $h \leq \frac{p-1}{2}$, leading to the following equation

$$\nu = x^h (1 - \zeta/x)^{\psi/q}.$$

The fact that $\nu \in \mathbb{R}$ has the important advantage that the rapidly converging binomial series expansion of the expression

$$(1-\zeta/x)^{\psi/q} = \prod_{\sigma\in G} (\sigma(1-\zeta/x))^{r_\sigma/q} \tag{11}$$

will in fact converge to ν/x^h (rather than to some number that differs from ν/x^h by a q-th root of unity, as would be true generally). With this, we obtain $\nu = x^h g(1/x) +$

$F(1/x)$, with $g \in \mathbb{K}[X]$ being a polynomial of degree h and $F(T) \in \mathbb{K}[[T]]$ a power series. Finally, appealing to some lower bounds on A which were obtained by Hyyrö, one can eventually show that, under the given arithmetic conditions, $F(1/x) = 0$. We thus have $v = x^h g(1/x)$, which leads to an arithmetic contradiction, completing the proof of Catalan's Conjecture.

2.3 The Fermat – Catalan equation

As previously mentioned, in the case that (x,y,z) is a non-trivial solution to the Fermat-Catalan equation (5) with odd exponents p and q, if we let $\alpha = \frac{x+\zeta y}{(1-\zeta)^e}$ and $\mathfrak{A} = (\alpha, z)$ – where $e = 1$ if $p \mid z$ and $e = 0$ otherwise, then we have $\mathfrak{A}^q = (\alpha)$, a situation which is reminiscent of both the Fermat and the Catalan equations, and a starting point for cyclotomic investigations of (5).

In this direction, the second author has tried to adapt the proof of Kummer's Theorem to the case of equation (5). It would take too long to explain here the main points in which this equation differs from (3), necessitating the introduction of additional methods. Let us only mention that it is possible to restrict our attention to six cases, depending on whether or not p or q divide any of the factors of $x \cdot y \cdot z \cdot (x \pm y)$. After proving a generalization of Kummer's descent to the $p \cdot q$-th cyclotomic field, it was often possible to either discard cases, or reduce them to conditions about the vanishing of some Fermat quotient – e.g. $2^{q-1} \equiv 1 \bmod q^2$ or $p^{q-1} \bmod q^2$. This approach succeeds in five cases. Unfortunately, in the sixth case, all classical Kummerian methods apparently fail. As a consequence, the second author was unable to find conditions on p and q which imply that (5) has no solutions. By symmetry, a set of such conditions could however be derived for the *rational Catalan* equation, i.e. the equation (6) in which $x, y \in \mathbb{Q}$ is allowed. Note that, after clearing denominators, this equation is equivalent to

$$X^p + Y^{pq} + Z^q = 0\,,$$

which may be viewed as a "symmetrized Fermat – Catalan" equation.

This first attempt to apply cyclotomic methods thus appears to confirm the somewhat pessimistic expectation that they are not sufficient for solving equation (5) in any generality.

3 Fermat's Last Theorem

In the previous section, we discussed the cyclotomic approach to the Fermat equation and its potential limitations. We now sketch the approach of Hellegouarch, Frey, Serre and Ribet which culminated in Wiles' proof of Fermat's Last Theorem.

Theorem 3 (Wiles [37]). *The only integer solutions to the Fermat equation*

$$x^n + y^n = z^n$$

with $n \geq 3$ *satisfy* $xyz = 0$.

Recall that we call a solution *trivial* if $xyz = 0$, otherwise it is called *non-trivial*. Thus the theorem states that all solutions to the Fermat equation are trivial. As we have seen, the theorem is true for exponents $n = 3$ and 4. Thus it is sufficient to show, for primes $p \geq 5$, that all solutions to

$$x^p + y^p + z^p = 0 \tag{12}$$

are trivial. Of course, if (x,y,z) is a solution, we may by scaling suppose that $\gcd(x,y,z) = 1$; we call such a solution *primitive*. The purpose of this section is to sketch the proof of Fermat's Last Theorem and the ideas leading to it. The proof is based on three main pillars:

(i) Mazur's Theorem on irreducibility of Galois representations of elliptic curves;
(ii) The modularity theorem, due to Wiles, Breuil, Conrad, Diamond and Taylor;
(iii) Ribet's level lowering theorem.

Explaining these pillars will involve a detour into some of the most fascinating areas of modern number theory: elliptic curves, Galois representations, modular forms and modularity.

3.1 Elliptic curves

There are many possible definitions of an elliptic curve. Let K be a field. An *elliptic curve* over K is a curve of genus 1 defined over K with a distinguished K-point. An alternative definition is: an *elliptic curve* over K is a 1-dimensional abelian variety over K. The simplest (though conceptually least enlightening) definition is: an *elliptic curve* E over K is a smooth curve in $\mathbb{P}^2$ given by an equation of the form

$$E \ : \ y^2z + a_1xyz + a_3yz^2 = x^3 + a_2x^2z + a_4xz^2 + a_6z^3,$$

with a_1, a_2, a_3, a_4 and $a_6 \in K$. This in fact is a curve of genus 1, and the distinguished K-point is $(x : y : z) = (0 : 1 : 0)$. If the characteristic of K is not 2 or 3, then we can transform to a much simpler model given in $\mathbb{A}^2$ by

$$E \ : \ Y^2 = X^3 + aX + b, \tag{13}$$

where a and $b \in K$. We call this equation a *Weierstrass model*. Let

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

which we call the *discriminant* of E (this is -16 times the discriminant of the polynomial on the right-hand side). The requirement that E is smooth is equivalent to

the assumption that $\Delta \neq 0$. The distinguished K-point is now the 'point at infinity', which we denote by ∞ or $\mathscr{O}$. Given a field $L \supseteq K$, the set of L-points on E is given by

$$E(L) = \{(x,y) \in L^2 \,:\, y^2 = x^3 + ax + b\} \cup \{\mathscr{O}\}.$$

It turns out that the set $E(L)$ has the structure of an abelian group with $\mathscr{O}$ as the identity element. The group structure is easy to describe geometrically: three points P_1, P_2, $P_3 \in E(L)$ add up to the identity element if and only if there is a line ℓ defined over L meeting E in P_1, P_2, P_3 (with multiplicities counted appropriately). The fact that $E(L)$ is an abelian group (where the group operation has a geometric interpretation) ties in with the fact that E is an abelian variety.

Theorem 4 (The Mordell–Weil Theorem). *Let K be a number field and E an elliptic curve over K. Then $E(K)$ is a finitely generated abelian group.*

When K is a number field we refer to the group $E(K)$ as the *Mordell–Weil group of E over K*.

Example 1. As an example, consider the Fermat degree 3 equation over $\mathbb{Q}$:

$$x^3 + y^3 = z^3. \tag{14}$$

Viewed as a curve in $\mathbb{P}^2$, this is in fact a curve of genus 1. Let us choose the point $(1:-1:0)$ to be the distinguished point. We now transform this into a Weierstrass model using the transformation

$$Y = \frac{36(x-y)}{x+y}, \qquad X = \frac{12z}{x+y}, \tag{15}$$

so that a solution to equation (14) corresponds to a rational point on the elliptic curve

$$E \,:\, Y^2 = X^3 - 432.$$

The solution $(1:-1:0)$ to (14) corresponds to the point $\infty = \mathscr{O}$ on E. The model E is the elliptic curve denoted by `27A` in Cremona's tables [8]. The group $E(\mathbb{Q})$ has rank zero and, in fact,

$$E(\mathbb{Q}) \simeq \mathbb{Z}/3\mathbb{Z}.$$

Indeed,

$$E(\mathbb{Q}) = \{\mathscr{O},\ (36,12),\ (36,-12)\}$$

where in the group law on E we have

$$2 \cdot (36,12) = (36,-12) \quad \text{and} \quad 3 \cdot (36,12) = \mathscr{O}\,.$$

Thus the degree 3 Fermat equation (14) has exactly three solutions, and we may obtain these by taking the three points belonging to $E(\mathbb{Q})$ and transferring them back to the model (14) using (15). Doing this, we find that the three solutions to (14) are $(1:-1:0)$, $(1:0:1)$ and $(0:1:0)$—that is, just the trivial solutions.

