Serre's conjecture.

Bas Edixhoven

April 1996

The aim of the first section is to state Serre's conjecture and to tell what is presently known about it, without proof. We start by recalling what modular forms are. Then we recall the result, due to Deligne, that to a mod p modular form one can associate a mod p Galois representation. After that we state Serre's conjecture and what we know about it. In Section 2 we will see which cases of it are actually needed in order to prove, following Wiles, that all semi-stable elliptic curves over $\mathbb{Q}$ are modular. In the last two sections we will sketch the proofs in those cases.

1 Serre's conjecture: statement and results

Let $N \geq 1$ and k be integers and let R be a $\mathbb{Z}[1/N]$-algebra. We will first recall Katz's definition [22] of modular forms of level N and weight k over R. This definition may seem more complicated than necessary, but it is very convenient in order to deal with modular forms over fields of positive characteristic p. For example, one has the Hasse invariant (see [24, §12.4]), which cannot be lifted to characteristic zero as a form of level 1 and weight $p-1$ for p equal to 2 or 3, and one has the derivation $q\,d/dq$ sending modular forms to modular forms, increasing the weight by $p+1$ (p being the characteristic of the finite field). A good reference for more details concerning modular forms in various settings is [14].

Let $[\Gamma_1(N)]_R$ denote the category whose objects are pairs $(E/S/R, \alpha)$, with S an R-scheme, E/S a generalized elliptic curve in the sense of [10] and $\alpha\colon (\mathbb{Z}/N\mathbb{Z})_S \to E[N]$ an embedding of group schemes such that the image of α meets all irreducible components of all geometric fibres of E/S. The morphisms from $(E'/S'/R, \alpha')$ to $(E/S/R, \alpha)$ in $[\Gamma_1(N)]_R$ are Cartesian diagrams:

$$(1.1)\qquad \begin{array}{ccc} E' & \to & E \\ \downarrow & & \downarrow \\ S' & \to & S \end{array}$$

which are compatible with α and α'. For a generalized elliptic curve E over a scheme S we have the invertible $\mathcal{O}_S$-module $\underline{\omega}_{E/S} := 0^*\Omega^1_{E/S}$, obtained by pulling back the sheaf of Kähler differentials of E over S by the zero section 0 in $E(S)$. A modular form f of level N and weight k over R is then a rule, that assigns to every object $(E/S/R, \alpha)$ of $[\Gamma_1(N)]_R$ an element

$f(E/S/R,\alpha)$ of $\underline{\omega}_{E/S}^{\otimes k}(S)$, compatible with morphisms in $[\Gamma_1(N)]_R$. The R-module of such modular forms will be denoted by $M(N,k)_R$. Let us give one example: for p a prime, the Hasse invariant is an element of $M(1,p-1)_{\mathbb{F}_p}$.

There are more down to earth ways to describe $M(N,k)_R$. For $N\geq 5$ the category $[\Gamma_1(N)]_R$ has a final object, called $(E_{\mathrm{univ}}/X_1(N)_R,\alpha_{\mathrm{univ}})$, and in that case one has simply $M(N,k) = \mathrm{H}^0(X_1(N)_R,\underline{\omega}^{\otimes k})$. Recall that $X_1(N)_R$ is a smooth projective curve over R whose fibres are geometrically irreducible. For $N\geq 1$ and $n\geq 3$ invertible in R one has the description:

$$M(N,k)_R = \mathrm{H}^0(\overline{\mathcal{M}}([\Gamma_1(N),\Gamma(n)]_R),\underline{\omega}^{\otimes k})^G, \tag{1.2}$$

where $\overline{\mathcal{M}}([\Gamma_1(N),\Gamma(n)]_R)$ is the moduli scheme parametrizing triples $(E/S/R,\alpha,\beta)$ of generalized elliptic curves with $\alpha\colon(\mathbb{Z}/N\mathbb{Z})_S\to E[N]$ and $\beta\colon(\mathbb{Z}/n\mathbb{Z})_S^2\to E[n]$ embeddings of group schemes such that the image of $\alpha+\beta$ meets all irreducible components of all geometric fibres, and where G is the group $\mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z})$. This $\overline{\mathcal{M}}([\Gamma_1(N),\Gamma(n)]_R)$ is a smooth projective curve over R, but its fibres are not geometrically irreducible.

In the special case $R=\mathbb{C}$ one easily shows that $M(N,k)_{\mathbb{C}}$ is naturally isomorphic to the space of modular forms defined as certain holomorphic functions on the upper half plane $\mathbb{H}$ (see Rohrlich's chapter in this book).

One can show, for example by using the modular form Δ in $M(1,12)_R$, that $M(N,k)_R=0$ for all $k<0$.

Let $R\to R'$ be a morphism of $\mathbb{Z}[1/N]$-algebras. Then we have a morphism of R'-modules $M(N,k)_R\otimes_R R'\to M(N,k)_{R'}$. Such a morphism is not always an isomorphism (consider for example $\mathbb{Z}\to\mathbb{F}_2$, $N=1$ and $k=1$). However, it is an isomorphism when $R\to R'$ is flat (use (1.2)).

Over $\mathbb{Z}[[q]]$ one has the Tate curve $\mathrm{Tate}(q)$, which is equipped with a basis dt/t of $\underline{\omega}_{\mathrm{Tate}(q)/\mathbb{Z}[[q]]}$. One has an isomorphism of groups:

$$(\mathbb{Z}/N\mathbb{Z})^2 \longrightarrow \mathrm{Tate}(q)[N](\overline{\mathbb{Q}((q))}),\quad (a,b)\mapsto \zeta_N^a(q^{1/N})^b, \tag{1.3}$$

where $\zeta_N\in\overline{\mathbb{Q}}^*$ is a fixed root of unity of order N. For R a $\mathbb{Z}[1/N,\zeta_N]$-algebra one has the standard q-expansion map $M(N,k)_R\to R[[q]]$ sending f to the series $\sum_{n\geq 0}a_n(f)q^n$ defined by:

$$f(\mathrm{Tate}(q),1\mapsto\zeta_N) = \left(\sum_{n\geq 0}a_n(f)q^n\right)(dt/t)^k. \tag{1.4}$$

This map is injective. The R-modules $M(N,k)_R$ can also be defined as follows. One replaces "generalized elliptic curve" in the definition we gave by "elliptic curve" and one demands that the q-expansions obtained from all points of order N of the Tate curve over $\mathbb{Z}[\zeta_N]((q^{1/N}))$ are power series in $q^{1/N}$ (a priori they are Laurent series). The R-module of cusp forms of level N and weight k is defined to be the submodule $M^0(N,k)_R$ of $M(N,k)_R$ of those f all of whose q-expansions have zero as constant coefficient. Equivalently, they are the forms that vanish on degenerate elliptic curves.

Suppose that $k \geq 1$. The R-modules $M(N,k)_R$ and $M^0(N,k)_R$ are equipped with certain endomorphisms. For $n \geq 1$ one has the Hecke operator T_n, defined in terms of isogenies of degree n. For a in $(\mathbb{Z}/N\mathbb{Z})^*$ one has the diamond operator $\langle a \rangle$, induced by the automorphism of $[\Gamma_1(N)_R]$ that sends $(E/S/R, \alpha)$ to $(E/S/R, a\alpha)$. The action of these operators is given by the usual formulas in terms of q-expansions (see Rohrlich's chapter).

In particular, one has $a_1(T_n(f)) = a_n(f)$. The construction of the T_n in this generality is a bit complicated, especially for $k = 1$. See for example [17, §4] for a construction of T_p on $M(N,1)_{\mathbb{F}_p}$. In general, one can construct the T_n as follows. It suffices to construct the T_p with p prime. If p divides N then p is invertible in R and the construction is easy. So suppose that p does not divide N. In this case one proceeds as in [17, §4]. Take f in $M(N,k)_R$, view it as a G-invariant section of $\underline{\omega}^{\otimes k}$ on $\overline{\mathcal{M}}([\Gamma_1(N), \Gamma(n)]_R)$, as in (1.2). Then restrict it to $\mathcal{M}([\Gamma_1(N), \Gamma(n)]_R)$, the complement of the cusps. This is affine, hence this restriction of f is a linear combination, with coefficients in R, of sections of $\underline{\omega}^{\otimes k}$ over $\mathcal{M}([\Gamma_1(N), \Gamma(n)]_{\mathbb{Z}[1/Nn]})$ and one knows what T_p does with those. One verifies that the $T_p(f)$ obtained in this way is again G-invariant and regular at the cusps.

The endomorphisms T_n and $\langle a \rangle$ of $M(N,k)_R$ all commute with each other. For ε a character of $(\mathbb{Z}/N\mathbb{Z})^*$ with values in R^* let $M(N,k,\varepsilon)_R$ denote the R-submodule of $M(N,k)_R$ of elements f such that $\langle a \rangle(f) = \varepsilon(a)f$ for all a; such f will be called forms of type (N,k,ε). If R is a field and f a non-zero element of $M(N,k)_R$ which is an eigenform for all T_n, then the formula $a_1(T_n(f)) = a_n(f)$ implies that the corresponding eigenspace for the T_n has dimension one, and that there is a unique character ε such that f is of type (N,k,ε).

1.5 Theorem. (Deligne) *Let p be a prime, N an integer prime to p, and $k \geq 1$. Let f be a non-zero eigenform in $M(N,k,\varepsilon)_{\overline{\mathbb{F}}_p}$, and let a_n be its eigenvalue for T_n. Then there is a unique semi-simple continuous representation $\rho_f \colon G_{\mathbb{Q}} := \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$, unramified outside pN, such that for all primes l not dividing pN the (arithmetic) Frobenius element $\rho_f(\mathrm{Frob}_l)$ has trace a_l and determinant $\varepsilon(l)l^{k-1}$.*

(The topology on $\mathrm{GL}_2(\overline{\mathbb{F}}_p)$ in this statement is the discrete one.) In fact, Deligne [8] showed that for an eigenform f of weight $k \geq 2$ and with coefficients in $\overline{\mathbb{Q}}_p$ one even has a p-adic representation ρ_f of $G_{\mathbb{Q}}$ with values in $\mathrm{GL}_2(\overline{\mathbb{Q}}_p)$. The case $k = 2$ had already been treated by Shimura in [40]. A detailed and reasonably elementary proof of Theorem 1.5 can be found in [17].

Let f be as in Theorem 1.5. The fact that every elliptic curve has the automorphism -1 implies that $\varepsilon(-1)f = \langle -1 \rangle(f) = (-1)^k f$, hence that $\varepsilon(-1) = (-1)^k$. It follows that, for σ in $G_{\mathbb{Q}}$ any complex conjugation, $\det(\rho_f(\sigma)) = \varepsilon(-1)(-1)^{k-1} = -1$. A representation of $G_{\mathbb{Q}}$ with this property will be called odd. Serre's conjecture [37, (3.2.4)$_?$] states that any continuous odd representation $\rho \colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ is isomorphic to some ρ_f, and for irreducible ρ it claims that such a f exists of a certain "minimal" type $(N(\rho), k(\rho), \varepsilon(\rho))$. Before we can state Serre's conjecture, we need some terminology.

Let p be a prime and $\rho \colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ a continuous representation. Then the image G of ρ is finite, let's say equal to $\mathrm{Gal}(K/\mathbb{Q})$ with K a finite Galois extension of $\mathbb{Q}$ in $\overline{\mathbb{Q}}$. It follows that

ρ is unramified at all but finitely many primes. The number $N(\rho)$ is by definition the Artin conductor of ρ, except that one doesn't take the prime p into account. More precisely, $N(\rho)$ will be a positive integer prime to p. Let l be a prime number different from p. Let I_l denote the inertia subgroup at l of G corresponding to a place of K above l, and let $I_l = I_{l,0} \supset I_{l,1} \supset \cdots$ be the higher ramification subgroups: $I_{l,i}$ is the subgroup of the decomposition group at the chosen place whose elements act trivially on the ring of integers modulo the $(i+1)$th power of the maximal ideal. The valuation of $N(\rho)$ at l, i.e., the number of factors l in it, is then given by the formula:

$$v_l(N(\rho)) = \sum_{i \geq 0} \frac{1}{[I_{l,0} : I_{l,i}]} \dim(V/V^{I_{l,i}}), \tag{1.6}$$

where V is a two-dimensional $\overline{\mathbb{F}}_p$-vector space with a G-action giving ρ, and where $V^{I_{l,i}}$ denotes its subspace of invariants under $I_{l,i}$. For a motivation for formula (1.6), see [38], pp. 171–174. Let us note, by the way, that replacing $V/V^{I_{l,i}}$ by the kernel of the map to the coinvariants $V \to V_{I_{l,i}}$ would give the same result. For $i > 0$ this is clear since then $I_{l,i}$ is an l-group, for $i = 0$ one uses that $V_{I_{l,0}}$ is obtained by first taking the coinvariants for the l-group $I_{l,1}$ and then the coinvariants for the cyclic group $I_{l,0}/I_{l,1}$.

We can now define the character $\varepsilon(\rho)$ and the image of $k(\rho)$ in $\mathbb{Z}/(p-1)\mathbb{Z}$. For each positive integer n let $\chi_n : G_\mathbb{Q} \to (\mathbb{Z}/n\mathbb{Z})^*$ denote the cyclotomic character of the nth roots of unity, i.e., the character such that for all σ in $G_\mathbb{Q}$ and all x in $\overline{\mathbb{Q}}$ with $x^n = 1$ we have $\sigma(x) = x^{\chi(\sigma)}$. We will often view a character of the group $(\mathbb{Z}/n\mathbb{Z})^*$ as a character of $G_\mathbb{Q}$, by composing it with χ_n. With these conventions, Theorem 1.5 implies that for f an eigenform in $M(N,k,\varepsilon)_{\overline{\mathbb{F}}_p}$ one has $\det \circ \rho_f = \varepsilon \chi_p^{k-1}$. This equality shows that $\det \circ \rho_f$ is a character of the group $(\mathbb{Z}/pN\mathbb{Z})^*$. This group is canonically isomorphic to $(\mathbb{Z}/N\mathbb{Z})^* \times \mathbb{F}_p^*$ via the corresponding isomorphism of rings. Under this decomposition, the restriction of $\det \circ \rho_f$ to $(\mathbb{Z}/N\mathbb{Z})^*$ is ε, and the restriction to $\mathbb{F}_p^*$ is χ_p^{k-1}. Let us now go back to the representation ρ. By the definition of $N(\rho)$, it is unramified away from $pN(\rho)$. Since the maximal abelian extension of $\mathbb{Q}$ is the cyclotomic extension, $\det \circ \rho$ can be considered as a character of $(\mathbb{Z}/p^n\mathbb{Z})^* \times (\mathbb{Z}/N(\rho)^m)^*$, for some $n, m \gg 0$. Since the character has values in $\overline{\mathbb{F}}_p^*$, one can take $n = 1$. Comparing formula (1.6) for ρ and $\det \circ \rho$, and using some class field theory, one sees that one can take $m = 1$. Then one defines $\varepsilon(\rho)$ to be the restriction of $\det \circ \rho$ to the factor $(\mathbb{Z}/N(\rho)\mathbb{Z})^*$, and the image of $k(\rho)$ in $\mathbb{F}_p^*$ such that the restriction of $\det \circ \rho$ to $\mathbb{F}_p^*$ is $\chi_p^{k(\rho)-1}$.

