A Proof of the Full Shimura-Taniyama-Weil Conjecture Is Announced

Henri Darmon[*]

September 9, 2007

On June 23, 1993, Andrew Wiles unveiled his strategy for proving the Shimura-Taniyama-Weil conjecture for *semistable* elliptic curves defined over the field $\mathbb{Q}$ of rational numbers. Thanks to the work of Gerhard Frey, Jean-Pierre Serre and Kenneth Ribet, this was known to imply Fermat's Last Theorem. Six years later, Christophe Breuil, Brian Conrad, Fred Diamond, and Richard Taylor have finally announced a proof of the full Shimura-Taniyama-Weil conjecture, for all elliptic curves over $\mathbb{Q}$.

The Conjecture

The Shimura-Taniyama-Weil conjecture relates *elliptic curves*—cubic equations in two variables of the form $y^2 = x^3 + ax + b$, where a and b are rational numbers—and *modular forms*—objects, to be defined below, arising as part of an ostensibly different circle of ideas.

An elliptic curve E can be made into an abelian group in a natural way, after adjoining to it an extra "solution at infinity" that plays the role of the identity element. This is what makes elliptic curves worthy of special study, for they alone, among all *projective curves* (equations in two variables,

[*] Henri Darmon is an associate professor of mathematics at McGill University and a member of CICMA (Centre Interuniversitaire en Calcul Mathématique Algébrique) and the CRM (Centre de Recherches Mathématiques). His e-mail address is darmonmath.mcgill.ca.

compactified by the adjunction of suitable points at infinity) are endowed with such a natural group law. If one views solutions geometrically as points in the (x, y)-plane, the group operation consists in connecting two points on the curve by a straight line, finding the third point of intersection of the line with the curve, and reflecting the resulting point about the x-axis.

After a change of variables is performed to bring it into the best possible form, the equation defining E can be reduced modulo any prime number p. If the resulting equation is nonsingular over the finite field with p elements $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, then E is said to have *good reduction* at p. All but finitely many primes are primes of good reduction for a given E. For example, the elliptic curve defined by the cubic equation

$$y^2 = x^3 - x^2 + 1/4, \qquad \text{or equivalently,} \qquad y^2 + y = x^3 - x^2, \tag{1}$$

has good reduction at all primes except 11.

Let N_p be the number of solutions (over $\mathbb{F}_p$) of the reduced equation and set $a_p(E) = p - N_p$. The sequence $\{a_p(E)\}_p$ (indexed by the primes p of good reduction) encodes basic arithmetic information on E. The sequence $a_p(E)$ for the elliptic curve of equation (1) is given in Table 1.

p	2	3	5	7	11	13	17	19	23	29	31	$\cdots$	10007
N_p	4	4	4	9	–	9	19	19	24	29	24	$\cdots$	9989
$a_p(E)$	-2	-1	1	-2	–	4	-2	0	-1	0	7	$\cdots$	18

Table 1. Sequence $a_p(E)$ for the elliptic curve (1).

It has been a long-standing concern of number theory to search for patterns satisfied by sequences of this sort. For example, in the simpler case of the quadratic equation in one variable $x^2 - d = 0$ with d an integer, any prime p that does not divide $2d$ is a prime of good reduction, and for such a p the integer N_p is equal to 2 or 0, depending on whether d is a square or not modulo p. Gauss's quadratic reciprocity law implies that this seemingly subtle property of p depends only on the residue class of p modulo $4d$, so that the sequence N_p obeys a simple periodicity law.

In the case of elliptic curves, a similar pattern arises. It is, however, a good deal more subtle—so much so that it emerged as a precise conjecture only in the 1950s through the work of Shimura, Taniyama, and Weil. This pattern involves the notion of a *modular form of weight two*—an analytic function on the complex upper half-plane $\{z \in \mathbb{C} \text{ with } \mathrm{Im}(z) > 0\}$, satisfying suitable growth conditions at the boundary as well as a transformation rule of the form