Example 2. One can similarly transform the equation

$$y^2 = x^4 + z^4 \tag{16}$$

into the elliptic curve

$$E \,:\, Y^2 = X^3 - 4X$$

which has Mordell–Weil group

$$E(\mathbb{Q}) = \{\mathscr{O},\ (0,0),\ (2,0),\ (-2,0)\} \simeq \mathbb{Z}/2\mathbb{Z} \times \mathbb{Z}/2\mathbb{Z}$$

and use this information to deduce that the only solutions to (16) are the trivial ones.

It turns out that the proofs by Fermat and Euler of the degree 4 and degree 3 cases of Fermat's Last Theorem are simply special cases of what are now standard Mordell–Weil group computations.

The degree p Fermat equation (12), viewed as defining a curve in $\mathbb{P}^2$, has genus $(p-1)(p-2)/2$, and thus does not define an elliptic curve for $p \geq 5$. We do mention in passing the following celebrated theorem of Faltings.

Theorem 5 (Faltings [13]). *Let C be a curve of genus ≥ 2 over a number field K. Then $C(K)$ is finite.*

Faltings' theorem tells us that for each $p \geq 5$ the Fermat equation (12) has finitely many primitive solutions. Faltings' theorem is *ineffective*, in the sense that the proof does not yield an algorithm that is guaranteed to find all solutions.

3.2 Modular forms

Let k and N be positive integers. We define

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \,:\, c \equiv 0 \pmod{N} \right\}.$$

It is easy to see that $\Gamma_0(N)$ is a subgroup of $\mathrm{SL}_2(\mathbb{Z})$ of finite index. Let $\mathbb{H}$ be the complex upper-half plane

$$\mathbb{H} = \{z \in \mathbb{C} \,:\, \mathrm{Im}(z) > 0\}.$$

The group $\Gamma_0(N)$ acts on $\mathbb{H}$ via fractional linear transformations

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \,:\, \mathbb{H} \to \mathbb{H}, \qquad z \mapsto \frac{az+b}{cz+d}.$$

The quotient $Y_0(N) = \Gamma_0(N)\backslash\mathbb{H}$ has the structure of a non-compact Riemann surface. This has a standard compactification denoted $X_0(N)$ and the difference $X_0(N) -$

$Y_0(N)$ is a finite set of points called the *cusps*. In fact the Riemann surface $X_0(N)$ has the structure of an algebraic curve defined over $\mathbb{Q}$ and is an example of what is known as a *modular curve*.

A *modular form f of weight k and level N* is a function $f : \mathbb{H} \to \mathbb{C}$ that satisfies the following conditions

(i) f is holomorphic on $\mathbb{H}$;
(ii) f satisfies the property

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^k f(z), \tag{17}$$

for all $z \in \mathbb{H}$ and $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma_0(N)$;
(iii) f extends to a function that is holomorphic at the cusps.

Observe that $\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right) \in \Gamma_0(N)$. Thus by (ii), the function f satisfies $f(z+1) = f(z)$. Letting $q(z) = \exp(2\pi i z)$, we see, from the periodicity, that f must have a Fourier expansion

$$f(z) = \sum_{n \geq N_0} c_n q^n$$

for some integer N_0. In fact, one of the cusps is the cusp at $i\infty$ which we can think of as being arbitrarily high up on the imaginary axis. Note that $q(i\infty) = 0$. We see that for f to be holomorphic at the cusp $i\infty$ we require $c_n = 0$ for $n < 0$. Thus we may write

$$f(z) = \sum_{n \geq 0} c_n q^n. \tag{18}$$

It turns out that set of modular forms of weight k and level N, denoted by $M_k(N)$, is a finite-dimensional vector space over $\mathbb{C}$.

A *cusp form* of weight k and level N is modular form f of weight k and level N that vanishes at all the cusps. As $q(i\infty) = 0$ we see in particular that a cusp form must satisfy $c_0 = 0$. The cusp forms naturally form a subspace of $M_k(N)$ which we denote by $S_k(N)$. Of particular interest are the weight 2 cusp forms of level N: these can be interpreted as regular differentials on the modular curves $X_0(N)$. It follows that the dimension of $S_2(N)$ as a $\mathbb{C}$-vector space is equal to the genus of the modular curve $X_0(N)$.

There is a natural family of commuting operators $T_n : S_2(N) \to S_2(N)$ (with $n \geq 1$) called the *Hecke operators*. The *eigenforms* of level N are the weight 2 cusp forms that are simultaneous eigenvectors for all the Hecke operators. Such an eigenform is called normalized if $c_1 = 1$ and thus its Fourier expansion has the form

$$f = q + \sum_{n \geq 1} c_n q^n.$$

3.3 Modularity

Let E be an elliptic curve over $\mathbb{Q}$. Such an elliptic curve has a model of the form

$$E \ : \ y^2 + a_1xy + a_3y = x^3 + a_2x^2 + a_4x + a_6, \tag{19}$$

where the $a_i \in \mathbb{Z}$, and a (non-zero) discriminant Δ_E which is an integer given by a complicated polynomial expression in terms of the a_i. It is possible to change the model by carrying out a suitable linear substitution in x, y, and and we generally work with a *minimal model*: that is one where the $a_i \in \mathbb{Z}$ with discriminant having the smallest possible absolute value. Associated to E is another, more subtle, invariant called the *conductor* N_E which we shall not define precisely, but we merely point that it is a positive integer sharing the same prime divisors as the discriminant; that it measures the 'bad behavior' of the elliptic curve E modulo primes; and that it can be computed easily through *Tate's algorithm* [31, Chapter IV].

Now let $p \nmid \Delta_E$ be a prime. Then we can reduce the equation (19) to obtain an elliptic curve $\tilde{E}$ over $\mathbb{F}_p$. The set $\tilde{E}(\mathbb{F}_p)$ is an abelian group as before, but now necessarily finite, and we denote its order by $\#\tilde{E}(\mathbb{F}_p)$. Let

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

We are now ready to state a version of the modularity theorem due to Wiles, Breuil, Conrad, Diamond and Taylor [37], [6]. This remarkable theorem was previously known as the Taniyama–Shimura conjecture.

Theorem 6 (The Modularity Theorem). *Let E be an elliptic curve over $\mathbb{Q}$ with conductor N. There exists a normalized eigenform $f = q + \sum c_n q^n$ of weight 2 and level N such that $c_n \in \mathbb{Z}$ for all n, and if $p \nmid \Delta_E$ is prime then $c_p = a_p(E)$.*

In fact, for an eigenform f the Fourier coefficients are determined by the coefficients c_p with prime indices. Thus from the elliptic curve E we can construct the Fourier expansion of the corresponding eigenform f. What is astonishing is that f then satisfies the transformation properties (17).

