Annals of Mathematics, **142** (1995), 443–551

Modular elliptic curves and Fermat's Last Theorem

By Andrew Wiles*

For Nada, Clare, Kate and Olivia

> *Cubum autem in duos cubos, aut quadratoquadratum in duos quadratoquadratos, et generaliter nullam in infinitum ultra quadratum potestatem in duos ejusdem nominis fas est dividere: cujus rei demonstrationem mirabilem sane detexi. Hanc marginis exiguitas non caperet.*
>
> *Pierre de Fermat*

Introduction

An elliptic curve over **Q** is said to be modular if it has a finite covering by a modular curve of the form $X_0(N)$. Any such elliptic curve has the property that its Hasse-Weil zeta function has an analytic continuation and satisfies a functional equation of the standard type. If an elliptic curve over **Q** with a given j-invariant is modular then it is easy to see that all elliptic curves with the same j-invariant are modular (in which case we say that the j-invariant is modular). A well-known conjecture which grew out of the work of Shimura and Taniyama in the 1950's and 1960's asserts that every elliptic curve over **Q** is modular. However, it only became widely known through its publication in a paper of Weil in 1967 [We] (as an exercise for the interested reader!), in which, moreover, Weil gave conceptual evidence for the conjecture. Although it had been numerically verified in many cases, prior to the results described in this paper it had only been known that finitely many j-invariants were modular.

In 1985 Frey made the remarkable observation that this conjecture should imply Fermat's Last Theorem. The precise mechanism relating the two was formulated by Serre as the ε-conjecture and this was then proved by Ribet in the summer of 1986. Ribet's result only requires one to prove the conjecture for semistable elliptic curves in order to deduce Fermat's Last Theorem.

*The work on this paper was supported by an NSF grant.

Our approach to the study of elliptic curves is via their associated Galois representations. Suppose that ρ_p is the representation of $\mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q})$ on the p-division points of an elliptic curve over $\mathbf{Q}$, and suppose for the moment that ρ_3 is irreducible. The choice of 3 is critical because a crucial theorem of Langlands and Tunnell shows that if ρ_3 is irreducible then it is also modular. We then proceed by showing that under the hypothesis that ρ_3 is semistable at 3, together with some milder restrictions on the ramification of ρ_3 at the other primes, every suitable lifting of ρ_3 is modular. To do this we link the problem, via some novel arguments from commutative algebra, to a class number problem of a well-known type. This we then solve with the help of the paper [TW]. This suffices to prove the modularity of E as it is known that E is modular if and only if the associated 3-adic representation is modular.

The key development in the proof is a new and surprising link between two strong but distinct traditions in number theory, the relationship between Galois representations and modular forms on the one hand and the interpretation of special values of L-functions on the other. The former tradition is of course more recent. Following the original results of Eichler and Shimura in the 1950's and 1960's the other main theorems were proved by Deligne, Serre and Langlands in the period up to 1980. This included the construction of Galois representations associated to modular forms, the refinements of Langlands and Deligne (later completed by Carayol), and the crucial application by Langlands of base change methods to give converse results in weight one. However with the exception of the rather special weight one case, including the extension by Tunnell of Langlands' original theorem, there was no progress in the direction of associating modular forms to Galois representations. From the mid 1980's the main impetus to the field was given by the conjectures of Serre which elaborated on the ε-conjecture alluded to before. Besides the work of Ribet and others on this problem we draw on some of the more specialized developments of the 1980's, notably those of Hida and Mazur.

The second tradition goes back to the famous analytic class number formula of Dirichlet, but owes its modern revival to the conjecture of Birch and Swinnerton-Dyer. In practice however, it is the ideas of Iwasawa in this field on which we attempt to draw, and which to a large extent we have to replace. The principles of Galois cohomology, and in particular the fundamental theorems of Poitou and Tate, also play an important role here.

The restriction that ρ_3 be irreducible at 3 is bypassed by means of an intriguing argument with families of elliptic curves which share a common ρ_5. Using this, we complete the proof that all semistable elliptic curves are modular. In particular, this finally yields a proof of Fermat's Last Theorem. In addition, this method seems well suited to establishing that all elliptic curves over $\mathbf{Q}$ are modular and to generalization to other totally real number fields.

Now we present our methods and results in more detail.