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^2 f(z), \quad \text{for all } \left(\begin{array}{cc} a & b \\ c & d \end{array}\right) \in \Gamma,$$

where Γ is an appropriate "congruence subgroup" of $\mathbf{SL}_2(\mathbb{Z})$. The main example of a congruence subgroup, sufficient for the formulation of the Shimura-Taniyama-Weil conjecture, is the group $\Gamma_0(N)$ of matrices in $\mathbf{SL}_2(\mathbb{Z})$ whose lower left entries are divisible by N. A *modular form of weight two* on $\Gamma_0(N)$ (also said to be *of level* N) is, in particular, invariant under translation by 1, and it can be expressed as a Fourier series

$$f(z) = \sum_{n=0}^{\infty} a_n(f) q^n, \quad \text{where } q = e^{2\pi i z}.$$

Of particular interest are the so-called "cusp forms", satisfying a more stringent growth condition at the boundary that implies in particular that $a_0(f) = 0$. The Shimura-Taniyama-Weil conjecture asserts that, if E is an elliptic curve over $\mathbb{Q}$, then there is an integer $N \geq 1$ and a weight-two cusp form f of level N, normalized so that $a_1(f) = 1$, such that

$$a_p(E) = a_p(f),$$

for all primes p of good reduction for E. When this is the case, the curve E is said to be *modular*. The conjecture also predicts the precise value of N: it should be equal to the "conductor" of E, an arithmetically defined quantity that measures the Diophantine complexity of the associated cubic equation. Its prime divisors are precisely the primes of bad reduction of E. If p divides N but p^2 does not, then E is said to have *semistable* reduction at p. In particular, E has semistable reduction at all primes p (i.e., is *semistable*), precisely when N is squarefree.

For instance, the elliptic curve of equation (1) has conductor 11 (and thus is an example of a semistable elliptic curve). It turns out that the space of weight two cusp forms of level 11 is one-dimensional, and is spanned by the function

$$\begin{aligned} q \prod_{n=1}^{\infty} & (1-q^n)^2 \cdot (1-q^{11n})^2 \\ = \quad & q - \mathbf{2q^2} - \mathbf{q^3} + 2q^4 + \mathbf{q^5} + 2q^6 - \mathbf{2q^7} \\ & -2q^9 - 2q^{10} + \mathbf{q^{11}} - 2q^{12} + \mathbf{4q^{13}} + 4q^{14} \\ & -q^{15} - 4q^{16} - \mathbf{2q^{17}} + 4q^{18} + 2q^{20} + 2q^{21} \\ & -2q^{22} - \mathbf{q^{23}} - 4q^{25} - 8q^{26} + 5q^{27} - 4q^{28} \\ & +2q^{30} + \mathbf{7q^{31}} + \cdots + \mathbf{18q^{10007}} + \cdots \end{aligned}$$

The reader will note that the Fourier coefficients of this function agree with the numbers computed—by wholly different methods—in Table 1.

The Shimura-Taniyama-Weil conjecture was widely believed to be unbreachable, until the summer of 1993, when Wiles announced a proof that every semistable elliptic curve is modular. A full proof of this result appeared in 1994 in the two articles [W] and [TW], the second joint with Taylor. Shortly afterwards, Diamond [Di1] was able to remove the semistability assumption in Wiles's argument at all the primes except 3 and 5. Then in 1998 Conrad, Diamond, and Taylor [CDT] refined the techniques still further, establishing the Shimura-Taniyama-Weil conjecture for all elliptic curves whose conductor is not divisible by 27. This is where matters stood at the start of the summer of 1999, before the announcement of Breuil, Conrad, Diamond, and Taylor.

The Importance of the Conjecture

The Shimura-Taniyama-Weil conjecture, and its subsequent, just-completed proof, stand as a crowning achievement of number theory in the twentieth century. This statement can be defended on (at least) three levels.

Fermat's Last Theorem

Firstly, the Shimura-Taniyama-Weil conjecture implies Fermat's Last Theorem. This is surprising at first, because the equation $x^n + y^n = z^n$ is not a cubic and bears, on the face of it, no relation with elliptic curves. But to a nontrivial solution $a^p + b^p = c^p$ of Fermat's equation with prime exponent $p > 5$, Frey associated the elliptic curve (now known as a "Frey curve") given by the equation $y^2 = x(x - a^p)(x + b^p)$. The conductor of E, when a, b and c are relatively prime, is the product of the primes dividing abc (so that in particular E is semistable). Ribet, guided by conjectures of Serre, proved that such an elliptic curve could not possibly correspond to a modular form in the way predicted by the Shimura-Taniyama-Weil conjecture.

Because the Frey curve is semistable, the original result of [W] and [TW] is enough to imply Fermat's Last Theorem, and the new result of Breuil, Conrad, Diamond, and Taylor yields nothing new on Fermat's equation. It does imply, however, other results of the same nature, such as the statement that a perfect cube cannot be written as a sum of two relatively prime nth powers with $n \geq 3$, generalizing Euler's result for $n = 3$. As in the case of Fermat's Last Theorem, a solution to the equation $a^p + b^p = c^3$ is used to construct an elliptic curve whose existence is shown in [DM] to contradict the Shimura-Taniyama-Weil conjecture. In many cases—whenever 3 does not divide ab—the conductor of this curve is divisible by 27, so that the full strength of the result of Breuil, Conrad, Diamond, and Taylor is needed to conclude the argument.