Example 3. We consider the following elliptic curve E over $\mathbb{Q}$:

$$E \ : \ y^2 + y = x^3 - x^2 - 10x - 20.$$

This has conductor 11, the smallest possible conductor, and discriminant -11^5. The space $S_2(11)$ is 1-dimensional. Naturally every non-zero element of $S_2(11)$ is an eigenform (i.e. an eigenvector for the Hecke operators), and we take as our basis the unique normalized eigenform which has the following Fourier expansion:

$$f(z) = q - 2q^2 - q^3 + 2q^4 + q^5 + 2q^6 - 2q^7 - \cdots$$

According to the modularity theorem the eigenform f corresponds to the elliptic curve E and we may check a few of the coefficients to convince ourselves that this is the case. For example,

$$\tilde{E}(\mathbb{F}_2) = \{\mathscr{O},\ (\overline{0},\overline{0}),\ (\overline{0},\overline{1}),\ (\overline{1},\overline{0}),\ (\overline{1},\overline{1})\}.$$

It follows that $a_2(E) = 2+1-\#E(\mathbb{F}_2) = -2$ agrees with the coefficient $c_2 = -2$ for q^2 in the Fourier expansion for f. The reader can easily verify the relation $a_p(E) = c_p$ for the primes $p = 3$, 5 and 7.

3.4 Galois representations

Let E be an elliptic curve over $\mathbb{C}$. The structure of the abelian group $E(\mathbb{C})$ is particularly easy to describe. There is a discrete lattice $\Lambda \subset \mathbb{C}$ of rank 2 (that is, as an abelian group $\Lambda \simeq \mathbb{Z}^2$) depending on E, and an isomorphism

$$E(\mathbb{C}) \simeq \mathbb{C}/\Lambda. \tag{20}$$

Let p be a prime. By the p-torsion of $E(\mathbb{C})$ we mean the subgroup

$$E[p] = \{Q \in E(\mathbb{C})\ :\ pQ = 0\}.$$

It follows from (20) that

$$E[p] \simeq (\mathbb{Z}/p\mathbb{Z})^2. \tag{21}$$

This can be viewed as 2-dimensional $\mathbb{F}_p$-vector space.

Example 4. Let

$$E\ :\ y^2 = x^3 + x. \tag{22}$$

It turns out that the corresponding lattice is $\Lambda = \mathbb{Z} + \mathbb{Z}i$. The p-torsion subgroup of $\mathbb{C}/\Lambda$ is

$$\left\{\frac{a+bi}{p} + \Lambda\ :\ a,b = 0,\ldots,p-1\right\}.$$

The reader will see that this a 2-dimensional $\mathbb{F}_p$-vector space with basis $1/p+\Lambda$ and $i/p+\Lambda$.

Now let E be an elliptic curve over $\mathbb{Q}$. Then we may view E as an elliptic curve over $\mathbb{C}$, and with the above definitions obtain an isomorphism $E[p] \simeq (\mathbb{Z}/p\mathbb{Z})^2$. However, in this setting the points of $E[p]$ have algebraic coordinates, and are acted on by $G_{\mathbb{Q}} := \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, *the absolute Galois group of the rational numbers*. Via the isomorphism (21), the group $G_{\mathbb{Q}}$ acts on $(\mathbb{Z}/p\mathbb{Z})^2$. As noted, the latter is a 2-dimensional $\mathbb{F}_p$-vector space. We obtain a 2-dimensional representation that depends on the elliptic curve E and the prime p:

$$\overline{\rho}_{E,p}\ :\ G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{F}_p). \tag{23}$$

Example 5. We continue looking at the elliptic curve (22) but now regard it as an elliptic curve over $\mathbb{Q}$. The 2-torsion subgroup is

$$E[2] = \{\mathscr{O},\ (0,0),\ (i,0),\ (-i,0)\}.$$

Recall that $\mathscr{O}$ is the identity element. The three other elements of $E[2]$ are points of order 2. Moreover, they satisfy the additional relation

$$(0,0) + (i,0) + (-i,0) = \mathscr{O}.$$

We can see this from the geometric description of the group law: the three points on the left-hand side are the intersection of the line $y = 0$ with E. As $2 \cdot (-i,0) = \mathscr{O}$ we have that $(-i,0) = -(i,0)$ and thus

$$(-i,0) = (0,0) + (i,0).$$

We now see that $E[2]$ is an $\mathbb{F}_2$-vector space with basis $(0,0)$ and $(i,0)$. Let us now use this to write down $\overline{\rho}_{E,2}$ explicitly. Let $\sigma \in G_{\mathbb{Q}}$. Then $\sigma(i) = i$ or $\sigma(i) = -i$. Suppose first that $\sigma(i) = i$. Then

$$\sigma(0,0) = (\sigma(0), \sigma(0)) = (0,0), \qquad \sigma(i,0) = (\sigma(i), \sigma(0)) = (i,0).$$

As σ leaves our chosen basis fixed, we have

$$\overline{\rho}_{E,2}(\sigma) = \begin{pmatrix} \overline{1} & \overline{0} \\ \overline{0} & \overline{1} \end{pmatrix} \in \mathrm{GL}_2(\mathbb{F}_2).$$

Suppose instead that $\sigma(i) = -i$. Then

$$\sigma(0,0) = (\sigma(0), \sigma(0)) = (0,0), \qquad \sigma(i,0) = (\sigma(i), \sigma(0)) = (-i,0) = (0,0) + (i,0).$$

Thus the action of σ with respect to our chosen basis is given by the matrix

$$\overline{\rho}_{E,2}(\sigma) = \begin{pmatrix} \overline{1} & \overline{1} \\ \overline{0} & \overline{1} \end{pmatrix} \in \mathrm{GL}_2(\mathbb{F}_2).$$

We record the image of the representation $\overline{\rho}_{E,2}$:

$$\overline{\rho}_{E,2}(G_{\mathbb{Q}}) = \left\{ \begin{pmatrix} \overline{1} & \overline{0} \\ \overline{0} & \overline{1} \end{pmatrix}, \begin{pmatrix} \overline{1} & \overline{1} \\ \overline{0} & \overline{1} \end{pmatrix} \right\}.$$

We note that the representation $\overline{\rho}_{E,2}$ is reducible, in the sense that all elements of the image share a common eigenvector $\begin{pmatrix} \overline{1} \\ \overline{0} \end{pmatrix}$.

We return to our general setting of an elliptic curve E over $\mathbb{Q}$ and a prime p. We say that the representation $\overline{\rho}_{E,p}$ is *reducible* if the matrices of the image $\overline{\rho}_{E,p}(G_{\mathbb{Q}})$ share some common eigenvector. Otherwise we say that $\overline{\rho}_{E,p}$ is irreducible.

We have now given enough definitions to be able to state Mazur's theorem; this is often considered as historically the first step in the proof of Fermat's Last Theorem.

Theorem 7 (Mazur [24]).

(i) Let E be an elliptic curve over $\mathbb{Q}$ and $p > 163$ be prime. Then $\overline{\rho}_{E,p}$ is irreducible.
(ii) Let E be an elliptic curve over $\mathbb{Q}$ with full 2-torsion (that is $E[2] \subseteq E(\mathbb{Q})$) and let $p \geq 5$ be prime. Then $\overline{\rho}_{E,p}$ is irreducible.

