REPRESENTATION THEORY AT INFINITY: AUTOMORPHIC FORMS AND THE SELBERG TRACE FORMULA

KENZ KALLAL

"Il a fallu Maass pour nous sortir du ghetto des fonctions holomorphes"

André Weil

Abstract. These notes are about automorphic forms on $GL_2(\mathbf{R})^+$, and the applications of their theory to problems in number theory and geometry. We cover three pieces of the theory: the connection to representation theory and the classification of $(\mathfrak{g}, K)$-modules for $GL_2(\mathbf{R})^+$, the application via the Selberg trace formula to prime geodesic theorems, and the further use of those to prove averaging results about class numbers of real quadratic fields. We follow the references of Bump and Langlands for the general theory of automorphic forms and representations, and Sarnak's articles and Hejhal's book for the applications of the trace formula.

Contents

1. Introduction

One extremely useful way to think about modular forms is as sections of line bundles on modular curves. In particular, a classical holomorphic modular form of weight $2k$ for a congruence subgroup $\Gamma \subset SL_2(\mathbf{Z})$ is the same thing as a meromorphic section of $(\Omega^1_{X(\Gamma)/\mathbf{C}})^{\otimes k}$ with certain restrictions on the poles (see [23, Ch. 1]). One constructs a line bundle ω which is the pushforward on $X(\Gamma)/\mathbf{C}$ of the sheaf of differentials on the universal elliptic curve,

Date: December 17, 2022.

and defines the modular forms of weight k to be the global sections of $\omega^{\otimes k}$ (for even k this coincides with the definition as k-fold differentials via the Kodaira–Spencer isomorphism). With this viewpoint, there is a systematic way to compute $\dim_{\mathbf{C}} M_k(\Gamma, \mathbf{C})$ or $\dim_{\mathbf{C}} S_k(\Gamma, \mathbf{C})$, by applying the Riemann–Roch theorem and computing the genus of $X(\Gamma)$ using Riemann–Hurwitz. This type of technique does not appear in this paper. Instead, this paper will partly be about the theory of *automorphic forms*, especially on $GL_2(\mathbf{R})$, of which the holomorphic modular forms are an example. If we decide that computing $\dim M_k$ and $\dim S_k$ is a good test of how useful a viewpoint on modular forms is, then the theory of automorphic forms certainly passes the test: one can derive formulae for $\dim S_k$ using the *Arthur-Selberg trace formula*, which in fact also gives formulae for the traces of all the Hecke operators acting on S_k (one modern reference is [13], but it was done by Selberg [21]; see also Duflo–Labesse [5]). This belongs to the theory of automorphic forms on $GL_2(\mathbf{A_Q})$, where the Hecke operators are most naturally defined. The theory of automorphic forms is the setting of one side of the conjectural Langlands correspondence. The Arthur–Selberg trace formula itself is moreover a crucial tool underpinning many results in the Langlands program: for instance, the Jacquet–Langlands correspondence, the base change lift, and the Langlands–Tunnel theorem, all of which are key features of Wiles' and Taylor–Wiles' proof of Fermat's last theorem. This is to say that when it comes to modular forms, there is more to life than sections of line bundles on modular curves.

This paper is not about the Langlands program, and in fact we will not discuss automorphic forms on adelic groups at all. Instead, we write about something which is more concrete in two ways:

(1) We will only consider the group GL_2, as opposed to arbitrary reductive algebraic groups.
(2) We will take $\mathbf{R}$-points rather than $\mathbf{A_Q}$-points.

Despite the simplicity of this situation compared to the general adelic context, it is still of obvious arithmetic significance: the Maass forms and holomorphic modular forms can still be seen as automorphic forms on $GL_2(\mathbf{R})$. In this paper, we explain the use of the Selberg trace formula in this setting, where it may be used to relate the spectral theory of the Laplace-Beltrami operator Δ on a compact Riemann surface of genus $g \geq 2$ to the asymptotic behavior of the number of (prime) geodesics of bounded length on X. This technique may also be extended to the case of noncompact finite-volume quotients of the upper half-plane (e.g. the uncompactified modular curves $Y(\Gamma)$), where one must take into account the continuous part of the spectrum of Δ. We will also discuss a fascinating application to arithmetic, also discovered by Sarnak: such a prime geodesic theorem for uncompactified modular curves results in an asymptotic formula for averages of class numbers of real quadratic fields, one of the most successful attempts to decouple the regulator from the class number in the classical Gauss–Siegel asymptotic formula.

In the interest of length, we have made the choice to include enough technical arguments from analysis to fully prove everything relating to the cuspidal part of the spectrum, and omit everything that requires knowledge of the Eisenstein series.

These notes are organized as follows. In the short Section 2, we will follow Bump [3, Ch. 2] in explaining how to view modular forms and Maass forms as automorphic forms on $GL_2(\mathbf{R})$, and set the stage for the application of representation theory. In Section 3, we follow Bump [3, Ch.2-3], Langlands [16], and Knapp [12] in carrying out the necessary representation-theoretic and analytic arguments to establish the necessary basic results in

infinite-dimensional representation theory of $GL_2(\mathbf{R})$, though our proofs are built to generalize to arbitrary reductive Lie groups. In that section, we will see that understanding the decomposition of the right regular representation of $GL_2(\mathbf{R})^+$ hinges on understanding the spectrum of the Laplace–Beltrami operator on quotients of the Poincaré upper half-plane. In Section 4, we discuss the Selberg trace formula following Hejhal [10], prove it fully under the assumption that Γ is hyperbolic and $\Gamma\backslash\mathbf{H}$ is compact, and explain how to apply it following Sarnak's thesis [19].

2. Automorphic forms on $GL_2(\mathbf{R})$

Life begins with the upper half-plane

$$\mathbf{H} = \{x + iy \in \mathbf{C} : y > 0\},$$

on which the holomorphic modular forms and Maass forms are functions. The way to view them as automorphic forms on $GL_2(\mathbf{R})$ is as follows. Acting on the left by fractional linear transformations, the connected component $GL_2(\mathbf{R})^+$ acts transitively on $\mathbf{H}$ (if the determinant is negative then $\mathbf{H}$ gets sent to the lower half-plane). The map

$$\begin{aligned} GL_2(\mathbf{R})^+ &\to \mathbf{H} \\ g &\mapsto g(i) \end{aligned}$$

is therefore surjective. One checks explicitly that the stabilizer of i is

$$Z^\circ K^\circ = \mathbf{R}^\times_{>0} \cdot SO_2(\mathbf{R}) \subset GL_2(\mathbf{R})^+,$$

where $Z \cong \mathbf{R}^\times$ is the center of $GL_2(\mathbf{R})$, i.e. the subgroup of scalar matrices, $K = O_2(\mathbf{R})$ is a maximal compact subgroup of $GL_2(\mathbf{R})$, and Z° and K° are their respective connected components. The subgroups $ZK, Z^\circ K^\circ \subset GL_2(\mathbf{R})$ are closed[1]. So as a 2-dimensional manifold, we have

$$\mathbf{H} \cong GL_2(\mathbf{R})^+/Z^\circ K^\circ \cong PGL_2(\mathbf{R})^+/SO_2(\mathbf{R}) \cong PGL_2(\mathbf{R})/O_2(\mathbf{R}) \cong GL_2(\mathbf{R})/ZK.$$

Since ZK is not normal in $GL_2(\mathbf{R})$, this does not equip $\mathbf{H}$ with any Lie group structure, but this isomorphism still preserves the action of $GL_2(\mathbf{R})$ on the left. And for any discrete subgroup $\Gamma \subset GL_2(\mathbf{R})$ acting nicely enough[2] on $\mathbf{H}$, the quotient $\Gamma\backslash\mathbf{H}$ can be seen as a double coset space

$$\Gamma\backslash GL_2(\mathbf{R})^+/Z^\circ K^\circ.$$

When Γ expresses a level structure on the set of isomorphism classes of elliptic curves, this double coset space is, compared to the modular curve $Y(\Gamma)$, an equally valid way of thinking about the corresponding moduli problem. For example, if $\Gamma = SL_2(\mathbf{Z})$, then this double coset space is clearly in natural bijection with the set of full-rank lattices in $\mathbf{R}^2 \cong \mathbf{C}$ up to isomorphism given by multiplication by a nonzero complex number. Recall

---

[1]For example, because it consists of the 2×2 invertible matrices of the form $\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$; being of this form is clearly a closed condition in $\mathbf{R}^4$, so after intersecting with $GL_2(\mathbf{R})$ it is closed.

[2]"nicely enough" just needs to be "so that the quotient actually is a manifold. We can take this to be the weak version of "properly discontinuous" in which each $x \in \mathbf{H}$ has a neighborhood U such that the set of $g \in \Gamma$ such that $gU \cap U \neq \emptyset$ is finite.

Definition 2.1 (Maass forms)**.** Let $k \geq 0$ be an integer and $\chi : \Gamma \to \mathbf{C}^\times$ a character. A *Maass form* of weight k and character χ is a smooth function

$$f : \mathbf{H} \to \mathbf{C}$$

satisfying

$$f(\gamma z) = \chi(\gamma) \left(\frac{cz + d}{|cz + d|} \right)^k f(z)$$

for all

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \gamma,$$

$$\Delta_k f = \lambda f$$

for some constant λ, where Δ_k is the weight-k Laplacian

$$\Delta_k = -y^2 \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right) + iky \frac{\partial}{\partial x},$$

and also satisfying a polynomial growth condition at the cusps of $\Gamma \backslash \mathbf{H}$.

The point of the above observations is that the Maass forms of weight[3] 0 can be considered as elements of the complex Hilbert space

$$L^2(\Gamma \backslash GL_2(\mathbf{R})^+ / Z^\circ K^\circ, \chi) = L^2(\Gamma \backslash PGL_2(\mathbf{R})^+ / K^\circ, \chi)$$

consisting of the measurable functions $f : GL_2(\mathbf{R})^+ \to \mathbf{C}$ with the property that[4]

$$f(\gamma g u \kappa) = \chi(\gamma) f(g)$$

for all $\gamma \in \Gamma$, $g \in GL_2(\mathbf{R})^+$, $u \in Z^\circ$, $\kappa \in K^\circ$ and

$$\int_{\Gamma \backslash PGL_2(\mathbf{R})} |f(x)|^2 \, dx < \infty$$

where dx is the Haar measure on $PGL_2(\mathbf{R})$. Since K, K° are compact, the choice of whether to quotient out by K° in the domain of integration is irrelevant.

This L^2 space is really only useful insofar as we can use representation theory to study it. We would want to define a left action of $GL_2(\mathbf{R})^+$, via right regular representation $(g \cdot f)(x) = f(xg)$. The problem with this is that K° is not in the center of $GL_2(\mathbf{R})^+$, so this action would not take $L^2(\Gamma \backslash GL_2(\mathbf{R}) / ZK, \chi)$ to itself.

This is maybe the motivation for removing the requirement of K-invariance, and considering the larger Hilbert space

$$\mathfrak{H} = L^2(\Gamma \backslash PGL_2(\mathbf{R})^+, \chi),$$

which admits a left-action of $GL_2(\mathbf{R})^+$ (namely the right regular action). The space of smooth vectors for this action is

$$C^\infty(\Gamma \backslash GL_2(\mathbf{R})^+ / Z^\circ, \chi),$$

defined in the obvious way. It is a general fact that the smooth vectors are dense:

[3]We are about to explain the representation-theoretic reason for the definition of Maass forms of general weight.

[4]We can also add a choice of "central character" $\omega : Z \to S^1$ with the obvious change to the definition, but this is not very relevant to the current discussion. The inclusion of the χ is just to reassure the reader that modular forms with Nebentypus character can be dealt with in this setting.

Lemma 2.2. *Let $\pi : G \to \mathrm{End}(\mathfrak{H})$ be a representation of a Lie group G into a Hilbert space $\mathfrak{H}$. The space of smooth vectors for this representation, $\mathfrak{H}^\infty$, is dense in $\mathfrak{H}$.*

Proof. The method is by convolution by a smooth function $\phi \in C_c^\infty(G)$. Let

$$\pi(\phi)v = \int_G \phi(g)\pi(g)v\, dg,$$

which is well-defined for all $v \in \mathfrak{H}$ because ϕ is compactly supported. In the toy model where π is the left regular representation and $\mathfrak{H} = L^2(G)$, this is the same as convolving a function in that L^2 space with ϕ.

Let $\mathfrak{g}$ be the Lie algebra of G. For any $v \in \mathfrak{H}$, $\phi \in C_c^\infty(G)$, and $X \in \mathfrak{g}$, we have

$$\begin{aligned}
\left.\frac{d}{dt}\right|_{t=0} \pi(\exp(tX))\pi(\phi)v &= \left.\frac{d}{dt}\right|_{t=0} \pi(\exp(tX)) \int_G \phi(g)\pi(g)v\, dg \\
&= \left.\frac{d}{dt}\right|_{t=0} \int_G \phi(g)\pi(\exp(tX)g)v\, dg \\
&= \left.\frac{d}{dt}\right|_{t=0} \int_G \phi(\exp(-tX)g)\pi(g)v\, dg \\
&= \int_G \left(\left.\frac{d}{dt}\right|_{t=0} \phi(\exp(-tX)g) \right) \pi(g)v\, dg
\end{aligned}$$

where the differentiation under the integral sign is okay because ϕ and $\left.\frac{d}{dt}\right|_{t=0}\phi(\exp(-tX)g)$ are compactly supported on G. So the action of $\mathfrak{g}$ on $\pi(\phi)v$ is well-defined and results in another thing of the form $\pi(\phi')v$, where $\phi' = \left.\frac{d}{dt}\right|_{t=0}\phi(\exp(-tX)g)$ is also in $C_c^\infty(G)$ and supported in the support of ϕ. It follows that the same argument we just did applies arbitrarily many times, which shows that $\pi(\phi)v \in \mathfrak{H}^\infty$.

Now the point is that we can approximate a given $v \in \mathfrak{H}$ with these smooth vectors $\pi(\phi)v$. Let $v \in \mathfrak{H}$, $\epsilon > 0$, and take an open set $U \subset G$ around the identity with the property that $|\pi(g)v - v| < \epsilon$ for all $g \in U$. This is possible because the function $g \mapsto |\pi(g)f - f|$ is continuous. By general theory of smooth manifolds, there exists a $\phi_\epsilon \in C^\infty(G)$ such that ϕ is supported on U, and $\int_G \phi_\epsilon = 1$. Then

$$|\pi(\phi_\epsilon)v - v| = \left| \int_G \phi_\epsilon(g)(\pi(g)v - v)\, dg \right| \leq \int_G \phi_\epsilon(g)\epsilon \leq \epsilon.$$

Since we showed that $\pi(\phi_\epsilon)v \in \mathfrak{H}^\infty$, this shows that $\mathfrak{H}^\infty$ is dense in $\mathfrak{H}$, as desired. □

So the Maass forms and modular forms of weight zero and character χ live in the dense subspace of smooth vectors

$$C^\infty(\Gamma\backslash GL_2(\mathbf{R})/ZK, \chi) \subset L^2(\Gamma\backslash GL_2(\mathbf{R})/ZK, \chi) \subset L^2(\Gamma\backslash GL_2(\mathbf{R})^+/Z^\circ, \chi).$$

In fact, $L^2(\Gamma\backslash GL_2(\mathbf{R})/ZK, \chi)$ is precisely the set of K-fixed vectors in $L^2(\Gamma\backslash GL_2(\mathbf{R})^+/Z^\circ, \chi)$, in other words the K°-isotypic subspace for the trivial representation of K°. To see the modular forms and Maass forms of arbitrary weight, we must look at all the K°-isotypic subspaces. This is a general technique in representation theory: when you have a Hilbert space representation of a group G, restrict it to a maximal compact subgroup K and use the representation theory of compact groups to your advantage. In our case, there are two reasons why it is more convenient to think about the connected component $G = GL_2(\mathbf{R})^+$ rather than $GL_2(\mathbf{R})$:

(1) It is more naturally connected to the upper half-plane, since the fractional linear transformations of negative determinant take the upper half-plane to the lower half-plane
(2) The maximal compact subgroup $K^\circ \subset GL_2(\mathbf{R})^+$ is abelian, whereas $K = O_2(\mathbf{R})$ is not.

These things don't make a big difference, because $PGL_2(\mathbf{R})/O_2(\mathbf{R}) \cong PGL_2(\mathbf{R})^+/SO_2(\mathbf{R})$, and it isn't hard to write down the irreducible representations of $O_2(\mathbf{R})$ by induction from $SO_2(\mathbf{R})$.

We started with the weight-0 modular and Maass forms of character χ, which are in the trivial K°-isotypic subspace of $L^2(\Gamma\backslash PGL_2(\mathbf{R})^+, \chi)$. The irreducible unitary representations of $K^\circ = SO_2(\mathbf{R}) \cong S^1$ are all 1-dimensional and given by the characters $\kappa_\theta \mapsto e^{ik\theta}$ for $k \in \mathbf{Z}$, where $\kappa_\theta \in SO_2(\mathbf{R})$ is a counterclockwise[5] rotation by $\theta \in [0, 2\pi)$. So by the Peter–Weyl theorem, we have a decomposition as a Hilbert space direct sum

$$L^2(\Gamma\backslash PGL_2(\mathbf{R})^+, \chi) = \bigoplus_{k \in \mathbf{Z}} L^2(\Gamma\backslash PGL_2(\mathbf{R})^+, \chi, k)$$

where $L^2(\Gamma\backslash PGL_2(\mathbf{R})^+, \chi, k)$ is the K-isotypic subspace corresponding to the character $\kappa_\theta \mapsto e^{-ik\theta}$ of K°. Generalizing the fact that $C^\infty(\Gamma\backslash PGL_2(\mathbf{R})^+/SO_2(\mathbf{R}), \chi)$ contains the Maass forms of weight 0, the smooth vectors in these K-isotypic components correspond to Maass forms of weight k. To get from a function on the upper half-plane to an element of the weight-k K-isotypic subspace, we can't simply transfer the function over using the isomorphism $\mathbf{H} \cong GL_2(\mathbf{R})^+/Z^\circ K^\circ$, since that function would always be in the isotypic subspace corresponding to $k = 0$. Instead, one must twist by the appropriate character of K°, using the Iwasawa decomposition. From this we recover the symmetry condition satisfied by Maass forms of weight k: let $L^2(\Gamma\backslash \mathbf{H}, \chi, k)$ be the subspace of $L^2(\mathbf{H})$ defined by the condition

$$f(\gamma z) = \chi(\gamma) \left(\frac{cz + d}{|cz + d|}\right)^k f(z), \qquad \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma.$$

Lemma 2.3. *The map*

$$\sigma_k : L^2(\Gamma\backslash \mathbf{H}, \chi, k) \to L^2(\Gamma\backslash PGL_2(\mathbf{R})^+, \chi, k)$$

given by

$$(\sigma_k f)(g) = e^{-ik\theta_g} f(x_g + iy_g),$$

where θ_g, x_g, y_g are defined via the Iwasawa decomposition

$$g = \begin{pmatrix} y^{1/2} & y^{-1/2}x \\ 0 & y^{-1/2} \end{pmatrix} \kappa_\theta,$$

is an isomorphism of Hilbert spaces.

