# SERRE'S CONJECTURE

First, I want to set some notation regarding $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$. We fix throughout an embedding $\overline{\mathbf{Q}} \hookrightarrow \overline{\mathbf{Q}}_\ell$ for each prime $\ell$, yielding a restriction map

$$\mathrm{G}_\ell = \mathrm{Gal}(\overline{\mathbf{Q}}_\ell/\mathbf{Q}_\ell) \hookrightarrow \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}).$$

This is an injection, so we may regard this as a subgroup. By acting on the residue field of $\mathcal{O}_{\overline{\mathbf{Q}}_\ell}$, we may produce a map

$$\mathrm{G}_\ell \to \mathrm{Gal}(\overline{\mathbf{F}}_\ell/\mathbf{F}_\ell).$$

Define $\mathrm{I}_\ell$, the inertia subgroup, to be the kernel of this map. Inside of $\mathrm{I}_\ell$, there is the wild inertia $\mathrm{I}_w$: this is the maximal pro-$\ell$ subgroup. The quotient by $\mathrm{I}_w$ is the tame inertia $\mathrm{I}_t$. By the initial observation, we can regard these all as sitting inside of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$.

To motivate Serre's conjecture, I want to first recall a bit about how modular forms have Galois representations attached to them. Take $k \geq 2$ and $N \geq 1$, and let $f = \sum_n a_n q^n$ be a weight $k$ normalized cuspidal eigenform in $S_k(\Gamma_1(N))$ ($\Gamma_1(N)$ consists of matrices which are unipotent modulo $N$). I'll usually break this up as

$$S_k(\Gamma_1(N)) = \bigoplus_{\varepsilon:(\mathbf{Z}/N\mathbf{Z})^\times \to \mathbf{C}^\times} S_k(N, \varepsilon)$$

according to the character $\varepsilon$ of $(\mathbf{Z}/N\mathbf{Z})^\times$.

Letting $\mathrm{E} = \mathbf{Q}(\ldots, a_n, \ldots)$, Deligne constructs a Galois representation

$$\rho_{f,\lambda} : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_2(\mathrm{E}_\lambda)$$

for each non-Archimedean prime $\lambda$ of E. For all primes $p$ not dividing $\ell N$ where $\ell$ is the residue characteristic of $\mathrm{E}_\lambda$, the trace $\mathrm{tr}(\mathrm{Frob}_p)$ recovers $a_p$.

You can pick a model of this Galois representation (via conjugation) such that we have

$$\tilde{\rho}_{f,\lambda} : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_2(\mathcal{O}_\lambda)$$

where $\mathcal{O}_\lambda$ is the ring of integers of $\mathrm{E}_\lambda$. Then, this induces a map to $\mathrm{GL}_2(\mathcal{O}_\lambda/\lambda\mathcal{O}_\lambda) \hookrightarrow \mathrm{GL}_2(\overline{\mathbf{F}}_\ell)$.

Thus, we see that we have a construction of Galois representations $\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_2(\overline{\mathbf{F}}_\ell)$ arising from modular forms. Note that it doesn't particularly matter if we do this in a canonical way.

Let us state precisely the sort of result we get from this. There is a space $S_k(N, \varepsilon, \overline{\mathbf{Z}})$ of cuspidal modular forms of weight $k$, level $N$, nebentype $\varepsilon$ with coefficients in $\overline{\mathbf{Z}}$. Upon reduction, we obtain $S_k(N, \varepsilon, \overline{\mathbf{F}}_\ell)$. The previous construction associates a mod $\ell$ Galois representation $\rho_f$ to eigenforms $f$ in $S_k(N, \varepsilon, \overline{\mathbf{F}}_\ell)$, with the following properties:

- It is semi-simple.
- $\rho$ is unramified outside of $N\ell$.
- We have $\operatorname{tr}(\rho(\operatorname{Frob}_p)) = a_p$, $p \nmid N\ell$. Here, $f = \sum_n a_n q^n$ in $S_k(N, \varepsilon, \overline{\mathbf{F}}_\ell)$.
- We have $\det(\rho(\operatorname{Frob}_p)) = p^{k-1}\varepsilon(p)$.

Is it possible that we produce all irreducible mod $\ell$ Galois representations $\rho$ from as some $\rho_f$? We can immediately see this is not the case, because there is already a necessary condition that holds for any Galois representations constructed this way: we need to have $\det(\rho(c)) = -1$, where $c$ denotes complex conjugation.

