NOTES ON THE TRACE FORMULA

RAHUL KRISHNA

Abstract. These are notes for MATH 482-2, taught in Spring 2018 at Northwestern university. They are very rough, so please let me know of any comments or corrections at krishna@math.northwestern.edu.

Contents

1. Number theoretic motivation

This first lecture will be mainly for motivation, so it may be a little all over the place. We will start from basics next time.

1.1. **Why should we care about the trace formula?** A fundamental (the fundamental?) problem in algebraic number theory is to find a good description of the absolute Galois group $\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$. Of course, this is some complicated profinite group, so what we mean by "a good description" can't really be as naive as a list of generators and relations. Rather, we would like to understand the category of all representations of $\mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$. For 1-dimensional representations, class field theory gives the entire story. Recall that the main statement of global CFT is the existence of the Artin reciprocity isomorphism

$$\mathrm{Art} : \pi_0(K^\times\backslash\mathbb{A}_K^\times) \to \mathrm{Gal}(\bar{K}/K)^{\mathrm{ab}}$$

for any number field K–thus, for instance, all characters $\chi : \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \to \mathbb{C}^\times$ can be thought of as finite order Hecke characters $\chi : \mathbb{Q}\backslash\mathbb{A}_{\mathbb{Q}}$ (which are essentially the same as multiplicative characters $\chi : (\mathbb{Z}/N)^\times \to \mathbb{C}^\times$).

For higher dimensional representations, the picture, as described by Langlands' notion of reciprocity, is still largely conjectural. Very roughly, this says that there is a natural bijection between

$$\left\{ \begin{array}{c} \text{irr. cont. representations} \\ \rho : \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_n(\mathbb{C}) \end{array} \right\} \leftrightarrow \left\{ \begin{array}{c} \text{cuspidal automorphic representations} \\ \pi \subset L^2(\mathrm{GL}_n(\mathbb{Q}) \backslash \mathrm{GL}_n(\mathbb{A}_{\mathbb{Q}})) \end{array} \right\}.$$

Let's not worry about what the objects on the RHS are for now–the precise statement of Langlands' general reciprocity conjecture is not our focus right now (in fact, the form I have given above is provably wrong, so is not a very good conjecture...).

Still, there is at least one relatively concrete and *known* case of this philosophy.

Theorem 1.1 (Deligne-Serre, Khare, Taylor, Langlands, Tunnell...)**.** *There is a canonical bijection between the two sets*

$$\left\{ \begin{array}{c} \textit{irr. cont. representations} \\ \rho : \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(\mathbb{C}) \\ \textit{s.t. } \det \rho(c) = -1 \textit{ and } \mathfrak{f}(\rho) = N \end{array} \right\} \leftrightarrow \left\{ \begin{array}{c} \textit{wt. 1 holomorphic eigen-cusp forms} \\ f(z) : \mathbb{H} \to \mathbb{C} \\ \textit{of level } \Gamma_1(N) \end{array} \right\}$$

where $c \in \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q})$ denotes complex conjugation, and $\mathfrak{f}(\rho)$ denotes the Artin conductor of ρ. This bijection matches Hecke eigenvalues on the RHS to Frobenius traces on the LHS.

Remark 1.2. Of course, both sides consist of objects up to isomorphism. On the LHS, this means up to equivalence, while on the RHS this means up to scaling by a constant.

Remark 1.3. Okay, maybe I should explain what some (but not all) of these terms mean.

- Let me skip the precise definition of the Artin conductor, and say it simply checks in each decomposition group $\mathrm{Gal}(\bar{\mathbb{Q}}_p/\mathbb{Q}_p)$ how far into inertia one has to go to make the representation trivial.
- $\mathbb{H} = \{z \in \mathbb{C} : \Im(z) > 0\}$ is the upper half plane, with the hyperbolic metric $ds^2 = \frac{1}{y^2}(dx^2 + dy^2)$. Its group of (orientiation preserving) isometries is $\mathrm{PSL}_2(\mathbb{R})$. Isometries act by fractional linear transformations.
- $\Gamma_1(N) := \{\gamma \in \mathrm{SL}_2(\mathbb{Z}) : \gamma \bmod N \text{ is upper uni-triangular}\}$. f is a holomorphic weight k form for $\Gamma_1(N)$ if
$$f(\frac{az+b}{cz+d}) = (cz+d)^k f(z) \text{ for all } \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_1(N)$$
- We say $f(z)$ is a cusp form if its local expansion around every cusp has no constant term.
- The notion of eigenform is roughly as follows: there is a commutative algebra of operators acting on modular forms, which commute with everything in sight. Eigenforms are those modular or cuspidal forms which are simulateous eigenvectors for all the Hecke operators (including the diamond operators).
- As for writing down the precise definition of the Hecke operators, let me skip this for now. We will have to talk about them very seriously later.

If you have never seen a statement like this, it should be striking–it relates certain 2-dimensional Galois representations to objects from complex analysis (or, if you prefer, complex algebraic geometry–holomorphic cusp forms are all sections of various bundles on algebraic curves).

One may ask what happens when we look at Galois representation with the opposite condition on $\rho(c)$. (Note since $c^2 = 1, \det \rho(c) = \pm 1$). There is an even crazier conjectural answer.

Conjecture 1.4 (Langlands)**.** *There is a canonical bijection between the two sets*

$$\left\{ \begin{array}{c} \textit{irr. cont. representations} \\ \rho : \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_2(\mathbb{C}) \\ \textit{s.t. } \det \rho(c) = 1 \textit{ and } \mathfrak{f}(\rho) = N \end{array} \right\} \leftrightarrow \left\{ \begin{array}{c} \textit{smooth eigen-cuspidal functions} \\ f(z) : \Gamma_1(N) \backslash \mathbb{H} \to \mathbb{C} \\ \textit{s.t. } \Delta f = \frac{1}{4} \textit{ for } \Delta = -y^2 \left(\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial x^2} \right) \end{array} \right\}$$

Once again, "eigen" means a simultaneous eigenfunction for the Hecke operators. This bijection matches Hecke eigenvalues on the RHS to Frobenius traces on the LHS.

Remark 1.5. The objects on the RHS are called *eigenvalue* 1/4 *Maass forms.*

This should be absolutely shocking! In Theorem 1.1, at least the objects on the RHS can be interpreted algebro-geometrically, which offers some connection with Galois representations through étale cohomology constructions. But in Conjecture 1.4, the objects on the RHS are *not even complex analytic*–they are weird real analytic eigenfunctions of the Laplacian! More stylishly: these are vibrating membranes on the Riemann surface $\Gamma_1(N)\backslash\mathbb{H}$. What do they have to do with Galois representations?

This should give some evidence for why, as a number theorist, it is worthwhile to understand these objects from harmonic analysis. The Selberg trace formula is the way to do this.

1.2. **The spectrum of the Laplacian on $\Gamma\backslash\mathbb{H}$.** So let us continue to be a bit informal. We will give proofs at a latter date (when we start to work systematically) for all of the following implicit results.

Suppose Γ is a discrete f.g. cofinite volume subgroup of $\mathrm{PSL}_2(\mathbb{R})$, e.g. $\Gamma = \Gamma_1(N)$. We can consider the quotient $Y(\Gamma) = \Gamma\backslash\mathbb{H}$; if Γ has no finite order elements besides the identity, then this happens to be a Riemann surface. If there are finite order elements, then it only just fails to be a Riemann surface–it is an orbifold.

In either case, we can look at the spectrum of the Laplacian $\Delta := -y^2\left(\frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial x^2}\right)$ on $L^2(\Gamma\backslash\mathbb{H})$. Here is a fundamental result we will prove in this class.

Theorem 1.6. *We have the following spectral decomposition of L^2.*

(1) When Γ is cocompact, i.e. $Y(\Gamma) = \Gamma\backslash\mathbb{H}$ is compact, then the Laplacian decomposes discretely. That is, there is a Hilbert space direct sum

$$L^2(Y(\Gamma)) = \bigoplus_{\lambda\in\sigma(\Delta)} H_\lambda$$

into eigenspaces, i.e. each $f \in H_\lambda$ satisfies $\Delta f = \lambda f$. Each eigenspace H_λ is finite dimensional, and the spectrum $\sigma(\Delta)$ has no accumulation points in $\mathbb{R}_{\geq 0}$.

(2) If Γ is finite covolume but not cocompact, then once again there is a decomposition

$$L^2(Y(\Gamma)) = \left(\bigoplus_{\lambda\in\sigma_{\mathrm{disc}}(\Delta)} H_\lambda\right) \oplus \bigoplus_{c\in c(\Gamma)} \mathcal{E}_c$$

where once again each H_λ is a finite dimensional space consisting of all L^2 eigenfunctions of Δ of eigenvalue λ. Every L^2 Laplace eigenfunction appears in $\bigoplus_{\lambda\in\sigma_{\mathrm{disc}}(\Delta)} H_\lambda$. $\sigma_{\mathrm{disc}}(\Delta)$ has no accumulation points in $\mathbb{R}_{\geq 0}$.

The set $c(\Gamma)$ is the set of cusps of Γ. Each $\mathcal{E}_c$ is infinite dimensional, and Δ decomposes continuously on each $\mathcal{E}_c$.

Here is a simple observation. If we are in case (1) of the above situations, i.e. Γ is cocompact, then there are actual Laplace eigenforms, i.e. $\sigma(\Delta) \neq 0$ (since otherwise the whole L^2 space would vanish). On the other hand, the groups $\Gamma_1(N)$ appearing in Conjecture 1.4 are all non-cocompact cofinite volume. But now it is not even clear that $\sigma_{\mathrm{disc}}(\Delta)$ is non-zero–it could be the case that the "continuous spectrum" is everything.[1]

Our first big use of the trace formula is the following fact, first proved by Selberg. In fact, it was the reason he developed the trace formula in the first place! This fact just says that our concerns in the previous paragraph does not come to pass for $\Gamma_1(N)$.

Theorem 1.7 (Selberg)**.** *If $\Gamma = \Gamma_1(N)$, or more generally is a congruence subgroup of $\Gamma(1) := \mathrm{PSL}_2(\mathbb{Z})$, then there exist Maass cusp forms.*

This may not seem that impressive. But it really is extremely striking, especially in light of the next conjecture. To state it, let me take a moment to make sense of what it would mean to pick a discrete cofinite volume group Γ with a certain number of cusps.

Fix two non-negative integers g, n and suppose that $2g + n > 2$, and that $n \geq 1$. Consider the moduli space $\mathscr{M}_{g,n}$ of genus g Riemann surfaces with n punctures; a point $x \in \mathscr{M}_{g,n}$ corresponds to a hyperbolic Riemann surface X with n cusps, and so to a discrete group Γ such that $X = X(\Gamma)$. Consider a measure μ on $\mathscr{M}_{g,n}$ coming from its structure as a smooth manifold.

[1]For experts: this is not exactly true. When there are cusps, there is always residual spectrum which appears discretely in L^2. What we are really talking about is the *cuspidal* spectrum.

Conjecture 1.8. *The set of all $X = X(\Gamma)$ which have Maass cusp forms is measure* 0.

There is some very nice work of Phillips and Sarnak which provides evidence for this. For a deeper discussion of the type of results needed to understand this work, go to Zelditch's class!

1.3. **Where do cusp forms come from?** In light of this conjecture, it may seem a bit surprising that congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$ admit Maass cusp forms at all–most groups which have cusps do not. Of course, in some sense, if there are no cusps to Γ, i.e. $Y(\Gamma) = \Gamma\backslash\mathbb{H}$ is a compact Riemann surface, then the Laplacian decomposes discretely, hence there are cusp forms for trivial reasons!

Here is an interesting observation, though: although no finite index subgroup of $\mathrm{SL}_2(\mathbb{Z})$ can be cocompact, some cocompact $\Gamma \subset \mathrm{PSL}_2(\mathbb{R})$ still come from arithmetic.

This works as follows: let $\mathrm{B}/\mathbb{Q}$ be a quaternion algebra. This means B is a simple (non-commutative) algebra over $\mathbb{Q}$ whose center is $\mathbb{Q}$. A consequence of the cohomological approach to class field theory is a classification of all such B. The classification is as follows: first, one notes that over $\mathbb{Q}_v$, there are only 2 quaternion algebras, the split algebra $\mathrm{Mat}_{2\times 2}$ or the unique quaternion division algebra. Then, one notes that given a quaternion algebra $\mathrm{B}/\mathbb{Q}$, we get a collection $\mathrm{B}_{\mathbb{Q}_v} := \mathrm{B} \otimes_{\mathbb{Q}} \mathbb{Q}_v$ of quaternion algebras over the local field $\mathbb{Q}_v$, and that

(1) B is determined by the set of places
$$\mathrm{Ram}(\mathrm{B}) := \{v : \mathrm{B}_{\mathbb{Q}_v} \text{ is a division algebra}\}$$
(2) Ram(B) is finite.
(3) #Ram(B) is even.
(4) Given a set S of places of $\mathbb{Q}$ satisfying (2) and (3), there exists a (unique) quaternion algebra B such that
$$S = \mathrm{Ram}(\mathrm{B}).$$

So now let $\mathrm{B}/\mathbb{Q}$ be a quaternion algebra. Let's assume that $\mathrm{Ram}(\mathrm{B}) \neq \emptyset$ and $\infty \notin \mathrm{Ram}(\mathrm{B})$. B has a reduced norm, denoted $\mathrm{Nrd} : \mathrm{B}^\times \to \mathbb{Q}^\times$, and has a ring of integers (a maximal order) $\mathscr{O}_\mathrm{B}$. One should think of B, $\mathrm{B}^\times$, $(\mathrm{B})^{\mathrm{Nrd}=1}$, $\mathscr{O}_\mathrm{B}$, and $(\mathscr{O}_\mathrm{B})^{\mathrm{Nrd}=1}$ as twisted versions of $\mathrm{Mat}_{2\times 2}(\mathbb{Q})$, $\mathrm{GL}_2(\mathbb{Q})$, $\mathrm{SL}_2(\mathbb{Q})$, $\mathrm{Mat}_{2\times 2}(\mathbb{Z})$, and $\mathrm{SL}_2(\mathbb{Z})$ respectively.

So let $\Gamma \subset (\mathscr{O}_\mathrm{B})^{\mathrm{Nrd}=1}$ be a congruence subgroup. Since $\Gamma \subset \mathrm{B}^{\mathrm{Nrd}=1} \subset (\mathrm{B} \otimes_{\mathbb{Q}} \mathbb{R})^{\mathrm{Nrd}=1} = \mathrm{SL}_2(\mathbb{R})$ we can think of Γ as acting on $\mathbb{H}$. It happens, essentially since for $p \in S$ we have that $(\mathrm{B} \otimes_{\mathbb{Q}} \mathbb{Q}_p)^{\mathrm{Nrd}=1}$ is compact, that $Y(\Gamma) = \Gamma\backslash\mathbb{H}$ is also compact. Thus, we can still construct some cocompact groups Γ coming from arithmetic, and these very easily have cusp forms.

Amazingly, this helps us find cusp forms on arithmetic groups like $\Gamma_1(N)$. This is the Jacquet-Langlands correspondence.

Theorem 1.9 (Jacquet-Langlands)**.** *Let Γ be as above, i.e. a cocompact discrete subgroup of $\mathrm{SL}_2(\mathbb{R})$ appearing as a congruence subgroup of a quaternion algebra. Suppose that $f \in L^2(Y(\Gamma))$ is a Laplace eigenfunction, with eigenvalue λ. Then there exists a congruence subgroup Γ' of $\mathrm{SL}_2(\mathbb{Z})$ and cusp form f' for Γ' of eigenvalue λ.*

This is yet another striking application of the trace formula!

2. Some degenerate versions of the trace formula

The Selberg trace formula is a generalization of the following simple identity: given a linear map $T \in \mathrm{End}(V)$ from a finite dimensional vector space V to itself, and given a basis $v_1, .., v_n$ and corresponding matrix $(a_{i,j})$ for T, we have

$$\sum_{i=1}^{n} \lambda_i = \mathrm{Tr}(T) = \sum_{i=1}^{n} a_{i,i}.$$

We will call the LHS of the above equality "the spectral side" and the RHS "the geometric side." Of course, in the trace formula, we only see certain spaces V and certain operators T.

Today, we will try to give a flavor for what we are looking at in two toy examples. The first captures the non-Abelian nature of the trace formula, but has no analytic difficulties. The second shows some of the analytic inputs, but works in an Abelian setting. Both should be somewhat familiar.

2.1. **Example 1: trace formula for finite groups.** Let H be a finite group, and let $\Gamma \subset H$ be a subgroup. Call $Y = \Gamma\backslash H$, and let $V = L^2(Y)$ be the space of complex valued functions ϕ on Y; we can also think of this as the space of ϕ on H such that $\phi(\gamma h) = \phi(h)$ for all $\gamma \in \Gamma$.

V has an obvious basis $\phi_{\Gamma h} := \mathbb{1}_{\Gamma h}$, with $h \in \Gamma\backslash H$. There is a left action of H on Y by right multiplication, which gives rise to a representation $R : H \to \mathrm{GL}(V)$ via

$$R(h) : \phi \mapsto \{x \mapsto \phi(xh)\}.$$

Of course, we can decompose R into irreducible representations of H: this gives an abstract decomposition

$$V = \bigoplus_{\pi \in \mathrm{Irr}(H)} m_\Gamma(\pi)\pi$$

where $m_\Gamma(pi)$ simply counts the multiplicity by which π, a representation of H, appears in R.

Fix a function $f : H \to \mathbb{C}$. We are interested in the operator $T = R(f)$, where we define

$$R(f)\phi(x) = \sum_{y \in H} \phi(xy)f(y).$$

Note that for any π a representation of H, we can similarly define

$$\pi(f)v = \sum_{y \in H} \pi(y).vf(y).$$

Let's compute $\mathrm{Tr}(T) = \mathrm{Tr}(R(f))$. On one hand, due to the spectral decomposition, we have

$$\mathrm{Tr}(R(f)) = \sum_{\pi \in \mathrm{Irr}(H)} m_\Gamma(\pi)\,\mathrm{Tr}(\pi(f)).$$

We can also manipulate

$$\begin{aligned} R(f)\phi(x) &= \sum_{y \in H} \phi(xy)f(y) \\ &= \sum_{y \in H} \phi(y)f(x^{-1}y) \\ &= \sum_{y \in \Gamma\backslash H} \left(\sum_{\gamma \in \Gamma} f(x^{-1}\gamma y) \right) \phi(y) \end{aligned}$$

so if we define

$$K_f(x, y) := \sum_{\gamma \in \Gamma} f(x^{-1}\gamma y)$$

then

$$R(f)\phi(x) = \sum_{y \in \Gamma\backslash H} K_f(x, y)\phi(y).$$

Note that this manipulation is taking $T = R(f)$ and writing its matrix for the basis $\phi_{\Gamma h}$ described above. Thus, taking the trace of $R(f)$ by summing down the diagonal gives

$$\mathrm{Tr}(R(f)) = \sum_{x \in \Gamma\backslash H} K_f(x,x) = \sum_{x \in \Gamma\backslash H} \sum_{\gamma \in \Gamma} f(x^{-1}\gamma x) = \sum_{\gamma \in \frac{\Gamma}{\Gamma}} \frac{\#H_\gamma}{\#\Gamma_\gamma} \sum_{x \in H_\gamma \backslash H} f(x^{-1}\gamma x)$$

In the above, H_γ and Γ_γ are the centralizers of γ in H and Γ respectively, while the notation $\frac{\Gamma}{\Gamma}$ means the quotient of Γ by its own action by conjugation.

We have derived

Proposition 2.1 (The trace formula for finite groups)**.** *Let notation be as above. Then*

$$\sum_{\pi \in \mathrm{Irr}(H)} m_\Gamma(\pi) \,\mathrm{Tr}(\pi(f)) = \sum_{\gamma \in \frac{\Gamma}{\Gamma}} \frac{\#H_\gamma}{\#\Gamma_\gamma} \sum_{x \in H_\gamma \backslash H} f(x^{-1}\gamma x).$$

Here is an amusing corollary.

Corollary 2.2 (Frobenius reciprocity)**.** *Let $\sigma \in \mathrm{Irr}(H)$. Then*

$$m_\Gamma(\sigma) = \dim \sigma^\Gamma.$$

Proof. Fix $\sigma \in \mathrm{Irr}(H)$, and let $f(h) = \overline{\mathrm{Tr}(\sigma(h))}$ be the complex conjugate of the character of σ. Put f into the trace formula above, and note that

$$\pi(f)v = \sum_{h \in H} \overline{\mathrm{Tr}(\sigma(h))}\pi(h)v$$

so that

$$\mathrm{Tr}(\pi(f)) = \sum_{h \in H} \overline{\mathrm{Tr}(\sigma(h))}\,\mathrm{Tr}(\pi(h)) = \begin{cases} \#H & \text{if } \sigma \cong \pi \\ 0 & \text{else} \end{cases}$$

by orthogonality of characters. Thus, the LHS of the trace formula becomes simply $m_\Gamma(\sigma)\#H$. On the other hand, the right hand side gives

$$\sum_{\gamma \in \frac{\Gamma}{\Gamma}} \frac{\#H_\gamma}{\#\Gamma_\gamma} \sum_{x \in H_\gamma \backslash H} f(x^{-1}\gamma x) = \sum_{\gamma \in \frac{\Gamma}{\Gamma}} \frac{\#H_\gamma}{\#\Gamma_\gamma} \frac{\#H}{\#H_\gamma} \overline{\mathrm{Tr}(\sigma(\gamma))}$$

which is clearly

$$\#H \sum_{\gamma \in \frac{\Gamma}{\Gamma}} \frac{1}{\#\Gamma_\gamma} \overline{\mathrm{Tr}(\sigma(\gamma))}.$$

Since

$$\sum_{\gamma \in \frac{\Gamma}{\Gamma}} \frac{1}{\#\Gamma_\gamma} \overline{\mathrm{Tr}(\sigma(\gamma))} = \dim \sigma^\Gamma$$

via

$$\#\Gamma \dim \sigma^\Gamma = \sum_{\gamma \in \Gamma} \mathrm{Tr}(\sigma(\gamma)) = \sum_{\gamma \in \frac{\Gamma}{\Gamma}} \frac{\#\Gamma}{\#\Gamma_\gamma} \mathrm{Tr}(\sigma(\gamma))$$

we are done. □

2.2. **Example 2: harmonic analysis on $\Lambda\backslash\mathbb{R}^n$.** Let's try to run the same picture in a less discrete (but still cocompact) world. Let Λ be a lattice in $\mathbb{R}^n$, so that abstractly $\Lambda \cong \mathbb{Z}^n$ and its $\mathbb{R}$ span is the whole space. We fix $\langle \cdot, \cdot \rangle$ the standard inner product on $\mathbb{R}^n$; this allows us to speak about $\Lambda^\vee$, the dual lattice to Λ under the pairing.

Throughout we will use $e(x) := e^{2\pi i x}$. We are interested in eigenfunctions of the Laplacian $\Delta = -\mathrm{divgrad}$ (this is the negative of what many people call the Laplacian) on $Y(\Lambda) := \Lambda\backslash\mathbb{R}^n$. Note that for $l \in \Lambda^\vee$ the functions $e(\langle x, l \rangle)$ are all eigenfunctions for the Laplacian, of eigenvalue $4\pi^2|l|^2$, and it is easy to see that these are all of them.

Let $f \in \mathscr{S}(\mathbb{R}^n)$ be a $\mathbb{C}$-valued Schwartz function, i.e. a smooth function f such that for every polynomial differential operator D on $\mathbb{R}^n$, Df is bounded. One can compute its Fourier transform

$$\hat{f}(\xi) = \int_{\mathbb{R}^n} f(x)e(-\langle x, \xi\rangle)dx$$

and we are going to prove the following fact.

Proposition 2.3 (Trace formula for $\Lambda\backslash\mathbb{R}^n$, i.e. Poisson summation)**.** *Let $f \in \mathscr{S}(\mathbb{R}^n)$. Then*

$$\mathrm{vol}(Y(\Lambda)) \sum_{\lambda\in\Lambda} f(\lambda) = \sum_{l\in\Lambda^\vee} \hat{f}(l)$$

Proof. We define, like in the finite group case, an operator on $Y(\Lambda) = \Lambda\backslash\mathbb{R}^n$, given by, for $f \in \mathscr{S}(\mathbb{R}^n)$,

$$R(f)\phi(x) = \int_{\mathbb{R}^n} \phi(x+y)f(y)dy = \int_{\Lambda\backslash\mathbb{R}^n} \sum_{\lambda\in\Lambda} f(-x+\lambda+y)\phi(y)dy.$$

The sum

$$K_f(x,y) = \sum_{\lambda\in\Lambda} f(-x+\lambda+y)$$

certainly converges, for any x and y, since the function f is Schwartz. Even better, the kernel K_f is Hilbert Schmidt, i.e.

$$\int_{Y(\Lambda)^2} |K_f(x,y)|^2 dxdy < \infty.$$

This is obvious–the space $Y(\Lambda)^2$ is compact, and $K_f(x,y)$ is a smooth function on this compact manifold, hence obviously L^2. It is also trace class (a definition I will be a bit vague about at present) which is also not hard to see–it follows from the fact that every $f \in \mathscr{S}(\mathbb{R}^n)$ can be written as $f = f_1 * f_2$, with f_i Schwartz.

So we can compute its trace in two ways. First we can "sum down the diagonal," i.e. compute

$$\mathrm{Tr}(R(f)) = \int_{Y(\Lambda)} K_f(x,x)dx = \int_{Y(\Lambda)} \sum_{\lambda\in\Lambda} f(-x+\lambda+x)dx = \mathrm{vol}(Y(\Lambda)) \sum_{\lambda\in\Lambda} f(\lambda).$$

On the other hand, the operator $R(f)$ commutes with the Laplacian, so it shares the eigenbasis given by functions $e(\langle x, l\rangle)$. We can also compute the trace of $R(f)$ by summing eigenvalues, i.e. we should compute each eigenvalue by

$$R(f)e(\langle\cdot, l\rangle)(x) = \int_{\mathbb{R}^n} e(\langle x+y, l\rangle)f(y)dy = \left(\int_{\mathbb{R}^n} e(\langle y, l\rangle)f(y)dy\right) e(\langle x, l\rangle) = \hat{f}(-l)e(\langle x, l\rangle)$$

Summing the eigenvalues gives

$$\mathrm{Tr}(R(f)) = \sum_{l\in\Lambda^\vee} \hat{f}(l)$$

so we are done. □

3. Introduction to symmetric spaces

Last time we discussed two toy versions of the Selberg trace formula. It's time to begin discussing the true trace formula.

Before we dive into this, though, let me say a few generalities about the general "analytic" setting for the trace formula. At some point we will restrict to the upper half plane, but for now it costs almost nothing to be more general.

3.1. **Riemannian symmetric spaces.** Let S be a connected and complete Riemannian mannifold, with metric $ds^2 = g_{ij}dx^i dx^j$. [2]

Definition 3.1. S is a *Riemannian symmetric space* if, for every point $x \in S$, there exists an isometry of S which fixes x and induces multiplication by -1 on $T_x S$.

We will write $\tilde{\mathrm{G}} = \mathrm{Isom}(S)$ for the group of isometries of S, and $\mathrm{G} = (\tilde{\mathrm{G}})^\circ$ for the connected component of the identity. Note that all $g \in \mathrm{G}$ are orientation preserving.

Here are some quick observations.

(1) The group of isometries $\tilde{\mathrm{G}}$ of a symmetric space S acts transitively. This is because, if x_1, x_2 are in S, then by virtue of the connectedness and (geodesic) completeness of S, there is a geodesic segment running from x_1 to x_2. If x_{mid} is the midpoint of this segment, then there is a $g \in \tilde{\mathrm{G}}$ acting by -1 on the tangent space at x_{mid}, thus g must invert this geodesic, hence flips x_1 and x_2.
(2) In fact, G acts transitively on S. See the exercises.
(3) G is a connected Lie group, by an old theorem of Myers and Steenrod.
(4) If we fix a base point $x_0 \in S$, then $\mathrm{K} := \mathrm{Stab}_{\mathrm{G}}(x_0)$ is a compact subgroup of G. This is because the action of K on $T_{x_0}S$ preserved the metric, hence gives rise to an embedding $\mathrm{K} \hookrightarrow \mathrm{O}(\dim S)$ (action on the tangent space is faithful).
(5) In fact, one can show that the data of Riemannian symmetric space is equivalent to giving a quadruple $(\mathrm{G}, \mathrm{K}, \sigma, g_0)$, where G is a Lie group, σ is an automorphism of G such that $\sigma^2 = 1$, K is an open compact contained in G^σ, and g_0 is a K-invariant inner form on $\mathfrak{g}/\mathfrak{k}$.