Proof. The fact that this map is a Hilbert space (unitary) map follows from the fact that the Haar measure on $PGL_2(\mathbf{R})^+$ pushes forward under the map $PGL_2(\mathbf{R})^+ \to \mathbf{H}$ to the measure coming from the hyperbolic metric on $\mathbf{H}$. To check this, one just needs to compute the left Haar measure on

$$\left\{ \begin{pmatrix} y^{1/2} & y^{-1/2}x \\ 0 & y^{-1/2} \end{pmatrix} : y > 0 \right\}.$$

[5]N.B.: this convention is the opposite of what is used in Bump, which explains why some of our formulas differ slightly from his.

To recover f from $\sigma_k f$, we just take $f(x+iy) = (\sigma_k f)\begin{pmatrix} y & x \\ 0 & 1 \end{pmatrix}$, which works because the K-component of this matrix in the Iwasawa decomposition is zero. It just remains to check that $\sigma_k f$ being a $e^{ik\theta}$-simultaneous eigenvector for the action of K is equivalent to f satisfying the symmetry property for Maass forms of weight k. This is because if $\sigma_k f = F \in L^2(\Gamma\backslash PGL_2(\mathbf{R})^+, \chi, k)$, then

$$f(\gamma\cdot(x+iy)) = F\left(\begin{pmatrix} y' & x' \\ 0 & 1 \end{pmatrix}\right) = F\left(\gamma\begin{pmatrix} y & x \\ 0 & 1 \end{pmatrix}\kappa_{\theta'}^{-1}\right) = e^{ik\theta'}\chi(\gamma)f(x+iy),$$

where $x'+iy' := \gamma\cdot(x+iy)$ and θ' is defined by the Iwasawa decomposition

$$\gamma\begin{pmatrix} y & x \\ 0 & 1 \end{pmatrix} = \sqrt{\frac{y}{y'}}\begin{pmatrix} y' & x' \\ 0 & 1 \end{pmatrix}\kappa_{\theta'}.$$

So we just need to compute θ' in terms of γ and $x+iy$. I don't know how to do it by pure thought, but the computation isn't that bad:

$$\begin{aligned}
\kappa_{\theta'} &= \sqrt{\frac{y'}{y}}\begin{pmatrix} y' & x' \\ 0 & 1 \end{pmatrix}^{-1}\begin{pmatrix} a & b \\ c & d \end{pmatrix}\begin{pmatrix} y & x \\ 0 & 1 \end{pmatrix} \\
&= (yy')^{-1/2}\begin{pmatrix} 1 & -x' \\ 0 & y' \end{pmatrix}\begin{pmatrix} ay & ax+b \\ cy & cx+d \end{pmatrix} \\
&= \frac{|cz+d|}{y}\begin{pmatrix} * & * \\ cyy' & cxy'+dy' \end{pmatrix} \\
&= \frac{1}{|cz+d|}\begin{pmatrix} * & * \\ cy & cx+d \end{pmatrix}
\end{aligned}$$

so

$$\cos\theta + i\sin\theta = \frac{cz+d}{|cz+d|},$$

and thus

$$f(\gamma\cdot(x+iy)) = \chi(\gamma)\left(\frac{cz+d}{|cz+d|}\right)^k f(x+iy)$$

as desired. □

So we have provided a natural explanation of the symmetry condition satisfied by the Maass forms of weight k in terms of the representation theory of $PSL_2(\mathbf{R})^+$. Where do the modular forms fit into this picture? Actually the answer is very simple.

Lemma 2.4. *Suppose that $f : \mathbf{H} \to \mathbf{C}$ is a modular form of weight k and character χ for Γ. Then*

$$y^{k/2} f \in C^\infty(\Gamma\backslash\mathbf{H}, \chi, k).$$

Proof. This is a straightforward observation about the relationship between the symmetry conditions satisfied by modular forms and Maass forms. In particular,

$$(\Im(\gamma\cdot z))^{k/2} f(\gamma\cdot z) = \frac{y^{k/2}}{|cz+d|^k}(cz+d)^k\chi(\gamma)f(z) = \left(\frac{cz+d}{|cz+d|}\right)^k\chi(\gamma)y^{k/2}f(z)$$

as required. □

So from the perspective of representation theory, the theory of modular forms is subsumed by the theory of Maass forms. Note that we haven't yet accounted for the entirety of the definition of a Maass form: we are still missing

(1) The growth condition at the cusps
(2) The requirement of being an eigenvalue for the Laplace operator

When there are no cusps (when $\Gamma\backslash\mathbf{H}$ is compact), (1) is irrelevant; and even then, it won't be necessary in our discussion to think about the growth condition because we will only need the discrete part of the spectrum, and probably only the cuspidal forms. (2) we can think of in the language of representation theory in terms of a decomposition into eigenspaces for the Casimir element.

3. The representation theory

Now that we have established the connection that Maass forms and modular forms have with the right regular representation of $GL_2(\mathbf{R})^+$ on $L^2(\Gamma\backslash PGL_2(\mathbf{R})^+, \chi)$, we should study that representation more systematically.

3.1. **The Lie algebra action and Maass–Shimura operators.** Since $C^\infty(\Gamma\backslash PGL_2(\mathbf{R})^+, \chi)$ are the smooth vectors in $L^2(\Gamma\backslash PGL_2(\mathbf{R})^+, \chi)$, they admit an action of the universal enveloping algebra $U(\mathfrak{g})$, induced from the right regular representation of $GL_2(\mathbf{R})$ on this L^2 space. The Laplacian on the upper half-plane, which we might originally justify as being the Laplace-Beltrami operator for the Poincaré upper half-plane, turns out to transfer over to this setting as the Casimir element of $U(\mathfrak{g})$. In fact, it is generally true that if you choose a bi-invariant metric on a Lie group G, then the Laplacian with respect to that metric coincides with the Casimir element corresponding to the induced inner product on the Lie algebra $\mathfrak{g}$.

From now on, the Lie group G is $GL_2(\mathbf{R})^+$, with maximal compact $K = SO_2(\mathbf{R})$ (this replaces the previous convention of K being a maximal compact of $GL_2(\mathbf{R})$). The center of $U(\mathfrak{g} \otimes \mathbf{C})$ is $\mathfrak{Z} = \mathbf{C}[\Delta, Z_\mathfrak{g}]$, where Δ is the Casimir element with respect to the Killing form. We can choose a Cartan subalgebra $\mathfrak{h}_\mathbf{C} \subset \mathfrak{g} \otimes \mathbf{C}$, given by the diagonal matrices. It is spanned over $\mathbf{C}$ by

$$Z_\mathfrak{g} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad H = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Note that here the center of G, $Z = \mathbf{R}^\times_{>0}$, is distinguished in our notation from the Lie algebra vector sitting on top of it, $Z_\mathfrak{g}$.

The real subspace consisting of real diagonal matrices is also spanned over $\mathbf{R}$ by Z and H, and is a Cartan subalgebra of $\mathfrak{g}$. We should be aware that the exponential map sends this choice of $\mathfrak{h}$ to the abelian subgroup of G given by diagonal matrices with positive entries. This is inconvenient for us, because we want this to contain a maximal compact of G (so that we can compare the action of H to the action of a maximal compact). It does contain such a maximal compact, but only of $GL_2(\mathbf{C})$: those elements are

$$\exp(i\theta H) = \begin{pmatrix} e^{i\theta} & 0 \\ 0 & e^{-i\theta} \end{pmatrix}.$$

We need to perform a change of variables to make the entries real. The canonical way of doing this is to conjugate by the Cayley transform

$$\mathcal{C} = -\frac{i+1}{2}\begin{pmatrix} i & 1 \\ i & -1 \end{pmatrix}$$

and if we set $\hat{H} = \mathcal{C}H\mathcal{C}^{-1}$ we have

$$\exp(i\theta\hat{H}) = \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} = \kappa_{-\theta} \in K = SO_2(\mathbf{R}).$$

Despite the fact that the actual matrix $\hat{H} = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$ is not as nice, we prefer to use this one because a decomposition of a K-finite representation of G into K-isotypic subspaces should correspond to a decomposition into eigenspaces of H. Since the compact abelian Lie group K is just S^1, such a K-finite representation V of G decomposes as an algebraic direct sum

$$V = \bigoplus_{k \in \mathbf{Z}} V(k)$$

where $V(k)$ is the isotypic component corresponding to the 1-dimensional representation of K given by the character

$$\kappa_\theta \mapsto e^{-ik\theta}.$$

If V is the space of K-finite vectors in $L^2(\Gamma\backslash G, \chi)$, then in the previous section we saw that Maass forms and modular forms of weight k and character χ for Γ can be thought of as elements of $V(k)$. By virtue of the way we changed variables via the Cayley transform, this decomposition is also an eigenspace decomposition for the action of $\hat{H}$, since

$$\begin{aligned} i\hat{H}v &= \frac{d}{dt}\Big|_{t=0} \exp(it\hat{H})v \\ &= \frac{d}{dt}\Big|_{t=0} e^{ikt}v \\ &= ikv \end{aligned}$$

for $v \in V(k)$, which means that $V(k)$ is exactly the k-eigenspace of the action of $\hat{H} \in \mathfrak{g} \otimes_{\mathbf{R}} \mathbf{C}$ on V.

Back in the setting of H and Z rather than their Cayley-transformed siblings, it is convenient that $\mathfrak{g} \otimes \mathbf{C}$ is reductive: it splits into

$$\mathfrak{sl}_2(\mathbf{C}) \oplus \mathbf{C} \cdot Z,$$

where we know that $\mathfrak{sl}_2(\mathbf{C})$ is simple (e.g. from the theory of its root system) and $\mathbf{C}Z$ is abelian. In particular, there is a maximal abelian subalgebra of $\mathfrak{sl}_2(\mathbf{C})$ spanned by H (or $\hat{H}$), and a root space decomposition

$$\mathfrak{sl}_2(\mathbf{C}) = \mathbf{C} \cdot H \oplus \mathbf{C} \cdot L \oplus \mathbf{C} \cdot R,$$

where

$$L := \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \qquad R := \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

span the -2 and $+2$ root spaces, respectively (since they are eigenvectors for the adjoint action of H with $[H, L] = -2L$ and $[H, R] = 2R$). This is just the standard root space decomposition for the semisimple Lie algebra $\mathfrak{sl}_2(\mathbf{C})$. If we conjugate by the Cayley transform (which is in $SL_2(\mathbf{C})$ which of course has a well-defined adjoint action on $\mathfrak{sl}_2(\mathbf{C})$) we get a slightly less standard root space decomposition

$$\mathfrak{sl}_2(\mathbf{C}) = \mathbf{C} \cdot \hat{H} \oplus \mathbf{C} \cdot \hat{L} \oplus \mathbf{C} \cdot \hat{R},$$

which has the advantage that the abelian subalgebra $\mathbf{C} \cdot \hat{H}$ acts nicely on the decomposition of V into K-isotypic subspaces. We still have

$$[\hat{H}, \hat{L}] = -2\hat{L}, \qquad [\hat{H}, \hat{R}] = +2\hat{R}$$

so since the decomposition of V into $V(k)$'s is a weight-space decomposition for V, the operator $\hat{L}$ decreases the H-eigenvalue by 2, and $\hat{R}$ increases it by 2. Translating to the language of functions on the upper half-plane, these produce differential operators which raise and lower the weight of Maass forms by 2.

When we think about $GL_2(\mathbf{R})^+$ instead, the only difference is that the Lie algebra has nontrivial center, namely $\mathbf{C} \cdot Z_{\mathfrak{g}}$. But since this is in the center, it necessarily acts on everything via the adjoint action by 0. And in the case we care about, namely the space of K-finite vectors in $L^2(\Gamma \backslash PGL_2(\mathbf{R})^+, \chi)$, the action of Z and thus $Z_{\mathfrak{g}}$ is also trivial. So these issues about the center will not be important for us, and all the important features that have to do with the Lie-algebra are contained in the subalgebra $\mathfrak{sl}_2\mathbf{C}$.

In the automorphic forms learning seminar, B. Lawrence asked the following question, which is really at the heart of the theory of this right regular representation, and what distinguishes the Maass forms that come from holomorphic modular forms (see Lemma 3.2) from the others.

Question 3.1. If you take a generic Maass form, and repeatedly apply $\hat{R}$ or $\hat{L}$, does it eventually vanish?

The answer to this question is key to the distinction between the two main types of infinitesimal equivalence classes of representations of G: the principal series and the discrete series representations. The basic idea is that the distinction is based on whether f comes from an (anti-)holomorphic modular form.

Lemma 3.2. *Let $f \in C^\infty(\Gamma \backslash G, \chi, k)$ be nonzero.*

(1) $\hat{L}f = 0$ if and only if $y^{-k/2}\sigma_k^{-1} f$ is a holomorphic modular form.
(2) $\hat{R}f = 0$ if and only if $y^{k/2}\sigma_k^{-1} f$ is an antiholomorphic modular form.

Proof. In the coordinates on G coming from the Iwasawa decomposition

$$g = \begin{pmatrix} u & \\ & u \end{pmatrix} \begin{pmatrix} y^{1/2} & y^{-1/2}x \\ & y^{-1/2} \end{pmatrix} \kappa_\theta,$$

the Maass–Shimura differential operators may be explicitly given by

$$\hat{R} = e^{-2i\theta}\left(iy\frac{\partial}{\partial x} + y\frac{\partial}{\partial y} - \frac{1}{2i}\frac{\partial}{\partial \theta}\right)$$
$$\hat{L} = e^{2i\theta}\left(-iy\frac{\partial}{\partial x} + y\frac{\partial}{\partial y} + \frac{1}{2i}\frac{\partial}{\partial \theta}\right)$$

so for a function $F \in C^\infty(\Gamma \backslash \mathbf{H}, \chi, k)$, we have

$$\sigma_{k+2}\hat{R}\sigma_k F = \left(iy\frac{\partial}{\partial x} + y\frac{\partial}{\partial y} + \frac{1}{2}k\right) F$$
$$\sigma_{k-2}\hat{L}\sigma_k F = \left(-iy\frac{\partial}{\partial x} + y\frac{\partial}{\partial y} - \frac{1}{2}k\right) F.$$

So as maps from $C^\infty(\Gamma\backslash\mathbf{H},\chi,k)$ to $k\pm 2$, the Maass differential operators are given by

$$\hat{R} = (z-\overline{z})\frac{\partial}{\partial z} + \frac{1}{2}k$$
$$\hat{L} = -(z-\overline{z})\frac{\partial}{\partial \overline{z}} - \frac{1}{2}k$$

$F \in C^\infty(\Gamma\backslash\mathbf{H},\chi,k)$ being killed by $\hat{R}$ is therefore equivalent to $y^{k/2}F$ (considered abstractly as a function on $\mathbf{H}$) being killed by $(z-\overline{z})\frac{\partial}{\partial z}$, since

$$(z-\overline{z})\frac{\partial}{\partial z}(y^{k/2}F) = y^{k/2}\hat{R}F.$$

By the Cauchy–Riemann equations, this is equivalent to $y^{k/2}F$ being antiholomorphic. Similarly, F being killed by $\hat{L}$ is equivalent to $y^{-k/2}F$ being holomorphic. □

We have now further understood the Lie algebra action on the right regular representation on the smooth vectors of $L^2(\Gamma\backslash G/Z,\chi)$. This will be necessary to understand how it decomposes into irreducible components.

3.2. **Decomposition into irreducible subspaces.** We want to analyze how the right regular representation $(\pi,\mathfrak{H})$ decomposes into irreducible components. Since it is unitary, by Zorn's lemma it suffices to show that any closed subrepresentation contains an irreducible subrepresentation. The full problem is quite difficult, because of the presence of the continuous spectrum, which is part of the spectrum of Δ given by a direct integral of Eisenstein series. So we only consider two cases which are actually relevant to us.

(1) In the easiest case, $\Gamma\backslash\mathbf{H}$ is compact, and there is no continuous spectrum. This excludes the case where Γ is a congruence subgroup of $SL_2(\mathbf{Z})$, so it seems like it is not very relevant to arithmetic. However, it is relevant to geometry: by the uniformization theorem for compact Riemann surfaces, the compact Riemannian manifolds $\Gamma\backslash\mathbf{H}$ (when this space is actually compact) account for all the compact Riemann surfaces of genus $g \geq 2$.
(2) In a somewhat more complicated case, Γ is a congruence subgroup of $SL_2(\mathbf{Z})$. In this case, $\Gamma\backslash\mathbf{H} = Y(\Gamma)$ is not compact, and has some finite number of cusps. The non-cusp part of $L^2(\Gamma\backslash G/Z,\chi)$ is given as a direct integral of Eisenstein series, and we will not discuss it. Instead, in this case, we will consider the right regular representation on the cuspidal part $L^2_{\text{cusp}}(\Gamma\backslash G/Z,\chi)$ which is defined by the further restriction of going to zero at the cusps. It is of course not enough to restrict to "vanishing at the cusps," since elements of L^2 are only considered up to equality almost everywhere, and besides they are not a priori functions on the compactified modular curve: one way to rigorously define a cuspidal element of L^2 is by using Fourier expansions at the cusps of $\Gamma\backslash G$.