Now we get to the exact definition of $k(\rho)$. The first thing to note is that $k(\rho)$ depends only on the restriction of ρ to some inertia subgroup $I_p \subset G_\mathbb{Q}$ at p. Hence the level $N(\rho)$ reflects the ramification away from p and the weight $k(\rho)$ the ramification at p. In order to state the definition of $k(\rho)$ we first have to classify the two-dimensional representations over $\overline{\mathbb{F}}_p$ of a decomposition subgroup $G_p \subset G_\mathbb{Q}$ at p (this subgroup G_p depends on the choice of a maximal ideal containing p, but since these are permuted transitively by $G_\mathbb{Q}$ the choice will not matter). For this classification, we need some terminology.

Let $\overline{\mathbb{Z}}$ be the integral closure of $\mathbb{Z}$ in $\overline{\mathbb{Q}}$. Let us choose a morphism of rings $\overline{\mathbb{Z}} \to \overline{\mathbb{F}}_p$. This gives us a decomposition group G_p (the stabilizer of the kernel). The action of G_p on $\overline{\mathbb{F}}_p$ gives a surjection $G_p \to G_{\mathbb{F}_p}$; its kernel I_p is the inertia group. We can identify G_p with $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p)$ and I_p with $\mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p^{\mathrm{unr}})$, where $\mathbb{Q}_p^{\mathrm{unr}}$ is the maximal unramified extension of $\mathbb{Q}_p$. Let $\mathbb{Q}_p^{\mathrm{tr}}$ be the maximal tamely ramified extension of $\mathbb{Q}_p$. Then the quotient $I_{p,\mathrm{t}} := \mathrm{Gal}(\mathbb{Q}_p^{\mathrm{tr}}/\mathbb{Q}_p^{\mathrm{unr}})$ of I_p is called the tame ramification group, and the subgroup $I_{p,\mathrm{w}} := \mathrm{Gal}(\overline{\mathbb{Q}}_p/\mathbb{Q}_p^{\mathrm{tr}})$ of I_p the wild ramification group. The tame extensions of $\mathbb{Q}_p^{\mathrm{unr}}$ are obtained by taking nth roots of elements p, with n prime to p. It follows that $I_{p,\mathrm{t}}$ can be identified with $\lim\limits_{\longleftarrow} \mathbb{F}_{p^n}^*$, where the limit is taken over the $n \geq 1$ and where the transition morphisms are norm maps. A character $\phi\colon I_{p,\mathrm{t}} \to \overline{\mathbb{F}}_p^*$ is called of level n, if n is the smallest integer such that ϕ factors through $\mathbb{F}_{p^n}^*$. The n characters $I_{p,\mathrm{t}} \to \mathbb{F}_{p^n}^* \to \overline{\mathbb{F}}_p^*$ that are induced by embeddings of fields $\mathbb{F}_{p^n} \to \overline{\mathbb{F}}_p$ are called the fundamental characters of level n.

Suppose now that $\rho_p\colon G_p \to \mathrm{GL}(V)$ is a continuous 2-dimensional representation on a $\overline{\mathbb{F}}_p$-vector space. Let V^{ss} be the semi-simplification of V for the action of G_p. we claim that $I_{p,\mathrm{w}}$ acts trivially on V^{ss}. To see that, let W be an irreducible factor of V^{ss}. Then the subspace $W^{I_{p,\mathrm{w}}}$ of W is stable under G_p because $I_{p,\mathrm{w}}$ is a normal subgroup of it, so either $W^{I_{p,\mathrm{w}}}$ is trivial or it is W. Since the image of G_p in $\mathrm{GL}(W)$ is finite, the representation of G_p on W can be realized over a finite extension of $\mathbb{F}_p$. For such a realization W', the number of fixed points for $I_{p,\mathrm{w}}$ is a multiple of p since the image of $I_{p,\mathrm{w}}$ is a p-group, hence W' has non-trivial $I_{p,\mathrm{w}}$-invariants. This proves that the action of I_p on V^{ss} is given by two characters $\phi, \phi'\colon I_{p,\mathrm{t}} \to \overline{\mathbb{F}}_p^*$. Since $\mathrm{Gal}(\overline{\mathbb{F}}_p/\mathbb{F}_p)$ acts by conjugation on $I_{p,\mathrm{t}}$, it follows that $\{\phi^p, \phi'^p\} = \{\phi, \phi'\}$. This means that there are two cases: either ϕ and ϕ' are both of level one, or ϕ and ϕ' are both of level two and $\phi^p = \phi'$, $\phi'^p = \phi$. In the first case ρ_p is reducible, whereas in the second case ρ_p is absolutely irreducible.

We need one more ingredient before we can define $k(\rho_p)$: the notion of "finiteness at p". Let $\mathbb{F}$ be a finite extension of $\mathbb{F}_p$ such that ρ_p can be realized over $\mathbb{F}$. An $\mathbb{F}$-vector space scheme V over a scheme S is then an S-scheme V, together with a structure of $\mathbb{F}$-vector space on all the sets of points $V(T)$ for S-schemes T, functorially in T. Equivalently, an $\mathbb{F}$-vector space scheme over S is a commutative group scheme over S given with an action of $\mathbb{F}$ on it by endomorphisms. An $\mathbb{F}$-vector space scheme V over $\mathbb{Q}_p$ gives rise to the representation of G_p on the $\mathbb{F}$-vector space $V(\overline{\mathbb{Q}}_p)$. This construction induces an equivalence between the category of $\mathbb{F}$-vector space schemes finite over $\mathbb{Q}_p$ and the category of representations of G_p on finite dimensional $\mathbb{F}$-vector spaces. Concretely, the $\mathbb{F}$-vector space scheme $V_{\mathbb{Q}_p}$ corresponding to ρ_p can be obtained as the quotient of the scheme $\mathbb{F}^2_{\mathbb{Q}_p} \times_{\mathrm{Spec}(\mathbb{Q}_p)} \mathrm{Spec}(\overline{\mathbb{Q}}_p)$ by the group G_p, with σ in G_p acting as $(\rho_p(\sigma)^{-1}, \mathrm{Spec}(\sigma))$. One says that ρ_p is finite at p if $V_{\mathbb{Q}_p}$ can be extended to a finite flat $\mathbb{F}$-vector space scheme over $\mathbb{Z}_p$. It is equivalent to demand that $V_{\mathbb{Q}_p}$ can be extended to a finite flat $\mathbb{F}$-vector space scheme over the ring of integers $\mathbb{Z}_p^{\mathrm{unr}}$ in $\mathbb{Q}_p^{\mathrm{unr}}$, or to demand that it can be extended over $\mathbb{Z}_p$ or $\mathbb{Z}_p^{\mathrm{unr}}$ as a finite flat group scheme. Finally, one can formulate the condition that ρ_p be finite at p in Galois theoretic terms, without using group schemes (see [15, Proposition 8.1], and its proof, and [37, (2.4.7)]).

We will now give the definition of $k(\rho_p)$. Let ψ and ψ' be the two fundamental characters of level 2.

1.7 Definition. Let $\rho_p\colon G_p \to GL(V)$, V^{ss}, ϕ and ϕ' be as above. We associate an integer

$k(\rho_p)$ to ρ_p as follows.

1. Suppose that ϕ and ϕ' are of level 2. we have:
$$\rho_p|_{I_p} \cong \begin{pmatrix} \phi & 0 \\ 0 & \phi' \end{pmatrix}.$$
After interchanging ϕ and ϕ' if necessary, we have (uniquely) $\phi = \psi^{a+pb} = \psi^a\psi'^b$ and $\phi' = \psi'^a\psi^b$ with $0 \leq a < b \leq p-1$. We set $k(\rho_p) = 1 + pa + b$.

2. Suppose that ϕ and ϕ' are of level 1.

   (a) If $\rho_p|_{I_{p,\mathrm{w}}}$ is trivial then we have:
$$\rho_p|_{I_p} \cong \begin{pmatrix} \chi^a & 0 \\ 0 & \chi^b \end{pmatrix}$$
   with $0 \leq a \leq b \leq p-2$. We set $k(\rho_p) = 1 + pa + b$.

   (b) Suppose that $\rho_p|_{I_{p,\mathrm{w}}}$ is not trivial. We have:
$$\rho_p|_{I_p} \cong \begin{pmatrix} \chi^\beta & * \\ 0 & \chi^\alpha \end{pmatrix}$$
   for unique α and β with $0 \leq \alpha \leq p-2$ and $1 \leq \beta \leq p-1$. We set $a = \min(\alpha, \beta)$, $b = \max(\alpha, \beta)$. If $\chi^{\beta-\alpha} = \chi$ and $\rho_p \otimes \chi^{-\alpha}$ is not finite at p then we set $k(\rho_p) = 1 + pa + b + p - 1$; otherwise we set $k(\rho_p) = 1 + pa + b$.

□

For $\rho\colon G_\mathbb{Q} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ continuous, we define $k(\rho)$ to be $k(\rho_p)$, where ρ_p is the restriction of ρ to a decomposition group at p.

1.8 Conjecture. (Serre) *Let p be a prime number and $\rho\colon G_\mathbb{Q} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ a continuous irreducible odd representation. Then there exists an eigenform f in $M^0(N(\rho), k(\rho), \varepsilon(\rho))_{\overline{\mathbb{F}}_p}$ such that ρ is isomorphic to ρ_f.*

Several remarks should be made, at this point. First of all, it is easy to see that for f an eigenform which is not cuspidal, the representation ρ_f is reducible (restrict f to the cusps, and study the action of the Hecke operators on the cusps). This explains that one can demand in the conjecture f to be a cuspform.

The second remark is more important. The conjecture, as we state it here (i.e., as suggested in [15, 4.3]), is not equivalent the one stated by Serre in [37, (3.2.4)$_?$]. The difference comes from the fact that the modular forms over $\overline{\mathbb{F}}_p$ considered in [37] are defined to be those obtained by reduction mod p of forms over $\overline{\mathbb{Q}}$. More precisely, in [37], a cuspidal modular form over $\overline{\mathbb{F}}_p$ of some type (N, k, ε), with N prime to p and $k \geq 2$, is an element in the image of the map $M^0(N, k, \widetilde{\varepsilon})_{\overline{\mathbb{Z}}_p} \to M^0(N, k, \varepsilon)_{\overline{\mathbb{F}}_p}$, where $\overline{\mathbb{Z}}_p$ is the integral closure of $\mathbb{Z}_p$ in $\overline{\mathbb{Q}}_p$ and $\widetilde{\varepsilon}\colon (\mathbb{Z}/N\mathbb{Z})^* \to$

$\overline{\mathbb{Z}}_p^*$ is the Teichmüller lift of ε, i.e., $\widetilde{\varepsilon}$ induces ε and they have the same order. The problem is that these reduction maps are not all surjective. Before we discuss what is known today about Serre's conjecture, we will discuss the differences between Conjecture 1.8 and [37, (3.2.4)$_?$]. It was suggested by Serre in his 1987–1988 course at the Collège de France to replace the mod p modular forms in [37] by those defined by Katz, i.e., the ones we are using here.

Let us first consider the problem of lifting modular forms from $\overline{\mathbb{F}}_p$ to $\overline{\mathbb{Z}}_p$, without paying attention to the character. Then we have the following result.

1.9 Lemma. *Let p be a prime, $N \geq 1$ prime to p.*

1. *Suppose that $k \geq 2$. Then the map $M^0(N,k)_{\overline{\mathbb{Z}}_p} \to M^0(N,k)_{\overline{\mathbb{F}}_p}$ is surjective if $N \neq 1$ or if $p > 3$.*
2. *The map $M^0(1,k)_{\overline{\mathbb{Z}}_2} \to M^0(1,k)_{\overline{\mathbb{F}}_2}$ is not surjective if and only if $k \geq 12$ and ($k \equiv 1$ mod 2 or $k \equiv 2$ mod 12).*
3. *The map $M^0(1,k)_{\overline{\mathbb{Z}}_3} \to M^0(1,k)_{\overline{\mathbb{F}}_3}$ is not surjective if and only if $k \geq 12$ and $k \equiv 2$ mod 12.*

Proof. Let us prove the first statement; the other two can be proved using the explicit descriptions of the rings of modular forms of level one over $\mathbb{Z}$, $\mathbb{F}_2$ and $\mathbb{F}_3$ found in [9, Proposition 6.2]. Suppose first that $N \geq 5$. Because $\overline{\mathbb{Z}}_p$ is flat over $\mathbb{Z}_p$ and $\overline{\mathbb{F}}_p$ is flat over $\mathbb{F}_p$, it suffices to prove that the reduction map induced by $\mathbb{Z}_p \to \mathbb{F}_p$ is surjective. Consider the long exact cohomology sequence arising from the short exact sequence of sheaves on $X_1(N)_{\mathbb{Z}_p}$:

$$0 \to \underline{\omega}^{\otimes k}(-\text{cusps}) \to \underline{\omega}^{\otimes k}(-\text{cusps}) \to i_*\underline{\omega}^{\otimes k}(-\text{cusps}) \to 0, \tag{1.9.1}$$

where the map $\underline{\omega}^{\otimes k}(-\text{cusps}) \to \underline{\omega}^{\otimes k}(-\text{cusps})$ is multiplication by p and $i\colon X_1(N)_{\mathbb{F}_p} \to X_1(N)_{\mathbb{Z}_p}$ denotes the closed immersion. To get the surjectivity we want, it is sufficient to show that $\mathrm{H}^1(X_1(N)_{\mathbb{F}_p}, \underline{\omega}^{\otimes k}(-\text{cusps})) = 0$, since by Nakayama's lemma and the long exact sequence this implies that $\mathrm{H}^1(X_1(N)_{\mathbb{Z}_p}, \underline{\omega}^{\otimes k}(-\text{cusps})) = 0$. The Kodaira-Spencer isomorphism (see [22, A1.4]) and Serre duality give isomorphisms:

$$\mathrm{H}^1(X_1(N)_{\mathbb{F}_p}, \underline{\omega}^{\otimes k}(-\text{cusps})) = \mathrm{H}^1(X_1(N)_{\mathbb{F}_p}, \Omega \otimes \underline{\omega}^{\otimes k-2}) = \mathrm{H}^0(X_1(N)_{\mathbb{F}_p}, \underline{\omega}^{\otimes 2-k})^\vee.$$

So if $k > 2$, this shows what we want, since the degree of $\underline{\omega}$ is positive. The case $k = 2$ is in fact easy, since the Kodaira-Spencer isomorphism identifies weight 2 cuspforms with differential forms, and those can be lifted. Another way to phrase the argument is to say that the dimensions of $M^0(N,k)_{\mathbb{F}_p}$ and $M^0(N,k)_{\mathbb{Q}_p}$ are given by the Riemann–Roch formula since the H^1-term vanishes, and that hence the reduction map is surjective.