Example 3.2. Here are some examples of symmetric spaces.

- $S = S^n$, the round n-sphere. The group of orientation preserving isometries of S^n is $\mathrm{G} = \mathrm{SO}(n+1)$, while we can take $\mathrm{K} = \mathrm{SO}(n)$
- $S = \mathbb{R}^n$, flat Euclidean space. Orientation preserving symmetries are $\mathrm{G} = \mathrm{SO}(n) \ltimes \mathbb{R}^n$, while the stabilizer of the origin is $\mathrm{K} = \mathrm{SO}(n)$.
- $S = \mathbb{H}^n$, the constant (sectional) curvature -1 hyperbolic space. $\mathrm{G} = \mathrm{SO}(1,n)^\circ$, and $\mathrm{K} = \mathrm{SO}(n)$. Since G acts transitively on one sheet of the hyperboloid $y^2 - (x_1^2 + ... + x_n^2) = 1$, and the stabilizer of $(1, 0, ..., 0)$ is K, this gives rise to the hyperboloid model of hyperbolic space. We can pass to more familiar models (the Poincare ball/upper half space, or the Klein ball) of hyperbolic space by various projections.

These are all of the constant sectional curvature symmetric spaces. There are more exotic examples, but for now we are really only concerned with one example on the list above: $\mathbb{H} = \mathbb{H}^2$, the hyperbolic plane. Once should note that when $n = 2$, $\mathrm{SO}(1,2) \cong \mathrm{PGL}_2(\mathbb{R})$, since $\mathrm{PGL}_2(\mathbb{R})$ acts faithfully on $\mathfrak{sl}_2(\mathbb{R})$ by the adjoint representation, and preserved the quadratic form given by det; thus, the identification $\mathbb{H} = \mathrm{SO}(1,2)^\circ / \mathrm{SO}(2)$ can be thought of as $\mathbb{H} = \mathrm{PSL}_2(\mathbb{R}) / \mathrm{SO}(2)$.

3.1.1. *Invariant differential operators on S.* Suppose S is a Riemannian symmetric space, i.e. $S = \mathrm{G} / \mathrm{K}$ for data $(\mathrm{G}, \mathrm{K}, \sigma, g_0)$ Given $g \in \mathrm{G}$, define the translation operator T_g acting on functions on S by

$$T_g \phi(x) = \phi(gx).$$

Definition 3.3. A (linear) differential operator D on S is said to be *invariant* if

$$DT_g = T_g D$$

. We write $\mathscr{D}(S)$ for the ring of invariant differential operators on S.

[2] Note that we are using, and will use, Einstein summation notation when convenient.

Example 3.4. Every translation invariant differential operator D has constant coefficients. If D is also invariant under $\mathrm{SO}(n)$, hence $D \in \mathscr{D}(\mathbb{R}^n)$, then D must be a polynomial in $\Delta = -\sum_i \frac{\partial^2}{\partial x_i^2}$ (check this yourself!)

Example 3.5 (Laplace-Beltrami)**.** On any Riemannian manifold, there is always an isometry invariant differential operator, denoted by Δ. It is given by

$$\Delta = -\mathrm{divgrad}$$

where

$$\mathrm{grad} f = g^{ij} \frac{\partial f}{\partial x_i} \frac{\partial}{\partial x_j}$$

and, for a vector field $X = X_i \frac{\partial}{\partial x^i}$, we have

$$\mathrm{div} X = \frac{1}{\sqrt{|\det g|}} \frac{\partial}{\partial x_i} (\sqrt{|\det g|} X_i)$$

It is true that Δ commutes with all isometries–in fact, a stronger statement is true.

Proposition 3.6. *Let X be a Riemannian manifold, and let $F : X \to X$ be a diffeomorphism. Then Δ commutes with F if and only if F is an isometry.*

See the exercises for the proof.

Often $\mathscr{D}(S)$ has more elements than just Δ, but sometimes not. For instance, in our examples $S^n, \mathbb{R}^n$, and $\mathbb{H}^n$, it will turn out that $\mathscr{D}(S)$ is exactly the ring of polynomials in Δ (the first and third are examples of symmetric spaces of rank 1, where this is always the case).

Proposition 3.7. *Suppose $S = \mathrm{G}/\mathrm{K}$ is a symmetric space, with S corresponding to $(\mathrm{G}, \mathrm{K}, \sigma, g_0)$. Then $\mathscr{D}(S)$ is finitely generated and commutative.*

This is very important to the problem of harmonic analysis. First of all, the fact that $\mathscr{D}(S)$ is commutative means that we can talk about simultaneous eigenfunctions for all differential operators. The fact that $\mathscr{D}(S)$ is finitely generated means that there is some hope that there exist simultaneous eigenfunctions for all differential operators.

Next time, we will prove this theorem.

In fact, there is a better and more refined version of this statement, relating $\mathscr{D}(S)$ and a particular sub-algebra of $\mathfrak{Z}(\mathfrak{g}_{\mathbb{C}})$, the center of the universal enveloping algebra. See Helgason's book.

4. Harmonic analysis on symmetric spaces

We would like to understand invariant differential operators on Riemannian symmetric spaces and their eigenfunctions. These will play the same role that $e_\xi(x) := e(\langle x, \xi\rangle)$ plays on $\mathbb{R}^n$.

Of fundamental importance to even the notion of eigenfunction is the fact that for S a symmetric space, $\mathscr{D}(S)$ is finitely generated and commutative, i.e Proposition 3.7.

4.1. Point-pair invariants.

Let $S = \mathrm{G} / \mathrm{K}$ be a symmetric space.

Definition 4.1. We say a smooth function $k : S \times S \to \mathbb{C}$ is a *point pair invariant* if

$$k(gx, gy) = k(x, y)$$

for all $g \in \mathrm{G}$ and if $k(x_0, y)$ is compactly supported, as a function of $y \in S$, for all fixed $x_0 \in S$. (If there is a notion of rapidly decaying along with derivatives, i.e. a notion of Schwartz function, this is also good enough.)

Example 4.2. Suppose that $h \in C_c^{\infty,+}(\mathbb{R})$ is an even compactly supported smooth function. Then if we set $k(x, y) = h(\mathrm{dist}(x, y))$, then k is a point pair invariant.

Example 4.3. On $\mathbb{R}^n$, all translation invariant functions $k(x, y)$ are of the form $k(x, y) = f(x - y)$ for some compactly supported smooth function f on $\mathbb{R}^n$. But, for $k(x, y)$ to preserve rotations, f must be invariant under rotation as well. Thus, every point pair invariant on $\mathbb{R}^n$ is of the form above.

The point pair invariants form an algebra $\mathscr{A}(S)$, with convolution given by

$$(k_1 * k_2)(x, y) = \int_S k_1(x, z) k_2(z, y) dz.$$

There is a representation of $\mathscr{A}(S)$ on $C^\infty(S)$, given by

$$k.\phi(x) = \int_S \phi(y) k(x, y) dy$$

One should note that all point pair invariants are symmetric, i.e.

$$k(x, y) = k(y, x).$$

This is because, for fixed x, y, there is an element of G which swaps them.

Lemma 4.4. *The ring $\mathscr{A}(S)$ is commutative.*

Proof. We simply manipulate, using symmetry of point pair invariants

$$\begin{aligned}(k_1 * k_2)(x, y) &= \int_S \int_S k_1(x, z) k_2(z, y) dz \\ &= \int_S \int_S k_2(y, z) k_1(z, x) dz \\ &= (k_2 * k_1)(y, x) = (k_2 * k_1)(x, y)\end{aligned}$$

□

Once we have this, it is easy to see half of Proposition 3.7: that $\mathscr{D}(S)$ is commutative.

Proof. What we will show is that every $D \in \mathscr{D}(S)$ will commute with every element of $\mathscr{A}(S)$ (where we view both inside endomorphisms of smooth functions) and then with each other. To do this, we need a particular family of point pair invariants.

Let h_δ on $\mathbb{R}$ be an approximation to the identity, i.e. a collection of even smooth functions on $\mathbb{R}$, with support in $[-\delta, \delta]$, and satisfying

$$\int_{\mathbb{R}} f(y) h_\delta(x - y) dy \to f(x)$$

for all $x \in \mathbb{R}$. Then if we set $k_\delta(x, y) = h_\delta(\mathrm{dist}(x, y))$, we have k_δ is obviously an approximation to the identity in on S.

Note that for any point pair invariant k, we can apply $D \in \mathscr{D}(S)$ to k in the first variable, which we write simply $(Dk)(x,y)$. This is a point pair invariant itself, since

$$(T_g Dk)(x,gy) = (DT_g k)(x,gy) = (Dk)(x,y)$$

Now we know

$$(k * (Dk_\delta))(x,y) = ((Dk) * k)(x,y)$$

so as $\delta \to 0$, we find that $kD = Dk$.

Even better, we know too that

$$D_1 k_{\delta_1} * D_2 k_{\delta_2} = D_2 k_{\delta_2} * D_1 k_{\delta_1}$$

and letting $\delta_1, \delta_2 \to 0$ shows that $\mathscr{D}(S)$ is commutative. □

What about the other half of Proposition 3.7: that $\mathscr{D}(S)$ is finitely generated?

Proof. (Sketch) The basic idea is not so complicated. Some details are, and will be omitted or reserved for the exercises.

The ring $\mathscr{D}(S)$ is most naturally thought of as a (commutative) filtered algebra, where filtration is by the total degree of the differential operator. We will construct a ring Symb which one can think of the as the ring of symbols (near our base point x_0) of these differential operators. This is naturally a graded algebra.

In fact, we can identify the ring $\text{Symb} = (\text{Sym}(\mathfrak{m}))^{\text{Ad(K)}}$, where we have used our usual notation and written $S = \text{G} / \text{K}$, thinking of K as the stabilizer of our base point x_0, and where we have written $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{m}$, with $\mathfrak{k} = \text{Lie}(K)$ is the $+1$ eigenspace for σ, and $\mathfrak{m}$ is the -1 eigenspace. Symb is finitely generated as it is the fixed points in a free polynomial algebra of a compact group.

We will also construct a linear map $L : \text{Symb} \to \mathscr{D}(S)$, which one can think of as symmetrization. The map L descends, on associated graded rings, to an isomorphism

$$\text{gr}(L) : \text{Symb} = \text{gr}(\text{Symb}) \to \text{gr}(\mathscr{D}(S)).$$

Thus, $\mathscr{D}(S)$ is commutative filtered algebra whose associated graded is finitely generated. This implies that $\mathscr{D}(S)$ is already finitely generated. □

4.2. **Eigenfunctions.** So now we have two commuting commutative algebras of endomorpisms $\mathscr{D}(S)$ and $\mathscr{A}(S)$. We can try to diagonalize the action on smooth functions simultaneously.

Definition 4.5. An *eigenparameter* λ for $\mathscr{D}(S)$ is a point $\lambda \in \text{Spec}(\mathscr{D}(S))(\mathbb{C})$, or equivalently a homomorphism $\lambda : \mathscr{D}(S) \to \mathbb{C}$, such that there exists a function $f \in C^\infty(S)$ such that

$$Df = \lambda(D) f$$

for all $D \in \mathscr{D}$. We call such an f an *eigenfunction* with eigenparameter λ.

Example 4.6. For $S = \mathbb{R}^n$, we know $\mathscr{D}(S) = \mathbb{C}[\Delta]$, so an eigenparameter simply corresponds to an eigenvalue for Δ. It is easy to see all eigenfunctions for Δ are of the form

$$e_\xi(x) := e(\langle x, \xi \rangle) = e^{2\pi i \langle x, \xi \rangle)}$$

for $\xi = (\xi_1, ..., \xi_n) \in \mathbb{C}^n$ and that the eigenvalue for e_ξ is simply $4\pi^2 \sum \xi_i^2 \in \mathbb{C}$. Here is an interesting point: not all of these eigenfunctions appear in the Fourier inversion formula, only those for $\xi \in \mathbb{R}^n$! This will be a common feature.

Remark 4.7. Although we described eigenfunctions as a priori smooth functions, something much more general can be done–indeed, we can talk about weak eigenfunctions, which are eigenfunctions in the distributional sense. Since $\Delta \in \mathscr{D}(S)$ is always elliptic, then any weak solution to $\Delta f = \lambda(\Delta) f$ must be smooth by elliptic regularity. Moreover, if S is analytic, then any such f must be analytic as well.

Definition 4.8. An eigenfunction f on S is a *zonal spherical function about* x_0 if, for all $k \in \text{K} = \text{Stab}_\text{G}(x_0)$, we have

$$f(kx) = f(x).$$

We often write w_λ for a zonal spherical function. We will also call these *spherical harmonics*.

Example 4.9. In S^n, the zonal spherical functions are the same spherical harmonics you see from classical representation theory. (More intuitively, these are the waves you see when you bounce a basketball.) In $\mathbb{R}^n$, you can see what happens:

$$e_\xi(kx) = e(\langle kx, \xi\rangle) = e(\langle x, k^{-1}\xi\rangle) = e_{k^{-1}\xi}(x)$$

and you can get invariant spherical harmonics by integrating:

$$w_\lambda := \int_{\mathrm{SO}(n)} e_{k\xi} dk.$$

We will go through some of these explicit formulas next time.

Our next goal is the following proposition.

Proposition 4.10. *For each λ, the space of zonal spherical functions is at most 1-dimensional. If it is non-vanishing, there is a unique spherical harmonic w_λ with $w_\lambda(x_0) = 1$.*

Proof. We know that S is always a (real) analytic manifold, and by elliptic regularity that every eigenfunction is real analytic. Thus, given a spherical harmonic w it is determined by its Taylor expansion around x_0. So fix normal coordinates $x_1, ..., x_n$ around the point. We will compute Taylor coefficients, i.e. for $\partial_i = \frac{\partial}{\partial x_i}$,

$$(\partial_1^{a_1}\partial_2^{a_2}...\partial_n^{a_n})\, f(x_0)$$

Let's call $D_{\underline{a}} = \partial_1^{a_1}\partial_2^{a_2}...\partial_n^{a_n}$.

Now note that given *any* differential operator D on S, we can define its invariant $\mathrm{inv}(D) \in \mathrm{Symb} = \mathrm{Sym}(\mathfrak{m})^{\mathrm{Ad(K)}} \cong \mathscr{D}(S)$ by

$$Df(x)|_{x=x_0} = \mathrm{inv}(D) f(x; x_0)|_{x=x_0}$$

where

$$f(x; x_0) = \int_{\mathrm{K}} f(kx) dk$$

is the average of f about x_0. (Here we normalize Haar measure on K so that $\mathrm{vol(K)} = 1$.) One finds that for a spherical harmonic w

$$D_{\underline{a}} w(x)|_{x=x_0} = \mathrm{inv}(D_{\underline{a}}) w(x, x_0)|_{x=x_0} = \mathrm{inv}(D_{\underline{a}}) w(x)_{x=x_0} = \lambda(\mathrm{inv}(D_{\underline{a}}))w(x)|_{x=x_0}$$

Thus, if $w(x_0) = 0$, all of its Taylor coefficients at x_0 vanish as well; while if it is not zero, then they are determined by λ. □

Next time we will explicitly describe some classical examples of spherical harmonics.

5. Explicit spherical harmonics on symmetric spaces

First, let's finish talking about the general theory. Then, some examples.

We had just defined zonal spherical functions (spherical harmonics) as smooth K-invariant eigenfunctions, and shown that for each eigenparameter $\lambda : \mathscr{D}(S) \to \mathbb{C}$, there is at most one zonal spherical function w_λ with eigenvalue λ. Moreover, the space of zonal spherical functions with a fixed eigenparameter vanishes if and only any one zonal spherical function w_λ satisfies $w_\lambda(x_0) = 0$; we can (and will) thus always normalize such spherical harmonics so that $w_\lambda(x_0) = 1$.

How do we construct such spherical harmonics, though? Given an eigenfunction f with eigenparameter λ, we can always average over K to get a K-invariant function

$$f(x; x_0) = \int_{\mathrm{K}} f(kx)dx$$

and it is clear that we can differentiate under the integral (why?!) to see that $Df(x; x_0) = \lambda(D)f(x; x_0)$ as well. Since the space of spherical harmonics with eigenparameter λ is at most one dimensional, we get

$$f(x; x_0) = f(x_0)w_\lambda(x). \tag{5.1}$$

This identity can be thought of as the mean value property of eigenfunctions. It provides, as obvious consequence, that all spherical harmonics w_λ can be constructed via this averaging practice, as long as the eigenfunction f we start with is non-vanishing at x_0.

5.1. The Selberg transform.

We can use the one-dimensionality of the space of spherical harmonics of eiegenparameter λ to define the following.

Definition 5.1. Let $k(x, y)$ be a point pair invariant, and let, for each eigenparameter λ, w_λ be the unique (if it exists) zonal spherical function, normalized so that $w_\lambda(x_0) = 1$. We call the function $\hat{k}(\lambda)$, defined by

$$\hat{k}(\lambda) = \int_S k(x_0, x)w_\lambda(x)dx$$

the *spectral transform* or the *Selberg transform* of k.

Remark 5.2. It should be clear that the Selberg transform is independent of choice of x_0. Namely, if x_1 is a different choice of basepoint with $gx_0 = x_1$, then $\mathrm{K}_{x_1} = g\,\mathrm{K}\,g^{-1}$ is the stabilizer of x_1; moreover, if w_λ is a zonal spherical function centered at x_0, then $T_{g^{-1}}w_\lambda = w_\lambda(g^{-1}\cdot)$ is a zonal spherical function centered at x_1, and all zonal spherical functions centered at x_1 arise in this way. Thus,

$$\begin{aligned}\hat{k}(\lambda) &= \int_S k(x_0, x)w_\lambda(x)dx \\ &= \int_S k(x_1, x)w_\lambda(g^{-1}x)dx\end{aligned}$$

shows the basepoint independence of the Selberg transform.

Proposition 5.3. *Let f be an eigenfunction for $\mathscr{D}(S)$ with eigenparameter λ, and let k be a point pair invariant. Then*

$$k.f(x) = \hat{k}(\lambda)f(x).$$

Proof. This is easy. First note that we need only check the identity

$$k.f(x_0) = \hat{k}(\lambda)f(x_0)$$

since for any fixed x and $g \in \mathrm{G}$ such that $x = gx_0$, $k.f(x) = (k.T_g f)(x_0)$ and $f(x) = T_g f(x_0)$.

But to see this, we need only compute using (5.1) that

$$k.f(x_0) = k.f(x_0; x_0) = f(x_0)k.w_\lambda(x_0) = \hat{k}(\lambda)f(x_0).$$

□

It's time to be a little more concrete and get into some formulas. Hopefully this helps to give some sense for what the Selberg transform is really doing. To simplify the formulas, we will stick only to 2-dimensional rank 1 examples for S; of course, most of the formulas we write down are extremely classical.

5.2. 2 dimensional examples.

5.2.1. *Example 1: $S = \mathbb{R}^2$.* Let's start small. Recall that for the plane, we have $\mathrm{G} = \mathrm{SO}(2) \ltimes \mathbb{R}^2$ and $\mathrm{K} = \mathrm{SO}(2)$, and that $\mathscr{D}(S) = \mathbb{C}[\Delta]$. We use polar coordinates around $x_0 = 0$, in which

$$\Delta = -\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r}\frac{\partial}{\partial r} + \frac{1}{r^2}\frac{\partial^2}{\partial \theta^2}\right).$$

Let's try to write down all the spherical harmonics on $\mathbb{R}^2$ in these polar coordinates. We know eigenfunctions are of the form $e_\xi(r,\theta) = e(\xi_1 r\cos\theta + \xi_2 r\sin\theta)$ where $\xi_i \in \mathbb{C}$ and so using the averaging trick (5.1) we find

$$w_\lambda(r,\theta) = \frac{1}{2\pi}\int_0^{2\pi} e(\xi_1 r\cos\theta + \xi_2 r\sin\theta)d\theta$$

It is worth noting that if $\xi_1 = 1$ and $\xi_2 = 0$, then this gives

$$\Re(w_\lambda(r,\theta)) = J_0(2\pi r)$$

where

$$J_0(r) = \frac{1}{\pi}\int_0^{\pi} \cos(r\cos\theta)d\theta = \frac{1}{2\pi}\int_0^{2\pi} \cos(r\cos\theta)d\theta$$

is the classical Bessel function.

We got to this formula for w_λ by starting with an eigenfunction and making it radially symmetric. On the other hand, we can try to look inside of radially symmetric functions and see which are Laplace eigenfunctions: this entails solving the second order ODE, for $\psi \in C^{\infty,+}(\mathbb{R})$,

$$r\frac{d^2\psi}{dr^2} + \frac{d\psi}{dr} + \lambda r\psi = 0.$$

[Figure out what the solutions to this linear ODE are.]

OK, why do we care about these computations besides the fact that they explain the appearance of many nice special functions? Well, consider the Fourier transform of a radially symmetric function $k(r) = k(r,\theta) \in \mathscr{S}(\mathbb{R}^2)$. One can think of such a radially symmetric function as what you get when you evaluate one variable of a point pair invariant to $x_0 = 0$. Thus, after one computes

$$\int_{\mathbb{R}^2} k(x)e_\xi(x)dx = 2\pi\int_0^\infty k(r)J_0(2\pi|\xi|r)dr := \tilde{k}(|\xi|).$$

This latter integral is called the *Hankel transform* of a function–of course, it is nothing more than the Selberg transform.

Since I can't resist, let me quickly remark that in this notation, the Poisson summation formula becomes

$$\sum_{l=0}^{\infty} r(l)k(l) = \sum_{l=0}^{\infty} r(l)\tilde{k}(l)$$

where $r(l)$ is the number ways to write l as the sum of two squares, i.e.

$$r(l) = \#\{(m,n) \in \mathbb{Z}^2 : m^2 + n^2 = l\}.$$

This can be used to get some control on the Gauss circle problem, i.e. one can use this to get the bound

$$\sum_{l\leq x} r(l) = \pi x^2 + O(x^{2/3})$$

We really expect the error to be $O_\varepsilon(x^{1/2+\varepsilon})$.

5.2.2. *Example 2: $S = S^2$.* For the sphere, we have $\mathrm{G} = \mathrm{SO}(3)$ and $\mathrm{K} = \mathrm{SO}(2)$. We will take x_0 to be the north pole. Once again, $\mathscr{D}(S) = \mathbb{C}[\Delta]$, so we have only to worry about the Laplacian.

We will need to work in coordinates. Write (ϕ,θ) for geodesic polar coordinates based at x_0–i.e. for usual spherical coordinates as we teach to Calculus students. One computes that in these coordinates

$$\Delta_{S^2} = -\left(\frac{\partial^2}{\partial\phi^2} + \frac{\cos\phi}{\sin\phi}\frac{\partial}{\partial\phi} + \frac{1}{\sin^2\phi}\frac{\partial^2}{\partial\theta^2}\right).$$

Now, we want to write down the zonal spherical functions in this setting. We can play the same game as we did in $\mathbb{R}^2$, and simply average eigenfunctions over K. In order to do this, we need to know (at least some!) eigenfunctions on S^2.

We can write a few eigenfunctions down by inspection, such as $f(\phi,\theta)=\cos\phi$, but here is a more general way. Using spherical coordinates on $\mathbb{R}^3$, we can write

$$\Delta_{\mathbb{R}^3}=-\frac{\partial^2}{\partial\rho^2}-\frac{2}{\rho}\frac{\partial}{\partial\rho}+\frac{1}{\rho^2}\Delta_{S^2}$$

which quickly shows that if $f\in C^\infty(\mathbb{R}^3)$ is actually a *harmonic* homogeneous polynomial of degree n, then $f|_{S^2}$ is an eigenfunction of Δ_{S^2} of eigenvalue $n(n+1)$.

Call H_n the space of restrictions of homogeneous degree n harmonic polynomials to S^2. Consider the algebra $H=\bigoplus_{n=0}^\infty H_n$ of functions on S generated by all of the H_n. One can easily show that H separates points on S^2, hence by the Stone-Weierstrass theorem is dense in $C(S^2)$, and so also in $L^2(S^2)$. In fact we can do better: it is easy to see that, since integration by parts gives

$$\int_S \Delta f g dx=\int_S f\Delta g dx$$

that the eigenspaces H_n for different eigenvalues are all orthogonal in L^2, and since H is dense, we can get quickly the decomposition

$$L^2(S^2)=\bigoplus_{n=0}^\infty H_n$$

To get zonal spherical functions, one needs simply by (5.1) to average a given $f_n\in H_n$ with $f_n(x_0)=1$ to find

$$w_{n(n+1)}(\phi,\theta)=w_{n(n+1)}(\phi)=\frac{1}{2\pi}\int_0^{2\pi} f_n(\phi,\theta)d\theta.$$

Example 5.4. Here are some examples:

- For $n=0$, we have $H_0=\mathbb{C}$.
- For $n=1$, the space H_1 is spanned by x,y,z. The restriction of z to S^2 is obviously zonal so we have $w_1=z=\cos\phi$.
- For $n=2$, the space H_2 is the restriction to S^2 of the five dimensional space spanned by xy,xz,yz,z^2-x^2,z^2-y^2. An eigenfunction f which is non-vanishing at x_0 is z^2-x^2; the averaging procedure gives $w_2=\frac{1}{2}(2z^2-(x^2+y^2))=\frac{1}{2}(3\cos^2\phi-1)$.
- For $n=3$, the space H_3 is 7 dimensional, and spanned by $xyz,x(y^2-z^2),y(x^2-z^2),z(x^2-y^2),x^3-3xy^2,y^3-3yx^2,z^3-3zx^2$. An eigenfunction f which is non-vanishing at x_0 is z^3-3zx^2; the averaging procedure gives $w_3=\frac{1}{2}(2z^3-3z(x^2+y^2))=\frac{1}{2}(5\cos^3\phi-3\cos\phi)$.

In general, solving $\Delta_{S^2}\psi=n(n+1)\psi$ for functions ψ which depend only on ϕ gives the equation

$$\sin\phi\frac{d^2\psi}{d\phi^2}+\cos\phi\frac{d\psi}{d\phi}+n(n+1)\sin\phi\psi=0,$$

which, if $\psi(\phi)=p(\cos\phi)$, becomes

$$\frac{d}{du}((1-u^2)\frac{dp}{du})+n(n+1)p=0.$$

Solutions (with value 1 at 1) to these are given by the Legendre polynomials P_n, and so we have the general formula

$$w_{n(n+1)}(\phi)=P_n(\cos\phi)$$

for the zonal spherical function.

Of course, there is also a nice representation theoretic way to come at this, using H_n to classify the irreducible representations of SO(3). This is reserved for the exercises.

Finally, the Selberg transform simply takes as input a function $k(\phi)$, symmetric about the north pole, and spits out

$$\hat{k}(n(n+1))=\frac{1}{\pi}\int_0^\pi k(\phi)P_n(\cos\phi)d\phi.$$

This is function on a discrete set.

5.2.3. *Example 3:* $S = \mathbb{H}^2$. Now $\mathrm{G} = \mathrm{PSL}_2(\mathbb{R})$ and $\mathrm{K} = \{\pm 1\}\backslash \mathrm{SO}(2)$. Actually, for simplicity of notation, let's just take $\mathrm{G} = \mathrm{SL}_2(\mathbb{R})$ and $\mathrm{K} = \mathrm{SO}(2)$. There will be no difference in the formulas.

Recall this identification $\mathbb{H}^2 = \mathbb{H} = \mathrm{G}\,/\,\mathrm{K}$ is given by writing every $z \in \mathbb{H}$ as $z = \frac{ai+b}{ci+d}$ for

$$g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{R}) = \mathrm{G}$$

and noting that $\mathrm{K} = \mathrm{SO}(2)$ stabilizes $x_0 = i$.

There is again only one differential operator, i.e. $\mathscr{D}(\mathbb{H}) = \mathbb{C}[\Delta]$, where is rectangular coordinates on $\mathbb{H}$,

$$\Delta = -y^2\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right).$$

We can rewrite this in geodesic polar coordinates around $x_0 = i$. In these coordinates, the metric becomes

$$ds^2 = dr^2 + \sinh^2 r d\theta^2$$

and Δ can be rewritten as

$$\Delta = -\left(\frac{\partial^2}{\partial r^2} + \frac{\cosh r}{\sinh r}\frac{\partial}{\partial r} + \frac{1}{\sinh^2 r}\frac{\partial^2}{\partial \theta^2}\right)$$

Note the similarities in the formulas to S^2.