It turns out that an elliptic curve E over $\mathbb{Q}$ such that $\overline{\rho}_{E,p}$ is reducible corresponds to a rational point on the modular curve $X_0(p)$ that is not a cusp. Mazur proved his theorem by determining the rational points on this infinite family of curves. In a sense, Mazur's theorem is not unlike Fermat's Last Theorem, which is also a statement about the rational points on an infinite family of curves.

We mention in passing the relationship between reducible mod p representations and isogenies. An *isogeny* of elliptic curves E, E' is a non-constant map $\phi : E \to E'$ defined by algebraic equations that takes the point at infinity on E to the point at infinity on E'. A non-trivial consequence of the Riemann–Roch theorem is that isogenies respect the group law, and so are in a sense algebro-geometric homomorphisms. A p-isogeny is an isogeny $\phi : E \to E'$ such that the kernel of ϕ has order p. Let E be defined over $\mathbb{Q}$. Then the existence of an p-isogeny $\phi : E \to E'$ defined over $\mathbb{Q}$ is equivalent to the representation $\overline{\rho}_{E,p}$ being reducible. In fact, if Q is a non-zero element of the kernel of ϕ, then Q is a non-zero eigenvector for all the elements of the image of $\overline{\rho}_{E,p}$. We can restate (i) of Mazur's theorem as saying that an elliptic curve E defined over $\mathbb{Q}$ has no p-isogenies for $p > 163$.

3.5 Ribet's level lowering theorem

Let E be an elliptic curve over $\mathbb{Q}$. We saw above that, for each prime p, the curve E gives rise to a mod p Galois representation $\overline{\rho}_{E,p} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{F}_p)$. Let f be an eigenform. Deligne and Serre showed that such an f gives rise, for each prime p, to a Galois representation $\overline{\rho}_{f,p} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{F}_{p^r})$, where $r \geq 1$ depends on f. If E corresponds to f via the Modularity theorem (Theorem 6) then, unsurprisingly, $\overline{\rho}_{E,p} \sim \overline{\rho}_{f,p}$ (the two representations are isomorphic). Thus the representation $\overline{\rho}_{E,p}$ is **modular** in the sense that it arises from a modular eigenform. Recall that if f corresponds to E via modularity then the conductor of E is equal to the level of f. Sometimes it is possible to replace f by another eigenform of smaller level which has the same mod p representation. This process is called *level lowering*. We now state a special case of Ribet's level lowering theorem. Ribet's theorem is in fact part of Serre's modularity conjecture [29] that was proved by Khare and Wintenberger [20], [21].

Theorem 8 (Ribet's level lowering theorem [28]). *Let E be an elliptic curve over $\mathbb{Q}$ with minimal discriminant Δ and conductor N. Let $p \geq 3$ be prime. Suppose*

(i) the curve E is modular;
(ii) the mod p representation $\overline{\rho}_{E,p}$ is irreducible.

Let

$$N_p = N \Big/ \prod_{\substack{\ell \| N, \\ p \mid \mathrm{ord}_\ell(\Delta)}} \ell. \tag{24}$$

Then $\overline{\rho}_{E,p} \sim \overline{\rho}_{g,p}$ *for some eigenform* g *of weight* 2 *and level* N_p.

Of course we now know, thanks to the Modularity theorem (Theorem 6) that all elliptic curves over $\mathbb{Q}$ are modular. Thus condition (i) in Ribet's theorem is automatically satisfied. But we still include it for historical interest.

We can make the relationship $\overline{\rho}_{E,p} \sim \overline{\rho}_{g,p}$ in Ribet's theorem more explicit. Write $g = q + \sum_{n \geq 1} d_n q^n$ for the Fourier expansion of g. It turns out that the d_n belong to the ring of integers $\mathfrak{O}_K$ of a number field K that depends on g. The relationship $\overline{\rho}_{E,p} \sim \overline{\rho}_{g,p}$ is equivalent to the existence of a prime ideal $\mathfrak{P}$ of $\mathfrak{O}_K$ that divides $p\mathfrak{O}_K$ such that $a_q(E) \equiv d_q \pmod{\mathfrak{P}}$ for all primes $q \nmid Np$.

Example 6. Consider the elliptic curve

$$E \; : \quad y^2 = x^3 - x^2 - 77x + 330$$

with Cremona reference `132B1`. Cremona's database [8] gives us the minimal discriminant and conductor

$$\Delta = 2^4 \times 3^{10} \times 11, \qquad N = 2^2 \times 3 \times 11. \tag{25}$$

The database also tells us that the only isogeny the curve E has is a 2-isogeny. Thus $\overline{\rho}_{E,p}$ is irreducible for $p \geq 3$. We apply Ribet's Theorem with $p = 5$. From the above recipe (24) for the level we find that $N_p = 44$. It is possible to check that $\overline{\rho}_{E,p} \sim \overline{\rho}_{g,p}$ where g is the following eigenform has weight 2 and level 44:

$$g = q + q^3 - 3q^5 + 2q^7 - 2q^9 - q^{11} + \cdots.$$

All the coefficients of g belong to $\mathbb{Z}$. We tabulate $a_q(E)$ and the coefficients d_q for primes $q < 50$. The reader will note that the relationship $a_q(E) \equiv d_q \pmod{5}$ holds for all primes q in the range except for $q = 3$ which does divide N.

q	2	3	5	7	11	13	17	19	23	29	31	37	41	43	47
$a_q(E)$	0	−1	2	2	−1	6	−4	−2	−8	0	0	−6	0	10	0
d_q	0	1	−3	2	−1	−4	6	8	−3	0	5	−1	0	−10	0

3.6 The proof of Fermat's Last Theorem

We are now able to sketch a proof of Fermat's Last Theorem. Suppose $p \geq 5$ is prime, and x, y and z are non-zero pairwise coprime integers such that $x^p + y^p + z^p =$

0. We may reorder (x,y,z) so that y is even and $x^p \equiv -1 \pmod 4$. We let E be the following elliptic curve which depends on the solution (x,y,z):

$$E : Y^2 = X \cdot (X - x^p) \cdot (X + y^p). \tag{26}$$

The curve E is called the *Frey–Hellegouarch curve*. The minimal discriminant and conductor of E are:

$$\Delta = \frac{x^{2p} y^{2p} z^{2p}}{2^8}, \qquad N = \prod_{\ell \mid \Delta} \ell.$$

The choice $2 \mid y$ and $x^p \equiv -1 \pmod 4$ ensures that $2 \,||\, N$.

We now consider $\overline{\rho}_{E,p}$. The 2-torsion subgroup for E is

$$E[2] = \{\mathscr{O},\ (0,0),\ (x^p,0),\ (-y^p,0)\}.$$

Note that $E[2] \subseteq E(\mathbb{Q})$. As $p \geq 5$ we know by Mazur's theorem (Theorem 7) that $\overline{\rho}_{E,p}$ is irreducible. Moreover E is modular by the Modularity theorem. The hypotheses of Ribet's theorem are satisfied. We compute $N_p = 2$ using the recipe in (24). It follows that $\overline{\rho}_{E,p} \sim \overline{\rho}_{g,p}$ where g has weight 2 and level 2. A simple computation shows that there are no eigenforms of weight 2 and level 2. This contradiction completes the proof of Fermat's Last Theorem.