Suppose now that $p > 3$. Then $M^0(N,k)_{\mathbb{F}_p} = \mathrm{H}^0(\overline{\mathcal{M}}([\Gamma_1(N), \Gamma(3)]_{\mathbb{F}_p}), \underline{\omega}^{\otimes k}(-\text{cusps}))^G$, with $G = \mathrm{GL}_2(\mathbb{F}_3)$. Since p does not divide the order of G, the functor $M \mapsto M^G$ from $\mathbb{Z}_p[G]$-modules to $\mathbb{Z}_p$-modules is exact. Combining this with the long exact sequence arising from the short exact sequence $0 \to \underline{\omega}^{\otimes k}(-\text{cusps}) \to \underline{\omega}^{\otimes k}(-\text{cusps}) \to i_*\underline{\omega}^{\otimes k}(-\text{cusps}) \to 0$ on $\overline{\mathcal{M}}([\Gamma_1(N), \Gamma(3)])_{\mathbb{Z}_p}$ gives the result.

Suppose now that $N = 2$ or $N = 4$. Then $p \neq 2$. In these cases, the category $[\Gamma_1(N)]_{\mathbb{Z}_p}$ is the quotient, in the sense of algebraic stacks, for the action of a subgroup G of $\mathrm{GL}_2(\mathbb{Z}/4\mathbb{Z})$ acting on $[\Gamma(4)]_{\mathbb{Z}_p}$. This gives a formula analogous to (1.2). The group G is a 2-group, hence or order prime to p. One can then apply an argument which is similar to the one used in the case $p > 3$ above.

Suppose now that $N = 3$. Then $p \neq 3$. Let $\mathbb{Z}_p[\zeta_3]$ be the subring of $\overline{\mathbb{Z}}_p$ generated by $\mathbb{Z}_p$ and a third root of unity ζ_3. Let $[\Gamma(3)^{\zeta_3-\mathrm{can}}]_{\mathbb{Z}_p[\zeta_3]}$ be the category of generalized elliptic curves over schemes over $\mathbb{Z}_p[\zeta_3]$ with an embedding α of the constant group scheme $(\mathbb{Z}/3\mathbb{Z})^2$, such that the Weil pairing of $\alpha(1,0)$ and $\alpha(0,1)$ equals ζ_3. Then $[\Gamma_1(N)]_{\mathbb{Z}_p[\zeta_3]}$ is the quotient of $[\Gamma(3)^{\zeta_3-\mathrm{can}}]_{\mathbb{Z}_p[\zeta_3]}$ for the action of a group of order 3. This means that one can again use the same argument. $\square$

The proof of this lemma indicates that the case $k = 1$ is very different, since the degree of $\underline{\omega}$ is, as one sees from the Kodaira-Spencer isomorphism, too small to make the H^1-term in the Riemann–Roch theorem vanish. Mestre has indeed found examples with $p > 3$ where the map $M^0(N,1)_{\mathbb{Z}_p} \to M^0(N,1)_{\mathbb{F}_p}$ is not surjective. In these examples one has an eigenform f in $M^0(N,1)_{\overline{\mathbb{F}}_p}$ such that the image of the representation ρ_f is too big to be embeddable in $GL_2(\mathbb{C})$; if ρ could be lifted to characteristic zero this would contradict the theorem of Deligne–Serre (Theorem 4.1 of [11]). Let us note that for a representation ρ as in Conjecture 1.8 it can very well happen that $k(\rho) = 1$; one can check that this is equivalent to ρ being unramified at p. This explains that the weight k_ρ for f that one finds in [37, (3.2.4)$_?$], is not in all cases the same as $k(\rho)$ defined above. The difference between k_ρ and $k(\rho)$ can be summarized as follows, in the notation of Definition 1.7. There are only two cases where they are different; in both cases the characters ϕ and ϕ' are of level 1. In the first case, the restriction of ρ to the wild inertia group $I_{p,\mathrm{w}}$ is trivial and $a = 0 = b$; then $k(\rho) = 1$ and $k_\rho = p$. In the second case $p = 2$, ρ is wildly ramified at 2, $\alpha = 0$, $\beta = 1$ and ρ is not finite at 2; then $k(\rho) = 3$ and $k_\rho = 4$.

Other problems arise if we take the character into account. In his course at the Collège de France, 1987–1988, Serre gave some counter examples against his conjecture [37, (3.2.4)$_?$]. These examples are found by considering the genus two curve $X_1(13)$. On this curve there are two eigenforms and of weight 2 over $\mathbb{Z}[\zeta_3]$, and the two corresponding characters are of order 6. The reductions mod 2 and 3 of these eigenforms have characters of order 3 and 2, respectively. One verifies that the Galois representations corresponding to these mod 2 and mod 3 forms are irreducible, and that the weights associated to them equal two (Definition 1.7 and [37, §2] coincide in these cases). In fact, these representations are dihedral, induced from $G_{\mathbb{Q}(\sqrt{-1})}$ and $G_{\mathbb{Q}(\sqrt{-3})}$, respectively. According to [37, (3.2.4)$_?$], the mod 2 and mod 3 reductions of the two eigenforms should have lifts to weight two eigenforms in characteristic zero on $X_1(13)$ with a character of the same order as the reduction. But since the genus is two, there are no such forms.

In the same course at the Collège de France, Serre showed that the only mod p eigenforms f of weight at least 2 that cannot be lifted to an eigenform with the same level, weight and order of character are among those in characteristic 2 or 3, whose representation ρ_f is induced from

$G_{\mathbb{Q}(\sqrt{-1})}$ or $G_{\mathbb{Q}(\sqrt{-3})}$, respectively. This result was obtained independently by Carayol, see [4, §4.4], and is usually called Carayol's Lemma. Serre's proof uses a result of Nakajima implying that for $n \geq 3$ prime to p and $k \geq 2$ the $\mathbb{Z}_p[\mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z})]$-module $\mathrm{H}^0(X(n)_{\mathbb{Z}_p}, \underline{\omega}^{\otimes k})$ is projective, if all stabilizers are of order prime to p.

Carayol's proof uses the realization of the Galois representation associated to modular forms in the first cohomology group of certain p-adic sheaves on modular curves over $\overline{\mathbb{Q}}$. His arguments can be adapted to the sheaves $\underline{\omega}^{\otimes k}$. This gives the following result, that we state without proof, and which can also be found in Serre's notes. Yet another version of it can be found in [12, §2].

1.10 Proposition. (Carayol's Lemma) *Let $N \geq 1$, p a prime not dividing N, and $k \geq 2$. Let $\varepsilon\colon (\mathbb{Z}/N\mathbb{Z})^* \to \overline{\mathbb{Z}}_p^*$ be a character with $\varepsilon(-1) = (-1)^k$, and let $\bar{\varepsilon}\colon (\mathbb{Z}/N\mathbb{Z})^* \to \overline{\mathbb{F}}_p^*$ be its reduction. Consider the map $\phi\colon M^0(N,k,\varepsilon)_{\overline{\mathbb{Z}}_p} \to M^0(N,k,\bar{\varepsilon})_{\overline{\mathbb{F}}_p}$. If $p \geq 5$, then ϕ is surjective. If $p = 3$ (resp., $p = 2$) and $f \in M^0(N,k,\bar{\varepsilon})_{\overline{\mathbb{F}}_p}$ is an eigenform with ρ_f irreducible and f not in the image of ϕ, then ρ_f is induced from $\mathbb{Q}(\sqrt{-3})$ (resp., $\mathbb{Q}(\sqrt{-1})$).* $\square$

In both proofs of this result it is quite clear where the $\mathbb{Q}(\sqrt{-3})$ and $\mathbb{Q}(\sqrt{-1})$ come from. Suppose for simplicity that $N \geq 5$. In Carayol's proof, it comes from the fact that the points of $X_0(N)_{\overline{\mathbb{Q}}}$ with an automorphism (i.e., an automorphism of the pair $(E/\overline{\mathbb{Q}}, G)$ corresponding to it) of order 3 (resp., order 4) are defined over abelian extensions of $\mathbb{Q}(\sqrt{-3})$ (resp., $\mathbb{Q}(\sqrt{-1})$). In Serre's proof, it comes from the fact that for primes $l \equiv -1 \bmod 3$ (resp., mod 4) there is no elliptic curve with an automorphism of order 6 (resp., 4) fixing a subgroup of order l, implying that if an eigenform is not liftable (in the sense of Proposition 1.10) then it is annihilated by T_l for such l; this implies that the character of ρ_f vanishes on Frob_l for such l; then it follows that ρ_f is induced as stated.

The statement of Carayol's Lemma in [4] is actually different from Proposition 1.10 above. It says that if an irreducible representation $\rho\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ arises from some eigenform f in $M(N,k,\varepsilon)_{\overline{\mathbb{Q}}_p}$ with $k \geq 2$, then for every character $\varepsilon'\colon (\mathbb{Z}/N\mathbb{Z})^* \to \overline{\mathbb{Q}}_p^*$ inducing the same character $(\mathbb{Z}/N\mathbb{Z})^* \to \overline{\mathbb{F}}_p^*$ as ε, there exists an eigenform f' inducing ρ. In this statement, which does not speak of modular forms mod p, one does not suppose that N is prime to p.

Let us now discuss what is known about Conjecture 1.8 and about its relation to [37, (3.2.4)$_?$].

1.11 Proposition. *Let p be a prime, and let $\rho\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ be continuous, irreducible and odd.*

1. *If $p = 2$ (resp., $p = 3$), suppose that ρ is not induced from $\mathbb{Q}(\sqrt{-1})$ (resp., $\mathbb{Q}(\sqrt{-3})$). Then if ρ satisfies Conjecture 1.8, it satisfies [37, (3.2.4)$_?$].*

2. *If $p = 2$, suppose that the restriction of ρ to a decomposition group at 2 is not an extension of some character by itself. Then if ρ satisfies [37, (3.2.4)$_?$], it satisfies Conjecture 1.8.*

1.12 Theorem. *Let $p > 2$ be a prime, and let $\rho\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ be continuous, irreducible and odd. Suppose that ρ comes from a modular form of some type. Then ρ satisfies Conjecture 1.8.*

Moreover, if ρ comes from a mod p modular form of some type (N,k,ε) with N prime to p, then N is a multiple of $N(\rho)$, $k \geq k(\rho)$ and ε is obtained from $\varepsilon(\rho)$ via composition by $\mathbb{Z}/N\mathbb{Z} \to \mathbb{Z}/N(\rho)\mathbb{Z}$.

The proof of these results is quite long and many people have contributed to it. A complete proof can be found by reading Diamond's article [12], and the references therein. A very good overview of the strategy of the whole proof is given in Ribet's report [32]. In the next section we will see which parts of these results are used in the proof of the conjecture of Shimura–Taniyama–Weil for semi-stable elliptic curves over $\mathbb{Q}$ and the proof of Fermat's Last Theorem. In Sections 3 and 4, we will then describe the proofs of those cases. To finish this section, we will briefly recall the history of the proofs of Proposition 1.11 and Theorem 1.12.

For the rest of this section, let p be prime and $\rho\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ be continuous, irreducible and odd. We will say that ρ is modular of type $(N,k,\varepsilon)_{\overline{\mathbb{F}}_p}$ (resp., $(N,k,\varepsilon)_{\overline{\mathbb{Q}}_p}$) if there is an eigenform f in $M^0(N,k,\varepsilon)_{\overline{\mathbb{F}}_p}$ (resp., $M^0(N,k,\varepsilon)_{\overline{\mathbb{Q}}_p}$) such that f gives the representation ρ. Serre formulated, in a letter to Mestre dated August 13, 1985, a part of his conjectures that, together with the Shimura–Taniyama–Weil conjecture, implies Fermat's Last Theorem. Mazur proved, in a letter to Mestre dated August 16, 1985, the following result: suppose moreover that $p > 2$, that ρ is modular of some type $(N,2,1)_{\overline{\mathbb{Q}}_p}$, that l is a prime not congruent to 1 mod p, that l divides N but l^2 does not, that ρ is unramified at l if $l \neq p$ and that ρ is finite at p if $l = p$; then ρ is modular of type $(N/l,2,1)_{\overline{\mathbb{Q}}_p}$. In 1987, Ribet removed the condition "$l \not\equiv 1 \bmod p$" from Mazur's result, under the assumption that p does not divide N (see [33]). These two results together imply already that Fermat's Last Theorem is a consequence of the Shimura–Taniyama–Weil conjecture. Together with Mazur [29], Ribet extended his result to the case where p divides N, but where p^2 does not.

Langlands, Deligne and Carayol have proved [5] that for f a newform the conductor of the system of l-adic representations associated to f is equal to the level of f. From this it follows easily that for f an eigenform in some $M^0(N,k,\varepsilon)_{\overline{\mathbb{Q}}_p}$, and ρ_f the mod p Galois representation that it gives, $N(\rho_f)$ divides N. Carayol [4] and Livné [26] classified, independently, in terms of the admissible irreducible representation of $\mathrm{GL}_2(\mathbb{Q}_l)$ associated to a newform f in some $M^0(N,k,\varepsilon)_{\overline{\mathbb{Q}}_p}$, the cases where the l-adic valuations of $N(\rho_f)$ and N are different (here ρ_f denotes the mod p representation associated to f, ρ_f is supposed to be irreducible and l is a prime different from p). Carayol [4] showed that if ρ is modular of type $(N,k,\varepsilon)_{\overline{\mathbb{Q}}_p}$ and not induced from $\mathbb{Q}(\sqrt{-1})$ (resp., $\mathbb{Q}(\sqrt{-3})$) if $p = 2$ (resp., $p = 3$), then it is modular of type $(N,k,\varepsilon')_{\overline{\mathbb{Q}}_p}$ for all ε' whose mod p reduction equals that of ε and such that $\varepsilon'(-1) = (-1)^k$ (this last condition is implied by the first if $p \neq 2$). This result shows that in order to prove Proposition 1.11 and Theorem 1.12 one need not pay attention anymore to the character, so we will drop it from the notation in what follows.

Suppose now that ρ is modular of some type $(N,k)_{\overline{\mathbb{Q}}_p}$. Then one wants to prove that ρ is modular of type $(N(\rho),k)$. In [4] Carayol reduces the proof of this, for $p \geq 5$ and $k \geq 2$, to the following two statements:

(A) There exists a prime number q not dividing Nl and a newform of type $(N'q,k,\varepsilon')_{\overline{\mathbb{Q}}_p}$ with

N' dividing N and ε' trivial mod q, whose associated mod p Galois representation is isomorphic to ρ.