It is easy to write down an eigenfunction for each eigenvalue. The Laplacian in rectangular coordinates is straightforward, and if we restrict to looking for eigenfunctions $f(z) = f(x, y) = f(y)$ which do not depend on x, we quickly end up with solving the ODE

$$y^2\frac{d^2}{dy^2}f + \lambda f = 0.$$

By inspection, we can write down solutions. First, note that the map $s \mapsto s(1-s)$ is surjective onto $\mathbb{C}$, i.e. we can always find an s s.t. $\lambda = s(1-s)$. Then note that if $\lambda \neq 1/4$, then the functions y^s, y^{1-s} are a set of linearly independent solutions for the ODE, while if $\lambda = 1/4$, then $y^{1/2}$ and $y^{1/2}\log y$ span all such eigenfunctions.

We take

$$f_s(z) = y^s$$

as a totally decent choice for an eigenfunction of eigenvalue $\lambda = s(1-s)$. These all have value 1 at i, and so we find, since $\Im(\frac{az+b}{cz+d}) = \frac{\Im z}{|cz+d|^2}$

$$w_\lambda(z) = \int_{\mathrm{K}} f_s(z) = \frac{(\Im z)^s}{2\pi}\int_{2\pi}\frac{1}{|\sin\theta z + \cos\theta|^s}d\theta.$$

While this is a formula, it may not seem very enlightening. So let's try our other approach of looking in radially symmetric functions for eigenfunctions.

Using our formula for Δ in geodesic polar coordinates, we find the ODE

$$\sinh r\frac{d^2\psi}{dr^2} + \cosh r\frac{d\psi}{dr} + s(1-s)\sinh r\psi.$$

Suppose $\psi(r) = F(\cosh r)$. Then this simplifies to

$$\frac{d}{du}((u^2-1)\frac{dF}{du}) + s(1-s)F = 0$$

Solutions to these are the Gauss hypergeometric functions. One can write these in integral form as

$$F_s(u) = \frac{1}{\pi}\int_0^\pi (2u + 1 + 2\sqrt{u(u+1)}\cos\theta)^{-s}d\theta$$

which you can also see from the above formula.

6. Kernels and the trace formula for compact quotients

We are finally ready to introduce a discrete group Γ.

Definition 6.1. A *locally symmetric space* Y is the quotient of a Riemannian symmetric space S by a discrete subgroup $\Gamma \subset \mathrm{G}$ acting without fixed points on S.

Remark 6.2. It probably better to give a more geometric definition. Here it is: a locally symmetric space Y is a connected complete Riemannian manifold s.t. every point $x \in Y$ has a normal neighborhood U s.t. geodesic inversion about x extends to an isometry of U. It then turns out, given this definition, that the universal cover of any such X is in fact a globally symmetric space S–this is a theorem of Cartan. See Kobayashi-Nomizu. We write $Y = Y(\Gamma)$, where $\Gamma \subset \mathrm{G} = \mathrm{Isom}(S)^\circ$ is such that $Y = \Gamma\backslash S$.

Remark 6.3. The condition that Γ acts fixed point freely is actually a bit restrictive for the class of quotients we care about, so we will omit it often. Instead, we will replace it with the condition that the discrete subgroup Γ acts on S with finite stabilizers, i.e. for all $x \in S$, $\#\Gamma_x < \infty$. This is equivalent to asking that Γ acts properly discontinuously.

Now, given a point pair invariant $k \in \mathscr{A}(S)$, we can construct an integral operator on $L^2(Y)$ by setting

$$K_k(x,y) = \sum_{\gamma\in\Gamma} k(x,\gamma y)$$

and considering, for $\phi \in L^2(Y)$,

$$T_k\phi(x) := \int_Y K_k(x,y)\phi(y)dy.$$

Observe that, since Γ acts properly discontinuously, that $K_k(x,y)$ is well defined for $x, y \in Y$ and smooth on $Y \times Y$.

This gives a map

$$\begin{aligned} \mathscr{A}(S) &\to \mathrm{End}(L^2(Y)) \\ k &\mapsto T_k \end{aligned}$$

and since we can write

$$T_k\phi(x) = \int_{\Gamma\backslash S} \sum_{\gamma\in\Gamma} k(x,\gamma y)\phi(y)dy = \int_S k(x,y)\phi(y)dy$$

by "unfolding the integral", it is clear that this morphism is really a homomorphism of algebras.

Finally, one should note that T_k as an operator on L^2Y satisfies

$$T_k^* = T_{\bar{k}}.$$

In particular, since the adjoint of any T_k also comes from a point-pair invariant and $\mathscr{A}(S)$ is commutative, every T_k is always normal, i.e. $T_kT_k^* = T_k^*T_k$.

6.1. **Spectral theory.** Recall that the trace formula is essentially a tool designed to give us access to information about the spectrum of $\mathscr{D}(S)$ in $L^2(Y)$, i.e. it is designed to describe eigenfunctions on the locally symmetric space $Y = \Gamma\backslash S$. Now, directly analyzing operators in $\mathscr{D}(S)$ such as the Laplacian is too technically hard, since these are all unbounded (densely defined) operators. However, since all of $\mathscr{A}(S)$ commutes with $\mathscr{D}(S)$ and acts on $L^2(Y)$ by integral kernels, we can hope to understand the spectrum of $\mathscr{D}(S)$ by looking at eigenvalues and eigenfunctions for T_k. This reduces down to using some very classical Hilbert space theory.

Here are some easy observations about T_k, stated as lemmas. Recall given a bounded operator $T : H \to H$ on a Hilbert space, we say T is *Hilbert-Schmidt* if $\sum_i ||T\phi_i||^2 < \infty$ for ϕ_i a fixed orthonormal basis of H. The quantity $||T||_2^2 := \sum_i ||T\phi_i||^2$ is independent of choice of basis, and is called the *Hilbert-Schmidt* norm of the operator T.

Lemma 6.4. *For all $k \in \mathscr{A}(S)$, the operator $T_k(x,y)$ is Hilbert-Schmidt on $L^2(Y)$.*

Proof. Since $K_k(x,y)$ is smooth on $Y \times Y$ and Y is compact, we have for free

$$\int_{Y\times Y} |K_k(x,y)|^2 dxdy < \infty.$$

This immediately implies T_k, since $||T_k||_2^2 = \int_{Y\times Y} |K_k(x,y)|^2 dxdy$

$$\begin{aligned}
||T_k||_2^2 &= \sum_i ||T_k\phi_i||^2 \\
&= \sum_i \sum_j |\langle T_k\phi_i, \phi_j\rangle|^2 \\
&= \sum_i \sum_j \left| \int_Y \int_Y K_k(x,y)\phi_i(y)\overline{\phi_j(x)} dxdy \right|^2 \\
&= \sum_{i,j} \left| \int_{Y\times Y} K_k(x,y)\overline{\overline{\phi_i}\otimes\phi_j(x,y)} dxdy \right|^2
\end{aligned}$$

and $\overline{\phi_i}\otimes\phi_j$ forms an orthonormal basis for $L^2(Y\times Y)$. Thus this last expression gives the L^2 norm of the function $K(x,y)$ on $Y\times Y$, hence

$$||T_k||_2^2 = \int_{Y\times Y} |K_k(x,y|^2 dxdy.$$

□

Lemma 6.5. *Every Hilbert-Schmidt operator is compact.*

Theorem 6.6 (Spectral theorem for compact normal operators)**.** *Let H be a Hilbert space and $T : H \to H$ a compact normal operator. Then there exists an orthonormal eigenbasis ϕ_i of H. If λ_i is the eigenvalue for ϕ_i, then $\lambda_i \to 0$ as $i \to \infty$. The eigenspace H_{λ_i} corresponding to $\lambda_i \neq 0$ is finite dimensional.*

The proofs of these last two statements are very standard, and left to the reader.

In any case, we now have the following spectral decomposition of $L^2(Y)$.

Proposition 6.7 ((Spectral decomposition for compact quotient)**.** *There is an orthonormal basis of joint eigenfunctions of $\mathscr{A}(S)$ and $\mathscr{D}(S)$ in $L^2(Y)$. Moreover, for each eigenparameter $\lambda : \mathscr{D}(S) \to \mathbb{C}$ the corresponding eigenspace is finite dimensional.*

Proof. The eigenspaces for $\mathscr{A}(S)$ are all finite dimensional: this follows because all operators K_k are compact, normal, and commute with one another. Thus they can be simultaneously diagonalized–the only danger comes from the possibility that they have an infinite dimensional common kernel. But this is impossible since there are approximations to the identity in $\mathscr{A}(S)$.

As for $\mathscr{D}(S)$, we know that if f is an eigenfunction for $\mathscr{D}(S)$ of eigenparameter λ, then f is an eigenfunction for $\mathscr{A}(S)$ since

$$Df = \hat{k}(\lambda)f.$$

Thus, finite dimensionality for eigenspaces of $\mathscr{A}(S)$ implies finite dimensionality for eigenspaces of $\mathscr{D}(S)$. □

6.2. **Traces.** We are almost ready to take traces. We need one last notion and a few easy lemmas from functional analysis before we can be sure that this is a well-defined thing to do.

Definition 6.8. Suppose $T : H \to H$ is a bounded operator on a Hilbert space. Then we say T is of *trace class* if for every orthonormal basis ϕ_i of H the sum

$$\sum_{i=1}^{\infty} |\langle T\phi_i, \phi_i\rangle|$$

converges.

Lemma 6.9. *If T is trace class, then*

$$\mathrm{Tr}(T) := \sum_{i=1}^{\infty} \langle T\phi_i, \phi_i\rangle$$

is independent of orthonormal basis ϕ_i.

Lemma 6.10. *Suppose $T : H \to H$ can be written as the composition of two Hilbert-Schmidt operators. Then T is trace class.*

The proofs of these two lemmas are hidden away in the exercises.

Finally, given an integral trace class operator, we can try to make sense of the trace as a "sum down the diagonal."

Lemma 6.11. *Suppose $k \in \mathscr{A}(S)$ satisfies $k = k_1 * k_2$ for $k_1, k_2 \in \mathscr{A}(S)$. Then*

(1) T_k is trace class.
(2) We have

$$\mathrm{Tr}(T_k) = \int_Y K_k(x,x)dk.$$

Proof. (1) follows since $k = k_1 * k_2$ implies that $T_k = T_{k_1}T_{k_2}$ is the composition of two Hilbert Schmidt operators, hence trace class. So it remains to show (2). To see this, let ϕ_i be an orthonormal basis of $L^2(Y)$, and compute

$$\begin{aligned}\mathrm{Tr}(T_k) &= \sum_{i=1}^{\infty}\langle T_k\phi_i, \phi_i\rangle \\ &= \sum_{i=1}^{\infty}\langle T_{k_2}\phi_i, T_{\bar{k_1}}\phi_i\rangle.\end{aligned}$$

If we write

$$T_{k_2}\phi_i = \sum_{j=1}^{\infty}\langle T_{k_2}\phi_i, \phi_j\rangle\phi_j$$

and similarly for $T_{\bar{k_1}}$, we find

$$\mathrm{Tr}(T_k) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\langle T_{k_2}\phi_i, \phi_j\rangle\overline{\langle T_{\bar{k_1}}\phi_i, \phi_j\rangle} = \langle K_{k_1}, K_{\bar{k}_2}\rangle_2$$

where this last expression is the inner product in $L^2(Y \times Y)$. Thus, we find

$$\mathrm{Tr}(T_k) = \int_{Y\times Y} K_{k_1}(x,y)\overline{K_{\bar{k_2}}(x,y)}dxdy = \int_{Y\times Y} K_{k_1}(x,y)K_{k_2}(x,y)dxdy = \int_Y K_k(x,x)dx$$

as desired. □

Remark 6.12. Note that the statement above is in some sense quite remarkable! Suppose $K(x,y) \in L^2(Y\times Y)$ is any integral kernel corresponding to a Hilbert-Schmidt operator. Then it would be a bit strange to talk about restricting $K(x,y)$ to the diagonal $Y^{\mathrm{diag}} \subset Y \times Y$ and integrating over this diagonal Y–the set we are integrating over is measure zero, and one can change $K(x,y)$ arbitrarily on this measure zero set and not affect the operator T corresponding to K.

6.3. **The trace formula for compact quotient.** OK, so we are essentially done. We can now happily compute $\mathrm{Tr}(T_k)$, for $k = k_1 * k_2$ a point pair invariant which is the convolution of two other point pair invariants, in two different ways. First, using the orthonormal eigenbasis ϕ_i described by Proposition 6.7, we have

$$\mathrm{Tr}(T_k) = \sum_i \langle T_k\phi_i, \phi_i\rangle = \sum_i \hat{k}(\lambda_i)$$

while on the other hand, we get by Lemma 6.11

$$\begin{aligned}\mathrm{Tr}(T_k) &= \int_{\Gamma\backslash S} \sum_{\gamma\in\Gamma} k(x,\gamma x)dx \\ &= \int_{\Gamma\backslash S} \sum_{\gamma\in\frac{\Gamma}{\Gamma}} \sum_{\delta\in\Gamma_\gamma\backslash\Gamma} k(x,\delta^{-1}\gamma\delta x)dx \\ &= \sum_{\gamma\in\frac{\Gamma}{\Gamma}} \int_{\Gamma\backslash S} \sum_{\delta\in\Gamma_\gamma\backslash\Gamma} k(\delta x,\gamma\delta x)dx \\ &= \sum_{\gamma\in\frac{\Gamma}{\Gamma}} \int_{\Gamma_\gamma\backslash S} k(x,\gamma x)dx\end{aligned}$$

Let's define the terms on the right to be *orbital integrals*, i.e. we write

$$\mathrm{Orb}(k,\gamma) := \int_{\Gamma_\gamma\backslash S} k(x,\gamma x)dx.$$

Note that if $\gamma = g^{-1}hg$, for some $h \in \mathrm{G}$, then

$$\mathrm{Orb}(k,\gamma) = \int_{\Gamma_\gamma\backslash S} k(x,\gamma x)dx = \int_{\Gamma_\gamma\backslash S} k(gx,\gamma' gx)dx = \int_{g\Gamma_\gamma g^{-1}\backslash S} k(x,hx)dx.$$

This will be important when we wish to be more explicit about orbital integrals next time.

Finally, note that these orbital integrals are over relatively simple domains $\Gamma_\gamma\backslash S$.

7. The trace formula for compact $\Gamma\backslash\mathbb{H}$

For today, and likely the immediate future, we restrict to the symmetric space $S = \mathbb{H} = \mathrm{SL}_2(\mathbb{R})/\mathrm{SO}(2)$. Recall that last time we had arrived, after much preamble, at a version of the Selberg trace formula valid when $Y = Y(\Gamma) = \Gamma\backslash\mathbb{H}$ is compact.

Proposition 7.1. *Suppose that Γ is a discrete subgroup of $\mathrm{G} = \mathrm{SL}_2(\mathbb{R})$ which is cocompact. Then, if $k \in \mathscr{A}(\mathbb{H})$ is a point pair invariant which can be written as the convolution of two other point pair invariants, then we have the equality*

$$\sum_{\lambda\in\sigma(\Delta)} \hat{k}(\lambda) = \sum_{\gamma\in\frac{\Gamma}{\Gamma}} \mathrm{Orb}(k,\gamma).$$

In the above, the orbital integrals on the RHS are given by

$$\mathrm{Orb}(k,\gamma) = \int_{\Gamma_\gamma\backslash\mathbb{H}} k(x,\gamma x)dx$$

and depend only on k and the conjugacy class of γ in Γ.

Today, our main goal is to make the expressions $\mathrm{Orb}(k,\gamma)$ more explicit.

7.1. **Conjugacy classes in $\mathrm{SL}_2(\mathbb{R})$.** First, let's think about what the possible conjugacy classes in $\mathrm{SL}_2(\mathbb{R})$ look like, in terms of their action on the upper half plane. Note that for $g \in \mathrm{SL}_2(\mathbb{R})$, we have the characteristic polynomial

$$T^2 - \mathrm{Tr}(g)T + 1 = 0$$

and of course this is conjugacy invariant. Behavior of eigenvalues breaks into three cases: whether the discriminant $\mathrm{Tr}(g)^2 - 4$ is negative, 0, or positive.

(1) $|\mathrm{Tr}(g)| < 2$. In this case we say g (or its conjugacy class) are *elliptic*.
(2) $|\mathrm{Tr}(g)| = 2$. In this case we say g (or its conjugacy class) are *parabolic*.
(3) $|\mathrm{Tr}(g)| > 2$. In this case we say g (or its conjugacy class) are *hyperbolic*.

[Draw picture of $\mathrm{SL}_2(\mathbb{R})$ and draw $|\,\mathrm{Tr}\,| = 2$.]

This rough classification of elements can also be explained in terms of the location of the fixed points of $g \in \mathrm{SL}_2(\mathbb{R})$ acting on $\mathbb{P}^1(\mathbb{C})$: to find fixed points, simply solve

$$az + b = (cz + d)z$$

to find that the fixed points are located at

$$\frac{a-d}{2c} \pm \frac{\sqrt{\mathrm{Tr}(g)^2 - 4}}{2c}.$$

Thus, we have that, if $g \neq \pm 1$,

(1) g is elliptic if and only if it has one fixed point in $\mathbb{H}$.
(2) g is parabolic if and only if it has one fixed point on the boundary $\mathbb{P}^1(\mathbb{R})$ of $\mathbb{H}$.
(3) g is hyperbolic if and only if it has two fixed points on the boundary $\mathbb{P}^1(\mathbb{R})$ of $\mathbb{H}$.

In terms of conjugacy classes this makes it very easy to visualize what is going on.

(1) If g is elliptic, let $z_0 \in \mathbb{H}$ be its fixed point in the upper half plane. Choose an element σ such that $\sigma z_0 = i$. Then $\sigma g \sigma^{-1} \in \mathrm{K} = \mathrm{SO}(2)$, so g is simply a rotation around z_0. If we are willing to talk about g up to $\mathrm{SL}_2(\mathbb{R})$, then we have a "model" for the conjugacy class, given by a rotation about i.
(2) If g is parabolic, let $z_0 \in \mathbb{P}^1(\mathbb{R})$ be its fixed point on the boundary of $\mathbb{H}$. Choose an element σ such that $\sigma z_0 = \infty$. Then $\sigma g \sigma^{-1} \in \pm\mathrm{N}$, where N is the group of upper unitriangular matrices in $\mathrm{SL}_2(\mathbb{R})$. These elements are our "model" for the conjugacy class of g, and act on $\mathbb{H}$ simply as $z \mapsto z + t$.
(3) If g is hyperbolic, let $z_0, z_1 \in \mathbb{P}^1(\mathbb{R})$ be the two distinct fixed point in the upper half plane. Draw the geodesic arc connecting them, and note that g must preserve this arc (not pointwise). Along the arc, one of the fixed points, say z_0, is repelling while the other is attracting. Choose an element σ such that $\sigma z_0 = 0$ and $\sigma z_1 = \infty$. Then $\sigma g \sigma^{-1} \in \pm\mathrm{A}^+$, where, A is the group of diagonal matrices in $\mathrm{SL}_2(\mathbb{R})$ and A^+ are those of the form

$$\begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix}$$

where $t > 0$. Our "model" is then simply scaling $z \mapsto e^t z$.

[Draw pictures.] Finally, let's say what the centralizers are.

(1) If g is elliptic, then G_g is conjugate to $\mathrm{K} = \mathrm{SO}(2)$.
(2) If g is parabolic, then G_g is conjugate to $\pm\mathrm{N}$.
(3) If g is hyperbolic, then G_g is conjugate to A, where A is diagonal matrices.

Note that $\Gamma_\gamma \subset \mathrm{G}_\gamma$ is always discrete. Thus we have

(1) If g is elliptic, then Γ_γ is finite.
(2) If g is parabolic, then Γ_γ is infinite cyclic, with generator γ_0.
(3) If g is hyperbolic, then Γ_γ is infinite cyclic, with generator γ_0.

Example 7.2. Let's look inside the discrete (but not cocompact!) group $\Gamma = \mathrm{SL}_2(\mathbb{Z})$.

(1) Convince yourself that the only elliptic elements, up to conjugation, are those that stabilize i and $\zeta_3 = e(1/3)$. So, for instance, the element
$$w_0 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$
is elliptic of order 2.
(2) The elements
$$\gamma = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}$$
are all parabolic. We can take
$$\gamma_0 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$
(3) Pick an element! Almost all are hyperbolic.

When Γ is cocompact, one of these types of conjugacy classes will not occur.

Lemma 7.3. *Suppose $\gamma \in \Gamma$ and Γ is cocompact. Then unless $\gamma = \pm 1$, γ cannot be parabolic.*

The proof is an easy exercise. One can also show

Lemma 7.4. *Suppose Γ is any discrete subgroup of G. Then there are only finitely many elliptic conjugacy classes on Γ.*

7.2. **Orbital integrals.**

7.2.1. *The identity element.* Suppose $\gamma = \pm 1 \in \mathrm{SL}_2(\mathbb{R})$ (equivalently is 1 in $\mathrm{PSL}_2(\mathbb{R})$). Then
$$\mathrm{Orb}(k, 1) = \int_{\Gamma\backslash\mathbb{H}} k(z, z)\frac{dxdy}{y^2} = k(0)\,\mathrm{vol}(\Gamma\backslash\mathbb{H})$$
where we have written $k(z, w) = k(u(z, w))$ where
$$u(z, w) = \frac{|z - w|^2}{4\Im z \Im w}$$
is the fixed point pair invariant we will write all others in terms of (this leads to nicer formulas than if we had used distance).

7.2.2. *Hyperbolic elements.* Write $\gamma = \gamma_0^l$, where γ_0 is a now fixed choice of generator of Γ_γ. We know $\Gamma_\gamma = \Gamma_{\gamma_0}$ are the same cyclic group. If we conjugate in G to send γ_0 to the map
$$z \mapsto pz$$
for $p = e^t > 1$ as above, then we find since then the generator of the conjugated Γ_{γ_0} sends $i \mapsto pi$
$$\begin{aligned}\mathrm{Orb}(k, \gamma) &= \int_{\Gamma_\gamma\backslash\mathbb{H}} k(z, \gamma_0^l z)\frac{dxdy}{y^2} \\ &= \int_1^p \int_{-\infty}^{\infty} k(z, p^l z)\frac{1}{y^2}dxdy\end{aligned}$$

Writing $k(z,w) = k(u(z,w))$ as above, we can, calling $d = \frac{1}{2}|(p^{l/2} - p^{-l/2}|$, simplify this to

$$\begin{aligned}\mathrm{Orb}(k,\gamma) &= \left(\int_1^p y^{-1}dy\right)\int_{-\infty}^{\infty} k(d^2(x^2+1))dx \\ &= \frac{\log p}{d}\int_{d^2}^{\infty}\frac{k(u)}{\sqrt{u-d^2}}du\end{aligned}$$

This can be massaged a little further, but for now let us just notice that this is a very explicit term involving $\log p$ and p^l. If one notices that $\log p$ is exactly the hyperbolic distance between i and pi, hence corresponds to the length of a closed geodesic in Y, this is quite striking!

7.2.3. *Elliptic elements.* Let γ be elliptic, with $\gamma = \gamma_0^l$, for γ_0 the generator of Γ_γ. Let m be the order in $\mathrm{PSL}_2(\mathbb{R})$ of γ_0. Since γ_0 is a rotation of angle $2\pi/m$, after conjugation γ_0 can be represented by

$$k(\theta) := \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

for $\theta = \pi/m$. A fundamental domain for $\Gamma_\gamma\backslash\mathbb{H}$ is a sector of angle 2θ and we find

$$\mathrm{Orb}(k,\gamma) = \frac{1}{m}\int_{\mathbb{H}} k(z, k(\theta l)z)\frac{dxdy}{y^2}$$

since m copies of this fundamental domain cover $\mathbb{H}$. After some manipulations, working in geodesic polar coordinates around i and once again writing $k(z,w) = k(u(z,w))$, we find

$$\mathrm{Orb}(k,\gamma) = \frac{\pi}{m}\int_0^{\infty} k(u\sin^2\frac{\pi l}{m})(u+1)^{-1/2}du.$$

8. Non-compact quotients of $\mathbb{H}$

Thus far, we have limited ourselves to thinking about the case of a compact hyperbolic Riemann surface $Y = \Gamma\backslash\mathbb{H}$ (and compact quotients of higher rank spaces). However, this rules out many important arithmetic example, for instance $\Gamma = \Gamma(1) := \mathrm{SL}_2(\mathbb{Z})$. So let's first be more careful about the class of discrete subgroups Γ we are interested in.

Definition 8.1. A *Fuchsian group* is a discrete subgroup Γ of $\mathrm{G} = \mathrm{SL}_2(\mathbb{R})$. If Γ is a Fuchsian group, then its *limit set* $L(\Gamma)$ is defined to be

$$L(\Gamma) = \{l \in \mathbb{P}^1(\mathbb{R}) : l \in \overline{\Gamma z_0} \text{ for some } z_0 \in \mathbb{H}\}.$$

We say Γ is *of the first kind* if $L(\Gamma) = \mathbb{P}^1(\mathbb{R})$, otherwise we say Γ is *of the second kind.*

Example 8.2. This definition is a bit opaque, but here is an example. Consider the group Γ generated by

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 4 \\ 0 & 1 \end{pmatrix}.$$

It is quite amusing to compute what $L(\Gamma)$ is–it is some sort of fractal, with Hausdorff dimension strictly less than 1 (think the Cantor set).

We often study discrete subgroups of isometries by drawing pictures of their fundamental domains: recall that a fundamental domain for Γ means a closed set $F \subset \mathbb{H}$ such that it is the closure of its interior $F = \overline{F^\circ}$ and so that for all $z \in F^\circ$, $F \cap \Gamma z = \{z\}$.

[Draw picture of fundamental domain for $\Gamma = \Gamma(1)$.] Note this fundamental domain is particularly nice: it is a (finite) hyperbolic polygon with vertices on the boundary. Unfortunately, not every Fuchsian group of the first kind has such a nice fundamental domain. However, we have the following fact, which we will not prove.

Proposition 8.3. *Every finitely generated Fuchsian group Γ of the first kind has finite covolume. Moreover, being finitely generated of the first kind is equivalent to the existence of a fundamental domain given by a hyperbolic polygon P whose intersection with $\mathbb{P}^1(\mathbb{R})$ consists only of vertices.*

Thus, from now on, *we will only work with Γ finitely generated Fuchsian groups of the first kind.* Since this is a mouthful, for this class let's just call such Γ "good Fuchsian groups."

Corollary 8.4. *If Γ is good, then any finite index subgroup Γ' of Γ is also good.*

Proof. Reserved for the exerices. Idea: pick coset representatives γ_i for Γ/Γ' cleverly so that $\cup_i \gamma_i P$, which is a fundamental domain for Γ', is connected, hence a hyperbolic polygon. □

8.1. **Cusps.** These vertices of the polygonal fundamental domain P on the boundary require special attention, since these are the things making $Y(\Gamma)$ non-compact.

Definition 8.5. A *cusp* of Γ is an element of

$$c(\Gamma) := \{c \in \mathbb{P}^1(\mathbb{R}) : \gamma c = c \text{ for some } \gamma \in \Gamma\}/\sim$$

where we say for $x, y \in \mathbb{P}^1(\mathbb{R})$ that $x \sim y$ if there exists $\gamma \in \Gamma$ such that $y = \gamma x$. That is, the set of cusps $c(\Gamma)$ is the set of elements stabilized by parabolic elements of Γ, up to the Γ action.

Example 8.6. For $\Gamma = \Gamma(1)$, it is easy to see that there is one cusp, since

$$c(\Gamma) = \mathbb{P}^1(\mathbb{Q})/\mathrm{SL}_2(\mathbb{Z} = \{\infty\}.$$

Compare this with the usual picture of the fundamental domain.

Lemma 8.7. *Suppose Γ is a good Fuchsian group, and P is a polygonal fundamental domain for Γ as above. Then the set of vertices of P on $\mathbb{P}^1(\mathbb{R})$ is a set of representatives for $c(\Gamma)$.*

Proof. Easy, but a bit tedious. I will omit this one. □

Remark 8.8. Although a cusp is properly an orbit of points on the boundary, we will often fix a polygonal fundamental domain in our minds and just think of the vertices of the polygon on the boundary as the cusps themselves. That is, we may abuse notation and use c both for the orbit of a point on the boundary, and for a fixed element on the boundary.

