Recent Work on Serre's Conjectures

Kenneth A. Ribet

UC Berkeley

UC Irvine
May 23, 2007

Prelude

In 1993–1994, I was among the number theorists who lectured to a variety of audiences about the proof of Fermat's Last Theorem that Andrew Wiles had announced in June, 1993. Seventeen summers ago, I spoke extensively with journalists who were preparing articles about the proof.

In our discussions, we stressed the theme that modularity bridges the worlds of algebra (elliptic curves) and analysis (modular forms).

Bridges

This talk will be about bridges, but only in passing about elliptic curves. We seek to link:

- Modular forms, holomorphic functions with many symmetries. Those that interest us can be written as sums $\sum_{n=1}^{\infty} c(n)q^n$, where the $c(n)$ are algebraic integers and where $q = e^{2\pi iz}$.
- The Galois group $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$, where $\overline{\mathbf{Q}}$ is the field of algebraic complex numbers. This group, which no one actually ever sees directly, is the limit of groups $\mathrm{Gal}(K/\mathbf{Q})$, where $K/\mathbf{Q}$ is a finite Galois extension and $K \subset \overline{\mathbf{Q}}$.

Modular forms and Galois representations

In his 1967–1968 DPP seminar on modular forms (“Une interprétation des congruences relatives à la a fonction τ de Ramanujan”), J-P. Serre proposed a new bridge of this kind. He sought to relate Galois representations to holomorphic modular forms that are eigenforms for Hecke operators.

Almost immediately, P. Deligne constructed the representations whose existence was conjectured by Serre.

For example, let

$$\Delta = q \prod_{n=1}^{\infty} (1 - q^n)^{24}, \quad q = e^{2\pi i z}$$

and expand Δ as a power series with integer coefficients:

$$\Delta = \sum_{n=1}^{\infty} \tau(n) q^n = q - 24q^2 + 252q^3 - 1472q^4 + \cdots .$$

For each prime p, there is a continuous representation

$$\rho_p : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathbf{GL}(2, \mathbf{F}_p)$$

whose arithmetic is tied up to that of the $\tau(n)$.

Concretely, to give ρ_p is to give a finite Galois extension $K/\mathbf{Q}$ along with an embedding $\mathrm{Gal}(K/\mathbf{Q}) \hookrightarrow \mathbf{GL}(2, \mathbf{F}_p)$.

The representation ρ_p is unramified at all primes different from p; in other words, the discriminant of K is a power of p. If $\ell \neq p$ is a prime and Frob_ℓ is a Frobenius element for ℓ in $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$, then the matrix $\rho_p(\mathrm{Frob}_\ell)$ has trace $\tau(\ell)$ mod p and determinant ℓ^{11} mod p.

These constraints determine ρ_p up to isomorphism once we replace ρ_p by its semisimplification in the rare situation where it is not already simple.

Serre and Swinnerton-Dyer studied the ρ_p for Δ and some other modular forms in the early 1970s. Using the ρ_p, they showed that the numbers $\tau(n)$ satisfy no congruences other than those that had been established by Ramanujan and others in the early part of the 20th century.

Around 1975, I extended their study to modular forms on the full modular group $\mathbf{SL}(2, \mathbf{Z})$ whose coefficients are not necessarily *rational* integers.

Meanwhile, Serre began asking whether there might be a converse to Deligne's construction. Suppose that $\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathbf{GL}(2, \mathbf{F})$ is a continuous homomorphism, where $\mathbf{F}$ is a finite field. To fix ideas, suppose that ρ is simple (i.e., that there is no basis in which the image of ρ is upper-triangular). Assume also that ρ is unramified at all primes other than p, the characteristic of $\mathbf{F}$.

Might we then guess that ρ is isomorphic to a representation associated with a form on $\mathbf{SL}(2, \mathbf{Z})$?

There is a salient necessary condition involving parity. Each non-zero form f on $\mathbf{SL}(2,\mathbf{Z})$ has a unique weight that describes the behavior of f under fractional linear transformations; for example, the weight of Δ is 12. Since forms are trivially invariant under $z \mapsto \frac{-z}{-1} = z$, it turns out that the weight of a form on $\mathbf{SL}(2,\mathbf{Z})$ is always even.

If ρ is a mod p representation associated to a form of weight k, the the determinant of $\rho(\mathrm{Frob}_\ell)$ is ℓ^{k-1} mod p; this is an odd power of ℓ mod p.

