Faltings plus epsilon, Wiles plus epsilon, and the Generalized Fermat Equation*

H. Darmon

September 9, 2007

Wiles' proof of Fermat's Last Theorem puts to rest one of the most famous unsolved problems in mathematics, a question that has been a wellspring for much of modern algebraic number theory. While celebrating Wiles' achievement, one also feels a twinge of regret at Fermat's demise. Is the Holy Grail of number theorists to become a mere footnote in the history books?

Hoping to keep some of the spirit of Fermat alive, I would like to discuss the *generalized Fermat equation*

$$x^p + y^q = z^r, \tag{1}$$

where p, q and r are fixed exponents. As in the case of Fermat's Last Theorem, one is interested in *integer solutions* (x, y, z), which are *non-trivial* in the sense that $xyz \neq 0$.

One might expect the equation above to have no such solutions if the exponents p, q, and r are large enough. But observe that, if $p = q$ is odd, and $r = 2$, then any solution to $a^p + b^p = c$ (of which there is an abundant supply!) yields the solution $(ac, bc, c^{\frac{p+1}{2}})$ to the equation $x^p + y^p = z^2$. A similar construction works whenever the exponents p, q, and r are pairwise coprime. However, the solutions produced in this way are not very interesting: the integers x, y and z have a large common factor.

*This is a transcription of the author's Aisenstadt prize lecture given at the CRM in March 1997. It is a pleasure to thank Andrew Granville and Loïc Merel for stimulating collaborations related to the topics of this essay, as well as Dan Abramovich for many helpful conversations over the years. This research was supported by CICMA and by grants from the Sloan Foundation, NSERC and FCAR.

Accordingly, one calls a solution (x, y, z) to the generalized Fermat equation *primitive* if $\gcd(x, y, z) = 1$.

Main Question: *What are the non-trivial primitive solutions to the generalized Fermat equation?*

In [DG], Andrew Granville and I made the following conjecture:

Generalized Fermat Conjecture: *If* $\frac{1}{p} + \frac{1}{q} + \frac{1}{r} < 1$, *then the generalized Fermat equation has no non-trivial primitive solutions except the following:*

$$1^n + 2^3 = 3^2, \quad 2^5 + 7^2 = 3^4, \quad 7^3 + 13^2 = 2^9, \quad 2^7 + 17^3 = 71^2, \quad 3^5 + 11^4 = 122^2,$$

$$17^7 + 76271^3 = 21063928^2, \quad 1414^3 + 2213459^2 = 65^7, \quad 9262^3 + 15312283^2 = 113^7,$$

$$43^8 + 96222^3 = 30042907^2, \quad 33^8 + 1549034^2 = 15613^3.$$

This conjecture is really more of a "provocation", to borrow a term from Barry Mazur. (The five larger solutions were found by a computer search by Beukers and Zagier, after I had conjectured that they did not exist!) But as a measure of the stock I now place in the conjecture, I will offer a reward of

$$300 \left(\frac{1}{\frac{1}{p} + \frac{1}{q} + \frac{1}{r}} - 1 \right)$$

(Canadian) dollars for a non-trivial primitive solution to $x^p + y^q = z^r$ which does not appear in the above list.

***M*-curves**: The solutions of the Fermat equation $x^n + y^n = z^n$ correspond to rational points on an algebraic curve of genus $(n-1)(n-2)/2$. Because equation (1) is not homogeneous in general, its non-trivial solutions correspond to integer points on an affine surface, which is frequently rational (whenever p, q and r are pairwise coprime, for example) and has a complicated singularity at the origin – the primitive solutions corresponding to points which are also integral relative to this singular point.

The multiplicative group also acts on the surface given by equation (1), by

$$\lambda(x, y, z) = (\lambda^{qr} x, \lambda^{pr} y, \lambda^{pq} z),$$

and so it is tempting to view the surface (1) as a curve in some kind of "weighted projective space". This can be done, but this curve is frequently rational, and the primitive solutions do not have a very natural interpretation. Nonetheless, in this diophantine study one is reluctant to abandon the well-tended landscape of curves for the untamed wilds of (singular) algebraic surfaces.

As it turns out, a better framework for discussing primitive solutions of the generalized Fermat equation is supplied by the notion of a *curve with multiplicities*, or an M-curve, which is defined as follows:

Definition *An M-curve over a field K is a smooth projective curve X/K, together with the assignment, for each point $P \in X(K)$, of a* multiplicity $m_P \in \{1, 2, 3, \ldots\} \cup \{\infty\}$, *such that $m_P = 1$ for all but finitely many P.*

Notation: We denote by

$$\underline{X} = (X; P_1, m_1; P_2, m_2; \ldots; P_r, m_r)$$

the M-curve whose underlying projective curve is X, and such that $m_{P_i} = m_i$, and $m_Q = 1$ if $Q \notin \{P_1, \ldots, P_r\}$.

