A brief overview of modular and automorphic forms

Kimball Martin

Original version: Fall 2010
Revised version: June 9, 2016

These notes were originally written in Fall 2010 to provide a very quick overview of some basic topics in modular forms, automorphic forms and automorphic representations. I have not made any significant changes since, or even proofread them completely (so some information may be outdated, and errors may remain), mostly just corrected some typos. If you spy any more errors, or have suggestions, please let me know.

The main sources used in the preparation of these notes were Zagier's notes in *The 1-2-3 of Modular Forms*, Kilford's book on modular forms, Cogdell's Fields Institute notes on automorphic forms and representations, and my brain. I've since written up course notes on modular forms, if you want to start studying these some of things in more detail. For automorphic forms, there are also some links to more sources on my automorphic representations course page. (I wrote some incomplete notes for that course, but they don't get to automorphic forms.)

1 Modular Forms

Let $\mathfrak{H} = \{z \in \mathbb{C} | \mathrm{Im}(z) > 0\}$ denote the upper half-plane. Imbued with the metric $ds^2 = \frac{dx^2+dy^2}{y^2}$, this is a standard model for the hyperbolic plane. We do not need a great understanding of the geometry of $\mathfrak{H}$ to say what modular forms are, but for your peace of mind here are some basic facts:

1. The distance between any two points in $\mathfrak{H}$ is finite.
2. Angles in $\mathfrak{H}$ are given by Euclidean angles.
3. The distance from any point in $\mathfrak{H}$ to the point at infinity $i\infty$ is infinite.
4. The distance from any point in $\mathfrak{H}$ to any point on the real line $\mathbb{R}$ is infinite. In fact the "points at infinity" for $\mathfrak{H}$ are precisely $\mathbb{R} \cup \{i\infty\}$.
5. The straight lines, or geodesics, in $\mathfrak{H}$ are precisely the Euclidean vertical lines and semicircles with center on $\mathbb{R}$ that meet $\mathbb{R}$ orthogonally.
6. Any $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R})$ defines an isometry of $\mathfrak{H}$ given by

$$z \mapsto \gamma z = \frac{az+b}{cz+d}. \tag{1}$$

Note $\begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} z = \frac{-z}{-1} = z$, i.e., $-I$ acts trivially on $\mathfrak{H}$. In fact, $\mathrm{PSL}_2(\mathbb{R}) = \mathrm{SL}_2(\mathbb{R})/\{\pm I\}$ is the group of all orientation-preserving isometries of $\mathfrak{H}$.

In number theory, the most important groups of isometries are the *congruence (or modular) subgroups*

$$\Gamma_0(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{PSL}_2(\mathbb{Z}) | c \equiv 0 \mod N \right\},$$

for $N \in \mathbb{N}$. (We view elements of $\mathrm{PSL}_2(\mathbb{R})$ or $\mathrm{PSL}_2(\mathbb{Z})$ as 2×2 matrices in $\mathrm{SL}_2(\mathbb{R})$ or $\mathrm{SL}_2(\mathbb{Z})$, up to a $\pm$ sign.) We call N the *level* of $\Gamma_0(N)$. Note the congruence subgroup of level 1, $\Gamma_0(1) = \mathrm{PSL}_2(\mathbb{Z})$, which is called the *full modular group*.

The $\Gamma_0(N)$ equivalence classes of the points at infinity $\mathbb{Q} \cup \{i\infty\}$ are called the *cusps* of $\Gamma_0(N)$. The number of cusps will always be finite. For $N = 1$, there is only one cusp, which we denote $i\infty$.

Definition 1.1. *Let $f : \mathfrak{H} \to \mathbb{C}$ be a holomorphic function and $k \in \mathbb{N} \cup \{0\}$. We say f is a* (holomorphic) modular form of weight k *and* level N *if*

$$f(\gamma z) = (cz+d)^k f(z) \text{ for } z \in \mathfrak{H},\ \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N), \tag{2}$$

and f is "holomorphic at each cusp" of $\Gamma_0(N)$. *Denote the space of modular forms of weight k and level N by $M_k(N)$.*

We will not explain precisely the notion of being holomorphic at a cusp, but simply say that it means there is a reasonable (in fact polynomial) growth condition on $f(z)$ as z tends to a point at infinity for $\mathfrak{H}$.

One may also define more general spaces of modular forms by generalizing the *modular transformation law* (2). Precisely, one may consider modular forms for an arbitrary discrete subgroup Γ of $\mathrm{PSL}_2(\mathbb{Z})$ by replacing $\Gamma_0(N)$ with Γ in (2). One may also consider *modular forms with character* χ as satisfying

$$f(\gamma z) = \chi(d)(cz+d)^k f(z) \text{ for } z \in \mathfrak{H},\ \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N),$$

where χ is a Dirichlet character modulo N.