Remark 3.3. If we want to keep going forward with the most general $\chi : \Gamma \to S^1$, we need to keep straight what the notion of cuspidal should be. The definition of cuspidal is obvious when $\chi|_{\Gamma\cap N} = 1$: in that case, f is periodic, so we say that f is cuspidal at ∞ if

$$\int_{(\Gamma\cap N)\backslash N} f(ng)\, dn = 0$$

for almost all g. The function $g \mapsto f(\gamma g) = \chi(\gamma) f(g)$ is also periodic, so f is said to be cuspidal at the cusp $\xi\infty$ if

$$\int_{(\Gamma \cap N)\backslash N} f\left(\xi^{-1} n g\right)\, dn = 0.$$

When χ has finite image (true of the most important case, when χ is a nebentypus character) and $\chi|_{\Gamma\cap N} \neq 1$, f is periodic with respect to horizontal translations, but not by everything in $\Gamma \cap N$: one must restrict to the kernel, which is the constant Fourier coefficient at ∞ is

$$\begin{aligned} f(\infty) &= \frac{1}{\mu((\ker\chi \cap N)\backslash N)} \int_{(\ker\chi\cap N)\backslash N} f(ng) \\ &= \frac{1}{\mu((\ker\chi \cap N)\backslash N)} \sum_{\gamma \in (\ker\chi\cap N)\backslash(\Gamma\cap N)} \int_{(\Gamma\cap N)\backslash N} f(\gamma n g) \\ &= \frac{1}{\mu((\ker\chi \cap N)\backslash N)} \left(\sum_{\gamma \in (\ker\chi\cap N)\backslash(\Gamma\cap N)} \chi(\gamma) \right) \int_{(\Gamma\cap N)\backslash N} f(ng) \\ &= 0. \end{aligned}$$

If χ does not have finite image, this exact argument doesn't work: there are convergence issues. A reasonable way to proceed is to use the Iwasawa decomposition $G = N \times A \times K$ and define

$$F(nak) = \chi(n)^{-1} f(nak),$$

which really is periodic with respect to the translations in $\Gamma \cap N$. Since χ has infinite image, any extension to $N \cong (\mathbf{R}, +)$ must be of the form

$$\chi : \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \mapsto e^{2\pi i \lambda_\chi},$$

where λ_χ is irrational. So f has a Fourier expansion which is $e^{2\pi i\lambda_\chi}$ times the Fourier expansion of f, and thus has no constant term (here we are transferring over to $\mathbf{H}$, i.e. fixing a coordinate in K, to be able to talk about Fourier expansions in elementary terms).

The above remark is the reason for

Definition 3.4. A function $f \in L^2(\Gamma\backslash G, \chi)$ is *cuspidal* at ∞ if $\chi|_{\Gamma\cap N} \neq 1$ or otherwise if

$$\int_{(\Gamma\cap N)\backslash N} f(ng)\, dn = 0$$

for almost all $g \in G$. It is *cuspidal* if $g \mapsto f(\xi^{-1} g)$ is cuspidal at ∞ as an element of $L^2(\xi\Gamma\xi^{-1}\backslash G/Z, \chi)$ for enough $\xi \in SL_2(\mathbf{Z})$ such that $\{\xi\infty\}$ exhausts all cusps of Γ.

We will establish, regardless of the compactness of $\Gamma\backslash\mathbf{H}$, that the relevant Hilbert space representation of $G = GL_2(\mathbf{R})^+$ decomposes as a Hilbert space direct sum of irreducible representations. The key point in these arguments is the judicious use of the spectral theorem for compact self-adjoint operators. These compact operators will be obtained by convolution by smooth functions.

From now on, we adapt the convention that $(\pi, \mathfrak{H})$ is the unitary Hilbert space right regular representation of G on the space $L^2(\Gamma\backslash G/Z, \chi)$ in the case of compact quotient, and on the cuspidal part of that space otherwise. The following two lemmas were proved in greater generality by Langlands [16], but the basic ideas are contained in these proofs for $GL_2(\mathbf{R})$.

The first is an obligatory estimate. The reader should feel free to skip it, but be aware that it is the only place where the cuspidality is an input. So the difficulties of the continuous spectrum are due to the failure of this estimate to hold when not restricted to the cuspidal part.

Lemma 3.5. *Suppose $\phi \in C_c^\infty(G)$ and $\Gamma \subset SL_2(\mathbf{Z})$ is a congruence subgroup. Then there exists a constant $C_{\phi,\Gamma,\chi}$ depending only on ϕ, χ and Γ such that*

$$\|\pi(\phi)f\|_{L^\infty} \leq C_{\phi,\Gamma,\chi}\|f\|_{L^2}$$

for all $f \in L^2_{\text{cusp}}(\Gamma\backslash G/Z, \chi)$.

Proof. The simplest way to carry out an estimate like this is to construct a crude approximation of a fundamental domain for $\Gamma\backslash G/Z$, from the standard knowledge of how $\Gamma\backslash\mathbf{H}$ works. The *Siegel set* defined via the Iwasawa decomposition

$$\mathcal{G}_{c,d} := \left\{ u \begin{pmatrix} y^{1/2} & y^{-1/2}x \\ 0 & y^{-1/2} \end{pmatrix} \kappa : 0 \leq x \leq d, y \geq c, u \in Z, \kappa \in K \right\}$$

contains a fundamental domain for $SL_2(\mathbf{Z})\backslash G$ if $c, d > 0$ are chosen correctly. Choose them correctly, and fix those values. Depending on the choice of Γ, there is a list of finitely many[6] $\xi_i \in SL_2(\mathbf{Z})$ which take ∞ to each of the cusps, and thus

$$\bigcup_i \xi_i \mathcal{G}_{c,d}$$

contains a fundamental domain for $\Gamma\backslash G$. Therefore, it suffices to show that

$$\sup_{g \in \mathcal{G}_{c,d}} |(\pi(\phi)f)(g)| \leq C_{\phi,\Gamma,\chi}\|f\|_{L^2}$$

for some $C_{\phi,\Gamma,\chi}$ only depending on ϕ, Γ, χ. This suffices, because for $\xi \in \{\xi_i\} \subset SL_2(\mathbf{Z})$, the function

$$F : g \mapsto f(\xi^{-1}g)$$

is in $L^2_{\text{cusp}}(\xi\Gamma\xi^{-1}\backslash G/Z, \chi)$. Applying the bound over $\mathcal{G}_{c,d}$ to this function, we have

$$\sup_{g \in \mathcal{G}_{c,d}} |(\pi(\phi)F)(g)| \leq C_{\phi,\xi\Gamma\xi^{-1},\chi}\|F\|_{L^2}.$$

The right hand side is equal to $C_{\phi,\xi\Gamma\xi^{-1},\chi}\|f\|_{L^2}$, and the left hand side is equal to $\sup_{g \in \xi\mathcal{G}_{c,d}} |(\pi(\phi)f)(g)|$. So if we can establish this inequality for the sup over $\mathcal{G}_{c,d}$ for arbitrary congruence subgroups Γ and $c, d > 0$, then we have

$$\sup_{g \in G} |(\pi(\phi)f)(g)| \leq (\max_i C_{\phi,\xi\Gamma\xi^{-1},\chi})\|f\|_{L^2}$$

as desired. For convenience, suppose that $\Gamma \cap N$ is generated by

$$\begin{pmatrix} 1 & n_0 \\ 0 & 1 \end{pmatrix}$$

where $n_0 \in \mathbf{Z}$. Now we have a canonical choice of fundamental domain for $(\Gamma \cap N)\backslash N$, namely

$$\mathcal{N}_\Gamma = \left\{ \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} : 0 \leq x < n_0 \right\}.$$

[6]since there are finitely many cusps

Using the Iwasawa decomposition, there is a fundamental domain for $(\Gamma \cap N)\backslash G/Z$ given by $\mathcal{N}_\Gamma \times A \times K$, where

$$A = \left\{ \begin{pmatrix} y & 0 \\ 0 & 1 \end{pmatrix} : y > 0 \right\}.$$

To carry out this estimate, we rewrite $\pi(\phi)$ as an integral operator and estimate the kernel.

In particular, for arbitrary $g \in G$,

$$\begin{aligned}
(\pi(\phi)f)(g) &= \int_G \phi(g^{-1}h) f(h)\, dh \\
&= \int_{(\Gamma\cap N)\backslash G/Z} \sum_{\gamma \in \Gamma \cap N} \int_Z f(\gamma h u)\phi(g^{-1}\gamma h u)\, dy = u\, dh \\
&= \int_{(\Gamma\cap N)\backslash G/Z} f(h) \sum_{\gamma \in \Gamma \cap N} \chi(\gamma) \int_Z \phi(ug^{-1}\gamma h)\, du\, dh \\
&= \int_{\mathcal{N}_\Gamma \times A \times K} f(h) \sum_{n \in \mathbf{Z}} \chi\left(\begin{pmatrix} 1 & n_0 n \\ 0 & 1 \end{pmatrix} \right) \int_Z \phi\left(g^{-1} \begin{pmatrix} 1 & n_0 n \\ 0 & 1 \end{pmatrix} h \right) du\, dh.
\end{aligned}$$

The integral over Z is necessary even in the absence of a central character, since we need it to be invariant under multiplication by elements of Z.

Since χ extends to a character of $N \cong (\mathbf{R}, +)$, we can by abuse of notation define

$$\Phi_{g,h}(t) = \chi\left(\begin{pmatrix} 1 & n_0 t \\ 0 & 1 \end{pmatrix} \right) \int_Z \phi\left(ug^{-1} \begin{pmatrix} 1 & n_0 t \\ 0 & 1 \end{pmatrix} h \right) du$$

a smooth function on $\mathbf{R}$. So we have written $\pi(\phi)$ as an integral operator

$$(\pi(\phi)f)(g) = \int_{\mathcal{N}_\Gamma \times A \times K} f(h) \sum_{n \in \mathbf{Z}} \Phi_{g,h}(n).$$

The main task is therefore to estimate

$$\sum_{n \in \mathbf{Z}} \Phi_{g,h}(n).$$

for $h \in \mathcal{N}_\Gamma \times A \times K$ and $g \in \mathcal{G}_{c,d}$.

Since ϕ is compactly supported on some compact set $\Omega \subset G$,

$$\phi_Z(g) = \int_Z \phi(ug)\, du$$

is supported on $Z\Omega$, where we can assume that $\Omega \subset SL_2(\mathbf{R})$. Since ϕ_Z is invariant under Z, we can also assume that $g, h \in SL_2(\mathbf{R})$. Therefore

$$g(Z\Omega)h^{-1} \cap N = g\Omega h^{-1}$$

since everything in $g\Omega h^{-1}$ and N has determinant 1. It follows that $\Phi_{g,h}$ is compactly supported. Hence $\Phi_{g,h} \in C_c^\infty(\mathbf{R})$, which means Poisson summation applies:

$$\sum_{n \in \mathbf{Z}} \Phi_{g,h}(n) = \sum_{n \in \mathbf{Z}} \widehat{\Phi}_{g,h}(n),$$

hence

$$(\pi(\phi)f)(g) = \int_{\mathcal{N}_\Gamma \times A \times K} f(h) \sum_{n \in \mathbf{Z}} \widehat{\Phi}_{g,h}(n).$$

Fourier transforms of smooth compactly-supported functions are nice, because they decay very fast, in fact faster than any polynomial[7]. So when $n \neq 0$ the terms in this sum can be controlled, which is what we do next. We need our bound to work over arbitrary y-coordinate for h and $y_g \geq c$, though, so we need to take this apart a little more. Writing

$$h = u_h \begin{pmatrix} y_h & x_h \\ 0 & 1 \end{pmatrix} \kappa_h, \qquad g = u_g \begin{pmatrix} y_g & x_g \\ 0 & 1 \end{pmatrix} \kappa_g$$

we have (sine χ is unitary and ϕ_Z is Z-invariant)

$$\begin{aligned}
|\widehat{\Phi}_{g,h}(n)| &= \left| \int_{-\infty}^{\infty} \chi\left(\begin{pmatrix} 1 & n_0 t \\ 0 & 1 \end{pmatrix} \right) \phi_Z \left(u_g^{-1} u_h \kappa_g^{-1} \begin{pmatrix} y_g^{-1} y_h & y_g^{-1}(x_h + n_0 t - x_g) \\ 0 & 1 \end{pmatrix} \kappa_h \right) e^{-2\pi i n t} \, dt \right| \\
&= \left| \int_{-\infty}^{\infty} \chi\left(\begin{pmatrix} 1 & n_0 t \\ 0 & 1 \end{pmatrix} \right) \phi_Z \left(\kappa_g^{-1} \begin{pmatrix} y_g^{-1} y_h & y_g^{-1}(x_h + n_0 t - x_g) \\ 0 & 1 \end{pmatrix} \kappa_h \right) e^{-2\pi i n t} \, dt \right| \\
&= \left| \int_{-\infty}^{\infty} \chi\left(\begin{pmatrix} 1 & n_0 t \\ 0 & 1 \end{pmatrix} \right) \phi_Z \left(\kappa_g^{-1} \begin{pmatrix} y_g^{-1} y_h & y_g^{-1} n_0 t \\ 0 & 1 \end{pmatrix} \kappa_h \right) e^{-2\pi i n t} \, dt \right| \\
&= |y_g| \left| \int_{-\infty}^{\infty} \chi\left(\begin{pmatrix} 1 & n_0 y_g t \\ 0 & 1 \end{pmatrix} \right) \phi_Z \left(\kappa_g^{-1} \begin{pmatrix} y_g^{-1} y_h & n_0 t \\ 0 & 1 \end{pmatrix} \kappa_h \right) e^{-2\pi i n y_g t} \, dt \right| \\
&= |y_g| \left| \int_{-\infty}^{\infty} \phi_Z \left(\kappa_g^{-1} \begin{pmatrix} y_g^{-1} y_h & n_0 t \\ 0 & 1 \end{pmatrix} \kappa_h \right) e^{-2\pi i (n - \lambda_\chi n_0) y_g t} \, dt \right|.
\end{aligned}$$

This quantity is $|y_g|$ times the Fourier transform, evaluated at $y_g(n - \lambda_\chi n_0)$, of the function

$$F_{g,h} : t \mapsto \phi_Z \left(\kappa_g^{-1} \begin{pmatrix} y_g^{-1} y_h & n_0 t \\ 0 & 1 \end{pmatrix} \kappa_h \right)$$

which is compactly supported and smooth for fixed g and h by the same argument as above. Also, $F_{g,h}$ is identically zero for $y_g^{-1} y_h$ outside of some compact set in $\mathbf{R}_{>0}$: ϕ_Z is supported on $Z\Omega$ for some compact $\Omega \subset SL_2(\mathbf{R})$, and

$$K(Z\Omega)K \cap \left\{ \begin{pmatrix} * & * \\ 0 & 1 \end{pmatrix} \right\}$$

is compact[8], which means that the set of $(y_g^{-1} y_h, t) \in \mathbf{R}_{>0} \times \mathbf{R}$ for which $F_{g,h}(t) \neq 0$ is contained in a compact set, hence the set of possible $y_g^{-1} y_h$ is contained in a compact set[9] as well. We have shown

$$|\widehat{\Phi}_{g,h}(n)| = |y_g| |\widehat{F}_{g,h}(y_g(n - \lambda_\chi n_0))|,$$

so since $F_{g,h} \in C_c^\infty(\mathbf{R})$ and only actually depends on $\kappa_g, \kappa_h, y_g^{-1} y_h$ and ϕ, Γ, for any N

$$|\widehat{\Phi}_{g,h}(n)| \ll_{\kappa_g, \kappa_h, y_g^{-1} y_h} |y_g| |y_g(n - \lambda_\chi n_0)|^{-N}$$

where the implicit constant varies continuously in $\kappa_g, \kappa_h, y_g^{-1} y_h$ (it also depends on N but we will only need one value of N). But we have shown that this constant may be chosen to be 0 when $y_g^{-1} y_h$ is outside of a compact subset $S \subset \mathbf{R}_{>0}$, which means (by taking the

[7]This is an exercise in integration by parts.

[8]It is the continuous image of $K \times \Omega \times K$ under the map that multiplies all the coordinates together and then normalizes so that the bottom-right coordinate is 1. One then intersects this with the closed condition that the bottom-left coordinate is 0, which is fine.

[9]The image of a compact set under the continuous projection map is compact

maximum of a continuous function on the compact set $K \times K \times S$) there is a constant $B_{\phi,\Gamma}$ depending only on ϕ, Γ such that

$$|\widehat{\Phi}_{g,h}(n)| \leq B_{\phi,\Gamma}|y_g|^{1-N}|n - \lambda_\chi n_0|^{-N}.$$

for all h, g (we haven't yet used any restriction to fundamental domains). As a result, choosing $N = 2$ so that the sum converges, there is a constant $B_{\phi,\Gamma.\chi}$ such that

$$\left| \sum_{\substack{n \in \mathbf{Z} \\ n - \lambda_\chi n_0 \neq 0}} \widehat{\Phi}_{g,h}(n) \right| \leq |y_g|^{-1} B_{\phi,\Gamma} \sum_{\substack{n \in \mathbf{Z} \\ n - \lambda_\chi n_0 \neq 0}} |n - \lambda_\chi n_0|^{-2} \leq B_{\phi,\Gamma,\chi}|y_g|^{-1}.$$

Since we are assuming $g \in \mathcal{G}_{c,d}$, we have $|y_g| \geq c$, so the contribution of this term to $(\pi(\phi)f)(g)$ is

$$\left| \int_{\mathcal{N}_\Gamma \times A \times K} f(h) \sum_{\substack{n \in \mathbf{Z} \\ n - \lambda_\chi n_0 \neq 0}} \widehat{\Phi}_{g,h}(n)\, dh \right| \leq c^{-1} B_{\phi,\Gamma,\chi} \int_{\mathcal{N}_\Gamma \times A \times K} |f(h)|\, dh.$$

Unfortunately, this is not enough to bound anything, since f is not compactly supported. However, we have already shown, for the purpose of controlling the constant, that the g, h such that $\Phi_{g,h}$ (and thus the same is true of $\widehat{\Phi}_{g,h}$) is nonvanishing must satisfy $y_g^{-1} y_h \in S$ for some compact set $S = [a, b] \subset \mathbf{R}_{>0}$. Since we are only considering g with $y_g \geq c$, this means that only h with

$$y_h \geq ac$$

contribute anything at all to the integral defining $(\pi(\phi)f)(g)$. So in fact we have the bound

$$\left| \int_{\mathcal{N}_\Gamma \times A \times K} f(h) \sum_{\substack{n \in \mathbf{Z} \\ n - \lambda_\chi n_0 \neq 0}} \widehat{\Phi}_{g,h}(n)\, dh \right| \leq c^{-1} B_{\phi,\Gamma,\chi} \int_{\substack{0 \leq x_h < n_0 \\ y_h \geq ac}} |f(h)|\, dh.$$

The domain of integration here can be covered by finitely many translates of a fundamental domain for $\Gamma \backslash G/Z$ (this is easily seen using the upper half-plane, and then taking products of everything with K which doesn't change the volume). So there is some positive integer N depending only on Γ such that this contribution is bounded by $c^{-1} B_{\phi,\Gamma,\chi} N \|f\|_{L^1}$. This L^1-norm is actually finite and bounded above by $\|f\|_{L^2} < \infty$, because the fundamental domain has finite volume (so it follows from Cauchy–Bunyakovski–Schwarz inquality).

