An Introduction to Modular Symbols

- This is a preparatory talk for Rob Harron's talk; he will talk about overconvergent modular symbols and families of p-adic modular forms.
- The goal of this talk is to review the relevant aspects of the classical theory of modular forms, and then discuss the basics of modular symbols.

1 The Classical Theory [in brief]

- Classical theory dates to early 1800s, connected with elliptic functions. Developed much more in the 1920s by Hecke and others. Classical modular forms live entirely over $\mathbb{C}$.
- Serre, 1970s, developed ad hoc theory of modular forms in a p-adic setting, by p-adically interpolating Eisenstein series.
- Hida expanded more in mid-1980s, but still had some limitations. Hida constructed families ("Hida families") of p-adic cuspforms varying with the weight k which were eigenforms for the (p-adic) Hecke operators.
  - In analogy with classical theory, also get Galois representations.

1.1 Modular Forms

- Holomorphic will always mean holomorphic on the upper half-plane.
- Let $\Gamma_0(N) = \left\{\gamma \in SL_2(\mathbb{Z}) : \gamma \equiv \begin{pmatrix} * & * \\ 0 & * \end{pmatrix} \bmod N\right\}$, $\Gamma_1(N) = \left\{\gamma \in SL_2(\mathbb{Z}) : \gamma \equiv \begin{pmatrix} 1 & * \\ 0 & 1 \end{pmatrix} \bmod N\right\}$. The latter are the <u>principal congruence subgroups</u>; in general we call $\Gamma \subseteq SL_2(\mathbb{Z})$ a congruence subgroup if it contains some $\Gamma_1(N)$.
- For $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL_2(\mathbb{Z})$, f holomorphic, the weight-k <u>slash operator</u> is $(f\,|_k\,\gamma)(z) = (cz+d)^{-k} f(\gamma z)$, where k, $N \in \mathbb{N}$, and $\gamma z = \dfrac{az+b}{cz+d}$ in the usual way.
- A <u>modular form</u> (resp. <u>cuspform</u>) of weight k and level N is a holomorphic f invariant under the weight k action of $\Gamma_1(N)$ and which is "holomorphic at ∞" (resp. vanishes at ∞) – i.e., such that $\lim_{y \to \infty}(f\,|_k\,\gamma)$ is finite (resp. 0) for all $\gamma \in SL_2(\mathbb{Z})$.
- Since in particular a modular form of any weight and level is invariant under translation by 1, it has a Fourier decomposition $f(\tau) = \sum_{n=0}^{\infty} c_n(f) \cdot q^n$where $q = e^{2\pi i \tau}$. Holomorphicity implies that the negative coefficients vanish; a cuspform also has zero constant term.
- Collection of modular forms of wt. k and level N is $\mathcal{M}_k(N)$, and of cuspforms is $\mathcal{S}_k(N)$. They are finite-dimensional vector spaces.
  - Constant functions are the only forms of weight 0. Generally we will assume that $k \geq 1$ since 0 is boring.
  - Eisenstein series $E_{2k} = \sum_{(m,n) \in \mathbb{Z}^2 \setminus 0} \dfrac{1}{(m+n\tau)^{2k}}$ for $k > 2$ give simple examples of positive weight.
  - I would be remiss if I didn't also write down $\Delta = g_2^3 - 27g_3^2$, where $g_2 = 60G_4$ and $g_3 = 140G_6$; Δ is a cuspform of weight 12.
- As an cultural remark, for us the weight of a modular form will always be an integer or (if you're more highbrow) something p-adic. But one can also talk about modular forms of non-integral weight (e.g., if you talk to the group of people who study stuff like mock theta functions: they often have modular forms or things like modular forms that have half-integral weight), and this theory is also interesting.

- The definition of the slash operator needs to be modified, in such cases: for example, if the congruence subgroup contains $-I$, then in the definition above one obtains $f(z) = (-1)^k f(z)$, which forces f to be identically zero. One can fix this by introducing a multiplicative character into the mix; this will take care of "root of unity" issues from the Γ-invariance