Fourier expansion

Let $f \in M_k(N)$. Note $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \in \Gamma_0(N)$ for all N. Then (2) with $\gamma = T$ simply says $f(z+1) = f(z)$, i.e., f is periodic. Hence it has a Fourier expansion

$$f(z) = \sum_{n \in \mathbb{Z}} a_n e^{2\pi i n z}.$$

We put $q = e^{2\pi i z}$ and $F(q) = f(z)$. Note as $z \to i\infty$, $q \to 0$. Hence the Fourier expansion (or q-expansion)

$$f(z) = F(q) = \sum a_n q^n$$

may be alternatively viewed as a "power series" expansion of $F(q)$ at $q = 0$, i.e., it is a "power series" expansion of $f(z)$ at the cusp $z = i\infty$. Here "power series" is in quotes because we allow negative exponents n also. In fact, a priori, from the Fourier expansion a_n may be nonzero for infinitely many negative n. However, the condition that $f(z)$ is holomorphic at the cusp $z = i\infty$ means that $a_n = 0$ for $n < 0$.[1] Thus the q-expansion above is indeed the power series expansion for $f(z)$ at $z = i\infty$.

As mentioned above, in the case $N = 1$, $z = i\infty$ is the only cusp of $\Gamma_0(1) = \mathrm{PSL}_2(\mathbb{Z})$. But for higher levels N, $\Gamma_0(N)$ has multiple cusps and there is a similar Fourier or q-expansion $\sum_{n\geq 0} a_n q^n$ about each cusp.

Definition 1.2. *Let $f \in M_k(N)$. We say f is a* cusp form *if f vanishes at each cusp, i.e., if $a_0 = 0$ in the q-expansion*

$$f(z) = \sum_{n\geq 0} a_n q^n$$

about any cusp of $\Gamma_0(N)$. The space of cusp forms in $M_k(N)$ is denoted $S_k(N)$.

Note for (most) $N > 1$, to check if f is a cusp form we need to check the constant term of multiple Fourier expansions, not just one. Cusp forms are the most interesting modular forms, and their Fourier coefficients provide arithmetic information, as we will see below.

Algebraic structure

Note that if $f, g \in M_k(N)$ and $c \in \mathbb{C}$, then $cf + g \in M_k(N)$. Hence $M_k(N)$ is a $\mathbb{C}$-vector space. An important fact is that for fixed k, N, the space $M_k(N)$ is finite dimensional.

If $f \in M_k(N)$ and $g \in M_\ell(N)$, then it is easy to see $f \cdot g \in M_{k+\ell}(N)$.

Note that if $M|N$, then $\Gamma_0(N) \subseteq \Gamma_0(M)$. So if $f \in M_k(M)$, the modular transformation law (2) for $\Gamma_0(M)$ automatically gives the transformation law for $\Gamma_0(N)$, i.e., we also know $f \in M_k(N)$. Hence we always have $\dim M_k(N) \geq M_k(M)$.

All of the above remarks apply equally to the space of cusp forms: $S_k(N)$ is a finite dimensional $\mathbb{C}$-vector space; the product of two cusp forms is a cusp form whose weight is a sum of the individual weights; and $S_k(M) \subseteq S_k(N)$ for $M|N$. When studying $S_k(N)$, one is often most interested in forms which don't come from a smaller level n in this trivial way. These "new forms" can be defined as follows.

One can make $S_k(N)$ a Hilbert space with the *Petersson inner product*

$$\langle f, g\rangle = \int\int_{\Gamma_0(N)\backslash\mathfrak{H}} f(z)\overline{g}(z) y^{2k-2} dx dy.$$

Let $S_k^{old}(N)$ be the subspace of $S_k(N)$ spanned by elements of $S_k(M)$ with $M|N$, $M \neq N$.[2] Using the Petersson inner product, we can define $S_k^{new}(N)$ to be its orthogonal complement, so that

$$S_k(N) = S_k^{old}(N) \oplus S_k^{new}(N).$$

Forms in $S_k^{old}(N)$ are called *old forms* and forms in $S_k^{new}(N)$ are called *new forms.*

Examples in level 1

[1] Some people also consider modular forms where the a_n's may be nonzero for finitely many $n < 0$. These are called *weakly holomorphic modular forms.*

[2] I'm being a bit imprecise here. The obvious embedding $S_k(M) \subseteq S_k(N)$ is not the only one. If $d|\frac{N}{M}$, then $f(z) \mapsto f(dz)$ is another such embedding, and the old space is the span of the images of these different embeddings.

For $k \geq 4$ even, the *Eisenstein series* (of weight k and level 1)

$$E_k(z) = \frac{1}{2} \sum_{\substack{c,d \in \mathbb{Z} \\ \gcd(c,d)=1}} \frac{1}{(cz+d)^k} \in M_k(1).$$

It has Fourier expansion

$$E_k(z) = \frac{(2\pi i)^k}{\zeta(k)(k-1)!} \left(-\frac{B_k}{2_k} + \sum_{n=1}^{\infty} \sigma_{k-1}(n) q^n \right),$$

where B_k is the k-th Bernoulli number and σ_{k-1} is the divisor function $\sigma_{k-1}(n) = \sum_{d|n} d^{k-1}$.
Fact: Any modular form of level 1, i.e., for $\mathrm{PSL}_2(\mathbb{Z})$, is a polynomial in E_4 and E_6. Further

$$\dim M_k(1) = \begin{cases} \lfloor k/12 \rfloor & k \equiv 2 \mod 12 \\ \lfloor k/12 \rfloor + 1 & k \equiv 0,4,6,8,10 \mod 12 \\ 0 & k \text{ odd}. \end{cases}$$

In particular, $\dim M_8(1) = 1$. But both $E_4^2, E_8 \in M_8(1)$ so they must be scalar multiples of each other. Comparing the first Fourier coefficient shows in fact $E_4^2 = E_8$. Consequently all their Fourier coefficients are equal, and this yields the following relation among divisor functions

$$\sum_{m=1}^{n-1} \sigma_3(m)\sigma_3(n-m) = \frac{\sigma_7(n) - \sigma_3(n)}{120}.$$

We also remark that because Eisenstein series are not cusp forms (the zero-th Fourier coefficient is nonzero), and the first cusp form does not occur until the first instance where $\dim M_k(1) = 2$, namely $k = 12$. One can check that

$$\Delta(z) = \frac{1}{1728} \left(E_4(z)^3 - E_6(z)^2 \right) \in S_{12}(1)$$

(and is nonzero).

