ℓ-ADIC MODULAR DEFORMATIONS AND WILES'S "MAIN CONJECTURE"

FRED DIAMOND AND KENNETH A. RIBET

1. Introduction

Let E be an elliptic curve over $\mathbf{Q}$. The Shimura-Taniyama conjecture asserts that E is modular, i.e., that there is a weight-two newform f such that $a_p(f) = a_p(E)$ for all primes p at which E has good reduction. Let ℓ be a prime, choose a basis for the Tate module $T_\ell(E)$ and consider the ℓ-adic representation

$$\rho_{E,\ell} : G_{\mathbf{Q}} \to \mathrm{Aut}(T_\ell(E)) \cong \mathbf{GL}_2(\mathbf{Z}_\ell).$$

Then E is modular if and only if $\rho_{E,\ell}$ is modular, i.e., if and only if $\rho_{E,\ell}$ is equivalent over $\mathbf{Q}_\ell$ to the representation $\rho_{f,\ell}$ for some f (see [22]).

We aim to prove a stronger result which characterizes ℓ-adic representations arising from modular forms. Since the coefficients of a newform lie in the ring of integers of a number field which is not necessarily $\mathbf{Q}$, we are led to consider representations

$$\rho : G_{\mathbf{Q}} \to \mathbf{GL}_2(A) \tag{1}$$

where A is the ring of integers of a finite extension of $\mathbf{Q}_\ell$, and ask which of these arise from modular forms. In fact, it turns out to be convenient to consider representations ρ as in (1) where now A is a complete local Noetherian ring with finite residue field k, and formulate a notion of modularity for such a representation. Roughly speaking, the main theorem of Wiles in [26] supposes we are given such a representation which is "plausibly modular" and concludes that under certain technical hypotheses:

$$\bar\rho \text{ modular} \Longrightarrow \rho \text{ modular}, \tag{2}$$

where $\bar\rho : G_{\mathbf{Q}} \to \mathbf{GL}_2(k)$ is the reduction of ρ.

Returning to the elliptic curve E, suppose we know that

$$\bar\rho_{E,\ell} : G_{\mathbf{Q}} \to \mathbf{GL}_2(\mathbf{F}_\ell)$$

is modular (by Langlands-Tunnell [16], for example, if $\ell = 3$ and the representation is irreducible). A result of the form (2) then implies $\rho_{E,\ell}$ is modular (assuming it is "plausibly modular" and satisfies the technical hypotheses), hence so is E!

To orient the reader, we recall the connection with Fermat's Last Theorem, which begins with an idea of G. Frey (see [14]): Suppose that we have

Date: Version of December 18, 1996.

a non-trivial solution $a^n + b^n = c^n$, where $n \geq 5$ is prime, b is even and $a \equiv -1 \bmod 4$. Consider the elliptic curve

$$E : Y^2 = X(X - a^n)(X + b^n).$$

The modularity of E implies in particular that the mod n representation $\bar{\rho}_{E,n}$ is modular. The main result of [21] (see [12]) then shows that $\bar{\rho}_{E,n}$ arises from a (non-zero) modular form of weight two and level two, a contradiction as there are no such forms.

2. Strategy

We begin our formal discussion by returning to the phrase "plausibly modular," which we used in connection with representations like (1). For simplicity, assume again that A is the ring of integers of a finite extension of $\mathbf{Q}_\ell$. Consider representations $\rho : G_{\mathbf{Q}} \to \mathbf{GL}_2(A)$ which arise from weight-two eigenforms on $\Gamma_0(N)$, where N is allowed to vary. The idea is to exhibit a list of conditions satisfied by these representations which encapsulates their modularity. Those ρ which arise from weight-two forms on $\Gamma_0(N)$ are irreducible with cyclotomic determinant. Further, they are unramified outside a finite set of prime numbers (namely, those not dividing ℓN). It is tempting to guess that any ρ satisfying these simple conditions is likely to be modular. In fact, however, one also needs to impose a further condition on the restriction of ρ to a decomposition group at ℓ. A sufficient condition in this direction is conjectured by Fontaine and Mazur in [13], but with the current technology we need to impose a stronger one in order to obtain results. We will introduce a condition which is both convenient to work with and sufficient for applications to elliptic curves with semistable reduction at ℓ. Namely, suppose now that the prime number ℓ is odd. Then we will assume that the representation ρ is "semistable" [2], i.e., that it is either finite flat or ordinary at ℓ in the terminology of [19].

Let us fix an irreducible mod ℓ representation

$$\bar{\rho} : G_{\mathbf{Q}} \to \mathbf{GL}_2(k)$$

whose determinant is the mod ℓ cyclotomic character. Denote by $\mathcal{R}$ the set of isomorphism classes of "plausibly modular" ρ as above with reduction $\bar{\rho}$. We denote by $\mathcal{T}$ the set of (genuinely) modular isomorphism classes in $\mathcal{R}$. Thus $\mathcal{T}$ is a set of modular forms giving rise to $\bar{\rho}$, and the goal is to prove $\mathcal{R} = \mathcal{T}$.

We suppose that $\mathcal{T}$ is not empty; we can paraphrase this condition by the statement that $\bar{\rho}$ is modular and semistable (locally at the prime ℓ). It is known that $\mathcal{T}$ is infinite [20], so that $\mathcal{R} \supseteq \mathcal{T}$ is infinite. This circumstance makes $\mathcal{R}$ rather unwieldy, so that we are led to filter $\mathcal{R}$ as follows: For each finite set of primes Σ, we will let $\mathcal{R}_\Sigma$ be the set of ρ in $\mathcal{R}$ which are of "type Σ," meaning they are well-behaved outside Σ, and set $\mathcal{T}_\Sigma = \mathcal{T} \cap \mathcal{R}_\Sigma$. Before making precise the notion of being "well-behaved," we remark that $\mathcal{R}$ will be the union of the $\mathcal{R}_\Sigma$ over all finite sets of primes Σ. Therefore, to prove that $\mathcal{R} = \mathcal{T}$, it will suffice to prove

$$\mathcal{R}_\Sigma = \mathcal{T}_\Sigma \tag{3}$$

for all Σ. With the definition we give, each set $\mathcal{R}_\Sigma$ corresponds, at least conjecturally, to an easily described finite set of modular forms.