1.2 Hecke Operators

- Have $\Gamma_1(N) \trianglelefteq \Gamma_0(N)$, and the map sending $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$ to d is an isomorphism of the quotient with $(\mathbb{Z}/N\mathbb{Z})^\times$. The wt-$k$ action of $\Gamma_0(N)$ preserves $\mathcal{M}_k(N)$ and $\mathcal{S}_k(N)$ hence we get an action of the quotient on each.
- For $d \in (\mathbb{Z}/N\mathbb{Z})^\times$ we write the corresponding automorphism of $\mathcal{M}_k(N)$ as $\langle d \rangle$, sometimes called the "diamond operator". (Probably nothing to do with Diamond.)
- If l is a prime not dividing N we define the automorphism $S_l = \langle l \rangle \cdot l^{k-2}$. [This is the l-diamond operator on f, times l^{k-2}.]
- The lth <u>Hecke endomorphism</u> T_l on $\mathcal{M}_k(N)$ is defined as $(T_l\, f)(\tau) = \sum_{n=0}^{\infty} c_{l \cdot n}(f) q^n + \sum_{n=0}^{\infty} l \cdot c_n(S_l f) q^{l \cdot n}$. [Not obvious that this actually preserves $\mathcal{M}_k(N)$, but it does, and it preserves $S_k(N)$ too.]
  - Lots of other ways to define the Hecke operator T_l: the geometric interpretation is that it's (essentially) a sum over all sublattices of index l.
  - Algebraically, it can be also written as a particular double coset operator.
  - Definition extends to composite values of l fairly easily, but skipped for brevity.
- The <u>Hecke algebra</u> $\mathbb{T}_k(N)$ of weight k and level N is the $\mathbb{Z}$-subalgebra of $\mathrm{End}(\mathcal{M}_k(N))$ generated by the operators $l\, S_l = \langle l \rangle\, l^{k-1}$ and T_l, as l ranges over all primes not dividing N.
  - $\mathbb{T}_k(N)$ is commutative, reduced, and free of finite rank over $\mathbb{Z}$. The tensor product $\mathbb{T}_k(N) \otimes \mathbb{C}$ also acts faithfully on $\mathcal{M}_k(N)$.
  - All eigenvalues of the Hecke operators are algebraic integers. The systems of simultaneous eigenvalues for all the Hecke operators on $\mathcal{M}_k(N)$ are closed under $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$. The space $\mathcal{M}_k(N)$ is the direct sum of simultaneous eigenspaces for Hecke operators.
- We call f a <u>Hecke eigenform</u> if it is an eigenvector for both $l\, S_l$ and T_l for all prime $l \nmid N$. This is equivalent to the existence of a ring hom $\lambda : \mathbb{T}_k(N) \to \mathbb{C}$ such that $T\, f = \lambda(T)\, f$ for all $T \in \mathbb{T}_k(N)$.
  - Hecke eigenforms are "nice" – for example, knowing the first-degree term in the Fourier expansion allows one to determine all the Fourier coefficients recursively. It would, for example, be generally useful if we had a recipe for writing down lots of Hecke eigenforms.
  - <u>Example</u>: Eisenstein series are Hecke eigenforms for $N = 1$. Their eigenvalues are $\lambda(lS_l) = l^{2k-1}$, $\lambda(T_l) = 1 + l^{2k-1}$.

1.3 Galois Representations (for cultural purposes only)

- <u>Theorem</u>: For any prime p, let Σ be the primes dividing pN and $\mathbb{Q}_\Sigma$ be the maximal extension unramified outside Σ. Then there is a continuous, semisimple representation $\rho_\lambda : \mathrm{Gal}(\mathbb{Q}_\Sigma/\mathbb{Q}) \to GL_2\left(\overline{\mathbb{Q}_p}\right)$ arising from $\mathbb{T}_k(N)$ uniquely determined up to equivalence by the condition that for each prime l not in Σ, the image of Frobenius – i.e., $\rho_\lambda(\mathrm{Frob}_l)$ – has characteristic polynomial given by $X^2 - \lambda(T_l)X + \lambda(lS_l)$.
  - The uniqueness of the representation is not very hard: Chebotarev implies that the union of the conjugacy classes is dense in $G_{\mathbb{Q},\Sigma}$, and so the characteristic polynomials of two such representations agree on a dense set (hence everywhere by continuity). Then, use the fact that a semisimple finite-dimensional representation of a group is uniquely determined up to equivalence by its characteristic polynomials.