We can always move our cusps to ∞. That is, given a cusp c corresponding to the vertex of a fundamental polygon, there exists a $g_c \in \mathrm{G}$ such that $g_c c = \infty$. Note then that the group $g_c \Gamma g_c^{-1}$ has a cusp at ∞. Indeed, the group $g_c \Gamma_c g_c^{-1}$ is a discrete subgroup of $\mathrm{G}_\infty = \mathrm{B}$ (the upper triangular matrices) and cannot act by scaling on the upper half plane, so must be of the form

$$g_c \Gamma_c g_c^{-1} = \begin{pmatrix} 1 & r\mathbb{Z} \\ 0 & 1 \end{pmatrix}$$

for some positive real number r. Replacing g_c by the matrix

$$\sigma_c = \begin{pmatrix} r^{-1/2} & 0 \\ 0 & r^{1/2} \end{pmatrix} g_c$$

we find

$$\sigma_c \Gamma_c \sigma_c^{-1} = \begin{pmatrix} 1 & \mathbb{Z} \\ 0 & 1 \end{pmatrix}.$$

We call such a σ_c a *scaling matrix* for c.

8.2. **Automorphic forms.** When talking about compact $\Gamma\backslash\mathbb{H}$, we were using the following implicit definition.

Definition 8.9. Suppose $Y(\Gamma) = \Gamma\backslash\mathbb{H}$ is compact. Then an *automorphic form* for Γ is a finite linear combination of smooth eigenfunctions on $Y(\Gamma)$, i.e. a smooth $f : \mathbb{H} \to \mathbb{C}$ such that

(1) $f(\gamma z) = f(z)$ for all $\gamma \in \Gamma$
(2) f is a finite linear combinations of (generalized) eigenfunctions for Δ.

Remark 8.10. The finite linear combination condition is often rephrased in more intimidating language: note that f is a finite linear combination of (generalized) eigenfunctions on a locally symmetric space if and only if f is killed by a finite-codimension ideal $I \subset \mathscr{D}(H)$. Furthermore, this is equivalent to saying that $\mathscr{D}(H)f$ is a finite dimensional space of functions–you will often see this "$\mathfrak{Z}(\mathfrak{g}_\mathbb{C})$-finiteness" condition.

What about for Γ an arbitrary good Fuchsian group? We must add a condition.

Definition 8.11. Suppose Γ is an arbitrary good Fuchsian group. An *automorphic form* on $Y(\Gamma)$ is a smooth function $f : \mathbb{H} \to \mathbb{C}$ such that

(1) $f(\gamma z) = f(z)$ for all $\gamma \in \Gamma$
(2) f is a finite linear combinations of (generalized) eigenfunctions for Δ.
(3) f is of moderate growth along every cusp.

This last condition requires some more explanation. We say that f is moderate growth along c if

$${}^c f(z) := f(\sigma_c^{-1} z)$$

which is invariant under $\mathrm{N}(\mathbb{Z}) = \sigma_c \Gamma_c \sigma_c^{-1}$, satisfies

$$|\,{}^c f(x + iy)| \ll |y|^M$$

for some $M > 0$. (Recall Vinogradov notation: $f(x) \ll g(x)$ iff $f(x) = O(g(x))$.)

We write $A(\Gamma\backslash\mathbb{H})$ for the space of automorphic forms for Γ.

Remark 8.12. In terminology more consistent with the first lecture, what we have defined is also known as the space $A(\Gamma\backslash\mathbb{H})$ of *Maass forms of weight* 0 *and level* Γ.

Proposition 8.13. *Suppose f is an automorphic form for Γ of eigenvalue $\lambda = s(1-s)$, $s \neq \frac{1}{2}$. Then, for each cusp c of Γ, f has a Fourier expansion*

$${}^c f(z) = {}^c\alpha y^s + {}^c\beta y^{1-s} + \sum_{\substack{n \in \mathbb{Z} \\ n \neq 0}} {}^c a_n(f) y^{\frac{1}{2}} K_{s-\frac{1}{2}}(2\pi|n|y) e(nx)$$

where ${}^c\alpha, {}^c\beta, {}^c a_n(f)$ are all constants, and $K_r(t)$ is the modified Bessel function of the second kind.

Proof. (Sketch) This is not too surprising. For fixed y, the function $f^c(x+iy)$ is obviously periodic in x, wtih period $w(c)$. Thus we can write

$$f^c(z) = a_0(y) + \sum_{\substack{n\in\mathbb{Z}\\ n\neq 0}} a_n(y) e(\frac{nx}{w(c)}).$$

The condition that $\Delta f = s(1-s)f$ forces $a_n(y)$ to satisfy a particular ODE. But this equation has two solutions, one rapidly decaying in y -this is the modified Bessel function of the second kind K- and one rapidly increasing in y - the modified Bessel function of the first kind I. By the moderate growth condition on f, we must be using the rapid decay solution for all n. □

9. Introduction to Eisenstein series

Let Γ be a good Fuchsian group. We are trying to derive the Selberg spectral decomposition of the space $L^2(\Gamma\backslash\mathbb{H})$–this should be a decomposition of this space "into eigenfunctions". We have a good notion of the sorts of eigenfunctions we may need to understand this decomposition. These are the automorphic forms, which recall were defined as follows.

Definition 9.1. An *automorphic form* for Γ is a smooth function $f : \mathbb{H} \to \mathbb{C}$ such that

(1) $f(\gamma z) = f(z)$ for all $\gamma \in \Gamma$
(2) f is a finite linear combinations of (generalized) eigenfunctions for Δ.
(3) f is of moderate growth along every cusp, i.e. there exists for each cusp $c \in c(\Gamma)$ an $M > 0$ s.t.

$$|\,^c f(x+iy)| \ll y^M$$

We will often replace (2) with the more restrictive condition that f is a true eigenfunction of Δ of eigenvalue λ, and not just a linear combination of such.

Remark 9.2. Note that there is no guarantee that automorphic forms live in $L^2(\Gamma\backslash\mathbb{H})$. However, the flexibility of the moderate growth condition is ideal for explaining the decomposition of the "continuous part of the spectrum." A toy example to think about: the Fourier transform on $\mathbb{R}$ gives a decomposition of $L^2(\mathbb{R})$ into eigenspaces, but none of the eigenfunctions $e(\lambda x)$, $\lambda \in \mathbb{R}$ that contribute are square-integrable! However, they are at least moderate growth (in fact, they are bounded in this example).

This is great and all, but we would like to know how to construct such automorphic forms! Even more basically, we could ignore the eigenfunction condition and simply try to construct smooth functions on the quotient $Y(\Gamma)$.

To be a little careful about the types of functions we want to talk about, let's take a moment to define some notation.

9.1. Growth conditions on various spaces of functions.

Of fundamental importance is the space of *automorphic functions* for Γ.

Definition 9.3. An *automorphic function* f on $Y(\Gamma) = \Gamma\backslash\mathbb{H}$ is a smooth function $f : \mathbb{H} \to \mathbb{C}$ such that

(1) $f(\gamma z) = f(z)$ for all $\gamma \in \Gamma$.

We denote the space of all automorphic functions by $C^\infty(\Gamma\backslash\mathbb{H})$.

We can further impose the condition

(2) $f(z)$ is of moderate growth at each cusp c, i.e. for all $c \in c(\Gamma)$ there exists $M > 0$ so that as $y \to \infty$,

$$|\,^c f(x+iy)| \ll y^M.$$

If we do so, we call the space of such functions the space of *tempered automorphic functions*, and denote it by $C^\infty_{\text{temp}}(\Gamma\backslash\mathbb{H})$.

Finally, one could impose instead of the moderate growth condition the more restrictive condition of rapid decay (together with all derivatives)

(2)' $f(z)$ is of rapid decay at each cusp c, i.e. $^c f(x+iy)$ is Schwartz in y.

We write the space of such *rapidly decaying automorphic functions* as $\mathscr{S}(\Gamma\backslash\mathbb{H})$.

Remark 9.4. Let's think about $L^2(\Gamma\backslash\mathbb{H})$ for a moment. We can easily see that $\mathscr{S}(\Gamma\backslash\mathbb{H}) \subset L^2\Gamma\backslash\mathbb{H}$. This is because given a nice polygonal fundamental domain P and a $T > 0$ sufficiently large, we can write

$$P = P(T) \cup \bigcup_{c \in c(\Gamma)} P_c(T)$$

where each $P_c(T)$ is a cuspidal domain, isomorphic to $[0,1] \times [T, \infty)$. [Draw picture.] The condition that f is rapidly decreasing along each c ensures that the integrals over $P_c(T)$

$$\int_P |f(z)|^2 \frac{dxdy}{y^2} = \int_{P(T)} |f(z)|^2 \frac{dxdy}{y^2} + \sum_{c \in c(\Gamma)} \int_{P_c(T)} |f(z)|^2 \frac{dxdy}{y^2}$$

all converge.

Moreover, $\mathscr{S}(\Gamma\backslash\mathbb{H})$ is dense in L^2–in fact, even the smaller space of smooth compactly supported functions on $\Gamma\backslash\mathbb{H}$ is dense.

Since these conditions of moderate growth and rapid decay are all about the behavior of functions at cusps, it is convenient to have a "model space" for functions near each cusp. Consider the space $\mathrm{N}\backslash\mathbb{H}$, where $\mathrm{N} \subset \mathrm{G}$ is the group of upper uni-triangular matrices. We can identify

$$\begin{aligned} \mathrm{N}\backslash\mathbb{H} &\xrightarrow{\sim} \mathbb{R}_{>0} \\ x+iy &\mapsto y. \end{aligned}$$

Since, given a cusp c we can use a scaling matrix[3] σ_c to move the cusp to ∞, we can similarly identify

$$\sigma_c^{-1}\,\mathrm{N}\,\sigma_c\backslash\mathbb{H} \xrightarrow{\sim} \mathbb{R}_{>0}.$$

Note that, for a cusp c and a fixed choice of scaling matrix σ_c with $\sigma_c c = \infty$ and with $\sigma_c\Gamma_c\sigma_c^{-1} = \pm\mathrm{N}(\mathbb{Z})$[4], we have the diagram

$$\begin{array}{ccc} & \Gamma_c\backslash\mathbb{H} & \\ \swarrow & & \searrow \\ \Gamma\backslash\mathbb{H} & & \sigma_c^{-1}\,\mathrm{N}\,\sigma_c\backslash\mathbb{H} \end{array} \quad . \tag{9.1}$$

If $c = \infty$ and $\Gamma_c = \mathrm{N}(\mathbb{Z})$ (e.g. this happens for the one cusp of $\Gamma = \Gamma(1)$) then this becomes the easy diagram

$$\begin{array}{ccc} & \mathrm{N}(\mathbb{Z})\backslash\mathbb{H} & \\ \swarrow & & \searrow \\ \Gamma\backslash\mathbb{H} & & \mathrm{N}\backslash\mathbb{H} \end{array} \quad .$$

The RHS of this diagram is our model space for the cusp; the space in the middle "is" the cusp (asymptotically).

The notions of moderate growth and rapid decay make perfect sense on the model space $\mathrm{N}\backslash\mathbb{H} \cong \mathbb{R}_{>0}$.

Definition 9.5. Let $\psi \in C^\infty(\mathbb{R}_{>0})$.

(1) We say ψ is of *moderate growth* if
$$|\psi(y)| \ll y^M$$
for some $M > 0$ as $y \to \infty$ and if
$$|\psi(y)| \ll y^{-N}$$
for some $N > 0$ as $y \to 0$. The space of smooth moderate growth functions will be denoted by $C^\infty_{\mathrm{temp}}(\mathbb{R}_{>0})$.

(2) We say ψ is *rapid decay* if ψ is Schwartz. The space of such functions will be denoted by $\mathscr{S}(\mathbb{R}_{>0})$.

We can use the diagram (9.1) and "push-pull" to construct various maps between spaces of functions.

9.2. **Constant terms and incomplete Eisenstein series.** Let $c \in c(\Gamma)$ be a cusp. Consider the map

$$\begin{aligned} \mathrm{const}_c : C^\infty(\Gamma\backslash\mathbb{H}) &\to C^\infty(\mathbb{R}_{>0}) \\ f &\mapsto \int_{\Gamma_c\backslash\sigma_c^{-1}\,\mathrm{N}\,\sigma_c} f(nz)dn = \int_0^1 f(\sigma_c^{-1}\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}\sigma_c z)dx \end{aligned}$$

This may look a little opaque, but think about what it means when $c = \infty$ and $\Gamma_c = \mathrm{N}(\mathbb{Z})$. Then

$$\mathrm{const}_\infty(f)(y) = \int_0^1 f(x+iy)dx.$$

In general if we identify $\sigma_c^{-1}\,\mathrm{N}\,\sigma_c\backslash\mathbb{H}$ with $\mathbb{R}_{>0}$ we have

$$\mathrm{const}_c f(y) = \int_0^1 {}^c f(x+iy)dx.$$

The terminology const_c is apt: this operator sends a function to the constant coefficient of ${}^c f$ in its Fourier expansion.

---

[3] If you want to write the RHS of the following identification without reference to a scaling matrix, then this is also doable: simply note that $\sigma_c^{-1}\,\mathrm{N}\,\sigma_c$ as $\mathrm{N}_c :=$ the unipotent radical of G_c.

[4] We are going to stop writing the ± 1, i.e. we may be a little blasé about the difference between $\mathrm{PSL}_2(\mathbb{R})$ and $\mathrm{SL}_2(\mathbb{R})$.

We can also play the "push pull" game in the other direction. Here we have to be careful because of convergence issues. Still, everything is easily seen to be absolutely convergent when we start with rapidly decaying functions

$$\begin{aligned}\mathrm{Eis}_c : \mathscr{S}(\mathbb{R}_{>0}) &\to \mathscr{S}(\Gamma\backslash\mathbb{H}) \\ \psi &\mapsto \sum_{\gamma\in\Gamma_c\backslash\Gamma} \psi(\Im(\sigma_c\gamma z)).\end{aligned}$$

Again, when $c=\infty$ and $\Gamma_c = \mathrm{N}(\mathbb{Z})$, this looks slightly less intimidating:

$$\mathrm{Eis}_c(\psi)(z) = \sum_{\gamma\in\pm\mathrm{N}(\mathbb{Z})\backslash\mathrm{SL}_2(\mathbb{Z})} \psi(\Im\gamma z).$$

Those functions which are of the form $\mathrm{Eis}_c(\psi)$ for some ψ are called *incomplete Eisenstein series*. We denote by

$$\mathcal{E}_c = \overline{\{\mathrm{Eis}_c(\psi) : \psi\in\mathscr{S}(\mathbb{R}_{>0})\}}$$

the closure in $L^2(\Gamma\backslash\mathbb{H})$ of the subspace of incomplete Eisenstein series coming from the cusp c.

Here are some basic relations between the operators const_c and Eis_c.

Proposition 9.6. *We have the following.*

(1) For a fixed $c\in c(\Gamma)$, the operators const_c and Eis_c are adjoint, i.e. for $\psi\in\mathscr{S}(\mathbb{R}_{>0})$ and $f\in\mathscr{S}(\Gamma\backslash\mathbb{H})$

$$\langle\psi, \mathrm{const}_c(f)\rangle_{L^2(\mathbb{R}_{>0},\frac{dy}{y^2})} = \langle\mathrm{Eis}_c(\psi), f\rangle_{L^2(\Gamma\backslash\mathbb{H})}$$

(2) If $c_1\neq c_2$ are two cusps of Γ, then for any $\psi_1,\psi_2\in\mathscr{S}(\mathbb{R}_{>0})$, we have

$$\langle\mathrm{Eis}_{c_1}(\psi_1), \mathrm{Eis}_{c_2}(\psi_2)\rangle_{L^2(\Gamma\backslash\mathbb{H})} = 0$$

Proof. We prove (1). For simplicity of notation, let's assume $c=\infty$ and $\sigma_c = 1$. Then

$$\begin{aligned}\langle\mathrm{Eis}_c(\psi), f\rangle &= \int_{\Gamma\backslash\mathbb{H}} \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} \psi(\Im(\gamma z)) f(z)\frac{dxdy}{y^2} \\ &= \int_{\Gamma_\infty\backslash\mathbb{H}} \psi(y) f(x+iy)\frac{dxdy}{y^2} \\ &= \int_0^\infty \psi(y)\int_0^1 f(x+iy)dx\frac{dy}{y^2} = \langle\psi, \mathrm{const}_c(f)\rangle_{L^2(\mathbb{R}_{>0},\frac{dy}{y^2})}\end{aligned}$$

as desired.

(2) is an exercise. □

A quick corollary of (1) is that

Corollary 9.7.

$$\ker\mathrm{const}_c = \mathscr{S}(\Gamma\backslash\mathbb{H})\cap(\mathcal{E}_c)^\perp$$

We will say

Definition 9.8. An automorphic function $f\in C^\infty(\Gamma\backslash\mathbb{H})$ is *cuspidal* if $\mathrm{const}_c(f) = 0$ for all $c\in c(\Gamma)$.

We get for free from the adjointness of Eis_c and const_c that

$$L^2(\Gamma\backslash\mathbb{H}) = L^2_{\mathrm{cusp}}(\Gamma\backslash\mathbb{H})\oplus\bigoplus_{c\in c(\Gamma)}\mathcal{E}_c$$

where we have set $L^2_{\mathrm{cusp}}(\Gamma\backslash\mathbb{H})$ to be the closure of those cuspidal automorphic functions in L^2.

Our goal now is two fold. First, we have to understand how the Laplacian decomposes on L^2_{cusp} and on each $\mathcal{E}_c$; second, we have to understand how to make sense of some version of the kernel function K_k these spaces. Note that, as we saw for compact quotient, understanding of these problems must happen simultaneously–we use the kernels K_k to diagonalize Δ.

9.3. **Eisenstein series, finally.** The spectral decomposition on each $\mathcal{E}_c$ is not so bad, but it crucially requires a particular sort of automorphic form: the Eisenstein series.

Where does this come from? Well, simply apply the averaging procedure we used to define the incomplete Eisenstein series to an eigenfunction y^s on $\mathbb{R}_{>0}$ (eigenfunction means eigenfunction for the differential operator $-y^2\frac{d^2}{dy^2}$). This gives the following definition.

Definition 9.9. For each cusp c, consider the subgroup Γ_c stabilizing c. The *Eisenstein series* corresponding to c is given by the sum

$$E_c(z,s) := \sum_{\gamma\in\Gamma_c\backslash\Gamma} \Im(\sigma_c\gamma z)^s$$

when it converges.

Here are the basic properties of this series.

Proposition 9.10. *$E_c(z,s)$ has the following properties.*

(1) $E_c(z,s)$ converges absolutely for $\Re(s)>1$. For $s\in K$, with K a compact subset of $\{\Re(s)>1\}$, the convergence is uniform in s.

For each fixed s with $\Re(s)>1$, the function $z\mapsto E(z,s)$ satisfies

(2) $E_c(\gamma z,s)=E_c(z,s)$ for all $\gamma\in\Gamma$.
(3) $\Delta E_c(z,s)=s(1-s)E_c(z,s)$.
(4) ${}^cE_c(z,s)=y^s+O(1)$ as $y\to\infty$.

Proof. Let us assume, without loss of generality, that $c=\infty$ and $\sigma_c=1$. Fix a polygonal fundamental domain P contained in the strip $\{z:0<\Re(z)<1\}$. Fix a set of representatives γ_i of $\Gamma_\infty\backslash\Gamma$ so that $\gamma_i P$ are all in $\{z:0<\Re(z)<1\}$ as well (we can do this by translating back into the strip). We can assume too that the representative for the identity coset is $\gamma_0=1$.

Fix $z=z_0$. We wish to show (1). We may assume $s=\sigma$ is real, since any imaginary component does not affect absolute convergence. We will use the mean value property of eigenfunctions. Pick a $\delta>0$ sufficiently small so that the geodesic balls $\gamma_i B(z_0,\delta)$ are all disjoint. We know by the mean value property for eigenfunctions (5.1) that

$$\int_{B(z_0,\delta)} \Im(z)^\sigma \frac{dxdy}{y^2} = \frac{\Im(z_0)^\sigma}{c(\delta,\sigma)}$$

where $c(\delta,\sigma)$ is simply the integral of the zonal spherical function centered at z_0 of eigenvalue $\sigma(1-\sigma)$ around $B(z_0,\delta)$ radially.

Thus, $\Im(z_0)^\sigma = c(\delta,\sigma)\int_{B(z_0,\delta)}\Im(z)^\sigma\frac{dxdy}{y^2}$. Therefore,

$$E(z_0,\sigma)=y^\sigma+\sum_{i\neq 0}\Im(\gamma_i z_0)^\sigma = y^\sigma + c(\delta,\sigma)\sum_{i\neq0}\int_{B(\gamma_i z_0,\delta)}\Im(z)^\sigma\frac{dxdy}{y^2} = y^\sigma + c(\delta,\sigma)\int_{\cup_{i\neq0}B(\gamma_i z_0,\delta)} y^\sigma\frac{dxdy}{y^2}.$$

But note that this last integral takes place within a domain $0\le x\le 1$ and $y\le D$ for some $D>0$. Thus, we have

$$E(z_0,\sigma)=y^\sigma+O(\int_0^D\int_0^1 y^{\sigma-2}dxdy)$$

and this last term

$$\int_0^D\int_0^1 y^{\sigma-2}dxdy = \int_0^D y^{\sigma-2}dy$$

converges when $\sigma>1$. □

10. The spectral expansion: cuspidal part

Last time we deduced a very coarse first approximation to the spectral decomposition. This says simply that for Γ a good Fuchsian group, the space of square integrable functions breaks into an orthogonal direct sum

$$L^2(\Gamma\backslash\mathbb{H}) = L^2_{\text{cusp}}(\Gamma\backslash\mathbb{H}) \oplus \bigoplus_{c\in c(\Gamma)} \mathcal{E}_c.$$

We came about this by defining, for every cusp $c \in c(\Gamma)$, the adjoint pair of operators "take constant coefficients" and "attach the incomplete Eisenstein series."

We also introduced the true Eisenstein series associated to each cusp

$$E_c(z,s) := \sum_{\gamma\in\Gamma_c\backslash\Gamma} \Im(\sigma_c\gamma z)^s.$$

Since this came about by averaging an eigenfunction y^s and not just a rapidly decaying function $\psi \in \mathscr{S}(\mathbb{R}_{>0})$, this has some convergence issues when $\Re(s) \leq 1$; however, it converges absolutely when $\Re(s) > 1$. When it converges, it is an automorphic form.

Among all automorphic forms, what makes $E_c(z,s)$ so special (i.e. why do we care about this)? Consider the formula for the incomplete Eisenstein series

$$\text{Eis}_c(\psi)(z) = \sum_{\gamma\in\Gamma_c\backslash\Gamma} \psi(\sigma_c\gamma z).$$

We can expand $\psi \in \mathscr{S}(\mathbb{R}_>0)$ in eigenfunctions for $\mathbb{R}_{>0}$. This means simply that we can use the Mellin inversion formula, i.e. for any fixed σ,

$$\psi(y) = \frac{1}{2\pi i}\int_{\Re(s)=\sigma} \mathcal{M}(\psi)(-s)y^s ds$$

where

$$\mathcal{M}(\psi)(s) = \int_0^\infty \psi(t)t^s\frac{dt}{t}.$$

If we apply this to the definition of the incomplete Eisenstein series, we get that for each $\sigma > 1$,

$$\text{Eis}_c(\psi)(z) = \frac{1}{2\pi i}\int_{\Re(s)=\sigma} \mathcal{M}(\psi)(-s)E_c(z,s)ds.$$

Thus, the importance of the Eisenstein series comes from the fact that they "span" the spaces $\mathcal{E}_c$ in this sense.

For each s with $\Re(s) > 1$, $E_c(z,s)$ is an automorphic form. It therefore enjoys a Fourier expansion

$${}^cE_c(z,s) = {}^c\alpha y^s + {}^c\beta y^{1-s} + \sum_{\substack{n\in\mathbb{Z}\\ n\neq 0}} {}^ca_n(f)y^{\frac{1}{2}}K_{s-\frac{1}{2}}(2\pi|n|y)e(nx).$$

By the growth conditions we already know on $E(z,s)$, it follows that the coefficient of y^s is 1. We use special notation for the other term in the constant coefficient, i.e. we write $\phi_{c,c}(s) := {}^c\beta$. By some easy manipulations, we can control the higher Fourier coefficients as well, and get using the rapid decay of the K-Bessel function

$${}^cE_c(z,s) = y^s + \phi_{c,c}(s)y^{1-s} + O((1+y^{\Re(s)})e^{-2\pi y}).$$

In general, we can take the c'-th Fourier expansion of the Eisenstein series corresponding to c. This gives

Proposition 10.1. *Let s be such that $\Re(s) > 1$. We have*

$${}^{c'}E_c(z,s) = \delta_{c,c'}y^s + \phi_{c,c'}(s)y^{1-s} + O((1+y^{\Re(s)})e^{-2\pi y})$$

where $\delta_{c,c'}$ is the Kronecker delta.

Remark 10.2. Note that the growth of $E_c(z,s)$ would be smallest when $s = \frac{1}{2}$, where we will have ${}^cE_c(z,s) \sim Cy^{\frac{1}{2}}$. This just barely fails to be in $L^2(\Gamma\backslash\mathbb{H})$–computing the L^2 norm is like integrating $\int_1^\infty y^{-1}dy$. This maybe helps to show why we would like to analytically continue the function $E(z,s)$.

To force square integrability, even rapid decay, we can truncate this. This will help us understand "the projection of a function $f \in L^2(\Gamma\backslash\mathbb{H})$ to $\mathcal{E}_c$. We will deal with this in the future.

For today, we focus attention on the cuspidal part of L^2.

10.1. **A rough analysis of $K_k(z,w)$.** We are still interested in the integral kernel

$$K_k(z,w) = \sum_{\gamma\in\Gamma} k(z,\gamma w).$$

Let's write our point pair invariant $k(z,w)$ as a function of

$$u(z,w) = \frac{|z-w|^2}{4\Im(z)\Im(w)} = \frac{(x-u)^2}{4yv} + \frac{y}{4v} + \frac{v}{4y} - \frac{1}{2}$$

where[5] $z = x+iy$ and $w = u+iv$. That is, we write

$$k(z,w) = k(u(z,w)).$$

Suppose for simplicity that k has compact support–this isn't so much stronger than the rapid decay condition from earlier. Suppose too, just to make notation clearer, that Γ is such that $c(\Gamma) = \{\infty\}$ and $\sigma_\infty = 1$.

If z is fixed and $w \to \infty$ in the cusp, then $K_k(z,w)$ becomes 0 and vice versa. The problem is if $z, w \to \infty$ while staying close to one another (e.g. in within translates of $\mathrm{Supp}(k)$ of one another).

So suppose that z, w are high up in the fundamental domain. [Draw picture.] The only translates which are close to one another, so have any hope of contributing to the sum $K_k(z,w)$ are those horizonal translates. So, for z, w high up enough, we have

$$K_k(z,w) = \sum_{m\in\mathbb{Z}} k(z, w+m)$$

We can use Poisson summation to rewrite this. That is, for z, w high enough in the cusp, we rewrite

$$K_k(z,w) = \int_{-\infty}^{\infty} k(z,w+t)dt + \sum_{m\neq 0}\int_{-\infty}^{\infty} k(\frac{(x-(u+t))^2}{4yv} + \frac{y}{4v} + \frac{v}{4y} - \frac{1}{2})e(mt)dt$$

Some simple examination of the Fourier coefficients shows that

$$\sum_{m\neq 0}\int_{-\infty}^{\infty} k(\frac{(x-(u+t))^2}{4yv} + \frac{y}{4v} + \frac{v}{4y} - \frac{1}{2})e(mt)dt \ll_N \frac{1}{(yv)^N}$$

so in total, we have

$$K_k(z,w) = \int_{-\infty}^{\infty} k(z,w+t)dt + O_N((yv)^{-N})$$

In particular, we have shown that for z, w large enough in the cusp, $K_k(z,w)$ is essentially given by its constant term (which grows like $\sqrt{yv}$, so fails to be L^2).

This explains why the kernel is not Hilbert-Schmidt. It also explains a way to modify it into a Hilbert-Schmidt kernel.