Remark: Our primary interest being Diophantine, we will mainly consider the case where K is a number field. In this case, we extend X to a smooth proper model $\mathcal{X}$ over $\mathcal{O}_{K,S}$, the ring of S-integers of K, where S is a finite set of primes containing the primes of bad reduction for X. For example, one could work with a minimal model for X, but this is not necessary: the statements that will be made later will be true for any choice of $\mathcal{X}$. Having fixed such a model $\mathcal{X}$ allows us to talk about $X(A) := \mathcal{X}(A)$ for any $\mathcal{O}_{K,S}$-algebra A. Note that since $\mathcal{X}$ is proper, $X(\mathcal{O}_{K,S}) = X(K)$.

Intersection numbers: If P and Q are distinct K-rational points of X (giving rise to sections on $\mathcal{X}$ over $\mathrm{Spec}(\mathcal{O}_{K,S})$) and $v \notin S$ is a place of K with associated prime ideal p_v, define the *arithmetic intersection number* $(P \cdot Q)_v$ as follows: it is the largest positive integer m such that P and Q have the same image in $\mathcal{X}(\mathcal{O}_{K,S}/p_v^m)$. Of course, this arithmetic intersection number depends on the model $\mathcal{X}$ for X chosen above. But given any two models $\mathcal{X}$ and $\mathcal{X}'$, the arithmetic intersection numbers $(P \cdot Q)_v$ agree for all but finitely many places v.

Definition: *An S-integral point on $\underline{X}$ is a point $Q \in X(K)$ satisfying:*

$$(Q \cdot P)_v \equiv 0 \pmod{m_P}, \qquad \forall\ P \in X(K), \quad v \notin S.$$

The set of S-integral points on $\underline{X}$ is denoted $\underline{X}(\mathcal{O}_{K,S})$.

This definition is completed by the following remarks:

1. We adopt the convention that a congruence modulo ∞ is an equality. In this way, the S-integral points on $\underline{X} = (X; P_1, \infty; P_2, \infty; \ldots; P_r, \infty)$ are just the S-integral points on $\mathcal{X}$, relative to the effective divisor $P_1 + \cdots + P_r$.

2. A smooth projective curve is a special case of an M-curve (where one assigns a multiplicity of 1 to every K-rational point). In that case, the set $\underline{X}(\mathcal{O}_{K,S})$ of S-integral points is equal to the set $X(K)$ of K-rational (or, equivalently, integral) points on the underlying projective curve.

3. *A caveat*: Note that the definition of $\underline{X}(\mathcal{O}_{K,S})$ depends on the choice of model $\mathcal{X}$ of X over $\mathcal{O}_{K,S}$. So one must take this model as part of the defining data for the M-curve $\underline{X}$. In what follows, we will often abuse notations and speak of the S-integral points on $\underline{X}$.

The motivating example: If $X = \mathbf{P}^1$, with its usual model over $\mathbf{Z}$, then the rational points on X are identified with the set $\mathbf{Q} \cup \{\infty\}$. If $t \in X(\mathbf{Q}) - \{\infty\}$, then, writing $t = \frac{a}{b}$ as a fraction in lowest terms, we have

$$(t \cdot 0)_v = \mathrm{ord}_v(a); \quad (t \cdot 1)_v = \mathrm{ord}_v(a-b); \quad (t \cdot \infty)_v = \mathrm{ord}_v(b). \qquad (2)$$

Let $\mathbf{P}^1_{p,q,r}$ denote the M-curve $(\mathbf{P}^1; 0, p; 1, q; \infty, r)$. Then an integral point on $\mathbf{P}^1_{p,q,r}$ corresponds to a rational number $t = \frac{a}{b}$ in lowest terms such that, for all rational primes v:

$$\mathrm{ord}_v(a) \equiv 0 \pmod{p}; \quad \mathrm{ord}_v(a-b) \equiv 0 \pmod{q}; \quad \mathrm{ord}_v(b) \equiv 0 \pmod{r}.$$

By the unique factorization in $\mathbf{Z}$, it follows that

$$a = \pm x^p, \quad a - b = \pm y^q, \quad b = \pm z^r,$$

so that (x, y, z) is a primitive solution to the equation

$$\pm x^p \pm y^q = \pm z^r.$$

Hence, the integral points on the M-curve $\mathbf{P}^1_{p,q,r}$ correspond to primitive solutions of the generalized Fermat equation (up to a small sloppiness in signs, which matters only when more than two of p, q and r are even).