- Existence is more work: part: the proof is due to Deligne ($k > 2$), Eichler-Shimura-Igusa ($k = 2$), and Deligne-Serre ($k = 1$). In the case $k = 2$ this representation really is familiar, as weight-2 modular forms correspond to elliptic curves, and the resulting Galois representation is just the usual one coming from the Tate module. For higher weights the construction involves the cohomology of particular moduli spaces.
- The polynomial $X^2 - \lambda(T_l)X + \lambda(lS_l)$ is called the lth Hecke polynomial of λ. It should be familiar from (e.g.,) elliptic curves; in that setting it is the characteristic polynomial of Frobenius and also the numerator of the curve's zeta function.

2 From Modular Forms to Modular Symbols

- Now I would like to explain how to go between modular forms as we usually think of them ("functions on the upper half-plane with a q-series") and modular symbols ("something else").
- From the basics of de Rham cohomology, we know that (up to the correct choice of coefficients), singular cohomology is "the same" as (algebraic) de Rham cohomology.
- In the particular case of the modular curve $X_0(N) = \Gamma_0(N)\backslash\mathfrak{h}$, this says $H_1(X_0(N), \mathbb{Q})\otimes\mathbb{C} \cong H^1_{dR}(X_0(N)/\mathbb{Q})\otimes\mathbb{C}$, the isomorphism being via Poincare duality, or, if you like being explicit, by the pairing $H_1(...)\times H^1_{dR}(...) \to \mathbb{C}$ given by integrating differential forms in the latter over cycles in the former.
  - As $\mathbb{Q}$-vector spaces, these two (co)homology groups $H_1(X_0(N), \mathbb{Q})$ and $H^1_{dR}(X_0(N)/\mathbb{Q})$ are isomorphic, but not the same (as subspaces of $\mathbb{C}^{\square}$), due to the presence of some transcendental periods.
- For weight 2 modular forms, the invariant differentials look like $\omega_f = f(z)\,dz$ – standard homework exercise: verify that the automorphy factor from f cancels the one from dz after applying $\gamma \in SL_2(\mathbb{Z})$. When we integrate ω, we land in $\mathbb{C}/\Lambda$, where $\Lambda = \langle\omega_1, \omega_2\rangle$ for some complex numbers ω_1 and ω_2 – this is just the usual story about complex tori and elliptic curves.
- For weight k, in order to make the transformations work out correctly, what is instead needed is to set $\omega_f = f(z)\cdot(zX+Y)^{k-2}\,dz$, where $X, Y \in Sym^{k-2}\mathbb{C}^2$.
  - Reminder: $Sym^g\mathbb{C}^2$ can be thought of as the space of homogeneous polynomials in 2 variables of degree g. The action of $SL_2(\mathbb{Z})$ is given by $(P|\gamma)(X,Y) = P(dX - cY, -bX + aY)$ – i.e., by the "adjugate" (sometimes called the "classical adjoint").
- For fun, let's check this works: for $\gamma = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in SL_2(\mathbb{Z})$, we have
  - $\gamma : f(z) \to (cz+d)^k f(z)$, by definition
  - $\gamma : dz \to (cz+d)^{-2}\,dz$, by "this is always true"
  - $\gamma : (zX+Y)^{k-2} \to \left[\frac{az+b}{cz+d}(\gamma X) + (\gamma Y)\right]^{k-2} = (cz+d)^{2-k}\cdot[(az+b)(dX-cY) + (cz+d)(-bX+aY)]^{k-2} = (cz+d)^{2-k}(zX+Y)^{k-2}$
  - and indeed if we multiply all of those things together, we see that $f(z)\cdot(zX+Y)^{k-2}\,dz$ is invariant under $SL_2(\mathbb{Z})$.
- So, if we denote $V = Sym^{k-2}\mathbb{C}^2$, we want to study elements in $H_1(\Gamma_0(N),\ V)$.
- Now, Eichler-Shimura almost says that $H_1(\Gamma_0(N),\ V) \cong M_k(\Gamma_0(N)) \oplus \bar{M}_k(\Gamma_0(N))$.
  - But... in Eichler-Shimura we actually want $M_k \oplus S_k$, so in fact we don't want to use the regular singular cohomology H_1: the correct cohomology theory is actually what is called "parabolic cohomology", the flavor of which is "the image of cohomology with compact support, in regular cohomology".
  - Explicitly, $H^1_p(\Gamma_0(N), Sym^{k-2}\mathbb{C}^2) = \text{image}\left[H^1_{cpt}(Y_0(N), V) \to H^1(Y_0(N), V)\right]$. (We're using cohomology with compact support, so we also compactify our space to $Y_0(N)$... for some reason.)