One can do similar Eisenstein series constructions in higher level as well.

Next, consider *Jacobi's theta function*

$$\theta(z) = \sum_{n \in \mathbb{Z}} q^{n^2} = 1 + 2q + 2q^4 + 2q^9 + \cdots, \; q = e^{2\pi i z}.$$

One can check that this is a modular form of "weight $\frac{1}{2}$" and level 4. (Though we haven't defined forms of half-integral weight, you can interpret this as $\theta^2 \in S_1(4)$.) Combinatorially it is easy to see that

$$\theta(z)^{2k} = \sum_{n \geq 0} r_{2k}(n) q^n,$$

i.e., $\theta^{2k} \in M_k(1)$ and the Fourier coefficients $a_n = r_{2k}(n)$ are precisely the number of ways one can write n as a sum of $2k$ squares. Now one can compute a basis for $M_k(1)$ in terms of Eisenstein series, and express θ^{2k} in terms of this basis simply by check the first few Fourier coefficients of θ^{2k} (how many depends upon the dimension of $M_k(1)$). For example, one can show

$$r_4(n) = 8\left(\sigma_1(n) - 4\sigma_1(n/4)\right) = 8\sum_{\substack{d|n \\ 4\nmid d}} d.$$

(Here we interpret $\sigma(n/4)$ to be 0 if $n \not\equiv 0 \mod 4$.)

Hecke operators

For each $m \in \mathbb{N}$, Hecke defined a linear operator on the space $M_k(N)$. If $f \in M_k(N)$ with $f(z) = \sum a_n q^n$, then the *m-th Hecke operator* T_m acts as

$$(T_m f)(z) = \sum b_n q^n$$

where

$$b_n = \sum_{d|\gcd(m,n)} d^{k-1} a_{mn/d^2},$$

assuming $\gcd(m, N) = 1$. (One defines different operators if $\gcd(m, N) > 1$.) Note that $T_1 f = f$, and for a prime $p \nmid N$,

$$(T_p f)(z) = \sum_{n\,:\,p\nmid n} a_{pn} + \sum_{n\,:\,p|n}^{\infty} (a_{pn} + p^{k-1} a_{n/p}) q^n.$$

One can check that Hecke operators T_m, T_n commute when m and n are relatively prime to the level N. It is a theorem that $S_k(N)$ has a basis of *Hecke eigenforms* $\{f\}$, meaning that for each such f, we have $T_m f = \lambda_m f$ for some $\lambda_m \in \mathbb{C}$ for all m with $\gcd(m, N) = 1$. Looking at the first Fourier coefficient, we see $b_1 = a_m = \lambda_m a_1$ whenever $\gcd(m, N) = 1$.

In particular, suppose $N = 1$ and f is a nonzero eigenform. Then $a_1 \neq 0$ for a nonzero eigenform f, and we may *normalize* f by assuming $a_1 = 1$, i.e., replace f with f/a_1. Then the Hecke operator T_m simply acts as $T_m f = a_m f$, where a_m is the m-th Fourier coefficient of f. One can then conclude that the Fourier coefficients of f are *multiplicative*, i.e., for m, n relatively prime, we have $a_m a_n = a_{mn}$. This is what makes Hecke eigenforms so nice. (Something similar is true for when $N > 1$ also, but one needs to introduce some additional operators when $\gcd(m, N) > 1$ to guarantee $a_1 \neq 0$—otherwise, one can have eigenforms f such that a_n is only nonzero when $n \equiv 0 \mod p$ for some $p|N$.)

L-functions

Let $f(z) = \sum a_n q^n \in S_k(N)$. We define its *L-function* (or *L-series*) to be

$$L(s, f) = \sum_{n=1}^{\infty} \frac{a_n}{n^s}.$$

Note the similarity to the Riemann zeta function $\zeta(s) = \sum_1^\infty \frac{1}{n^s}$. By Hecke's bound $a_n = O(n^{k/2})$, one can show $L(s, f)$ converges for $Re(s) > \frac{k}{2} + 1$. Furthermore, it extends uniquely to an entire function on $\mathbb{C}$ and satisfies the functional equation

$$L(s, f) = i^k N^{k/2-s} (2\pi)^{2s-k} \frac{\Gamma(k-s)}{\Gamma(s)} L(k-s, g)$$

where $g(z) = N^{-k/2} z^{-k} f(-1/Nz)$. (If $N = 1$, i.e., $\Gamma = \mathrm{SL}_2(\mathbb{Z})$, then $g = f$.)

