A FEARSOME FOURSOME: LANGLANDS, TUNNELL, WILES AND FERMAT

KIMBALL MARTIN

Modulaire! Modulaire!
Those words are so unfair!
Many meetings, many seatings,
Many meanings, many gleanings.

Yet so obtusive, so elusive,
Is there nothing more conducive?
Ah, here's a friend by far more fair!
Though rough and rugged for the wear.

Seldom was a longer name so seemly,
Or came functoriality so dreamy,
Than when I turned from modulaire,
And found that automorphy in the air.

These notes are from a presentation for Ma 162b taught by Edray Goins at Caltech in Winter 2004. I attempt here to give a rough sketch of the role of automorphic forms and representations in the proof of Fermat's last theorem (that is, the proof that all (semistable) elliptic curves are modular). I am really not at all following Gelbart's article in the Cornell-Silverman-Stevens volume, except perhaps in Section 4. In Section 3, I attempted to follow Cogdell's lecture notes from a course at the Fields Institute (available on their website) in Winter 2003. I claim absolutely no responsibility to the veracity of the words which follow.

Notation: $G_{\mathbb{Q}} = \mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, $\mathfrak{H}$ is the upper half-plane, tr is the trace map, and Fr_p denotes a Frobenius conjugacy class for p in an appropriate finite quotient of $G_{\mathbb{Q}}$.

1. L-Functions

We've talked about a correspondence between two-dimensional Galois representations and modular forms, but I'd like to rephrase things in terms of L-functions, though I suppose I don't actually need to. However it will be much more convenient for stating things more generally. Let f be a eigen-cusp-new-form of weight $w \geq 1$ and character ε. By Deligne, Serre, Eichler and Shimura, one can attach to f an odd, continuous Galois representation $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(F)$ such that for almost all primes p,

$$\mathrm{tr}(\rho(\mathrm{Fr}_p)) = a_p \ , \ \det(\rho(\mathrm{Fr}_p)) = \varepsilon(p)p^{w-1}, \tag{1}$$

where $F = \mathbb{C}$ if $w = 1$ and F can be $\overline{\mathbb{Q}}_l$ for any prime l if $w \geq 2$.

In fact for $F = \mathbb{C}$ or $\overline{\mathbb{F}}_l$, it's conjectured that any odd, continuous irreducible Galois representation $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(F)$ should correspond to a modular form f (defined over F) in the above sense. (I'm told things are more delicate when $F = \overline{\mathbb{Q}}_l$.) In this case, we'll say that ρ is *modular*. Let's reformulate the weight-one case with L-functions. Write $f = \sum_{n\geq 1} a_n q^n$. Define the L-function

$$L(s,f) = \sum_{n\geq 1} \frac{a_n}{n^s} = \prod_p L_p(s,f) \ , \ L_p(s,f) = \frac{1}{1 - a_p p^{-s} + \varepsilon(p)p^{w-1-2s}} \ \ (p \not| lN)$$

Let $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{C})$ be a continuous Galois representation. Define the Artin L-function by

$$L(s,\rho) = \prod_p L_p(s,\rho),$$

where at the unramified places for ρ (so at almost all places),

$$L_p(s,\rho) = \frac{1}{\det(I - \rho(\mathrm{Fr}_p)p^{-s})} = \frac{1}{1 - \mathrm{tr}(\rho(\mathrm{Fr}_p))p^{-s} + \det(\rho(\mathrm{Fr}_p))p^{-2s}}.$$

Thus f corresponds to ρ if and only if $L_p(s,f) = L_p(s,\rho)$ for almost all p. This can only happen when ρ is odd. I'll remark that if ρ is even, ρ should correspond to something called a *Maass form*. Similarly, you can define an L-function $L(s,E)$ for an elliptic curve E so that E is *modular* if and only if $L(s,E) = L(s,f)$, but we'll do something a little different.