Another example: Let $f(x,y) = (x - \alpha_1 y)\cdots(x - \alpha_r y)$ be a square-free homogeneous polynomial of degree r with coefficients in $\mathbf{Z}$. Let K be the extension of $\mathbf{Q}$ generated by the α_i, and let S be the set of primes of K dividing $\mathrm{Disc}(f(x,1))$. Then a solution of the equation $z^m = f(x,y)$ studied in [DG] gives rise to an S-integral point $t = \frac{x}{y}$ on the M-curve over K

$$(\mathbf{P}^1; \alpha_1, m; \alpha_2, m; \ldots; \alpha_r, m; \infty, m/\gcd(m,r)).$$

Maps between M-curves: The M-curves form a category, which ought to be thought of as a natural "enlargement" of the category of curves. We describe now what are the morphisms in this category.

Let $\underline{X}$ and $\underline{Y}$ be M-curves over K, and let X and Y be the underlying smooth projective curves. If π is any morphism (defined over K) from X to Y, and P is a closed point of X, we denote by $e_\pi(P)$ the ramification index of π at the point P.

Definition *A morphism $\underline{\pi} : \underline{X} \longrightarrow \underline{Y}$ is a smooth proper morphism $\pi : X \longrightarrow Y$ with the property that, for all closed points $P \in X$,*

$$m_{\pi(P)} \text{ divides } e_\pi(P) m_P.$$

The ratio $e_\pi(P) m_P / m_{\pi(P)}$ is called the ramification index *of $\underline{\pi}$ at P, and is denoted $e_{\underline{\pi}}(P)$.*

The morphism $\underline{\pi}$ is called *unramified* if $e_{\underline{\pi}}(P) = 1$ for all closed points P. The degree of $\underline{\pi}$ is simply defined to be the degree of the underlying curve morphism π.

By enlarging the set S if necessary, we can assume (and always will, from now on) that π extends to a smooth proper morphism over $\mathcal{O}_{K,S}$ between our chosen S-integral models for X and Y. Once this is done, we have the following statement which justifies our definition of morphisms:

Proposition: *Let $\underline{\pi} : \underline{X} \longrightarrow \underline{Y}$ be a morphism of M-curves over K. Then*

$$\underline{\pi}(\underline{X}(\mathcal{O}_{K,S})) \subset \underline{Y}(\mathcal{O}_{K,S}).$$

The fact that the morphism $\underline{\pi} : \underline{X} \longrightarrow \underline{Y}$ sends S-integral points to S-integral points follows directly from the behaviour of the intersection number under smooth proper morphisms.

The Chevalley-Weil theorem: If $\underline{\pi} : \underline{X} \longrightarrow \underline{Y}$ is a morphism of M-curves, and P is an S-integral point on $\underline{Y}$, denote by $K(\underline{\pi}^{-1}(P))$ the smallest field extension of K over which the points in the inverse image of P by π are defined. The "'lifting problem", broadly stated, is the question of controlling the field $K(\underline{\pi}^{-1}(P))$ - say, by bounding a priori its degree, ramification, or discriminant. For example, when does an S-integral point of $\underline{Y}$ necessarily lift to an S-integral point on $\underline{X}$, by $\underline{\pi}$? The following is the classical theorem of Chevalley-Weil (cf. [La], ch. 2, §8) for M-curves:

Chevalley-Weil theorem: *If $\underline{\pi}$ is unramified, then $K(\underline{\pi}^{-1}(P))$ is unramified outside of S for all $P \in \underline{Y}(\mathcal{O}_{K,S})$.*

This theorem is quite familiar in the case of curves:

1. If $\pi : E \longrightarrow E$ is an isogeny of elliptic curves over K, then π is unramified. If S is a set of places containing the bad reduction primes for E and those dividing the degree of π, and P is any point in $E(K)$, then $\pi^{-1}(P)$ is an extension of K which is unramified outside S. This theorem plays a key role in the proof of the (weak) Mordell-Weil theorem.

2. The group of S-units in a number field K give rise to S-integral points on the M-curve $\mathbf{G}_m := (\mathbf{P}^1; 0, \infty; \infty, \infty)$. The morphism $\mathbf{G}_m \longrightarrow \mathbf{G}_m$ which sends x to x^m is unramified, and indeed the field obtained by adjoining to K an m-th root of an S-unit is unramified outside the places in S and those dividing m. (For which the morphism has "bad reduction".)

For further discussion, and a proof of the Chevalley-Weil theorem for M-curves when the underlying curve is $\mathbf{P}_1$, see [Be].