> Lemma 0.1. We have $\det(\rho(c)) = -1$ when $\rho$ arises from a modular form.

*Proof.* We can actually just figure out what $\det \rho$ is in general. Indeed, at Frobenius elements one compatibility of Deligne's construction is

$$\det(\rho(\operatorname{Frob}_p)) = p^{k-1}\varepsilon(p)$$

where $\varepsilon : (\mathbf{Z}/N\mathbf{Z})^\times \to \mathbf{C}^\times$ is given by $f|\langle d\rangle = \varepsilon(d)f$ for a diamond operator $\langle d\rangle$.

Now Chebatorev density tells us that $\det \rho : \operatorname{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \overline{\mathbf{F}}_\ell^\times$ is given by $\chi^{k-1}\varepsilon$ where $\chi$ is the mod $\ell$ cyclotomic character and $\varepsilon$ is now interpreted as landing in $\overline{\mathbf{F}}_\ell^\times$ (this makes sense as before it landed in $\mathcal{O}_\lambda$). We compose with the mod $N$ cyclotomic character to get $\varepsilon : \operatorname{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \overline{\mathbf{F}}_\ell^\times$.

Then using $\varepsilon(-1) = (-1)^k$ (by applying the diamond operator $\langle -1\rangle$ to $f$), we see on this new incarnation of $\varepsilon$ the value at $c$ is $\varepsilon(-1) = (-1)^k$. Since $\chi(c) = -1$ (it has a nontrivial value and squares to one), the result follows. □

Thus, we cannot expect to get all mod $\ell$ Galois representations from modular forms. Serre's conjecture says that this is the only real condition.

> Conjecture 0.2 (Serre's conjecture, weak form). Assume that $\rho : \operatorname{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \operatorname{GL}_2(\overline{\mathbf{F}}_\ell)$ is irreducible and odd ($\rho(c) = -1$). Then $\rho$ is modular.

Here, modular means that we can produce it from the previous construction. We have already seen enough to produce some small amount of evidence for this: the Langlands-Tunnell result is enough to prove this for $\ell = 2, 3$.

In fact, Serre made a stronger conjecture which is that you can read off from the Galois representation an exact minimal level and weight.

> CONJECTURE 0.3 (Serre's conjecture, strong form). Let $\rho$ be an irreducible and odd mod $\ell$ Galois representation. Then $\rho$ is modular, and the associated modular form $f$ can be chosen to have weight $k(\rho)$ and level $N(\rho)$.

> REMARK 0.4. There is a reason I didn't say anything about the character. If you do it in the obvious way, there is a counterexample when $\ell = 2$ for example.

It was known earlier that these conjectures are in fact equivalent (which is most of what I will be saying today). Both are now theorems.

What are the optimal level and weight? The recipe for $k(\rho)$ is a bit more complicated, so we'll delay this for the moment. The weight $N(\rho)$ is a bit easier to motivate. In the construction of Galois representations from modular forms, we see already that the representation is unramified for all $p \nmid N\ell$.

In particular, we expect the optimal level to be of the form of the conductor

$$N(\rho) = \prod_{p \neq \ell} p^{n_p(\rho)},$$

where $n_p(\rho) = 0$ if $\rho$ is unramified at $p$ and is $> 0$ otherwise.

The basic idea is that we should then consider what happens locally at $p$, that is consider the representation

$$\mathrm{G}_p = \mathrm{Gal}(\overline{\mathbf{Q}}_p/\mathbf{Q}_p) \to \mathrm{GL}_2(\overline{\mathbf{F}}_\ell).$$

Now let G be the quotient of $\mathrm{G}_p$ by the kernel of this representation, which is some finite Galois group. Then we have a ramification filtration on the finite group G: we have

$$\mathrm{G} = \mathrm{G}_{p,-1} \supset \bar{\mathrm{I}}_p = \mathrm{G}_{p,0} \supset \dots.$$

Here, the bar denotes that we take the image in $\mathrm{GL}_2(\overline{\mathbf{F}}_\ell)$. Note that we can't do this for all of $\mathrm{G}_p = \mathrm{Gal}(\overline{\mathbf{Q}}_p/\mathbf{Q}_p)$, since the filtration only makes sense for a finite extension.

It's natural to then guess that having higher ramification corresponds to $n_p(\rho)$ being larger.