2. There and Back Again

Let it be known that E is a semistable elliptic curve over $\mathbb{Q}$. The goal is to prove that E is modular. Recall we have associated to the l-torsion points of E a Galois representation $\rho_{E,l} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{Z}_l)$. This gives a residual representation $\overline{\rho}_{E,l} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{F}_l)$. We'll say that $\rho_{E,l}$ is *residually modular (of weight two)* if $\rho_{E,l}$ (more or less) corresponds to a weight-two normalized eigenform f mod l, i.e., that Equation (1) holds mod l for nearly all p. In this case we'll say that $\overline{\rho}_{E,l}$ is *modular (of weight two)*.

Theorem 1. *(Wiles) If $\overline{\rho}_{E,3}$ is irreducible and modular (of weight two), then ρ_{E_3} (and hence E) is modular.*

(Due to Conrad, et al., you probably don't even need that E is semistable.) Pretty much, either $\overline{\rho}_{E,3}$ or $\overline{\rho}_{E,5}$ is irreducible. Using his unpatented "3–5 switch", Wiles shows it suffices to assume $\overline{\rho}_{E,3}$ is irreducible. A theorem of Langlands and Tunnell then applies to show that $\overline{\rho}_{E,3}$ is actually modular. This is wherein lies the connection with automorphic forms and what we shall discuss in the final section.

3. Why eat modular when you can have automorphic every day of the week?

The annoying thing about modular forms is their modularity. Say $f : \mathfrak{H} \to \mathbb{C}$ is a modular form on $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ of weight w. Let

$$j(g,z) = \det(g)^{-1/2}(cz+d) \ , \ g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

The modularity condition then means $f(\gamma z) = j(\gamma;z)^w f(z)$ for $\gamma \in \Gamma$. This isn't too bad if $w = 0$, but I think you'll agree we'd all be better off without this j term. So let's get rid of it.

Not only does $\mathrm{SL}_2(\mathbb{Z})$ act on $\mathfrak{H}$, so does $\mathrm{GL}_2(\mathbb{R})^+$. Note

$$Stab_{\mathrm{GL}_2(\mathbb{R})^+}\{i\} = Z \cdot K \ \ , Z = Z(\mathrm{GL}_2(\mathbb{R})^+) \ , \ K = \mathrm{SO}(2).$$

So $\mathfrak{H} \simeq Z\backslash \mathrm{GL}_2(\mathbb{R})^+/K$. Lift f to a function F on $\mathrm{GL}_2(\mathbb{R})^+$ so that

$$F(g) = f(g \cdot i) \ , \ F(zgk) = F(g), \ \ z \in Z, \ \ k \in K.$$

Let $\varphi(g) = j(g;i)^{-w}F(g)$. Then

(i) $\varphi(\gamma g) = \varphi(g)$, $\gamma \in \Gamma$

(ii) $\varphi(zg) = \varphi(g)$, $z \in Z$

(ii) $\varphi(gk_\theta) = e^{i\pi w\theta}\varphi(g)$, $k_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \in K$

(iiii) $\varphi(g)$ is an eigenfunction for the invariant differential operators $\mathcal{Z}$ on $\mathrm{GL}_2(\mathbb{R})$.

(v) for any norm on $\mathrm{GL}_2(\mathbb{R})^+$, $|\varphi(g)| \leq C||g||^r$ for some C, r.

Then $\varphi : Z\Gamma\backslash\mathrm{GL}_2(\mathbb{R})^+ \to \mathbb{C}$ is an *automorphic form* on $\mathrm{GL}_2(\mathbb{R})^+$. Condition (iiii) corresponds to holomorphy of f and (v) to holomorphy of f at ∞. If you have a good imagination, I'm sure you can guess that things go similarly for Γ a discrete subgroup of $\mathrm{GL}_2(\mathbb{R})^+$.

Since we claim to be doing number theory, we should probably get some other fields involved now. Let $\mathbb{A}$ be the adèles of $\mathbb{Q}$ so we have a restricted direct product decomposition $\mathrm{GL}_2(\mathbb{A}) = \mathrm{GL}_2(\mathbb{R}) \times \prod' \mathrm{GL}_2(\mathbb{Q}_p)$. Let $K = K_\infty K_f \subseteq \mathrm{GL}_2(\mathbb{A})$ where $K_\infty = \mathrm{O}(2)$ and $K_f = \prod \mathrm{GL}_2(\mathbb{Z}_p)$. ($K, K_\infty, K_f$ are maximal compact subgroups in their respective GL_2 ambient groups, and K_f is open.) As every Japanese 3rd grader knows,