(B) If $l \neq p$ divides N, l^2 does not divide N and l does not divide $N(\rho)$ (i.e., ρ is unramified at l), then ρ is modular of type $(N/l, k)_{\overline{\mathbb{Q}}_p}$.

The first of these two statements is used to switch in certain cases from modular curves to Shimura curves associated to indefinite quaternion algebras over $\mathbb{Q}$, via the Jacquet–Langlands correspondence. The main part of Ribet's article [32] is about establishing some geometric integral version of this correspondence in the case of weight two and trivial character. Statement (A) for weight two and trivial character was proved first by Ribet in [34] and more generally by Diamond for $2 \leq k \leq p+1$ and arbitrary character in [13]. Note that statement (B) for weight two, trivial character and p^2 not dividing N, is the result of Mazur and Ribet above. A crucial point in their proof is that ρ has multiplicity one in the p-torsion of the jacobian $J_0(N)$, in some sense (see Section 3.3). A semi-simplicity result in [2] made it possible for Ribet to prove statement (B) for weight two and trivial character, but without the condition that p^2 does not divide N (see [35] and [32]). In [15] it was shown, using work of Gross [17] and of Coleman and Voloch [7], that one can always adapt the weight, in the following sense: if $p \neq 2$ and ρ is modular of some type $(N, k, \varepsilon)_{\overline{\mathbb{F}}_p}$ with N prime to p, then ρ is modular of type $(N, k(\rho), \varepsilon)_{\overline{\mathbb{F}}_p}$. The definition of $k(\rho)$ makes it clear that the mechanism behind the proof of this result was known to Serre; in Section 4 we will discuss a part of it. This mechanism includes the fact that if ρ is modular of some type $(N, k, \varepsilon)_{\overline{\mathbb{F}}_p}$, with N prime to p, then for some integer a, $\rho \otimes \chi_p^a$ is modular of type $(N, k', \varepsilon)_{\overline{\mathbb{F}}_p}$ with $2 \leq k' \leq p+1$, and ρ is modular of type $(Np^2, 2)_{\overline{\mathbb{Q}}_p}$. It follows from this that in order to prove statement (B), one may assume that the weight is two. This is used in [32] to show statement (B) for $p \geq 3$ and ρ with $\det(\rho) = \chi_p$; Ribet also remarks that he expects his proof to extend without difficulty to $\det(\rho)$ arbitrary. Finally, Diamond [12] proved statement (B) for $p \geq 3$, following [32]. Another proof of statement (B), not using the reduction to weight two, but extending the arguments of [33] to weights k between 2 and $p+1$, was suggested by Jordan and Livné in [21]. The multiplicity one result needed for that is proved in [16].

2 The cases we need

Special cases of Theorem 1.12 are used at three different places in the proof of the Shimura–Taniyama–Weil conjecture and of Fermat's Last Theorem. First of all, Ribet's proof that the Shimura–Taniyama–Weil conjecture implies Fermat's Last Theorem is a special case of Theorem 1.12. We briefly recall the situation. One supposes that Fermat's Last Theorem is not true. Then there exist a prime $p > 3$ and non-zero integers a, b and c that are pairwise relatively prime and satisfy $a^p + b^p + c^p = 0$. This leads, via a construction of Hellegouarch (see [19] and [20]), to a semi-stable elliptic curve E over $\mathbb{Q}$ that is usually called the Frey curve associated to (a^p, b^p, c^p). Let ρ_p be the representation $G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ given by the p-torsion of E. It follows from Mazur's work on isogenies between elliptic curves over $\mathbb{Q}$ (see [27] and

[28]) that ρ_p is irreducible. Moreover, E has the miraculous property that ρ_p is unramified away from $2p$ and that its ramification at 2 and p is very well-behaved: one has $N(\rho_p) = 2$, $k(\rho_p) = 2$ and $\varepsilon(\rho_p) = 1$; see [37, §4]. The conductor N of E is the product of all primes dividing abc; note that it is square free. If E is modular, i.e., if the Shimura–Taniyama–Weil conjecture is true for E, then ρ_p is modular of type $(N, 2, 1)_{\mathbb{Q}_p}$. So in this case it suffices to have Theorem 1.12 for ρ_p that are modular of some type $(N, 2, 1)_{\overline{\mathbb{Q}}_p}$ with N square free and with $p > 3$.

Let us now look where Theorem 1.12 intervenes in Wiles's proof of the Shimura–Taniyama–Weil conjecture for semi-stable elliptic curves. One has the following proposition, that was given by Oesterlé in the seminar held in Paris on the work of Wiles.

2.1 Proposition. *Let E be a semi-stable elliptic curve over $\mathbb{Q}$, p be a prime number and $\rho_p\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{F}_p)$ the representation obtained by choosing some basis of the p-torsion of E. Then either ρ_p is surjective, or it is reducible and its semi-simplification is isomorphic to the direct sum $1 \oplus \chi_p$ of the trivial character with the cyclotomic character.*

Proof. If $p > 5$ or if the image G of ρ_p has order divisible by p, this is part of Proposition 21 of [39]. Suppose that we are not in this case. We use the same arguments as Serre does. If ρ_p is reducible Serre's argument shows that its semi-simplification is $1 \oplus \chi_p$. So we suppose that ρ_p is irreducible and we have to derive a contradiction. The morphism $\det\colon G \to \mathbb{F}_p^*$ is surjective because $\det \circ \rho_p = \chi_p$. We have $p = 2$, 3 or 5 and ρ_p is unramified outside p. The image under ρ_p of inertia at p is a non-split Cartan subgroup (i.e., cyclic of order $p^2 - 1$) or "half" of a split Cartan (i.e., a conjugate of the subgroup of elements of the form $\left(\begin{smallmatrix} * & 0 \\ 0 & 1 \end{smallmatrix}\right)$). The first case arises if and only if E has good, supersingular reduction at p.

Suppose that $p = 2$. Then G has order 3 since ρ_p is irreducible. It follows that $\mathbb{Q}$ has a Galois extension of degree 3 which is ramified only at 2; there is no such extension. This finishes the proof for $p = 2$.

Suppose that $p = 3$. Then G has order dividing 16. The absolute irreducibility of ρ_p implies that G has order 8 or 16. But then G has a quotient of order 4, leading to a Galois extension of $\mathbb{Q}$ of degree 4 which is unramified outside 3; again, such an extension does not exist.

Suppose that $p = 5$. By §2.6 of [39], G is contained in the normalizer of a Cartan subgroup or the image $\overline{G}$ of G in $\mathrm{GL}_2(\mathbb{F}_p)/\mathbb{F}_p^*$ is isomorphic to A_4, S_4 or A_5. Since G is of order prime to 5, we cannot have $\overline{G}$ isomorphic to A_5. If $\overline{G}$ is isomorphic to A_4 or S_4 we get a Galois extension of degree 3 of $\mathbb{Q}$ or of $\mathbb{Q}(\sqrt{5})$ which is unramified outside 5, and such extensions do not exist. Hence G is contained in the normalizer N of a Cartan subgroup C. The absolute irreducibility of ρ_p implies that G is not contained in C. This gives us a degree 2 extension K of $\mathbb{Q}$. We claim that K is unramified, which gives us the desired contradiction. The extension K is clearly unramified outside p, so we have to show that $\rho_p(I_p)$ is contained in C. If $\rho_p(I_p)$ is a non-split Cartan subgroup then this Cartan subgroup has to be C (look at the action on the projective line), hence we get what we need. If $\rho_p(I_p)$ is "half" of a split Cartan subgroup, then again $\rho_p(I_p)$ has to be contained in C because of its action on the projective line. □

Let E be a semi-stable elliptic curve over $\mathbb{Q}$. If for a prime number p the representation ρ_p is not surjective, then E admits a $\mathbb{Q}$-rational isogeny of degree p. Hence, if ρ_3 and ρ_5 are both not sujective, E admits a $\mathbb{Q}$-rational isogeny of degree 15, hence defines a $\mathbb{Q}$-rational non-cuspidal point of $X_0(15)$. According to the tables in [1], $X_0(15)$ is an elliptic curve with exactly eight rational points, four of which are cusps. The four non-cuspidal points correspond (up to twist) to elliptic curves of conductor 50 of which one can easily see that no twist of them is semi-stable. So we conclude that at least one among ρ_3 and ρ_5 is surjective.

Suppose that ρ_3 is surjective. Results of Langlands [25] and Tunnell [43] show that ρ_3 is modular of type $(3^m N(\rho_3)^2, 1)_{\mathbb{Q}_3}$, for some $m \geq 0$. This will be discussed in Section 4. In that same section, we will show that ρ_3 is then modular of type $(N(\rho_3), 2, 1)_{\overline{\mathbb{Q}}_3}$ if ρ_3 is finite at 3, and of type $(3N(\rho_3), 2, 1)_{\overline{\mathbb{Q}}_3}$ if ρ_3 is not finite at 3. We will not need the results on companion forms, and we can use an older version of Carayol's Lemma dating back to Mazur's [28]. Carayol's reductions in [4] imply that ρ_3 is modular of type $(N(\rho_3), 2, 1)_{\overline{\mathbb{Q}}_3}$ if ρ_3 is finite at 3, and of type $(3N(\rho_3), 2, 1)_{\overline{\mathbb{Q}}_3}$ if ρ_3 is not finite at 3. Starting at this point, Wiles and Taylor prove (see [44] and [42]) that all deformations of ρ_3 that are ramified at only finitely many primes, that are "semi-stable" at 3 and whose determinant is the cyclotomic character are modular. Hence E is modular. (Let us note that the restriction of ρ_3 to $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q}(\sqrt{-3}))$ is absolutely irreducible since its image is $\mathrm{SL}_2(\mathbb{F}_3)$.) So in this case we don't need Ribet's part of the proof of Theorem 1.12.

Suppose that ρ_3 is not surjective. Then ρ_5 is surjective. In this case, Wiles [44] shows that there exists an elliptic curve E' over $\mathbb{Q}$ such that its representation ρ'_3 is irreducible and with ρ'_5 isomorphic to ρ_5 (this follows from the fact that $X(5)_{\overline{\mathbb{Q}}}$ is a disjoint union of four projective lines). The curve E' is semi-stable since this can be read off from ρ'_5, even at the prime 5 itself. The previous argument shows that E' is modular, hence ρ_5 is modular of type $(N', 2, 1)_{\mathbb{Q}_5}$, where N' is the conductor of E'. Note that N' is square free. Ribet's part of the proof of Theorem 1.12 shows that ρ_5 is modular of type $(N(\rho_5), 2, 1)_{\overline{\mathbb{Q}}_5}$ if ρ_5 is finite at 5, and of type $(5N(\rho_5), 2, 1)_{\overline{\mathbb{Q}}_5}$ if ρ_5 is not finite at 5. Wiles and Taylor ([44] and [42]) show that all deformations of ρ_5 that are ramified at only finitely many primes, that are "semi-stable" at 5 and whose determinant is the cyclotomic character are modular. Hence E is modular.

The final conclusion of this section is the following. We need Ribet's part of the proof of Theorem 1.12 in the case where $p > 3$ and ρ is modular of some type $(N, 2, 1)_{\overline{\mathbb{Q}}_p}$ with N square free. We also need Carayol's reductions for $p = 3$ and ρ modular of type $(N^2, 2, 1)_{\overline{\mathbb{Q}}_3}$ with N square free. It should be noted that Carayol's reductions use Mazur's result, so in the next section we will explain that result for $p \geq 3$.

3 Weight two, trivial character and square free level

The aim of this section is to describe a proof of the following theorem.

3.1 Theorem. *Let $p \geq 3$ be prime. Let $\rho\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ be irreducible and modular of some type $(N, 2, 1)_{\overline{\mathbb{Q}}_p}$ with N square free. If ρ is finite at p it is modular of type $(N(\rho), 2, 1)_{\overline{\mathbb{Q}}_p}$. If ρ is not finite at p it is modular of type $(pN(\rho), 2, 1)_{\overline{\mathbb{Q}}_p}$.*

The proof of this theorem has two parts. The first part, due to Mazur, deals with the primes l dividing $N/N(\rho)$ that are not congruent to 1 modulo p. The second part, due to Ribet, deals with an arbitrary $l \neq p$ at the cost of introducing a prime q in the level that one can get rid of by the first part. These two results will be described in the following two sections. For a detailed proof of in fact more general results we refer to [32, §6–§8]; our aim is just to give a good idea of what happens in [32].

3.2 Mazur's result

3.2.1 Theorem. *Let $p \geq 3$ be prime. Let $\rho: G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ be irreducible and modular of some type $(N, 2, 1)_{\overline{\mathbb{Q}}_p}$. Suppose that l is a prime not congruent to 1 mod p, that l divides N but l^2 does not, that ρ is unramified at l if $l \neq p$ and that ρ is finite at p if $l = p$. Then ρ is modular of type $(N/l, 2, 1)_{\overline{\mathbb{Q}}_p}$.*

The proof of this result is by contradiction: let $M := N/l$ and suppose that ρ is not modular of type $(M, 2, 1)_{\overline{\mathbb{Q}}_p}$. Let $X_0(N)$ denote the model over the localization $\mathbb{Z}_{(l)}$ of $\mathbb{Z}$ at l which is constructed in [10] as a coarse moduli space (see also [24]). Let $J_0(N)_{\mathbb{Q}}$ be the jacobian of $X_0(N)_{\mathbb{Q}}$, and let $J_0(N)$ be its Néron model over $\mathbb{Z}_{(l)}$. Then $J_0(N)$ represents in fact the degree zero part of the relative Picard functor of $X_0(N)$ over $\mathbb{Z}_{(l)}$ by [30]; this implies that one can describe $J_0(N)$ in terms of $X_0(N)$. By [10, V–VI], the fibre $X_0(N)_{\mathbb{F}_l}$ of $X_0(N)$ over $\mathbb{F}_l$ is the union of two copies of the smooth curve $X_0(M)_{\mathbb{F}_l}$, which intersect transversally at the supersingular points. This gives the following "devissage" of $J_0(N)_{\mathbb{F}_l}$. One has the connected component of the identity element $J_0(N)^0_{\mathbb{F}_l}$ of $J_0(N)_{\mathbb{F}_l}$. The quotient $J_0(N)_{\mathbb{F}_l}/J_0(N)^0_{\mathbb{F}_l}$ is a finite etale group scheme over $\mathbb{F}_l$, which is in fact a constant group scheme that we will denote by $\Phi_0(N)_{\mathbb{F}_l}$. By construction we have an exact sequence:

$$0 \to J_0(N)^0_{\mathbb{F}_l} \to J_0(N)_{\mathbb{F}_l} \to \Phi_0(N)_{\mathbb{F}_l} \to 0. \tag{3.2.2}$$

The normalization map $X_0(M)_{\mathbb{F}_l} \coprod X_0(M)_{\mathbb{F}_l} \to X_0(N)_{\mathbb{F}_l}$ induces another short exact sequence:

$$0 \to T_0(N)_{\mathbb{F}_l} \to J_0(N)^0_{\mathbb{F}_l} \to J_0(M)^2_{\mathbb{F}_l} \to 0, \tag{3.2.3}$$

where $T_0(N)_{\mathbb{F}_l}$ is a torus whose character group $\mathrm{Hom}_{\overline{\mathbb{F}}_l}(T_0(N)_{\overline{\mathbb{F}}_l}, \mathrm{G}_{m\overline{\mathbb{F}}_l})$ can be identified with the group of degree zero divisors on $X_0(M)_{\overline{\mathbb{F}}_l}$ with support in the supersingular points. Let $\mathbb{T}_0(N)$ be the Hecke algebra of level N: it is the subring of $\mathrm{End}(J_0(N)_{\mathbb{Q}})$ generated by the Hecke operators T_n, $n \geq 1$. By the Néron property of $J_0(N)$, $\mathbb{T}_0(N)$ acts on it. This action can be described explicitly by correspondences over $\mathbb{Z}_{(l)}$. The short exact sequences (3.2.2) and (3.2.3) are compatible with the action of $\mathbb{T}_0(N)$.