> DEFINITION 0.5. Set
> $$n_p(\rho) := \sum_{i \geq 0} \frac{1}{[\mathrm{G}_{p,0} : \mathrm{G}_{p,i}]} \dim(\mathrm{V}/\mathrm{V}^{\mathrm{G}_{p,i}}).$$

This defines the level $N(\rho)$, in a way which is fairly natural: the deeper into the ramification filtration $\ker \rho$ goes, the higher the multiplicity of $p$ in $N$ needs to be. However, it is non-trivial that we actually get an integer out of this!

The story for $k(\rho)$ is a bit more complicated. We'll first explicitly study the situations that can arise for the restriction to inertia when we look at the Galois representation attached to $f = \sum_n a_n q^n$ where $2 \leq k(\rho) \leq \ell + 1$.

> Theorem 0.6. Assume $2 \leq k \leq \ell + 1$ and let $f$ be a cuspidal eigenform of some level $N$.
>
> Assuming $a_\ell \neq 0$ in $f$, called the ordinary case, we have
> $$\rho_{f,\ell}|_{\mathrm{I}_\ell} = \begin{pmatrix} \chi_\ell^{k-1} & * \\ 0 & 1 \end{pmatrix}$$
> by a result of Deligne.
>
> If $a_\ell = 0$, or the supersingular case, Fontaine gives a different description. Namely, we have
> $$\rho_{f,\ell}|_{\mathrm{I}_\ell} = \begin{pmatrix} \psi^{k-1} & 0 \\ 0 & \phi^{k-1} \end{pmatrix}$$
> where $\psi$ and $\phi$ are two fundamental characters of level two.

The tame inertia can be identified with $\varprojlim \mathbf{F}_{\ell^n}^\times$, since it is generated by the extensions $\mathbf{Q}_\ell^{\mathrm{nr}}(\sqrt[n]{\ell})$ for $n$ not divisible by $\ell$. These each have Galois group $\mu_n$ over $\mathbf{Q}_\ell^{\mathrm{nr}}$; we can therefore identify tame inertia with $\varprojlim \mathbf{F}_{\ell^n}^\times$, where the maps in the inverse limit are the norm maps.

A fundamental character of level $n$ is a representation on the tame inertia
$$\mathrm{I}_t = \varprojlim \mathbf{F}_{\ell^n}^\times \to \overline{\mathbf{F}}_\ell$$
of $\mathrm{G}_\ell$ given by projecting down to $\mathbf{F}_{\ell^n}^\times \subseteq \mathbf{F}_{\ell^n}$ and using one of the $n$ field embeddings into $\overline{\mathbf{F}}_\ell$. We can of course trivially extend to $\mathrm{G}_\ell \supseteq \mathrm{I}_t$ and then to $\mathrm{G}_{\mathbf{Q}} = \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$.

> Remark 0.7. From a modern viewpoint, these results are coming from the fact that $\rho_f$ is crystalline when $\ell$ is prime to the level: when we're in the ordinary case, we can read off from $p$-adic Hodge theory that the Hodge-Tate weights are $\{0, k-1\}$. If $\rho_f$ is reducible mod $\ell$, results of Berger for two-dimensional crystalline representations tell us that we get exactly the form Deligne says.

> Precisely, Berger's result tells us that for a two dimensional crystalline representation which is reducible modulo $\ell$, we have
> $$\mathrm{V} \simeq \chi^r \xi_1 \oplus \xi_2$$
> for unramified characters $\chi_1, \chi_2$ and the Hodge-Tate weights are $\{0, r\}$.

The main strategy will be to try to reduce to these cases with twists by a cyclotomic character. When we are already in this case, we can see the following definition is then well-motivated:

> Definition 0.8. If $\rho$ already lands in one of these cases, then define $k(\rho) = k$.

Now, we'll want to see how to do the reduction so we can get a general formula. This begins by studying how Katz's $\Theta$ operator interacts with the Galois representation.

> Theorem 0.9. Let $\Theta = q\frac{\mathrm{d}}{\mathrm{d}q}$, so we get a map
> $$\Theta : S_k(N, \varepsilon, \overline{\mathbf{F}}_\ell) \to S_{k+\ell+1}(N, \varepsilon, \overline{\mathbf{F}}_\ell).$$
> If $f \in S_k(N, \varepsilon, \overline{\mathbf{F}}_\ell)$ is a normalized eigenform, the Galois representation associated to $\Theta(f)$ is given by
> $$\rho_{\Theta(f)} = \chi_\ell \otimes \rho_f.$$

In particular, we can produce modular forms realizing all higher twists using $\Theta$. One can check this by verifying the Frobenius traces and determinants are the same, and then using that this determines the representation if it is semisimple.