- That last guy is basically $H^1_{cpt}(\Gamma\backslash\bar{\mathfrak{h}}, \tilde{V})$, where $\tilde{V}$ is the locally constant sheaf associated to V on $\Gamma\backslash\mathfrak{h}$, and $\bar{\mathfrak{h}} = \mathfrak{h} \cup \mathbb{P}^1(\mathbb{Q})$, the compactification of the upper-half plane (note $\bar{\mathfrak{h}}$ is just the closure under adjoining ∞ and all of its images under $SL_2(\mathbb{Z})$).
- Now excision more or less says that $H_1(\bar{\mathfrak{h}}, \mathbb{P}^1(\mathbb{Q}), \mathbb{Z}) \cong \mathrm{Div}^0(\mathbb{P}^1(\mathbb{Q}))$, because $\mathbb{P}^1(\mathbb{Q})$ is the boundary of the compactified upper half-plane.

- So, if that long chain of hazy statements about cohomology groups made any sense, the result is that we can study modular forms by studying things in $\mathrm{Hom}\left[\mathrm{Div}^0(\mathbb{P}^1(\mathbb{Q})),\ V\right]$ which behave correctly under the action of Γ.
  - Roughly, the idea is: elements of the divisor group are just sums of $[P]-[Q]$ where P and Q are rational numbers (or ∞).
  - If we are given a modular form wrt Γ, it gives a Γ-invariant element of $\mathrm{Hom}\left[\mathrm{Div}^0(\mathbb{P}^1(\mathbb{Q})),\ V\right]$, by integrating the modular form over the paths which sum to an element of $\mathrm{Div}^0(\mathbb{P}^1(\mathbb{Q}))$.

3 Modular Symbols, in general

- Okay, so, we've just seen that we want to study stuff in $\mathrm{Hom}\left[\mathrm{Div}^0(\mathbb{P}^1(\mathbb{Q})),\ V\right]$. Let's get a little bit more formal.
- Let $\Delta_0 = \mathrm{Div}^0(\mathbb{P}^1(\mathbb{Q}))$ denote the set of degree-zero divisors on $\mathbb{P}^1(\mathbb{Q})$.
  - It has the structure of a left $\mathbb{Z}[GL_2(\mathbb{Q})]$-module where $GL_2(\mathbb{Q})$ acts via fractional linear transformations.
- Let Γ be a finite-index subgroup of $PSL_2(\mathbb{Z})$ and V be a right $\mathbb{Z}[\Gamma]$-module with an addition structure of $S_0(p) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} : (a,p)=1,\ p|c,\ ad-bc \neq 0 \right\}$. In particular note that $V = Sym^{k-2}\mathbb{C}^2$ has such a structure, as noted earlier.
  - We give $\mathrm{Hom}(\Delta_0, V)$ the structure of a right Γ-module by defining $(\phi|\gamma)(D) = \phi(\gamma D)|\gamma$ for $\phi : \Delta_0 \to V$, $D \in \Delta_0$, and $\gamma \in \Gamma$.
- For $\phi \in \mathrm{Hom}(\Delta_0, V)$, we say that ϕ is a <u>V-valued modular symbol on Γ</u> if $\phi|\gamma = \phi$ for all $\gamma \in \Gamma$, and we denote the space of all V-valued modular symbols by $\mathrm{Symb}_\Gamma(V)$.
  - So, if $\phi : \Delta_0 \to V$, $\phi \in \mathrm{Symb}_\Gamma(V)$ iff $\phi(\gamma D) = \phi(D)|\gamma^{-1}$ for all $\gamma \in \Gamma$.
  - As we saw earlier, we can view modular symbols (canonically) as elements of the cohomology group $H^1_c(\mathcal{H}/\Gamma, \tilde{V})$. But we would like to get our hands on them, so we will prefer to think of them as actual explicit maps rather than cohomology classes.
- The addition structure gives us a Hecke action on the modular symbols via (e.g.,) double cosets.
  - Example: if $\Gamma = \Gamma_0(N)$ and $l \nmid N$ then $\phi|T_l = \phi|\begin{pmatrix} l & 0 \\ 0 & 1 \end{pmatrix} + \sum_{a=0}^{l-1} \phi|\begin{pmatrix} 1 & a \\ 0 & l \end{pmatrix}$.
- What we would like is to be able to describe Γ-invariant functions on Δ_0 with values in V. This is accomplished by the Manin relations.
  - Let $GL_2^+(\mathbb{Q})$ be the positive-determinant matrices. If $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ then we denote by $[\gamma]$ the oriented geodesic path in the upper half-plane which goes from $\frac{a}{b}$ to $\frac{c}{d}$. In other words, it is a semicircular path in the upper half-plane whose endpoints are $\frac{a}{b}, \frac{c}{d}$. In the usual way we think of these as 1-chains in $\mathbb{P}^1(\mathbb{Q})$, and take closures under finite formal sums of such things.
  - We denote by $Z_1 = Z_1(\mathcal{H}^*, \mathbb{P}^1(\mathbb{Q}))$ the $\mathbb{Z}$-module of such modular 1-chains.