10.2. **Truncation.** So now fix a $T > 0$ large. Let $\alpha_T(y)$ be a smooth function with

$$\alpha_T(y) = \begin{cases} 1 & \text{if } y > T \\ 0 & \text{if } y < T-1 \end{cases}$$

and satisfying

$$0 \leq \alpha_T(y) \leq 1.$$

Take $\alpha_T(z) := \alpha_T(\Im(z))$. We define

Definition 10.3. The *smooth truncation* of $K_k(z,w)$ with respect to α_T is

$$\widetilde{K}_k(z,w) = K_k(z,w) - \alpha_T(z)\int_{-\infty}^{\infty} k(z,w+t)dt$$

Note that if we view this as a function on the fundamental domain, it descends to an automorphic function on $\Gamma\backslash\mathbb{H}$ in both variables.

[5]Sorry about all the u's!

Remark 10.4. It should be clear that this truncated kernel really is a Hilbert-Schmidt kernel on $\Gamma\backslash\mathbb{H}$–by our analysis above, $\widetilde{K}_k$ is rapidly decaying along the cusp.

Remark 10.5. One can also do a non-smooth truncation, where we replace α_T with $\mathbb{1}_{[T,\infty)]}$. This introduces annoying analytic difficulties into the story.

Proposition 10.6. *Consider the operator $\widetilde{T}_k$. We have*

$$\widetilde{T}_k|_{L^2_{\text{cusp}}(\Gamma\backslash\mathbb{H})} = T_k|_{L^2_{\text{cusp}}(\Gamma\backslash\mathbb{H})}$$

Proof. Compute that for $f \in C^\infty_{\text{cusp}}(\Gamma\backslash\mathbb{H})$, we have

$$\begin{aligned}\int_{\Gamma\backslash\mathbb{H}} f(w)\alpha_T(z)\int_{-\infty}^{\infty} k(z,w+t)dt\frac{dudv}{v^2} &= \int_{\Gamma\backslash\mathbb{H}} f(w)\alpha_T(z)\sum_{m\in\mathbb{Z}}\int_0^1 k(z,w+m+t)dt\frac{dudv}{v^2}\\ &= \int_{\Gamma\backslash\mathbb{H}} \alpha_T(z)\sum_{m\in\mathbb{Z}} k(z,w+m)\int_0^1 f(w+t)dt\frac{dudv}{v^2}\\ &= 0\end{aligned}$$

since this inner integral vanishes by cuspidality of f. □

This completely gives the spectral decomposition in L^2_{cusp}. We know that

- $\widetilde{K}_k$ is a Hilbert-Schmidt kernel, hence $\widetilde{T}_k$ is a compact operator on $L^2(\Gamma\backslash\mathbb{H})$. Thus, $\widetilde{T}_k|_{L^2_{\text{cusp}}(\Gamma\backslash\mathbb{H})}$ is compact.
- $\widetilde{T}_k|_{L^2_{\text{cusp}}(\Gamma\backslash\mathbb{H})} = T_k|_{L^2_{\text{cusp}}(\Gamma\backslash\mathbb{H})}$, hence normal.

and so we can once again apply the spectral theorem to conclude that L^2_{cusp} has a basis of eigenfunctions for $T_k|_{L^2_{\text{cusp}}}$. But these all commute with the Laplacian! That is, we have shown

$$L^2_{\text{cusp}}(\Gamma\backslash\mathbb{H}) = \bigoplus_{\lambda\in\sigma_{\text{cusp}}(\Delta)} H_\lambda$$

where, given λ, $H_\lambda = \{f : \Delta f = \lambda f\}$, and with each H_λ finite dimensional. $\sigma_{\text{cusp}}(\Delta)$ has no accumulation points.

We can say epsilon more. The restriction $\widetilde{T}_k|_{L^2_{\text{cusp}}}$ is again trace class, by the same proof as before, and so once again we have a discrete spectral contribution

$$\text{Tr}(\widetilde{T}_k|_{L^2_{\text{cusp}}}) = \sum_{\lambda\in\sigma_{\text{cusp}}(\Delta)} \hat{k}(\lambda)$$

to what will be the trace formula.

Remark 10.7. This whole analysis has never shown that the space L^2_{cusp} is non-zero. Indeed, we actually expect that for most Γ with a cusp, this space vanishes.

11. Analytic continuation of Eisenstein series

Today, we must confront the analytic (really, meromorphic) continuation and functional equation of Eisenstein series. To simplify the discussion, we will assume throughout today that Γ is a good Fuchsian group, with only one cusp ∞, and that $\sigma_\infty = 1$.

In this context, the result we are after is the following.

Theorem 11.1. *The Eisenstein series $E(z,s)$ has meromorphic continuation to the whole plane. It satisfies*

(1) $E(z,1-s) = \phi(1-s)E(z,s)$
(2) $\phi(s)\phi(1-s) = 1$
(3) $|\phi(\frac{1}{2}+i\tau)| = 1$, *hence ϕ has no poles on* $\Re(s) = \frac{1}{2}$.

In the above, given the meromorphy of $E(z,s)$, $\phi(s)$ is the meromorphic function which appears as the coefficient of y^{1-s} in the constant term of $E(z,s)$.

We will spend today proving this result. Actually, we will only prove the meromorphy–all other statements essentially follow from the proof.

11.1. **An easy uniqueness result.** The heart of the proof of Theorem 11.1 comes from exploiting various characterizing features of $E(z,s)$. In the end, we want to show, after meromorphic continuation, that $\phi(s)E(z,1-s)$ has the same unique property as $E(z,s)$. What is the sort of property we are looking for? Here is a warm up

Lemma 11.2. *Suppose f is an automorphic form on $\Gamma\backslash\mathbb{H}$. Suppose f is an eigenfunction of Δ, of eigenvalue $\lambda = s(1-s)$ for s with $\Re(s) > 1$. Then*

$$f(z) = \alpha E(z,s)$$

for some α.

Proof. Look at the Fourier expansion of f, given by

$$f(z) = \alpha y^s + \beta y^{1-s} + O(1)$$

and note that $f(z) - \alpha E(z,s) = (\beta - \alpha\phi(s))y^{1-s} + O(1)$ must be in L^2. It is also an eigenfunction of eigenvalue $\lambda = s(1-s)$. But integration by parts shows any L^2 eigenfunction of Δ must have positive eigenvalue, which would force s to either have $\Re(s) = \frac{1}{2}$ or $s \in \mathbb{R}$. Since we are assuming $\Re(s) > 1$, the first case cannot occur, and for the second, note that $s(1-s) < 0$ if s is real and greater than 1. Hence $f(z) - \alpha E(z,s)$ must be 0. □

Remark 11.3. Explain the analogue of this proposition for when Γ has many cusps. What changes?

Remark 11.4. This characterization of Eisenstein series is on paper nice, but also a bit worthless. It very crucially relies on $\Re(s) > 1$, which is exactly the region we already understand! So we need a more sophisticated characterization which has hope of saying more. This will come from Fredholm theory

11.2. **Fredholm theory, in brief.** Fredholm theory is concerned with the following problem: let X be a space, and $K(x,y)$ an integral kernel on $L^2(X)$, with T its corresponding operator. How can you find solutions ψ to an equation of the form

$$(T-\lambda)\psi = f$$

where $f \in L^2(X)$.

Proposition 11.5 (The Fredholm alternative)**.** *Suppose we are in the above setting. Fix λ. If K is a Hilbert-Schmidt kernel, then either*

(1) The equation

$$(T-\lambda)\psi = 0$$

has a solution, or

(2) The operator

$$(T-\lambda)^{-1}$$

which may be a priori defined on only a dense subspace of $L^2(X)$, is in fact bounded and so extends everywhere. That is, for every f, the equation

$$(T-\lambda)\psi = f$$

has a unique solution. Moreover, that unique solution depends analytically on $\lambda \in \mathbb{C} - \{\text{eigenvalues of } T\}$

I will not prove this here–the proof is a bit delicate, but again not particularly hard (once you know the trick).

11.3. **A not quite Fredholm equation.** Recall that we have defined a smoothly truncated kernel function via

$$\begin{aligned}\widetilde{K}_k(z,w) &= K_k(z,w) - \alpha_T(z)\int_{\mathbb{R}} k(z,w+t)dt \\ &= K_k(z,w) - \alpha_T(z)\int_0^1 \sum_{n\in\mathbb{Z}} k(z,w+n+t)dt\end{aligned}$$

where $\alpha_T(y)$ is a smooth function which is 1 for $y > T$ and 0 for $y < T-1$. Let's call $K_0(z,w) := \alpha_T(z)\int_{\mathbb{R}} k(z,w+t)dt$.

We will use $\widetilde{K}_k = K_k - K_0$ to help establish analytic continuation of $E(z,s)$. The strategy is as follows: we know $\widetilde{K}(z,s)$ is a Hilbert-Schmidt kernel, and we want to solve the Fredholm equation we get when we apply this operator to $E(z,s)$. This is obviously no good, since $E(z,s)$ is not in $L^2(\Gamma\backslash\mathbb{H})$, so Fredholm's theorems don't seem to apply (at least not as we have explained them). But let's work formally and see what we have.

Note that the operator defined by K_0 acts on $E(z,s)$ only through its constant term, i.e.

$$\begin{aligned}T_0E(z,s) &= \int_{\Gamma\backslash\mathbb{H}} \alpha_T(w)\int_0^1 \sum_{n\in\mathbb{Z}} k(z,w+n+t)dtE(w,s)dw \\ &= \alpha_T(z)\int_{\mathrm{N}(\mathbb{Z})\backslash\mathbb{H}} \sum_{n\in\mathbb{Z}} k(z,w+n+t)dtE(w,s)dw \\ &= \alpha_T(z)\int_{\mathbb{H}} k(z,w)\int_0^1 E(w-t,s)dtdw \\ &= \alpha_T(z)\hat{k}(s(1-s))(y^s+\phi(s)y^{1-s})\end{aligned}$$

and so computing formally how $\widetilde{K}_k$ acts on $E(z,s)$, we find

$$\widetilde{T}_kE(z,s) = \hat{k}(s(1-s))E(z,s) - \alpha_T(z)\hat{k}(s(1-s))(y^s+\phi(s)y^{1-s})$$

which we rewrite as

$$(\widetilde{T}_k - \hat{k}(s(1-s)))E(z,s) = -\alpha_T(z)\hat{k}(s(1-s))(y^s+\phi(s)y^{1-s}). \tag{11.1}$$

OK, this is a totally fine equation, valid for $\Re(s) > 1$, but we have some issues. First of all neither $E(z,s)$ nor the function $\alpha_T(z)\hat{k}(s(1-s))(y^s+\phi(s)y^{1-s})$ are in L^2, so we can't simply apply Fredholm theory to win. Even worse, we know nothing about the function $\phi(s)$ as a function of s, so we couldn't say much using this equation even if both sides were in L^2 since we don't know for which s $\phi(s)$ can be continued to. Still, this suggests the following principle:

The analytic continuation of Eisenstein series follows from the analytic continuation of its constant term. This idea is Langlands' *principle of the constant term*.

OK, with that philosophy squared away, what are we really going to do? We will remove $\phi(s)$ from the "Fredholm" equation (11.1) and see what we get.

11.4. **Auxiliary equations.** Let's try to solve the equation

$$(\widetilde{T}_k - \hat{k}(s(1-s)))E^*(z,s) = -\alpha_T(z)\hat{k}(s(1-s))y^s \tag{11.2}$$

where we simply ignore the $\phi(s)y^{1-s}$ term in (11.1). It is worth noting that the RHS of (11.2) is still not in L^2, so this is still not attackable using Fredholm theory. Still, we can do better. Consider the even more degenerate

$$(\widetilde{T}_k - \hat{k}(s(1-s)))E^{**}(z,s) = \widetilde{K}_k.\alpha_T(z)y^s. \tag{11.3}$$

Notation is getting a little confusing, but the RHS in (11.3) means apply $\widetilde{T}_k$ to the function $\alpha_T(z)y^s$. We write it as we did above to emphasize that we are not trying to invert this appearance of $\widetilde{T}_k$–view the RHS as simply an L^2 function.

Where did equation (11.3) even come from? Well, suppose you want to solve (11.2). It's not clear such an $E^*(z,s)$ exists for most s, but if it did, we could try to make it L^2 by cutting off, i.e. replacing it by

$$E^*(z,s) - \alpha_T(z)y^s$$

checking to see how $\widetilde{K}(z,w)$ acts on this function shows this is an $E^{**}(z,s)$ satisfying (11.3).

Good, so now (11.3) is an honest Fredholm equation. We can therefore solve it for s outside of the spectrum of $\widetilde{K}(s(1-s))$, which is a countable set. This means that for almost all s, there is a unique solution $E^{**}(z,s) \in L^2$–moreover, this is meromorphic in s, with location of poles completely independent of z. Taking $E^*(z,s) = E^{**}(z,s) + \alpha_T(z)y^s$, it follows that we have a unique meromorphic solution to our first auxiliary equation.

Let's now try to put back $\phi(s)y^{1-s}$ to get at the true equation characterizing the Eisenstein series. Observe that

$$(\widetilde{T}_k - \hat{k}(s(1-s)))(E^*(z,s) + \phi(s)E^*(z,1-s)) = -\alpha_T(z)\hat{k}(s(1-s))(y^s + \phi(s)y^{1-s}). \tag{11.4}$$

This is the same differential equation as (11.1) ! We want then to conclude that

$$E(z,s) = E^*(z,s) + \phi(s)E^*(z,1-s)$$

when $\Re(s) > 1$ so that $E(z,s)$ has meromorphic continuation to all of $\mathbb{C}$ if and only if $\phi(s)$ does. To conclude this, we muscle our way back to L^2. Look at the difference

$$F(z,s) := E(z,s) - (E^*(z,s) + \phi(s)E^*(z,1-s))$$

which very clearly satisfies the homogeneous Fredholm equation

$$(\widetilde{T}_k - \hat{k}(s(1-s)))F(z,s) = 0$$

for $\Re(s) > 1$. Observe that $F(z,s)$ has no constant coefficient, and thus that $\widetilde{T}_k F = T_k F$, so that F is an eigenfunction for all the T_k operators. But then F is an honest L^2 eigenfunction for Δ of eigenvalue $s(1-s)$ with $\Re(s) > 1$ and so by our easy uniqueness lemma, it must be 0.

Thus, we need only understand the meromorphy of $\phi(s)$.

Proposition 11.6. *For $\Re(s) > 1$ and s not a pole of $E^*(z,s)$ or $E^*(z,1-s)$, the function*

$$E_\lambda(z,s) := E^*(z,s) + \lambda E^*(z,1-s)$$

is an eigenfunction for Δ with eigenvalue $s(1-s)$ if and only if $\lambda = \phi(s)$.

Proof. We already essentially saw and used one direction (that it is an eigenfunction if $\lambda = \phi(s)$). For the other direction, suppose it is an eigenfunction. Take the difference and apply the easy uniqueness lemma to win. □

Finally, to see meromorphy of $\phi(s)$, consider, for each fixed $z \in P$, P the fundamental domain, the equation

$$\Delta(E^*(\cdot,s) + \lambda E^*(\cdot,1-s))|_{\cdot=z} = s(1-s)(E^*(z,s) + \lambda E^*(z,1-s)).$$

By the proposition above, there is one value of λ which solves this equation for all $z \in P$ simultaneously, namely $\lambda = \phi(s)$. But there is a $z_0 \in P$ for which there is a unique λ for which

$$\Delta(E^*(\cdot,s) + \lambda E^*(\cdot,1-s))|_{\cdot=z_0} = s(1-s)(E^*(z_0,s) + \lambda E^*(z_0,1-s))$$

If we solve this linear ODE (in s) for such λ, if follows that $s \mapsto \lambda(s)$ is meromorphic. But this function is $\phi(s)$ by the proposition above!

12. Residues of Eisenstein series and the spectral expansion

Last lecture we deduced the analytic continuation of Eisenstein series from some careful application of Fredholm theory. We did this under the notationally simplifying assumption that Γ had a single cusp ∞, and that $\sigma_\infty = 1$.

For completeness, let me state the theorem on Eisenstein series in the general context.

Once again, throughout today we will assume that Γ is a good Fuchsian group with a single cusp ∞, and that $\sigma_\infty = 1$.

Today, we want to use that knowledge to complete our understanding of the spectral expansion. There is one final tool we need: the (sharply) truncated Eisenstein series.

Definition 12.1. Let $T \gg 0$ be a large fixed postive number. Let

$$E^T(z,s) := \begin{cases} E(z,s) - y^s - \phi(s)y^{1-s} & \text{if } y \geq T \\ E(z,s) & \text{if } y < T \end{cases}$$

be the *(sharply) truncated Eisenstein series.*

Remark 12.2. It is worth noting that this is the same procedure we did earlier when we defined $\widetilde{K}_k$–we subtract away the constant term when $\Im(z)$ is large–but here we are doing this sharply, without using our smooth cutoff function α_T.

The sharp cutoff makes E^T live in L^2, and we can compute the following

Proposition 12.3 (Maass-Selberg relation)**.** *Suppose $s_1, s_2 \in \mathbb{C}$ are both regular points of $E(z,s)$. Suppose further that $s_1 \neq \overline{s_2}$ and $s_1 + \overline{s_2} \neq 1$. Then*

$$\langle E^T(\cdot, s_1), E^T(\cdot, s_2)\rangle = (s_1 - \overline{s_2})^{-1}\left(\overline{\phi(s_2)}T^{s_1-\overline{s_2}} - \phi(s_1)T^{\overline{s_2}-s_1}\right) + (s_1 + \overline{s_2} - 1)^{-1}\left(T^{s_1+\overline{s_2}-1} - \phi(s_1)\overline{\phi(s_2)}T^{1-s_1-\overline{s_2}}\right)$$

Proof. (Sketch) By analytic continuation, this only needs to be proved for $\Re(s_1), \Re(s_2) > 1$.

Work over the truncated fundamental domain $P^T = \{z \in P : \Im(z) < T\}$. We can apply Green's theorem to reduce the LHS to an integral over the boundary of P^T, and then compute. The integral over $\Im(z) = T$ can be computed using the Fourier expansion of $E(z,s)$, and only the zeroth terms end up surviving the integration.

Roughly, this is because if we were integrating over all of P just $E(z,s)$ -which we are not allowed to do since nothing converges- we would get 0. But E^T is just E after we have fiddled only with the constant term, thus, only the constant terms can contribute to the RHS. □

This result has some nice corollaries. Write $s = \sigma + i\tau$ and apply the proposition above with $s_1 = s_2 = \sigma + i\tau$. Then we find

$$||E^T(\cdot, \sigma + i\tau)||^2 = (2i\tau)^{-1}\left(\overline{\phi(\sigma+i\tau)}T^{2i\tau} - \phi(\sigma+i\tau)T^{-2i\tau}\right) + (2\sigma-1)^{-1}\left(T^{2\sigma-1} - |\phi(\sigma+i\tau)|^2 T^{1-2\sigma}\right). \tag{12.1}$$

Apply this for $\sigma \to \frac{1}{2}$.

Corollary 12.4. *We have*

$$||E^T(\cdot, \sigma+i\tau)||^2 = 2\log T - \frac{\phi'}{\phi}(\frac{1}{2}+i\tau) + (2i\tau)^{-1}\left(\overline{\phi(\frac{1}{2}+i\tau)}T^{2i\tau} - \phi(\frac{1}{2}+i\tau)T^{-2i\tau}\right)$$

The appeal of this expression is that it gives us some understanding of ϕ on the line $\Im(s) = \frac{1}{2}$.

We can also take (12.1) and fix $\sigma \neq 1/2$ and let $\tau \to 0$ to see

Corollary 12.5.

$$||E^T(\cdot, \sigma+i\tau)||^2 = (2\sigma-1)^{-1}\left(T^{2\sigma-1} - |\phi(\sigma)|^2 T^{1-2\sigma}\right) + \phi(\sigma)\log T - \phi'(\sigma).$$

If we think a bit, this gives, together with $|\phi(\frac{1}{2}+i\tau)| = 1$

Proposition 12.6. *$\phi(s)$ is holomorphic for $\Re(s) \geq 1/2$ except for a finite number of simple poles in the segment $(1/2, 1]$. The residue of such a pole is real and positive.*

Proof. (Sketch) Let s_0 be a pole of $\phi(s)$ with $\Re(s_0) \geq 1/2$. Since $|\phi(\frac{1}{2} + i\tau)| = 1$ we know $\Re(s_0) > 1/2$. Use (12.1) to conclude that s_0 must be a pole of $E(z, s)$. But then the leading term of $E(z, s)$ at s_0 is an L^2 automorphic form of eigenvalue $s_0(1 - s_0)$, so s_0 must be real. We know $s_0 \leq 1$. To see that s_0 must be simple, carefully apply Corollary 12.5. □

Since we know that the poles of $E(z, s)$ are a subset of the poles of $\phi(s)$ (counted with multiplicity), we get from this that

Proposition 12.7. *$E(z, s)$ is holomorphic for $\Re(s) \geq 1/2$ except for a finite number of simple poles in the segment $(1/2, 1]$. If s_0 is such a pole, $\operatorname{res}_{s=s_0} E(z, s)$ is a square integrable automorphic form of eigenvalue $s_0(1 - s_0)$.*

Proof. Suppose $u(z)$ is such a residue. We have not really justified why $u(z)$ is an L^2 form, although we used this above. Let's do this now: we have

$$u(z) = \lim_{s \to s_0} (s - s_0) E(z, s)$$

and it is clear that $u(z)$ is moderate growth. We can use its Fourier expansion along the cusp to write

$$u(z) = \operatorname{res}_{s=s_0} \phi(s) y^{1-s_0} + O(1).$$

Since $s_0 > 1/2$ this is L^2. □

12.1. **The spectral expansion.** We are finally able to state the spectral expansion carefully. We have essentially proved it already.

Theorem 12.8. *The space $\mathcal{E} = \mathcal{E}_\infty$ of incomplete Eisenstein series splits orthogonally into Δ invariant subspaces*

$$\mathcal{E} = \mathcal{R} \oplus \mathcal{E}_{\text{cont}}$$

where the spectrum of Δ is discrete and consists of a finite number of $\lambda_j = s_j(1 - s_j)$ where $1/2 < s_j \leq 1$. It is spanned by the residues u_j of the Eisenstein series. The spectrum of Δ on $\mathcal{E}_{\text{cont}}$ is absolutely continuous and covers $\lambda \in [1/4, \infty)$ uniformly with multiplicity 1. We can expand any $f \in \mathcal{E}$ as

$$f(z) = \sum_j \langle f, u_j \rangle u_j(z) + \frac{1}{4\pi} \int_{-\infty}^{\infty} \langle f, E(\cdot, \frac{1}{2} + it) \rangle E(z, \frac{1}{2} + it) dt$$

Proof. Simply write f as an incomplete Eisenstein series in terms of the Mellin transform, and compute. Moving the integration over a vertical $\Re(s) = \sigma$ for $\sigma > 1$ to $\Re(s) = 1/2$ picks up the residual forms. □

13. The spectral side of the Selberg trace formula

The analysis thus far has been in pursuit of the following formula. To state it cleanly, recall some notation.

- $k \in \mathscr{A}(\mathbb{H})$ is a point pair invariant.
- $h(t) = \hat{k}((\frac{1}{2}+it)(\frac{1}{2}-it))$ is the Selberg transform, evaluated at an eigenvalue $\lambda = s(1-s)$ corresponding to $s = \frac{1}{2}+it$.
- $h(t)$ can be derived from $k(t)$ either by an integration next to the Gauss hypergeometric function (using the eigenfunction characterization of $\hat{k}$) or by the following steps:
  (1) For $k(z,w) = k(u(z,w)$ with $u(z,w) = \frac{|z-w|^2}{4\Im z \Im w}$, consider
  $$q(v) = \int_v^\infty k(u)(u-v)^{-1/2}du.$$
  (2) Set
  $$g(r) = 2q((\sinh\frac{r}{2})^2).$$
  (3) Take
  $$h(t) = \int_{-\infty}^\infty e^{irt}g(r)dr.$$

  This procedure can be inverted to derive k from h. This goes as follows:
  (1) Fourier invert to get
  $$g(r) = \frac{1}{2\pi}\int_{-\infty}^\infty e^{irt}h(t)dt.$$
  (2) Take
  $$q(v) = \frac{1}{2}g(2\log(\sqrt{v+1}+\sqrt{v})).$$
  (3) Note that
  $$k(u) = -\frac{1}{\pi}\int_u^\infty (v-u)^{-1/2}dq(v).$$

In practice, I will try to use these formulas as little as possible, but you should not be afraid of them. They are all a consequence of the eigenfunction characterization of $\hat{k}$ and some simple manipulations of integrals.

We put the following restrictions on h, hence on k.

(1) For all sufficiently small $\varepsilon > 0$, $h(t)$ is holomorphic in the strip $|\Im t| \le \frac{1}{2}+\varepsilon$ and
(2) $h(t) \ll_\varepsilon (|t|+1)^{-2-\varepsilon}$ in this strip.

If k is Schwartz, it is easy to check that h satisfies this. In addition, h will always be even.

Theorem 13.1 (Selberg trace formula)**.** *Let Γ be a good Fuchsian group, and let $k \in \mathscr{A}(\mathbb{H})$ be as above. Write the eigenvalues λ_j of Δ on $L^2_{\text{disc}}(\Gamma\backslash\mathbb{H})$ as $\lambda_j = \frac{1}{4}+t_j^2$, with $t_j \ge 0$ if $\lambda_j \ge \frac{1}{4}$ and $\Re(t_j) > 0$ if $0 \le \lambda_j \le \frac{1}{4}$. Let $\varphi(s) = \det\Phi(s)$ be the determinant of the scattering matrix. Then*

$$\begin{aligned}\sum_j h(t_j) + \frac{1}{4\pi}\int_{-\infty}^\infty h(r)\frac{-\varphi'}{\varphi}(\frac{1}{2}+ir)dr + \frac{h(0)}{4}\operatorname{Tr}(\Phi(\frac{1}{2}))&\\ = \frac{\operatorname{vol}(\Gamma\backslash\mathbb{H})}{4\pi}\int_{-\infty}^\infty h(r)r\tanh(\pi r)dr&\\ + \sum_P\sum_{l=1}^\infty \frac{\log \mathrm{N}(P)}{\mathrm{N}(P)^{l/2}-\mathrm{N}(P)^{-l/2}}g(l\log\mathrm{N}(P))&\\ + \sum_R\sum_{0<l<m}\frac{1}{2m\sin\frac{\pi l}{m}}\int_{-\infty}^\infty h(r)\frac{\cosh(\pi(1-2l/m)r)}{\cosh(\pi r)}dr&\\ + \#c(\Gamma)\left(\frac{h(0)}{4} - g(0)\log 2 - 2\pi\int_{-\infty}^\infty h(r)\frac{\Gamma'}{\Gamma}(1+ir)dr\right).&\end{aligned}$$

This formula is undoubtedly a bit intimidating. However, many terms in it are not so crazy, once we take a closer look. On the LHS we have the spectral contribution, and on the RHS the geometric contribution. Each line on the RHS corresponds to a contribution from a different type of conjugacy class in Γ; the trivial class, hyperbolic classes, elliptic classes, and parabolic classes.

To show such a thing, we would like to interpret both sides as the trace of the operator coming from the kernel $K_k(z,w)$. As we are well aware, this is meaningless since $K_k(z,w)$ fails to be trace class, or even Hilbert Schmidt. So instead what we will do is compute "the truncated trace"

$$\operatorname{Tr}^T K_k = \int_{P^T} K_k(z,z)\frac{dxdy}{y^2}$$

where P^T is the "core" of the polygonal fundamental domain P, i.e. what remains after we cut off each cusp at height T (draw picture). What will occur is that the truncated spectral trace will compute

$$\operatorname{Tr}^T K_k \sim S_{\text{spec}} + R_{\text{spec}} \log T$$

while computing the geometric side will give

$$\operatorname{Tr}^T K_k \sim S_{\text{geom}} + R_{\text{geom}} \log T.$$

The trace formula will be the equality

$$S_{\text{spec}} = S_{\text{geom}}.$$

13.1. **The truncated spectral trace.** To access the "spectral trace" we will use the spectral decomposition. Applying it to K_k itself gives:

Proposition 13.2. *Suppose $k \in \mathscr{A}(\mathbb{H})$, so in particular satisfies the conditions on h above. Then*

$$K_k(z,w) = \sum_j h(t_j)u_j(z)\overline{u_j(w)} + \sum_{c\in c(\Gamma)} \frac{1}{4\pi}\int_{-\infty}^{\infty} h(r)E_c(z,\frac{1}{2}+ir)\overline{E_c(w,\frac{1}{2}+ir)}dr$$

where u_r are an orthonormal eigenbasis for L^2_{disc}. All sums and integrals above are absolutely convergent.