It follows from the Cebotarev density theorem that $\det \rho = \chi_p^{k-1}$, where χ_p is the mod p cyclotomic character: the homomorphism $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathbf{F}_p^*$ giving the action of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ on the pth roots of unity in $\overline{\mathbf{Q}}$. What's important is that $\det \rho$ is always odd in the sense that it takes the value -1 on the element "complex conjugation" of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. Indeed, it's an odd power of the mod p cyclotomic character; this character is odd because complex conjugation inverts roots of unity.

Serre's conjecture for modular forms of level 1 (i.e., those on $\mathbf{SL}(2,\mathbf{Z})$) states: *if a continuous odd irreducible representation* $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathbf{GL}(2,\mathbf{F}_{p^\nu})$ *is unramified outside p, then it arises from a modular form on* $\mathbf{SL}(2,\mathbf{Z})$.

This conjecture was advanced as a question to J. T. Tate in 1972 and was discussed, for instance, in Serre's article for the Journées Arithmétiques de Bordeaux ("Valeurs propres des opérateurs de Hecke modulo *l*").

Using discriminant bounds, Tate proved the conjecture for mod 2 representations. Later, Serre proved the conjecture for $p = 3$ in a similar manner. In 1999, S. Brueggeman treated the case $p = 5$ modulo the generalized Riemann hypothesis.

For those values of p, there simply are no representations ρ of the type contemplated by the conjecture! As Serre explained in Bordeaux, the conjecture predicts in fact that there are no irreducible ρ as in the conjecture for $p < 11$.

In 1987, Serre published his Duke Journal article "Sur les représentations modulaires de degré 2 de $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$," which extended his conjecture to continuous irreducible representations $\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathbf{GL}(2, \mathbf{F})$ that are not necessarily unramified outside p. If such a representation satisfies the necessary parity condition, it should come from a holomorphic modular form.

The level and weight of the modular form are predicted by Serre's conjecture from the local behavior of the representation: The level of the form depends on the restriction of ρ to inertia subgroups of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ for primes $\ell \neq p$. The weight depends on the restriction to the inertia subgroup at p.

In this respect, Serre's conjecture is analogous to the modularity conjecture that predicted that the Galois representations of an elliptic curve should arise from a specific space of modular forms—namely, the space of weight-2 forms of level equal to the conductor of the elliptic curve.

Around the time that Serre published his conjecture, I proved a "level-lowering" theorem that reduced Fermat's Last Theorem to the modularity conjecture for semistable elliptic curves. My "level-lowering" theorem, plus work of Edixhoven, Carayol, Diamond, Buzzard and others, implies that if a Galois representation comes from some modular form, it actually comes from a form of the predicted weight and level (at least for $p > 2$). Said differently: the mere modularity of Galois representations (the "weak conjecture" of Serre) implies the precise statement that Serre made in his article (the "strong conjecture").

In his 1987 article, conjecture(s) were viewed as far beyond the horizon. Serre pointed out that his conjectures implied Fermat's Last Theorem (essentially because of the level-lowering built into the statement of the strong conjecture), the modularity of elliptic curves over $\mathbf{Q}$, and a number of other statements. The strength of the conjectures made them appear well beyond the mathematical horizon.

One notable consequence of Serre's conjectures is the modularity of abelian varieties of $\mathbf{GL}(2)$-type over $\mathbf{Q}$: if the endomorphism algebra of an abelian variety A over $\mathbf{Q}$ contains a number field of degree equal to $\dim A$, then A is a quotient of the Jacobian of some classical modular curve.

The landscape has changed considerably in the last 2–3 years and even in the last several months! In an exciting series of preprints, a group of mathematicians including Chandrashekhar Khare, Jean-Pierre Wintenberger and Mark Kisin have established Serre's conjecture! In what follows, I will sketch some of their ideas. Many of the key themes emerge already in studying the original situation considered by Serre and Tate, namely the modularity of odd mod p representations that are unramified outside p.

For citations, see the most recent preprint available from `http://www.math.uchicago.edu/~kisin/` and the articles by Khare and others that are listed in the references of this preprint.

Review of FLT

To understand the work of Khare et. al., it is useful to recall how the modularity lifting theorem of Taylor–Wiles comes into the proof of Fermat's Last Theorem: The proof starts with a Fermat counterexample

$$a^p + b^p = c^p, \quad p \geq 5$$

and hones in on the mod p representation

$$\rho_p : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathbf{GL}(2, \mathbf{F}_p)$$

that gives the action of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$ on the p-division points on $E : y^2 = x(x - a^p)(x + b^p)$.

This representation is odd and it is irreducible (Mazur). According to Serre's conjecture, ρ_p should come from a space of modular forms that is observably 0.