Some Historical Remarks. In the early 1970s, Hellegouarch (e.g. [19]) had the idea of associating to a non-trivial solution of the Fermat equation the elliptic curve (26); he noted that the number field generated by its p-torsion subgroup $E[p]$ has surprisingly little ramification. In the early 1980s, Frey [18] observed that this elliptic curve enjoys certain remarkable properties that should rule out its modularity. Motivated by this, in 1985 Serre [29] made precise his modularity conjecture and showed that it implies Fermat's Last Theorem. Serre's remarkable paper also uses several variants of the Frey–Hellegouarch curve to link modularity to other Diophantine problems. Ribet announced his level-lowering theorem 1987, thereby proving that modularity of the Frey–Hellegouarch curve implies Fermat's Last Theorem. The proof of the Modularity theorem was completed around 1999 by Breuil, Conrad, Diamond and Taylor [6]. A *semistable* elliptic curve is one with squarefree conductor. We note from (25) that the Frey–Hellegouarch curve is semistable. In 1994 Wiles [37], with some help from Taylor [33], proved modularity of semistable elliptic curves over $\mathbb{Q}$, thereby proving Fermat's Last Theorem.

4 The (more) generalized Fermat equation

We now return to the *generalized Fermat equation*

$$x^p + y^q = z^r, \tag{27}$$

where x, y and z are integers, and the exponents p, q and r are (potentially distinct) positive integers. We restrict our attention to *primitive* solutions, i.e. those with $\gcd(x,y,z) = 1$, since, without such a restriction, it is easy to concoct uninteresting solutions in a fairly trivial fashion. Indeed, if we assume, say, that p, q and r are pairwise relatively prime, then we can choose integers u, v and w such that

$$uqr \equiv -1 \pmod{p}, \;\; vpr \equiv -1 \pmod{q} \;\text{ and }\; wpq \equiv -1 \pmod{r}.$$

If we are given any integers a, b and c with $a + b = c$, multiplying this equation by $a^{uqr} b^{vpr} c^{wpq}$, we thus have that

$$\left(a^{(uqr+1)/p} b^{vr} c^{wq}\right)^p + \left(a^{ur} b^{(vpr+1)/q} c^{wp}\right)^q = \left(a^{uq} b^{vp} c^{(wpq+1)/r}\right)^r.$$

We call (p,q,r) the *signature* of equation (27). The behaviour of primitive solutions depends fundamentally upon the size of the quantity

$$\sigma(p,q,r) = \frac{1}{p} + \frac{1}{q} + \frac{1}{r},$$

in particular, whether $\sigma(p,q,r) > 1$, $\sigma(p,q,r) = 1$ or $\sigma(p,q,r) < 1$. If we set $\chi = \sigma(p,q,r) - 1$, then χ is the Euler characteristic of a certain stack associated to equation (27). It is for this reason that the cases $\sigma(p,q,r) > 1$, $\sigma(p,q,r) = 1$ and $\sigma(p,q,r) < 1$ are respectively termed *spherical*, *parabolic* and *hyperbolic*.

4.1 The spherical case $\sigma(p,q,r) > 1$

In this case, we may assume that (p,q,r) is one of $(2,2,r)$, $(2,q,2)$, $(2,3,3)$, $(2,3,4)$, $(2,4,3)$ or $(2,3,5)$. In each of these situations, the (infinitely many) relatively prime integer solutions to (27) come in finitely many two parameter families (the canonical model to bear in mind here is that of Pythagorean triples); in the (most complicated) $(2,3,5)$ case, there are precisely 27 such families, as proved by Johnny Edwards [11] in 2004 via an elegant application of classical invariant theory. In the case $(p,q,r) = (2,4,3)$, by way of example, we find that the relatively prime solutions x, y and z satisfy one of the following four parametrizations :

$$\begin{cases} x = 4ts(s^2 - 3t^2)(s^4 + 6t^2s^2 + 81t^4)(3s^4 + 2t^2s^2 + 3t^4), \\ y = \pm(s^2 + 3t^2)(s^4 - 18t^2s^2 + 9t^4), \\ z = (s^4 - 2t^2s^2 + 9t^4)(s^4 + 30t^2s^2 + 9t^4), \end{cases}$$

where $s \not\equiv t \pmod 2$ and $3 \nmid s$,

$$\begin{cases} x = \pm(4s^4 + 3t^4)(16s^8 - 408t^4s^4 + 9t^8), \\ y = 6ts(4s^4 - 3t^4), \\ z = 16s^8 + 168t^4s^4 + 9t^8, \end{cases}$$

where t is odd and $3 \nmid s$,

$$\begin{cases} x = \pm(s^4 + 12t^4)(s^8 - 408t^4s^4 + 144t^8), \\ y = 6ts(s^4 - 12t^4), \\ z = s^8 + 168t^4s^4 + 144t^8, \end{cases}$$

where $s \equiv \pm 1 \pmod 6$, or

$$\begin{cases} x = 2(s^4 + 2ts^3 + 6t^2s^2 + 2t^3s + t^4)(23s^8 - 16ts^7 - 172t^2s^6 - 112t^3s^5 \\ \qquad -22t^4s^4 - 112t^5s^3 - 172t^6s^2 - 16t^7s + 23t^8), \\ y = 3(s-t)(s+t)(s^4 + 8ts^3 + 6t^2s^2 + 8t^3s + t^4), \\ z = 13s^8 + 16ts^7 + 28t^2s^6 + 112t^3s^5 + 238t^4s^4 \\ \qquad +112t^5s^3 + 28t^6s^2 + 16t^7s + 13t^8, \end{cases}$$

where $s \not\equiv t \pmod 2$ and $s \not\equiv t \pmod 3$. Here, s and t are relatively prime integers. Details on these parametrizations (and much more besides) can be found in Cohen's exhaustive work [7].

4.2 The parabolic case $\sigma(p,q,r) = 1$

If we have $\sigma(p,q,r) = 1$, then, up to reordering,

$$(p,q,r) = (2,6,3), (2,4,4), (4,4,2), (3,3,3) \text{ or } (2,3,6).$$

As in Examples 1 and 2, each equation now corresponds to an elliptic curve of rank 0 over $\mathbb{Q}$; the only primitive non-trivial solution comes from the signature $(p,q,r) = (2,3,6)$, corresponding to the Catalan solution $3^2 - 2^3 = 1$.

4.3 The hyperbolic case $\sigma(p,q,r) < 1$

It is the *hyperbolic case*, with $\sigma(p,q,r) < 1$, where most of our interest lies. Here, we are now once again considering the equation and hypotheses (4). As mentioned previously, it is expected that the only solutions to (4) are with (x,y,z,p,q,r) corresponding to the identity $1^p + 2^3 = 3^2$, for $p \geq 6$, or to

$$2^5 + 7^2 = 3^4, \quad 7^3 + 13^2 = 2^9, \quad 2^7 + 17^3 = 71^2, \quad 3^5 + 11^4 = 122^2,$$
$$17^7 + 76271^3 = 21063928^2, \; 1414^3 + 2213459^2 = 65^7, \; 9262^3 + 15312283^2 = 113^7,$$
$$43^8 + 96222^3 = 30042907^2 \quad \text{and} \quad 33^8 + 1549034^2 = 15613^3.$$

A less ambitious conjecture would be that (4) has at most finitely many solutions (where we agree to count those coming from $1^p + 2^3 = 3^2$ only once). In the rest of this section, we will discuss our current knowledge about this equation.

4.4 The Theorem of Darmon and Granville

What we know for sure in the hyperbolic case, is that, for a fixed signature (p,q,r), the number of solutions to equation (4) is at most finite :

Theorem 9 (Darmon and Granville [9]). *If A,B,C,p,q and r are fixed positive integers with*

$$\frac{1}{p}+\frac{1}{q}+\frac{1}{r}<1,$$

then the equation

$$Ax^p+By^q=Cz^r$$

has at most finitely many solutions in coprime non-zero integers x,y and z.