The Arithmetic of Elliptic Curves

Secondly, and more centrally perhaps, the Shimura-Taniyama-Weil conjecture lies at the heart of the theory of elliptic curves.

A theorem of Mordell asserts that the abelian group, denoted $E(\mathbb{Q})$, of points of E with rational coordinates is *finitely generated*, so that it is isomorphic as an abstract group to $\mathbb{Z}^r \oplus T$, where T is finite. It is known how to determine T explicitly. The integer $r \geq 0$, called the *rank* of E over $\mathbb{Q}$, is a more subtle invariant: no algorithm is known at present to calculate r as a function of E.

It has been a longstanding feeling that much information on the arithmetic of E (such as the invariant r) can be gleaned from the sequence $N_p(E)$, or equivalently $a_p(E)$ as p varies. A convenient way to package the information contained in this sequence is to form the L-series of E, a function of the complex variable s defined initially by the Euler product

$$L(E, s) := \prod_{p \nmid N} (1 - a_p(E)p^{-s} + p^{1-2s})^{-1}.$$

(In the later parts of the theory, elementary factors are included in the product for the finitely many primes p dividing N.) This product converges when Re(s)$> 3/2$. But the Shimura-Taniyama-Weil conjecture gives a strong control on the (arithmetically defined) sequence $a_p(E)$, and implies through the work of Hecke that $L(E,s)$ extends to an analytic function on the entire complex plane. In particular, it makes sense to study the behavior of $L(E,s)$ in a neighborhood of $s=1$. Note that formally,

$$L(E,1) = \prod_p \frac{p}{N_p+1}.$$

It is believed that the size of r might affect the size of N_p on average, which may in turn be reflected in the analytic behavior of $L(E,s)$ near $s=1$. Indeed, Birch and Swinnerton–Dyer in the 1960s conjectured that the *order of vanishing* of $L(E,s)$ at $s=1$ is equal to r:

$$\mathrm{ord}_{s=1} L(E,s) = r.$$

This conjecture is of fundamental importance for the arithmetic of elliptic curves, and is still far from being settled, although the work of Gross-Zagier and Kolyvagin shows that it is true when $\mathrm{ord}_{s=1}(L(E,s)) \leq 1$.

Knowing that E is modular also gives control on the arithmetic of E in other ways, by allowing the construction of certain global points on E defined over abelian extensions of quadratic imaginary fields, via the theory of complex multiplication. Such analytic constructions of global points on E actually play an important role in studying the Birch and Swinnerton–Dyer conjecture, through the work of Gross-Zagier and of Kolyvagin.

The Langlands Program

A *Galois representation* is a (finite-dimensional) representation

$$\rho : G_{\mathbb{Q}} \longrightarrow \mathbf{GL}_n(F),$$

where $G_{\mathbb{Q}} := \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ is the absolute Galois group of $\mathbb{Q}$, and F is any field. (Of special interest are the case where $F = \mathbb{C}$, $\bar{\mathbb{Q}}_\ell$, or $\bar{\mathbb{F}}_\ell$.)

Wiles's work can be viewed in the broader perspective of establishing connections between automorphic forms—objects arising in the (infinite-dimensional) representation theory of adelic groups—and Galois representations. Viewed in this light, it becomes part of a vast conjectural edifice put together by Langlands, based on earlier insights of Tate, Shimura, Taniyama, and many others. In this setting, Wiles's discoveries have enriched the theory with a powerful new method that should keep the experts occupied well into the new millenium. Indeed, the impact of Wiles's ideas has only started being felt in many diverse aspects of the Langlands program:

Two-dimensional complex representations of $G_{\mathbb{Q}}$: Emil Artin associated to a Galois representation $\rho : G_{\mathbb{Q}} \longrightarrow \mathbf{GL}_n(\mathbb{C})$ an L-function $L(\rho, s)$ and conjectured that it has an analytic continuation to the whole complex plane. Via work of Deligne and Serre, the Langlands program relates such representations, when $n = 2$, to certain "cusp forms of weight one" on a group slightly different from $\Gamma_0(N)$. This relation implies the analytic continuation of $L(\rho, s)$, just as the modularity of an elliptic curve implies the analytic continuation of its L-series through the work of Hecke. Before Wiles, the only cases that could be attacked with any generality were the case where ρ is reducible, by work of Hecke, and where the image of ρ is solvable, thanks to the work of Langlands and Tunnell. (It should be noted that for a specific ρ, the modular form attached to it can in principle be found after a finite amount of computation, so that the Langlands conjecture could be checked for a finite number of ρ with nonsolvable image; the first such example was produced by Joe Buhler in his Harvard PhD thesis.)