So far this is analogous to the Riemann zeta function (except that $\zeta(s)$ has a pole at $s = 1$), but $\zeta(s)$ has another very important feature, the *Euler product expansion*,

$$\zeta(s) = \prod_p \frac{1}{1 - p^{-s}},$$

which is valid for $Re(s) > 1$. In order to do the same trick for $L(s, f)$, we would want the Fourier coefficients a_n to be multiplicative. Well, they are if f is a Hecke eigenform, and we know such elements span $S_k(N)$. In this case, there is an Euler product expansion

$$L(s, f) = \prod_{p \nmid N} \frac{1}{1 - a_p p^{-s} + p^{k-1-2s}} \cdot \prod_{p|N} (\text{“bad factors”}).$$

Like the Riemann zeta function, L-functions occupy a central role in modern number theory. For one, L-functions allow you to compare objects of different types: for an elliptic curve E[3], we can also associate an L-function $L(E, s) = \sum \frac{b_n}{n^s}$ where the b_n's are essentially the number of points on the elliptic curve mod n (at least for n prime). Then we can say E and f are correspond if $L(E, s) = L(f, s)$. The fact that every elliptic curve (over $\mathbb{Q}$) corresponds to a modular form (of weight 2) (the Taniyama–Shimura conjecture[4], or now, Modularity Theorem) was one of the most spectacular mathematical accomplishments of the 20th century. (It's still not known in general for elliptic curves over other number fields). Moreover, the analytic properties of $L(f, s)$ give important information on the a_n's. Another important feature is that the values $L(f, s)$ at certain special values of s (e.g., the *central value* $L(f, k/2)$) carry interesting arithmetic information (e.g., about the a_n's).

A useful variant that is often studied is the *twisted L-function*. If χ is a Dirichlet character, one can consider the twist

$$L(f, s; \chi) = L(f \otimes \chi, s) = \sum \frac{\chi(n) a_n}{n^s}.$$

This is a sort of hybrid between Dirichlet's L-functions and $L(f, s)$.

Generalizations. One can generalize the notion of modular forms to functions on higher-dimensional analogues of the upper-half plane $\mathfrak{H}$. There are different ways to do this, and one ends up with different kinds of generalized modular forms such as Hilbert modular forms, Siegel modular forms and Jacobi forms. Additionally, one can consider “anti-holomorphic” analogues of modular forms called Maass forms. To differentiate the original notion of modular forms from these generalizations, one sometimes calls the modular forms we've defined *elliptic modular forms* (this terminology does not mean they all correspond to elliptic curves, however).

2 Automorphic Forms

Classical automorphic forms

[3]An elliptic curve is a (smooth) cubic curve of the form $y^2 = x^3 + ax + b$. They arise in many number theory problems. See for instance the section of sums of cubes in my notes *Sums of squares, sums of cubes, and modern number theory*, which explain part of the role elliptic curves and modular forms play in determining what numbers are sums of two (rational) cubes.

[4]A way to construct elliptic curves from weight 2 modular forms with integral Fourier coefficients was known earlier by Eichler–Shimura. The Taniyama–Shimura conjecture essentially says one gets all elliptic curves this way.

Recall the group of orientation-preserving isometries of $\mathfrak{H}$ is $G = \mathrm{PSL}_2(\mathbb{R})$ (and the action was given above). Let K be the subgroup of G stabilizing $i \in \mathfrak{H}$. It is easy to see that $K = \mathrm{SO}(2)/\{\pm I\}$. Since G acts transitively on $\mathfrak{H}$, we may in fact identify $\mathfrak{H}$ with the quotient space

$$\mathfrak{H} = G/K.$$

Let $f \in M_k(N)$ and $\Gamma = \Gamma_0(N)$. Since $f : \mathfrak{H} \to \mathbb{C}$ we may lift $f : G \to \mathbb{C}$, and f satisfies

$$f\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix} g\right) = (cz+d)^k f(g), \ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma.$$

Consider the function

$$\phi : G \to \mathbb{C}$$

given by

$$\phi\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) = (cz+d)^{-k} f\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right).$$

It is evident that $\phi(\gamma) = f(1)$ is constant for $\gamma \in \Gamma$. Moreover, one can check that

(i) [automorphy] $\phi(\gamma g) = \phi(g)$ for $\gamma \in \Gamma$; and

(ii-a) $\phi(gk_\theta) = e^{k\pi i\theta}\phi(g)$ for $k_\theta = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \in K$ if $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$.

If $\Gamma \neq \mathrm{PSL}_2(\mathbb{Z})$, the analogue of (ii-a) is more complicated, but can abstractly be described as

(ii) [K-finiteness] the vector-space $\langle \varphi_k(g) := \phi(gk) | k \in K\rangle \subset C^\infty(G)$ is finite dimensional.

Note that (ii-a) implies the vector-space $\langle \varphi_k(g) | k \in K\rangle$ is 1-dimensional, so (ii) is a generalization of (ii-a).

Definition 2.1. *Let $\Gamma \subset G$ be a discrete subgroup, e.g., $\Gamma = \Gamma_0(N)$. An* automorphic form for Γ *is a smooth function $\phi : G \to \mathbb{C}$ satisfying conditions (i) and (ii) above, as well as (iii) a differential condition and (iv) "moderate growth" condition.*

We will not worry about the details of (iii) and (iv), but just remark that they essentially correspond to the conditions of (iii') holomorphy of f on $\mathfrak{H}$ and (iv') holomorphy of f at the cusps. In fact (iii) is more general that (iii'), so that the non-holomorphic analogues of modular forms, *Maass forms*, are included in the definition of automorphic forms.