We now give a preliminary definition of "type Σ" in terms of the conductors[1] $N(\rho)$ and $N(\bar\rho)$: We say that ρ is *type* Σ if Σ contains the set of prime divisors of $N(\rho)/N(\bar\rho)$ (an integer by [1] or [18]).

We shall assume below that $N(\bar\rho)$ is square-free, in which case this preliminary definition of type Σ turns out to be suitable (cf. remark 3.3), but we shall have to extend it (along with the notions of plausibly modular and modular) to representations ρ as in (1) where A is a complete local Noetherian ring with finite residue field. The purpose is to work in the context of Mazur's deformation theory [19], thereby introducing more structure into the problem and enabling us to use tools from commutative algebra. The desired equality (3) is subsumed by Wiles's "Main Conjecture," a precise version of (2) which takes the following form: A certain ring homomorphism

$$R_\Sigma \to \mathbf{T}_\Sigma \tag{4}$$

is an isomorphism, where

- R_Σ is a universal deformation ring which parametrizes representations of type Σ with a fixed residual representation $\bar\rho$;
- $\mathbf{T}_\Sigma$ is a Hecke algebra which parametrizes the newforms of weight two giving rise to such representations.

Wiles's strategy is to prove this first in the case $\Sigma = \emptyset$, and then deduce the result for arbitrary Σ. The aim of this article is to explain the statement of the conjecture and the reduction to the case $\Sigma = \emptyset$.

3. The "Main Conjecture"

3.1. **The Hecke algebra.** Once again, we fix a representation

$$\bar\rho : G_{\mathbf{Q}} \to \mathbf{GL}_2(k)$$

where k is a finite field of characteristic ℓ. We assume:

(a) ℓ is odd;
(b) $\bar\rho$ is irreducible;
(c) the restriction of $\bar\rho$ to a decomposition group at ℓ is finite flat or ordinary;
(d) $\bar\rho$ has cyclotomic determinant;
(e) $\bar\rho$ has square-free conductor.

[1] Use the following ad hoc definition of the exponent of ℓ in the conductor of a representation which is semistable at ℓ and has cyclotomic determinant: it is trivial if the representation is finite flat and 1 otherwise.

Remark 3.1. While hypotheses (a)–(c) are needed for the existing methods, the last two are made to simplify the exposition. Their presence causes no problem in the application to Fermat's Last Theorem, since they are satisfied by the mod ℓ representations coming from semistable elliptic curves. See [8] for a discussion of how to work without them — hypothesis (d) is not so serious; hypothesis (e) was removed in [7].

For a reason fundamental to Wiles's method, we must make the further assumption that $\bar\rho$ is modular. Roughly speaking, this means that $\bar\rho$ is equivalent to the reduction of a representation $\rho_{f,\lambda}$ arising from a modular form. To make this assumption precise, let us fix embeddings $\overline{\mathbf{Q}} \hookrightarrow \overline{\mathbf{Q}}_\ell$, $\overline{\mathbf{Q}} \hookrightarrow \mathbf{C}$. We also fix an embedding of k in $\overline{\mathbf{F}}_\ell$, where $\overline{\mathbf{F}}_\ell$ is the residue field of the ring of integers of $\overline{\mathbf{Q}}_\ell$. Suppose that f is a newform of weight two, level N_f and trivial character, and let K_f denote the number field generated by its coefficients $a_n(f)$. The chosen embeddings determine a prime λ of $\mathcal{O}_f$, the ring of integers of K_f. We write simply ρ_f for $\rho_{f,\lambda}$ (see [22] where this representation is denoted ρ_λ and defined in the course of the proof of theorem 4). Thus ρ_f is the absolutely irreducible representation

$$G_{\mathbf{Q}} \to \mathbf{GL}_2(K_{f,\lambda})$$

characterized up to isomorphism by the following property:

$$\text{If } p \text{ is a prime not dividing } \ell N_f, \text{ then } \rho_f \text{ is unramified at } p, \\ \mathrm{tr}\rho_f(\mathrm{Frob}_p) = a_p(f) \text{ and } \det\rho_f(\mathrm{Frob}_p) = p. \tag{5}$$

One can choose a basis so that the image of ρ_f is contained in $\mathbf{GL}_2(\mathcal{O}_{f,\lambda})$, and the reduction

$$\bar\rho_f : G_{\mathbf{Q}} \to \mathbf{GL}_2(\mathcal{O}_f/\lambda)$$

is well-defined up to semi-simplification.

We assume $\bar\rho$ is equivalent over $\overline{\mathbf{F}}_\ell$ to $\bar\rho_f$ for some f as above. It turns out that if this assumption holds, then in fact there are infinitely many f to choose from. (As we mentioned above, this follows from the results of [20].) Given a finite set of primes Σ, we can then ask which of these f give rise to representations of type Σ, in the sense that $\rho_{f,\lambda}$ is semistable at ℓ and Σ contains the set of primes dividing $N(\rho_{f,\lambda})/N(\bar\rho)$. A sufficient condition is that the level of f divide N_Σ where

$$N_\Sigma = N(\bar\rho)\prod_{p\in\Sigma} p^{m_p}$$

and the m_p are defined as follows:

- $m_p = 2$ if p does not divide $\ell N(\bar\rho)$;
- $m_p = 1$ if $p \neq \ell$ and p divides $N(\bar\rho)$;
- $m_\ell = 1$ if $\bar\rho$ is finite flat and ordinary at ℓ;
- $m_\ell = 0$ otherwise.