Using Wiles's method, Taylor has formulated a novel strategy [Ta] for proving the Artin conjecture in the remaining (most interesting) case where the image of ρ in $\mathbf{PGL}_2(\mathbb{C})$ is isomorphic to A_5—the so-called *icosahedral case.* Enough of Taylor's program has now been carried out in joint work with Kevin Buzzard, Mark Dickinson, and Nicholas Shepherd–Barron to establish the truth of the Artin conjecture for infinitely many icosahedral Galois representations.

Generalizations to other number fields. A number of ingredients in Wiles's method have been significantly simplified, by Diamond and Fujiwara among others. Fujiwara, Skinner, and Wiles have been able to extend Wiles's results to the case where the field $\mathbb{Q}$ is replaced by a totally real number field K. In particular, this yields analogues of the Shimura-Taniyama-Weil conjecture for a large class of elliptic curves defined over such a field.

n-dimensional generalizations. Michael Harris and Richard Taylor have explored generalizations of the main results of [W] and especially [TW] to the

context of n-dimensional representations of $G_{\mathbb{Q}}$. (This work, and the proof of the local Langlands conjecture for $\mathbf{GL}_n$ by Harris and Taylor, is expected to be covered in a future *Notices* article.)

The Work of Breuil, Conrad, Diamond, and Taylor

The space $S_2(N)$ of weight-two cusp forms on $\Gamma_0(N)$ is a finite-dimensional complex vector space, equipped with an action of a natural family of commuting self-adjoint operators, the so-called "Hecke operators". A *normalized newform on* $\Gamma_0(N)$ is a simultaneous eigenvector for these operators, normalized so that its first Fourier coefficient is equal to 1, and not already arising in the space of cusp forms on $\Gamma_0(D)$ for any D dividing N. A construction of Eichler and Shimura associates to a normalized newform of level N with *rational* Fourier coefficients an elliptic curve over $\mathbb{Q}$ of conductor N. The original Shimura-Taniyama-Weil conjecture states that this construction yields a bijection from the set of normalized newforms on $\Gamma_0(N)$ with rational Fourier coefficients to the set of elliptic curves of conductor N, taken modulo an equivalence, weaker than isomorphism, known as *isogeny.* It appears difficult, even now, to give an *a priori* estimate for the size of either set, as a function of N; in fact, the question of the rationality of the Fourier coefficients of an eigenform is a subtle one that seems hard to come to terms with.

In general, the Fourier coefficients of a normalized eigenform f are *algebraic numbers*, defined over a finite extension $K_f \subset \bar{\mathbb{Q}}$ of $\mathbb{Q}$. Fix a prime ℓ, and an inclusion $\iota : \bar{\mathbb{Q}} \longrightarrow \bar{\mathbb{Q}}_\ell$, where $\mathbb{Q}_\ell$ is the field of ℓ-adic numbers. By a generalization of the Eichler-Shimura construction, f gives rise to an ℓ-adic Galois representation

$$\rho_f : \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \longrightarrow \mathbf{GL_2}(\bar{\mathbb{Q}}_\ell)$$

satisfying $\mathrm{trace}(\rho_f(\mathrm{Frob}_p)) = \iota(a_p(f))$, for all primes p not dividing $N\ell$. Here Frob_p is the "Frobenius element" at p. A notion of *conductor* can be defined for an ℓ-adic Galois representation, and it follows from the work of Carayol, Deligne, Igusa, Langlands, and Shimura, that the conductor of ρ_f is equal to the level of f.

When f is an eigenform with rational Fourier coefficients, corresponding to an elliptic curve E_f under the original Eichler-Shimura construction, then ρ_f is simply obtained by piecing together the natural action of $G_{\mathbb{Q}}$ on the space of ℓ^n-torsion points of $E_f(\bar{\mathbb{Q}})$, as n varies.

It becomes natural to formulate a more general version of the Shimura-Taniyama-Weil conjecture, replacing elliptic curves with two-dimensional representations of $G_{\mathbb{Q}}$ with coefficients in $\bar{\mathbb{Q}}_\ell$. This more general version would have the virtue of avoiding the subtleties associated with fields of definition of Fourier coefficients of eigenforms.