Orbifolds, and the topology of M-curves: If X is a projective curve, its complex points $X(\mathbf{C})$ form a compact Riemann surface in a natural way. For each $P \in X(\mathbf{C})$, let t_P denote a uniformizing element for the local ring of $X(\mathbf{C})$ at P.

One can associate to the data $\underline{X}$ an *orbifold*, denoted $\underline{X}(\mathbf{C})$. Its underlying set is the same as that of the Riemann surface $X(\mathbf{C})$, but its sheaf of analytic functions is defined differently: a function is now said to be locally analytic at P on the orbifold $\underline{X}(\mathbf{C})$ if its image in the local ring $\mathbf{C}[[t_P]]$ belongs to the subring $\mathbf{C}[[t_P^{m_P}]]$. One denotes by $\mathcal{O}_{\underline{X},P} = \mathbf{C}[[t_P^{m_P}]]$ the ring of locally analytic functions on $\underline{X}$ at P.

A morphism $\underline{\pi} : \underline{X}(\mathbf{C}) \longrightarrow \underline{Y}(\mathbf{C})$ of orbifolds is simply an analytic morphism $\pi : X(\mathbf{C}) \longrightarrow Y(\mathbf{C})$ of the underlying Riemann surfaces with the property that $\pi^*(\mathcal{O}_{\underline{Y},\pi(P)}) \subset \mathcal{O}_{\underline{X},P}$, where $\pi^*(f) := f\pi$ is the pullback of f by π.

With these definitions, the reader will check that the assignment $\underline{X} \mapsto \underline{X}(\mathbf{C})$ defines a functor from the category of M-curves to the category of orbifolds which extends the usual functor sending a curve X to its underlying Riemann surface $X(\mathbf{C})$.

The Euler characteristic of a Riemann surface is defined for orbifolds by the more general formula:

$$\chi(\underline{X}(\mathbf{C})) = 2 - 2g(X(\mathbf{C})) - \sum_P \left(1 - \frac{1}{m_P}\right),$$

where $g(X(\mathbf{C}))$ is the genus of the Riemann surface $X(\mathbf{C})$, and the sum is taken over all points of $X(\mathbf{C})$, with the obvious convention that $\frac{1}{\infty} = 0$. Note that almost all the terms in the sum are equal to 0. If $\underline{X}$ is a projective curve, then this is the usual Euler characteristic; in general, it is a rational number.

Riemann-Hurwitz theorem: *If $\underline{\pi} : \underline{X} \longrightarrow \underline{Y}$ is a degree d morphism of M-curves, then*

$$\chi(\underline{X}(\mathbf{C})) = d\chi(\underline{Y}(\mathbf{C})) - \sum_P (e_{\underline{\pi}}(P) - 1),$$

where the sum is taken over the points of $X(\mathbf{C})$.

For Riemann surfaces, this is the usual Riemann-Hurwitz formula. The proof in the case of orbifolds proceeds by a direct reduction to the case of Riemann surfaces.

A covering lemma: The following lemma allows us to reduce diophantine questions about M-curves to similar questions about curves, for which they have been more studied.

Covering lemma: *If $\underline{X}$ is an M-curve over K with $\chi(\underline{X}) < 0$, then there exists a curve $\tilde{X}$ defined over some number field M, and an unramified morphism of M-curves $\underline{\pi} : \tilde{X} \longrightarrow \underline{X}$ defined over M.*

Proof: This result follows directly from Riemann's existence theorem: the issue is to produce a covering of the curve X, with "'prescribed ramification data". See for example [Se1].

Faltings plus epsilon: We remind the reader of Faltings' theorem for curves, formerly known as the Mordell conjecture:

Faltings' theorem: *If X is a projective curve over K with $\chi(X) < 0$, then $X(K)$ is finite.*

The theorem "Faltings plus epsilon" alluded to in the title is simply the Mordell conjecture for M-curves.

Theorem (Faltings plus epsilon): *If $\underline{X}$ is an M-curve over K with $\chi(\underline{X}) < 0$, then $\underline{X}(\mathcal{O}_{K,S})$ is finite.*

Proof: (Cf. [DG], sec. 3.)
1. By Riemann's existence theorem, there is an unramified morphism

$$\underline{\pi} : \tilde{X} \longrightarrow \underline{X},$$

where $\tilde{X}$ is a curve defined over some number field $M \supset K$. We extend this morphism to a smooth proper morphism over $\mathcal{O}_{M,S'}$, where S' is some finite set of places of M containing all the places above those in S.