The motivation for this definition of N_Σ is that this condition is known to be necessary as well as sufficient, as long as we restrict our attention to forms f with trivial character and level not divisible by ℓ^2. (The proof of the necessity relies on the Deligne-Langlands-Carayol theorem, an analysis of possible values of $N(\rho)/N(\bar\rho)$ which is due independently to Carayol and Livné, and well-known results on the reduction of modular curves, abelian varieties and ℓ-divisible groups. We shall not, however, make use of this fact; indeed it turns out to be a consequence of what follows.)

Let Φ_Σ denote the set of newforms f of weight two, trivial character and level dividing N_Σ. As explained in [12], it follows from the results of [21] and others on Serre's conjecture that the set $\Phi_\emptyset$ is non-empty. The analogous statement holds *a fortiori* for each Φ_Σ. We can then consider the ring

$$\tilde{\mathbf{T}}_\Sigma := \prod_{f\in\Phi_\Sigma} \mathcal{O}_{f,\lambda}.$$

Recall that for each f, the prime λ of K_f is determined by our choices of embeddings and note that $\tilde{\mathbf{T}}_\Sigma$ is semilocal and finitely generated as a $\mathbf{Z}_\ell$-module. For each prime p not in Σ, we let T_p denote the element $(a_p(f))_{f\in\Phi_\Sigma}$. We define the Hecke algebra $\mathbf{T}_\Sigma$ as the $\mathbf{Z}_\ell$-subalgebra of $\tilde{\mathbf{T}}_\Sigma$ generated by the elements T_p for p not in $\Sigma\cup\{\ell\}$.

We can give another description of $\mathbf{T}_\Sigma$ in terms of the subring $\mathbf{T}$ of the ring of endomorphisms of $S = S_2(\Gamma_0(N_\Sigma))$ generated by the operators T_p for all primes p. We suppose that f is in Φ_Σ and we define f_Σ as a certain $\mathbf{T}$-eigenform in S for which f is the associated newform. The eigenform f_Σ is characterized by this together with the properties:

- if p is in $\Sigma\setminus\{\ell\}$, then $a_p(f_\Sigma) = 0$;
- if ℓ divides N_Σ, then $a_\ell(f_\Sigma)$ is an ℓ-adic unit.

The map sending T_p to the reduction of $a_p(f_\Sigma)$ defines a homomorphism $\mathbf{T}\to\overline{\mathbf{F}}_\ell$, and we write $\mathbf{T}_\mathfrak{m}$ for the completion of $\mathbf{T}$ at the kernel $\mathfrak{m}$ of this homomorphism. We then have the lemma:

Lemma 3.2. *If ℓ is not in Σ, then the element T_ℓ of $\tilde{\mathbf{T}}_\Sigma$ is in $\mathbf{T}_\Sigma$. For arbitrary Σ there is an isomorphism $\mathbf{T}_\Sigma \xrightarrow{\sim} \mathbf{T}_\mathfrak{m}$ such that $T_p\mapsto T_p$ for all p not in Σ.*

We refer the reader to section 2.3 of [26] or section 4.2 of [2] for the verification, which is tedious, unilluminating, and ultimately unnecessary (see the discussion in section 5 below).

3.2. **The universal deformation ring.** We appeal to Mazur's deformation theory of Galois representations, discussed in [4] and [19], to define a certain universal deformation ring R_Σ.

Keep the technical hypotheses imposed on $\bar\rho$ at the beginning of the preceding section, and consider deformations of $\bar\rho$ of the form

$$\rho : G_{\mathbf{Q}} \to \mathbf{GL}_2(A),$$

where A is a complete local Noetherian ring with residue field k. We say such a deformation is of type Σ if the following statements are true:

(a) $\det\rho$ is cyclotomic.

(b) ρ is finite flat or ordinary at ℓ.

(c) let p be a prime not in Σ. Then:
- if $p \neq \ell$ and $\bar{\rho}$ is unramified at p, then so is ρ;
- if $p \neq \ell$ and $\bar{\rho}$ is of type A at p, then so is ρ;
- if $p = \ell$ and $\bar{\rho}$ is finite flat, then so is ρ.

Remark 3.3. Suppose that A is the ring of integers of a finite extension of $\mathbf{Q}_\ell$ and the first two conditions are satisfied. In that case, one can check that condition (c) is equivalent to the equality $\operatorname{ord}_p N(\rho) = \operatorname{ord}_p N(\bar{\rho})$. We shall not use this fact.

The results discussed in [4] and [19] furnish a universal deformation ring R_Σ and a universal deformation

$$\rho_\Sigma^{\text{univ}} : G_{\mathbf{Q}} \to \mathbf{GL}_2(R_\Sigma)$$

of $\bar{\rho}$ of type Σ.

Suppose now that $\bar{\rho}$ is modular. For a newform f in Φ_Σ, we let A_f denote the subring of $\mathcal{O}_{f,\lambda}$ consisting of those elements whose reduction mod λ is in k. One checks that with respect to some basis, we have

$$\rho_f : G_{\mathbf{Q}} \to \mathbf{GL}_2(A_f),$$

a deformation of $\bar{\rho}$ of type Σ. The universal property of R_Σ therefore furnishes a unique homomorphism $\pi_{f,\Sigma} : R_\Sigma \to A_f$ such that the composite

$$G_{\mathbf{Q}} \to \mathbf{GL}_2(R_\Sigma) \to \mathbf{GL}_2(A_f)$$

is equivalent to ρ_f. Since R_Σ is topologically generated by the traces of $\rho_\Sigma^{\text{univ}}(\text{Frob}_p)$ for p not in $\Sigma \cup \{\ell\}$, we conclude that the image of the map

$$\begin{array}{ccc} R_\Sigma & \longrightarrow & \tilde{\mathbf{T}}_\Sigma \\ r & \longmapsto & (\pi_{f,\Sigma}(r))_{f \in \Phi_\Sigma} \end{array}$$

is precisely $\mathbf{T}_\Sigma$. We define ϕ_Σ to be the resulting surjective ring homomorphism $R_\Sigma \to \mathbf{T}_\Sigma$.