- We have an action of $PGL_2^+(\mathbb{Q})$ on Z_1 via FLTs, and if $\beta, \gamma \in GL_2^+(\mathbb{Q})$ then $\beta \cdot [\gamma] = [\beta\gamma]$.
- Taking boundaries gives a surjective $PGL_2^+(\mathbb{Q})$-morphism $\partial : Z_1 \to \Delta_0$. Two modular chains are homologous if their images under ∂ are equal.

- Let $G = PSL_2(\mathbb{Z})$. A modular path of the form $[\gamma]$ with $\gamma \in G$ is called a unimodular path, and a formal sum of them is called a unimodular 1-chain.
  - Every modular chain is homologous to a unimodular chain. Reason: the matrix whose entries are the numerators and denominators of any consecutive terms in a continued fraction approximation to any real number always has determinant ± 1. (This is called the "Manin trick".)
    - Alternatively, to get from p/q to r/s, one could write down the portion of the Farey sequence of rationals with denominators $\leq \max(q, s)$, and use the wonderful fact that the determinants of adjacent terms are always 1.
  - Moreover, G acts transitively on unimodular paths, so we get a surjective map from $\mathbb{Z}[G] \to Z_1 \xrightarrow{\partial} \Delta_0$.
  - The kernel of this map is the left ideal $I = \mathbb{Z}[G](1 + \tau + \tau^2) + \mathbb{Z}[G](1 + \sigma)$, where $\sigma = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ and $\tau = \begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}$ of orders 2 and 3 respectively.
  - These are the so-called Manin relations, and they tell us that $\Delta_0 \cong \mathbb{Z}[G]/I$.

- Ergo, we can describe what Δ_0 looks like in the following way: write down a set of right coset representatives of $\Gamma \backslash G$, say $g_1, \cdots, g_r$. Then the relations between the generators come from I: for example, $g_i(1 + \sigma) = g_i + g_i\sigma = g_i + \gamma_{i,j} g_j \in I$ for some j and some $\gamma_{i,j} \in \Gamma$.
  - This is especially pleasant to do geometrically: all that is needed is to draw a fundamental region $\mathcal{F}$ and all of the paths joining each of the coset representatives on the real axis (and ∞), and then reduce to find a minimal set of generators by using $\{a \to b\} + \{b \to c\} = \{a \to c\}$ and the other relations.
  - The end result will be (one can show) if Γ is torsion-free, then after doing this process of reduction, essentially the only relation will be "$\partial F = 0$" (where this is summing over the paths making up the boundary of $\mathcal{F}$).

4 Now what?

- I don't want to tread too much on what Rob is going to do in his talk, but let me just outline the ideas of what comes next:
  - Above, I told the story of what happens with V-valued modular symbols when $V = Sym^{k-2}\mathbb{C}^2$. But there is no reason only to consider just these V.
  - One can, for example, study instead the V-valued modular symbols where $V = Sym^{k-2}\mathbb{Q}_p^2$. This leads to the theory of "overconvergent modular symbols", which are particular modular symbols taking values in certain spaces of (p-adic) distributions.
  - One cares about overconvergent modular symbols because, roughly, they say things about p-adic modular forms, in the same way that ordinary modular symbols say things about classical modular forms.
  - p-adic families of modular forms ("Hida families") arose from the observation that there existed families of modular forms of varying weights whose coefficients satisfied congruences modulo powers of p: the particular congruences were shown to be compatible with taking an inverse limit, hence "p-adic family".
  - It would be nice (read as: "we want and are trying to make computable") if one could say enough things about overconvergent modular symbols to be able to compute lots of Hida families.