There is still a possibility that the restriction to $n \in \mathbf{Z}$ such that $n - \lambda_\chi n_0 \neq 0$ has forced us to leave out a term. This is where cuspidality is used. There are two cases:

(1) $\chi|_{\Gamma \cap N}$ has finite image
(2) $\chi|_{\Gamma \cap N}$ has infinite image. This case is not relevant, because then λ_χ is irrational, so $n - \lambda_\chi n_0$ cannot vanish, and the contribution we have already estimated accounts for everything.

If $\chi|_{\Gamma \cap N}$ is trivial, then $\lambda_\chi = 0$ and this just means we have left out the $n = 0$ term. That term is

$$\int_{(\Gamma \cap N) \backslash G/Z} f(h) \widehat{\Phi}_{g,h}(0) = \int_{(\Gamma \cap N) \backslash G/Z} f(h) \int_N \phi_Z(g^{-1} n h)\, dn\, dh$$

$$\begin{aligned}
&= \int_{(\Gamma\cap N)\backslash G/Z} f(h) \int_N \phi_Z(g^{-1}n^{-1}h)\,dn\,dh \\
&= \int_{\mathcal{N}_\Gamma} \int_{(\Gamma\cap N)\backslash G/Z} f(h) \sum_{\gamma\in\Gamma\cap N} \phi_Z(g^{-1}n^{-1}\gamma^{-1}h)\,dh\,dn \\
&= \int_{\mathcal{N}_\Gamma} (\pi(\phi)f)(ng) \\
&= 0
\end{aligned}$$

since $\pi(\phi)f$ is assumed cuspidal at ∞. The same argument works as long as χ has finite image. In that case, we may replace Γ with $\ker\chi$ and repeat the same argument (from the very beginning). In real life, where χ is a Nebentypus character, $\ker\chi$ is a congruence subgroup, but we have not depended on Γ actually being a congruence subgroup anywhere in this argument. □

Proposition 3.6. *Let $\phi \in C_c^\infty(G)$. Then the convolved operator $\pi(\phi)$ is a compact operator on $\mathfrak{H}$.*

Proof. First, we consider the case of compact quotient. In that case, for $f \in L^2(\Gamma\backslash G/Z, \chi)$ and $h \in G$, we have

$$\begin{aligned}
(\pi(\phi)f)(h) &= \int_G \phi(g)(\pi(g)f)(h)\,dg \\
&= \int_G \phi(g)f(hg)\,dg \\
&= \int_G \phi(h^{-1}g)f(g)\,dg \\
&= \int_{\mathcal{F}} \sum_{\gamma\in\Gamma} \int_Z \phi(h^{-1}\gamma g u)\chi(\gamma)f(g)\,du\,dg \\
&= \int_{\mathcal{F}} K(g,h)f(g)\,dg,
\end{aligned}$$

where

$$K(g,h) = \sum_{\gamma\in\Gamma} \int_Z \phi(h^{-1}\gamma g u)\chi(\gamma)\,du$$

and $\mathcal{F}$ is a fundamental domain[10] in G for $\Gamma\backslash G/Z$. The fact that $\phi \in C_c^\infty(G)$ means that $K(g,h)$ is smooth in g and h, and $\Gamma\backslash G/Z$ being compact therefore implies that

$$K \in L^2(\mathcal{F}\times\mathcal{F}).$$

So $\pi(\phi)$ is a Hilbert–Schmidt operator on $L^2(\Gamma\backslash G/Z,\chi) \cong L^2(\mathcal{F})$ [where the isomorphism is as Hilbert spaces], and is therefore compact.

[10]These fundamental domains are already familiar from the theory of $SL_2(\mathbf{Z})$ acting on $\mathbf{H}$. Starting with a fundamental domain $\mathcal{F}_{\mathbf{H}}$ for the action of Γ on $\mathbf{H}$, the construction of which is well-known, you can just translate over to G using the Iwasawa decomposition. This is the same reason why there is no question that if $\Gamma\backslash\mathbf{H}$ is compact, so is $\Gamma\backslash G/Z$.

In the case of noncompact quotients, to prove the statement, we need to check that $\pi(\phi)$ restricts to a well-defined operator on $L^2_{\text{cusp}}(\Gamma\backslash PGL_2(\mathbf{R}), \chi)$. In other words, if

$$\int_{(\Gamma\cap N)\backslash N} f\left(\gamma n g\right)\, dn = 0$$

for all $g \in G$ and $\gamma \in SL_2(\mathbf{Z})$, then we need to check that

$$\int_{(\Gamma\cap N)\backslash N} (\pi(\phi)f)\left(\gamma n g\right)\, dn = 0.$$

This is not hard to check:

$$\begin{aligned}\int_{(\Gamma\cap N)\backslash N} (\pi(\phi)f)\left(\gamma n g\right)\, dn &= \int_{(\Gamma\cap N)\backslash N}\int_G \phi(h) f\left(\gamma n g h\right)\, dh\, dn \\ &= \int_G \phi(h)\int_{(\Gamma\cap N)\backslash N} f\left(\gamma n g h\right)\, dn\, dh \\ &= 0\end{aligned}$$

where the Fubini/Tonelli justification can be made using the fact that $(\Gamma\cap N)\backslash N$ is compact and ϕ is compactly supported.

The argument we have written down so far is not a priori a valid argument for why $\pi(\phi)|_{L^2_{\text{cusp}}}$ is compact (indeed, if it worked without modification, then there would be no need to restrict to the cuspidal part). The reason is that when $\Gamma\backslash G/Z$ is not compact, $K(\cdot,\cdot)$ is not guaranteed to be in $L^2(\mathcal{F}\times\mathcal{F})$. The additional technical observation that must be made is that there is a constant C_ϕ depending only on ϕ such that

$$\|\pi(\phi)f\|_{L^\infty} \leq C_\phi \|f\|_{L^2}$$

for all $f \in L^2_{\text{cusp}}(\Gamma\backslash G/Z, \chi)$. This is what we did in Lemma 3.5, and it is where the assumption of cuspidality is used.

There are two ways of establishing the compactness of $\pi(\phi)_{L^2_{\text{cusp}}}$ from here. The first, which I learned from Lang, involves more functional analysis. The basic point is that for any $x \in G$, Lemma 3.5 says that the linear functional

$$T_x : L^2_{\text{cusp}}(\Gamma\backslash G/Z, \chi) \to \mathbf{C}$$

given by

$$f \mapsto (\pi(\phi)f)(x)$$

is bounded. By the Riesz representation theorem, it follows that for all such x, there exists a $q_x \in L^2_{\text{cusp}}(\Gamma\backslash G/Z, \chi)$ such that $T_x(f) = \langle f, q_x\rangle$. The map $x \mapsto q_x$ from G to L^2_{cusp} has bounded image, because by Lemma 3.5

$$\|q_x\|_{L^2} = \sqrt{\langle q_x, q_x\rangle} = \sqrt{T_x(q_x)} = \sqrt{(\pi(\phi)q_x)(x)} \leq \sqrt{C_\phi \|q_x\|_{L^2}}$$

so $\|q_x\|_{L_2} \leq C_\phi$ for all x. Also, since $L^2_{\text{cusp}}(\Gamma\backslash G/Z, \chi) \cong L^2_{\text{cusp}}(\mathcal{F})$ has a countable orthonormal basis[11] $\{u_i\}$, we can write

$$q_x = \sum_{i\geq 0} g_i(x) u_i,$$

[11] $L^2(\mathcal{F})$ is separable by general theory, and L^2_{cusp} is a closed subspace and thus separable too.

where g_i is a priori just a map of sets $G \to \mathbf{C}$. The fact that $g_i(x) = \langle q_x, u_i \rangle = u_i(x)$ means that the functions $g_i(x)$ are actually measurable functions on G and, since measurability respects products and limits,

$$x \mapsto \langle q_x, q_x \rangle = \sum_i g_i(x)^2$$

is a measurable bounded function on G. Restricting it to a fundamental domain $\mathcal{F}$ for $\Gamma\backslash G/Z$, which has finite volume, and using the Hilbert space isomorphism $L^2_{\text{cusp}}(\Gamma\backslash G/Z, \chi) \cong L^2_{\text{cusp}}(\mathcal{F})$, the function $x \mapsto \langle q_x, q_x \rangle$ is therefore in $L^1(\mathcal{F})$. When $g(x,y)$ is the characteristic function of $U \times V$ the product of measurable sets in X, we have

$$\begin{aligned}\int_{\mathcal{F}}\int_{\mathcal{F}} g(x,y)\overline{q_x(y)}\,dy\,dx &= \int_{\mathcal{F}} \chi_U(x) \int_{\mathcal{F}} \chi_V(y)\overline{q_x(y)}\,dy\,dx \\ &= \int_{\mathcal{F}} \chi_U(x)\langle \chi_V, q_x\rangle\,dx \\ &= \int_{\mathcal{F}} \chi_U(x)(\pi(\phi)\chi_V)(x)\,dx \\ &< \infty\end{aligned}$$

so this iterated integral is well-defined as long as $g(x,y)$ is a step function on $\mathcal{F}\times\mathcal{F}$. By the Cauchy–Bunyakovsky–Schwarz inequality[12] and Lemma 3.5, we have (still only as long as g is a step function, which is the only case in which we have established the left hand side is a real thing)

$$\left|\int_{\mathcal{F}}\int_{\mathcal{F}} g(x,y)\overline{q_x(y)}\,dy\,dx\right| \leq \|g\|_{L^2}\sqrt{\int_{\mathcal{F}}\int_{\mathcal{F}} |q_x(y)|^2\,dy\,dx}$$

where the right hand side is well-defined from our previous observation that $x \mapsto \langle g_x, g_x\rangle$ is in $L^1(\mathcal{F})$. So the linear map

$$L^2(\mathcal{F}\times\mathcal{F}) \to \mathbf{C}$$

densely defined on the step functions and given by

$$g \mapsto \int_{\mathcal{F}}\int_{\mathcal{F}} g(x,y)\overline{q_x(y)}\,dy\,dx$$

is continuous where it is defined and is therefore extends to all of $L^2(\mathcal{F}\times\mathcal{F})$. By the Riesz representation theorem, there exists a $Q(\cdot,\cdot) \in L^2(\mathcal{F}\times\mathcal{F})$ such that

$$\int_{\mathcal{F}}\int_{\mathcal{F}} g(x,y)\overline{q_x(y)}\,dy\,dx = \int_{\mathcal{F}}\int_{\mathcal{F}} g(x,y)\overline{Q(x,y)}\,dy\,dx$$

for all step functions g. If we choose the step function g correctly, we see that this implies that $q_x = Q(x,-)$ in $L^2_{\text{cusp}}(\mathcal{F})$ for almost all $x \in \mathcal{F}$. Therefore, we really can write

$$(\pi(\phi)f)(x) = T_x f = \int_{\mathcal{F}} f(y)\overline{Q(x,y)}\,dy$$

for almost all $x \in \mathcal{F}$. Since Q is by definition an element of $L^2(\mathcal{F}\times\mathcal{F})$, this means that $\pi(\phi)$ is Hilbert–Schmidt and therefore compact.

Note that the only time the assumption of cuspidality was used was to establish the estimate $\|\pi(\phi)f\|_{L^\infty} \leq C_\phi \|f\|_{L^2}$, and this was only used to show that the evaluation-at-x

[12]technically speaking, one has to repeat the proof to deduce what follows.

functional was bounded. The rest of the proof is not dependent on the specifics of the situation at all, and is a general technique in functional analysis. □

Once we have our hands on these compact operators, we may prove that $\mathfrak{H}$ decomposes as a finite-multiplicity Hilbert space direct sum of irreducible subspaces. The proof of the decomposition from the compactness of these operators is not long, but is more technically involved than most proofs of this type. It is reproduced here from [3, Theorem 2.3.3], but according to [14], the argument originally appeared in [6]. I have tried to elucidate a little bit more than the standard references what the motivation behind the argument is.

Theorem 3.7 (Gelfand–Graev–Pjateckii-Shapiro, 1966)**.** *Let $(\pi, \mathfrak{H})$ be the right regular representation of G on $\mathfrak{H} = L^2_{\text{cusp}}(\Gamma\backslash G/Z, \chi)$. Then we have a discrete decomposition of $\mathfrak{H}$ as a Hilbert space orthogonal direct sum of irreducible representations of G*

$$\mathfrak{H} = \bigoplus_i \pi_i^{m_i}.$$

Proof. The basic technique of the proof is the same as usual: let $\mathfrak{H}'$ be a nonzero closed subspace of $\mathfrak{H}$ which is closed under the action of G. We will show that $\mathfrak{H}'$ contains a nontrivial irreducible representation of G, which will show by Zorn's lemma[13] (via the fact that π is unitary) that the desired decomposition exists (though not a priori with finite multiplicity).

There exists a choice of $\phi \in C_c^\infty(G)$ such that $\pi(\phi)$ is not only compact but also self-adjoint on $\mathfrak{H}$. Such a ϕ just needs to have

$$\phi(g^{-1}) = \overline{\phi(g)}$$

for all $g \in G$, since then (again using the fact that π is unitary)

$$\begin{aligned}
\langle \pi(\phi)v, w\rangle &= \int_G \phi(g)\langle \pi(g)v, w\rangle \, dg \\
&= \int_G \phi(g)\langle v, \pi(g^{-1})w\rangle \, dg \\
&= \int_G \overline{\phi(g)}\langle v, \pi(g)w\rangle \, dg \\
&= \langle v, \pi(\phi)w\rangle.
\end{aligned}$$

A ϕ satisfying this condition is easily cooked up using the usual theory of bump functions on manifolds, for instance, by taking a bump function ϕ_0 supported on a compact set $U \subset G$ and then letting

$$\phi(g) = \phi_0(g) + \overline{\phi_0(g^{-1})}.$$

The fact that the multiplicities are finite does not lie deeper than the rest of the statement: by the spectral theorem for compact self-adjoint operators, $\pi(\phi)$ diagonalizes and has eigenvalues going to zero, so each eigenspace with nonzero eigenvalue is finite-dimensional. Also, from its definition, $\pi(\phi)$ restricts to a well-defined G-intertwining operator on each irreducible component π_i, where it must act as a scalar by Schur's lemma. This scalar only

[13]By Zorn's lemma, there is a maximal set of mutually orthogonal closed subrepresentations of π. Since π is unitary, the orthogonal complement of the Hilbert space direct sum of all those subrepresentations is also a closed subrepresentation, and showing that it has a nontrivial irreducible closed subrepresentation contradicts the maximality statement from Zorn's lemma; it follows that the orthogonal complement is zero, and thus the desired orthogonal decomposition into irreducible Hilbert space representations exists.

depends on i, so the fact that the nonzero eigenvalues have finite multiplicity implies $m_i < \infty$ as well, as long as $\pi(\phi)$ doesn't act by zero on π_i. This could technically happen, but can be avoided easily, by choosing some nonzero $f \in \pi_i$ and then choosing[14] $\phi \in C_c^\infty(G)$ such that $|\pi(\phi)f - f|$ is small enough that $\pi(\phi)f$ cannot vanish.

Now for the construction of the nontrivial irreducible subspace of $\mathfrak{H}'$. By assumption, there exists some $0 \neq f \in \mathfrak{H}'$, so by choosing ϕ such that $\pi(\phi)$ is compact and self-adjoint and $\pi(\phi)f \neq 0$ (which we have already shown how to do), the operator

$$\pi(\phi)|_{\mathfrak{H}'}$$

is also compact, nonzero, and self-adjoint. By the spectral theorem for compact self-adjoint operators, it therefore has some nonzero eigenvalue λ with finite dimensional eigenspace $V_\lambda \subset \mathfrak{H}'$. It is true from the definition of $\pi(\phi)$ that $\pi(\phi)$ has a well-defined restriction to any subrepresentation, but it is *not* true that V_λ is G-invariant: the action of G does not actually commute with $\pi(\phi)$. Still, V_λ is useful in the construction, because $\pi(\phi)$ is supposed to restrict to each $\pi_i \subset \mathfrak{H}'$ to something diagonalizable, where the λ-eigenspace is $V_\lambda \cap \pi_i$. Motivated by this, the trick is to take $L_0 \subset V_\lambda$ to be the minimal nonzero subspace of V_λ of the form $V_\lambda \cap \mathfrak{H}_0'$ where $\mathfrak{H}_0'$ is a closed subrepresentation of $\mathfrak{H}'$ (this is well-defined because V_λ is finite-dimensional). The minimal $\mathfrak{H}_0'$ such that $L_0 = V_\lambda \cap \mathfrak{H}_0'$ ought to be irreducible and nonzero. To construct it, just take the intersection of all such $\mathfrak{H}_0'$:

$$\mathfrak{V} := \bigcap_{\substack{\mathfrak{H}_0' \subset \mathfrak{H}' \\ L_0 = V_\lambda \cap \mathfrak{H}_0'}} \mathfrak{H}_0'.$$

Since $0 \neq L_0 \subset \mathfrak{V}$, the definition guarantees that $\mathfrak{V} \neq 0$, and it remains to show that $\mathfrak{V}$ is irreducible. It is irreducible because of the minimal nature of its construction: if it had a proper subrepresentation $\mathfrak{V}_1$, then $\mathfrak{V}_1 \cap V_\lambda$ has to be properly contained in L_0 by the minimality of $\mathfrak{V}$. Unless $\mathfrak{V}_1 = 0$, this contradicts the minimality of L_0. So we just need to show that $\mathfrak{V}_1$ can be chosen so that $\mathfrak{V}_1 \cap V_\lambda \neq 0$. Since π is unitary, we actually have $\mathfrak{V} = \mathfrak{V}_1 \oplus \mathfrak{V}_2$ for closed subrepresentations $\mathfrak{V}_i$. Taking intersections with V_λ, we have

$$L_0 = \mathfrak{V} \cap V_\lambda = (\mathfrak{V}_1 \oplus \mathfrak{V}_2) \cap V_\lambda = (\mathfrak{V}_1 \cap V_\lambda) \oplus (\mathfrak{V}_2 \cap V_\lambda).$$

The key point is the last equality, which is because $f_1 + f_2 \in \mathfrak{V} \cap V_\lambda$, for $f_i \in \mathfrak{V}_i$, means that $\pi(\phi)f_1 + \pi(\phi)f_2 = \lambda f_1 + \lambda f_2$. Since the $\mathfrak{V}_i$ are acted on by $\pi(\phi)$, this implies $f_i \in V_\lambda$ too, as desired. Since $L_0 \neq 0$, at least one of $\mathfrak{V}_i \cap V_\lambda$ is nonzero, so we are done. □

The reason why this technical analysis-heavy argument (which uses in a crucial way the existence of these compact operators and thus the fact that we are restricting to the cuspidal part) is necessary is that one cannot simply construct an irreducible subrepresentation by taking the G-span of a nonzero vector: the resulting subspace is not necessarily closed. So one must take the closure to obtain a bona-fide Hilbert space subrepresentation, but this new object is not necessarily irreducible. One needs to show that this closure has the Artinian descending chain condition. The canonical way to do this is to intersect with V_λ and use finite-dimensionality of V_λ, which is essentially the same strategy as the version of the proof we have written down.