3.3. **Statement of the conjecture.** We keep the hypotheses on $\bar{\rho}$ imposed in section 3.1 (the hypothesis of modularity as well as the technical conditions listed at the beginning of the section). We suppose that Σ is a finite set of primes and we consider the map ϕ_Σ defined at the end of the section 3.2. In our setting, Wiles's conjecture 2.16 in [26] becomes:

Conjecture 3.4. *The map ϕ_Σ is an isomorphism.*

We briefly recall how the conjecture implies the Shimura-Taniyama Conjecture for semistable elliptic curves (see also [3]).

Theorem 3.5. *Suppose conjecture 3.4 holds and E is an elliptic curve with square-free conductor N_E. If there is an odd prime ℓ such that $\bar{\rho}_{E,\ell}$ is irreducible and modular, then E is modular.*

To prove this, one checks that under these hypotheses, the mod ℓ representation $\bar{\rho}_{E,\ell}$ satisfies the technical conditions of §3.1. Moreover the ℓ-adic representation $\rho_{E,\ell}$ is a deformation of $\bar{\rho}_{E,\ell}$ of type Σ for some Σ; for example, one can take Σ to be the set of primes dividing $N_E\ell$. On taking $\mathcal{O} = \mathbf{Z}_\ell$, we obtain from the universal property of R_Σ a homomorphism $\theta : R_\Sigma \to \mathbf{Z}_\ell$ where

$$\mathrm{tr}(\rho_\Sigma^{\mathrm{univ}}(\mathrm{Frob}_p)) \quad \longmapsto \quad a_p(E) = p+1 - \#E(\mathbf{F}_p)$$

for $p \neq \ell$ not in Σ. Now the conjecture implies ϕ_Σ is an isomorphism, so we may consider the composite

$$\mathbf{T}_\Sigma \xrightarrow{\phi_\Sigma^{-1}} R_\Sigma \xrightarrow{\theta} \mathbf{Z}_\ell.$$

One sees from the definition of $\mathbf{T}_\Sigma$ that such a homomorphism is necessarily a projection $\pi_{f,\Sigma}$ for some f in Φ_Σ. It follows that $\rho_{E,\ell}$ is isomorphic to ρ_f, or equivalently that

$$a_p(f) = a_p(E) \quad \text{for all } p \notin \Sigma.$$

Therefore E is modular.

If $\bar{\rho}_{E,3}$ is irreducible, then the Langlands-Tunnell theorem [16] shows that $\bar{\rho}_{E,3}$ is modular, hence conjecture 3.4 implies that E is modular. If $\bar{\rho}_{E,3}$ is reducible, then Wiles argues (see [23]) that $\bar{\rho}_{E,5}$ is isomorphic to $\bar{\rho}_{E',5}$ for some semistable E' with irreducible $\bar{\rho}_{E',3}$. Since E' is modular, so is $\bar{\rho}_{E',5} \approx \bar{\rho}_{E,5}$, so we may apply the preceding theorem with $\ell = 5$.

4. Reduction to the case $\Sigma = \emptyset$

4.1. **Commutative algebra.** Suppose now that $f = \sum a_n q^n$ is a newform in $\Phi_\emptyset$ (recall from [12] that such an f exists), hence in Φ_Σ for every finite set of primes Σ. Consider the commutative triangle

$$\begin{tikzcd} R_\Sigma \arrow[r, "\phi_\Sigma"] \arrow[dr] & \mathbf{T}_\Sigma \arrow[d] \\ & A_f \end{tikzcd}$$

where the downwards arrow is $\pi_{f,\Sigma}$ and the diagonal one corresponds to ρ_f via the universal property of R_Σ.

In order to apply the commutative algebra results explained in [5], it will be convenient to work with $\mathcal{O}$-algebras where $\mathcal{O} = \mathcal{O}_{f,\lambda}$. Note that ϕ_Σ is an isomorphism if and only if

$$\phi_\Sigma \otimes_{W(k)} \mathcal{O} : R_\Sigma \otimes_{W(k)} \mathcal{O} \to \mathbf{T}_\Sigma \otimes_{W(k)} \mathcal{O}$$

is an isomorphism. From now on, we replace ϕ_Σ, R_Σ and $\mathbf{T}_\Sigma$ by their tensor products over $W(k)$ with $\mathcal{O}$. The resulting representation $G_\mathbf{Q} \to \mathbf{GL}_2(R_\Sigma)$ is universal for type-Σ deformations of $\bar{\rho}$ as in (1), where now A is a complete local Noetherian $\mathcal{O}$-algebra with residue field $\mathcal{O}/\lambda$. Writing $\mathcal{O}'_g$ for the $\mathcal{O}$-subalgebra of $\overline{\mathbf{Q}}_\ell$ generated by the Fourier coefficients of f, we may identify

$\mathbf{T}_\Sigma$ with the $\mathcal{O}$-subalgebra of

$$\prod_{g\in\Phi_\Sigma}\mathcal{O}'_g$$

generated by the operators T_p for p not in $\Sigma\cup\{\ell\}$. Our commutative triangle becomes

$$\begin{tikzcd} R_\Sigma \arrow[r, "\phi_\Sigma"] \arrow[dr] & \mathbf{T}_\Sigma \arrow[d] \\ & \mathcal{O} \end{tikzcd}.$$

Write π_Σ for the map $\mathbf{T}_\Sigma\to\mathcal{O}$, let $\mathfrak{p}_\Sigma$ denote the kernel of $R_\Sigma\to\mathcal{O}$, and let η_Σ denote the ideal $\pi_\Sigma(\mathrm{Ann}_{\mathbf{T}_\Sigma}\ker\pi_\Sigma)$ (which is non-zero). According to Criterion I of [5], we have

$$\text{length}_{\mathcal{O}}(\mathfrak{p}_\Sigma/\mathfrak{p}_\Sigma^2)\geq\text{length}_{\mathcal{O}}(\mathcal{O}/\eta_\Sigma) \tag{6}$$

and the following are equivalent:

- ϕ_Σ is an isomorphism between complete intersections over $\mathcal{O}$.
- Equality holds in (6).