Edixhoven proved the following result:

> Theorem 0.10 (Edixhoven). Let $f \in S_k(N, \varepsilon, \overline{\mathbf{F}}_\ell)$ be an eigenform. Then there exists another eigenform $g \in S_{k'}(N, \varepsilon, \overline{\mathbf{F}}_\ell)$ where $2 \leq k' \leq \ell + 1$ such that $f$ and $\Theta^i g$ have the same eigenvalues for all Hecke operators $\mathrm{T}_p$ (away from $\ell$) and $0 \leq i \leq \ell - 1$.

The game is now as follows: we'll start with some Galois representation $\rho$, and to guess $k(\rho)$ we'll start to look at $\rho|_{\mathrm{I}_\ell}$. Then, we'll want to figure out how to pull out a twist of a cyclotomic character to use Edixhoven's result. Once we do this, we can use the previous theorem to safely extend $k(\rho)$ from the case where we already mostly understood it.

Let $\rho_\ell := \rho|_{\mathrm{G}_\ell}$.

> LEMMA 0.11. The wild inertia $\mathrm{I}_w$ acts trivially on $\rho_\ell^{\mathrm{ss}}$.

*Proof.* Because we are semisimplifying, without loss of generality assume we are in the irreducible case. Let the vector space for the representation be V, which we can assume is some $\mathbf{F}_{\ell^i}$ vector space. Because $\mathrm{I}_w$ is a pro-$\ell$ subgroup, when we look at the $\mathrm{I}_w$ action the $\overline{\mathbf{F}}_{\ell^i}$ vector space must break into orbits whose sizes are powers of $\ell$. Due to the size of the entire vector space as a finite set, there cannot be just 0 as a fixed point: the number of orbits of size one must be divisible by $\ell$. This means that $\mathrm{V}^{\mathrm{I}_w}$ is nontrivial, hence all of V by irreducibility. □

Hence, there is only a nontrivial component coming from the tame inertia. As this is abelian, it splits into characters. We have

$$\rho_\ell^{\mathrm{ss}}|_{\mathrm{I}_t} = \psi \oplus \phi$$

where $\psi$ and $\phi$ are characters of some level.

> PROPOSITION 0.12. Taking the $p$th power of $\rho_\ell^{\mathrm{ss}}|_{\mathrm{I}_t}$ yields a conjugate representation. We then have
> $$\{\psi, \phi\} = \{\psi^p, \phi^p\}.$$

It follows there are two cases:

- (1) We have $\psi = \phi^p$ and vice versa. Both are of level 2.
- (2) We have $\psi = \psi^p$ and $\phi = \phi^p$. Then both characters are of level 1.

**Case 1**. Let $\eta, \eta'$ be the fundamental characters of level 2. These generate level two characters of $\mathrm{I}_t$, and so we write

$$\psi = \eta^a \eta'^b, \phi = \eta'^a \eta^b.$$

Here, $0 \leq a, b \leq \ell - 1$.

One can show in this case that $\rho$ is necessarily irreducible, and therefore agrees with the semisimplification. It follows

$$\rho_\ell|_{\mathrm{I}_\ell} = \begin{pmatrix} \eta^a\eta'^b & 0 \\ 0 & \eta'^a\eta^b \end{pmatrix} = \chi_\ell^a \otimes \begin{pmatrix} \eta'^{b-a} & 0 \\ 0 & \eta^{b-a} \end{pmatrix}.$$

In particular, we reduce to the form in the supersingular case up to a Frobenius twist. Edixhoven's result tells us to define

$$k(\rho) = (b - a + 1) + a(\ell + 1).$$

**Case 2**. This is a bit more subtle. First, assume that $\mathrm{I}_w$ acts trivially. Then we have $\{\psi, \phi\} = \{\chi_\ell^a, \chi_\ell^b\}$ for $0 \leq a, b \leq \ell - 2$. Then

$$\rho_\ell|_{\mathrm{I}_\ell} = \begin{pmatrix} \chi_\ell^b & 0 \\ 0 & \chi_\ell^a \end{pmatrix} = \chi_\ell^a \otimes \begin{pmatrix} \chi_\ell^{b-a} & 0 \\ 0 & 1 \end{pmatrix}$$

when we assume without loss of generality that $a \leq b$. Then we've produced a Galois representation matching Deligne's result in the ordinary case up to a twist. We again define $k(\rho) = (b - a + 1) + a(\ell - 1)$, except now if $a = b = 0$ we assign $\ell$: we didn't attach Galois representations when the weight is 1. We're allowed to modify by multiples of $\ell - 1$ by looking at the determinant, so the correct thing is $\ell$.