[14]To do that, just use the fact that π is continuous, so there exists a neighborhood U of the identity in G such that $|\pi(g)f - f| < \epsilon$ for all $g \in U$. We may take ϕ compactly supported in U such that $\int_G \phi = 1$, which is enough.

Remark 3.8. The results of this section hold with essentially the same proofs if we add also a choice of central character $\omega : Z \to S^1$, according to which f must also tranform upon translation by elements of Z. The more difficult generalization to arbitrary reductive Lie groups has also been done: see [16]. The basic ideas in those proofs are the same as the ones for $GL(2)$.

3.3. **Explicit decomposition into irreducible components.** Now that we know $\mathfrak{H}$ splits as a Hilbert space direct sum of irreducible unitary representations of G, we ought to be very interested in the following question.

Question 3.9. Which isomorphism classes of irreducible representations appear in the decomposition of $\mathfrak{H}$?

To deal with the general problem of classifying irreducible unitary representations of reductive Lie groups, Harish-Chandra [7, 8, 9] and others developed the theory of $(\mathfrak{g}, K)$-modules and infinitesimal equivalence.

Since the right regular representation is unitary, we may restrict our attention to that case. It turns out to also be useful to restrict to the category of *admissible* representations.

Definition 3.10. A Hilbert space representation V of G is *admissible* if, in its decomposition into irreducible representations of a maximal compact K, each irreducible representation has finite multiplicity.

This restriction does no harm, because of

Theorem 3.11 (Harish-Chandra, 1953, 1954)**.** *Every irreducible unitary representation of a connected reductive Lie group G is admissible.*

Proof. See [12, Theorem 8.1]. □

Harish-Chandra first proved this theorem in 1953 [7], and later improved it with an explicit bound on the dimension of $\mathfrak{H}(k)$ [9]. In our case, we only really need the special case

Theorem 3.12. *Every irreducible subrepresentation of $\mathfrak{H}$ is admissible, where $\mathfrak{H}$ is the right regular representation of $G = GL_2(\mathbf{R})^+$ on $L^2_{\text{cusp}}(\Gamma\backslash G/Z, \chi)$.*

Proof. A proof for $GL(2)$ can be found in Bump [3, Theorem 2.4.3]. There are two main inputs, only one of which we have justified:

(1) The compactness of the integral operators $\pi(\phi)$.
(2) The commutativity of the Hecke algebra[15] $\mathcal{H} = C_c^\infty(K\backslash G/K, \sigma)$ for any character $\sigma : K \to S^1$.

The first step is to prove that the K-isotypic subspace $\mathfrak{H}(k)$ is irreducible as an $\mathcal{H}_k$-module, where $\mathcal{H}_k$ is the Hecke algebra $C_c^\infty(K\backslash G/K, e^{ik\theta})$. This is proved directly. Once this is done, points (1) and (2) come into play. By the usual technique, we may select a $\phi \in C_c^\infty(G)$ such that $\pi(\phi)$ is a compact self-adjoint operator on $\mathfrak{H}$ and is nonzero on $\mathfrak{H}(k)$. It may further be chosen to actually be in $\mathcal{H}_k$ By the spectral theorem, $\pi(\phi)$ has a nonzero eigenvector in $\mathfrak{H}(k)$, with nonzero eigenvector and finite-dimensional eigenspace. By the commutativity

[15]This Hecke algebra is the set of compactly-supported smooth functions $f : G \to \mathbf{C}$ satisfying $f(h_1 g h_2) = \sigma(h_1 h_2) f(g)$. The ring structure is given by convolution. One shows quite directly in the case of $GL(2)$ that this algebra is commutative (see [3, Proposition 2.2.8]). Without the character, it is true by a technique due to Gelfand (see [3, Theorem 2.2.3] and [4, Ch. 47], on the general theory of Gelfand pairs).

of $\mathcal{H}_k$ and the fact that $\pi(\phi_1)\pi(\phi_2) = \pi(\phi_1 * \phi_2)$, it follows that this eigenspace of $\pi(\phi)$ is $\mathcal{H}_k$-invariant. Since $\mathfrak{H}(k)$ is supposed to be irreducible as an $\mathcal{H}_k$-module, this means that everything in $\mathcal{H}_k$ acts as a scalar on $\mathfrak{H}(k)$, and that $\mathfrak{H}(k)$ is finite-dimensional, as desired.

In fact, $\mathfrak{H}(k)$ cannot be more than 1-dimensional, since we may repeat the same argument on the span of a single eigenvector. □

So the unitary admissible irreducible representations of $GL_2(\mathbf{R})^+$ are all we need to classify if we want to understand $\mathfrak{H}$. One extremely useful way to carry out such a classification is to use

Definition 3.13. Let G be a Lie group with Lie algebra $\mathfrak{g}$ and maximal compact subgroup K. A $(\mathfrak{g}, K)$-*module* is a Hilbert space V with a Lie algebra action of $\mathfrak{g}$ and a Lie group action of K. It must be K-finite, in the sense that V is the algebraic direct sum of finite-dimensional K-subrepresentations (i.e. every element of V is K-finite), and satisfy a compatibility condition between the actions of K and $\mathfrak{g}$:

(1) For $X \in \mathfrak{k} \subset \mathfrak{g}$ the Lie algebra of K and $f \in V$,
$$X \cdot f = \frac{d}{dt}\bigg|_{t=0} \exp(tX) \cdot f.$$
(2) $g \cdot X \cdot g^{-1} \cdot f = (\mathrm{Ad(g)}X) \cdot f$ for $g \in K$.

Given a Hilbert space representation of a Lie group G, one obtains a $(\mathfrak{g}, K)$-module by taking the K-finite[16] vectors in that representation. It is a theorem that the K-finite vectors are smooth and dense [3].

Definition 3.14. Two admissible Hilbert space representations of a Lie group G are *infinitesimally equivalent* if their corresponding $(\mathfrak{g}, K)$-modules are isomorphic.

Langlands [15] proved the *Langlands classification*, a classification up to infinitesimal equivalence of irreducible representations, for a broad class of reductive Lie groups.

Infinitesimal equivalence in the category of admissible representations is useful for our purposes because of the following two-step strategy for classifying irreducible admissible unitary representations of a Lie group.

step 1: Determine the isomorphism classes of irreducible $(\mathfrak{g}, K)$-modules.
step 2: Determine which of those isomorphism classes contain a $(\mathfrak{g}, K)$-module coming from a unitary representation of G.

This strategy holds water in a very general setting because of the following results.

Theorem 3.15 (Harish-Chandra, 1953)**.** *Let G be an arbitrary connected reductive Lie group. Suppose two admissible unitary irreducible representations $(\pi, \mathfrak{H})$, $(\pi', \mathfrak{H}')$ are infinitesimally equivalent. Then they are isomorphic.*

Proof. The fact that π, π' are unitary is used via the fact that $\mathfrak{g}$ acts via skew-adjoint operators. This proof is essentially the same as [12, Corollary 9.2]. The special case for $GL_2(\mathbf{R})^+$ may be found in [3, Theorem 2.6.6].

The main part of the proof is in seeing that one can renormalize the isomorphism of $(\mathfrak{g}, K)$-modules $\varphi : \mathfrak{H}_{K-\mathrm{fin}} \to \mathfrak{H}'_{K-\mathrm{fin}}$ to force it to be an isometry (this uses the fact that these are unitary representations). To do this, renormalize φ such that $|\varphi(f)| = 1$ for some

[16] $v \in \pi$ is K-finite if the space spanned by $\pi(g)v$ for $g \in K$ is finite-dimensional

$0 \neq f \in \mathfrak{H}_{K-\text{fin}}(\sigma)$ with length 1. Since $(\mathfrak{g}, K)$-modules are orthogonal algebraic direct sums of finite-dimensional K-isotypic subspaces, and π is unitary, there exists an abstract linear operator

$$B : \mathfrak{H}_{K-\text{fin}} \to \mathfrak{H}_{K-\text{fin}}$$

acting independently on the K-isotypic components such that

$$\langle \varphi v, \varphi w \rangle = \langle Bv, w \rangle$$

for all $v, w \in \mathfrak{H}_{K-\text{fin}}$ (this is the usual trick from finite-dimensional linear algebra on each finite-dimensional K-isotypic orthogonal component). If we can show that B acts by a scalar λ_B, we are done, since then

$$|\varphi(v)| = \sqrt{\langle \varphi(v), \varphi(v) \rangle} = \sqrt{\lambda_B}|v|,$$

and $\lambda_B = 1$ from setting $v = f$.

To prove that B acts by a scalar λ_B, it suffices to show that B commutes with the action of $\mathfrak{g}$: once this is established, we know B commutes with $\mathfrak{k} \subset \mathfrak{g}$ and thus with K; this means that B restricts to K-intertwining maps on the finite-dimensional K-isotypic components. Choose an arbitrary such component $\mathfrak{H}(\sigma)$. Since it is finite-dimensional, B has an eigenvector in $\mathfrak{H}(\sigma)$, so $B - \lambda : \mathfrak{H}_{K-\text{fin}} \to \mathfrak{H}_{K-\text{fin}}$ has nontrivial kernel and is $\mathfrak{g}$-intertwining. Since $\mathfrak{H}_{K-\text{fin}}$ is irreducible as a $\mathfrak{g}$-module, it follows that $B = \lambda =: \lambda_B$, as desired.

Finally, the fact that B commutes with the action of $\mathfrak{g}$ is because

$$\begin{aligned}
\langle B\pi(X)v, w \rangle &= \langle \varphi(\pi(X)v), \varphi(w) \rangle \\
&= \langle \varphi(\pi(X)v), \varphi(w) \rangle \\
&= \langle \pi(X)\varphi(v), \varphi(w) \rangle \\
&= -\langle \varphi(v), \varphi(\pi(X)w) \rangle \\
&= -\langle Bv, \pi(X)w \rangle \\
&= \langle \pi(X)Bv, w \rangle
\end{aligned}$$

for any $v, w \in \mathfrak{H}_{K-\text{fin}}$ and $X \in \mathfrak{g}$.

So (and this is the key point) the abstract isomorphism φ of K-finite vectors can be chosen to be an isometry, and thus can be extended to a $\mathfrak{g}$-intertwining isometry $\mathfrak{H} \to \mathfrak{H}'$ because the K-finite vectors are dense.

For the second part, the fact that the extension of φ is G-intertwining comes from the fact that it is $\mathfrak{g}$-intertwining, using the fact that exp is surjective and

$$\pi(\exp(X))v = \sum_{n=0}^{\infty} \frac{1}{n!} X^n v.$$

This part doesn't even need the unitary assumption (and in fact we already used this argument to deduce that $\mathfrak{g}$-intertwining operators are K-intertwining). □

So a given infinitesimal equivalence class can only have at most one unitary representation in it up to isomorphism. In other words, the map

$$\frac{\{\text{Admissible unitary representations of } G\}}{\text{isomorphism}} \to \frac{\{\text{Admissible unitary representations of } G\}}{\text{infinitesimal equivalence}}$$

is a bijection. This fact is not true without the inclusion of the word "unitary."

In fact, we may work directly with admissible $(\mathfrak{g}, K)$ modules because of the following general result.

Theorem 3.16 (Harish-Chandra, ??)**.** *Let G be a connected reductive Lie group. Then every admissible $(\mathfrak{g}, K)$-module is isomorphic to $\mathfrak{H}_{K-fin}$ for some admissible representation $\mathfrak{H}$ of G.*

Apology. I have been assured that this is true, but I do not know a reference. In the special case of $GL_2(\mathbf{R})^+$, this is done in [3, §2.6] via parabolic induction. The same method surely works in the general context. □

Corollary 3.17. *The canonical map*

$$\frac{\{\textit{Admissible (unitary) representations of } G\}}{\textit{infinitesimal equivalence}} \to \frac{\textit{Admissible (unitary)}(\mathfrak{g}, K) - \textit{modules}}{\textit{infinitesimal equivalence}}$$

is a bijection.

The reality is that for the purposes of studying the decomposition of the right regular representation of G on $L^2_{\text{cusp}}(\Gamma\backslash G/Z, \chi)$, Theorem 3.16 is not necessary in practice, because once we classify the $(\mathfrak{g}, K)$-modules, we just need to ask how they can show up in this representation. So the only thing on which we logically depend is Theorem 3.15, which we have proved in full generality.

Now we restrict our attention exclusively to $G = GL_2(\mathbf{R})^+$. Step 1 of the strategy is a purely algebraic exercise. It is an "exercise" precisely because of the way that we have chosen K and $\mathfrak{h} \subset G$ to be compatible, meaning that the $(\mathfrak{g}, K)$-module structure is determined completely by how the four generating matrices, two of which are already completely determined, act.

Theorem 3.18. *An irreducible admissible $(\mathfrak{g}, K)$-module for $GL_2(\mathbf{R})^+$ is determined by the following information:*

1. *The $Z_{\mathfrak{g}}$-eigenvalue and the Δ-eigenvalue (this is the same as "the action of $\mathfrak{Z} = \mathbf{C}[\Delta, Z_{\mathfrak{g}}] = Z(U(\mathfrak{g}))$")*
2. *The set of characters of K that appear in the decomposition as an algebraic direct sum $V = \bigoplus V(k)$ (the "K-type").*

In particular, $\dim_{\mathbf{C}} V(k) \leq 1$ *for all* $k \in \mathbf{Z}$.

Given what we know about $\mathfrak{g}$ and its root space decomposition, it's clear what the possible K-types are: either the set of all $k \geq k_0$ or $k \leq k_0$ of some given parity $\epsilon \in \mathbf{Z}/2\mathbf{Z}$ (these are called the *discrete series representations*, and they appear in $L^2(\Gamma\backslash G/Z, \chi)$ in the first case as being generated by $y^{k_0/2} f$, where f is a holomorphic modular form of weight k_0, or in the second case as being generated by $y^{-k_0/2} f$, where f is an antiholomorphic modular form of weight k_0), the set of all $k \in \mathbf{Z}$ of parity ϵ (these are called *principal series representations* and are generated by appropriate Maass forms), or the set of all $k \in \mathbf{Z}$ of parity ϵ bounded above by k_0 and below by $-k_0$. Things must look this way because of the relations

$$\hat{R}\hat{L} = \frac{k}{2}\left(1 - \frac{k}{2}\right) - \Delta$$

and

$$\hat{L}\hat{R} = -\frac{k}{2}\left(1 - \frac{k}{2}\right) - \Delta.$$

This further implies that things can only die if the Δ-eigenvalue is of the form $\pm\frac{k}{2}\left(1-\frac{k}{2}\right)$ for some $k \in \mathbf{Z}$.

It is easy to pin down how many times the discrete series and principal series show up in $L^2(\Gamma\backslash G/Z, \chi)$, even without constructing examples of arbitrary $(\mathfrak{g}, K)$-modules. The principal series inside there are in bijection with the linearly independent Maass forms which are not eventually killed by raising or lowering. The pairs of discrete series are in bijection with the holomorphic modular forms. In fact, for arbitrary eigenvalues of Δ and $Z_{\mathfrak{g}}$, not just the ones that occur for Maass forms, it is possible to construct these $(\mathfrak{g}, K)$-modules via parabolic induction, and to find unitary examples, but this is not relevant.

What about the finite dimensional ones? The next thing we do is construct them explicitly, and show more abstractly that they cannot be unitary unless they are one-dimensional.

Since Z and Δ are in the center of $U(\mathfrak{g})$, by Schur's lemma, they act by scalars on any irreducible $(\mathfrak{g}, K)$-module. Call them μ (the Z-eigenvalue) and λ (the Δ-eigenvalue). The classification of $(\mathfrak{g}, K)$-modules for G says that a finite-dimensional irreducible $(\mathfrak{g}, K)$-module V is determined by the following pieces of data:

(1) its dimension
(2) μ

If the dimension is d, then (by the classification) in fact we must have

$$V = \bigoplus_{\substack{n \equiv k \bmod 2 \\ -k < n < k}} V(n)$$

where $k = d + 1$, and the Δ-eigenvalue must be

$$\lambda = \frac{k}{2}\left(1 - \frac{k}{2}\right).$$

It turns out that given any choice of d and μ, we can construct a $(\mathfrak{g}, K)$-module with those parameters. One way to do this is by induction from the Borel subgroup of upper-triangular matricies. At least this is the only way I know how to construct the discrete series representations for arbitrary[17] λ and μ. But for the finite-dimensional $(\mathfrak{g}, K)$-modules, there is a slightly more concrete way to construct them directly. The first observation is that as soon as we have constructed a $(\mathfrak{g}, K)$-module (π, V) of dimension $d = k - 1$ and Z-eigenvalue μ, we can immediately construct $(\mathfrak{g}, K)$-modules of the same dimension and arbitrary Z-eigenvalue. That is because of the one-dimensional representation of G, called $(\det)^r$ given by

$$gv = (\det g)^r \cdot v.$$

This is well-defined whenever r is an arbitrary complex number, since $\det g$ is always a positive real number. We will view $(\det)^r$ as its induced $(\mathfrak{g}, K)$-module.