It is actually the following strengthening of conjecture 3.4 whose proof we reduce to the case of $\Sigma=\emptyset$.

Conjecture 4.1. *The surjection ϕ_Σ is an isomorphism between complete intersections over $\mathcal{O}$.*

In the remaining sections we sketch the proof that if

$$\text{length}_{\mathcal{O}}(\mathfrak{p}_\Sigma/\mathfrak{p}_\Sigma^2)=\text{length}_{\mathcal{O}}(\mathcal{O}/\eta_\Sigma),$$

then

$$\text{length}_{\mathcal{O}}(\mathfrak{p}_{\Sigma'}/\mathfrak{p}_{\Sigma'}^2)\leq\text{length}_{\mathcal{O}}(\mathcal{O}/\eta_{\Sigma'})$$

where $\Sigma'=\Sigma\cup\{p\}$. This implies:

Theorem 4.2. *If conjecture 4.1 holds in the case $\Sigma=\emptyset$, then it holds for all Σ.*

The proof of conjecture 4.1 in the case $\Sigma=\emptyset$ is explained in [3].

4.2. **Selmer groups.** Recall from Chapter VI of [19] (see also §2.7 of [2]) that the $\mathcal{O}$-module $\mathfrak{p}_\Sigma/\mathfrak{p}_\Sigma^2$ has a natural description in terms of Galois cohomology. Write M_f for $\mathcal{O}^2$ with Galois action defined by ρ_f and E_f for ad^0M_f, the $\mathcal{O}[G_{\mathbf{Q}}]$-module of trace-zero $\mathcal{O}$-endomorphisms of M_f. Let $E_{f,n}$ denote $E_f\otimes_{\mathcal{O}}\lambda^{-n}\mathcal{O}/\mathcal{O}$ and set

$$E_{f,\infty}=\varinjlim_n E_{f,n}\cong E_f\otimes_{\mathcal{O}}K/\mathcal{O}\cong E_f\otimes_{\mathbf{Z}_\ell}\mathbf{Q}_\ell/\mathbf{Z}_\ell.$$

According to §28 of [19] (or more precisely its analogue in the case of fixed determinant; cf. Proposition 3 of §26), we have a canonical isomorphism

$$\mathrm{Hom}(\mathfrak{p}_\Sigma/\mathfrak{p}_\Sigma^2,K/\mathcal{O})\cong H^1_{\mathcal{D}}(G_{\mathbf{Q},\Sigma\cup\{\ell\}},E_{f,\infty})$$

where $\mathcal{D}$ is the type-Σ deformation condition. Appealing to the descriptions in §§29–31 of the resulting conditions on local cohomology classes, we find that the latter $\mathcal{O}$-module is the "generalized Selmer group"

$$\varinjlim_{n} H^1_{\mathcal{L}}(\mathbf{Q}, E_{f,n})$$

in the notation of §6 of [25] (with the appropriate choice of local condition at ℓ from §7).

The analogous statements hold with Σ replaced by Σ' (in which case we write $\mathcal{D}'$ and $\mathcal{L}'$), so to compare the lengths of $\mathfrak{p}_\Sigma/\mathfrak{p}_\Sigma^2$ and $\mathfrak{p}_{\Sigma'}/\mathfrak{p}_{\Sigma'}^2$ we consider the cokernel of the natural inclusion

$$H^1_{\mathcal{D}}(G_{\mathbf{Q},\Sigma\cup\{\ell\}}, E_{f,\infty}) \to H^1_{\mathcal{D}'}(G_{\mathbf{Q},\Sigma\cup\{\ell\}}, E_{f,\infty}).$$

Suppose now that p does not divide $\ell N(\bar{\rho})$. From the definitions of the generalized Selmer groups, we see that our cokernel embeds naturally in

$$H^1(I_p, E_{f,\infty})^{G_p/I_p}.$$

Since the action of I_p on $E_{f,\infty}$ is trivial, the cohomology group can be identified with

$$\mathrm{Hom}(I_p, E_{f,\infty}) \cong \mathrm{Hom}(\mathbf{Z}_\ell(1), E_{f,\infty})$$

as a module for G_p/I_p. We are therefore reduced to computing the length of

$$H^0\left(G_p/I_p, E_{f,\infty}(-1)\right).$$

This is just the kernel of the endomorphism $1 - \mathrm{Frob}_p^{-1}$ of

$$E_f(-1) \otimes_{\mathcal{O}} K/\mathcal{O}.$$

This kernel is finite if and only if $1 - \mathrm{Frob}_p^{-1}$ defines an automorphism of the K-vector space $E_f(-1) \otimes_{\mathcal{O}} K$, in which case its $\mathcal{O}$-length is simply the valuation of the determinant. We compute this determinant using the fact that the characteristic polynomial of Frob_p on $M_f \otimes_{\mathcal{O}} K$ is $X^2 - a_pX + p$ (see (5)). The result is that our determinant is

$$(1 - p)\left((1 + p)^2 - a_p^2\right) \tag{7}$$

(non-zero by [22], theorem 5), so we conclude that

$$\mathrm{length}_{\mathcal{O}}(\mathfrak{p}_{\Sigma'}/\mathfrak{p}_{\Sigma'}^2) \leq \mathrm{length}_{\mathcal{O}}(\mathfrak{p}_\Sigma/\mathfrak{p}_\Sigma^2) + v_\lambda(c_p) \tag{8}$$

where c_p is given by (7).

The cohomological calculation above is a special case of the following general result, whose proof is left as an exercise.