Let us now look what all this has to do with our representation ρ. The fact that ρ is modular of type $(N, 2, 1)_{\overline{\mathbb{Q}}_p}$ means that there exist a maximal ideal m of $\mathbb{T}_0(N)$, an embedding of $k := \mathbb{T}_0(N)/m$ into $\overline{\mathbb{F}}_p$, an integer $d \geq 1$ and a 2-dimensional k-vector space V with an action by $G_{\mathbb{Q}}$ with $\overline{\mathbb{F}}_p \otimes_k V$ giving ρ, such that $J_0(N)(\overline{\mathbb{Q}})[m]$ is isomorphic to the direct sum V^d of d copies of V as $k[G_{\mathbb{Q}}]$-modules. (Here $J_0(N)(\overline{\mathbb{Q}})[m]$ denotes the kernel of m, i.e., the elements x such that $tx = 0$ for all t in m.) The fact that $J_0(N)(\overline{\mathbb{Q}})[m]$ is semi-simple is proved in [2].

We will now first treat the case where $l \neq p$, which is technically simpler than the case $l = p$. So we suppose that $l \neq p$. Then ρ is unramified at l by hypothesis, and there is a unique finite etale k-vector space scheme W over $\mathbb{Z}_{(l)}$ such that $V \cong W(\overline{\mathbb{Q}})$ as $k[G_{\mathbb{Q}}]$-modules. We choose an injection of V into $J_0(N)(\overline{\mathbb{Q}})[m]$. This gives us an injection of $W_{\mathbb{Q}}$ into $J_0(N)_{\mathbb{Q}}$. The Néron property of $J_0(N)$ implies that this injection extends to a morphism $W \to J_0(N)$, which must be injective since $J_0(N)[p]$ is etale. Consider the image of $W_{\mathbb{F}_l}$ in $\Phi_0(N)_{\mathbb{F}_l}$ under (3.2.2). It was proved in [33] that the action of $\mathbb{T}_0(N)$ on $\Phi_0(N)_{\mathbb{F}_l}$ is "Eisenstein", in the sense that for q prime to N the operator T_q acts as multiplication by $q+1$. Since ρ is irreducible, it follows that the image of $W_{\mathbb{F}_l}$ in $\Phi_0(N)_{\mathbb{F}_l}$ is zero. Hence $W_{\mathbb{F}_l}$ lands in $J_0(N)^0_{\mathbb{F}_l}$. Since we suppose that ρ is not modular of type $(M,2,1)_{\overline{\mathbb{Q}}_p}$, the image of $W_{\mathbb{F}_l}$ in $J_0(M)^2_{\mathbb{F}_l}$ is zero. So $W_{\mathbb{F}_l}$ lands in the torus $T_0(N)_{\mathbb{F}_l}$. This has strong consequences for the Frobenius element $\rho(\mathrm{Frob}_l)$. Namely, Frob_l acts on $T_0(N)_{\mathbb{F}_l}(\overline{\mathbb{F}}_l)$ simultaneously as lT_l and as $-lw_l$, with w_l the Atkin–Lehner involution of level l. This implies that $\rho(\mathrm{Frob}_l)$ is in k^* and it follows that $l = \det(\rho(\mathrm{Frob}_l)) = \rho(\mathrm{Frob}_l^2) = l^2$. This contradicts the assumption that l is not congruent to 1 modulo p.

Let us now assume that $l = p$. We have the information that ρ is finite at p. Let $W_{\mathbb{Q}}$ be finite k-vector space scheme over $\mathbb{Q}$ such that $W_{\mathbb{Q}}(\overline{\mathbb{Q}})$ gives V. Then W can be extended to a finite flat k-vector space scheme over $\mathbb{Z}_{(p)}$. Such an extension is unique by results of Raynaud in [31] (here we use that $p \neq 2$). We choose an injection of V into $J_0(N)(\overline{\mathbb{Q}})[m]$. This gives an injection of $W_{\mathbb{Q}}$ into $J_0(N)_{\mathbb{Q}}$. Ribet showed in [33, Lemma 6.2] that this injection extends to an injection of W into $J_0(N)$. Note that this is not just a consequence of the Néron property of $J_0(N)$, since W is not smooth as $\det(\rho) = \chi_p$. Once one has this result the rest of the proof of Theorem 3.2.1 proceeds as in the case $l \neq p$. So it remains to explain the proof of [33, Lemma 6.2]. We may replace $\mathbb{Z}_{(p)}$ by its completion $\mathbb{Z}_p$. Then W sits in a short exact sequence:

$$0 \to W^0 \to W \to W^{\mathrm{et}} \to 0, \tag{3.2.4}$$

where W^0 denotes the connected component of the zero section of W, and where W^{et} is the largest finite etale quotient of W. Grothendieck proved, see [18, 7, IX, §11.6], that the action of $G_{\mathbb{Q}_p}$ on $J_0(N)(\overline{\mathbb{Q}}_p)/J_0(N)^0(\overline{\mathbb{Z}}_p)$ is unramified. This implies that the injection of $W^0_{\mathbb{Q}_p}$ into $J_0(N)_{\mathbb{Q}_p}$ extends to a morphism of W^0 into $J_0(N)^0$, which is a closed immersion by [31]. Then one can finish by applying the Néron property to the amalgamated sum $(W \oplus J_0(N)^0)/W^0$ (this trick is due to Grothendieck, see [18, 7, IX, 5.9.2]).

3.3 Ribet's result

3.3.1 Theorem. *Let $p \geq 3$ be prime. Let $\rho\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ be irreducible and modular of some type $(N,2,1)_{\overline{\mathbb{Q}}_p}$. Suppose that $l \neq p$ is a prime, that l divides N but l^2 does not, and that ρ is unramified at l. Then there exists a prime number q, not dividing N and congruent to -1 mod p, such that ρ is modular of type $(qN/l,2,1)_{\overline{\mathbb{Q}}_p}$.*

We define $M := N/l$. Let q be a prime number not dividing Np with the property that $\rho(\mathrm{Frob}_q) = \rho(c)$, with c a complex conjugation. This implies that q is congruent to -1 modulo p.

In what follows we will consider the modular curves $X_0(Mlq)_{\mathbb{Q}}$, $X_0(Ml)_{\mathbb{Q}}$, $X_0(Mq)_{\mathbb{Q}}$, and a certain Shimura curve $C_{\mathbb{Q}}$ that we will now define (see [33, §4] for more details). Let $B_{\mathbb{Q}}$ be a quaternion algebra over $\mathbb{Q}$ of discrminant lq and let B be a maximal order in it. Then $C_{\mathbb{Q}}$ is the Shimura curve associated to $B_{\mathbb{Q}}$ of level $\Gamma_0(M)$. More precisely, this means that $C_{\mathbb{Q}}$ is the coarse moduli scheme for objects $(A/S, \alpha, G)$, with S a $\mathbb{Q}$-scheme, A/S an abelian scheme of relative dimension two, with $\alpha\colon B \to \mathrm{End}_S(A)$ a morphism of rings and $G \subset A$ a finite flat closed subgroup scheme of rank M^2, which is killed by M and which is stable for the action by B. We define $J_{\mathbb{Q}}$ to be the jacobian of $C_{\mathbb{Q}}$. From the moduli interpretation of $C_{\mathbb{Q}}$ it is clear that one can define Hecke correspondences on $C_{\mathbb{Q}}$, inducing endomorphisms T_n, $n \geq 1$, of $J_{\mathbb{Q}}$. Let $\mathbb{T}$ be the subring of $\mathrm{End}_{\mathbb{Q}}(J_{\mathbb{Q}})$ generated by these T_n.

We have the usual pairs of degeneracy morphisms from $X_0(Mlq)_{\mathbb{Q}}$ to $X_0(Ml)_{\mathbb{Q}}$ and to $X_0(Mq)_{\mathbb{Q}}$, inducing a morphism $J_0(Mlq)_{\mathbb{Q}} \to J_0(Ml)^2_{\mathbb{Q}} \times J_0(Mq)^2_{\mathbb{Q}}$. The lq-new subvariety $J_0(Mlq)^{l,q-\mathrm{new}}_{\mathbb{Q}}$ of $J_0(Mlq)_{\mathbb{Q}}$ is defined as the connected component of the identity element of the kernel of this last morphism. One knows that $J_{\mathbb{Q}}$ is isogeneous to $J_0(Mlq)^{l,q-\mathrm{new}}_{\mathbb{Q}}$ (this results from trace formula calculations by Eichler, Shimizu, Jacquet–Langlands and Faltings's isogeny theorem). Ribet has given a more precise version of this in terms of the character groups of the torus parts of the reductions mod l and q of the jacobians of the curves under consideration. For G a commutative algebraic group over a field k, let $X(G) := \mathrm{Hom}_{\overline{k}}(T_{\overline{k}}, \mathrm{G}_{m\overline{k}})$ be the character group of the maximal torus $T_{\overline{k}}$ of $G_{\overline{k}}$. Then Ribet constructed a short exact sequence:

$$0 \to X(J_{\mathbb{F}_q}) \to X(J_0(Mlq)_{\mathbb{F}_l}) \to X(J_0(Ml)^2_{\mathbb{F}_l}) \to 0 \tag{3.3.2}$$

which is Hecke-equivariant in the sense that for each $n \geq 1$, the element T_n in $\mathbb{T}_0(Mlq)$ induces the element T_n of $\mathbb{T}$ on $X(J_{\mathbb{F}_q})$. The induced action of T_n on $X(J_0(Ml)^2_{\mathbb{F}_l})$ can be described in terms of a two by two matrix with coefficients in $\mathbb{T}_0(Ml)$. Since l and q play symmetric roles, we also have the following exact sequence:

$$0 \to X(J_{\mathbb{F}_l}) \to X(J_0(Mlq)_{\mathbb{F}_q}) \to X(J_0(Mq)^2_{\mathbb{F}_q}) \to 0. \tag{3.3.3}$$

To construct these sequences, Ribet relies heavily on work of Cerednik, Drinfeld and Jordan–Livné concerning the q-adic uniformization of $C_{\mathbb{Q}_q}$. A detailed account of this uniformization can be found in [3]. An amazing feature of these sequences is that they compare character groups of tori over fields of distinct characteristics. Since $J^0_{\mathbb{F}_q}$ is its maximal torus, it follows from (3.3.2) that $\mathbb{T}_0(Mlq)$ acts on $J_{\mathbb{Q}}$ via a (necessarily unique) morphism of rings $\mathbb{T}_0(Mlq) \to \mathbb{T}$ that sends T_n to T_n.

Let η_q be the element $T_q^2 - 1$ of $\mathbb{T}_0(Mlq)$, and let $\Phi_{\mathbb{F}_q}$ be the group of connected components of $J_{\mathbb{F}_q}$. For M a finite abelian group, let $M^* := \mathrm{Hom}_{\mathbb{Z}}(M, \mathbb{Q}/\mathbb{Z})$ be its Pontrjagin dual. Theorem 4.3 of [33] asserts that there is a Hecke equivariant exact sequence:

$$0 \to K_l \to X(J_0(Ml)^2_{\mathbb{F}_l})/\eta_q X(J_0(Ml)^2_{\mathbb{F}_l}) \to \Phi^*_{\mathbb{F}_q} \to C_q \to 0, \tag{3.3.4}$$

with K_l and C_q "Eisenstein" in the sense we saw in the previous section. Likewise, one has an exact sequence:

$$0 \to K_q \to X(J_0(Mq)^2_{\mathbb{F}_q})/\eta_q X(J_0(Mq)^2_{\mathbb{F}_q}) \to \Phi^*_{\mathbb{F}_l} \to C_l \to 0 \tag{3.3.5}$$

with K_q and C_l "Eisenstein". Since ρ is modular of type $(Ml,2,1)_{\overline{\mathbb{Q}}_p}$, ρ arises from a maximal ideal of $\mathbb{T}_0(Ml)$, in the way we have seen in Section 3.2. It is not hard to see that then ρ also arises from a maximal ideal m of $\mathbb{T}_0(Mlq)$ (note that we do not claim that ρ arises from a newform whose level is divisible by q). More precisely, we have a maximal ideal m of $\mathbb{T}_0(Mlq)$, an embedding of $k := \mathbb{T}_0(Mlq)/m$ into $\overline{\mathbb{F}}_p$, a two-dimensional k-vector space V with an action by $G_{\mathbb{Q}}$ with $\overline{\mathbb{F}}_p \otimes_k V$ giving ρ, such that $J_0(Mlq)(\overline{\mathbb{Q}})[m]$ is isomorphic to V^λ for some positive integer λ (this λ is called the multiplicity at m of ρ in $J_0(Mlq)$). Let μ be the multiplicity at m of ρ in $J_{\mathbb{Q}}$: $J_{\mathbb{Q}}(\overline{\mathbb{Q}})[m] \cong V^\mu$ (it follows again from [2] that $J_{\mathbb{Q}}(\overline{\mathbb{Q}})[m]$ is semi-simple). It is clear that $\mu \geq 0$.