Moreover, by level-lowering, once ρ_p can be associated to some modular form, we can conclude with a contradiction. Thus, as noted before, to prove Fermat's Last Theorem we need "only" verify Serre's conjecture for ρ_p.

An important point is that ρ_p is part of the *family* of ℓ-adic representations coming from E; these are Galois representations $\tilde{\rho}_\ell$, one for each prime ℓ. The reduction ρ_ℓ of $\tilde{\rho}_\ell$ is the mod ℓ representation attached to E. We no longer really care about E—but rather only about the system $(\tilde{\rho}_\ell)$!

Because all the representations $\tilde{\rho}_\ell$ are linked through the behavior of Frobenius elements, a single $\tilde{\rho}_\ell$ is modular if and only if all $\tilde{\rho}_\ell$ are modular. In particular, if one $\tilde{\rho}_\ell$ is modular, then the reduction ρ_p of $\tilde{\rho}_p$ is modular, and we are done.

Furthermore, the *Modularity lifting theorems* of Taylor–Wiles, Skinner–Wiles, Kisin and others (including Breuil, Diamond, Conrad, Fujiwara, and Savitt) tend to show that $\tilde{\rho}_\ell$ is modular if its reduction ρ_ℓ is modular.

Wearing our rose-colored glasses, we put 2 and 2 together and conclude that the target representation ρ_p is modular once there is a single prime ℓ such that ρ_ℓ is modular.

In Wiles's case, ρ_3 was known be modular because its image is contained in a solvable $\mathbf{GL}(2)$. Indeed, Langlands and Tunnell had essentially proved the modularity of two-dimensional Galois representations with solvable image using Langlands's base-change machine that he constructed after seeing the work of Saito and Shintani.

So, to everyone's astonishment, Wiles's strategy of proof established the modularity of ρ_p (and thus FLT) from the modularity of ρ_3.

Side Remark

The proof of modularity that Wiles used does not generalize readily to the case of an abelian variety of $\mathbf{GL}(2)$-type. The problem is that we might encounter a non-solvable $\mathbf{GL}(2, \mathbf{F}_{3^\nu})$.

As you may recall, even for FLT, the situation was more complicated because of the possibility that ρ_3 might be reducible. This led Wiles to his "3–5 trick" and required him to consider two elliptic curves (and thus two families of representations) in some circumstances.

In 1993–1994, I lectured about FLT and often described modularity as "contagious." If two different systems of Galois representations have a common reduction, then one can often deduce modularity of one whole system from modularity of the other.

In proving Serre's conjecture, an immediate obstacle is the lack of a support system $(\tilde{\rho}_\ell)$; we have only ρ_p.

The existence of $(\tilde{\rho}_\ell)$ is a logical consequence of Serre's conjecture. As soon as the conjecture was promulgated, several people asked whether it would be possible to construct this system, or at least a p-adic lift of the given mod p Galois representation.

In the late 1990s, Ravi Ramakrishna lifted $\rho_p : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathbf{GL}(2, \mathbf{F})$ to $\tilde{\rho}_p : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathbf{GL}(2, W)$ where W is the ring of Witt vectors of $\mathbf{F}$, but only at the expense of allowing $\tilde{\rho}_p$ to be ramified at some primes where ρ_p is unramified. However, the goal was to build the system of lifts so that it would retain the same arithmetic properties of the original mod p representation.

In their work, Khare–Wintenberger rely on results of Böckle and others to find a $\tilde{\rho}_p : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathbf{GL}(2, \mathcal{O})$ with the desired "minimal" ramification properties. Here $\mathcal{O}$ is the ring of integers of a (possibly ramified!) finite extension of the field of fractions of W.

To insert $\tilde{\rho}_p$ into a family $(\tilde{\rho}_\ell)$ of representations that are compatible with $\tilde{\rho}_p$. K. and W. use "potential modularity" theorems of Richard Taylor ("Remarks on a conjecture of Fointaine and Mazur," "On the meromorphic continuation of degree two L-functions," c. 1999). Taylor proved, for example, that there are totally real number fields F so that ρ_p becomes modular after restriction to $\mathrm{Gal}(\overline{\mathbf{Q}}/F)$. One can select F so that p is unramified in $F/\mathbf{Q}$ and so that the group $\rho_p(\mathrm{Gal}(\overline{\mathbf{Q}}/F))$ is not too different from the image of ρ_p.

A guiding philosophy (due to Nigel Boston) is to manipulate deformations of mod p Galois representations in a way that mimics mod p congruences among modular forms. The theory of congruences among modular forms is quite rich and includes a plethora of results about level-raising and tradeoffs between the levels and weights of mod p forms that give the same mod p Galois representation.