Proof. The proof by Darmon and Granville is extremely elegant and we cannot resist giving a brief sketch. The hypothesis $1/p+1/q+1/r<1$ is used to show the existence of a cover $\phi : D\to\mathbb{P}^1$ that is ramified only above 0, 1, ∞, where the curve D has genus ≥ 2. Moreover, this cover has the property that the ramification degrees above 0 are all divisors of p, above 1 are all divisors of q, and above ∞ are all divisors of r. Now let (x,y,z) be a non-trivial primitive solution to the equation $Ax^p+By^p=Cz^r$. The above properties of the cover ϕ imply that the points belonging to the fiber $\phi^{-1}(Ax^p/Cz^r)$ are defined over a number field K that is unramified away from the primes dividing $2ABCpqr$. It follows from a classical theorem of Hermite that there are only finitely many such number fields K. Moreover, by Faltings' theorem, for each possible K there are only finitely many K-points on D. It follows that the equation $Ax^p+By^p=Cz^r$ has only finitely many primitive solutions.

It is worth noting that the argument used in the proof is ineffective, due to its dependence upon Faltings' theorem; it is not currently known whether or not there exists an algorithm for finding all rational points on an arbitrary curve of genus ≥ 2.

4.5 A brief survey of what we know

What we would really like to do goes rather further than what the theorem of Darmon and Granville tells us. Indeed, we would like to obtain finiteness results for equation (4) where we allow the exponent triples (p,q,r) to range over infinite families. In the following tables, we list all known (as of 2015) instances where equation (4) has been completely solved. For references to the original papers we recommend the exhaustive survey [4]. The first table collects all known infinite families treated to date :

(p,q,r)	reference(s)
(n,n,n)	Wiles, Taylor-Wiles
$(n,n,k), k \in \{2,3\}$	Darmon-Merel, Poonen
$(2n,2n,5)$	Bennett
$(2,4,n)$	Ellenberg, Bennett-Ellenberg-Ng, Bruin
$(2,6,n)$	Bennett-Chen, Bruin
$(2,n,4)$	Bennett-Skinner, Bruin
$(2,n,6)$	Bennett-Chen-Dahmen-Yazdani
$(3j,3k,n),\ j,k \geq 2$	immediate from Kraus
$(3,3,2n)$	Bennett-Chen-Dahmen-Yazdani
$(3,6,n)$	Bennett-Chen-Dahmen-Yazdani
$(2,2n,k),\ k \in \{9,10,15\}$	Bennett-Chen-Dahmen-Yazdani
$(4,2n,3)$	Bennett-Chen-Dahmen-Yazdani
$(2j,2k,n),\ j,k \geq 5$ prime, $n \in \{3,5,7,11,13\}$	Anni-Siksek

Our second table lists "sporadic" triples where the solutions to (4) have been determined, and infinite families of exponent triples where the (p,q,r) satisfy certain additional local conditions.

(p,q,r)	reference(s)
$(3,3,n)^*$	Chen-Siksek, Kraus, Bruin, Dahmen
$(2,2n,3)^*$	Chen, Dahmen, Siksek
$(2,2n,5)^*$	Chen
$(2m,2n,3)^*$	Bennett-Chen-Dahmen-Yazdani
$(2,4n,3)^*$	Bennett-Chen-Dahmen-Yazdani
$(3,3n,2)^*$	Bennett-Chen-Dahmen-Yazdani
$(2,3,n),\ n \in \{6,7,8,9,10,15\}$	Poonen-Schaefer-Stoll, Bruin, Zureick-Brown, Siksek, Siksek-Stoll
$(3,4,5)$	Siksek-Stoll
$(5,5,7),\ (7,7,5)$	Dahmen-Siksek

The asterisk here refers to conditional results. For instance, in case $(p,q,r) = (3,3,n)$, we have no solutions if either $3 \leq n \leq 10^9$, or $n \equiv \pm 2$ modulo 5, or $n \equiv \pm 17$ modulo 78, or

$$n \equiv 51, 103, 105 \text{ modulo } 106,$$

or for n (modulo 1296) one of

$$43,49,61,79,97,151,157,169,187,205,259,265,277,295,313,367,373,385,403,$$
$$421,475,481,493,511,529,583,601,619,637,691,697,709,727,745,799,805,$$
$$817,835,853,907,913,925,943,961,1015,1021,1033,1051,1069,1123,$$
$$1129,1141,1159,1177,1231,1237,1249,1267,1285.$$

The results mentioned here have been proved by essentially two distinct methods. For a number of fixed triples, the problem has been reduced (via arguments similar to those of Darmon and Granville, or otherwise) to one of determining $\mathbb{Q}$-rational points on certain curves of genus 2 or higher. These points were subsequently found via Chabauty-type methods and appeal to a version of the Mordell-Weil sieve. In

each case where equation (4) has been solved for an infinite family of triples (p,q,r), however, a different approach has been utilized, relying upon Frey–Hellegouarch curves and connections between them and modular forms.

5 The modular approach and the generalized Fermat equation

It is natural to ask if the proof of Fermat's Last Theorem can be adapted to resolve (4), at least for certain signatures (p,q,r). Roughly speaking a *Frey–Hellegouarch curve* is an elliptic curve E over $\mathbb{Q}$, attached to a solution of a Diophantine equation satisfying two conditions:

(i) the discriminant of E has the form $A \cdot B^p$ where A is a known (small) integer and p is a prime;
(ii) every prime $q \mid B$ divides the conductor exactly once.

Examining the recipe (24) in Ribet's theorem the reader will note that the level N_p depends only on the known quantity A. For example, in the proof of Fermat's Last Theorem, A is a power of 2 and the level $N_p = 2$.

Alas, only a few signatures have workable Frey–Hellegouarch curves. In the following table we record some of the known ones.

Equation	Frey–Hellegouarch Curve
$a^p + b^p = c^2$	$Y^2 = X^3 + 2cX^2 + a^pX$
$a^p + b^p = c^3$	$Y^2 = X^3 + 3cX^2 - 4b^p$
$a^3 + b^3 = c^p$	$Y^2 = X^3 + 3(a-b)X^2 + 3(a^2 - ab + b^2)X$
$a^2 + b^3 = c^p$	$Y^2 = X^3 + 3bX + 2a$

These and similar Frey–Hellegouarch curves have been used to prove many of the results surveyed in Section 4.5.

5.1 A sample signature : $(p,p,2)$

To illustrate the approach we look specifically at the equation $x^p + y^p = z^2$ where $p \geq 7$ is prime. Here we follow the paper of Darmon and Merel [10] who showed that the only primitive solutions are $(\pm 1, \mp 1, 0)$, $(1,0,\pm 1)$, $(0,1,\pm 1)$. Let $(x,y,z) = (a,b,c)$ be a primitive solution satisfying $ab \neq 0$. As in the preceding table, we associate to this the Frey–Hellegouarch curve

$$E \; : \; Y^2 = X^3 + 2cX^2 + a^pX.$$

This is modular by Theorem 6. By a variant of Mazur's theorem (Theorem 7) the mod p representation $\overline{\rho}_{E,p}$ is irreducible. Now an application of the Ribet's theorem shows that $\overline{\rho}_{E,p} \sim \overline{\rho}_{g,p}$ where g is an eigenform of weight 2 and level 32. This is

where we diverge from the proof of Fermat's Last Theorem: there is an eigenform of weight 2 and level 32. It turns however that this eigenform is rather special as it corresponds to an elliptic curve with complex multiplication. It follows from this fact that $\overline{\rho}_{g,p} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{F}_p)$ is not surjective. To complete the resolution of the equation $x^p + y^p = z^2$, Darmon and Merel needed to show that if $ab \neq -1$ then $\overline{\rho}_{E,p}$ is in fact surjective and hence cannot be isomorphic to $\overline{\rho}_{g,p}$. To do this, they showed that if $\overline{\rho}_{E,p}$ is not surjective then it gives rise to a rational point on one of a family of certain modular curves, and completed the proof by determining the rational points on this family. This last step is somewhat similar to the proof of Mazur's theorem.