2. If P is a point of $\underline{X}(\mathcal{O}_{M,S'})$, then P lifts to a point in $\tilde{X}(M_P)$, where M_P is an extension of M of degree at most d, which is unramified outside S', by the Chevalley-Weil theorem. By a theorem of Minkowski, there are finitely many such fields M_P. Let L be the compositum of all of them. It is of finite degree over K, and

$$\underline{\pi}(\tilde{X}(L)) \supset \underline{X}(\mathcal{O}_{M,S'}) \supset \underline{X}(\mathcal{O}_{K,S}). \tag{3}$$

3. By the Riemann-Hurwitz formula,

$$\chi(\tilde{X}) = d\chi(\underline{X}) < 0.$$

Therefore $\tilde{X}(L)$ is finite by Faltings' theorem. Hence so is $\underline{X}(\mathcal{O}_{K,S})$, by (3). The theorem follows.

Remarks:
1. Note that Faltings plus epsilon, applied to the M-curve

$$(\mathbf{P}^1; 0, \infty; 1, \infty; \infty, \infty)$$

gives Siegel's theorem on the finiteness of S-integral points on $\mathbf{P}^1 - \{0, 1, \infty\}$. Siegel's proof is more difficult than the one given above, because Siegel did not have the luxury of invoking Faltings' theorem. But unramified coverings of $\mathbf{P}^1 - \{0, 1, \infty\}$ also play an important role in Siegel's original proof.

2. Of course, the deepest ingredient in the proof of "Faltings plus epsilon" is Faltings' theorem invoked in step 3. The reduction to Faltings' theorem in the case of curves exploits the Chevalley-Weil theorem and the finiteness theorem of Minkowski in much the same way that it is used by Weil in his proof of the weak Mordell-Weil theorem for elliptic curves and abelian varieties. Weil's proof has its roots directly in Fermat's method of descent. Another connection between Fermat and "Faltings plus epsilon" is given by the following corollary:

Corollary: *If $\frac{1}{p}+\frac{1}{q}+\frac{1}{r} < 1$, then the generalized Fermat equation $x^p+y^q = z^r$ has only finitely many primitive integer solutions.*

Proof: The primitive solutions to the generalized Fermat equation give rise to integral points on the M-curve $\mathbf{P}^1_{p,q,r}$. But

$$\chi(\mathbf{P}^1_{p,q,r}) = \frac{1}{p} + \frac{1}{q} + \frac{1}{r} - 1 < 0,$$

so that $\mathbf{P}^1_{p,q,r}(\mathbf{Z})$ is finite, by "Faltings plus epsilon".

Beyond Faltings? Admittedly, mere finiteness of the solution set for a given (p, q, r) is not all that is wanted. But the proof of "Faltings plus epsilon" suggests a *general program* for studying the generalized Fermat equation $x^p + y^q = z^r$.

1. The geometric step: Find an "explicit" unramified covering

$$\underline{\pi} : X \longrightarrow \mathbf{P}^1_{p,q,r},$$

where X is a projective curve. By explicit, we mean one whose field of definition, and set of primes of bad reduction, can be controlled, and are not too large.

2. The arithmetic step: Understand the lifting problem, i.e, show that a point in $\mathbf{P}^1_{p,q,r}(\mathbf{Z})$ necessarily lifts to a point in $X(\mathbf{Q})$, or in $X(K)$ where K is a specific extension of the rationals which is not too large.

3. The diophantine step: Analyze carefully, and bound, the rational points on X, or those defined over the field K obtained in step 2.

The Fermat equation: To apply this general program to the study of the usual Fermat equation $x^p + y^p = z^p$, (where p is, say, an odd prime) one needs to start with an unramified covering of the M-curve $\mathbf{P}^1_{p,p,p}$, i.e., a covering $X \longrightarrow \mathbf{P}^1$ which is unramified over $\mathbf{P}^1 - \{0, 1, \infty\}$ and such that the ramification indices of all the points lying above 0, 1 and ∞ are equal to p.

The first example of such a covering that comes to mind is, of course, the Fermat curve itself. If F is the curve defined by the equation $x^p + y^p = 1$, then the map $\pi : F \longrightarrow \mathbf{P}^1_{p,p,p}$ which sends (x, y, z) to $t = \frac{x^p}{z^p}$ is an unramified morphism of degree p^2. It has good reduction outside p, and it even becomes a Galois covering over the field $\mathbf{Q}(\zeta_p)$ of pth roots of unity, with Galois group $\mathbf{Z}/p\mathbf{Z} \times \mathbf{Z}/p\mathbf{Z}$. The lifting problem in this case is tautological: a point on $\mathbf{P}^1_{p,p,p}(\mathbf{Z})$ lifts to a rational point on the Fermat curve. This leads (in a roundabout way) to the traditional geometric approach to Fermat's Last Theorem.