Lemma 3.19. *Suppose (π, V) is an irreducible $(\mathfrak{g}, K)$-module with Δ-eigenvalue λ and Z-eigenvalue μ. Then $(\det)^r \otimes \pi$ is also an irreducible $(\mathfrak{g}, K)$-module, and has Δ-eigenvalue λ and Z-eigenvalue $\mu + 2r$.*

Proof. This boils down mostly to thinking about what $(\det)^r \otimes \pi$ is as a $(\mathfrak{g}, K)$-module. The elements of K all have determinant 1, so the action of K is the same as that on V. In our situation, the $(\mathfrak{g}, K)$-module V will always be induced by a representation of G. But we

[17] If λ is the Laplace eigenvalue of a Maass form, then of course you can construct the representation that way. But those eigenvalues only account for a discrete subset of $\mathbf{R}$.

might as well think about this tensor product abstractly in the category of $(\mathfrak{g}, K)$-modules. In this category, I think one is supposed to define the tensor product of $(\mathfrak{g}, K)$-modules using the product rule. For this reason, in the special case where ψ has dimension 1 and π is arbitrary, we define the Lie algebra action by

$$(\psi \otimes \pi)(X)v = (\psi(X) + \pi(X))v.$$

So for the action of $\mathfrak{g}$ in the $(\mathfrak{g}, K)$-module $\pi \otimes (\det)^r$, we may compute explicitly

$$\begin{aligned} Zv &= \mu v + \frac{d}{dt}\Big|_{t=0} \det(I + tZ)^r v \\ &= \mu v + \frac{d}{dt}\Big|_{t=0} (1+t)^{2r} v \\ &= (\mu + 2r)v \end{aligned}$$

which shows at least that Z still acts by a scalar, only this time equal to $\mu + 2r$. Also,

$$\begin{aligned} Rv &= \pi(R)v + \frac{d}{dt}\Big|_{t=0} \det(I + tR)^r v \\ &= \pi(R)v \end{aligned}$$

and similarly for L, H, and all elements of K. So, as claimed, nothing except the scalar by which Z acts is changed (in particular the action of Δ is determined by that of H, R, L). Also, by looking at the decomposition of $\pi \otimes (\det)^r$ into K-isotypic subspaces, we see that the raising and lowering operators behave the same way, and since Z still acts by a scalar, $\pi \otimes (\det)^r$ is still irreducible. □

Remark 3.20. I haven't thought about how generally Lemma 3.19 can be made to work. I think it should work for GL_n, but I'm not sure about general reductive Lie groups.

So it suffices to construct a single finite-dimensional $(\mathfrak{g}, K)$-module of every dimension, and we will see that every possible (according to the classification) equivalence class actually occurs. To do this construction, just let V be the two-dimensional $\mathbf{C}$-vector space spanned by the symbols X and Y, with the standard action of $GL(V) \cong GL_2(\mathbf{C})$. Consider the symmetric $(d-1)$-th power

$$\mathrm{Sym}^{d-1} V = V^{\otimes(d-1)} / \langle v_1 \otimes \cdots \otimes v_{d-1} - v_{\sigma(1)} \otimes \cdots \otimes v_{\sigma(d-1)} \rangle_{\sigma \in S_{d-1}}.$$

In this situation, there is no reason to view $\mathrm{Sym}^{d-1} V$ as being anything other than the d-dimensional $\mathbf{C}$-vector space of homogeneous degree-d polynomials in the formal variables X and Y, with the standard action of $GL_2(\mathbf{C})$. This is finite-dimensional, so all the vectors are smooth and K-finite. So the associated $(\mathfrak{g}, K)$-module is the same as a vector space, and has a weight-space decomposition

$$\mathrm{Sym}^{d-1} V = \mathbf{C} \cdot \pi(\mathcal{C})^{-1} X^{d-1} \oplus \mathbf{C} \cdot \pi(\mathcal{C})^{-1} X^{d-1} Y \oplus \cdots \oplus \mathbf{C} \cdot \pi(\mathcal{C})^{-1} Y^{d-1}.$$

To check that this is indeed a weight space decomposition, we just need to check that $X^r Y^{d-1-r}$ is an H-eigenvector (with eigenvalue depending on r). In fact, we might as well manually write down the action of

$$\mathfrak{g} = \mathbf{C} \cdot Z + \mathbf{C} \cdot H + \mathbf{C} \cdot L + \mathbf{C} \cdot R.$$

In particular,

$$Z \cdot X^r Y^{d-1-r} = \frac{d}{dt}\Big|_{t=0} (1+t)^{d-1} X^r Y^{d-1-r} = (d-1) X^r Y^{d-1-r},$$

so as expected Z acts by a constant, which happens to be $d-1$. Similarly,

$$\begin{aligned} H \cdot X^r Y^{d-1-r} &= \frac{d}{dt}\bigg|_{t=0} \begin{pmatrix} 1+t & 0 \\ 0 & 1-t \end{pmatrix} X^r Y^{d-1-r} \\ &= \frac{d}{dt}\bigg|_{t=0} (1+t)^r (1-t)^{d-1-r} X^r Y^{d-1-r} \\ &= (2r-d+1) X^r Y^{d-1-r} \end{aligned}$$

so this is, as claimed, an eigenspace decomposition for the action of H (and we also see that the weights range from $-d+1$ to $d-1$, taking on only values which have the same parity as $d-1$). The theory of root systems already shows from this that the raising and lowering operators are supposed to go between the weight spaces of weights differing by 2 without killing anything (unless one is lowering a lowest-weight or raising a highest-weight vector), which shows already that this is an irreducible $(\mathfrak{g}, K)$-module of desired dimension. Just for kicks, we might as well compute the action:

$$\begin{aligned} R \cdot X^r Y^{d-1-r} &= \frac{d}{dt}\bigg|_{t=0} \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} X^r Y^{d-1-r} \\ &= \frac{d}{dt}\bigg|_{t=0} X^r (tX+Y)^{d-1-r} \\ &= (d-1-r) X^{r+1} Y^{d-1-(r+1)} \end{aligned}$$

and

$$\begin{aligned} L \cdot X^r Y^{d-1-r} &= \frac{d}{dt}\bigg|_{t=0} \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix} X^r Y^{d-1-r} \\ &= \frac{d}{dt}\bigg|_{t=0} (X+tY)^r Y^{d-1-r} \\ &= r X^{r-1} Y^{d-1-(r-1)}. \end{aligned}$$

Notice that this is consistent with the fact that R should vanish on heighest-weight (i.e. $r = d-1$) and L should vanish on lowest-weight (i.e. $r = 0$). Recall also that general computations already tell us that the Δ-eigenvalue is

$$\lambda = \frac{k}{2}\left(1 - \frac{k}{2}\right)$$

where $k = d+1$. So we have shown

Theorem 3.21. *The equivalence class of irreducible $(\mathfrak{g}, K)$-modules of dimension d and Z-eigenvalue μ contains*

$$\mathrm{Sym}^{d-1}(\mathbf{C}^2) \otimes (\det)^{\frac{\mu-d+1}{2}}.$$

However, we should be aware that few of these, if any, can appear in the right regular representation, because they are not unitarizable. In particular, for $d > 1$, none of these finite-dimensional $(\mathfrak{g}, K)$-modules come from a unitary representation of G (and thus cannot contribute to the regular representation, since that representation is unitary).

Lemma 3.22. *No irreducible unitary representation of $GL_n(\mathbf{R})^+$ has dimension > 1.*

Bump's proof. Such a representation is a continuous open group homomorphism

$$\pi : GL_n(\mathbf{R})^+ \to U_n(\mathbf{R})$$

with compact image. It restricts to another map

$$SL_n(\mathbf{R}) \to U_n(\mathbf{R})$$

with the same properties, so that $SL_n(\mathbf{R})/\ker(\pi|_{SL_n(\mathbf{R})})$ is compact. It is a general result in group theory that $SL_n(\mathbf{R})$ is simple (as a group, not just a Lie group; I don't know how to prove this). So π is trivial on $SL_n(\mathbf{R})$. It follows that π must factor through the determinant map, since that map provides a splitting

$$GL_n(\mathbf{R})^+ \cong SL_n(\mathbf{R}) \times \mathbf{R}^\times_{>0}.$$

The only unitary irreducible representations of the (abelian!) group $\mathbf{R}^\times_{>0}$ are 1-dimensional, so we are done. □

Bump's proof is incomplete, however: it is not true a priori that the image is compact (homomorphisms of Lie groups, even compact Lie groups, are not necessarily closed). Later I will add a corrected version.

The upshot is that only three kinds of irreducible subrepresentations can show up in the right regular representation we are interested in:

(1) The 1-dimensional irreducible representation on which g acts by an imaginary power of $\det g$. This occurs exactly once, and with multiplicity 1. The imaginary power depends on the central character, so since we have been assuming there is trivial central character, this is just the subspace of functions which are equal to a constant function almost everywhere on a fundamental domain for $\Gamma\backslash G$.
(2) The principal series representations, generated by a Maass form which never vanishes upon repeated applications of the Shimura–Maass weight-raising and lowering operators.
(3) The (limits of) discrete series representations, generated by the holomorphic modular forms and their conjugates.

Items (1) and (3) are much easier in terms of representation theory than (2). (1) is clearly not worth saying much more about – it just accounts for essentially constant functions. The discrete series representation π coming from a modular form of weight k is easy for us to understand, because the highest weight vector in π is a modular form of weight k, and is killed by the weight-lowering operator, and therefore also by RL. But

$$RL = -\Delta + \frac{k}{2}\left(1 - \frac{k}{2}\right)$$

in $U(\mathfrak{g})$ (this is for example by explicit computation with the Casimir element), which means that the action of Δ is by $\frac{k}{2}\left(1-\frac{k}{2}\right)$. The $Z_{\mathfrak{g}}$-eigenvalue is 1 since we have chosen the convention of trivial central character, and the K-type is determined by the weight of f.

The conjugate of a holomorphic modular form with character $\overline{\chi}$ leads in the same way to the discrete series representations which are eventually killed by iterates of the weight-raising operator.

On the other hand, the Laplacian eigenvalues of Maass forms are not well-understood. In fact, this is a very deep part of the theory. For instance, according to the Langlands conjectures, Maass forms of eigenvalue $1/4$ are supposed to be attached to even 2-dimensional Galois representations. The exact statement of such a conjecture looks like

Conjecture 3.23. *Let f be a Maass cusp form of weight 0 and Δ-eigenvalue $1/4$ and Fourier coefficients $\{a_n\}$. In other words,*

$$f(x+iy) = \sum_{n=1}^{\infty} a_n y^{1/2} K_0(2\pi n y)\cos(2\pi n x).$$

Then there is an even Galois representation

$$\rho : \mathrm{Gal}(\bar{\mathbf{Q}}/\mathbf{Q}) \to GL_2(\mathbf{C})$$

such that

$$a_p = \mathrm{Tr}(\rho(\mathrm{Frop}_p))$$

for all rational primes p such that ρ is unramified at p.

The extra assumption of weight 0 is reduntant, since it is a general fact that Maass forms of weight 1 not coming from modular forms have Δ-eigenvalue strictly greater than $1/4$.

Compared to the case of modular forms, fewer statements relating Galois representations to Maass forms are currently known unconditionally (for example, Booker–Strömbergsson–Venkatesh [2] were able to attach Maass forms to even Galois representations, but this work was conditional on Artin's conjecture).

This conjecture is not known to be true, and the list of eigenvalues for Δ itself is not well-understood. For example, it is not known in general what the multiplicity is of the eigenvalue $1/4$ is.

So the next step in understanding the decomposition of the representation $L^2_{\mathrm{cusp}}(\Gamma\backslash G/Z, \chi)$ is to try to derive asymptotics on the eigenvalues $\{\lambda_i\}$. As we saw above, the list of these eigenvalues is the exact piece of information remaining in order to fully understand the list of irreducible representations that actually appear in this decomposition. Since Δ acts by a scalar λ on these irreducible principal series subrepresentations π, which all have a K-finite (and thus smooth) vector of weight $\epsilon_\pi \in \{0,1\}$ (obtained by raising or lowering an arbitrary one). So it suffices to study the eigenvalues of the Laplacian acting on weight-0 and weight-1 Maass forms.

In the next subsection, we will follow Bump in deriving a basic quantitative result in this direction, namely Bessel's inequality

$$\sum \lambda_i^{-2} < \infty,$$

which implies that $\lambda_i \to \infty$ because the negative eigenvalues are only finitely many (they must come from modular forms of weight 0 or 1).

The last section of these notes are about the Selberg trace formula, which can be used to prove more precise asymptotics on the asymptotic growth of the λ_i. As far as I know, it is much more difficult to control the multiplicities of these eigenvalues than it is to control their growth.

3.4. **Green's functions and the spectrum of the Laplacian.** The useful technical input in this section towards proving some quantitative results about the Δ-eigenvalues is the construction of the appropriate *Green's function*. It is easier to deal with this type of issue by dealing directly with the upper half-plane rather than with representation theory, because this allows us to think concretely about geodesics and their lengths. There is no real difference between this setting and the setting of weight-k Maass forms, since in that case one is assuming some kind of invariance under Γ, K and Z anyway

Let $\epsilon \in \{0, 1\}$. As remarked in the last section, to understand what Laplace eigenvalues can appear for Maass forms of any weight not coming from holomorphic modular forms, it suffices to understand the eigenvalues of Δ acting on the Maass forms of weight ϵ.

The insight is actually something we have used before, but in a slightly less abstract way. Since $\Delta|_{\mathfrak{H}(\epsilon)}$ might not be positive (it may have nontrivial 0-eigenspace when $\epsilon = 0$), we first add a positive constant $s > 0$. The plan of attack is to construct a Hilbert–Schmidt operator on $\mathfrak{H}(\epsilon)$ whose eigenvalues contain[18] the $\{(\lambda_i + s)^{-1}\}$ (presumably with the eigenvectors the same as those of Δ_ϵ), so that we are guaranteed that

$$\sum (\lambda_i + s)^{-2} < \infty$$

and thus also

$$\sum \lambda_i^{-2} < \infty$$

where the sum omits all i such that $\lambda_i = 0$. The goal is therefore to construct a Hilbert-Schmidt kernel with the right properties, that is a function $G_{s,\epsilon}(\cdot, \cdot) \in L^2(\Gamma\backslash\mathbf{H}, \Gamma\backslash\mathbf{H})$ with the property that if $f : \mathbf{H} \to \mathbf{C}$ is a Maass form of weight ϵ and Δ_ϵ-eigenvalue λ, then

$$\int_{\Gamma\backslash\mathbf{H}} G_{s,\epsilon}(z, \zeta) f(\zeta)\, d\mu(\zeta) = (\lambda + s)^{-1} f(z).$$

In other words, we need to construct G such that

$$\int_{\Gamma\backslash\mathbf{H}} G_{s,\epsilon}(z, \zeta)(\Delta_\epsilon + s) f(\zeta)\, d\mu(\zeta) = f(z) \tag{3.24}$$

for all smooth functions f on $\Gamma\backslash\mathbf{H}$.

The key insight is that in order for this to happen, we expect that for all ζ,

$$(\Delta_\epsilon + s)_z G_{s,\epsilon}(z, \zeta) = 0.$$

Why do we have this expectation? The reason is not rigorous (because we haven't proved the conclusion yet, and because G is a priori not smooth), but this is the heuristic given by Bump [3, §2.3]. If we actually have $\sum(\lambda_i + s)^{-2} < \infty$, then we may define an operator called $(\Delta_\epsilon + s)^{-1}$ which acts by $(\lambda_i + s)^{-1}$ on the λ_i-eigenvector of Δ_i; and our result means that this operator and its inverse extend to all of $L^2(\Gamma\backslash\mathbf{H}, \chi)$. If the orthogonal eigenbasis for Δ_ϵ we have been using the whole time is called $\{\phi_i\}$, then Equation (3.24) amounts to saying that

$$G_{s,\epsilon}(z, \zeta) = \sum (\lambda_i + s)^{-1} \phi_i(z) \overline{\phi_i(\zeta)}$$

and thus

$$(\Delta_\epsilon + s)_z G_{s,\epsilon}(z, \zeta) = \sum \phi_i(z) \overline{\phi_i(\zeta)}.$$

(though in reality this is nonsense — this is not a valid element of $L^2(\Gamma\backslash\mathbf{H} \times \Gamma\backslash\mathbf{H})$) is supposed to act on functions like the Dirac delta distribution. In other words,

$$\int (\Delta_\epsilon + s) G_{s,\epsilon}(z, \zeta) f(\zeta) d\mu(z)\, d\mu(\zeta) = f(z).$$

Though the reasoning is nonsense, it makes sense to ask that $G_{s,\epsilon}(z, \zeta)$ satisfy the differential equation

$$(\Delta_\epsilon + s)_z G_{s,\epsilon}(z, \zeta) = 0$$

[18] though in reality they will be exactly equal to this list

for all ζ, and to blow up when $|z - \zeta| \to 0$. The actual reasoning for why this would imply Equation (3.24) is more technically involved, but at its core is really the same thing in disguise: integration by parts. Before doing anything, we make the simplifying assumption that $G_{s,\epsilon}$ only depends on the hyperbolic distance between z and ζ, i.e. that

$$G_{s,\epsilon}(z, \zeta) = g(r),$$

where

$$r = \left| \frac{z - \zeta}{z - \overline{\zeta}} \right|.$$

We will get the blowing-up of g as $r \to 0$ to be logarithmic in r. The rigorous technical justification for asking for these conditions is

Proposition 3.25. *Suppose that $g : (0, 1) \to \mathbf{R}$ is smooth with logarithmic growth as $r \to 0$ in the sense that $g(r) = c \log(r) + O(1)$ as $r \to 0$ for some constant c, $f : \mathbf{H} \to \mathbf{C}$ smooth and compactly supported, and*

$$(\Delta_\epsilon + s)_z g \left(\frac{z - \zeta}{z - \overline{\zeta}} \right) = 0$$

for all $z \neq \zeta$. Then for all $z \in \mathbf{H}$,

$$\int_{\mathbf{H}} G_{s,\epsilon}(z, \zeta)(\Delta_\epsilon + s)_z f(\zeta) \, d\mu(\zeta) = f(z).$$

Proof. This is an exercise in changing coordinates and applying Stokes' theorem. See [3, p. 182-183]. The proof in Bump also applies when $\epsilon = 1$, with an extra integration by parts. □

Constructing the function g with the required properties is an exercise in applying the theory of regular singular points of second-order differential equations. The condition $(\Delta_\epsilon + s)g\left(\frac{z-\zeta}{z-\bar{\zeta}}\right)$ enforces the differential equation

$$g''(r) + \frac{1}{r} g'(r) - \frac{4s}{(1 - r^2)^2} g(r) = 0$$

when $\epsilon = 0$, and something similar when $\epsilon = 1$. Either way, if we enforce some boundary condition as $r \to 1$, then the theory of regular singular points provides us with a solution with a logarithmic singularity near 0.