$$\begin{aligned} \Gamma\backslash\mathrm{GL}_2(\mathbb{R})^+ &= \mathrm{GL}_2(\mathbb{Q})\backslash\mathrm{GL}_2(\mathbb{A})/K_f, \text{ so} \\ Z(\mathbb{R})\Gamma\backslash\mathrm{GL}_2(\mathbb{R})^+ &= Z(\mathbb{A})\mathrm{GL}_2(\mathbb{Q})\backslash\mathrm{GL}_2(\mathbb{A})/K_f, \end{aligned}$$

where $Z(F)$ means $Z(\mathrm{GL}_2(F))$. So our automorphic form φ is actually a function of the quotient on the right.

Pictorially, we have a parallelo-diagram

$$\begin{tikzcd}
 & \mathrm{GL}_2(\mathbb{Q})\backslash\mathrm{GL}_2(\mathbb{A}) \arrow[dl] \arrow[r, "\varphi"] & \mathbb{C} \\
Z(\mathbb{A})\mathrm{GL}_2(\mathbb{Q})\backslash\mathrm{GL}_2(\mathbb{A})/K_f \arrow[r, equal] & Z(\mathbb{R})\Gamma\backslash\mathrm{GL}_2(\mathbb{R})^+ \arrow[ur, "\varphi"'] &
\end{tikzcd}$$

Thus we may think of φ as a function of $\mathrm{GL}_2(\mathbb{A})$ such that

(o) $\varphi(zg) = \omega(z)\varphi(g)$, $z \in Z(\mathrm{GL}_2(\mathbb{A}))$, $\omega(z) = 1$
(i) [automorphy] $\varphi(\gamma g) = \varphi(g)$, $\gamma \in \mathrm{GL}_2(\mathbb{Q})$
(ii) [K-finite] $\varphi(gk_\theta k_f) = e^{i\pi w\theta}\varphi(g)$, $k_\theta \in K_\infty^+ = \mathrm{SO}(2)$, $k_f \in K_f$; and in fact, $\langle \varphi(gk) | k \in K\rangle$ is finite dimensional
(iii) [$\mathcal{Z}$-finite] $\langle X\varphi(g) | X \in \mathcal{Z}\rangle$ is finite dimensional
(iiii) [moderate growth] for any norm on $\mathrm{GL}_2(\mathbb{A})$, $|\varphi(g)| \le C||g||^r$ for some C, r.

Note φ is *smooth*, i.e., C^∞ at ∞ and locally constant at the finite places. Any smooth function $\varphi : \mathrm{GL}_2(\mathbb{A})$ satisfying these conditions (i)–(iiii) is called a *(K-finite) automorphic form* on $\mathrm{GL}_2(\mathbb{A})$. Generally, the *central character* ω in condition (o) might not be 1, just as there are modular forms with nontrivial character. We will say φ is a *cusp form* if

(v) [cuspidality]

$$\int_{\mathbb{Q}\backslash\mathbb{A}} \varphi\left(\begin{pmatrix} 1 & x \\ 0 & 1\end{pmatrix} g\right) dx = 0.$$

(Recall that classically, cuspidality states

$$a_0 = \int_0^1 f(x+iy)dx = \int_0^1 f\left(\begin{pmatrix} 1 & x \\ 0 & 1\end{pmatrix} \cdot iy\right) dx = 0.)$$

Denote the vector space of K-finite automorphic (resp., cusp) forms by $\mathcal{A}$ (resp., $\mathcal{A}_0$). Unfortunately, we don't quite get "automorphic" representations of $\mathrm{GL}_2(\mathbb{A})$ on $\mathcal{A}$ but we do get ones of a Hecke algebra. On the other hand, one can define *smooth automorphic forms* and *L^2 automorphic forms* which relax the condition of K-finiteness which do afford "automorphic" representations of $\mathrm{GL}_2(\mathbb{A})$. Using L^2 automorphic forms you can get representations of $\mathrm{GL}_2(\mathbb{A})$ on the space of K-finite cusp forms, but we won't worry about this.