Note that any modular form f corresponds to an automorphic form ϕ, and what is going on is the following. We may view f as a function on G which is invariant under K and satisfies some transformation property for Γ. We can exchange f for a function ϕ which is invariant under Γ and satisfies some transformation property for K. Often ϕ is more convenient to work with than f, and automorphic forms generalize more naturally than modular forms, and are amenable to study with adèles (and therefore a local-global approach) and representation theory.

One can define classical automorphic forms for symmetric spaces G/K (generalizations of $\mathfrak{H}$) as follows. Let G be an algebraic group[5] over $\mathbb{R}$, which you can think of a group of matrices defined by polynomial equations. For example $G = \mathrm{GL}_n(\mathbb{R})$, $\mathrm{SL}_n(\mathbb{R})$, $\mathrm{SO}(n)$ or

$$\mathrm{Sp}_{2n}(\mathbb{R}) = \left\{ g \in \mathrm{SL}_{2n}(\mathbb{R}) : {}^t g \begin{pmatrix} & I \\ -I & \end{pmatrix} g = \begin{pmatrix} & I \\ -I & \end{pmatrix} \right\}.$$

[5]For those who know about algebraic groups, we assume affine, connected, reductive here.

This latter example is called the *symplectic group* of rank n. Let K be a maximal compact subgroup of G, and Γ a discrete subgroup of G. Then $\phi \in C^\infty(G)$ is an automorphic form for Γ if (i), (ii), (iii) and (iv) hold. *Siegel modular forms* are (essentially) automorphic forms on $\mathrm{Sp}_4(\mathbb{R})$ (or more generally $\mathrm{Sp}_{2n}(\mathbb{R})$). They are important in the theory of quadratic forms.

Adelic automorphic forms

There are basically two ways of looking at automorphic forms, the classical way described above, and the *adelic* approach. The adelic approach, while considerably more involved, has a number of advantages.

For a number field F, recall the adèles of F are the ring

$$\mathbb{A}_F = \left\{ (x_v) \in \prod_v F_v | x_v \in \mathcal{O}_{F_v} \text{ for almost all (finite) places } v \right\},$$

where $\{v\}$ is the set of places of F and F_v denotes the completion of F with respect to v. Let G be an algebraic group over F, e.g., $G = \mathrm{GL}(n)$, $G = \mathrm{PSL}(n)$, $G = \mathrm{SO}(n)$ or $G = \mathrm{Sp}(n)$. For example if $G = \mathrm{GL}(n)$, then $G(F)$ denotes $\mathrm{GL}_n(F)$ and $G(\mathbb{A}_F)$ means $\mathrm{GL}_n(\mathbb{A}_F)$.

Definition 2.2. *Let K be a maximal compact subgroup of $G(\mathbb{A}_F)$. A smooth function $\phi : G(\mathbb{A}_F) \to \mathbb{C}$ is a* (K-finite) automorphic form *if*

(i) [automorphy] $\phi(\gamma g) = \phi(g)$ for all $\gamma \in G(F)$
(ii) [K-finiteness] the space $\langle \phi(gk) | k \in K \rangle$ is finite dimensional
(iii) ϕ satisfies a differential condition
(iv) ϕ is of moderate growth.

One also looks at larger classes of *smooth automorphic forms* or L^2 *automorphic forms*, depending on the application and/or tools one wants to use.

Example: classical modular forms

Let $f \in M_k(1)$. Then we saw above this can be transformed into a classical automorphic form ϕ on $\mathrm{PSL}_2(\mathbb{R})$,

$$\phi : \Gamma \backslash \mathrm{PSL}_2(\mathbb{R}) \to \mathbb{C}$$

where $\Gamma = \mathrm{SL}_2(\mathbb{Z})$. Let $F = \mathbb{Q}$, $G = \mathrm{GL}(2)$ and $Z \simeq \mathrm{GL}(1)$ denote the center of G. A maximal compact open subgroup of $G(\mathbb{A}_\mathbb{Q})$ is $K = K_f \times \mathrm{SO}(2)$, where $K_f = \prod_{p<\infty} G(\mathbb{Z}_p)$. One has the isomorphism

$$Z(\mathbb{A})G(\mathbb{Q})\backslash G(\mathbb{A})/K_f \simeq \Gamma \backslash \mathrm{PSL}_2(\mathbb{R}),$$

from whence it follows that ϕ lifts to a (smooth) function on

$$\phi : G(\mathbb{A}) \to \mathbb{C}$$

which is left-invariant under $G(\mathbb{Q})$, i.e., ϕ satisfies the automorphy condition (i) above. As you might guess, it satisfies (ii)–(iv) also, and this provides the passage from modular forms of level 1 to adelic automorphic forms. The passage for modular forms of level N is similar: it follows because there is a subgroup $K_N \subset K_f$ such that

$$Z(\mathbb{A})G(\mathbb{Q})\backslash G(\mathbb{A})/K_N \simeq \Gamma_0(N) \backslash \mathrm{PSL}_2(\mathbb{R}).$$

One generalization of classical modular forms is that of *Hilbert modular forms.* They are easier to describe in adelic language than classical language: namely, they are just automorphic forms on $\mathrm{GL}_2(\mathbb{A}_F)$ where F is a totally real number field.