An important insight that emerged over the last decades through the work of Alexander Grothendieck, Pierre Cartier, Jean Dieudonné, and finally Jean-Marc Fontaine and his school, is that it should be possible to characterize the ℓ-adic representations arising from modular forms entirely in Galois-theoretic terms—or, more precisely, in terms of their restriction to a "decomposition group" $\mathrm{Gal}(\bar{\mathbb{Q}}_\ell/\mathbb{Q}_\ell)$ at ℓ, or even an "inertia group" at this prime. Such representations are called "potentially semistable", and this notion is a key ingredient for generalizing the Shimura-Taniyama-Weil conjecture to ℓ-adic Galois representations. Around 1990, Fontaine and Mazur conjectured that the ℓ-adic Eichler-Shimura construction yields a bijection from the set $\Lambda_{mod}(N)$ of normalized eigenforms of level N to the set $\Lambda_{gal}(N)$ of ℓ-adic Galois representations of conductor N that are potentially semistable at ℓ. Wiles's proof in essence amounts to a sophisticated counting argument in which these two sets are compared and found to be of the same size.

The main tools in controlling the size of Λ_{mod} are supplied by the theory of "Hecke rings" and congruences between modular forms, a rich body of techniques developed by Mazur, Hida, and Ribet, and used to great effect by Ribet to derive Fermat's Last Theorem from the Shimura-Taniyama-Weil conjecture.

The set $\Lambda_{gal}(N)$ is, in many ways, the more subtle object of the two, about which there is *a priori* the least explicit information. There are two major ingredients used to estimate the size of $\Lambda_{gal}(N)$ and relate it to $\Lambda_{mod}(N)$.

- The theory of "base change", and in particular the work of Langlands and Tunnell on solvable base change.

- The theory of deformations of Galois representations pioneered by Mazur and Hida.

The second ingredient is extremely general and flexible, and is being honed into a powerful tool in the arithmetic study of automorphic forms. The first ingredient, by contrast, is available only when the image of ρ_f is a (pro)-*solvable* group. Number theory has, over the last hundred years, developed an arsenal of techniques for understanding abelian and solvable extensions, as evidenced in Class Field Theory, which gives a precise description of all the abelian extensions of a given number field, as well as the behavior of the Frobenius elements in these extensions. Arriving at an understanding of nonsolvable extensions on the same terms has proved far more elusive.

Unfortunately, the image of ρ_f is rarely solvable. But it is when the prime ℓ is 3, by a fortuitous accident of group theory: the group $\mathbf{GL}_2(\mathbb{F}_3)$, and hence $\mathbf{GL_2}(\mathbb{Z}_3)$, is solvable, a fact that ceases to be true as soon as 3 is replaced by any larger prime. It is for this reason that in the application to the Shimura-Taniyama-Weil conjecture it is indispensable in Wiles's strategy to work with the prime $\ell = 3$.

The last obstacle to carrying out Wiles's program to a complete proof of the Shimura-Taniyama-Weil conjecture arose from a technical difficulty: the 3-adic Galois representations of conductor N, when 27 divides N, have an intricate behavior when restricted to the inertia group at 3—and a precise description, and understanding, of this behavior is required to control the set $\Lambda_{gal}(N)$ when $\ell = 3$. Overcoming this difficulty required some new insights into the structure of 3-adic representations of $G_{\mathbb{Q}}$ that are "highly ramified" at the prime 3. A number of these key insights were provided by the work of Breuil strengthening Fontaine's theory.

References

[CDT] Conrad, Brian; Diamond, Fred; Taylor, Richard. *Modularity of certain potentially Barsotti-Tate Galois representations.* J. Amer. Math. Soc. **12** (1999), no. 2, 521–567.

[DM] Darmon, Henri; Merel, Loïc. *Winding quotients and some variants of Fermat's last theorem.* J. Reine Angew. Math. **490** (1997), 81–100.

[Di1] Diamond, Fred. *On deformation rings and Hecke rings.* Ann. of Math. (2) **144** (1996), no. 1, 137–166.

[Ta] Taylor, Richard. *Icosahedral Galois representations.* Olga Taussky-Todd: in memoriam. Pacific J. Math. 1997, Special Issue, 337–347.

[TW] Taylor, Richard; Wiles, Andrew. *Ring-theoretic properties of certain Hecke algebras.* Ann. of Math. (2) **141** (1995), no. 3, 553–572.

[W] Wiles, Andrew. *Modular elliptic curves and Fermat's last theorem.* Ann. of Math. (2) **141** (1995), no. 3, 443–551.