$\mathrm{GL}_2(\mathbb{A})$ acts by right translation on the space of cusp forms. Given a cusp form φ which is an eigenform in some sense, let $\pi = V_\varphi$ be the representation of $\mathrm{GL}_2(\mathbb{A})$ spanned by φ. Any such representation π is called a *cuspidal automorphic representation* of $\mathrm{GL}_2(\mathbb{A})$. More generally[1], any irreducible representation of $\mathrm{GL}_2(\mathbb{A})$ on the space of cusp forms is a cuspidal automorphic representation π, but it's a big deal (called Multiplicity One) that (for GL_n) $\pi = V_\varphi$ for some cusp form φ.

When I started off writing this, I thought I could define some things and present a bit of the relevant theory, but somehow things degenerated and chaos ensued, like a Chesterton novel (or so I'm told). So don't feel bad if none of this makes sense, and if perhaps automorphy doesn't sound like such a great idea anymore. But the point is that things called automorphic forms can be defined on $\mathrm{GL}_n(\mathbb{A}_F)$ (or other algebraic groups more generally) and over any number field F, and (for GL_n) they correspond to other things called automorphic representations of $\mathrm{GL}_n(\mathbb{A}_F)$, which have meromorphic L-functions (actually entire for cuspidal representations). Langlands conjectured that any irreducible Galois representation $\rho : G_F \to \mathrm{GL}_n(\mathbb{C})$ corresponds to a cuspidal automorphic representation π of $\mathrm{GL}_n(\mathbb{A}_F)$ on some space of cusp forms (in the sense that they have L-functions which agree almost everywhere). This is called, among other things, the strong Artin conjecture and does indeed imply Artin's conjecture that $L(s,\rho)$ is entire for $\rho \ne 1$ irreducible. The Langlands-Tunnell theorem stated in the next section (and what we need) is a special case of the strong Artin conjecture.

Note that modular forms and Maass forms are essentially automorphic forms (or representations) for $n = 2$, $F = \mathbb{Q}$. In fact, an irreducible two-dimensional Galois representation ρ should correspond to a modular form if ρ is odd and a Maass form if ρ is even.

[1]By the end of this sentence, I seem to say that it's not more general at all, so I don't know why I wrote any of this.

4. Hurray hurray! Automorphy saves the day!

Theorem 2. *(Langlands-Tunnell) Let* $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{C})$ *be a continuous representation. If the image of* ρ *is solvable, then* ρ *corresponds to an automorphic representation* π *of* $\mathrm{GL}_2(\mathbb{A})$ *in the sense that* $L_p(s,\rho) = L_p(s,\pi)$ *for almost all primes* p.

This is a great theorem, and if I had time to prove it, you'd reprimand yourself for ever having doubted automophy. See for example Rogawski's article "Functoriality and the Artin Conjecture," *Proc. Symp. Pure Math.* **61** (1997). It's also available on his website.

(For those who know the background, here's a recap of Langlands's proof of the tetrahedral case. Let $\sigma : G_F \to \mathrm{GL}_2(\mathbb{C})$ be a tetrahedral representation. Then there is a normal cubic extension K/F such that σ_K is modular. Say $\sigma_K \leftrightarrow \Pi$. There are three representations π_0, π_1, π_2 of $\mathrm{GL}_2(\mathbb{A}_F)$ whose base change $\pi_{i,K}$ to K is Π. One of these should actually correspond to σ. There is a unique $\pi = \pi_i$ whose central character matches with the determinant of σ. Then one proves $\mathrm{Sym}^2(\sigma) \leftrightarrow \mathrm{Sym}^2(\pi)$. This combined with the correspondence $\sigma_K \leftrightarrow \pi_K$ allows one to conclude that, at any unramified place v, either $\sigma_v \leftrightarrow \pi_v$ or $\overline{\sigma}(\mathrm{Fr}_v) \in A_4$ has order divisible by 6. But A_4 has no elements of order 6, so in fact $\sigma \leftrightarrow \pi$.)

We want to deduce that $\overline{\rho}_{E,3}$ is modular when it is irreducible. If it is irreducible, then it is absolutely irreducible, i.e., irreducible over $\overline{\mathbb{F}}_3$. Furthermore, it is odd. Then the following result applies.