Cusp forms

The notion of a classical cusp form generalizes naturally to automorphic forms. We will skip the motivation and explain how things work for $G = \mathrm{GL}(n)$. We call a subgroup $P \subset G$ a *parabolic subgroup* if, up to conjugation, it is of the form

$$P = \left\{ \begin{pmatrix} g_1 & * & \cdots & * \\ 0 & g_2 & \cdots & * \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & g_r \end{pmatrix} | g_i \in \mathrm{GL}(n_i) \right\}$$

where $n_1 + \cdots + n_r = n$. We can decompose $P = MN$ where $M \simeq \mathrm{GL}_{n_1} \cdots \mathrm{GL}_{n_r}$ is the *Levi subgroup* and N is the *unipotent subgroup*

$$U = \left\{ \begin{pmatrix} I_{n_1} & * & \cdots & * \\ 0 & I_{n_2} & \cdots & * \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & I_{n_r} \end{pmatrix} \right\}.$$

For example, if $G = \mathrm{GL}(2)$, then up to conjugation there is one proper parabolic subgroup with corresponding Levi and unipotent

$$P = \left\{ \begin{pmatrix} * & * \\ & * \end{pmatrix} \right\}, M = \left\{ \begin{pmatrix} * & \\ & * \end{pmatrix} \right\}, N = \left\{ \begin{pmatrix} 1 & * \\ & 1 \end{pmatrix} \right\}.$$

(Technically $P = G$ is also a parabolic subgroup, in which case $M = G$ and $U = I$.) If $G = \mathrm{GL}(3)$, then up to conjugation there are two proper parabolic subgroups

$$P_1 = \left\{ \begin{pmatrix} * & * & * \\ * & * & * \\ & & * \end{pmatrix} \right\}, M_1 = \left\{ \begin{pmatrix} * & * & \\ * & * & \\ & & * \end{pmatrix} \right\}, N_2 = \left\{ \begin{pmatrix} 1 & 0 & * \\ 0 & 1 & * \\ & & 1 \end{pmatrix} \right\}.$$

and

$$P_2 = \left\{ \begin{pmatrix} * & * & * \\ & * & * \\ & & * \end{pmatrix} \right\}, M_2 = \left\{ \begin{pmatrix} * & & \\ & * & \\ & & * \end{pmatrix} \right\}, N_2 = \left\{ \begin{pmatrix} 1 & * & * \\ & 1 & * \\ & & 1 \end{pmatrix} \right\}.$$

Note $P_2 \subset P_1$, so P_1 is called a maximal parabolic. One can similarly define parabolic and unipotent subgroups for other algebraic groups G.

Definition 2.3. *Let G be an algebraic group and ϕ be an automorphic form on $G(\mathbb{A}_F)$. We say ϕ is a* cusp form *if*

$$\int_{N(F)\backslash N(\mathbb{A})} \phi(n) dn = 0$$

for all nontrivial unipotent subgroups $N \subset G$.[6]

[6] By this, I mean all N appearing in the decomposition of a proper parabolic $P = MN$, not arbitrary subgroups of unipotent matrices. Technically these N are called unipotent radicals of parabolics.

Note the integral over the quotient $N(F)\backslash N(\mathbb{A})$ makes sense since ϕ is left invariant by any $g \in G(F)$, hence $n \in N(F)$. Furthermore $N(F)\backslash N(\mathbb{A})$ is compact, so the integral necessarily converges.

Automorphic representations

Let G be an algebraic group over F and $\mathcal{A}(G(F)\backslash G(\mathbb{A}_F))$ be the space of automorphic forms on $G(\mathbb{A}_F)$. It is essentially true that $G(\mathbb{A}_F)$ acts on $\mathcal{A}(G(F)\backslash G(\mathbb{A}_F))$ by right translation.[7] Namely,

$$(g \cdot \phi)(x) = \phi(xg).$$

(This is true for L^2 automorphic forms, but there are some technicalities at the infinite places for K-finite or smooth automorphic forms. So if you want the statements to be more-or-less technically correct, just assume that we are talking about L^2 automorphic forms.) For each ϕ, we can consider the representation (π_ϕ, V_ϕ) where $V_\phi = G(\mathbb{A}_F) \cdot \phi$. These are *automorphic representations* of $G(\mathbb{A}_F)$. They are in general infinite dimensional. If ϕ is cuspidal, we say π_ϕ is a *cuspidal automorphic representation.*

One way to think of this is that we can decompose the space of cusp forms $\mathcal{A}_{\mathrm{cusp}}(G(F)\backslash G(\mathbb{A}_F))$ as

$$\mathcal{A}_{\mathrm{cusp}}(G(F)\backslash G(\mathbb{A}_F)) = \bigoplus_{\pi} V_\pi,$$

where the V_π's are the irreducible constituents under the action of $G(\mathbb{A}_F)$. Then the cuspidal automorphic representations are these V_π's. (A similar statement is true for the non-cuspidal representations.) Looking at automorphic representations is essentially the same as looking at automorphic forms, and it allows for the use of representation theory.