If $\mathrm{I}_w$ does not act trivially, we are in a bit of trouble. It is now possible for $k = \ell + 1$, so we'll need to tell apart weight 2 and weight $\ell + 1$ modular forms. We'll get in general

$$\rho_\ell|_{\mathrm{I}_\ell} = \begin{pmatrix} \chi_\ell^\beta & * \\ 0 & \chi_\ell^\alpha \end{pmatrix}$$

where $1 \leq \beta \leq \ell - 1$ and $0 \leq \alpha \leq \ell - 2$. When $\beta \neq \alpha + 1$, we can proceed as in the previous subcase to get for $a = \min(\alpha, \beta)$ and $b = \max(\alpha, \beta)$ the weight $k(\rho) = (b - a + 1) + a(\ell - 1)$ by pulling out a power of the mod $\ell$ cyclotomic character. Otherwise, we need to tell apart weight 2 and $\ell + 1$. These can be told apart by looking at the difference in wild ramification, and then we proceed as before: we get 2 if it is finite flat, and $\ell + 1$ otherwise. The difference is that in weight two we can produce the Galois representation $\mathrm{A}[\lambda]$ from an abelian variety arising from the Jacobian $J_1(N)$. This abelian variety has a good model, which lets us see the representation is finite flat for $p \nmid N$.

With this, modulo Edixhoven's result we have explained how to show a modular mod $\ell$ Galois representation arises from a modular form of the minimal weight $k(\rho)$.

> Theorem 0.13 (Edixhoven). Assume $\rho$ is a modular mod $\ell$ Galois representation. Then it can be chosen to have weight $k(\rho)$.

To finish, I'll talk a bit about how to get the optimal level once we know this result. Doing this would show that the strong and weak conjectures are the same.

> Proposition 0.14 (Serre). Given some modular mod $\ell$ representation $\rho$ coming from a level $N$ and weight $k$ form $f$, we can produce $\rho$ from a modular form of level prime to $N$ and the same weight.

With this, we can then reduce to weight $2 \leq k \leq \ell + 1$ using Edixhoven's result: the $\Theta$ operator does not change the level, so it suffices to find the optimal level in only this case.

We can then further reduce to the weight two case: there is a correspondence for mod $\ell$ Galois representations associated to eigenforms

$$\{2 < k \leq \ell + 1, \text{level } N\} \leftrightarrow \{k = 2, \text{level } \ell N\}.$$

Note that we have allowed only a single power of $\ell$ back into the level, and also on the left side two can be excluded as we are already in weight two in that case. The advantage of this technique is that the weight two case gives simpler geometry to work with: the Galois representations in this case can be produced by A[$\lambda$], for an abelian variety A arising from $J_1(N)$. This makes optimizing the level easier.

Carayol showed that $N(\rho)|N$. In particular, Carayol reduced it to the following key case:

> Theorem 0.15. Let $\rho : \mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q}) \to \mathrm{GL}_2(\overline{\mathbf{F}}_\ell)$ be a Galois representation that arises from a weight 2 newform $f$ of level $pN$ , with $p \nmid \ell N$, and character $\varepsilon : (\mathbf{Z}/pN\mathbf{Z})^\times \to \mathbf{C}^\times$.
>
> Assume that $\rho$ is unramified at $p$, and that $\varepsilon$ factors through the natural map $(\mathbf{Z}/pN\mathbf{Z})^\times \to (\mathbf{Z}/N\mathbf{Z})^\times$. Then $\rho$ arises from a form of level $N$.

This is what is called the epsilon conjecture, and was proven by Ribet. If we know all elliptic curves over $\mathbf{Q}$ are modular, this suffices to prove FLT.

## References

[Bes15] Alex J Best, *Serre's conjecture.*

[Edi97] Bas Edixhoven, *Serre's conjecture*, Modular forms and Fermat's last theorem, Springer, 1997, pp. 209–242.

[RS99] Kenneth A Ribet and William A Stein, *Lectures on serre's conjectures*, Arithmetic algebraic geometry (Park City, UT, 1999) **9** (1999), 143–232.