From now on we suppose that ρ is not modular of type $(Mq,2,1)_{\overline{\mathbb{Q}}_p}$. From this assumption and the exact sequences above Ribet then derives that $2\mu \leq \lambda$ and that $2\lambda \leq 2\mu$, which gives a contradiction because we know that $\lambda > 0$. So let us describe the arguments of Ribet. One starts by localizing the exact sequence (3.3.5) at m; this shows that $\Phi_{\mathbb{F}_l,m}$ is zero, because m is not in the support of $X(J_0(Mq)^2_{\mathbb{F}_q})$. This implies that $J^0_{\mathbb{F}_l}(\overline{\mathbb{F}}_l)[m]$ is isomorphic to V^μ (as k-vector spaces), so that:

(3.3.6) $$\dim_k(X(J_{\mathbb{F}_l}) \otimes_{\mathbb{T}_0(Mlq)} k) = 2\mu$$

The exact sequence (3.3.3) shows that:

(3.3.7) $$\dim_k(X(J_0(Mlq)_{\mathbb{F}_q}) \otimes_{\mathbb{T}_0(Mlq)} k) = \dim_k(X(J_{\mathbb{F}_l}) \otimes_{\mathbb{T}_0(Mlq)} k).$$

Next we have the following exact sequence, obtained by replacing N by Mlq in (3.2.2):

(3.3.8) $$0 \to J_0(Mlq)^0_{\mathbb{F}_l} \to J_0(Mlq)_{\mathbb{F}_l} \to \Phi_0(Mlq)_{\mathbb{F}_l} \to 0.$$

Since $\Phi_0(Mlq)_{\mathbb{F}_l}$ is "Eisenstein", it follows that $J_0(Mlq)(\overline{\mathbb{Q}})[m]$ specializes into $J_0(Mlq)^0_{\mathbb{F}_l}(\overline{\mathbb{F}}_l)$. As in (3.2.3), the normalization of $X_0(Mlq)_{\mathbb{F}_l}$ induces a short exact sequence:

(3.3.9) $$0 \to T_0(Mlq)_{\mathbb{F}_l} \to J_0(Mlq)^0_{\mathbb{F}_l} \to J_0(Mq)^2_{\mathbb{F}_l} \to 0.$$

Since ρ is not modular of type $(Mq,2,1)_{\overline{\mathbb{Q}}_p}$, it follows that:

(3.3.10) $$\dim_k(X(J_0(Mlq)_{\mathbb{F}_l}) \otimes_{\mathbb{T}_0(Mlq)} k) = 2\lambda.$$

Ribet shows that the Frobenius endomorphism of $J^0_{\mathbb{F}_q}$ is equal to qT_q. It follows from this that $\rho(\mathrm{Frob}_q)$ acts as a scalar (i.e., an element of k) on $J^0_{\mathbb{F}_q}(\overline{\mathbb{F}}_q)[m]$. But by the choice of q, $\rho(\mathrm{Frob}_q)$ is in the conjugacy class of $\left(\begin{smallmatrix}1 & 0\\ 0 & -1\end{smallmatrix}\right)$. It follows that:

(3.3.11) $$\dim_k(X(J_{\mathbb{F}_q}) \otimes_{\mathbb{T}_0(Mlq)} k) \leq \mu.$$

The same argument applied to the maximal torus $T_0(Mlq)_{\mathbb{F}_q}$ in $J_0(Mlq)_{\mathbb{F}_q}$ gives:

(3.3.12) $$\dim_k(X(J_0(Mlq)_{\mathbb{F}_q}) \otimes_{\mathbb{T}_0(Mlq)} k) \leq \lambda.$$

3.3.13 Lemma. *We have* $\dim_k(X(J_0(Ml)^2_{\mathbb{F}_l}) \otimes_{\mathbb{T}_0(Mlq)} k) \leq \mu$.

Proof. First note that η_q is in m. The exact sequence (3.3.4) shows that:

$$\dim_k(X(J_0(Ml)^2_{\mathbb{F}_l})\otimes_{\mathbb{T}_0(Mlq)} k) = \dim_k(\Phi^*_{\mathbb{F}_q})\otimes_{\mathbb{T}_0(Mlq)} k = \dim_k(\Phi_{\mathbb{F}_q}[m]).$$

Grothendieck's description [18, 7, IX, §11] of $\Phi_{\mathbb{F}_q}$ gives an exact sequence:

$$0 \to X(J'_{\mathbb{F}_q}) \to X(J_{\mathbb{F}_q})^\vee \to \Phi_{\mathbb{F}_q} \to 0, \tag{3.3.14}$$

where for M a $\mathbb{Z}$-module $M^\vee$ denotes its $\mathbb{Z}$-dual, and where $J'_{\mathbb{F}_q}$ is just $J_{\mathbb{F}_q}$, but with the dual $\mathbb{T}_0(Mlq)$-action: t in $\mathbb{T}_0(Mlq)$ acts as t^*, the dual of the endomorphism given by t (this uses the natural autoduality of jacobians). This makes the sequence (3.3.14) Hecke equivariant. Consider multiplication by p on the exact sequence (3.3.14). Applying the snake Lemma and then taking kernels for m gives an injection of $\Phi_{\mathbb{F}_q}[m]$ into $(X(J'_{\mathbb{F}_q})\otimes_{\mathbb{Z}}\mathbb{F}_p)[m]$. Ribet shows, using more results from [18, 7, IX, §11] and a description of the action of Frob_q on $\mathrm{Hom}(X(J_{\mathbb{F}_q}),\mu_p)$ and on $X(J'_{\mathbb{F}_q})$, that one has an exact sequence:

$$0 \to \mathrm{Hom}(X(J_{\mathbb{F}_q}),\mu_p)[m] \to J(\overline{\mathbb{Q}})[m] \to (X(J'_{\mathbb{F}_q})\otimes\mathbb{F}_p)[m] \to 0. \tag{3.3.15}$$

The middle term in this sequence has dimension 2μ by the definition of μ. The fact that $\rho(\mathrm{Frob}_q)$ has the two distinct eigenvalues 1 and -1 then implies that the other two terms both have dimension μ. $\square$

We can now finish our description of the proof of Theorem 3.3.1. The sequence (3.3.2) gives:

$$\dim_k(X(J_0(Mlq)_{\mathbb{F}_l})\otimes_{\mathbb{T}_0(Mlq)} k) \le \dim_k(X(J_{\mathbb{F}_q})\otimes_{\mathbb{T}_0(Mlq)} k) + \dim_k(X(J_0(Ml)^2_{\mathbb{F}_l})\otimes_{\mathbb{T}_0(Mlq)} k). \tag{3.3.16}$$

Combining (3.3.6), (3.3.7) and (3.3.12) gives $2\mu \le \lambda$. On the other hand, combining (3.3.10), (3.3.16), (3.3.11) and Lemma 3.3.13 gives $2\lambda \le 2\mu$.

4 Dealing with the Langlands–Tunnell form

Let E be a semi-stable elliptic curve over $\mathbb{Q}$, with ρ_3 irreducible. It follows, Proposition 2.1, that $\rho_3 \colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{F}_3)$ is surjective. The aim of this section is to show that ρ_3 is modular of type $(N(\rho_3),2,1)_{\overline{\mathbb{Q}}_3}$ if ρ_3 is finite at 3 and of type $(3N(\rho_3),2,1)_{\overline{\mathbb{Q}}_3}$ if ρ_3 is not finite at 3.

4.1 The theorem of Langlands and Tunnell

The first step is to show that ρ_3 is modular, by lifting ρ_3 to a continuous representation of $G_{\mathbb{Q}}$ with values in $\mathrm{GL}_2(\mathbb{C})$, i.e., a continuous representation $\rho\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{C})$ (for the discrete topology on $\mathbb{C}$) such that ρ_3 is a reduction modulo 3 of ρ, and applying a theorem of Langlands and Tunnell to ρ. To do this, we compose ρ_3 with a two-dimensional complex representation of $\mathrm{GL}_2(\mathbb{F}_3)$. The irreducible representations of the groups $\mathrm{GL}_2(k)$, where k is a finite field, are well-known, see for example [6]. If q denotes the cardinality of k, then the irreducible representations have dimensions 1, $q-1$, q or $q+1$. Those of dimension one factor via the determinant. Hence for $q > 3$ there are no faithful two-dimensional representations. In the case of interest for us, i.e., $q = 3$, there are three irreducible two-dimensional representations. One of them can be realized in $\mathrm{GL}_2(\mathbb{Z})$, but it is not faithful (it factors through the quotient S_3). The other two are Galois conjugates and can be realized in $\mathrm{GL}_2(\mathbb{Z}[\sqrt{-2}])$. Let $\pi\colon \mathrm{GL}_2(\mathbb{F}_3) \to \mathrm{GL}_2(\mathbb{Z}[\sqrt{-2}])$ be one of these, and let $\rho := \pi\circ\rho_3\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}[\sqrt{-2}])$. From the character table given in [6], one sees immediately that π is faithful (our π is one of the $\pi(\Lambda)$ with Λ of order eight), and that the restriction of π to the subgroup $\left(\begin{smallmatrix} * & 0 \\ 0 & 1 \end{smallmatrix}\right)$ is the regular representation. It follows that ρ is odd and that $\det\circ\rho$ is the composition of $\chi_3\colon G_{\mathbb{Q}} \to \mathbb{F}_3^*$ with the embedding of $\mathbb{F}_3^*$ into $\mathbb{C}^*$. Looking again at the character table shows that reduction of ρ modulo one of the two maximal ideals containing 3 gives ρ_3. According to [37, §5.3] the conductor $N(\rho)$ of ρ is $3^m N(\rho_3)^2$, for some $m \geq 0$. This is explained as follows. Let $l \neq 3$ be a prime. If ρ_3 is unramified at l then ρ is unramified too. Suppose that ρ_3 is ramified at l. Since E is semi-stable, $\rho_3(I_l)$ is a conjugate of the subgroup $\left(\begin{smallmatrix} 1 & * \\ 0 & 1 \end{smallmatrix}\right)$. From the character table one sees that $\pi\left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ has trace -1, hence its eigenvalues are the two roots of unity of order 3. This means that the dimension of the space of I_l-invariants for ρ_3 is one, and that it is zero for ρ. Hence the exponent of l in $N(\rho)$ is twice that of l in $N(\rho_3)$. It will be of use for us to determine the exponent m of 3 in $N(\rho)$ in the case where $\rho_3|_{I_3}$ is isomorphic to the direct sum $\chi_3 \oplus 1$. In this case, in which the ramification is tame, (1.6) shows that $m = 1$. One can see easily that in all other cases $m > 1$.

A deep result of Langlands and Tunnell ([25] and [43]) says that ρ is modular of type $(N(\rho), 1, \det(\rho))_{\mathbb{Z}[\sqrt{-2}]}$, i.e., there exists a cuspidal eigenform f of level $N(\rho)$, of weight 1 and with character $\det(\rho)$ such that $\rho \cong \rho_f$, where $\rho_f\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{C})$ is the representation associated to f by Deligne–Serre in [11] (see Gelbart's chapter). At this point we know that ρ_3 arises from the modular form f.

4.2 Some purely mod 3 arguments

The second step is now to show that ρ_3 is modular of type $(N(\rho_3)^2, k(\rho_3), 1)_{\overline{\mathbb{F}}_3}$. First we want to see that ρ is modular of type $(N(\rho_3)^2, k)_{\overline{\mathbb{F}}_3}$, for some $k \geq 1$. This is folklore. Ribet has given a proof of this in [32, §2] for $p > 2$. We remark that there are (unpublished) conceptual short proofs, valid for all p, using either the results of Katz and Mazur in [24] on the reduction mod p of modular curves (the idea is just to restrict the form to a suitable irreducible component of this reduction), or the representation theory of $\mathrm{GL}_2(\mathbb{F}_p)$ on $\mathbb{F}_p$-vector spaces and some group cohomology or etale cohomology.

We can now assume that ρ_3 is modular of type $(N(\rho_3)^2, k)_{\overline{\mathbb{F}}_3}$, for some $k \geq 1$. We want to show that it is modular of type $(N(\rho_3)^2, k(\rho_3))_{\overline{\mathbb{F}}_3}$, i.e., we want to adjust the weight. Theorem 4.5 of [15] says that this adjustment of the weight can always be done; in [15] one can find a detailed proof. We will now describe only the main features of that proof, and indicate a simplification, due to Taylor, in the case of our ρ_3.

The proof of Theorem 4.5 of [15] runs as follows. Let p be a prime number and $\rho\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_p)$ be irreducible, continuous and modular of some level N which is prime to p. Then one knows that there exist a in $\mathbb{Z}/(p-1)\mathbb{Z}$ and k in $\{2,\ldots,p+1\}$ such that $\rho \cong \rho_f \otimes \chi_p^a$ for some eigenform f of type $(N,k)_{\overline{\mathbb{F}}_p}$. Since f has low weight, i.e., $k \leq p+1$, one knows that $\rho_f|_{I_p}$ is an extension of the trivial character by χ_p^{k-1} if f is ordinary (i.e., $T_p f \neq 0$) and that it is the direct sum of ψ^{k-1} and ψ'^{k-1} if f is supersingular (i.e., $T_p f = 0$); here ψ and ψ' are the two fundamental characters of level two that occur in Definition 1.7. The first part of this result is due to Deligne (see [17] for a proof). The second part was first proved by Fontaine (see [15] for a proof). It follows that a and k are completely determined by $\rho|_{I_p}$, except if $\rho|_{I_p}$ is a direct sum of two distinct powers of χ_p or if it is a non-split extension of χ_p^a by χ_p^{a+1}. In the first case there are two candidates for a; in the second case one cannot decide between the values 2 and $p+1$ for k. In the second case Serre conjectured, and Mazur showed under the assumption that $p > 2$, that one can take $k = 2$ if and only if $\rho \otimes \chi_p^{-a}$ is finite at p. Recall that Mazur's argument was described in §3.2. In the first case, Serre conjectured that both candidates for a should work and called the two forms corresponding to the two values of a companions of each other. The fact that such companion forms exist was proved in [17], [7] and [16], except for some cases when $p = 2$. With these results, it remains to construct a form f' of type $(N,k')_{\overline{\mathbb{F}}_p}$, with k' as small as possible, such that $\rho \cong \rho_{f'}$. This last equality is equivalent to: $a'_l = l^a a_l$ for all primes l not dividing Np, where the a'_l and a_l are the eigenvalues of f' and f for T_l. There is a operation θ on modular forms over $\overline{\mathbb{F}}_p$ that has the effect of the derivation qd/dq on q-expansions. This derivation θ was first constructed, more or less by hand, for forms of level one by Swinnerton-Dyer and Serre in [41], and in general, in terms of the Gauss–Manin connection, by Katz in [23]. The properties proved about θ in [23] imply that one knows exactly how to construct f': one should take $f' := A^{-n}\theta^a(f)$ with n as large as possible (here A denotes the Hasse invariant), or the same expression with f replaced by its companion if it exists. The results in [23] imply that one knows the value of n.

Let us now look at our ρ_3. We want to show that we do not need the results on companion forms to show that ρ_3 is modular of type $(N(\rho_3)^2, k(\rho_3))$ (this is an observation of Taylor). We know that ρ_3 or $\rho_3 \otimes \chi_3$ arises from a form of weight k with $2 \leq k \leq 4$. If E has bad or good and ordinary reduction at 3, then $\rho_3|_{I_3}$ is an extension of the trivial character 1 by χ_3; $k(\rho_3) = 2$ if ρ_3 is finite at 3 (this is the case if and only if the exponent of 3 in the discriminant of E is a multiple of 3), otherwise $k(\rho_3) = 4$. If E has good and supersingular reduction at 3, then $\rho_3|_{I_3}$ is the direct sum of ψ and ψ', the two fundamental characters of level two.