In this business, one typically wants to decompose global (adelic) objects into products of local objects. If π is an automorphic representation, then it decomposes into a product of local representations

$$\pi = \otimes_v \pi_v$$

where π_v is a representation of $G(F_v)$.

Note if $G = \mathrm{GL}(1)$, then $G(F)\backslash G(\mathbb{A}_F) = \mathrm{GL}_1(F)\backslash \mathrm{GL}_1(\mathbb{A}_F) = F^\times \backslash \mathbb{A}_F^\times = C_F$, the idèle class group of F! Hence automorphic forms on G are just functions on C_F, and automorphic representations are precisely the idèle class characters χ for F, which can be decomposed into a tensor product of local representations $\chi_v : F_v \to \mathbb{C}^\times$.

If $G = \mathrm{GL}(2)$, $F = \mathbb{Q}$, and ϕ comes from a classical modular eigenform $f \in S_k(N)$, consider the associated automorphic representation $\pi = \pi_\phi$. In the decomposition $\pi = \otimes_v \pi_v = \otimes_p \pi_p \otimes \pi_\infty$, one can determine π_∞ just from the weight k of f, and π_p is determined by the eigenvalue λ_p of the Hecke operator T_p acting on f: $T_p f = \lambda_p f$.

Local representations

To study the automorphic representation $\pi = \otimes_v \pi_v$ above, one wants to understand the local representations π_v. Assume v is finite, so $K_v = G(\mathcal{O}_{F_v})$ is a maximal compact subgroup of $G(F_v)$. Each π_v is an *admissible* representation of $G(F_v)$, meaning $\pi_v|_{K_v}$ has a finite-dimensional invariant subspace. Further π_v is *unramified* for almost all v, meaning $\pi_v|_{K_v}$ as a 1-dimensional invariant subspace, i.e., there is a vector $\phi_v \in \pi_v$ which is fixed under the action of K_v.

[7] One typically mods out by the center (possibly with a character) if the center is not compact, but for simplicity I'll ignore this.

To understand the local representations π_v, one first wants a classification of the irreducible admissible representations of $G(F_v)$. Let us suppose $G = \mathrm{GL}(2)$. The simplest way to construct representations of $\mathrm{GL}_2(F_v)$ is via principal series. Namely, let χ_1 and χ_2 be two characters of $F_v^\times$. Then one can define a character of the standard parabolic subgroup P by

$$\begin{pmatrix} a & \\ & b \end{pmatrix} \begin{pmatrix} 1 & x \\ & 1 \end{pmatrix} \mapsto \chi_1(a)\chi_2(b).$$

Let $\pi(\chi_1, \chi_2)$ be the induction of this character from P to $\mathrm{GL}_2(F_v)$, which is called a *principal series* representation.

The irreducible admissible representations of $\mathrm{GL}_2(F_v)$ are

- 1-dimensional representations — these never occur as local components π_v
- special representations — an irreducible component of $\pi(\chi, \chi|\cdot|^{\pm 1})$, which is not irreducible
- irreducible principal series — $\pi(\chi_1, \chi_2)$ where $\chi_1 \neq \chi_2|\cdot|^{\pm 1}$
- supercuspidal representations — irreducible representations not occurring in any principal series

The special and supercuspidal representations are ramified, and the irreducible principal series may be ramified or unramified ($\pi(\chi_1, \chi_2)$ is unramified if χ_1 and χ_2 are). All of these types of representations occur as local components π_v of an automorphic representation, and they are all infinite dimensional. (This does not mean every irreducible principal series occurs as the component of a global automorphic representation, but it is a theorem that all special and supercuspidals do.) The classification for $\mathrm{GL}(n)$ is similar, but is more complicated for other groups.

The way to define L-functions for automorphic representations $\pi = \otimes \pi_v$ is to define local L-functions $L(s, \pi_v)$ for each π_v, and set

$$L(s, \pi) = \prod L(s, \pi_v),$$

which will converge in some right-half plane. To define $L(s, \pi)$ and prove it has the desired properties (e.g., meromorphic continuation, functional equation) is much more complicated than for L-functions of classical modular forms, even for $G = \mathrm{GL}(2)$. The case of $G = \mathrm{GL}(1)$ is done in Tate's thesis, which gives one a good indication of what needs to be done for the case of $\mathrm{GL}(2)$.

In the case of unramified principal series, for $\mathrm{GL}_2(\mathbb{F}_v)$, one associates to $\pi_v = \pi(\chi_1, \chi_2)$ the *Satake parameters* $t_v = (a_v, b_v) = (\chi_1(\varpi), \chi_2(\varpi))$ where ϖ is a uniformizer for $\mathcal{O}_{F_v}$. (E.g., if $F_v = \mathbb{Q}_p$, then one can take $\varpi = p$.) Then the local L-function is defined to be

$$L(s, \pi_v) = \frac{1}{(1 - a_v q^{-s})(1 - b_v q^{-s})}$$

where q is the size of the residue field $\mathcal{O}_{F_v}/\varpi\mathcal{O}_{F_v}$.