6 Modularity over number fields

Even when we are interested in Diophantine equations in rational integer unknowns, factorization arguments often force us to consider Diophantine equations where the coefficients or unknowns lie in a number field. Consider for example the equation

$$a^4 + b^2 = c^p, \qquad \gcd(a,b,c) = 1 \tag{28}$$

where the exponent p is prime. This equation is not known to have a Frey–Hellegouarch curve defined over $\mathbb{Q}$. We can, however, factor the left-hand side as $(a^2+bi)(a^2-bi)$ where $i = \sqrt{-1}$. It is not hard to show using the arithmetic of the Gaussian ring $\mathbb{Z}[i]$ that

$$a^2 + bi = \alpha^p, \qquad a^2 - bi = \overline{\alpha}^p.$$

where $\alpha \in \mathbb{Z}[i]$ and $\overline{\alpha}$ is its conjugate. Eliminating b we obtain the equation

$$\alpha^p + \overline{\alpha}^p = 2a^2.$$

This is an $(p,p,2)$ equation. However, unlike the equations we met in Section 5.1, some of the unknowns belong to $\mathbb{Z}[i]$. We ignore this uncomfortable fact for now and simply imitate the approach in the previous section to associate a Frey–Hellegouarch curve to this equation. The Frey–Hellegouarch curve is

$$E : y^2 = x^3 + 2ax^2 + 2\alpha^p x \tag{29}$$

which has discriminant $2^9(\alpha\overline{\alpha}^2)^p$. We note that the discriminant is close to being a perfect p-th power. To solve our original generalized Fermat equation (28) with signature $(4,2,p)$ and unknowns belonging to $\mathbb{Z}$, we need to consider an elliptic curve that is defined over $\mathbb{Q}(i)$. *It is natural to ask how much of modularity and level lowering carry over to the setting of number fields.* If we ask these questions for elliptic curves over general number fields then the answers are conjectural with almost no satisfactory theorems. However there are two situations where there are

satisfactory theorems and these have been applied to certain generalized Fermat equations: $\mathbb{Q}$-curves and elliptic curves over totally real fields.

6.1 $\mathbb{Q}$-curves

A $\mathbb{Q}$-*curve* is an elliptic curve E over a number field K that is isogenous to all its conjugates. The Frey curve (29) is an example of a $\mathbb{Q}$-curve: it is defined over the number field $\mathbb{Q}(i)$ and it happens to be isogenous to its conjugate $y^2 = x^3 + 2ax^2 + 2\overline{\alpha}^p x$ (the conjugate is obtained simply by conjugating the coefficients of the elliptic curve).

A consequence of the proof of Khare and Wintenberger of Serre's modularity conjecture is that $\overline{\mathbb{Q}}$-curves are modular. The modularity of the $\mathbb{Q}$-curve E given by (29) was used by Ellenberg [12] and by the first author, Ellenberg and Ng [5] to completely solve $a^4 + b^2 = c^p$ showing in fact that there no non-trivial primitive solutions with $n \geq 4$ (here n does not have to be prime). The first author and Chen [3] have used modularity of Frey–Hellegouarch $\mathbb{Q}$-curves to show that the equation $a^2 + b^6 = c^n$ has no non-trivial solutions with $\gcd(a,b,c) = 1$ and $n \geq 3$.

6.2 Elliptic curves over totally real fields

A number field K of degree n has n embeddings into the complex numbers $\iota_j : K \hookrightarrow \mathbb{C}$ with $j = 1, \ldots, n$. For example if $K = \mathbb{Q}(\theta)$ is a number field of degree n then θ is the root of an irreducible degree n polynomial with rational coefficients. Such a polynomial has n distinct complex roots $\theta_1, \ldots, \theta_n$, and the embedding ι_j satisfies $\iota_j(\theta) = \theta_j$. The embedding ι_j is real if $\iota_j(K) \subset \mathbb{R}$. Equivalently if $\theta_j \in \mathbb{R}$. If all the embeddings are real then we say that K is a totally real (number) field. An example of a totally real field is the cubic field $K = \mathbb{Q}(\theta)$ where $\theta = \zeta_7 + \zeta_7^{-1}$. Here θ is a root of the polynomial $x^3 + x^2 - 2x - 1$. The roots θ_j of the polynomial are $2\cos(2\pi j/7)$ with $j = 1, 2, 3$ which are all real.

Elliptic curves over totally real fields are conjecturally expected to be modular in the sense that they correspond to what are known as Hilbert modular forms (the classical modular forms of Section 3.2 are a special case of Hilbert modular forms). There has been substantial progress towards proving modularity for elliptic curves over totally real fields thanks to the work of Barnet-Lamb, Breuil, Diamond, Gee, Geraghty, Kisin, Skinner, Wiles and many others (many of these results in fact integrate level lowering with modularity). Building on this work, modularity of elliptic curves over real quadratic fields was recently proved by Freitas, Le Hung and Siksek [14]. To solve Diophantine problems via Frey curves that are defined over totally real fields one needs not only modularity and level lowering, but also irreducibility theorems for mod p representations of elliptic curves. Over the rationals, as we saw in Section 3.4, this is provided by Mazur's theorem. No such theorem is known

over any number field other than $\mathbb{Q}$. However Frey curves are almost semistable and this fact can usually be used [17] together with the celebrated uniform boundedness theorem of Merel [25] to supply the required irreducibility result.

As an example we mention the equation

$$a^{2\ell} + b^{2m} = c^p, \qquad \gcd(a,b,c) = 1. \tag{30}$$

studied by Anni and Siksek [1]. Here $\ell, m \geq 5$ and $p \geq 3$ are primes. A complicated factorization argument is used to reduce this to a Fermat equation with signature (ℓ,ℓ,ℓ) with coefficients and unknowns belonging to the totally real field $\mathbb{Q}(\zeta_p + \zeta_p^{-1})$. The corresponding Frey curves over this field are shown to be modular using the above-mentioned works for $p = 3$, 5, 7, 11 and 13. This is then used to show that the only solutions to (30) are the trivial ones.

7 A way forward: Darmon's program

The Frey–Hellegouarch approach used in the proof of Fermat's Last Theorem and in the resolution of many other equations (as sketched in previous sections) attaches an elliptic curve to a hypothetical solution of the equation in question and then uses modularity to make deductions about this solution. It is natural to ask:

(i) Are there geometric objects other than elliptic curves that are somehow modular?
(ii) If so, can these be used as an alternative, perhaps to add flexibility and tackle generalized Fermat equations for which no Frey–Hellegouarch elliptic curve is known?

An **abelian variety** is a connected and projective algebraic group. Roughly speaking this means that it is defined by algebraic equations in projective space and carries a group structure (that happens to be abelian). An abelian variety has a dimension $d \geq 1$, and abelian varieties of dimension 1 are simply elliptic curves. Are abelian varieties over $\mathbb{Q}$ modular? The answer should be 'yes', except that the precise meaning of word modular in this generality is not yet resolved.