Remark 13.3. The condition on growth of h should now be quite apparent: it is to ensure that the integral above will converge.

Strictly speaking, to prove this we can't just blindly apply the spectral decomposition to $K_k(z,w)$, since it does not lie in L^2. To properly derive this proposition from the spectral decomposition, one has to first work with the truncated kernel, spectrally expand, and then add back in the constant term. Since this is very close to arguments we have already seen ad nauseam, we omit this careful derivation.

We can now compute. We get

$$\operatorname{Tr}^T K_k = \sum_j h(t_j)\int_{P^T}|u_j(z)|^2\frac{dxdy}{y^2} + \frac{1}{4\pi}\int_{-\infty}^{\infty} h(r)\sum_{c\in c(\Gamma)}\int_{P^T}|E_c^T(z,\frac{1}{2}+ir)|^2\frac{dxdy}{y^2}dr.$$

Note that we have replaced E_c by the truncated Eisenstein series since they agree on the core P^T. We can bound this above by what we get by applying the appropriate integration over all of P. Note that since we have replaced E_c by the truncated Eisenstein series E_c^T, the integration in question now converges. This gives the upper bound

$$\operatorname{Tr}^T K_k \leq \sum_j h(t_j) + \frac{1}{4\pi}\sum_c \int_{-\infty}^{\infty} ||E_c^T(\cdot,\frac{1}{2}+ir)dr||^2$$

and we can write

$$\begin{aligned}\sum_c \int_{-\infty}^{\infty} ||E_c^T(\cdot,\frac{1}{2}+ir)dr||^2 &= \operatorname{Tr}\langle \mathcal{E}^T(\cdot,\tfrac{1}{2}+ir), {}^t\mathcal{E}^T(\cdot,\tfrac{1}{2}+ir)\rangle \\ &= \frac{1}{2ir}\operatorname{Tr}(\Phi(\tfrac{1}{2}-ir)Y^{2ir} - \Phi(\tfrac{1}{2}+ir)Y^{-2ir}) + 2\#c(\Gamma)\log T - \frac{\varphi'}{\varphi}(\tfrac{1}{2}+ir).\end{aligned}$$

In the above, we have used the Maass-Selberg relations and linear algebra trickery, i.e. the fact that

$$\frac{\varphi'}{\varphi}(s) = \operatorname{Tr}\Phi'(s)\Phi(s)^{-1}$$

(to see this linear algebra trick, diagonalize $\Phi(s)$ by a unitary matrix).

For now, let's focus on the terms not involving $\log T$. For instance, consider the matrix valued

$$I(T) := \frac{1}{4\pi}\int_{-\infty}^{\infty} \frac{h(r)}{2ir}\left(\Phi(\frac{1}{2}-ir)Y^{2ir} - \Phi(\frac{1}{2}+ir)Y^{-2ir}\right)dr.$$

Using that $h(r) = h(-r)$ and adding and subtracting $\Phi(\frac{1}{2}$, we get

$$I(T) = \frac{1}{4\pi i}\int_{-\infty}^{\infty} \frac{h(r)}{r}\left(\Phi(\frac{1}{2}-ir)Y^{2ir} - \Phi(\frac{1}{2})\right).$$

If we shift contours up to $\Im(r) = \varepsilon$, then since $\Phi(s)$ is bounded in this small strip, we get

$$I(Y) = -\Phi(\frac{1}{2})\frac{1}{4\pi i}\int_{\Im(r)=\varepsilon} \frac{h(r)}{r}dr + O_\varepsilon(Y^{-2\varepsilon})$$

and since

$$\frac{1}{2\pi i}\int_{\Im(r)=\varepsilon} \frac{h(r)}{r}dr = -\frac{1}{2}h(0)$$

by contour integration (move to $\Im(r) = -\varepsilon$ and use symmetry), we get

$$I(Y) = \frac{1}{4}\Phi(\frac{1}{2})h(0) + O_\varepsilon(Y^{-2\varepsilon}).$$

Taking the trace and going back to our inequality gives

$$\mathrm{Tr}^T K_k \leq \sum_j h(t_j) + \frac{1}{4\pi}\int_{-\infty}^{\infty} \frac{-\varphi'}{\varphi}(\frac{1}{2}+ir)h(r)dr + \frac{1}{4}h(0)\,\mathrm{Tr}\,\Phi(\frac{1}{2}) + g(0)\#c(\Gamma)\log T + O_\varepsilon(T^{-\varepsilon})$$

Now, this inequality fails to be an equality only because we have not subtracted away those terms coming from integrations near the cusps. We can show these are negligible as functions of T, i.e. contribute at most $O_\varepsilon(T^{-\varepsilon})$. That is,

Lemma 13.4. *Suppose u_j is cusp form, normalized to have L^2-norm 1.*

$$\int_{P-P^T} |u_j(z)|^2 \frac{dxdy}{y^2} \ll |s_j| Y^{-2}.$$

Lemma 13.5. *Suppose u_j is a residue of an Eisenstein series, corresponding to $\frac{1}{2} < s_j \leq 1$ and normalized to have L^2-norm 1. Then*

$$\int_{P-P^T} |u_j(z)|^2 \frac{dxdy}{y^2} \ll Y^{1-2s_j}.$$

Lemma 13.6. *We have*

$$\int_{-R}^{R} \sum_c ||E_c^T(\cdot, \frac{1}{2}+ir)||^2_{P-P^T} dr \ll R^3 Y^{-1}.$$

Remark 13.7. Unfortunately Lemma 13.6 does not seem to be enough to conclude the bound

$$\frac{1}{4\pi}\int_{-\infty}^{\infty} h(r) \sum_c ||E_c^T(\cdot, \frac{1}{2}+ir)||^2_{P-P^T} dr \ll Y^{-\varepsilon}$$

from the growth constraint

$$h(r) \ll (|t|+1)^{-2-\varepsilon}$$

that we know. To run the game, we need the stronger

$$h(r) \ll (|t|+1)^{-4-\varepsilon}$$

but we can assume this, and then deduce the full result for the weaker condition on h by absolute convergence of everything in sight and analytic continuation.

For some measure of completeness, let us show Lemma 13.4. TO BE ADDED- sketch: use standard bound on cusp forms (or residual forms) coming from Fourier expansion and amplification.

14. The parabolic contribution and a first application

We were deriving the Selberg trace formula last time. The basic strategy was as follows: given a smooth point pair invariant k, we compute the truncated trace

$$\operatorname{Tr}^T K_k = \int_{P^T} K_k(x,x)dx$$

in two different ways: by applying the spectral expansion of $K_k(x,y)$, and by unfolding $K_k(x,x)$ into a sum over conjugacy classes. We carried out the spectral computation last time: it gave

$$\operatorname{Tr}^T K_k = \sum_j h(t_j) + \frac{1}{4\pi}\int_{-\infty}^{\infty} \frac{-\varphi'}{\varphi}(\frac{1}{2}+ir)h(r)dr + \frac{1}{4}h(0)\operatorname{Tr}\Phi(\frac{1}{2}) + g(0)\#c(\Gamma)\log T + O_\varepsilon(T^{-\varepsilon}).$$

That is, as a function of T,

$$\operatorname{Tr}^T K_k \sim S_{\text{spec}} + R_{\text{spec}}\log T$$

where

$$S_{\text{spec}} = \sum_j h(t_j) + \frac{1}{4\pi}\int_{-\infty}^{\infty} \frac{-\varphi'}{\varphi}(\frac{1}{2}+ir)h(r)dr + \frac{1}{4}h(0)\operatorname{Tr}\Phi(\frac{1}{2})$$

and

$$R_{\text{spec}} = g(0)\#c(\Gamma)$$

Today, we will show that the geometric expansion of K_k into a sum over conjugacy classes also gives that

$$\operatorname{Tr}^T K_k \sim S_{\text{geom}} + R_{\text{geom}}\log T$$

for some constants S_{geom} and R_{geom}, and confirm that $R_{\text{geom}} = g(0)\#c(\Gamma) = R_{\text{spec}}$. This will give an equality

$$S_{\text{spec}} = S_{\text{geom}}$$

which, after some simplification, yields the trace formula.

Something quite surprising happens in the computation of the geometric truncated trace. We will proceed similarly to how things worked for the compact quotient case, first noting that

$$K_k(z,z) = \sum_{\gamma\in\frac{\Gamma}{\Gamma}} K_{k,\gamma}(z) = K_{k,\text{triv}}(z) + K_{k,\text{hyp}}(z) + K_{k,\text{ell}}(z) + K_{k,\text{par}}(z)$$

where

$$K_{k,\gamma}(z) = \sum_{\delta\in\Gamma_\gamma\backslash\Gamma} k(\delta z, \gamma\delta z)$$

and

$$K_{k,\text{triv}} = K_{k,1} = k(z,z)$$
$$K_{k,\text{hyp}} = \sum_{\substack{\gamma\in\frac{\Gamma}{\Gamma}\\ \gamma\text{hyperbolic}}} K_{k,\gamma}$$
$$K_{k,\text{ell}} = \sum_{\substack{\gamma\in\frac{\Gamma}{\Gamma}\\ \gamma\text{elliptic}}} K_{k,\gamma}$$
$$K_{k,\text{par}} = \sum_{\substack{\gamma\in\frac{\Gamma}{\Gamma}\\ \gamma\text{parabolic}}} K_{k,\gamma}$$

breaks up the sum over $\frac{\Gamma}{\Gamma}$ into the four types of conjugacy classes.[6] Then we have

$$\operatorname{Tr}^T K_k = \operatorname{Tr}^T K_{k,\text{triv}} + \operatorname{Tr}^T K_{k,\text{hyp}} + \operatorname{Tr}^T K_{k,\text{ell}} + \operatorname{Tr}^T K_{k,\text{par}}$$

and a real miracle occurs: when we compute $\operatorname{Tr}^T K_{k,\text{par}}$, we will find that as a function of T is asymptotically looks like

$$\operatorname{Tr}^T K_{k,\text{par}} \sim S_{\text{par}} + g(0)\#c(\Gamma)\log T. \tag{14.1}$$

[6] We call these expressions the *partial kernels*.

Since the constant next to $\log T$ is the *same* as the coefficient of $\log T$ in the spectral side it follows that, at least for positive k, the other partial kernels cannot have $\lim \mathrm{Tr}^T \to \infty$ as $T \to \infty$, hence we can simply compute their contribution by integrating over all of P, not just P^T (i.e. take trace not truncated trace). For general k, the same holds by writing k as the difference of two positive functions.

14.1. **The truncated trace of the parabolic partial kernel.** Let's show (14.1). We can compute by first parameterizing those parabolic conjugacy classes. Since a parabolic element of Γ must fix a (representative of a) cusp $c \in c(\Gamma)$, and since the stabilizer of a cusp is infinite cyclic, it follows that every parabolic conjugacy class $\gamma \in \frac{\Gamma}{\Gamma}$ is indexed by a pair (c, l) where $c \in c(\Gamma)$ and $l \in \mathbb{Z} - \{0\}$. For each parabolic conjugacy class γ, we have

$$\mathrm{Tr}^T K_{k,\gamma} = \int_{P^T} \sum_{\delta \in \Gamma_\gamma \backslash \Gamma} k(\delta z, \gamma \delta z) \frac{dxdy}{y^2} = \int_{\Gamma_\gamma \backslash \mathbb{H}^T} k(z, \gamma z) \frac{dxdy}{y^2}$$

where $\mathbb{H}^T$ is simply $\mathbb{H}$ but with the T-neighborhoods of the cusps removed. [Draw picture for a $\Gamma(1)$.]

If we conjugate γ by the scaling matrix σ_c, where c is the cusp corresponding to γ, then we find

$$\mathrm{Tr}^T K_{k,\gamma} = \int_{\mathrm{N}(\mathbb{Z}) \backslash \sigma_c^{-1} \mathbb{H}^T} k(z, z + l) \frac{dxdy}{y^2}$$

Note that we can think of $\mathrm{N}(\mathbb{Z}) \backslash \sigma_c^{-1} \mathbb{H}^T$ as a truncated fundamental domain contained in the box $\{z : 0 < \Re(z) \leq 1, 0 < \Im(z) \leq T\}$. If we define $w(c)$ to be the (Euclidean) radius of the largest isometric circle in the fundamental polygon for $\sigma_c \Gamma \sigma_c^{-1}$ (draw a picture) then it also contains the box $\{z : 0 < \Re(z) \leq 1, w(c)^2 T^{-1} < \Im(z) \leq T\}$. We can also think of this number as

$$w(c)^{-1} := \min \left\{ c > 0 : \begin{pmatrix} * & * \\ c & * \end{pmatrix} \in \sigma_c \Gamma \sigma_c^{-1} \right\}.$$

This gives, at least for positive k,

$$\int_0^1 \int_{w(c)^2 T^{-1}}^T k(z, z + l) \frac{dxdy}{y^2} \leq \mathrm{Tr}^T K_{k,\gamma} \leq \int_0^1 \int_0^T k(z, z + l) \frac{dxdy}{y^2}. \tag{14.2}$$

The RHS of this equality gives, writing $k(z, w) = k(u(z, w))$,

$$\int_0^1 \int_0^T k(z, z + l) \frac{dxdy}{y^2} = \int_0^T k((\frac{l}{2y})^2) y^{-2} dy = \frac{1}{|l|} \int_{(l/2T)^2}^\infty k(u) u^{-1/2} du.$$

If we sum over all l to get the contribution from a cusp c, we find

$$\sum_{l \neq 0} \int_0^1 \int_0^T k(z, z + l) \frac{dxdy}{y^2} = 2 \sum_{l=1}^\infty \int_{(l/2T)^2}^\infty \frac{1}{l} k(u) u^{-1/2} du = 2 \int_{(1/2T)^2}^\infty k(u) u^{-1/2} \left(\sum_{1 \leq l \leq 2T\sqrt{u}} \frac{1}{l} \right) du.$$

This inner sum is

$$\sum_{1 \leq l \leq 2T\sqrt{u}} \frac{1}{l} = \log(2T\sqrt{u}) + \gamma + O(\frac{1}{T\sqrt{u}})$$

where γ is the Euler-Mascheroni constant. All together we get

$$\sum_{l \neq 0} \int_0^1 \int_0^T k(z, z + l) \frac{dxdy}{y^2} = 2 \int_0^\infty k(u) u^{-1/2} (\log(2T\sqrt{u}) + \gamma) du + O(T^{-1} \log T)$$

Now, this was the sum over l in the upper bound part of the inequality (14.2). The difference between the upper bound and the lower bound is obviously

$$\sum_{l \neq 0} \int_0^1 \int_0^{w(c)^2 T^{-1}} k(z, z + l) \frac{dxdy}{y^2} = 2 \sum_{l=1}^\infty \frac{1}{l} \int_{(lT/2w(l))^2}^\infty k(u) u^{-1/2} du = O(T^{-1})$$

and so asymptotically,

$$\sum_{l \neq 0} \mathrm{Tr}^T K_{k,\gamma_0^l} \sim 2 \int_0^\infty k(u) u^{-1/2} (\log(2T\sqrt{u}) + \gamma) du.$$

In the above, we are thinking of a fixed cusp c and a γ_0 generating Γ_c, which gives the contribution from parabolic conjugacy classes coming from c.

Call

$$L(T) := 2\int_0^\infty k(u)u^{-1/2}(\log(2T\sqrt{u}) + \gamma)du.$$

We eventually want to understand

$$\mathrm{Tr}^T K_{k,\mathrm{par}} = \sum_{c\in c(\Gamma)}\sum_{l\neq 0} \mathrm{Tr}^T K_{k,\gamma_0^l} \sim \#c(\Gamma)L(T)$$

But note

$$L(T) = g(0)(\log(2T) + \gamma) + \int_0^\infty k(u)u^{-1/2}\log u du.$$

where we have used that

$$\int_0^\infty k(u)u^{-1/2}du = q(0) = \frac{1}{2}g(0).$$

Observe that this already gives that

$$\mathrm{Tr}^T K_{k,\mathrm{par}}\ S_{\mathrm{par}} + R_{\mathrm{par}}\log T$$

where $R_{\mathrm{par}} = g(0)\#c(\Gamma) = R_{\mathrm{spec}}$ as desired.

Some further manipulations give the precise formula for the parabolic contribution to the trace formula, but this is not so important.

14.2. **A first application: Weyl's law.** So, modulo some calculus computations, we have shown the trace formula. Let's now apply it to get some sense of "how many" eigenfunctions there are. For a positive number X, let

$$N_\Gamma(X) = \#\{j : |t_j| \leq X\}.$$

This counts the number of discrete eigenforms of eigenvalue less than $\frac{1}{4} + X^2$ (plus some finite error which asymptotically does not matter coming from residual spectrum). We can also count "the contribution from the continuous spectrum analogously"–if we look at the trace formula, the corresponding term looks like

$$M_\Gamma(X) := \frac{1}{4\pi}\int_{-X}^{X} \frac{-\varphi'}{\varphi}(\frac{1}{2} + it)dt.$$

There is a beautiful asymptotic for these quantities.

Theorem 14.1 (Selberg)**.** *As $X \to \infty$, we have*

$$N_\Gamma(X) + M_\Gamma(X) \sim \frac{\mathrm{vol}(\Gamma\backslash\mathbb{H})}{4\pi}X^2.$$

We will show this next time.

15. Weyl's law and the existence of cusp forms

Let's continue describing our first application of the trace formula. This is the asymptotic count of eigenfunctions, also known as Weyl's law.

Definition 15.1. Let

$$N_\Gamma(X) := \#\{j : |t_j| \leq X\}$$

be the number of L^2-eigenforms of eigenvalue up to size $\frac{1}{4} + X^2$. Similarly, let

$$M_\Gamma(X) := \frac{1}{4\pi} \int_{-X}^{X} \frac{-\varphi'}{\varphi}(\frac{1}{2} + it)dt$$

be the "continuous" contribution to that count.

We have the following asymptotic.

Theorem 15.2.

$$N_\Gamma(X) + M_\Gamma(X) \sim \frac{\mathrm{vol}(\Gamma\backslash\mathbb{H})}{4\pi} X^2$$

Let's sort of informally convince ourselves of this fact. We will describe the full argument in a minute.

To see the asymptotic, we can input an obvious choice of test function into the trace formula. Let f be a smooth cutoff for $[-1, 1]$, i.e. which satisfied $0 \leq f(x) \leq 1$ for all x and with $f(x) = 0$ if $|x| \geq 1 + \varepsilon$ and $f(x) = 1$ for $|x| \leq 1$. We will choose ε later.

Then consider $h(t) = f(t/X)$. Plug this into the trace formula, and find that the "intertwining operator term" $h(0)/4 \operatorname{Tr}(\Phi(1/2))$ contributes only a constant in R. Meanwhile, the identity element offers the correct size. The hyperbolic elements contribute only a constant times R, the elliptic terms offer only a constant.

All that remains is to understand the main part of the parabolic contribution. This is a term of the form

$$-\#c(\Gamma)2\pi \int_{-\infty}^{\infty} h(r) \frac{\Gamma'}{\Gamma}(1 + ir)dr.$$

We will use that

$$\frac{\Gamma'}{\Gamma}(1 + it) + \frac{\Gamma'}{\Gamma}(1 - it) = \log(1 + t^2) + O((1 + t^2)^{-1})$$

which follows from

$$\frac{\Gamma'}{\Gamma}(s) = -\gamma - \sum_{n=0}^{\infty} \left(\frac{1}{n + s} - \frac{1}{n + 1} \right).$$

This formula is nothing more than the Weierstrass product for the Gamma function. All together, we find that

$$\int_{-\infty}^{\infty} h(r) \frac{\Gamma'}{\Gamma}(1 + ir)dr = \int_{X}^{X+X\varepsilon} h(r) \log(1 + r^2)dr + \int_{0}^{X} \log(1 + r^2)dr + O(1) = O(X \log X).$$

Thus, we know that the main term comes from the identity element, which os Weyl's law.

Here is a slightly different proof. One can show the auxiliary

Proposition 15.3. *For any $\delta > 0$,*

$$\sum_j e^{-\delta t_j^2} + \frac{1}{4\pi} \int_{-\infty}^{\infty} \frac{-\varphi'}{\varphi}(\frac{1}{2} + it) e^{-\delta t^2} dt = \frac{\mathrm{vol}(\Gamma\backslash\mathbb{H})}{4\pi} + \frac{h \log \delta}{4\sqrt{\pi\delta}} - \frac{\gamma h}{4\sqrt{\pi\delta}} + O(1).$$

Proof. (Sketch) This is very similar to what we explained above, but with the Gaussian test function. That is, apply the trace formula for $h(t) = e^{-\delta t^2}$ and hence $g(t) = \frac{1}{2\sqrt{\pi\delta}} e^{-t^2/4\delta})$. Note that the hyperbolic and elliptic terms contribute a bounded quantity, i.e. with no dependency on δ. The intertwining operator term $\frac{1}{4} h(0) \operatorname{Tr}(\Phi(1/2))$ is also always bounded. So it remains to understand the contribution from the identity and from the parabolic terms. The identity class gives

$$\frac{\mathrm{vol}(\Gamma\backslash\mathbb{H})}{4\pi} \int_{-\infty}^{\infty} e^{-\delta t^2} t \tanh(\pi t)dt = \frac{\mathrm{vol}(\Gamma\backslash\mathbb{H})}{4\pi} 2 \int_{0}^{\infty} e^{-\delta t^2} t \tanh(\pi t)dt = \frac{\mathrm{vol}(\Gamma\backslash\mathbb{H})}{4\pi\delta} + O(1)$$

since hyperbolic tangent looks a lot like 1 for large t. We can bound the parabolic classes as above, and with a bit more care, deduce the result. □

The Tauberian theorem allows us to conclude Weyl's law.

Remark 15.4. For some reason, this Tauberian theorem approach seems to be the way most proofs of Weyl's law are presented. I have no idea why–the smooth cutoff argument above seems better in almost every sense! My suspicion: the original proof of Weyl's law was through some heat kernel analysis, which roughly corresponds to this particular choice of test function in the proposition above.

15.1. **Existence of cusp forms.** Let's finally show what I said on day one would be our first application of the trace formula.

Theorem 15.5. *Let $\Gamma \subset \mathrm{PSL}_2(\mathbb{Z})$ be a congruence subgroup. Then Γ has infinitely many linearly independent cusp forms.*

The proof will essentially be an application of Weyl's law. What we will show is that

$$M_\Gamma(X) \ll_\varepsilon X^{1+\varepsilon}$$

which implies, since $N_\Gamma(X) + M_\Gamma(X) \sim \frac{\mathrm{vol}(\Gamma\backslash\mathbb{H}}{4\pi} X^2$, that $N_\Gamma(X) \sim \frac{\mathrm{vol}(\Gamma\backslash\mathbb{H}}{4\pi} X^2$, so in particular goes to ∞ as $X \to \infty$. Since there are only finitely many residual forms contributing to $N_\Gamma(X)$, we find that there must be infinitely many cusp forms.

Thus, we must control $\frac{\varphi'}{\varphi}(\frac{1}{2} + it)$. We claim that when Γ is congruence, we have

$$|\frac{\varphi'}{\varphi}(s)| \ll_\varepsilon |s|^\varepsilon.$$

This will be a corollary of the fact that φ must be a meromorphic function of order 1.

Recall this definition from complex analysis.

Definition 15.6. The *order* at ∞ of a function f meromorphic on $\mathbb{C}$ is

$$\inf\{\rho : |f(z)| \ll_\rho e^{|z|^\rho} \text{ as } z \to \infty\}.$$

There is the following easy lemma from complex analysis.

Lemma 15.7. *If f is meromorphic of order 1, then $\frac{f'}{f}$ satisfies*

$$|\frac{f'}{f}(s)| \ll_\varepsilon |s|^\varepsilon.$$

Proof. Suppose for simplicity that f is entire. Apply the Weierstrass factorization theorem to f, and take its logarithmic derivative. Since f is entire of order 1, we have that the sum of the reciprocals of the zeros converges absolutely, and the result follows. □

Thus, it remains to show that $\varphi(s)$ is of order 1 when Γ is congruence. This is where we crucially use the congruence condition.

Proposition 15.8. *Suppose Γ is a congruence subgroup. Then $\varphi(s)$ is a product of ratios of (completed) Dirichlet L-functions.*

Proof. To simplify the computation, let me only present the case $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$. I will say a few words about the general case in a minute.

First note that, via $\mathrm{SL}_2(\mathbb{Z})$ acting on the right on $\mathbb{Z}^2$, we have

$$\begin{aligned} \mathrm{N}(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{Z}) &\xrightarrow{\sim} \{(c,d) \in \mathbb{Z}^2 : \gcd(c,d) = 1\} \\ \begin{pmatrix} a & b \\ c & d \end{pmatrix} &\mapsto \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} c & d \end{pmatrix}. \end{aligned}$$

Thus, the Eisenstein series is given by

$$E(z,s) := \sum_{\gamma \in \mathrm{N}(\mathbb{Z})\backslash \mathrm{SL}_2(\mathbb{Z})/\{\pm 1\}} \Im(\gamma z)^s = \sum_{\substack{c \geq 0, d \in \mathbb{Z} \\ \gcd(c,d)=1}} \frac{y^s}{|cz+d|^2 s}.$$

Let us compute its constant term to see

$$\int_0^1 E(x+iy,s)dx = y^s + \int_0^1 \frac{y^s}{|x+iy|^{2s}}dx + \sum_{c=1}^{\infty} \sum_{\substack{d\neq 0 \\ \gcd(c,d)=1}} \int_0^1 \frac{y^s}{|cx+icy+d|^{2s}}dx$$

Let's compute the latter sum from $c=2$ to ∞. This is

$$\begin{aligned}
\sum_{c=2}^{\infty} \sum_{\substack{d\neq 0 \\ \gcd(c,d)=1}} \int_0^1 \frac{y^s}{|cx+icy+d|^{2s}}dx &= y^s \sum_{c=2}^{\infty} \sum_{l\in(\mathbb{Z}/c)^\times} \sum_{k\in\mathbb{Z}} \int_0^1 \frac{dx}{|cx+ck+icy+l|^{2s}} \\
&= y^s \sum_{c=2}^{\infty} \sum_{l\in(\mathbb{Z}/c)^\times} \int_{-\infty}^{\infty} \frac{dx}{|cx+icy+l|^{2s}} \\
&= y^s \sum_{c=2}^{\infty} \sum_{l\in(\mathbb{Z}/c)^\times} \frac{1}{c^{2s}} \int_{-\infty}^{\infty} \frac{dx}{|x+iy+\frac{l}{c}|^{2s}} \\
&= y^s \sum_{c=2}^{\infty} \frac{\varphi(c)}{c^{2s}} \int_{-\infty}^{\infty} \frac{dx}{|x+iy|^{2s}}
\end{aligned}$$

where $\varphi(c)$–unfortunate notation–is not the determinant of the scattering matrix but rather the Euler totient function.[7]

We can now use, if we take as convention $\varphi(1)=1$,

$$y^s \int_{-\infty}^{\infty} \frac{dx}{|x+iy|^{2s}} = y^{1-s}\pi^{\frac{1}{2}} \frac{\Gamma(s-\frac{1}{2})}{\Gamma(s)}$$

Finally, we get

$$\begin{aligned}
\int_0^1 E(x+iy,s)dx &= y^s + \int_0^1 \frac{y^s}{|x+iy|^{2s}}dx + y^s \sum_{d\neq 0} \int_0^1 \frac{dx}{|x+iy+d|^{2s}} + y^s \sum_{c=2}^{\infty} \frac{\varphi(c)}{c^{2s}} \int_{-\infty}^{\infty} \frac{dx}{|x+iy|^{2s}} \\
&= y^s + y^{1-s} \frac{\pi^{-\frac{2s-1}{2}}}{\pi^{-\frac{2s}{2}}} \frac{\Gamma(s-\frac{1}{2})}{\Gamma(s)} \sum_{c=1}^{\infty} \frac{\varphi(c)}{c^{2s}}
\end{aligned}$$

and noting that

$$\sum_{c=1}^{\infty} \frac{\varphi(c)}{c^{2s}} = \prod_p \sum_{k=0}^{\infty} \frac{\varphi(p^k)}{p^{2ks}} = \prod_p \sum_{k=0}^{\infty} \frac{p^{k-1}(p-1)}{p^{2ks}} = \frac{\zeta(2s-1)}{\zeta(2s)}$$

we find

$$\int_0^1 E(x+iy,s)dx = y^s + \frac{\Lambda(2s-1)}{\Lambda(2s)} y^{1-s}$$

where $\Lambda(s) = \pi^{-s/2}\Gamma(s/2)\zeta(s)$ is the completed zeta function.