Proposition 4.3. *Suppose that $p \neq \ell$ and X a finitely generated free $\mathcal{O}$-module with a continuous action of G_p. Let $X_\infty = X \otimes_{\mathcal{O}} K/\mathcal{O}$. Then the Fitting ideal of the $\mathcal{O}$-module $H^1(I_p, X_\infty)^{D_p/I_p}$ is generated by the determinant of the endomorphism $1 - \mathrm{Frob}_p^{-1}p$ of $(X \otimes_{\mathcal{O}} K)_{I_p}$.*

Applying it in the case that $p \neq \ell$, but p divides $N(\bar{\rho})$, we find that the space $(E_f \otimes_{\mathcal{O}} K)_{I_p}$ is one-dimensional, Frob_p acts by the inverse of the cyclotomic character, and (8) holds with $c_p = 1 - p^2$.

Finally, in the case $p = \ell$, the groups H^1_Σ and $H^1_{\Sigma'}$ coincide unless $\bar{\rho}$ is ordinary and finite flat. In that case, bounding the cokernel is more subtle and one uses the calculations in [25] to show that the length increases by at most the valuation of $1 - \alpha_\ell^2$, where α_ℓ is the unit root of $X^2 - a_\ell X + \ell = 0$.

Summing up, we have

Lemma 4.4. *In all of the cases above,*

$$\mathrm{length}_{\mathcal{O}}(\mathfrak{p}_{\Sigma'}/\mathfrak{p}_{\Sigma'}^2) \leq \mathrm{length}_{\mathcal{O}}(\mathfrak{p}_\Sigma/\mathfrak{p}_\Sigma^2) + v_\lambda(c_p)$$

where

$$c_p = \begin{cases} (1-p)\left((1+p)^2 - a_p^2\right) & \text{if } p \nmid N(\bar{\rho}) \\ 1 - p^2 & \text{otherwise.} \end{cases}$$

4.3. **Congruence modules.** Now we have to prove the inequality

$$\mathrm{length}_{\mathcal{O}}(\mathcal{O}/\eta_{\Sigma'}) \geq \mathrm{length}_{\mathcal{O}}(\mathcal{O}/\eta_\Sigma) + v_\lambda(c_p),$$

or equivalently:

$$\eta_{\Sigma'} \subset c_p \eta_\Sigma, \tag{9}$$

where c_p is defined above. Before sketching the proof, we describe the general strategy and consider an example.

The first observation to make is that the ideal η_Σ measures congruences between f and other forms of level N_Σ. Suppose for simplicity that $\mathcal{O}$ contains the coefficients of all newforms in Φ_Σ. Then $\mathbf{T}_\Sigma$ is contained in a product of copies of $\mathcal{O}$ indexed by these newforms and π_Σ is the projection onto the coordinate corresponding to f. The ideal η_Σ consists of those x such that

$$(x, 0, 0, \ldots, 0) \in \mathbf{T}_\Sigma,$$

where the first coordinate corresponds to f. If there are just two forms, f and g, in Φ_Σ, then η_Σ is the ideal generated by $a_p(f) - a_p(g)$ for p not dividing $N_\Sigma \ell$. So η_Σ measures how congruent are the coefficients of f and g at "good" primes. More generally, one finds that $\eta_\Sigma \subset \lambda^n$ if $f \equiv g \bmod \lambda^n$ (in the sense above) for some g which is a linear combination over $\mathcal{O}$ of the forms different from f.

Consider for example the unique newform f of weight 2 and level 11. Its Fourier coefficients are rational, and the associated L-function is that of (the isogeny class of) the elliptic curve E defined by

$$Y^2 + Y = X^3 - X.$$

The 3-adic representation attached to f is equivalent to

$$\rho_{E,3} : G_{\mathbf{Q}} \to \mathrm{Aut}(\mathcal{T}_3(E)) \cong \mathbf{GL}_2(\mathbf{Z}_3),$$

and we let $\bar{\rho}$ denote the reduction. (We leave it to the reader to verify that $\bar{\rho}$ satisfies our running hypotheses.) Since $\Phi_\emptyset$ is the singleton set $\{f\}$ (as

$N_\emptyset = 11$), we have $\eta_\emptyset = \mathcal{O}$. Take $\Sigma = \{3\}$. Then $N_\Sigma = 33$, and Φ_Σ consists of f and the unique newform g of (weight 2, trivial character and) level 33. To prove this statement, one needs to check that f and g are congruent mod 3; for this, it suffices to compare the Galois action on the 3-division points of the corresponding elliptic curves. Therefore, $\eta_\Sigma = 3\mathcal{O}$ is indeed generated by c_3.

Now consider the problem of comparing η_Σ and $\eta_{\Sigma'} = \eta_\Sigma \cup \{p\}$ when $p = 7$. Then $N_{\Sigma'} = 1617$ and c_p is 9 times a unit. So we could verify this case of the desired inclusion by finding congruent newforms among the levels 77, 231, 539 and 1617, and then writing down a linear combination of g with these forms which is congruent to f mod $27\mathcal{O}$.

This type of problem ("raising the level") was first addressed in [20], where the general strategy is as follows: Rather than produce such congruences directly, one detects them using the cohomology of modular curves[2], or in our case, their Jacobians. The problem of comparing these "cohomological congruences" at different levels is then reduced to studying the kernel of a certain homomorphism of Jacobians induced by degeneracy maps on the curves. This last issue is then resolved by a result of Ihara.

The method and results of [20] were sharpened and generalized in various ways in such articles as [6], [11], [26] and [7]. We now sketch Wiles's proof ([26], section 2.2) of (9), with some modifications taken from [2].