Corollary 1. *Let* $\overline{\rho} : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{F}_3)$ *be an odd, absolutely irreducible representation. Then* $\overline{\rho}$ *corresponds to a weight-two normalized eigenform* f.

I'll now try to outline how this goes. It's fortunate that $\mathrm{GL}_2(\mathbb{F}_3)$ embeds inside $\mathrm{GL}_2(\mathbb{C})$, and in a way that (more or less) respects trace and determinant. Specifically, we can define a faithful honomorphism $\psi : \mathrm{GL}_2(\mathbb{F}_3) \to \mathrm{GL}_2(\mathbb{C})$ by

$$\psi\begin{pmatrix}-1 & 1\\ -1 & 0\end{pmatrix} = \begin{pmatrix}-1 & 1\\ -1 & 0\end{pmatrix} \quad , \quad \psi\begin{pmatrix}1 & -1\\ 1 & 1\end{pmatrix} = \begin{pmatrix}1 & -1\\ -i\sqrt{2} & -1+i\sqrt{2}\end{pmatrix}.$$

Then in fact $\psi : \mathrm{GL}_2(\mathbb{F}_3) \to \mathrm{GL}_2(\mathbb{Z}(i\sqrt{2}))$. Note $\mathbf{3} = (1 - i\sqrt{2})$ is a prime of $\mathbb{Z}(i\sqrt{2})$ above 3 (since $(1-i\sqrt{2})(1+i\sqrt{2}) = 3$) and you can check that

$$\mathrm{tr}(\psi(g)) \equiv \mathrm{tr}(g) \mod \mathbf{3} \quad , \quad \det(\psi(g)) \equiv \det(g) \mod 3.$$

Now we can extend $\overline{\rho}$ to a representation $\rho : G_{\mathbb{Q}} \to \mathrm{GL}_2(\mathbb{C})$ as

$$\begin{tikzcd} G_{\mathbb{Q}} \arrow[r, "\rho"] \arrow[dr, "\overline{\rho}"'] & \mathrm{GL}_2(\mathbb{C}) \\ & \mathrm{GL}_2(\mathbb{F}_3) \arrow[u, "\psi"'] \end{tikzcd}$$

Note that ρ is an odd, continuous, irreducible Galois representation with solvable image. It is odd because $\overline{\rho}$ is odd and ψ preserves determinants mod 3. It is continuous because it evidently has finite image. It's irreducible because its image is non-abelian (or else $\overline{\rho}$ would not be absolutely irreducible). It has solvable image because $\mathrm{PGL}_2(\mathbb{F}_3) \simeq S_4$ (and hence $\mathrm{GL}_2(\mathbb{F}_3)$) is solvable.

By the Langlands-Tunnell theorem, ρ corresponds to some cuspidal automorphic representation π of $\mathrm{GL}_2(\mathbb{A})$. So in fact ρ corresponds to a weight-one eigenform f. So $\overline{\rho}$ corresponds to f mod $\mathbf{3}$. We want to show that $\overline{\rho}$ corresponds to a normalized eigenform of weight two. The idea is to multiply f by an Eisenstein series of weight one. Let χ be the "mod 3" character, and

$$E(z) = E_{1,\chi}(z) = 1 + 6\sum_{n=1}^{\infty}\sum_{d|n}\chi(d)e^{2\pi i n z}.$$

Then $E \equiv 1 \mod \mathbf{3}$ (i.e., each Fourier coefficient except for the constant term is 0 mod $\mathbf{3}$), so $g = fE$ is a normalized weight-two form. However, it's highly unlikely that g is actually an eigenform, but it will be a "mod $\mathbf{3}$ eigenform," meaning that $T_n g \equiv T_n f \equiv a_n f \equiv a_n g \mod \mathbf{3}$ for all n. A result of Deligne and Serre, which I won't state, applies in this case to say there's another normalized weight-two form h which is an eigenform and $h \equiv g \mod \mathbf{3}$ (i.e., their Fourier coefficients are the same mod $\mathbf{3}$). Then h is the desired modular form.