We did this over $\mathbf{H}$, but we need the same thing to hold if we replace $\mathbf{H}$ with $\Gamma\backslash\mathbf{H}$. Doing this is a matter of averaging

$$G_{s,\epsilon,\Gamma}(z, \zeta) = \sum_{\Gamma \in \Gamma} \chi(\gamma)^{-1} \left(\frac{cz + d}{|cz + d|} \right)^{-\epsilon} G_{s,\epsilon}(z, \gamma\zeta)$$

so that $G_{s,\epsilon,\Gamma}(z, \zeta) \in C^\infty(\Gamma\backslash\mathbf{H}, \epsilon)$ and we can view it as an element of $C^\infty(\Gamma\backslash\mathbf{H})$ in the same way we would view a Maass form f. In this case, the same kind of argument shows that for $f \in C^\infty(\Gamma\backslash\mathbf{H})$,

$$\int_{\Gamma\backslash\mathbf{H}} G_{s,\epsilon,\Gamma}(z, \zeta)(\Delta_\epsilon + s) f(\zeta) \, d\mu(\zeta) = f(z).$$

So $G_{s,\epsilon,\Gamma}$ is a Hilbert–Schmidt (because the logarithmic singularity on the diagonal is not enough to make its L^2 norm diverge) kernel for the resolvent, which proves that

$$\sum_{\lambda_i \neq 0} \lambda_i^{-2} \neq 0.$$

4. The trace formula

The most general formulation of the Arthur–Selberg trace formula is a computation of the traces of the compact operators $\pi(\phi)$ which were so useful in studying the right regular representation on $L^2(\Gamma\backslash G/Z,\chi)$ (here G can go much farther than GL_2). In the context of $G = GL_2$, where it was first discovered by Selberg [21], it is usually described in slightly less general terms (often using the language of the upper half-plane). We will give essentially complete proofs for compact quotient, but only statements in the noncompact finite-volume case, because of the difficulties posed by the continuous spectrum. The full details of the finite-volume case have been explained in the literature, for example in Hejhal's book [10] in the language of the upper half-plane. The general Arthur–Selberg trace formula and its applications are described in much greater generality in Arthur's notes [1].

4.1. **Compact quotient.** Suppose that $\Gamma\backslash G$ is compact and let $\phi \in C_c^\infty(G)$. Here we are mostly thinking about $G = PSL_2(\mathbf{R})$. The trace formula is a computation of the trace of the Hilbert–Schmidt operator

$$\pi(\phi) = \int_G \phi(g)\pi(g)\,dg.$$

Recall that it is Hilbert–Schmidt because of the identity

$$\pi(\phi)f(g) = \int_{\mathcal{F}} f(h)\left(\sum_{\gamma\in\Gamma}\chi(\gamma)\phi(g^{-1}\gamma h)\right)\,dh,$$

where $\mathcal{F}$ is a fundamental domain for $\Gamma\backslash G$. So the relevant Hilbert–Schmidt kernel is

$$K_\phi(g,h) = \sum_{\gamma\in\Gamma}\chi(\gamma)\phi(g^{-1}\gamma h).$$

Lemma 4.1. *If the Hilbert–Schmidt operator with kernel $K_\phi(g,h)$ is trace-class, then its trace is*

$$\int_{\Gamma\backslash G} K_\phi(g,g)\,dg.$$

Proof. This is a general fact. Since this is a Hilbert–Schmidt kernel, i.e.

$$K_\phi(\cdot,\cdot) \in L^2(\Gamma\backslash G\times\Gamma\backslash G)$$

(we are grateful for the compactness assumption here), we may write

$$K_\phi(g,h) = \sum a_i\phi_i(g)\overline{\phi_i}(h)$$

for a Hilbert space orthonormal basis $\{\phi_i\}$ of $L^2(\mathcal{F})$ (namely, the one that diagonalizes the compact operator $\pi(\phi)$). So its action on $L^2(\mathcal{F}) \cong L^2(\Gamma\backslash G,\chi)$ is by $\phi_i \mapsto a_i\phi_i$, and we have (under the trace-class assumption)

$$\begin{aligned}
\mathrm{Tr}\pi(\phi) &= \sum a_i \\
&= \sum a_i\langle\phi_i,\phi_i\rangle \\
&= \int_{\mathcal{F}}\sum a_i\phi_i(g)\overline{\phi_i(g)}\,dg \\
&= \int_{\mathcal{F}} K_\phi(g,g)\,dg
\end{aligned}$$

as desired. □

The most basic form of the Arthur–Selberg trace formula is simply a computation of this integral, and the observation that it is equal to the trace of this operator. Suppose that $L^2(\Gamma\backslash G, \chi)$ decomposes into irreducibles as

$$L^2(\Gamma\backslash G, \chi) \cong \bigoplus \pi_i^{m_i}.$$

The Arthur–Selberg trace formula in this situation is

Theorem 4.2. *If $\pi(\phi)$ is trace-class, then*

$$\underbrace{\sum_i m_i \mathrm{Tr}\left(\int_G \phi(g)\pi_i(g)\,dg\right)}_{\textit{spectral side}} = \underbrace{\sum_{\gamma\in\{\Gamma\}} \mu(\Gamma_\gamma\backslash G_\gamma)\int_{G_\gamma\backslash G} f(g^{-1}\gamma g)\,dg}_{\textit{geometric side}}$$

where both are equal to $\mathrm{Tr}\pi(\phi)$, $\{\Gamma\}$ is the set of representatives of conjugacy classes of Γ, and and subscripts denote centralizers.

Proof. The fact that the spectral side equals $Tr\pi(\phi)$ is immediate from the spectral decomposition of $L^2(\Gamma\backslash G, \chi)$ and the fact that the integral operator $\pi(\phi)$ restricts to a well-defined operator on any G-invariant subspace.

The main part of the proof is the computation of the geometric side, which is, by the previous lemma, a computation of the integral

$$\int_G K_\phi(g,g)\,dg.$$

This computation is reproduced from [1, §1]:

$$\begin{aligned}
\int_G K_\phi(g,g)\,dg &= \int_{\Gamma\backslash G}\sum_{\gamma\in\Gamma}\phi(g^{-1}\gamma g)\,dg \\
&= \int_{\Gamma\backslash G}\sum_{\gamma\in\{\Gamma\}}\sum_{\delta\in\Gamma_\gamma\backslash\Gamma}\phi(g^{-1}\delta^{-1}\gamma\delta g)\,dg \\
&= \sum_{\gamma\in\{\Gamma\}}\int_{\Gamma_\gamma\backslash G}\phi(g^{-1}\gamma g)\,dg \\
&= \sum_{\gamma\in\{\Gamma\}}\int_{\Gamma_\gamma\backslash G_\gamma}\int_{G_\gamma\backslash G}\phi((g_1g_2)^{-1}\gamma g_1 g_2)\,dg_1\,dg_2.
\end{aligned}$$

Of course, elements of G_γ act by conjugation by the identity on γ, so the outer integral is the integral of a constant function, hence we can ignore g_1 and the integral, simply multiplying by $\mu(\Gamma_\gamma\backslash G_\gamma)$. □

If ϕ is selected to have the appropriate decay properties, guaranteeing convergence, then the trace-class assumption is satisfied. This is the reason for the technical conditions that appear in specific instances of the trace formula. In most cases where this point of view on the trace formula is taken, the conjugacy classes in Γ are separated into parabolic, hyperbolic, elliptic, and identity terms (in each case the centralizer looks different). This is why the explicit forms of Selberg's trace formula in the literature have so many terms, and why

there is always the work of computing or bounding terms of all different types, even under compactness assumptions.

From the "spectral side equals geometric side" form of the trace formula, there are basically two ways to directly leverage it in this setting:

(1) Cook up a test function ϕ such that the spectral side is something you want to know about the Laplace eigenvalues, and the geometric side is something you can estimate.
(2) Cook up a test function ϕ such that the geometric side is something you want to know about lengths of prime geodesics on a hyperbolic Riemannian surface, and the spectral side is something you can estimate. This is reasonable because conjugacy classes in the fundamental group are supposed to correspond to closed geodesics.

Before we execute on these two promises, we should state the form of the trace formula typically used when dealing directly with the upper half-plane. In that case, the analysis is really the same, but we replace G with $G/K \cong \mathbf{H}$ (now one can repeat everything with $\mathbf{H}$ instead of G and replace the Haar measure on G with the measure induced by the appropriate normalization of the hyperbolic metric on $\mathbf{H}$; the fact that dg is invariant under G is replaced by the fact that the hyperbolic measure is invariant under isometries).

At first, it might be difficult to understand how Selberg's trace formula for the upper half-plane is the same thing as what I have explained above. However, the relationship is actually quite direct. The test function $\phi \in C_c^\infty(G)$ has the property that

$$\phi((\sigma g)^{-1}(\sigma h)) = \phi(g^{-1}h)$$

for all $\sigma \in G$. In particular, the map $G \times G \to \mathbf{R}$ given by

$$(g, h) \mapsto \phi(g^{-1}h)$$

is invariant under multiplying both coordinates by the same element of $PSL_2(\mathbf{R})$. So when we restrict to the setting of the upper half-plane where the K-coordinate is trivial, we replace the test function ϕ with a smooth function

$$\phi : \mathbf{H} \times \mathbf{H} \to \mathbf{R}$$

invariant under $PSL_2(\mathbf{R}) = \mathrm{Isom}(\mathbf{H})$ acting on the diagonal; the Hilbert–Schmidt kernel $K_\phi(\cdot, \cdot)$ is therefore replaced with

$$K_\phi(z, \zeta) = \sum_{\gamma \in \Gamma} \chi(\gamma)\phi(z, \gamma\zeta).$$

Remark 4.3. Though we haven't formulated everything yet, this formula alone should seem like promising progress towards using the trace formula to describing the asymptotics of the Laplace eigenvalues of Maass forms, or at least the trace of the resolvent of $\Delta + s$; in this setting one would use the Green's function $G_{s,\epsilon}$ for ϕ, so that the resolvent is given by the Hilbert–Schmidt kernel $G_{s,\epsilon,\Gamma}$.

The proof of Theorem 4.2 translates over without change to this situation, but now it is easier to write things down explicitly[19]. Due to the desire for explicitness, it is useful to think about these things as integrals on fundamental domains.

Theorem 4.4. *Suppose $\Gamma \subset PSL_2(\mathbf{R})$ such that $\Gamma\backslash\mathbf{H}$ is compact, and let $\phi \in C^\infty(\mathbf{H}\times\mathbf{H})$ be such that*

$$\phi(z,\zeta) = \Phi_0\left(d_{\mathbf{H}}(z,\zeta)\right) = \Phi\left(\frac{|z-\zeta|^2}{\Im(z)\Im(\zeta)}\right)$$

for some $\Phi \in C_c^\infty(\mathbf{R})$. Let L_ϕ be the Hilbert–Schmidt operator on $L^2(\Gamma\backslash\mathbf{H},\chi)$ given by the kernel $K_\phi(z,\zeta) = \sum_{\gamma\in\Gamma}\phi(z,\gamma\zeta)$. Then

$$\mathrm{Tr}L_\phi = \sum_{\gamma\in\{\Gamma\}}\int_{\mathcal{F}[\Gamma_\gamma\backslash\mathbf{H}]}\phi(z,\gamma z)\,d\mu(z)$$

where $\mathcal{F}[\Gamma_\gamma\backslash\mathbf{H}]$ is a fundamental domain for Γ_γ acting on $\mathbf{H}$.

Proof. The proof is identical to that of Theorem 4.2, except it stops right before the last step and is in the language of the hyperbolic measure on $\mathbf{H}$ rather than the Haar measure on $PSL_2(\mathbf{R})$:

$$\begin{aligned}
\mathrm{Tr}L_\phi &= \int_{\Gamma\backslash\mathbf{H}} K_\phi(z,z)\,d\mu(z)\\
&= \int_{\Gamma\backslash\mathbf{H}}\sum_{\gamma\in\Gamma}\phi(z,\gamma z)\,d\mu(z)\\
&= \int_{\Gamma\backslash\mathbf{H}}\sum_{\gamma\in\{\Gamma\}}\sum_{\delta\in\Gamma_\gamma\backslash\Gamma}\phi(z,\delta^{-1}\gamma\delta z)\,d\mu(z)\\
&= \int_{\Gamma\backslash\mathbf{H}}\sum_{\gamma\in\{\Gamma\}}\sum_{\delta\in\Gamma_\gamma\backslash\Gamma}\phi(\delta z,\gamma\delta z)\,d\mu(z)\\
&= \sum_{\gamma\in\{\Gamma\}}\int_{\Gamma\backslash\mathbf{H}}\sum_{\delta\in\Gamma_\gamma\backslash\Gamma}\phi(\delta z,\gamma\delta z)\,d\mu(z)\\
&= \sum_{\gamma\in\{\Gamma\}}\int_{\Gamma_\gamma\backslash\mathbf{H}}\phi(z,\gamma z)\,d\mu(z)
\end{aligned}$$

as desired. □

Making Theorem 4.4 explicit requires expanding both the geometric and spectral sides further. Since we are already on this topic, let's continue expanding the geometric side.

We have assumed that Γ is hyperbolic, so the main term is for the hyperbolic elements of Γ (the only other one is the identity).

[19]The "reason" why the Selberg trace formula is more "explicit" than the full Arthur–Selberg trace formula is because now ϕ may be written as a function of a single real variable, namely the hyperbolic distance between the two coordinates. The reality is that the more general version can be treated explicitly as well: in applications to the Langlands program when one is exploiting the Arthur–Selberg trace formula on adelic groups, one chooses the test function at each place, typically choosing a matrix coefficient of some kind at the infinite place.

Lemma 4.5. *If $\gamma \in \Gamma \subset PSL_2(\mathbf{R})$ is hyperbolic, its centralizer is cyclic, and it is conjugate in $PSL_2(\mathbf{R})$ to a unique element of the form $z \mapsto N(\gamma)z$, where*

$$\log N(\gamma) = \inf d_{\mathbf{H}}(\gamma z, z)$$

Proof. It's a standard fact that every element of $SL_2(\mathbf{R})$ is conjugate (in $SL_2(\mathbf{R})$) to an element of the form

$$\begin{pmatrix} x & 0 \\ 0 & x^{-1} \end{pmatrix}, \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}, \kappa_\theta.$$

From knowledge of the upper half-plane, only the first kind of conjugacy class is hyperbolic. One computes explicitly that the centralizer of such a diagonal matrix is just the group of diagonal matrices in $SL_2(\mathbf{R})$. So the centralizer of γ in Γ is equal to the subgroup of Γ consisting of diagonal matrices. This subgroup embeds as a discrete subgroup of $\mathbf{R}^\times$, so it must be cyclic. Finally, we observe that for

$$\gamma = \begin{pmatrix} x^{1/2} & 0 \\ 0 & x^{-1/2} \end{pmatrix},$$

we have

$$\begin{aligned} \inf_{z \in \mathbf{H}} d_{\mathbf{H}}(\gamma z, z) &= \inf_{z \in \mathbf{H}} d_{\mathbf{H}}(xz, z) \\ &= \inf_{z \in \mathbf{H}} \int_{\Im(z)}^{x\Im(z)} \frac{1}{y}\, dy \\ &= \log(x) \end{aligned}$$

as desired. □

So when $\gamma \in \Gamma$ is hyperbolic, there is an $\eta \in PSL_2(\mathbf{R})$ such that $\eta^{-1}\gamma\eta$ acts by $z \mapsto N(\gamma)z$, and the centralizer of $\eta^{-1}\gamma\eta$ is generated by $\eta^{-1}\gamma_0\eta$, where without loss of generality, $N(\eta^{-1}\gamma_0\eta) = N(\gamma_0) > 1$. The orbital integral corresponding to γ is

$$\begin{aligned} \int_{\mathcal{F}[\Gamma_\gamma \backslash \mathbf{H}]} \phi(z, \gamma z)\, d\mu(z) &= \int_{\eta^{-1}\mathcal{F}[\Gamma_\gamma \backslash \mathbf{H}]} \phi(\eta z, \gamma\eta z)\, d\mu(z) \\ &= \int_{\mathcal{F}[\eta^{-1}\Gamma_\gamma \eta \backslash \mathbf{H}]} \phi(\eta z, \gamma\eta z)\, d\mu(z) \\ &= \int_{\mathcal{F}[\Gamma_{\eta^{-1}\gamma\eta} \backslash \mathbf{H}]} \phi(\eta z, \gamma\eta z)\, d\mu(z) \\ &= \int_{\mathcal{F}[\Gamma_{\eta^{-1}\gamma\eta} \backslash \mathbf{H}]} \phi(z, \eta^{-1}\gamma\eta z)\, d\mu(z) \\ &= \int_{\mathcal{F}[\Gamma_{\eta^{-1}\gamma\eta} \backslash \mathbf{H}]} \phi(z, \eta^{-1}\gamma\eta z)\, d\mu(z) \\ &= \int_{1 \le \Im(z) \le N(\gamma_0)} \phi(z, N(\gamma)z)\, d\mu(z) \\ &= \int_1^{N(\gamma_0)} \int_{-\infty}^{\infty} \Phi\left(\frac{|(x+iy) - N(\gamma)(x+iy)|^2}{N(\gamma) y^2} \right) \frac{dx\, dy}{y^2} \\ &= 2 \int_1^{N(\gamma_0)} \int_0^{\infty} \Phi\left(\frac{(N(\gamma)-1)^2}{N(\gamma)} \frac{x^2+y^2}{y^2} \right) \frac{dx\, dy}{y^2} \end{aligned}$$

$$= \log N(\gamma_0) \frac{\sqrt{N(\gamma)}}{N-1} \int_{(N(\gamma)-1)^2/N(\gamma)}^{\infty} \frac{\Phi(t)}{\sqrt{t - \frac{(N(\gamma)-1)^2}{N(\gamma)}}} \, dt.$$

It's nice that these numbers $N(\gamma)$ show up: these are related to lengths of geodesics on $\Gamma\backslash\mathbf{H}$, and hence the trace formula will allow us to study the distribution of those lengths.

The identity term is easier. If $\gamma = I \in \Gamma$, then

$$\int_{\mathcal{F}[\Gamma_\gamma\backslash\mathbf{H}]} \phi(z, Iz) \, d\mu(z) = \int_{\Gamma\backslash\mathbf{H}} \phi(z,z) \, d\mu(z) = \Phi(0)\mu(\Gamma\backslash\mathbf{H}).$$

Now we should go back and deal with the spectral side in more explicit terms.