Suppose first that E is good and supersingular at 3. Then $\rho_3 \otimes \chi_3$ cannot arise from a form of weight between 2 and 4, because its restriction to I_3 is not of the right form. Hence ρ_3 itself arises from an eigenform f of weight k with $2 \leq k \leq 4$. Because $\det(\rho_3) = \chi_3$, we must have $k = 2$ or $k = 4$. It remains to see that we can take $k = 2$. So suppose that f has weight 4. We claim that then $f = Ag$, where A is the Hasse invariant and g an eigenform of weight two. To prove this, it suffices to show that f vanishes at the supersingular points of the modular curve on which it lives. Now if f would not vanish at all the supersingular points, then the regular differential form $\omega(f)$ on $X_0(3N(\rho_3)^2)_{\overline{\mathbb{F}}_3}$ that will be constructed in the next section has a non-zero residue at some supersingular point. A study of the action of T_3 then shows that $T_3 f = \pm f$, but we have $T_3 f = 0$. The eigenform g has the same eigenvalues as f, hence

$\rho_3 \cong \rho_g$.

Suppose now that E has bad or good and ordinary reduction. Then $\rho_3|_{I_3}$ is an extension of 1 by χ_3. Let us first deal with the case where this extension is not split. Then $\rho_3 \otimes \chi_3$ cannot arise from a form of weight between 2 and 4, hence there is a form f of weight 2 or 4 such that $\rho_3 \cong \rho_f$. If $k(\rho_3) = 4$ we have $k = 4$ and there is nothing to prove. So suppose that $k(\rho_3) = 2$. If $k = 2$ there is nothing to prove, so suppose moreover that $k = 4$. Let $\omega(f)$ be the differential form on $X_0(3N(\rho_3)^2)_{\overline{\mathbb{F}}_3}$ as constructed in the next section. Since $\omega(f)$ can be lifted this shows that ρ_3 is modular of type $(3N(\rho_3)^2, 2, 1)_{\overline{\mathbb{Q}}_3}$. Since ρ_3 is finite at 3, Mazur's result (Theorem 3.2.1) shows that ρ_3 is modular of type $(N(\rho_3)^2, 2, 1)_{\overline{\mathbb{Q}}_3}$. It remains to deal with the split case. So suppose that $\rho_3|_{I_3} \cong \chi_3 \oplus 1$. Then we have seen that the weight one form f given by Langlands and Tunnell has level $3N(\rho_3)^2$. Then $fE_{1,\chi}$, with $\chi\colon \mathbb{F}_3^* \to \mathbb{C}^*$ non-trivial, shows that ρ_3 is modular of type $(3N(\rho_3)^2, 2, 1)_{\overline{\mathbb{Q}}_3}$. Theorem 3.2.1 shows that ρ_3 is modular of type $(N(\rho_3)^2, 2, 1)_{\overline{\mathbb{Q}}_3}$.

4.3 Differential forms and a version of Carayol's lemma

In the last section we have seen that ρ_3 is modular of type $(N(\rho_3)^2, k(\rho_3), 1)_{\overline{\mathbb{F}}_3}$. Now we want to deduce that ρ_3 is modular of type $(N(\rho_3)^2, 2, 1)_{\overline{\mathbb{Q}}_3}$ if ρ_3 is finite at 3 and of type $(3N(\rho_3)^2, 2, 1)_{\overline{\mathbb{Q}}_3}$ otherwise. It is well known that cuspidal modular forms of a given type $(N, 2, 1)_K$, where K is a field of characteristic zero, correspond via the Kodaira–Spencer isomorphism to differential forms on the modular curve $X_0(N)_K$, i.e, to sections of the sheaf Ω^1 on $X_0(N)_K$. This correspondence is compatible with the action of Hecke operators. We define $N := N(\rho_3)^2$ if ρ_3 is finite at 3 and $N := 3N(\rho_3)^2$ otherwise. We will show in fact that there exists a non-zero differential form ω on $X_0(N)_{\overline{\mathbb{Q}}_3}$ which is an eigenform for the Hecke algebra and whose associated Galois representation $\rho_\omega\colon G_{\mathbb{Q}} \to \mathrm{GL}_2(\overline{\mathbb{F}}_3)$ is isomorphic to ρ_3. We will construct this ω by first constructing it over $\overline{\mathbb{F}}_3$, which is in fact the hardest part (it involves a version of Carayol's lemma); a standard argument from commutative algebra shows that it can be lifted to an eigenform (see for example [17, §9]).

Let us first do this construction in the case where ρ_3 is finite at 3. So then N is prime to 3 and we have a cuspidal eigenform f of type $(N, 2, 1)_{\overline{\mathbb{F}}_3}$. We want to construct a differential form ω on $X_0(N)_{\overline{\mathbb{F}}_3}$ which is an eigenform for all Hecke operators, with the same eigenvalues as f. The problem is that in general we cannot view f as a section of a sheaf $\underline{\omega}^{\otimes 2}$ on $X_0(N)_{\overline{\mathbb{F}}_3}$ and apply a Kodaira–Spencer isomorphism, because $X_0(N)_{\overline{\mathbb{F}}_3}$ does not carry a universal family of generalized elliptic curves with a level structure of type $\Gamma_0(N)$. We will see that this problem is not just artificial, in the sense that for certain eigenforms of type $(N, 2, 1)_{\overline{\mathbb{F}}_3}$ there does not exist a differential form as we want; it will exist for our f because ρ_3 is not induced from $G_{\mathbb{Q}(\sqrt{-3})}$. Let $n \geq 3$ be an integer prime to 3. According to (1.2) we can view f as a section of the sheaf $\underline{\omega}^{\otimes 2}$ on the smooth projective curve $X := \overline{\mathcal{M}}([\Gamma_1(N), \Gamma(n)]_{\overline{\mathbb{F}}_3})$ over $\overline{\mathbb{F}}_3$ that carries a universal family of generalized elliptic curves with a point of order N and a trivialization of the n-torsion. Let Y denote the curve $X_0(N)_{\overline{\mathbb{F}}_3}$. Then Y is the quotient of X by the action of the group $G := (\mathbb{Z}/N\mathbb{Z})^* \times \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z})$ acting on X, and f is a G-invariant section of $\underline{\omega}^{\otimes 2}$. Let $\omega(f)$ be the global differential form on X obtained by applying the Kodaira–Spencer isomorphism to f. The functorial properties of this isomorphism imply that $\omega(f)$ is G-invariant. We would like to conclude that $\omega(f)$ is the pullback of a unique differential form, that we will also call $\omega(f)$, on Y. The uniqueness is guaranteed by the fact that the morphism $\pi\colon X \to Y$ is separable, which is implies that the pullback morphism π^* on differential forms is injective. The problem is the existence. The rest of this section is motivated by [28, II, Lemma 4.4], and its proof. Let V be the biggest open part of Y over which π is etale. Note that G acts on X via its quotient $\overline{G}$ by the subgroup generated by $(-1,-1)$. The group $\overline{G}$ acts faithfully, hence V is the complement in Y of the image of under π of those points of X with non-trivial stabilizer in $\overline{G}$. Since $\pi\colon \pi^{-1}V \to V$ is etale, the restriction of $\omega(f)$ to $\pi^{-1}V$ is the pullback of a unique differential form $\omega(f)$ on V. So we have to show that this $\omega(f)$ has no poles in the complement of V. The information we have is that the pullback of $\omega(f)$ to X has no poles. Let y be a point of Y over which π is ramified, and let x be in $\pi^{-1}y$. Let s and t be uniformizers at x and y, respectively. Let e be the ramification index at x, i.e., $t = s^e u$ with u in $\mathcal{O}_{X,x}^*$. Let r be the valuation $v_x(dt)$ at x of dt. Since $dt = s^{e-1}(e+u's)ds$, where $u' = du/ds$, one has $r \geq e-1$ with equality if and only if π is tamely ramified at x. We have $v_x(\omega(f)) = ev_s(\omega(f)) + r$. It follows that $\omega(f)$ is regular at y if π is tamely ramified over y, so it remains to look at those y over which π is wildly ramified. Such points y all correspond to the elliptic curve E of j-invariant zero over $\overline{\mathbb{F}}_3$. The automorphism group of E is isomorphic to $\mathbb{F}_3^* \times \mathrm{GL}_2(\mathbb{F}_2)$; the projection to $\mathbb{F}_3^*$ comes from the action of $\mathrm{Aut}(E)$ on the tangent space at zero of E, the other projection comes from the action on the 2-torsion of E. The invariants e and r are the same at all x at which π is wildly ramified; note that we have $e = 3$ or $e = 6$. A global calculation as in [28, II, §2], using the Hurwitz formula for π, shows that we have $r = e$, which means that the wild ramification is of the mildest form. (In fact, this can also be derived from Table 1 of [28, II, §2].) It follows that $\omega(f)$ has a pole of order at most one at y. Now the number of y over which π is wildly ramified can be easily computed: it is 1 if $N = 1$, it is 0 if N is divisible by a prime number congruent to -1 modulo 3, and otherwise it is $2^{\nu-1}$ where ν is the number primes dividing N. In the situation of [27, II, Lemma 4.4] this number is zero or one, and it follows that in fact $\omega(f)$ is regular because the sum of its residues must be zero. But if all primes dividing N are 1 modulo 3 and $\nu > 1$, there really exist eigenforms f such that $\omega(f)$ has poles, so we have to show why our $\omega(f)$ is regular. So suppose that $\omega(f)$ is not regular. Let D be the set of y at which π is wildly ramified, viewed as an effective divisor on Y, and let V be the $\overline{\mathbb{F}}_3$-vector space of $\overline{\mathbb{F}}_3$-valued functions on D. Consider the map $R\colon \mathrm{H}^0(Y, \Omega^1(D)) \to V$ which sends a form to its residues at the y in D. Then the image of $\omega(f)$ is not zero. There is a natural action of the Hecke operators on V which is compatible with R. For $n \geq 1$ the Hecke operator T_n acts on V via isogenies of degree n between elements y of D, hence by endomorphisms of degree n of E. Let σ denote one of the two automorphisms of order 3 of E. An endomorphism ϕ of E of degree n that does not commute with σ contributes zero to the action of T_n on V, since the correspondence inducing T_n is wildly ramified at ϕ. Let l be prime and congruent to -1 mod 3. Then E has no endomorphism of degree l commuting with σ, hence T_l acts as zero on V. It follows that $T_l f = 0$ for all such l, which contradicts that ρ_3 is not induced from $G_{\mathbb{Q}(\sqrt{-3})}$.

Let us now consider the case where ρ_3 is not finite at 3. Then $N = 3N'$, with $N' = N(\rho_3)^2$ prime to 3, and $k(\rho_3) = 4$. In this case $X_0(N)_{\overline{\mathbb{F}}_3}$ has two irreducible components, both isomorphic to $X_0(N')_{\overline{\mathbb{F}}_3}$, which intersect transversally at the supersingular points. The sheaf of Kähler differentials on $X_0(N)_{\mathbb{Z}_3}$ is not locally free of rank one at the double points over $\overline{\mathbb{F}}_3$, and it is better to work with the dualizing sheaf Ω on it. This sheaf Ω can be obtained as follows: let $X_0(N)^{\rm sm}_{\mathbb{Z}_3}$ be the smooth locus of $X_0(N)_{\mathbb{Z}_3}$ (i.e., the complement of the double points), let $j\colon X_0(N)^{\rm sm}_{\mathbb{Z}_3} \to X_0(N)_{\mathbb{Z}_3}$ be the inclusion and let Ω^1 be the sheaf of Kähler differentials on $X_0(N)^{\rm sm}_{\mathbb{Z}_3}$; then $\Omega = j_*\Omega^1$. The dualizing sheaf Ω is locally free of rank one and it is dualizing in the sense of Serre duality. For a more detailed description of Ω in the context of modular curves see [29, §§6–7], [17, §§8–9] and references therein. Let $Y := X_0(N)_{\overline{\mathbb{F}}_3}$, and let Y_0 (resp., Y_∞) be the irreducible component of Y containing the cusp 0 (resp., ∞). The restrictions of Ω to Y_0 and Y_∞ are the sheafs of Kähler differentials with poles of order at most one in the supersingular points. Let $n \geq 3$ be prime to 3, let $X := \overline{\mathcal{M}}([\Gamma_1(N'), \Gamma(n)]_{\overline{\mathbb{F}}_3})$ and let $G := (\mathbb{Z}/N'\mathbb{Z})^* \times \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z})$. Then f is a G-invariant section of $\underline{\omega}$ on X. It follows that f/A (recall that A is the Hasse invariant) is a rational section of $\underline{\omega}^{\otimes 2}$ with poles of at most order one in the supersingular points. Applying the Kodaira–Spencer isomorphism to f/A gives a G-invariant rational differential form $\omega(f)$ on X which has poles of order at most one in the supersingular points. We identify the quotient of X by G with Y_∞. Since $\omega(f)$ is G-invariant, we can view $\omega(f)$ as a rational differential form on Y_∞, with poles only at the supersingular points and at the points where $X \to Y_\infty$ is ramified. A calculation as above shows that there are only poles of order at most one in the supersingular points. The proof of [17, Prop. 9.3] shows that there exists a section of Ω on $X_0(N)_{\overline{\mathbb{Z}}_3}$ which us an eigenform for the Hecke algebra and whose restriction to Y_∞ is $\omega(f)$.

We end this section with some remarks. First of all, one can show that the Hecke action on the $\overline{\mathbb{F}}_3$-vector space V that occurs in the arguments for ρ_3 finite at 3, is "Eisenstein": for l prime and not dividing $3N$, T_l acts as $l+1$ on V. Hence it would have been sufficient to use that ρ_3 is irreducible. For more general quotients of X by subgroups of the form $H \times \mathrm{GL}_2(\mathbb{Z}/n\mathbb{Z})$ of G, the vector space V is not necessarily "Eisenstein". This is related to results on groups of connected components of Néron models in [36].

Secondly, instead of studying in detail the wild ramification in the morphism $X \to Y$, we could have used Theorem 3.2.1 as follows. Let q be any prime number that is congruent to -1 mod 3 and that does not divide N. Then, replacing N by qN, one gets $X \to Y$ tamely ramified, hence a differential form on $X_0(Nq)_{\overline{\mathbb{F}}_3}$. This shows that ρ_3 is modular of type $(Nq, 2, 1)_{\overline{\mathbb{Q}}_3}$. Then Mazur's result shows that ρ_3 is modular of type $(N, 2, 1)_{\overline{\mathbb{Q}}_3}$. One reason to give the argument above is to illustrate the problems one gets when interpreting modular forms as differential forms.