Let $\mathfrak{m}$ denote the maximal ideal of $\mathbf{T} \otimes \mathcal{O}$ containing the kernel of the homomorphism to $\mathcal{O}$ determined by f_Σ. Define M_Σ as the localization at $\mathfrak{m}$ of $\mathcal{T}_\ell(J_0(N)) \otimes_{\mathbf{Z}_\ell} \mathcal{O}$. One deduces from lemma 3.2 that the $\mathcal{O}$-algebra $\mathbf{T}_\Sigma$ is isomorphic to $(\mathbf{T} \otimes \mathcal{O})_\mathfrak{m}$, so we may regard M_Σ as a module for $\mathbf{T}_\Sigma$, hence R_Σ. Recall from [24] that Wiles proves the following generalization of results of Mazur and others:

Theorem 4.5. *The $\mathbf{T}_\Sigma$-module M_Σ is free of rank two.*

The module M_Σ is equipped with an alternating pairing, $\langle\, , \rangle_\Sigma$ that induces an isomorphism

$$M_\Sigma \cong \mathrm{Hom}_\mathcal{O}(M_\Sigma, \mathcal{O})$$

of $\mathbf{T}_\Sigma$-modules. Since M_Σ is free of rank two over $\mathbf{T}_\Sigma$, the submodule $M_\Sigma[\mathfrak{p}_\Sigma]$ (the set of elements annihilated by every t in $\mathfrak{p}_\Sigma$) is free of rank two over $\mathbf{T}_\Sigma/\mathfrak{p}_\Sigma = \mathcal{O}$. On combining these facts, one shows easily that if $\{x, y\}$ is a basis for $M_\Sigma[\mathfrak{p}_\Sigma]$, then

$$\eta_\Sigma = \langle x, y\rangle_\Sigma.$$

(See [2] for the details.)

To compare η_Σ and $\eta_{\Sigma'}$, one defines a $\mathbf{T}_{\Sigma'}$-equivariant map

$$M_{\Sigma'} \to M_\Sigma.$$

[2] This approach is suggested by work of Hida [17], which also establishes a relation between congruences and certain values of L-functions; see also section 4.4 of [2] and [9].

Its definition employs the degeneracy maps $X_0(N_{\Sigma'}) \to X_0(N_\Sigma)$ induced by $\tau \mapsto p^i\tau$ for $i \le m_p$. These induce by Albanese functoriality maps $J_0(N_{\Sigma'}) \to J_0(N_\Sigma)$, hence maps

$$\mathcal{T}_\ell(J_0(N_{\Sigma'})) \otimes_{\mathbf{Z}_\ell} \mathcal{O} \to \mathcal{T}_\ell(J_0(N_\Sigma)) \otimes_{\mathbf{Z}_\ell} \mathcal{O} \to M_\Sigma$$

which we denote δ_i. A suitable $\mathbf{T}_\Sigma$-linear combination of these, namely

- $\delta_0 - p^{-1}T_p\delta_1 + p^{-1}\delta_2$ if $p \ne \ell$ and $m_p = 2$;
- $\delta_0 - p^{-1}T_p\delta_1$ if $p \ne \ell$ and $m_p = 1$;
- $\delta_0 - \tau_\ell^{-1}\delta_1$ if $p = \ell$ and $m_\ell = 1$, where τ_ℓ is the unit root in $\mathbf{T}_\Sigma$ of
$$X^2 - T_\ell X + \ell = 0;$$
- δ_0 if $p = \ell$ and $m_\ell = 0$;

induces the desired homomorphism

$$\beta : M_{\Sigma'} \to M_\Sigma$$

of $\mathbf{T}_\Sigma$-modules (see p. 119 of [2]).

Write β' for the adjoint of β with respect to the pairings $\langle\,,\,\rangle_\Sigma$ and $\langle\,,\,\rangle_{\Sigma'}$. A straightforward computation shows that the composite $\beta\beta'$ is an endomorphism in $\mathbf{T}_\Sigma$ which is a unit times

- $(1-p)\left((1+p)^2 - T_p^2\right)$ if $p \nmid N(\bar\rho)$,
- $1 - p^2$ otherwise

(see p. 121 of [2]). Note that this operator is c_p mod $\mathfrak{p}_\Sigma$. (Moreover this holds with $\mathfrak{p}_\Sigma$ replaced by any minimal prime of $\mathbf{T}_\Sigma$ and c_p replaced by its analogue defined using the corresponding newform. Since these are nonzero, $\beta\beta'$, and hence β', is injective.) To obtain (9), it suffices to prove that β' has torsion-free cokernel, for then a basis $\{x, y\}$ for M_Σ yields a basis $\{\beta'(x), \beta'(y)\}$ for $M_{\Sigma'}$ and we conclude that

$$\eta_{\Sigma'} = (\langle\beta'(x), \beta'(y)\rangle_{\Sigma'}) = c_p(\langle x, y\rangle)_\Sigma = c_p\eta_\Sigma.$$

Since β' is injective, it has torsion-free cokernel if and only if β is surjective (or equivalently, β' mod λ is injective). Recall that β is defined using the maps on homology (or Tate modules) induced by the degeneracy maps $X_0(N_{\Sigma'}) \to X_0(N_\Sigma)$. We therefore wish to analyze the cokernel of the homomorphism

$$H_1(X_0(N'_\Sigma), \mathcal{O}) \to H_1(X_0(N_\Sigma), \mathcal{O})^{m_p+1}$$

gotten from the degeneracy maps. This map is not surjective in general, but it is enough to prove:

Lemma 4.6. *Suppose that $N_\Sigma > 3p$ if p divides N_Σ. Then the map*

$$H_1(X_0(N'_\Sigma), \mathbf{Z}_\ell) \to H_1(X_0(N_\Sigma), \mathbf{Z}_\ell)^{m_p+1}_{\mathfrak{m}'}$$

is surjective, where $\mathfrak{m}'$ is the intersection of $\mathfrak{m}$ with the subalgebra of $\mathbf{T} \otimes \mathbf{Z}_\ell$ generated by the operators T_r for primes $r \ne p$.

We sketch the proof of the lemma in the generic (and most difficult) case of $m_p = 2$ and then explain what changes are needed to treat the remaining cases.

First recall that a similar problem is solved in [20]. Let $X_1(N_\Sigma, p)$ be the modular curve associated to $\Gamma_1(N_\Sigma) \cap \Gamma_0(p)$ and consider the map

$$(\pi_{1,*}, \pi_{2,*}) : H_1(X_1(N_\Sigma, p), \mathbf{Z}_\ell) \to H_1(X_1(N_\Sigma), \mathbf{Z}_\ell)^2, \tag{10}$$

where here and below, $\pi_{1,*}$ (resp. $\pi_{2,*}$) will denote the map induced by $\tau \mapsto \tau$ (resp. $\tau \mapsto p\tau$). In section 4 of [20], the surjectivity of this map is proved by a group-theoretic argument using results of Ihara. The surjectivity of (10) is a key ingredient in the proof of the lemma.