An abelian variety of dimension d is said to be of GL_2-type if its endomorphism ring is an order in a number field of degree d. A Theorem of Khare and Wintenberger [20] states that abelian varieties of GL_2-type over $\mathbb{Q}$ are modular in a very precise sense (that is in fact very close to that of elliptic curves in Section 3.3). Abelian varieties of GL_2-type over totally real fields are expected to be modular in the sense that they correspond to Hilbert modular forms. Darmon exploits this idea to attack the equation $x^p + y^p = z^r$, where p and r are primes and $\gcd(x,y,z) = 1$ as usual. He attaches a hypothetical non-trivial solution to an abelian variety of GL_2-type over the totally real field $\mathbb{Q}(\zeta_r + \zeta_r^{-1})$. Using this he proves several beautiful theorems about possible solutions, though all are dependent on yet unproven conjectures. The biggest obstruction to Darmon's program is the absence of a Mazur-style irreducibility theorem for mod p representations of abelian varieties of GL_2-type.

The Darmon program holds the greatest promise for further substantial progress on the generalized Fermat equation. Just as Frey's original work was the spark that led to the formulation of Serre's modularity conjecture, and the proofs of Ribet's theorem and the Modularity theorem, so we hope that Darmon's program will supply the impetus for new theorems for abelian varieties of GL_2-type that in turn allow us to make deductions about the generalized Fermat equation.

References

1. S. Anni and S. Siksek, *On the generalized Fermat equation $x^{2\ell}+y^{2m}=z^p$ for $3 \leq p \leq 13$*, preprint, arXiv 1506.02860.
2. K. Belabas, F. Beukers, P. Gaudry, H. Lenstra, W. McCallum, B. Poonen, S. Siksek, M. Stoll, M. Watkins, *Explicit Methods in Number Theory: Rational Points and Diophantine Equations*, Panoramas et synthèses **36**, Société Mathématique de France, Paris, 2012.
3. M. A. Bennett and I. Chen, *Multi-Frey $\mathbb{Q}$-curves and the Diophantine equation $a^2+b^6=c^n$*, Algebra Number Theory **6** (2012), 707–730.
4. M. A. Bennett, I. Chen, S. R. Dahmen and S. Yazdani, *Generalized Fermat equations: a miscellany*, Int. J. Number Theory **11** (2015), 1–28.
5. M. A. Bennett, J. S. Ellenberg and N. Ng, *The Diophantine equation $A^4+2^\delta B^2=C^n$*, Int. J. Number Theory **6** (2010), 311-338.
6. C. Breuil, B. Conrad, F. Diamond and R. Taylor, *On the modularity of elliptic curves over $\mathbb{Q}$: wild 3-adic exercises*, J. Amer. Math. Soc. **14** (2001), 843–939.
7. H. Cohen, *Number theory. Vol. II. Analytic and modern tools*, Graduate Texts in Mathematics, 240, Springer, New York, 2007.
8. J. Cremona, Elliptic Curve Data, September 2015.
9. H. Darmon and A. Granville, *On the equations $z^m=F(x,y)$ and $Ax^p+By^q=Cz^r$*, Bull. London Math. Soc. textbf27 (1995), 513–543.
10. H. Darmon and L. Merel, *Winding quotients and some variants of Fermat's last theorem*, J. Reine Angew. Math. **490** (1997), 81–100.
11. J. Edwards, *A complete solution to $X^2+Y^3+Z^5=0$*, J. Reine Angew. Math.**571** (2004), 213–236.
12. J. S. Ellenberg, *Galois representations attached to $\mathbb{Q}$-curves and the generalized Fermat equation $A^4+B^2=C^p$*, Amer. J. Math. **126** (2004), 763–787.
13. G. Faltings, *Endlichkeitssätze für abelsche Varietäten über Zahlkörpen*, Invent. Math. **73** (1983), 349–366.
14. N. Freitas, B. Le Hung and S. Siksek, *Elliptic curves over real quadratic fields are modular*, Invent. Math. **201** (2015), 159–206.
15. N. Freitas and S. Siksek, *The asymptotic Fermat's Last Theorem for five-sixths of real quadratic fields*, Compos. Math. **151** (2015), 1395–1415.
16. N. Freitas and S. Siksek, *Fermat's last theorem over some small real quadratic fields*, Algebra Number Theory **9** (2015), 875–895.
17. N. Freitas and S. Siksek, *Criteria for irreducibility of mod p representations of Frey curves*, J. Théor. Nombres Bordeaux **27**, 2015, 67–76.
18. G. Frey, *Links between stable elliptic curves and certain Diophantine equations*, Ann. Univ. Sarav. Ser. Math. **1** (1986), iv+40.
19. Y. Hellegouarch, *Points d'ordre $2p^h$ sur les courbes elliptiques*, Acta Arith. **26** (1974/75), 253–263.
20. C. Khare and J.-P. Wintenberger, *Serre's modularity conjecture. I*, Invent. Math. **178** (2009), 485–504.
21. C. Khare and J.-P. Wintenberger, *Serre's modularity conjecture. II*, Invent. Math. **178** (2009), 505–586.

22. A. Kraus, *Majorations effectives pour l'équation de Fermat généralisée*, Canad. J. Math. **49** (1997), 1139–1161.
23. S. Lang, *Cyclotomic fields I and II*, Graduate Texts in Mathematics, 121, Springer, New York, 1990.
24. B. Mazur, *Rational isogenies of prime degree*, Invent. Math. **44** (1978), 129–162.
25. L. Merel, *Bornes pour la torsion des courbes elliptiques sur les corps de nombres*, Invent. Math. **124** (1996), 437–449.
26. M. Th. Rassias, *Problem-solving and selected topics in number theory. In the spirit of the mathematical olympiads. With a foreword by Preda Mihilescu.*, Springer, NewYork, 2011.
27. P. Ribenboim, 13 *Lectures on Fermat's Last Theorem*, Springer, 1979.
28. K. Ribet, *On modular representations of* $Gal(\overline{\mathbb{Q}}/\mathbb{Q})$ *arising from modular forms*, Invent. Math. **100** (1990), 431–476.
29. J.-P. Serre, *Sur les représentations modulaires de degré* 2 *de* $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$, Duke Math. J. **54** (1987), 179–230.
30. S. Siksek, *The modular approach to Diophantine equations*, pages 151–179 of [2].
31. J. H. Silverman, *Advanced topics in the arithmetic of elliptic curves*, Graduate Texts in Mathematics, 151, Springer-Verlag, New York, 1994.
32. S. Sitaraman, *Vandiver revisited*, J. Number Theory **57** (1996), 122–129.
33. R. Taylor and A. Wiles, *Ring-theoretic properties of certain Hecke algebras*, Ann. of Math. **141** (1995), 553–572.
34. G. Terjanian, *Sur l'équation* $x^{2p} + y^{2p} = z^{2p}$, Comptes Rendus Académie de Sciences Paris, **285** (1977), 973–975.
35. F. Thaine, *On the ideal class groups of real abelian number fields.*, Ann. of Math. **128** (1988), 1–18.
36. L. Washington, *Introduction to Cyclotomic Fields*, Graduate Texts in Mathematics, Springer, New York, 1996.
37. A. Wiles, *Modular elliptic curves and Fermat's Last Theorem*, Ann. Math. **141** (1995), 443–551.