One can also consider the curve X whose function field is the fraction field of

$$\mathbf{Q}[X, Y, (X + \zeta_p^j Y)^{\frac{1}{p}}]/(X^p + Y^p - 1),$$

where ζ_p is a primitive p-th root of unity. It is an extension of $\mathbf{Q}(\zeta_p, X^p/Y^p)$ of degree p^{p+1} which has a solvable Galois group. A point in $\mathbf{P}^1_{p,p,p}(\mathbf{Z})$ lifts to a point on this curve defined over a field K which is an abelian extension of $\mathbf{Q}(\zeta_p)$ unramified outside of p. A more careful analysis allows one to analyze the ramification at p precisely, and in understanding the possible fields K that could arise in the lifting problem one is led to questions about the p-part of the ideal class group of the cyclotomic fields $\mathbf{Q}(\zeta_p)$. This is the basis for the attack on Fermat's Last theorem initiated by Kummer via the theory of cyclotomic fields. To this day, a proof of Fermat's Last Theorem based on the covering $X \longrightarrow \mathbf{P}^1_{p,p,p}$ remains elusive, although a number of deep results in this direction have been obtained. The proof of Fermat's Last Theorem

completed by Andrew Wiles had to rely on a completely different covering arising from modular curves. Here are the main lines of the proof of Fermat's last theorem, following the 3-step program described above.

1. The geometric step: modular curves, and the Hellegouarch-Frey trick: Explicit unramified covering of $\mathbf{P}^1_{p,p,p}$ can be obtained from modular curves. More precisely, the modular curve $X(2)$ which classifies elliptic curves together with a basis of points of order 2 is isomorphic to $\mathbf{P}^1$, and is given by the classical λ-line of Legendre, the universal elliptic curve over it being described by the equation

$$y^2 = x(x-1)(x-\lambda).$$

The curve $X(2)$ has three cusps associated to elliptic curves with degenerate reduction, which are given by the values $\lambda = 0, 1, \infty$. Let $X(2)_{p,p,p}$ denote the M-curve whose underlying curve is $X(2)$, with a multiplicity of p attached to each of the three cusps. Let $X(2p)$ be the usual modular curve which classifies elliptic curves with a basis of $2p$-division points. Then the natural projection

$$\underline{\pi}_{Frey} : X(2p) \longrightarrow X(2)_{p,p,p}$$

is unramified, and has good reduction outside of $2p$.

The M-curve $X(2)_{p,p,p}$ is a model for $\mathbf{P}^1_{p,p,p}$, and its integral points correspond to Frey curves via the moduli interpretation. More precisely, the point $\lambda = -a^p/b^p$ of $X(2)_{p,p,p}$ (where $a^p + b^p = c^p$) corresponds to the curve $y^2 = x(x-1)(x+\frac{a^p}{b^p})$, which is a twist of the Frey curve

$$y^2 = x(x-a^p)(x+b^p).$$

The field $K = \underline{\pi}^{-1}_{Frey}(\lambda)$ is closely related to the field of definition of the p-division points of this Frey curve. One is thus led to consider the p-division field of the Frey curve. (In fact, it is better to rigidify the situation somewhat, and consider the mod p Galois representation attached to the p-torsion points of the Frey curve, instead of merely the field cut out by this representation.)

2. The arithmetic step: the Ribet-Wiles theorem: Because the covering $\underline{\pi} : X(2p) \longrightarrow X(2)_{p,p,p}$ is unramified and has good reduction outside of $2p$, the field $\underline{\pi}^{-1}(\lambda)$ is unramified[1] outside of $2p$, for all $\lambda \in X(2)_{p,p,p}$.

[1]This can also be seen by analyzing directly the field of p-division points of the Frey curve, using Tate's analytic theory.

The work of Frey, Mazur, Serre, Ribet, and finally Wiles was directly concerned with the lifting problem associated to this covering. Let $X_0(2,p)$ be the modular curve which classifies elliptic curves with full level 2 structure and a rational subgroup of order p. This curve has 6 cusps, of which 3 are unramified for the natural projection $X_0(2,p) \longrightarrow X(2)$. Let $X_0(2,p)_{p,p,p}$ be the M-curve obtained from $X_0(2,p)$ by assigning a multiplicity of p to each of these three cusps. Then the covering $X_0(2,p)_{p,p,p} \longrightarrow X(2)_{p,p,p}$ is unramified.

Ribet-Wiles theorem *A point in $X(2)_{p,p,p}(\mathbf{Z})$ lifts to an integral point on $X_0(2,p)_{p,p,p}(\mathbf{Z})$.*

It seems tempting to tackle this statement head on, and try to supply a direct proof. Yet a staggering amount of difficult mathematics is involved in Ribet and Wiles' argument, which rests on the deep interplay between Galois representations and modular forms. We will not even begin to scratch the surface here! An expository account of parts of their proof (described along more conventional lines) can be found in [DDT], [Se2], [Ri1] and [Wi].