Next consider the sequence of homology groups of non-compact modular curves

(11)
$$\begin{array}{ccccc} H_1(Y_1(N_\Sigma p, p^2), \mathbf{Z}_\ell) & \longrightarrow & H_1(Y_1(N_\Sigma p), \mathbf{Z}_\ell)^2 & \longrightarrow & H_1(Y_1(N_\Sigma), \mathbf{Z}_\ell) \\ x & \longmapsto & (\pi_{1,*}x, \pi_{2,*}x); \quad (y, z) & \longmapsto & \pi_{2,*}y - \pi_{1,*}z. \end{array}$$

Wiles proves the exactness of this sequence by an elementary group-theoretic argument (see lemma 2.5 of [26]). If we could replace X_1 and Y_1 with X_0 in (10) and (11), we could now deduce that

$$H_1(X_0(N_\Sigma p^2), \mathbf{Z}_\ell) \to H_1(X_0(N_\Sigma), \mathbf{Z}_\ell)^3$$

is surjective. A minor complication arises when we try to make this change, and one finds instead that the cokernel is supported only at "Eisenstein" maximal ideals of the Hecke algebra. (See the last half of section 4.5 of [2] for more details). The irreducibility hypothesis on $\bar{\rho}$ ensures that $\mathfrak{m}'$ is not Eisenstein, from which we deduce the lemma.

Suppose now that $m_p = 1$. If $p \neq \ell$ (in which case p divides N_Σ), then one uses the exactness of (11) with N_Σ replaced by N_Σ/p. If $p = \ell$, then one just uses the surjectivity of (10). In either case, the lemma follows as above since $\mathfrak{m}'$ is not Eisenstein, using also that $\mathfrak{m}'$ is not in the support of $H_1(X_0(N_\Sigma/p), \mathbf{Z}_\ell)$ in the case $p \neq \ell$. There is nothing to prove if $m_p = 0$.

5. Epilogue

Some parts of the exposition above, especially the last section, draw from [2]. As explained there, theorem 4.5 is actually only used in the case $\Sigma = \emptyset$.

In a recent article [10], the first author has presented a modification of the method of Taylor-Wiles-Faltings which makes no appeal to lemma 3.2 and theorem 4.5.[3] Instead, in the modified approach, one deduces these two results as *by-products* of the proof of the Main Conjecture. The new idea is to use tools from commutative algebra to prove directly, without any initial reference to the Hecke algebra $\mathbf{T}_\emptyset$, that $M_\emptyset$ is free over $R_\emptyset$.

[3] A similar method was found independently by Fujiwara [15].

References

[1] H. Carayol, *Sur les représentations galoisiennes modulo ℓ attachées aux formes modulaires,* Duke Math. J. **59** (1989), 785–801.

[2] H. Darmon, F. Diamond, R. Taylor, *Fermat's Last Theorem,* in Current Developments in Mathematics, 1995, International Press, 1–154.

[3] E. de Shalit, *Hecke rings and universal deformation rings,* this volume.

[4] B. de Smit, H. Lenstra, *Explicit construction of universal deformation rings,* this volume.

[5] B. de Smit, K. Rubin, R. Schoof, *Criteria for complete intersections,* this volume.

[6] F. Diamond, *On congruence modules associated to Λ-adic forms,* Comp. Math. **71** (1989), 49–83.

[7] F. Diamond, *On deformation rings and Hecke rings,* Annals of Math. **144** (1996), 137–166.

[8] F. Diamond, *An extension of Wiles' results,* this volume.

[9] F. Diamond, *Congruences between modular forms: Raising the level and dropping Euler factors,* to appear in Proc. NAS.

[10] F. Diamond, *The Taylor-Wiles construction and multiplicity one,* to appear in Invent. Math.

[11] F. Diamond, R. Taylor, *Non-optimal levels of mod ℓ modular representations,* Invent. Math. **115** (1994) 435–462.

[12] B. Edixhoven, *Serre's conjecture,* this volume.

[13] J.-M. Fontaine, B. Mazur, *Geometric Galois representations,* in: Elliptic Curves, Modular Forms and Fermat's Last Theorem, International Press, Cambridge (1995), 41–78.

[14] G. Frey, *On ternary equations of Fermat type and relations with elliptic curves,* this volume.

[15] K. Fujiwara, *Deformation rings and Hecke algebras in the totally real case,* preprint.

[16] S. Gelbart, Three lectures on the modularity of $\bar{\rho}_{E,3}$ and the Langlands reciprocity conjecture, this volume.

[17] H. Hida, *Congruences of cusp forms and special values of their zeta functions,* Inv. Math. **63** (1981), 225–261.

[18] R. Livné, *On the conductors of mod ℓ representations coming from modular forms*, J. Number Theory **31** (1989), 133–141.

[19] B. Mazur, *An introduction to the deformation theory of Galois representations,* this volume.

[20] K. Ribet, *Congruence relations between modular forms,* Proc. ICM, 1983, 503–514.

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

[22] D. Rohrlich, *Modular functions and modular curves,* this volume.

[23] K. Rubin, *Modularity of mod 5 representations,* this volume.

[24] J. Tilouine, *Hecke algebras and the Gorenstein property,* this volume.

[25] L. Washington, *Galois cohomology,* this volume.

[26] A. Wiles, *Modular elliptic curves and Fermat's Last Theorem,* Annals of Math. **141** (1995), 443–551.

Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139

E-mail address: fdiamond@math.mit.edu

Department of Mathematics 3840, University of California, Berkeley, CA 94720-3840

E-mail address: ribet@math.berkeley.edu