For general congruence groups, one has to the entries of the scattering matrix in terms of a double coset decomposition, where each double coset contributes a certain coprimality condition combined with a congruence condition. The computation itself is not particularly enlightening! □

[7]Note too we have been a little sloppy, identifying $l \in (\mathbb{Z}/c)$ with $l = 0, 1, ..., c-1$.

16. An application of the theory of Eisenstein series

Today will be a bit of a mixed bag. Next time, I will give one of the most beautiful applications of the classical trace formula, namely the prime geodesic theorem. This is essentially the statement of the prime number theorem, but for closed geodesics in a Riemann surface.

To gain some appreciation for this type of result, I want to spend some time today to explain the classical prime number theorem. Amusingly enough, this can be seen as an application of the theory of Eisenstein series that we have developed thus far in the course.

Thus, our goal today is the following theorem.

Theorem 16.1 (The prime number theorem; Hadamard, de la Vallée-Poussin)**.** *Let* $\pi(x) = \#\{p \text{ prime} : p \leq x\}$*. Then*

$$\pi(x) \sim \frac{x}{\log x}$$

To prove this, we need to introduce some analytic trickery which will prove very useful for the prime geodesic theorem as well. Really, we have already seen this idea in some guise before.

16.1. **Smoothing sums.** Suppose that $a(n)$ is some complex valued function. Many problems in analytic number theory have to do with understanding asymptotics or bounds on sums

$$\sum_{n \leq x} a(n).$$

For instance, the Gauss circle problem, the prime number theorem, Dirichlet's theorem, and Vinogradov's theorem on odd Goldbach all have this flavor.

The smoothing sums trick is an analytic tool for reducing such problems to an understanding of the complex analytic behavior of the associated Dirichlet series

$$D(s) := \sum_{n=1}^{\infty} \frac{a(n)}{n^s}.$$

This is particularly powerful when $a(n)$ is (completely) multiplicative, as then this Dirichlet series will have an Euler product

$$D(s) = \prod_p (1 - a(p)p^{-s})^{-1})$$

which can frequently be useful.

The smoothing sums game works as follows. Pick a smooth even function $f(t)$ such that

- $0 \leq f(t) \leq 1$
- $$f(t) = \begin{cases} 1 & \text{if } t \leq 1 \\ 0 & \text{if } t \geq 1 + \varepsilon \end{cases}$$

We pick ε later. Note that

$$\sum_{n \leq x} a(n) \sim \sum_n a(n) f(\frac{n}{x})$$

and that if we write using Mellin inversion

$$f(\frac{n}{x}) = \frac{1}{2\pi i} \int_{\Re(s)=\sigma} \widetilde{f}(s) |\frac{n}{x}|^{-s} ds$$

we get

$$\sum_n a(n) f(\frac{n}{x}) = \frac{1}{2\pi i} \int_{\Re(s)=\sigma} D(s) \widetilde{f}(s) x^s ds$$

Now, $\widetilde{f}(s)$ is meromorphic, with its only pole a simple pole at $s = 0$ of residue 1. It also has rapid decay on vertical lines. So if we know the complex analytic properties of $D(s)$, including its growth on vertical lines, then we can push the contour to the left to gain saving in x of the smoothed sum.

16.2. **The prime number theorem.** Let's now reduce the prime number theorem to a problem we can solve easily. The first step is to show that

$$\pi(x) \sim \frac{x}{\log x} \text{ if } \psi(x) \sim x \tag{16.1}$$

where

$$\psi(x) = \sum_{n \leq x} \Lambda(n)$$

where

$$\Lambda(n) = \begin{cases} \log p & \text{if } n = p^k, k \geq 1 \\ 0 & \text{else} \end{cases}$$

is the von-Mangoldt function.

Actually, (16.1) is really an equivalence-but we don't need that. It is easy to see. It follows simply from summation by parts!

$$\begin{aligned} \pi(x) &= \sum_{p \leq x} 1 \\ &= \sum_{2 \leq n \leq x} \frac{\Lambda(n)}{\log n} - \sum_{\substack{n = p^k \leq x \\ k \geq 2}} \frac{\Lambda(n)}{\log n} \\ &= \frac{\psi(x)}{\log x} + o(\frac{x}{\log x}) \end{aligned}$$

So we have only to show that $\psi(x) \sim x$. This is exactly what the smoothing sums trick can help us with.

Proposition 16.2. $\psi(x) \sim x$.

By the smoothing sums trick, we have to only prove

$$\lim_{\varepsilon \to 0} \psi_f(x) \sim x$$

where f is our smoothing function and

$$\psi_f(x) := \sum_n \Lambda(n) f(\frac{n}{x}).$$

But note that the Dirichlet series associated to the arithmetic function $\Lambda(n)$ is simply

$$\frac{-\zeta'}{\zeta}(s) = \sum_{n=1}^{\infty} \frac{\Lambda(n)}{n^s}$$

and so

$$\sum_n \Lambda(n) f(\frac{n}{x}) = \frac{1}{2\pi i} \int_{\Re(s) = \sigma} \frac{-\zeta'}{\zeta}(s) \widetilde{f}(s) x^s ds$$

for any $\sigma > 1$. Let's play the game of pushing this contour as far to the right as we can. This seems hard, since every pole and zero of $\zeta(s)$ will contribute. However, we can push the contour to something a little bit to the left of $\Re(s) = 1$. (Draw picture of an appropriate new contour C avoiding the pole and running up to $\Re(s) = 1$.) This picks up a contribution from the pole at $s = 1$, which gives

$$\sum_n \Lambda(n) f(\frac{n}{x}) = \widetilde{f}(1) x + \frac{1}{2\pi i} \int_C \frac{-\zeta'}{\zeta}(s) \widetilde{f}(s) x^s ds$$

for our contour C. Of course, this works only if there are no zeros on the line $\Re(s) = 1$, so that we pick up no further contributions.

The integral over C that we have left over can quickly be seen to contribute $o(x)$–this follows from decay in vertical strips of $\widetilde{f}(s)$ and similar control for the logarithmic derivative of ζ (simply use that the function is meromorphic of order 1). Thus, it remains to conclude that $\zeta(s)$ has no zeros on $\Re(s) = 1$.

Lemma 16.3. *$\zeta(s)$ has no zeros on $\Re(s) = 1$.*

Proof. We simply use the analytic properties of the Eisenstein series for $\mathrm{PSL}_2(\mathbb{Z})$.

We computed last time that

$$\int_0^1 E(x+iy,s)dx = y^s + \frac{\Lambda(2s-1)}{\Lambda(2s)} y^{1-s}.$$

We can run a similar computation for the higher order Fourier coefficients. This gives

$$\int_0^1 E(x+iy,s)e(-mx)dx = \frac{4\sigma_{s-\frac{1}{2}}(m)}{\Lambda(2s)} y^{1/2} K_{s-\frac{1}{2}}(2\pi m y)$$

where

$$\sigma_r(m) = \sum_{d|m} d^r$$

is a divisor sum.

Now we know that the poles of $E(z,s)$ are contained in the poles of $\varphi(s) = \frac{\Lambda(2s-1)}{\Lambda(2s)}$ and that $|\varphi(\frac{1}{2}+it)| = 1$ hence, $E(z,s)$ has no poles on $\Re(s) = \frac{1}{2}$. But then by the Fourier expansion this implies that $\Lambda(s)$ has no zeros on this line, hence that $\zeta(s)$ has no zeros on this line. □

Remark 16.4. What number theorists often call "the prime number theorem" for an L-function is simply a non-vanishing result $L(\frac{1}{2}+it,\pi) \neq 0$ for $t \in \mathbb{R}$. If you are looking for such a non-vanishing result for more exotic π, the trick above is pretty much the only game in town.

Namely, for every π such that $L(s,\pi)$ appears in the constant term of an Eisenstein series, we can run the same sort of argument as above to get a non-vanishing result, *and this is the only game in town.* It would be extremely interesting to find a more general approach!

17. The prime geodesic theorem

Last time we proved the prime number theorem. Actually, we got slightly more than that–if we took the smoothed sum

$$\psi_f(x) = \frac{1}{2\pi i}\int_{\Re(s)=\sigma} \widetilde{f}(s)\frac{-\zeta'}{\zeta}(s)x^s ds = \sum_{n=1}^{\infty}\Lambda(n)f(\frac{n}{x})$$

we could shift this contour as far to the left as would like and pick up an explicit formula relating primes and zeros of ζ.

$$\psi_f(x) = \sum_p\sum_{k=1}^{\infty}\log(p)f(\frac{p^k}{x}) = \widetilde{f}(1)x - \sum_{s_j}(\widetilde{f}(s_j)x_j^s + \widetilde{f}(1-s_j)x^{1-s_j}) + \text{junk}$$

where the sum is over zeros of ζ lying in the critical strip and with $\Re(s) \geq 1/2$, and satisfying $\Im(s) > 0$ if $\Re(s) = \frac{1}{2}$ (it is known that $\zeta(\frac{1}{2}) \neq 0$, e.g. by numerical calculations). The term "junk" corresponds to contributions coming from the trivial zeros.

We can look at similar contributions appearing the Selberg trace formula. Note that Mellin transform (e.g. of our smoothing function f) is really just Fourier transform after a change of variables. It is therefore tempting to think of the hyperbolic contribution below (think $\widetilde{f} \approx \widehat{g} = h$)

$$\begin{aligned}\sum_j h(t_j) + \frac{1}{4\pi}\int_{-\infty}^{\infty} h(r)\frac{-\varphi'}{\varphi}(\frac{1}{2}+ir)dr + \frac{h(0)}{4}\operatorname{Tr}(\Phi(\frac{1}{2}))& \\ = \frac{\operatorname{vol}(\Gamma\backslash\mathbb{H})}{4\pi}\int_{-\infty}^{\infty} h(r)r\tanh(\pi r)dr& \\ + \sum_P\sum_{l=1}^{\infty}\frac{\log \mathrm{N}(P)}{\mathrm{N}(P)^{l/2} - \mathrm{N}(P)^{-l/2}}g(l\log \mathrm{N}(P))& \\ + \sum_R\sum_{0<l<m}\frac{1}{2m\sin\frac{\pi l}{m}}\int_{-\infty}^{\infty} h(r)\frac{\cosh(\pi(1-2l/m)r)}{\cosh(\pi r)}dr& \\ + \#c(\Gamma)\left(\frac{h(0)}{4} - g(0)\log 2 - 2\pi\int_{-\infty}^{\infty} h(r)\frac{\Gamma'}{\Gamma}(1+ir)dr\right).&\end{aligned}$$

as one side of an explicit formula, namely the ψ part. The spectral side of the trace formula should correspond to the RHS of a this explicit formula.

This motivates the following definition.

Definition 17.1. The Selberg zeta function $Z_\Gamma(s)$ is defined by

$$Z_\Gamma(s) = \prod_P\prod_{k=0}^{\infty}(1 - \mathrm{N}\,P^{-s-k})$$

In the definition, and indeed in the trace formula above, P runs over primitive hyperbolic conjugacy classes in Γ. It is clear that this set is the same as the set of closed geodesics on $\Gamma\backslash\mathbb{H}$ which are primitive in the sense that as maps from $[0,1) \to \Gamma\backslash\mathbb{H}$ they are injective.

The raison d'être behind $Z_\Gamma(s)$ lies in the following easy computation.

Lemma 17.2. *Let $\Re(s) > 1$. Then*

$$\frac{Z_\Gamma'}{Z_\Gamma}(s) = \sum_P\sum_{l=1}^{\infty}\frac{\log \mathrm{N}\,P}{\mathrm{N}\,P^{ls} - \mathrm{N}\,P^{-ls}}$$

Proof. Take a log and differentiate. □

It is worth noting that this logarithmic derivative can be written as a Dirichlet series of sorts. Namely, if we expand out the denominator in the above expression, we find

$$\begin{aligned} -\frac{Z_\Gamma'}{Z_\Gamma}(s) &= \sum_P \sum_{l=1}^\infty \sum_{k=1}^\infty \frac{\log \mathrm{N}\, P}{\mathrm{N}\, P^{l(2k-1)s}} \\ &= \sum_n \frac{\Lambda_\Gamma(n)}{n^s} \end{aligned}$$

where the sum over n is over the multiplicative submonoid of $\mathbb{R}_{>0}$ generated by all $\mathrm{N}\, P$, which is a discrete set in $\mathbb{R}_{>0}$, and where

$$\Lambda_\Gamma(n) = \begin{cases} 2^{\nu^{\mathrm{odd}}(r)} \log \mathrm{N}\, P & \text{if } n = \mathrm{N}\, P^r \\ 0 & \text{else} \end{cases}.$$

In the above, for $r \in \mathbb{N}$, if we write r in terms of its prime factorization

$$r = \prod_p p^{e(p)} = 2^{e(2)} \prod_{p \text{ odd}} p^{e(p)}$$

$$\nu^{\mathrm{odd}}(r) = \sum_{p \text{ odd}} e(p)$$

is number of odd prime factors of r counted with multiplicity.

Now, the trace formula allows us to access information about this. We can take a particular test function and input it into the trace formula. Set, for some $\alpha, \beta \in \mathbb{C}$,

$$h(t) = \frac{1}{t^2+\alpha^2} - \frac{1}{t^2+\beta^2}.$$

Now, this is not Schwartz, but it has sufficient decay that we can put it into the trace formula (although this requires some careful analysis of the terms, which I am too lazy to do). Computing the Fourier transform gives

$$g(r) = \frac{1}{2\alpha} e^{-\alpha r} - \frac{1}{2\beta} e^{-\beta r}$$

Set $\alpha = s - \frac{1}{2}$ and $\beta = b - \frac{1}{2}$, for some fixed b, and look at the hyperbolic contribution to the trace formula. Some easy but tedious computation shows that this gives

$$\frac{1}{2s-1}\frac{Z_\Gamma'}{Z_\Gamma}(s) - \frac{1}{2b-1}\frac{Z_\Gamma'}{Z_\Gamma}(b).$$

We can exploit this to get analytic continuation properties for Z_Γ through its logarithmic derivative. If we stare at the remainder of the trace formula e.g. at the discrete spectral contribution

$$\sum_j \frac{1}{t_j^2 + (s-\frac{1}{2})^2} - \frac{1}{t_j^2 + (b-\frac{1}{2})^2}$$

we can read off meromorphicity! There are poles corresponding to $s_j = \frac{1}{2} + it)j$ of multiplicity equal to the dimension of the corresponding eigenspace. The continuous contribution is absolutely convergent, so does not contribute any poles, and the intertwining operator contribution is also fine as a function of s except at $s = 1/2$.

One can combine the contributions from the parabolic, elliptic, identity motion, and intertwining operators to pick out the remaining poles and their orders. In total, this analysis yields

Theorem 17.3. *The Selberg zeta function $Z(s) = Z_\Gamma(s)$ admits meromorphic continuation to all of $\mathbb{C}$. It has the following properties:*

1. *In $\Re(s) \geq 1/2$, it has zeros at the points s_j and $\overline{s_j}$, where $s_j(1-s_j)$ is a discrete eigenvalue of Δ. If $s_j \neq 1/2$, the zero is of order equal to the dimension of the eigenspaces.*
2. *At $s = 1/2$, $Z(s)$ has a zero or pole equal to twice the dimension of $1/4$ eigenspace minus the number of inequivalent cusps.*

(3) *$Z(s)$ is a meromorphic function of Hadamard order 2. This means that on vertical strips $\sigma_0 \leq x \leq \sigma_1$, we have*

$$\frac{Z'}{Z}(x+iy) \ll_\varepsilon |y|^{1+\varepsilon}$$

as $|y| \to \infty$.

If the hand-waviness above made you unhappy (I did skip a number of technical calculations!) think only of the case when $\Gamma\backslash\mathbb{H}$ is a compact Riemann surface. This corresponds to there being no parabolic contribution, no elliptic contribution, so all that is needed is an analysis of the identity term, which is easy.

In any case, now that we know some of the analytic properties of $Z_\Gamma(s)$, we can get for free the prime geodesic theorem. In fact, we can do much better–since we know the location of the zeros of $Z(s)$, and we know that they must lie, with only finitely many exceptions coming from residual spectrum, on the line $\Re(s) = 1/2$, we in fact know the "Riemann hypothesis" for this zeta function.

Let's apply the smoothed sums argument now to the sum

$$\psi_\Gamma(x) = \sum_{n \leq x} \Lambda_\Gamma(n).$$

It is worth noting that this sum is not particularly different from

$$\sum_{\{P : \mathrm{N}\, P \leq x\}} \log \mathrm{N}\, P.$$

The smoothed version of this sum is

$$\psi_{\Gamma,f}(x) = \sum_n \Lambda_\Gamma(n) f\left(\frac{n}{x}\right) = \frac{1}{2\pi i} \int_{\Re s = \sigma} \widetilde{f}(s) \frac{-Z'}{Z}(s) x^s ds$$

And shifting the contour as far to the left as we can, we find

$$\sum_{\{P : \mathrm{N}\, P \leq x\}} \log \mathrm{N}\, P = \sum_{1/2 < s_j < 1} s_j^{-1} x^{s_j} + O(x^{3/4})$$

where here the error term is not $1/2$ because Z is of order 2. (Equivalently, apply the Weyl law to see the spacing of zeros of $Z(s)$ and see what that gets you in the smoothed sum).

18. Adelic theory: motivations

It is time to move away from the classical language of locally symmetric spaces and towards the more modern language of reductive groups over the adeles. The starting point here is the following simple observation: given an automorphic form f (of weight 0)[8] on $\mathbb{H}$ for a congruence group Γ, there is a corresponding representation of $\mathrm{SL}_2(\mathbb{R})$. This arises by looking at the space V_f of (finite linear combinations) of g translates of f for $g \in \mathrm{SL}_2(\mathbb{R})$. V_f is typically infinite dimensional and irreducible.

It's natural to ask which representations π_∞ of $\mathrm{SL}_2(\mathbb{R})$ can occur in this manner. There is an easy invariant of such representation–the center of the universal enveloping algebra $\mathfrak{Z}(\mathfrak{g})$ acts on any smooth representation, hence by a character on any irreducible smooth representation. For our $\mathrm{G} = \mathrm{SL}_2(\mathbb{G})$, $\mathfrak{Z}(\mathfrak{g})$ is generated by one element and this element acts on V_f acts by the Laplacian.

That is, assuming f is an eigenform for Δ, we can tell the representations V_f apart crudely by looking at the eigenvalue λ of Δ on f.

Selberg conjectured that such eigenvalues were bounded below.

Conjecture 18.1 (Selberg)**.** *Let f be a cusp form on $\mathbb{H}$ for a congruence subgroup, and suppose that f is a Laplace eigenform of eigenvalue λ Then*

$$\lambda \geq 1/4.$$

Remark 18.2. Note that this is bounding below the smallest non-zero eigenvalue of Δ, since congruence subgroups have no residual spectrum except that coming from the constant function (check location of poles of Eisenstein series using $\varphi(s)$ which is a ratio of Dirichlet L-functions).

Remark 18.3. It's worth noting too that this conjecture is very false for non-congruence subgroups. One can construct non-congruence Γ for which the smallest non-zero eigenvalue goes to 0.

So we expect to see only very particular representations π_∞ of $\mathrm{SL}_2(\mathbb{R})$ which occur in this manner for congruence subgroups.

But from the point of view of our discussion of the trace formula from the perspective of locally symmetric spaces, there is nothing particularly special about congruence subgroups anywhere! That is, pretty much anything we can prove with the TF as we have done it should be true for all Γ, not just congruence groups.

This is our next step: *to build the condition of Γ being congruence into the setup of the trace formula.*

This will have some lovely consequences.

(1) First, if we develop such a thing, we can explain the existence of some cusp forms in terms quaternion algebras, like I sketched on the first day. This is the Jacquet-Langlands correspondence.
(2) Second, we can begin to see the more refined structure in the classification of automorphic forms on GL_2 [9], or for a general reductive group G. This is most often described as the phenomenon of *endoscopy*, which is itself a more specific instance of Langlands' *principle of functoriality*. While the theory of endoscopy is now very developed and quite well understood, the story for functoriality is very far from complete. Some beautiful problems come out of this dream.

So let's begin.

18.1. Lifting Maass forms to GL_2.

It is worth expanding our definition of Maass form slightly. This will allow us to include classical holomorphic (and anti-holomorphic) modular forms of Hecke.

Definition 18.4. A Maass form f of weight k, level $\Gamma_0(N)$ (for Γ a congruence group), and nebentypus χ is a smooth function $f : \mathbb{H} \to \mathbb{C}$ such that

(1) We have

$$f(\gamma z) = \chi(d) \left(\frac{cz+d}{|cz+d|} \right)^k f(z)$$

for all

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma.$$

Here χ is a Dirichlet character modulo N (our nebentypus).

[8] Thus far, with the exception of the first lecture, I have never spoken about any higher weight automorphic forms. I will define them later on today.

[9] I shifted from SL_2 to GL_2 but for many practical purposes this is inconsequential.

(2) f is moderate growth at all cusps.
(3) f is a finite linear combination of eigenfunctions for Δ_k, where

$$\Delta_k = -y^2\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) + iky\frac{\partial}{\partial x}$$

Remark 18.5. Given a holomorphic modular form F of weight k, level $\Gamma_0(N)$, and nebetypus, we can attach a Maass form by taking

$$f(z) = y^{k/2}F(z).$$

Thus this very general definition of Maass form encompasses both the non-holomorphic eigenfunction we have been discussing and the classical holomorphic modular forms of Hecke.

Let us compare and contrast this with a modern definition. Let $\mathbb{A} = \mathbb{A}_{\mathbb{Q}}$ be the ring of adeles of $\mathbb{Q}$.

Definition 18.6. An *adelic automorphic form* of $\mathrm{GL}_2(\mathbb{A})$ is a function $\phi : \mathrm{GL}_2(\mathbb{A}) \to \mathbb{C}$ such that

(1) ϕ is *smooth.* This means that for every $g_0 \in \mathrm{GL}_2(\mathbb{A})$, there exists an open subset $U \subset \mathrm{GL}_2(\mathbb{A})$ and smooth function ϕ_∞^U on $\mathrm{GL}_2(\mathbb{R})$ such that $\phi(g) = \phi_\infty^U(g_\infty)$.
(2) ϕ is *moderate growth.* This means that it bounded by a polynomial in $|g|$, where

$$|g| := \prod_v \max\left\{|a|_v, |b|_v, |c|_v, |d|_v, |\det|_v^{-1}\right\}.$$

(3) ϕ is K-*finite,* where $\mathrm{K} = \prod_{v<\infty}\mathrm{GL}_2(\mathbb{Z}_v) \times \mathrm{O}_2(\mathbb{R})$. This means that the space of K translates of ϕ is finite dimensional.
(4) ϕ is $\mathfrak{Z}(\mathfrak{g})$-*finite.* This mean that if we act on ϕ by invariant differential operators at the infinite place, that $D\phi$ all lie in a finite dimensional space.

We can lift Maass forms on the upper half plane to adelic automorphic forms ϕ. I will explain this in the simplest case.

First, a warmup: how do we lift Dirichlet characters to characters of the idele class group? This is a useful trick in CFT. The idea is easy: every Dirichlet character

$$\chi : (\mathbb{Z}/N)^\times \to \mathbb{C}^\times$$

can be viewed as a character of $\widehat{\mathbb{Z}}^\times \cong \prod_p \mathbb{Z}_p^\times$ via

$$\widehat{\mathbb{Z}} \to \mathbb{Z}/N.$$

But if we consider the isomorphism of the idele class group

$$\mathbb{Q}^\times\backslash\mathbb{A}^\times \cong \left(\prod_p \mathbb{Z}_p^\times\right) \times \mathbb{R}_{>0}$$

(this isomorphism is stupid: to get an inverse to the obvious map right to left, simply move all denominators to ∞) then we can lift χ to a character of $\mathbb{Q}^\times\backslash\mathbb{A}^\times$ by simply ignoring the Archimedean factor.

So how do we proceed for Maass forms? The idea is identical. For simplicity, let's assume that $\Gamma = \Gamma(1)$ is the full modular group, that $k = 0$, and that $\chi = 1$. Assume too that $f(x+iy) = f(-x+iy)$, i.e. f is an even function $f : \mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H} \to \mathbb{C}$. We can consider the adelic quotient

$$\mathrm{GL}_2(\mathbb{Q})\backslash\mathrm{GL}_2(\mathbb{A}) = \left(\prod_p \mathrm{GL}_2(\mathbb{Z}_p)\right) \times D_\infty$$

where D_∞ is a fundamental domain for $\mathrm{GL}_2(\mathbb{Z})\backslash\mathrm{GL}_2(\mathbb{R})$. f can viewed as a function on D_∞, and hence on the quotient.

To lift more general forms, we have to be a little careful, but this is not hard–k corresponds to how $\mathrm{SO}_2(\mathbb{R})$ should act upstairs, while level corresponds to invariance on the right by the appropriate subgroup of the finite adeles. The nebentypus corresponds to central character. The basic point is the *strong approximation property:* for every $N = \prod p_i^{e_i}$, we can write

$$\mathrm{GL}_2(\mathbb{A}) = \mathrm{GL}_2(\mathbb{Q})\,\mathrm{GL}_2(\mathbb{R})\prod_p \mathrm{K}_0(p_i^e)$$

where

$$\mathrm{K}_0(p^{e_i}) := \left\{ g_p \in \mathrm{GL}_2(\mathbb{Z}_p) : g_p \equiv \begin{pmatrix} * & * \\ 0 & * \end{pmatrix} \bmod p^{e_i} \right\}$$

Thus this decomposition gives

$$\mathrm{GL}_2(\mathbb{Q}) \backslash \mathrm{GL}_2(\mathbb{A}) = \left(\prod_p] K_0(p^{e_i}) \right) \times D_\infty(N)$$

where $D_\infty(N)$ is a fundamental domain for $\mathrm{GL}_2(\mathbb{Q}) \cap \prod_p \mathrm{K}_0(p^{e_i})$ acting on $\mathrm{GL}_2(\mathbb{R})$ on the left.

So we can interpret all of our automorphic forms now as adelic objects.

19. Automorphic representations

Last time, we began the passage from the world of classical Maass forms to the world of automorphic representations of GL_2. The key observation was the following simple lifting procedure:

$$\{\text{functions on } \mathrm{PSL}_2(\mathbb{Z})\backslash\mathbb{H}\} \rightsquigarrow \{\text{functions on } \mathrm{GL}_2(\mathbb{Z})\backslash \mathrm{GL}_2(\mathbb{R})\} \rightsquigarrow \{\text{functions on } \mathrm{GL}_2(\mathbb{Q})\backslash \mathrm{GL}_2(\mathbb{A})\}$$

and its versions for higher level and weight. Since classical Maass forms have various restrictions built into their definition (e.g. moderate growth, finiteness under differential operators) we arrive at the following definition. This is designed to encompass all functions on $\mathrm{GL}_2(\mathbb{Q})\backslash \mathrm{GL}_2(\mathbb{A})$ arising from classical Maass forms via lifting.

Definition 19.1. An *automorphic form* on $\mathrm{GL}_2(\mathbb{A})$ is a function $\phi : \mathrm{GL}_2(\mathbb{A}) \to \mathbb{C}$ such that

(1) ϕ is *smooth.* (Roughly, uniformly locally constant times smooth at ∞.)
(2) ϕ is *moderate growth.*
(3) ϕ is K-*finite,* where $\mathrm{K} = \prod_{v<\infty} \mathrm{GL}_2(\mathbb{Z}_v) \times \mathrm{O}_2(\mathbb{R})$.
(4) ϕ_∞ is $\mathfrak{Z}(\mathfrak{g})$-*finite.*

We write $A(\mathrm{GL}_2)$ for the space of automorphic forms on GL_2.

Definition 19.2. If $\phi \in A(\mathrm{GL}_2)$ is an automorphic form which in addition satisfies the condition

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

for all g, then we say ϕ is a *cusp form.* We write $A_0(\mathrm{GL}_2)$ for the space of cusp forms.