3. The diophantine step: Mazur's theorem: It turns out that the Diophantine step (step 3) had been handled earlier by Mazur in his fundamental papers [Ma1] and [Ma2] on the Eisenstein ideal. In particular, it follows from Mazur's results that

Theorem (Mazur): *A point in $X(2)_{p,p,p}(\mathbf{Z})$ does not lift to an integral point on $X_0(2,p)_{p,p,p}(\mathbf{Z})$.*

Proof: The Frey curve associated to $\lambda = -a^p/b^p$ is a twist of a semistable elliptic curve. Mazur shows that such a curve cannot have a rational subgroup of order p for $p \geq 5$. The result follows.

Combining the theorems of Ribet-Wiles and Mazur gives a contradiction, and Fermat's Last Theorem follows.

Modular curves and the generalized Fermat equation: Since coverings coming from modular curves have been so effective in proving Fermat's Last Theorem, it is natural to ask the following question:

Question: *What are the unramified coverings of $\mathbf{P}^1_{p,q,r}$ arising from modular curves?*

The modular curve $X_0(2)$ has two cusps, and a special point P_{1728} corresponding to an elliptic curve with invariant $j = 1728$, at which the natural

map to the j-line is unramified. Let $X_0(2)_{2,p,p}$ be the M-curve whose underlying curve is $X_0(2) \simeq \mathbf{P}^1$, and where a multiplicity of 2 has been assigned to P_{1728}, and a multiplicity of p to each of the two cusps. There is an isomorphism of $X_0(2)_{2,p,p}$ with $\mathbf{P}^1_{2,p,p}$ defined over $\mathbf{Z}[\frac{1}{2}]$. If $X(2,p)$ is the curve which classifies elliptic curves with a point of order 2 and full level p structure, then the natural projection

$$X(2,p) \longrightarrow X_0(2)_{2,p,p}$$

is unramified and has good reduction outside of $2p$.

Likewise, the modular curve $X_0(3)$ has two cusps, and a special point P_0 corresponding to an eliptic curve with invariant $j = 0$, at which the natural map to the j-line is unramified. We define the M-curve $X_0(3)_{3,p,p}$ by assigning a multiplicity of 3 to P_0, and p to each of the cusps; using the same notation as before, we find that the covering

$$X(3,p) \longrightarrow X_0(3)_{3,p,p}$$

is unramified and has good reduction outside of $3p$.

By exploiting these two coverings, and following the Mazur-Ribet-Wiles approach, Loïc Merel and I proved the following theorem [DM] towards the generalized Fermat conjecture, which is the theorem"Wiles plus epsilon" referrred to in the title:

Theorem (Wiles plus epsilon):
1. The equation $x^n + y^n = z^2$ has no non-trivial primitive solution for $n \geq 4$.
2. If the Shimura-Taniyama conjecture is true, then the equation $x^n + y^n = z^3$ has no non-trivial primitive solution for $n \geq 3$.

The Shimura-Taniyama conjecture needs to be assumed for part 2 because the elliptic curves that arise in the proof are not known to be modular, in spite of the recent work of Conrad, Diamond and Taylor: their conductor is frequently divisible by 27.

Can one go further than this? Here is a table listing the exponents (p,q,r) for which one might tackle the generalized Fermat equation by exploiting a "Frey curve" construction.

(p,q,r)	Frey curve for $a^p + b^q = c^r$	Δ
$(2,3,p)$	$y^2 = x^3 + 3bx + 2a$	$-2^6 3^3 c^p$
$(3,3,p)$	$y^2 = x^3 + 3(a-b)x^2 + 3(a^2 - ab + b^2)x$	$-2^4 3^3 c^{2p}$
$(4,p,4)$	$y^2 = x^3 + 4acx^2 - (a^2 - c^2)^2 x$	$2^6(a^2 - c^2)^2 b^{2p}$
$(5,5,p)$	$y^2 = x^3 - 5(a^2 + b^2)x^2 + 5\frac{a^5+b^5}{a+b}x$	$2^4 5^3 (a+b)^2 c^{2p}$
$(7,7,p)$	$y^2 = x^3 + (a^2 + ab + b^2)x^2$ $-(2a^4 - 3a^3 b + 6a^2 b^2 - 3ab^3 + 2b^4)x$ $-(a^6 - 4a^5 b + 6a^4 b^2 - 7a^3 b^3 + 6a^2 b^4 - 4ab^5 + b^6)$	$2^4 7^2 (\frac{a^7+b^7}{a+b})^2$
$(p,p,2)$	$y^2 = x^3 + 2cx^2 + a^p x$	$2^6(a^2 b)^p$
$(p,p,3)$	$y^2 + cxy = x^3 - c^2 x^2 - \frac{3}{2}cb^p x + b^p(a^p + \frac{5}{4}b^p)$	$3^3(a^3 b)^p$
(p,p,p)	$y^2 = x(x - a^p)(x + b^p)$	$2^4(abc)^{2p}$