This philosophy is exploited big time by Khare–Wintenberger. To give one example: if ρ_p is unramified outside p and has Serre weight $p+1$, they can choose the family $(\tilde{\rho}_\ell)$ as if it came from a modular form of weight 2 on the subgroup $\Gamma_0(p)$ of $\mathbf{SL}(2,\mathbf{Z})$, or alternatively so it seems to come from a form of weight $p+1$ and level 1. Here, they are guided by a theorem of Serre that relates mod p forms of weight $p+1$ with forms of weight 2 on $\Gamma_0(p)$.

Once the family $(\tilde{\rho}_\ell)$ is built, the goal is to find some particular ℓ for which the reduction ρ_ℓ of $\tilde{\rho}_\ell$ is already known to be modular and for which a modularity lifting theorem can prove that $\tilde{\rho}_\ell$ is necessarily modular.

If ρ_p has Serre weight k, then the most natural system $(\tilde{\rho}_\ell)$ has weight k in the sense that each ℓ-adic representation is (locally at ℓ) potentially semistable with Hodge-Tate weights 0, $k-1$. A key hypothesis in modularity lifting is that k should be no larger than $\ell+1$. By twisting ρ_p by a power of the mod p cyclotomic character, we can ensure $k \leq p+1$. Thus if ρ_ℓ is modular for some $\ell \geq p$, ρ_p will be modular. This is already one amazing principle of the proof—that the conjecture for any given prime (and a fixed level) implies the conjecture for all smaller primes (and the same level).

In fact, we need to deal with a system of representations that is indexed naturally by the set of prime ideals of some number field, which we can view as being embedded in a fixed $\overline{\mathbf{Q}}$. For each prime ℓ, we choose and fix a prime of $\overline{\mathbf{Q}}$ that lies over ℓ and use this choice to determine one prime ideal of the number field for each prime number. This allows the shorthand notation $\tilde{\rho}_\ell$; the "ℓ" is really a prime ideal of a number field. As we move around in the system of representations, we never need two prime ideals that belong to the same rational prime ℓ. On the other hand, we have to move around from prime to prime quite deftly.

Once we have the idea that we need to prove the conjecture only for an infinite sequence of primes p, then we think inductively. Given that the conjecture is true mod p, we seek a prime $P > p$ for which we can prove the conjecture. It turns out to be sufficient to find a prime P that is not a Fermat prime (!) and that is no bigger than a complicated bound that is a tad smaller than $2p$. Grosso modo, for this we need only Bertrand's "postulate," or more precisely a refinement of the arguments that were supplied by Chebyshev in 1850 to prove the postulate.

Assume that we know the conjecture mod p and seek to deduce it mod P, where (roughly)

$$p < P < 2p.$$

We need to establish the modularity of a mod P representation ρ_P; we can and do assume that its weight k satisfies $p + 1 < k \leq P + 2$.

If ρ_P does come from a modular form of weight k (and level 1, say—for simplicity, we pretend now that we are proving the original conjecture for forms on $\mathbf{SL}(2, \mathbf{Z})$), then it comes from a weight-2 form of level P whose associated Dirichlet character is $\epsilon = \omega^{k-2}$, where ω is the character of order $P - 1$ that's often referred to as the Teichmüller character.

We don't yet have any forms on the table, but we can construct a system $(\tilde{\rho}_\ell)$ that is like the system coming from such a form; we insist that the mod P reduction of $\tilde{\rho}_P$ be the given representation ρ_P.

We select wisely a prime q dividing $P-1$ and consider the mod q representation ρ_q. In optimal circumstances, ρ_q would come from a weight-2 form f of level P whose character is ϵ. If ϵ' is a second power of ω with the same mod q reduction as ϵ, ρ_q would be attached also to a weight-2 form f' with character ϵ'. We need to take ϵ' to be the product of ϵ and some power of $\omega^{(P-1)/q}$. We choose this power so that $\epsilon' = \omega^i$, where i is around $P/2$, which is around p.

There is as of yet no f and no f'. Nonetheless, we (i.e., Khare–Wintenberger) can construct $(\tilde{\rho}'_\ell)$ as if it came from f'! Keeping track of things, they show that (some twist of) ρ'_P now has low weight (and level 1), so it is modular.

Then a cascade: by modularity lifting, $(\tilde{\rho}'_\ell)$ is modular, so that ρ'_q, in particular, is modular. But $\rho'_q \approx \rho_q$, so ρ_q is modular. Applying modularity lifting again, we get that $(\tilde{\rho}_\ell)$ is modular. Hence, in particular, the original ρ_P is modular!