4.4 Carayol's reductions

At this point we know that ρ_3 is modular of type $(N(\rho_3)^2, 2, 1)_{\overline{\mathbb{Q}}_3}$ if ρ_3 is finite at 3 and of type $(3N(\rho_3)^2, 2, 1)_{\overline{\mathbb{Q}}_3}$ otherwise. We want to show that ρ_3 is modular of type $(N(\rho_3), 2, 1)_{\overline{\mathbb{Q}}_3}$

if ρ_3 is finite at 3 and of type $(3N(\rho_3),2,1)_{\overline{\mathbb{Q}}_3}$ otherwise. Before explaining how this is done, it is good to recall some results of Langlands, Deligne and Carayol (see [5]). Let p be a prime number and let f be a newform of some type $(N,k,\varepsilon)_{\overline{\mathbb{Q}}_p}$ with $k\geq 2$. Then this gives us a representation $\rho_f\colon G_{\mathbb{Q}}\to \mathrm{GL}_2(\overline{\mathbb{Q}}_p)$, determined by the property that it is unramified outside Np and that for l not dividing Np the Frobenius element $\rho_f(\mathrm{Frob}_l)$ has trace $a_l(f)$. On the other hand, there is also a representation $\pi_f\colon \mathrm{GL}_2(\hat{\mathbb{Z}}\otimes\mathbb{Q})\to\mathrm{GL}(V)$, with V an infinite dimensional $\overline{\mathbb{Q}}_p$-vector space, associated to f in the following way. Let W be the direct limit, taken over all multiples $n\geq 1$ of N, of the $\overline{\mathbb{Q}}_p$-vector spaces $\mathrm{H}^0(\overline{\mathcal{M}}([\Gamma(n)]_{\overline{\mathbb{Q}}_p}),\underline{\omega}^{\otimes k})$. It is clear that $\mathrm{GL}_2(\hat{\mathbb{Z}})$ acts on W, and it is not hard to see that this action extends to one of $\mathrm{GL}_2(\hat{\mathbb{Z}}\otimes\mathbb{Q})$. Then V is the subspace of W that is generated by the $g(f)$, for g in $\mathrm{GL}_2(\hat{\mathbb{Z}}\otimes\mathbb{Q})$. One knows that V is an irreducible representation of $\mathrm{GL}_2(\hat{\mathbb{Z}}\otimes\mathbb{Q})$ and that $V=\otimes_l' V_l$ is the restricted tensor product, over all primes l, of irreducible admissible representations $\pi_{f,l}\colon\mathrm{GL}_2(\mathbb{Q}_l)\to\mathrm{GL}(V_l)$. The result alluded to describes, for all $l\neq p$, the restriction $\rho_{f,l}$ of ρ_f to some decomposition group at l in terms of $\pi_{f,l}$. It follows from this result that the conductor $N(\overline{\rho}_f)$ of the reduction $\overline{\rho}_f\colon G_{\mathbb{Q}}\to\mathrm{GL}_2(\overline{\mathbb{F}}_p)$ divides N (here we suppose that $\overline{\rho}_f$ is irreducible, since otherwise it is not well defined). Carayol [4] and Livné [26] have classified, in terms of the $\pi_{f,l}$, the $l\neq p$ dividing $N/N(\overline{\rho}_f)$.

The strategy for proving that ρ_3 is modular of type $(N(\rho_3),2,1)_{\overline{\mathbb{Q}}_3}$ if ρ_3 is finite at 3 and of type $(3N(\rho_3),2,1)_{\overline{\mathbb{Q}}_3}$ otherwise will be the following. Suppose that we know that ρ_3 is modular of type $(N,2,1)_{\overline{\mathbb{Q}}_3}$ for some N dividing $N(\rho_3)^2$. Let $l\neq 3$ be a prime number and suppose that l^2 divides N. Then we want to show that ρ_3 is modular of type $(N',2,1)_{\overline{\mathbb{Q}}_3}$ for some N' dividing N with l^2 not dividing N'. Of course, this suffices.

So suppose that f is a cuspidal eigenform of type $(N,2,1)_{\overline{\mathbb{Q}}_3}$ with N dividing $N(\rho_3)^2$ and that $l\neq 3$ is a prime number such that l^2 divides N. The newform associated to f has level dividing N, so we may in fact assume that f is a newform. The classification of Carayol and Livné then says that $\rho_{f,l}$ is of one of the following two types:

1. $\rho_{f,l}$ is a direct sum of two ramified characters $\alpha,\beta\colon G_{\mathbb{Q}_l}\to\overline{\mathbb{Q}}_3^*$ whose reductions $\overline{\alpha},\overline{\beta}\colon G_{\mathbb{Q}_l}\to\overline{\mathbb{F}}_3^*$ are unramified,
2. $\rho_{f,l}=\mathrm{Ind}_{G_K}^{G_{\mathbb{Q}_l}}\psi$ with K the unique unramified quadratic extension of $\mathbb{Q}_l$ and $\psi\colon G_K\to\overline{\mathbb{Q}}_3^*$ a ramified character with unramified reduction $\overline{\psi}\colon G_K\to\overline{\mathbb{F}}_3^*$.

Let us deal with the first case first. Let $\tilde{\chi}_3\colon G_{\mathbb{Q}}\to\mathbb{Z}_3^*$ be the character giving the action on all roots of unity of 3-power order. Recall that since f has trivial character and weight two we have $\det(\rho_f)=\tilde{\chi}_3$. This implies that $\alpha\beta$ is unramified. There is a unique character ϵ of $\mathbb{F}_l^*=\mathrm{Gal}(\mathbb{Q}(\zeta_l)/\mathbb{Q})$ with values in the kernel of $\overline{\mathbb{Z}}_3^*\to\overline{\mathbb{F}}_3^*$ such that $\alpha\varepsilon$ is unramified. Let f' be the newform corresponding to the twist $f\otimes\varepsilon$ of f by ε. One way to express this is to say that $a_n(f')=a_n(f)\varepsilon(n)$ for all n prime to l. Another way is to say that $\rho_{f'}=\rho_f\otimes\varepsilon$. Anyway, f' is a newform of type $(N/l,2,\varepsilon^2)_{\overline{\mathbb{Q}}_3}$ giving ρ_3. Since $\overline{\varepsilon}=1$ and ρ_3 is not induced from $\mathbb{Q}(\sqrt{-3})$ Carayol's Lemma implies that ρ_3 is modular of type $(N/l,2,1)_{\overline{\mathbb{Q}}_3}$.

Let us now say something about case 2. The analog of this case with 3 replaced by a prime $p\geq 5$ is treated in [4, §5], and uses the Jacquet–Langlands correspondence to switch to a certain

Shimura curve. The generalization to the case $p = 3$, using that ρ_3 is not induced from $\mathbb{Q}(\sqrt{-3})$, is explained in [12, §5]. One might also say that this generalization is done in [4], if one admits that the remarks in [4, §4.4] concerning modular curves also hold for the Shimura curves used in [4, §5]. We will now sketch the argument. Let q be a prime number not dividing $3N$, such that $\rho_3(\mathrm{Frob}_q)$ is conjugated to $\rho_3(c)$, with c a complex conjugation. Then a result of Ribet (see [34]) says that there exists a newform f' of type $(N'q, 2, 1)_{\overline{\mathbb{Q}}_3}$, with N' dividing N, such that $\overline{\rho}_{f'} \cong \rho_3$ and with $\rho_{f',q}$ special, i.e., $\rho_{f',q}$ is a non-split extension of an unramified character α with $\alpha^2 = 1$ by $\alpha\tilde{\chi}_3$ (see also [44, II, Lemma 2.3]). If $\rho_{f',l}$ is not in case 2, then one applies the method to deal with case 1 to get rid of the l^2 in the level and then one applies Mazur's Theorem 3.2.1 to get rid of q. So we may assume that $\rho_{f',l}$ is in case 2. Let B be the quaternion algebra over $\mathbb{Q}$ with discriminant pq. By the Jacquet–Langlands correspondence and the results of [5], $\rho_{f'}$ can be constructed from the 3-adic Tate module of the jacobian of a Shimura curve of a certain level associated to B. On this Shimura curve one has an action by the group $\mathbb{F}_{l^2}^* \times \mathbb{F}_{q^2}^*$ which is analogous to the action of the diamond operators on modular curves. A version of Carayol's Lemma then shows that ρ_3 actually arises from the quotient of this Shimura curve by that group. Switching back to modular curves by the Jacquet–Langlands correspondence then shows that ρ_3 is modular of type $(N''q, 2, 1)_{\overline{\mathbb{Q}}_3}$, with N'' dividing N/l. Then one finishes by applying Theorem 3.2.1.

References

[1] B.J. Birch and W. Kuyk (editors). *Modular functions of one variable IV.* Springer Lecture Notes in Mathematics 476 (1975).

[2] N. Boston, H.W. Lenstra and K.A. Ribet. *Quotients of group rings arising from two-dimensional representations.* C.R. Acad. Sci. Paris, t. 312, Série I, p. 323–328 (1991).

[3] J.-F. Boutot and H. Carayol. *Uniformisation p-adique des courbes de Shimura: les théorèmes de Cerednik et Drinfeld.* Astérisque 196–197, 45–158 (1991).

[4] H. Carayol. *Sur les représentations galoisiennes mod l attachées aux formes modulaires.* Duke Math. Journal **59** (1989), No. 3, 785–801.

[5] H. Carayol. *Sur les représentations l-adiques associées aux formes modulaires de Hilbert.* Ann. Sci. Ecole Norm. Sup. (4) 19, 409–468 (1986).

[6] P. Cartier. *Détermination des caractères des groupes finis simples: travaux de Lusztig.* Séminaire Bourbaki, Exposé 658 (1986).

[7] R.F. Coleman and J.F. Voloch. *Companion forms and Kodaira-Spencer theory.* Invent. Math. **110** (1992), 263–281.

[8] P. Deligne. *Formes modulaires et représentations l-adiques.* Séminaire Bourbaki, exposé 355, Springer Lecture Notes in Mathematics 179, 139–172, (1971).

[9] P. Deligne. *Courbes elliptiques: formulaire d'après John Tate.* Modular functions of one variable IV, pages 53–73. Springer Lecture Notes in Mathematics 476 (1975).

[10] P. Deligne and M. Rapoport. *Les schémas de modules des courbes elliptiques.* In Modular Functions of One Variable II. Springer Lecture Notes in Mathematics 349 (1973).

[11] P. Deligne and J-P. Serre. *Formes modulaires de poids 1.* Ann. Sci. Ecole Norm. Sup. (4) 7, 507–530 (1974).

[12] F. Diamond. *The refined conjecture of Serre.* In: Elliptic Curves, Modular Forms and Fermat's Last Theorem, J. Coates, S.T. Yau, eds., International Press, Cambridge, pages 22–37 (1995).

[13] F. Diamond. *Congruence primes for cuspforms of weight $k \geq 2$.* Astérisque 196–197, 205–213 (1991).

[14] F. Diamond and J. Im. *Modular forms and modular curves.* CMS Conf. Proc. Vol. 17, AMS Publ., Providence, "Seminar on Fermat's Last Theorem", edited by V.K. Murty.

[15] S.J. Edixhoven. *The weight in Serre's conjectures on modular forms.* Invent. Math. **109** (1992), 563–594.

[16] G. Faltings and B. Jordan. *Crystalline cohomology and* $\mathrm{GL}_2(\mathbb{Q})$. Israel Journal of Mathematics **90**, 1–66 (1995).

[17] B.H. Gross. *A tameness criterion for Galois representations associated to modular forms (mod p).* Duke Math. Journal **61** (1990). No. 2, 445–517.

[18] A. Grothendieck. *Séminaire de géométrie algébrique.* Springer Lecture Notes in Mathematics 151, 152, 153, 224, 225, 269, 270, 288, 305, 340, 589.

[19] Y. Hellegouarch. *Courbes elliptiques et équation de Fermat.* Thesis, Besançon, 1972.

[20] Y. Hellegouarch. *Points d'ordre $2p^h$ sur les courbes elliptiques.* Acta Arithmetica XXVI, 253–263 (1975).

[21] B. Jordan and R. Livné. *Conjecture "epsilon" for weight $k > 2$.* Bull. AMS 21, 51–56 (1989).

[22] N.M. Katz. *p-adic properties of modular schemes and modular forms.* In Modular Functions of One Variable III. Springer Lecture Notes in Mathematics 350 (1973).

[23] N.M. Katz. *A result on modular forms in characteristic p.* Springer Lecture Notes in Mathematics 601, 53–61 (1976).

[24] N. Katz and B. Mazur. *Arithmetic moduli of elliptic curves.* Annals of Mathematics studies, study 108. Princeton University Press, 1985.

[25] R.P. Langlands. *Base change for* GL_2. Princeton University Press (1980).

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

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

[28] B. Mazur. *Modular curves and the Eisenstein ideal.* Publications Mathématiques de l'I.H.E.S. 47 (1977).

[29] B. Mazur and K.A. Ribet. *Two-dimensional representations in the arithmetic of modular curves.* Astérisque 196–197, 215–255 (1991).

[30] M. Raynaud. *Spécialisation du foncteur de Picard.* Publications Mathématiques de l'I.H.E.S. 38 (1970).

[31] M. Raynaud. *Schémas en groupes de type* $(p,...,p)$. Bull. Soc. Math. France 102, 241-280 (1974).

[32] K.A Ribet. *Report on mod l representations of* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. Proceedings of Symposia in Pure Mathematics, **55** (1994), Part 2.

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

[34] K.A. Ribet. *Congruence relations between modular forms.* Proc. Int. Congr. of Math., Warsaw, 503–514 (1983).

[35] K.A. Ribet. *Lowering the levels of modular representations without multiplicity one.* Internat. Math. Res. Notices, 15–19 (1991).

[36] K.A. Ribet. *Irreducible Galois representations arising from component groups of jacobians.* In: Elliptic curves, modular forms and Fermat's Last Theorem (eds: Coates, Yau), International Press, 1995, pages 131–147.

[37] J-P. Serre. *Sur les représentations de degré 2 de* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. Duke Mathematical Journal 54, No. 1, (1987).

[38] J-P. Serre. *Représentations linéaires des groupes finis.* Collection Méthodes, Hermann, Paris (1978).

[39] J-P. Serre. *Propriétés galoisiennes des points d'ordre fini des courbes elliptiques.* Invent. Math. **15** (1972), 259–331.

[40] G. Shimura. *Introduction to the theory of automorphic functions.* Iwanami Shoten and Princeton University Press, Princeton, 1971.

[41] H.P.F. Swinnerton-Dyer. *On l-adic representations and congruences for coefficients of modular forms.* Springer Lecture Notes in Mathematics 350, 1–55 (1973).

[42] R. Taylor and A. Wiles. *Ring theoretic properties of certain Hecke algebras.* Annals of Math. **142**, 553–572 (1995).

[43] J. Tunnell. *Artin's conjecture for representations of octahedral type.* Bull. A.M.S. **5**, 173–175 (1981).

[44] A. Wiles. *Modular elliptic curves and Fermat's Last Theorem.* Annals of Math. **142**, 443–551 (1995).