The cases of exponents $(p,p,2)$ and $(p,p,3)$ are disposed of in [DM], and the results proved there also imply that the equation with exponents $(4,p,4)$ has no non-trivial primitive solution for $p > 2$. (Cf. [Da].) But the methods used by Frey, Serre, Mazur, Ribet and Wiles to eventually resolve Fermat's Last Theorem are extremely delicate, particularly as concerns the Ribet-Wiles lifting theorem. For the other triples of exponents, one seems to run into difficulties caused by the presence of modular forms. In spite of this, A. Kraus [Kr] has obtained some partial results in the case of exponent $(3,3,p)$, which imply in particular that the associated generalized Fermat equation has no non-trivial primitive solution when $17 \leq p \leq 10000$.

In conclusion, here are two questions:

1. Can one refine the existing techniques based on elliptic curves, modular forms, and Galois representations to prove the generalized Fermat conjecture for all the exponent listed in the above table?

2. Can one find other examples of unramified coverings $X \longrightarrow \mathbf{P}^1_{p,q,r}$ (admitting, perhaps, a nice moduli-theoretic interpretation) for which a program of attack similar to the one of Mazur, Ribet and Wiles can be carried out?

These questions may appear ambitious, especially the second. Of course, the generalized Fermat equation fits right into the body of questions adressed by the famous abc conjecture, which has received much recent attention

although a proof seems nowhere in sight. Thus we can hope that the Queen of Mathematics will hold on to the mystery of the generalized Fermat equation for at least a few more decades, to the bafflement (and delight) of number theorists, amateur and professional alike.

References

[Be] S. Beckmann, *On extensions of number fields obtained by specializing branched coverings*, Jour. für die reine und ang. Math. **419** (1991), 27–53.

[Da] H. Darmon, *The equation* $x4 - y^4 = z^p$, C.R. Math. Rep. Acad. Sci. Canada. **XV** No. 6 (1993) pp. 286-290.

[DDT] H. Darmon, F. Diamond, R. Taylor, *Fermat's Last Theorem*, Current Developments in Math, Vol. **1**, pp. 1–157, International Press, 1996.

[DG] H. Darmon, A. Granville, *On the equations* $x^p + y^q = z^r$ *and* $z^m = f(x, y)$, Bulletin of the London Math. Society, no 129, **27** part 6, November 1995, pp. 513–544.

[DM] H. Darmon, L. Merel, *Winding quotients and some variants of Fermat's Last Theorem*, Journal für die Reine und Angewandte Mathematik, to appear.

[Kr] A. Kraus, *Sur l'équation* $a^3 + b^3 = c^p$, J. of Experimental Math., to appear.

[La] S. Lang, Fundamentals of Diophantine Geometry, Springer-Verlag, 1983.

[Ma1] B. Mazur, *Modular curves and the Eisenstein ideal*, Publ. Math. IHES **47** (1977) 33–186.

[Ma2] B. Mazur, *Rational isogenies of prime degree*, Inv. Math. **44** (1978), 129–162.

[Ma3] B. Mazur, *Questions about number*, in: New Directions in Mathematics, to appear.

[Ri1] K. Ribet, *On modular representations of Gal($\bar{\mathbf{Q}}/\mathbf{Q}$) arising from modular forms*, Invent. Math. **100** (1990), 431–476.

[Ri2] K. Ribet, *On the equation* $a^p + 2^\alpha b^p + c^p = 0$, Acta Arithmetica, to appear.

[Se1] J.-P. Serre, Topics in Galois Theory, Jones and Bartlett, 1992.

[Se2] J.-P. Serre, *Sur les représentations modulaires de degré* 2 *de Gal*($\bar{\mathbf{Q}}/\mathbf{Q}$), Duke Math. J. Vol. **54** no. 1 (1987), 179–230.

[TW] R. Taylor and A. Wiles, *Ring theoretic properties of certain Hecke algebras*, Annals of Math. **141**, No. 3, 1995, pp. 553-572.

[Wi] A. Wiles, *Modular elliptic curves and Fermat's last theorem*, Annals of Math. **141**, No. 3, 1995, pp. 443-551.