Remark 19.3. Fix a character $\omega : \mathbb{Q}^\times\backslash\mathbb{A}^\times \to \mathbb{C}^\times$. We can look at the spaces $A(\mathrm{GL}_2, \omega)$ and $A_0(\mathrm{GL}_2, \omega)$ of automorphic forms which have central character ω, i.e. the subspace of forms ϕ satisfying

$$\phi(zg) = \omega(z)\phi(g)$$

where

$$z = \begin{pmatrix} z & 0 \\ 0 & z \end{pmatrix}.$$

If ω is unitary, then one can show that $A_0(\mathrm{GL}_2, \omega) \subset L^2(\mathrm{GL}_2(\mathbb{Q})\backslash \mathrm{GL}_2(\mathbb{A}))$. (Roughly, this is because cusp forms have to be rapidly decaying, just as in the classical case.)

Example 19.4. We have already seen that every classical Maass form gives rise to an automorphic form. But let's explain a construction that starts and ends in the adelic world, without ever working classically. EISENSTEIN SERIES.

So far, this seems quite unmotivated. The passage to adelic automorphic forms has, however, one great benefit–we can rephrase many properties of Maass forms in terms of representation theory.

To a given automorphic form ϕ, we can attach an infinite dimensional representation $\pi = \pi_\phi$ of $\mathrm{GL}_(\mathbb{A})$ by

$$\pi := R(\mathrm{GL}_2(\mathbb{A})).\phi$$

as the space of finite linear combinations of right translates of ϕ.

Remark 19.5. There is an annoying subtle point at play here. There are vectors in $R(\mathrm{GL}_2(\mathbb{A})).\phi$ which are not K-finite! That is, the representation we attached is not a subspace of $A(\mathrm{GL}_2)$. We could restrict to those K-finite vectors, but then this is not a representation of $\mathrm{GL}_2(\mathbb{A})$ since $\mathrm{GL}_2(\mathbb{R})$ will not preserve the K-finiteness–indeed, $A(\mathrm{GL}_2)$ is not a representation of $\mathrm{GL}_2(\mathbb{A})$. A common solution is to make this restriction anyways, and replace the representation at infinity with the notion of a $(\mathfrak{g}, \mathrm{K})$-module.

Also bad: the representation of $\mathrm{GL}_2(\mathbb{R})$ at ∞ is not on a complete topological vector space (such as a Hilbert space) since we have not completed.

In total it is probably better to define π as a certain "moderate growth" Fréchet completion of the space we described above, or of the K-fixed vectors therein. The miracle is that this procedure can be done canonically in such a way that it only depends on the underlying $(\mathfrak{g}, \mathrm{K})$-module structure. This is the remarkable work of Casselman and Wallach.

In any case, let us elide this point. The construction $\phi \mapsto \pi_\phi$ motivates the following definition.

Definition 19.6. An *automorphic representation* π is an irreducible representation of $\mathrm{GL}_2(\mathbb{A})$ which appears as a subquotient of (the moderate growth Fréchet completion of) the space of automorphic forms.

Since we often work with classical cusp forms, we need a version of the cuspidality condition for adelic automorphic forms. For GL_2 this is

Definition 19.7. An automorphic form ϕ is *cuspidal* if

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

for all $g \in \mathrm{GL}_2(\mathbb{A})$. We write $A_0(\mathrm{GL}_2)$ for the space of cusp forms, and say that an automorphic representation π is cuspidal if it appears as a subquotient of (the completion of) $A_0(\mathrm{GL}_2)$.

The automorphic condition is where all of the number theory lives. Without it, we can define the following weaker notion.

Definition 19.8. An *admissible representation* π of $\mathrm{GL}_2(\mathbb{A})$ is a smooth representation such that every irredicible constituent in the restriction to the maximal compact subgroup has finite multiplicity.

By virtue of the K-finiteness and $\mathfrak{Z}(\mathfrak{g})$ finiteness, automorphic representations are all admissible (see Borel-Jacquet).

Every irreducible admissible representation factors as a tensor product $\pi = \otimes' \pi_v$, where v runs over all places of $\mathbb{Q}$, where each π_v is an admissible representation of $\mathrm{GL}_2(\mathbb{Q}_v)$. This is a fundamental but not so unexpected result of Flath[10] However, a word needs to be said about the $'$ in the tensor product. Implicit in the expression

$$\pi = \otimes'_v \pi_v$$

is the statement that for almost all v, the irreducible π_v has a $\mathrm{K}_v = \mathrm{GL}_2(\mathbb{Z}_v)$-fixed vector. It is easily shown that for an irreducible local representation, such a K_v-fixed vector must be unique up to scaling, and the $'$ means that in every pure tensor, at almost all places we must take the K_v-fixed vector. Those smooth irreducible representations of $\mathrm{GL}_2(\mathbb{Q}_p)$ with a K_p fixed vector are called *spherical*. It is easy to classify all of them.

Theorem 19.9. *Every spherical representation π_p of $\mathrm{GL}_2(\mathbb{Q}_\mathbb{p})$ is of the form*

$$\pi_p \cong \mathrm{Ind}_{\mathrm{B}(\mathbb{A})}^{\mathrm{GL}_2(\mathbb{A})}(\chi\delta_{\mathrm{B}}^{\frac{1}{2}})$$

where $\chi = (\chi_1, \chi_2)$ is character of the Borel subgroup B *of upper triangular matrices via*

$$\chi\left(\begin{pmatrix}a & b\\ 0 & d\end{pmatrix}\right) = \chi_1(a)\chi_2(d)$$

and where

$$\delta_{\mathrm{B}}\left(\begin{pmatrix}a & b\\ 0 & d\end{pmatrix}\right) = \frac{|a|}{|d|}$$

is the modular quasi-character of B.

The fundamental question in automorphic forms is the question of which admissible representations occur as automorphic representations (really automorphic cuspidal representations). There is a conjectural restriction imposed at infinity, namely Selberg's eigenvalue 1/4 conjecture. But infinity is no different than any other place of $\mathbb{Q}$. The analogous condition is

Conjecture 19.10 (Generalized Ramanujan conjecture)**.** *Suppose $\pi = \otimes'_v \pi_v$ is an irreducible automorphic cuspidal representation of $\mathrm{GL}_2/\mathbb{Q}$. Then every local representation π_v appearing in π is tempered.*

Remark 19.11. The notion of tempered is entirely representation-theoretic, and, most importantly, local. It can be made very explicit, however we don't have time to dive into the local representation theory in enough detail to explain this carefully.

[10]Don't think of this as deep–the underlying reason behind it is the fact that if π is an irreducible representation of a product of groups $G_1 \times G_2$, then $\pi \cong \sigma_1 \otimes \sigma_2$ for σ_i irreducible representations of G_i.

Remark 19.12. This directly generalizes Ramanujan's original conjecture bounding the Fourier coefficients of the modular form Δ to all Maass forms, while also incorporating Selberg's conjecture at ∞. There is a lot hiding here!

20. Langlands' reciprocity conjecture

Last time we began looking at the problem of which local representations could occur as components of a global automorphic representation. Langlands noticed something extremely striking–that this question seemed to be related to the question of which local Galois representations appear as restrictions of global Galois representations.

Today, I would like to explain these conjectures a little more carefully than I did on day one.

We work in a slightly more general context than GL_2. So, for now, let $\mathrm{G} = \mathrm{GL}_n$. One can take, word for word, the same definition of automorphic form in this context. The cuspidality condition becomes a bit more serious: for every P a standard parabolic subgroup of G (corresponding to a partition $n = n_1 + ... + n_k$), and every unipotent radical $\mathrm{N_P}$, we require

$$\int_{\mathrm{N_P}(\mathbb{Q})\backslash \mathrm{N_P}(\mathbb{A})} \phi(ng)dn = 0$$

for all g. However, beyond the appearance of multiple types of parabolic subgroups, there is little formal difference in the definitions.

We would like to explain the reciprocity conjecture of Langlands. Recall that we said, on the first day of class, that this is roughly a bijection of the form

$$\left\{ \begin{array}{c} \text{irr. cont. representations} \\ \rho : \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_n(\mathbb{C}) \end{array} \right\} \leftrightarrow \left\{ \begin{array}{c} \text{cuspidal automorphic representations} \\ \pi \subset L^2(\mathrm{GL}_n(\mathbb{Q})\backslash \mathrm{GL}_n(\mathbb{A}_{\mathbb{Q}})) \end{array} \right\}.$$

There are few problems with formulating the conjecture this way.

(1) First, there can be no such bijection for stupid reasons! The right hand side is easily seen to be uncountable (simply by twisting by $|\det(\cdot)|^{it}$). However, the left hand side is countable, as this is a countable union of sets of representations of finite groups.
(2) Second, even if we had a bijection, this is meaningless–until we impose some conditions on this, any two sets of the same cardinality have a bijetion between them. So we need to refine our conjecture and ask for some additional compatibility with local data.

We address the second point first. This requires some local theory.

20.1. **A local prerequisite: the Satake isomorphism.** Suppose that $\pi = \pi_p$ is representation of $\mathrm{G} = \mathrm{GL}_n(\mathbb{Q}_p)$ which is spherical, i.e. which has a fixed $\mathrm{K} = \mathrm{GL}_n(\mathbb{Z}_p)$ vector. The goal in this section is to attach to such a spherical representation some combinatorial data. We will need the following easy lemma.

Lemma 20.1. *Suppose π is an irreducible admissiible spherical representation. Then the one dimensional $\mathbb{C}$-vector space π^{K} is a module over $\mathscr{H}_K := C_c^\infty(\mathrm{K}\backslash \mathrm{G} / \mathrm{K})$. π is determined uniquely by the $\mathscr{H}_K$-module structure on π^{K}.*

Proof. (Sketch) That π^{K} is a module over $\mathscr{H}_{\mathrm{K}}$ is obvious. We showed last time that $\mathscr{H}_{\mathrm{K}}$ is commutative and that π^{K} was an irreducible module, hence one dimensional.

Thus, all that remains is to show that if π_1, π_2 are two irreducible spherical representations, then $\pi_1^{\mathrm{K}} \cong \pi_2^{\mathrm{K}}$ implies $\pi_1 \cong \pi_2$. To see this, one needs only to see that for all $\phi \in C_c^\infty(\mathrm{G})$, the action on matrix coefficients is the same. In the case that π_i are spherical, it is easy to see that $\widetilde{\pi}_i$ are also spherical, and that we can reduce to showing

$$\langle \pi_1(\phi)v_1, \widetilde{v_1}\rangle = \langle \pi_1(\phi)v_2, \widetilde{v_2}\rangle$$

where $v_i, \widetilde{v_i}$ are spherical vectors normalized so that $\langle v_i, \widetilde{v_i}\rangle = 1$. To see this, work with the cases $\phi \in \mathscr{H}_K$ and ϕ in the kernel of the projection

$$\begin{aligned} \mathscr{C}_c^\infty(\mathrm{G}) &\to \mathscr{H}_K \\ \phi &\mapsto \mathbb{1}_K * \phi * \mathbb{1}_K \end{aligned}$$

separately to conclude. □

Lemma 20.2. *Given any character $a : \mathscr{H}_{\mathrm{K}} \to \mathbb{C}$, there is a corresponding spherical π with* K*-fixed vector v so that $\pi(\phi)v = a(\phi)v$ for all $\phi \in \mathscr{H}_{\mathrm{K}}$.*

Remark 20.3. This relies on the construction of spherical representations as certain induced representations. Since we have not developed the appropriate technology, I will not prove this here.

Thus we have

$$\left\{ \begin{array}{c} \text{irr. spherical representations} \\ \pi \text{ of } \mathrm{GL}_n(\mathbb{Q}_p) \end{array} \right\} \leftrightarrow \left\{ \begin{array}{c} \text{homomorphisms} \\ a : \mathscr{H}_{\mathrm{K}} \to \mathbb{C} \end{array} \right\} \leftrightarrow \mathrm{Spec}(\mathscr{H}_{\mathrm{K}})(\mathbb{C}).$$

We can even be more explicit about the ring structure on $\mathscr{H}_{\mathrm{K}}$.

Theorem 20.4 (Satake isomorphism for GL_n.)**.** *Let* A *denote the usual diagonal maximal torus of* GL_n*. Then the map*

$$S : \phi \mapsto \{a \mapsto \delta_{\mathrm{B}}(a)^{1/2} \int_{\mathrm{N_B}(\mathbb{Q}_p)} \phi(an)dn\}$$

gives rise to an isomorphism of $\mathbb{C}$*-algebras*

$$S : \mathscr{H}_{\mathrm{K}} \to \mathscr{H}_{\mathrm{A_K}}^W$$

where

$$\mathscr{H}_{\mathrm{A_K}} = C_c^\infty(\mathrm{A}(\mathbb{Z}_p)\backslash \mathrm{A}(\mathbb{Q}_p)/\mathrm{A}(\mathbb{Z}_p)) = C_c^\infty(\mathrm{A}(\mathbb{Z}_p)\backslash \mathrm{A}(\mathbb{Q}_p)) \cong \mathbb{C}[X_*(\mathrm{G},\mathrm{A})]^W$$

and the Weyl group $W \cong \Sigma_n$ *acts by permutation of coordinates on* $\mathscr{H}_{\mathrm{A_K}}$*.*

Proof. (Sketch) The argument proceeds in three steps.

First, show that the image of S lies in the correct space. That it lands in $\mathscr{A}(\mathbb{Z}_p)$ invariant functions is apparent. To see Weyl group invariance, note that you only have to show this for regular elements. Then compute.

Second, show that this is an algebra homomorphism. This essentially comes down to judicious application of the Iwasawa decomposition and computation. Be careful about measure normalizations.

Finally, show that the map is a bijection. This is not so hard: the source and target both have obvious bases. Simply compute that the "main" terms go to one another, and proceed by induction. (Show that the matrix representing S in these bases is upper triangular.) □

What does this give us? Well if we combine this isomorphism with the above observations about spherical representations, we find, since

$$\mathscr{H}_{\mathrm{A_K}} \cong \mathbb{C}[X_*(\mathrm{G},\mathrm{A})]^W$$

that spherical representations of $\mathrm{GL}_n(\mathbb{Z}_p)$ correspond to elements of

$$(\mathbb{C}^\times)^n/W \cong \frac{\mathrm{GL}_n}{\mathrm{GL}_n}(\mathbb{C}).$$

That is, every unramified representation corresponds uniquely to a semi-simple conjugacy class in $\mathrm{GL}_n(\mathbb{C})$.

20.2. **Langlands reciprocity.** Let's return to our discussion of Langlands' conjectures. We were saying that we expect something roughly like a bijection

$$\left\{ \begin{array}{c} \text{irr. cont. representations} \\ \rho : \mathrm{Gal}(\bar{\mathbb{Q}}/\mathbb{Q}) \to \mathrm{GL}_n(\mathbb{C}) \end{array} \right\} \leftrightarrow \left\{ \begin{array}{c} \text{cuspidal automorphic representations} \\ \pi \subset L^2(\mathrm{GL}_n(\mathbb{Q})\backslash \mathrm{GL}_n(\mathbb{A}_{\mathbb{Q}})) \end{array} \right\}.$$

We can now pin down some appropriate local restrictions to make this a meaningful statement. Since at almost all places v the local component π_v of π is spherical, we can use Satake to assign a semisimple conjugacy class $a(\pi_v) \in \frac{\mathrm{GL}_n(\mathbb{C})}{\mathrm{GL}_n(\mathbb{C})}$ to π_v, called the *Satake parameter* associated to π_v.

On the other hands, at almost every place a Galois representation ρ must have its restriction ρ_v to the decomposition group (which is only defined up to conjugation) be unramified, i.e. trivial on inertia. For such a representation, we can meaningfully talk about the image of the Frobenius element Frob_v. But $\rho(\mathrm{Frob_v})$ is most naturally thought of as a conjugacy class in $\mathrm{GL}_n(\mathbb{C})$.

Thus, the compatibility condition we are after is the following: given $\rho \leftrightarrow \pi$, then for every v such that ρ is unramified, π_v is spherical, and

$$\rho(\mathrm{Frob}_v) = a(\pi_v).$$

Conversely, if v is such that π_v is spherical, then ρ is unramified at v and the equality above holds.

Wonderful, this pins down the isomorphism completely (this is since any π is determined by π_v for all but finitely many places, and similarly for ρ by Chebotarëv density). However, the issue remains: there can't be a bijection between these two sets, since they don't even have the same cardinality! The problem: the Galois groups is not big enough for $\mathbb{C}$-valued representations to account for all automorphic representations.

Conjecture 20.5 (Langlands)**.** *There exist topological groups $\mathscr{L}_{\mathbb{Q}}$, together with local versions $\mathscr{L}_{\mathbb{Q}_v}$. These are extensions of the Weil groups $W_{\mathbb{Q}}$, resp. $W_{\mathbb{Q}}$, by compact groups and come with canonical embeddings and morphisms*

$$\begin{array}{ccccc} \mathscr{L}_{\mathbb{Q}_v} & \longrightarrow & W_{\mathbb{Q}_v} & \longrightarrow & \mathrm{Gal}(\overline{\mathbb{Q}_v}/\mathbb{Q}_v) \\ \downarrow & & \downarrow & & \downarrow \\ \mathscr{L}_{\mathbb{Q}} & \longrightarrow & W_{\mathbb{Q}} & \longrightarrow & \mathrm{Gal}(\overline{\mathbb{Q}_v}/\mathbb{Q}_v) \end{array}$$

where all embeddings are defined up to conjugation. There is a canonical bijection

$$\left\{ \begin{array}{c} \textit{irr. cont. representations} \\ \rho : \mathscr{L}_{\mathbb{Q}} \to \mathrm{GL}_n(\mathbb{C}) \end{array} \right\} \leftrightarrow \left\{ \begin{array}{c} \textit{cuspidal automorphic representations} \\ \pi \subset L^2(\mathrm{GL}_n(\mathbb{Q}) \backslash \mathrm{GL}_n(\mathbb{A}_{\mathbb{Q}})) \end{array} \right\}$$

which preserves the local compatibility with the Satake isomorphism.

$\mathscr{L}_{\mathbb{Q}}$ is called the *global Langlands group.*

Remark 20.6. Note that this conjecture implies that every $\mathbb{C}$-valued Galois representation (such things are called *Artin representations*) is automorphic. Since L-functions for automorphic representations are known to have meromorphic continuation, functional equation, etc., this conjecture implies Artin's conjecture!

There is a better way to describe this conjecture. We can try to upgrade the bijection above into an equivalence of categories instead of what we have, which is essentially a description of isomorphism classes. That is, at first approximation, we can remove the word "cuspidal" on the RHS and the word "irreducible" on the LHS. Then we can talk about the structure of category of representations of $\mathscr{L}_{\mathbb{Q}}$ and what that induces on the RHS.

Namely, the category of representations of any group has the structure of a Tannakian category. We should expect the same on the right hand side. This offers a different perspective on the existence of $\mathscr{L}_{\mathbb{Q}_p}$–a sort of construction.

Conjecture 20.7 (Langlands)**.** *The category* $\mathrm{RepAut}_{\mathbb{Q}}$ *of (isobaric) automorphic representations of* $\mathrm{GL}_n/\mathbb{Q}$*, for all n varying, has the structure of a neutral Tannakian category. In particular, it admits*

1. *A tensor product, which attaches to two automorphic representations π_n and π_m of GL_n and GL_m respectively an automorphic representation $\pi_n \boxtimes \pi_m$ of GL_{nm}.*
2. *An exact faithful tensor functor* $\mathrm{Fib} : \mathrm{RepAut}_{\mathbb{Q}} \to \mathrm{Vect}_{\mathbb{C}}$*. This should attach to an automorphic representation the underlying space of its associated "Galois" representation.*

Remark 20.8. Even the direct sum structure here is not trivial–one has to use the theory of Eisenstein series to explain a meaninful way to take direct sums. This construction is the reason for the word isobaric above.

21. Langlands' functoriality conjecture

Last time, we discussed some of the fundamentals of the reciprocity conjecture of Langlands. The key point was the existence of a group, the global Langlands group $\mathscr{L}_{\mathbb{Q}}$, whose semi-simple finite dimensional representations corresponded to (isobaric) automorphic representations.

We are very far from an understanding of such an object. However, one can propose a similar local statement. Recall that part of the conjecture of Langlands asks also for local groups $\mathscr{L}_{\mathbb{Q}_v}$. The following is an amazing theorem of Langlands and Harris-Taylor.

Theorem 21.1 (Local Langlands correspondence for GL_n)**.** *Let*

$$\mathscr{L}_{Q_v} = \begin{cases} W_{\mathbb{R}} & \text{if } v = \infty \\ W'_{\mathbb{Q}_v} := W_{\mathbb{Q}_v} \times \mathrm{SU}_2(\mathbb{R}) & \text{if } v < \infty \end{cases}.$$

There is a canonical bijection

$$\left\{ \begin{array}{c} \text{s.s. cont. representations} \\ \varphi_v : \mathscr{L}_{\mathbb{Q}_v} \to \mathrm{GL}_n(\mathbb{C}) \end{array} \right\} \leftrightarrow \left\{ \begin{array}{c} \text{(isobaric) smooth adm. representations} \\ \pi_v \text{ of } \mathrm{GL}_n(\mathbb{Q}_{\mathrm{v}} \end{array} \right\}.$$

When φ_v is an unramified representation of $\mathscr{L}_{\mathbb{Q}_v}$ this bijection assigns to it a spherical representation π_v such that

$$\varphi_v(\mathrm{Frob}_v) = a(\pi_v)$$

where $a(\pi_v)$ is the Satake parameter of π_v. When $n = 1$ this bijection is given by local class field theory. For all n it is compatible with the formation of ε-factors and twisting by characters.

Remark 21.2. This was shown by Langlands by explicit classification of both sides in the case $v = \infty$ for $n = 2$. For general n and $v = \infty$, this is a consequence of work of Langlands and Shelstad. In the non-Archimedean case $n < \infty$ this was first shown by Harris and Taylor by studying the geometry of some simple Shimura varieties. There is a more representation theoretic proof due to Henniart. Finally, there is a remarkable proof of this for $p < \infty$ due to Scholze.

21.1. **A consequence of the reciprocity conjecture.** Let us be optimistic and pretend that the global reciprocity conjecture is true. Then given a homomorphism $\rho : \mathrm{GL}_n(\mathbb{C}) \to \mathrm{GL}_m(\mathbb{C})$, we can post-compose a $\varphi : \mathscr{L}_{\mathbb{Q}} \to \mathrm{GL}_n(\mathbb{C})$ with ρ.

This should give rise to an assignment

$$\{ \text{ aut. representations of } \mathrm{GL}_n \} \to \{ \text{ aut. representations of } \mathrm{GL}_m \}$$

called the *functorial transfer along ρ*. This can occur more generally between autmorphic representations on arbitrary reductive groups. To state this more carefully, we need some language.

21.2. **Reductive groups.** This will be a very fast recollection of some facts from the theory of reductive groups over a field. For proofs, see Springer's text.

For a moment, let k be an arbitrary field and $\bar{k}$ its separable closure.

An *algebraic group* G over k is simply a group scheme G/k of finite type. We say that G is *linear algebraic* if it is affine. This can be shown to be equivalent to the existence of a faithful finite-dimensional representation of G, i.e. it can be shown that G is affine if and only if it can be realized as a closed subgroup scheme of some GL_n.

Some linear algebraic groups have particularly strange representation theory. We say a linear algebraic group G is *unipotent* if every finite dimensional representation has a non-zero fixed vector. This is equivalent to saying that every element of $\mathrm{G}(\bar{k})$ is unipotent in every finite dimensional representation. The representation theory of unipotent groups is essentially based on the (wild) problem of classifying iterated extensions of the trivial representation by itself.

The *unipotent radical* of a group G is its maximal connected normal unipotent subgroup. We say G is *reductive* if it has trivial unipotent radical.

Example 21.3. Here are some examples.

- Take $\mathrm{G} = \mathbb{G}_a$. This is always unipotent. More interestingly, take the group N of upper-uni-triangular matrices in GL_n.

- Consider B the group of upper-triangular matrices in GL_n. This is neither unipotent nor reductive. Its unipotent radical is N.
- GL_n is reductive. More generally, any of the classical groups $\mathrm{GL}_n, \mathrm{SL}_n, \mathrm{PGL}_n, \mathrm{SO}_n, \mathrm{Sp}_{2n}, \ldots$ are all reductive.
- All tori are reductive. Recall that group G is a *torus* if $\mathrm{G}_{\bar{k}} \cong \mathbb{G}_{m,k}^n$ for some n. We say G is split if such an isomorphism exists over k.
- Suppose k has characteristic p. Consider the group scheme $\alpha_p = \operatorname{Spec} k[x]/x^p$. This is a connected and unipotent dimension 1 algebraic group.

Reductive groups over algebraically closed fields are classified by simple combinatorial data. To make this work over a non-algebraically closed field such as $k = \mathbb{Q}$, we need some additional conditions on G.

A subgroup B of a reductive G is a *Borel* subgroup if B is a maximal connected solvable subgroup. A subgroup A is a *maximal split torus* if A is maximal among all split tori. Of course, given a reductive G/k, its base change $\mathrm{G}_{\bar{k}}$ has a Borel and even a split torus. When such subgroups can be defined over k, we are in business. That is, we say that G is *quasi-split* if G has a Borel subgroup defined over k, and that G is *split* if G has a split torus maximal among all tori.

Root data. Action on based root data given by Galois. Pinnings, etc.

Split groups are relatively rare. However, given a general reductive G, there always exists a unique quasi-split inner form. This is problem in Galois cohomology: namely inner forms of G are those forms of G (i.e. other groups over k which become isomorphic after extension of scalars to $\bar{k}$) which arise from a conjugation over $\bar{k}$. They are classified by

$$H^1(k, \mathrm{G}/\mathrm{Z}).$$

21.3. **The L-group.** A *pinning* of a reductive group is a choice $\{n_\alpha\}$ of non-zero unipotent elements, $n_\alpha \in \mathrm{N}_\alpha - \{1\}$ for all $\alpha \in \Delta$. Outer automorphisms are the same as pinned automorphisms.

22. The simplest case of functoriality

Recall that last time we spent some time defining the Langlands L-group of a reductive group over a field k. This was defined as follows. Given a reductive group G over k which is quasi-split, we took the appropriate based root datum $\Psi_0 = \Psi_0(\mathrm{G}, \mathrm{B}) := (X^*(A_{\bar{k}}), \Delta, X_*(A_{\bar{k}}), \Delta^\vee)$. We could then flip this data (flip $X^* \leftrightarrow X^*$ and $\Delta \leftrightarrow \Delta^\vee$) to construct a root datum $\Psi_0^\vee$ which is the based root datum of a group $\widehat{\mathrm{G}}/\mathbb{C}$.

If we fix a pinning $\{u_\alpha\}_\alpha$ for $\widehat{\mathrm{G}}$, this gives rise to a splitting of the exact sequence

$$1 \to \widehat{\mathrm{G}}/\widehat{\mathrm{Z}} \to \mathrm{Aut}(\widehat{\mathrm{G}}) \to \mathrm{Out}(\widehat{\mathrm{G}}) \to 1.$$

References

[Art05] James Arthur. An introduction to the trace formula. In *Harmonic analysis, the trace formula, and Shimura varieties*, volume 4 of *Clay Math. Proc.*, pages 1–263. Amer. Math. Soc., Providence, RI, 2005.

[GJ79] Stephen Gelbart and Hervé Jacquet. Forms of GL(2) from the analytic point of view. In *Automorphic forms, representations and L-functions (Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977), Part 1*, Proc. Sympos. Pure Math., XXXIII, pages 213–251. Amer. Math. Soc., Providence, R.I., 1979.

[Hel01] Sigurdur Helgason. *Differential geometry, Lie groups, and symmetric spaces*, volume 34 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2001. Corrected reprint of the 1978 original.

[Iwa02] Henryk Iwaniec. *Spectral methods of automorphic forms*, volume 53 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI; Revista Matemática Iberoamericana, Madrid, second edition, 2002.

[Lap10] Erez M. Lapid. Introductory notes on the trace formula. In *Automorphic forms and the Langlands program*, volume 9 of *Adv. Lect. Math. (ALM)*, pages 135–175. Int. Press, Somerville, MA, 2010.

[MgW95] C. Mœ glin and J.-L. Waldspurger. *Spectral decomposition and Eisenstein series*, volume 113 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1995. Une paraphrase de l'Écriture [A paraphrase of Scripture].

[Sar03] Peter Sarnak. Spectra of hyperbolic surfaces. *Bull. Amer. Math. Soc. (N.S.)*, 40(4):441–478, 2003.