Let f be an eigenform associated to the congruence subgroup $\Gamma_1(N)$ of $\mathrm{SL}_2(\mathbf{Z})$ of weight $k \geq 2$ and character χ. Thus if T_n is the Hecke operator associated to an integer n there is an algebraic integer $c(n, f)$ such that $T_n f = c(n, f)f$ for each n. We let K_f be the number field generated over $\mathbf{Q}$ by the $\{c(n, f)\}$ together with the values of χ and let $\mathcal{O}_f$ be its ring of integers. For any prime λ of $\mathcal{O}_f$ let $\mathcal{O}_{f,\lambda}$ be the completion of $\mathcal{O}_f$ at λ. The following theorem is due to Eichler and Shimura (for $k = 2$) and Deligne (for $k > 2$). The analogous result when $k = 1$ is a celebrated theorem of Serre and Deligne but is more naturally stated in terms of complex representations. The image in that case is finite and a converse is known in many cases.

THEOREM 0.1. *For each prime $p \in \mathbf{Z}$ and each prime $\lambda \mid p$ of $\mathcal{O}_f$ there is a continuous representation*

$$\rho_{f,\lambda}\colon \mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q}) \longrightarrow \mathrm{GL}_2(\mathcal{O}_{f,\lambda})$$

which is unramified outside the primes dividing Np and such that for all primes $q \nmid Np$,

$$\operatorname{trace} \rho_{f,\lambda}(\operatorname{Frob} q) = c(q, f), \qquad \det \rho_{f,\lambda}(\operatorname{Frob} q) = \chi(q)q^{k-1}.$$

We will be concerned with trying to prove results in the opposite direction, that is to say, with establishing criteria under which a λ-adic representation arises in this way from a modular form. We have not found any advantage in assuming that the representation is part of a compatible system of λ-adic representations except that the proof may be easier for some λ than for others.

Assume

$$\rho_0 : \mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q}) \longrightarrow \mathrm{GL}_2(\bar{\mathbf{F}}_p)$$

is a continuous representation with values in the algebraic closure of a finite field of characteristic p and that $\det \rho_0$ is odd. We say that ρ_0 is modular if ρ_0 and $\rho_{f,\lambda} \bmod \lambda$ are isomorphic over $\bar{\mathbf{F}}_p$ for some f and λ and some embedding of $\mathcal{O}_f/\lambda$ in $\bar{\mathbf{F}}_p$. Serre has conjectured that every irreducible ρ_0 of odd determinant is modular. Very little is known about this conjecture except when the image of ρ_0 in $\mathrm{PGL}_2(\bar{\mathbf{F}}_p)$ is dihedral, A_4 or S_4. In the dihedral case it is true and due (essentially) to Hecke, and in the A_4 and S_4 cases it is again true and due primarily to Langlands, with one important case due to Tunnell (see Theorem 5.1 for a statement). More precisely these theorems actually associate a form of weight one to the corresponding complex representation but the versions we need are straightforward deductions from the complex case. Even in the reducible case not much is known about the problem in the form we have described it, and in that case it should be observed that one must also choose the lattice carefully as only the semisimplification of $\overline{\rho_{f,\lambda}} = \rho_{f,\lambda} \bmod \lambda$ is independent of the choice of lattice in $K^2_{f,\lambda}$.

If $\mathcal{O}$ is the ring of integers of a local field (containing $\mathbf{Q}_p$) we will say that $\rho : \mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q}) \longrightarrow \mathrm{GL}_2(\mathcal{O})$ is a lifting of ρ_0 if, for a specified embedding of the residue field of $\mathcal{O}$ in $\bar{\mathbf{F}}_p$, $\bar{\rho}$ and ρ_0 are isomorphic over $\bar{\mathbf{F}}_p$. Our point of view will be to assume that ρ_0 is modular and then to attempt to give conditions under which a representation ρ lifting ρ_0 comes from a modular form in the sense that $\rho \simeq \rho_{f,\lambda}$ over $\overline{K_{f,\lambda}}$ for some f, λ. We will restrict our attention to two cases:

(I) ρ_0 is ordinary (at p) by which we mean that there is a one-dimensional subspace of $\bar{\mathbf{F}}_p^2$, stable under a decomposition group at p and such that the action on the quotient space is unramified and distinct from the action on the subspace.