Functoriality

Langlands proposed the general theory of automorphic representations as a way solve many problems in number theory and geometry. Two main conjectures in the *Langlands program* are (i) *modularity*, that Galois representations should correspond to automorphic representations, and (ii) *functoriality.* We will just say a little bit about functoriality. The naive idea is that if there is a homomorphism from G into H, then automorphic representations of G should transfer to automorphic representations of H. A less naive (and more correct) idea is that if there is a homomorphism from the *dual group* of G to the *dual group* of H, automorphic representations of G transfer to H. However in our examples, each group will be its own dual, so we can temporary delude ourselves into believing the more naive idea.

Let

$$G = \mathrm{GSp}(4) = \left\{ g \in \mathrm{GL}(4) : {}^t g \begin{pmatrix} & I \\ -I & \end{pmatrix} g = \lambda(g) \begin{pmatrix} & I \\ -I & \end{pmatrix} \text{ for some } \lambda(g) \in \mathrm{GL}(1) \right\}$$

and $H = \mathrm{GL}(4)$. Siegel modular forms, in addition to being viewed as automorphic representations of $\mathrm{Sp}_4(\mathbb{A}_F)$, maybe viewed as automorphic representations of $\mathrm{GSp}_4(\mathbb{A}_F)$, which is in some ways a nicer group to work with. (Unlike GSp(4), Sp(4) is not its own dual group, rather its dual group is SO(5).) Here functoriality says that embedding $\mathrm{GSp}(4) \hookrightarrow \mathrm{GL}(4)$ should yield a transfer of automorphic representations π of $\mathrm{GSp}_4(\mathbb{A}_F)$ to automorphic representations of $\mathrm{GL}_4(\mathbb{A}_F)$. This transfer is known, by Asgari and Shahidi (2006), for *generic* representations π, but is still not known for all non-generic π.[8] A consequence of this transfer would be that one can apply results about GL(4) (which are easier to prove) to representations of GSp(4).

Similarly, transfer of generic representations of classical groups (e.g., $\mathrm{SO}(n)$, $\mathrm{Sp}(2n)$) to an appropriate $\mathrm{GL}(n)$ is known (Cogdell, Kim, Piatetski-Shapiro and Shahidi, 2004). There are several applications of these cases of functoriality.

Another interesting case of functoriality of a different flavor comes from the symmetric power lifts. Take $G = \mathrm{GL}(2)$. Take a 2-dimensional vector space $V = \langle v, w \rangle$ so $\mathrm{GL}(2) \simeq \mathrm{GL}(V)$ is the group of linear isomorphisms of V with itself. Any $g \in \mathrm{GL}(2)$ acting on V also acts on the 3-dimensional vector space $\mathrm{Sym}^2(V) = \langle v \otimes v, v \otimes w, w \otimes w \rangle$, i.e, we can view $g \in \mathrm{GL}(Sym^2(V)) \simeq \mathrm{GL}(3)$. The map

$$\mathrm{Sym}^2 : \mathrm{GL}(2) \to \mathrm{GL}(3)$$

obtained in this way is called the *symmetric square* representation of GL(2). Similarly there is a *symmetric n-th power* representation

$$\mathrm{Sym}^n : \mathrm{GL}(2) \to \mathrm{GL}(n+1),$$

for each $n \in \mathbb{N}$.

Here functoriality predicts that Sym^n induces a transfer, called the *symmetric power lift*, of automorphic representations π of $\mathrm{GL}_2(\mathbb{A}_F)$ to automorphic representations, denoted $\mathrm{Sym}^n(\pi)$, of $\mathrm{GL}_{n+1}(\mathbb{A}_F)$. This is known for $n = 2, 3, 4$. (It is easy to see what the local components of $\mathrm{Sym}^n(\pi) = \otimes \mathrm{Sym}^n(\pi_v)$ should be, but the difficulty lies in showing the tensor product on the right actually occurs as an automorphic representation.) Suppose π corresponds to a classical eigen cusp form $f(z) = \sum a_n q^n$ of weight k. Then Ramanujan conjectured

$$|a_p| \leq 2p^{(k-1)/2}.$$

This was proved by Deligne (1974, for $k \geq 2$) but generalizations, such as to Maass forms or Hilbert modular forms, are still not known. However, it would follow from knowing functoriality of all symmetric powers for GL(2). What can be currently shown is

$$|a_p| \leq 2p^{(k-1)/2+7/64}$$

[8]Update: it's essentially known now by Arthur, at least for π with trivial central character.

using Sym^4 (for classical modular forms and the analogue for Maass forms and Hilbert modular forms). These bounds on Fourier coefficients have many applications in number theory.

It is perhaps worth mentioning two other important cases of functoriality: *base change* and *automorphic induction*. Suppose π is an automorphic representation of $\mathrm{GL}_n(\mathbb{A}_F)$ and E/F is a Galois extension of degree d. Then base change says π should lift to an automorphic representation of $\mathrm{GL}_n(\mathbb{A}_E)$. Conversely, if π' is an automorphic representation of $\mathrm{GL}_n(\mathbb{A}_E)$, automorphic induction says π' should lift to an automorphic representation of degree $\mathrm{GL}_{nd}(\mathbb{A}_F)$. These are known if E/F is cyclic of prime degree. Base change and automorphic induction can be used to show certain Galois representations are modular, e.g., the Langlands–Tunnell Theorem, which played a key role in Wiles' proof of Fermat's Last Theorem.