(II) ρ_0 is flat (at p), meaning that as a representation of a decomposition group at p, ρ_0 is equivalent to one that arises from a finite flat group scheme over $\mathbf{Z}_p$, and $\det \rho_0$ restricted to an inertia group at p is the cyclotomic character.

We say similarly that ρ is ordinary (at p) if, viewed as a representation to $\bar{\mathbf{Q}}_p^2$, there is a one-dimensional subspace of $\bar{\mathbf{Q}}_p^2$ stable under a decomposition group at p and such that the action on the quotient space is unramified.

Let $\varepsilon : \mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q}) \longrightarrow \mathbf{Z}_p^\times$ denote the cyclotomic character. Conjectural converses to Theorem 0.1 have been part of the folklore for many years but have hitherto lacked any evidence. The critical idea that one might dispense with compatible systems was already observed by Drinfeld in the function field case [Dr]. The idea that one only needs to make a geometric condition on the restriction to the decomposition group at p was first suggested by Fontaine and Mazur. The following version is a natural extension of Serre's conjecture which is convenient for stating our results and is, in a slightly modified form, the one proposed by Fontaine and Mazur. (In the form stated this incorporates Serre's conjecture. We could instead have made the hypothesis that ρ_0 is modular.)

CONJECTURE. *Suppose that* $\rho\colon \mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q}) \longrightarrow \mathrm{GL}_2(\mathcal{O})$ *is an irreducible lifting of* ρ_0 *and that* ρ *is unramified outside of a finite set of primes. There are two cases:*

(i) *Assume that* ρ_0 *is ordinary. Then if* ρ *is ordinary and* $\det \rho = \varepsilon^{k-1}\chi$ *for some integer* $k \geq 2$ *and some* χ *of finite order,* ρ *comes from a modular form.*

(ii) *Assume that* ρ_0 *is flat and that* p *is odd. Then if* ρ *restricted to a decomposition group at* p *is equivalent to a representation on a* p*-divisible group, again* ρ *comes from a modular form.*

In case (ii) it is not hard to see that if the form exists it has to be of weight 2; in (i) of course it would have weight k. One can of course enlarge this conjecture in several ways, by weakening the conditions in (i) and (ii), by considering other number fields in place of $\mathbf{Q}$ and by considering groups other than GL_2.

We prove two results concerning this conjecture. The first includes the hypothesis that ρ_0 is modular. Here and for the rest of the paper we will assume that p is an odd prime.

THEOREM 0.2. *Suppose that ρ_0 is irreducible and satisfies either* (I) *or* (II) *above. Suppose also that ρ_0 is modular and that*

(i) *ρ_0 is absolutely irreducible when restricted to* $\mathbf{Q}\left(\sqrt{(-1)^{\frac{p-1}{2}}p}\right)$.

(ii) *If $q \equiv -1 \bmod p$ is ramified in ρ_0 then either $\rho_0|_{D_q}$ is reducible over the algebraic closure where D_q is a decomposition group at q or $\rho_0|_{I_q}$ is absolutely irreducible where I_q is an inertia group at q.*

Then any representation ρ as in the conjecture does indeed come from a modular form.

The only condition which really seems essential to our method is the requirement that ρ_0 be modular.

The most interesting case at the moment is when $p = 3$ and ρ_0 can be defined over $\mathbf{F}_3$. Then since $\mathrm{PGL}_2(\mathbf{F}_3) \simeq S_4$ every such representation is modular by the theorem of Langlands and Tunnell mentioned above. In particular, every representation into $\mathrm{GL}_2(\mathbf{Z}_3)$ whose reduction satisfies the given conditions is modular. We deduce:

THEOREM 0.3. *Suppose that E is an elliptic curve defined over $\mathbf{Q}$ and that ρ_0 is the Galois action on the 3-division points. Suppose that E has the following properties*:

(i) *E has good or multiplicative reduction at* 3.

(ii) *ρ_0 is absolutely irreducible when restricted to* $\mathbf{Q}\left(\sqrt{-3}\right)$.

(iii) *For any $q \equiv -1 \bmod 3$ either $\rho_0|_{D_q}$ is reducible over the algebraic closure or $\rho_0|_{I_q}$ is absolutely irreducible.*

Then E is modular.

We should point out that while the properties of the zeta function follow directly from Theorem 0.2 the stronger version that E is covered by $X_0(N)$