Lemma 4.6. *Suppose $f : \mathbf{H} \to \mathbf{C}$ is such that $\Delta_0 f = \lambda f$ for some $\lambda \in \mathbf{C}$. Then*

$$\int_{\mathbf{H}} \phi(z,\zeta) f(\zeta) \, d\mu(\zeta) = \Lambda(\lambda) f(z),$$

where $\Lambda(\lambda)$ depends only on λ and Φ (and in particular not on z). In fact, Λ is an entire function of λ.

This proof is taken from [10, Proposition 3.1]

Proof. Since Δ_0 respects the action of $PSL_2(\mathbf{R})$, $z \mapsto f(\sigma z)$ is also a λ-eigenfunction. So if we can prove the lemma when $z = i$, then we are done, since then we may choose $\sigma \in PSL_2(\mathbf{R})$ such that $\sigma i = z$, and then we have

$$\begin{aligned}
\int_{\mathbf{H}} \phi(z,\zeta) f(\zeta) \, d\mu(\zeta) &= \int_{\mathbf{H}} \phi(\sigma i, \zeta) f(\zeta) \, d\mu(\zeta) \\
&= \int_{\mathbf{H}} \phi(i, \sigma^{-1}\zeta) f(\zeta) \, d\mu(\zeta) \\
&= \int_{\mathbf{H}} \phi(i, \zeta) f(\sigma\zeta) \, d\mu(\zeta) \\
&= \Lambda(\lambda) f(\sigma i) \\
&= \Lambda(\lambda) f(z).
\end{aligned}$$

By the same argument we may actually choose $\sigma \in PSL_2(\mathbf{C})$, and transform the situation to be situated on the unit disc model of hyperbolic space, where $z = 0$. By the standard formulas for how the hyperbolic metric on $\mathbf{H}$ translates over to this situation, it suffices to show that

$$4 \int_{|z|<1} \Phi(|z|) f(z) \frac{dx\, dy}{(1-|z|^2)^2} = \Lambda(\lambda) f(0)$$

where $f : \{|z| < 1\} \to \mathbf{C}$ is a λ-eigenfunction for the Laplacian on the the unit disc model. By averaging and the fact that the measure only depends on $|z|$, we have

$$\int_{|z|<1} \Phi(|z|) f(z) \frac{dx\, dy}{(1-|z|^2)^2} = \int_{|z|<1} \Phi(|z|) F(z) \frac{dx\, dy}{(1-|z|^2)^2}$$

where $F(z) := \frac{1}{2\pi} \int_0^{2\pi} f(ze^{i\theta}) \, d\theta$. The new function F is useful because it only depends on $|z|$. So we can rewrite the integral we are interested in as

$$2\pi \int_0^1 \Phi(r) F(r) r \frac{dr}{(1-r^2)^2}.$$

The trick is now to use the same differential equation that let us to Green's functions: by differentiation under the integral sign, $\Delta_0 F = \lambda F$, which in the language of functions on the disc model we have already seen equates to

$$F''(r) + \frac{1}{r}F'(r) - \frac{4\lambda}{(1-r^2)^2}F(r) = 0$$

with initial conditions $F(0) = f(0)$ and $F'(0) = 0$ (both of these follow directly from the definition of $F(r)$ as the average of f on the circle of radius r). By the theory of regular singular points, we see that there exists a function $G_\lambda(r)$ [depending only on λ, the only other thing coming up in the differential equation] such that

$$F(r) = f(0) \cdot G(r).$$

This proves the result, because

$$2\pi \int_0^1 \Phi(r)F(r)r\frac{dr}{(1-r^2)^2} = f(0) \cdot 2\pi \int_0^1 \Phi(r)G_\lambda(r)r\frac{dr}{(1-r^2)^2}$$

and the integral on the right hand side only depends on λ and Φ. □

It's useful that this works without requiring that K_ϕ is a Green's function. The cost is that we need to understand the function Λ. Also, note that $G_\lambda(r)$ is different from the Green's function we used in the previous section: this one does not blow up near $r = 0$.

It isn't obvious (to me) that Λ is entire from its definition as an integral involving a function satisfying a differential equation depending on λ. Instead, one uses the result of Lemma 4.6: to compute Λ, we may choose any test function f we want[20], as long as it has $\Delta_0 f = \lambda f$. This is what allows for

Lemma 4.7. *For $r \in \mathbf{C}$, we have*

$$\Lambda\left(\frac{1}{4} + r^2\right) = h(r),$$

where

$$h(r) := \int_{\mathbf{R}} e^{iru} \int_{e^u+e^{-u}-2}^{\infty} \frac{\Phi(t)}{\sqrt{t-(e^u+e^{-u}-2)}}\, dt\, du.$$

In fact, Λ is entire.

Proof. The point is to take the test function (on the upper half-plane, not the disc) $f(x+iy) = \Im(y)^s$, where $s \in \mathbf{C}$. Then

$$\Delta_0 f = s(1-s)f.$$

So by Lemma 4.6,

$$\Lambda(s(1-s)) = \int_{\mathbf{H}} \phi(i,\zeta)\Im(\zeta)^s\, d\mu(\zeta).$$

Taking

$$s = \frac{1}{2} + ir,$$

with $r \in \mathbf{C}$ so that

$$s(1-s) = \frac{1}{4} + r^2,$$

[20]this is not a typo. For a brief time, now f will be a "test function" rather than Φ.

we may use this to compute (substituting $t = \frac{x^2+(y-1)^2}{y}$ and then $u = \log y$)

$$\begin{aligned}
\Lambda\left(\frac{1}{4} + r^2\right) &= \Lambda(s(1-s)) \\
&= \int_{\mathbf{H}} \phi(i,\zeta)\Im(\zeta)^s \, d\mu(\zeta) \\
&= 2\int_0^\infty \int_0^\infty \Phi\left(\frac{x^2+(y-1)^2}{y}\right) y^{s-2} \, dx \, dy \\
&= \int_0^\infty \int_{(y-1)^2/y}^\infty \Phi\left(t\right) y^{s-2} \frac{dt}{(\sqrt{ty-(y-1)^2})/y} \, dy \\
&= \int_{-\infty}^\infty \int_{e^u+e^{-u}-2}^\infty \Phi(t) e^{u(s-2)} \frac{dt}{(\sqrt{te^u-(e^u-1)^2})e^{-u}} \frac{du}{e^{-u}} \\
&= \int_{-\infty}^\infty e^{u\left(s-\frac{1}{2}\right)} \int_{e^u+e^{-u}-2}^\infty \frac{\Phi(t)}{\sqrt{t-e^u-e^{-u}+2}} \, dt \, du \\
&= \int_{-\infty}^\infty e^{iru} \int_{e^u+e^{-u}-2}^\infty \frac{\Phi(t)}{\sqrt{t-e^u-e^{-u}+2}} \, dt \, du
\end{aligned}$$

as claimed. This at least makes $s \mapsto \Lambda(s(1-s))$ an entire function, hence $\lambda \mapsto \Lambda(\lambda)$ is holomorphic away from $\lambda = 1/4$. But Λ is defined and continuous at $\lambda = 1/4$, so in fact Λ is entire. □

The useful thing about this[21] is that the function h that shows up here is the inverse Fourier transform of something that shows up in the geometric side. Let $g \in C^\infty(\mathbf{R})$ be the Fourier transform (appropriately normalized) of h. In other words,

$$g(u) = \frac{1}{2\pi}\int_{\mathbf{R}} h(r)e^{-iru} \, dr = \int_{e^u+e^{-u}-2}^\infty \frac{\Phi(t)}{\sqrt{t-(e^u+e^{-u}-2)}} \, dt.$$

Recall that for γ hyperbolic, we computed the relevant orbital integral

$$\int_{\mathcal{F}[\Gamma_\gamma\backslash\mathbf{H}]} \phi(z,\gamma z) \, d\mu(z) = \log N(\gamma_0) \frac{\sqrt{N(\gamma)}}{N(\gamma)-1} \int_{(N(\gamma)-1)^2/N(\gamma)}^\infty \frac{\Phi(t)}{\sqrt{t-\frac{(N(\gamma)-1)^2}{N(\gamma)}}} \, dt.$$

In our new notation, this becomes

Lemma 4.8. *If $\gamma \in SL_2(\mathbf{Z})$ is hyperbolic, then*

$$\int_{\mathcal{F}[\Gamma_\gamma\backslash\mathbf{H}]} \phi(z,\gamma z) \, d\mu(z) = \frac{\log N(\gamma_0)}{N(\gamma)^{1/2}-N(\gamma)^{-1/2}} g(\log N(\gamma)).$$

The identity term of the geometric side can also be simplified using this new notation. This is thanks to

Lemma 4.9. *If $\Phi \in C_c^\infty(\mathbf{R})$, then for any $t \in \mathbf{R}$,*

$$\Phi(t) = -\frac{1}{\pi}\int_t^\infty \frac{\frac{d}{dx}\int_x^\infty \frac{\Phi(v)}{\sqrt{v-x}} \, dv}{\sqrt{x-t}} \, dx$$

[21] I don't know why this isn't a coincidence.

Proof. The proof of this formula is the main part of [10, Proposition 4.1]. I don't know why it is not a coincidence. The fact that Φ is compactly supported makes all of our manipulations below kosher. First, observe that

$$\begin{aligned}\frac{d}{dx}\int_x^\infty \frac{\Phi(v)}{\sqrt{v-x}}\,dv &= 2\frac{d}{dx}\int_0^\infty \Phi(x+u^2)\,du \\ &= 2\int_0^\infty \Phi'(x+u^2)\,du \\ &= \int_x^\infty \frac{\Phi'(v)}{\sqrt{v-x}}\,dv.\end{aligned}$$

So we may compute

$$\begin{aligned}\int_t^\infty \frac{\frac{d}{dx}\int_x^\infty \frac{\Phi(v)}{\sqrt{v-x}}\,dv}{\sqrt{x-t}}\,dx &= \int_t^\infty \frac{\int_x^\infty \frac{\Phi'(v)}{\sqrt{v-x}}\,dv}{\sqrt{x-t}}\,dx \\ &= \int_t^\infty \int_x^\infty (x-t)^{-\frac{1}{2}}(v-x)^{-\frac{1}{2}}\Phi'(v)\,dv\,dx \\ &= \int_t^\infty \Phi'(v)\int_t^v (x-t)^{-\frac{1}{2}}(v-x)^{-\frac{1}{2}}\,dx\,dv \\ &= \int_t^\infty \Phi'(v)\int_0^1 x^{-\frac{1}{2}}(1-x)^{-\frac{1}{2}}\,dx\,dv \\ &= \int_t^\infty \Phi'(v)B\left(\frac{1}{2},\frac{1}{2}\right)\,dv \\ &= \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma(1)}\int_t^\infty \Phi'(v)\,dv \\ &= -\pi\Phi(t),\end{aligned}$$

again using the fact that Φ is compactly supported on $\mathbf{R}$ and in reality replacing t with $t+\epsilon$ while taking $\epsilon \to 0$ to deal with the improper integrals. □

Now we can deal with the identity term of the geometric side.

Lemma 4.10. *The identity term on the geometric side is*

$$\int_{\mathcal{F}[\Gamma\backslash\mathbf{H}]} \phi(z,Iz)\,d\mu(z) = \frac{\mu(\Gamma\backslash\mathbf{H})}{2\pi}\int_0^\infty rh(r)\tanh(\pi r)\,dr,$$

where h is the holomorphic function defined as above.

Proof. Recall from above that

$$\int_{\mathcal{F}[\Gamma\backslash\mathbf{H}]} \phi(z,Iz)\,d\mu(z) = \Phi(0)\mu(\Gamma\backslash\mathbf{H}),$$

and from Lemma 4.9 that (substituting $e^u + e^{-u} - 2$ for x)

$$\begin{aligned}\Phi(0) &= -\frac{1}{\pi}\int_0^\infty \frac{\frac{d}{dx}\int_x^\infty \frac{\Phi(v)}{\sqrt{v-x}}\,dv}{\sqrt{x}}\,dx \\ &= -\frac{1}{\pi}\int_0^\infty \frac{\frac{d}{dx}g(u)}{\sqrt{e^u+e^{-u}-2}}\frac{dx}{du}\,du\end{aligned}$$

$$= -\frac{1}{\pi}\int_0^\infty \frac{g'(u)}{e^{u/2} - e^{-u/2}}\, du$$

Since the spectral side of the trace formula we only have in terms of the inverse Fourier transform h of g, it is convenient to simplify this further using the formula for the derivative of the Fourier transform. In fact, using the symmetry of the integral involved,

$$g(u) = \frac{1}{2\pi}\int_{\mathbf{R}} h(r)e^{-iru}\, dr = \frac{1}{\pi}\int_0^\infty h(r)\cos(ru)\, dr$$

so differentiating under the integral sign yields

$$g'(u) = -\frac{1}{\pi}\int_0^\infty rh(r)\sin(ru)\, dr$$

hence

$$\begin{aligned}\Phi(0) &= \frac{1}{\pi^2}\int_0^\infty rh(r)\int_0^\infty \frac{\sin(ru)}{e^{u/2} - e^{-u/2}}\, du\, dr \\ &= \frac{1}{2\pi}\int_0^\infty rh(r)\tanh(\pi r)\, dr,\end{aligned}$$

as desired. □

Putting everything together on both the spectral and geometric side, we finally have the first version of the Selberg trace formula:

Theorem 4.11 (Selberg, 1956)**.** *Assume that Γ is hyperbolic, and that $\Gamma\backslash\mathbf{H}$ is compact. Let $\Phi \in C_c^\infty(\mathbf{R})$, and let $\{\lambda_i\}$ be the Δ_0-eigenvalues, counted with multiplicity. Defining g as above, h to be its Fourier transform (normalized as above), and $r_n = \sqrt{\lambda_n - 1/4}$, we have*

$$\sum_{n=0}^\infty h(r_n) = \frac{\mu(\Gamma\backslash\mathbf{H})}{2\pi}\int_0^\infty rh(r)\tanh(\pi r)\, dr + \sum_{I\neq\gamma\in\{\Gamma\}} \frac{\log N(\gamma_0)}{N(\gamma)^{1/2} - N(\gamma)^{-1/2}} g(\log N(\gamma))$$

where γ_0 denotes a generator of the centralizer of γ in Γ.

Proof. Both sides are equal to the trace of the operator L_ϕ, as defined in Theorem 4.4. The left hand side (the spectral side) is equal to this trace by Lemma 4.6 and Lemma 4.7. The first term on the geometric side is the identity term and was computed in Lemma 4.10. The second term is the hyperbolic term and its computation ends with Lemma 4.8. □

Of key importance for applications is that one can choose h instead of Φ. Once h is chosen, we may obtain g by taking the Fourier transform, and Φ from that via Lemma 4.9. In particular, one sets

$$g(u) = \frac{1}{2\pi}\int_{\mathbf{R}} h(r)e^{-iru}\, dr$$

and

$$\Phi(t) = -\frac{1}{\pi}\int_t^\infty \frac{g'(u)}{\sqrt{e^u + e^{-u} - 2 - t}}\, du$$

As long as the resulting Φ is well-defined and compactly supported, this is kosher. Of course, we want a bit more freedom in choosing h (Φ being compactly supported is too hard to satisfy), which can be done simply by approximating by compactly supported functions and carrying out the same arguments as in the proof of Theorem 4.11. This analysis is

carried out in [10, §1.7], where it is shown that Theorem 4.11 is true as stated for any test function $h(r)$ satisfying the properties

(1) h is analytic on $|\Im(r)| \leq \frac{1}{2} + \delta$ (the point being that the r_n live in this region) for some positive constant δ.
(2) $h(r) = h(-r)$.
(3) $|h(r)| \ll |1 + |\Re(r)||^{-2-\delta}$.

In fact, thanks to condition (3), both sides of the trace formula stated in Theorem 4.11 are absolutely convergent. This uses the strengthening of Bessel's inequality (see Section 3 for the proof using the technique of Green's functions) to the effect that for any $\epsilon > 0$,

$$\sum \lambda_i^{-(1+\epsilon)} < \infty.$$

With only the form of Bessel's inequality we know (with $\epsilon = 1$), condition (3) must be strengthened to $|h(r)| \ll |1 + |\Re(r)||^{-4-\delta}$.

One application is to further pinning down the asymptotic growth of the λ_i. This proof is taken from [18, Proposition 10].

Theorem 4.12 (Weyl's law for compact surfaces)**.** *Under the same hypotheses on* Γ *as above,*

$$\#\{i : \lambda_i \leq T\} \sim \frac{\mu(\Gamma\backslash\mathbf{H})}{4\pi} T$$

as $T \to \infty$*.*

Proof. Fix $\beta > 0$, and take the test function $h(t) = e^{-\beta t^2}$. The spectral side of the trace formula is the *heat kernel*

$$\sum_n e^{-\beta r_n^2}.$$

If we can understand the asymptotic behavior of this quantity as $\beta \to 0$, then we can expect to understand the λ_i better. The Fourier transform of h (using the nonstandard normalization convention we have been using thus far) is

$$g(t) = \frac{1}{2\sqrt{\pi\beta}} e^{-\frac{t^2}{2\beta}}.$$

Plugging this into the trace formula, we get

$$\sum_n e^{-\beta r_n^2} = \frac{\mu(\Gamma\backslash\mathbf{H})}{2\pi} \int_0^\infty r e^{-\beta r^2} \tanh(\pi r)\, dr + \frac{1}{2\sqrt{\pi\beta}} \sum_{I \neq \gamma \in \{\Gamma\}} \frac{\log N(\gamma_0) e^{-(\log N(\gamma))^2/(2\beta)}}{N(\gamma)^{1/2} - N(\gamma)^{-1/2}}.$$

Multiplying by $e^{-\beta/4}$, we obtain

$$\sum_n e^{-\beta\lambda_n^2} = \frac{\mu(\Gamma\backslash\mathbf{H})}{2\pi} \int_0^\infty r e^{-\beta\left(\frac{1}{4}+r^2\right)} \tanh(\pi r)\, dr + \frac{e^{-\beta/4}}{2\sqrt{\pi\beta}} \sum_{I \neq \gamma \in \{\Gamma\}} \frac{\log N(\gamma_0) e^{-(\log N(\gamma))^2/(2\beta)}}{N(\gamma)^{1/2} - N(\gamma)^{-1/2}}.$$

As $\beta \to 0$, the negative exponential in the hyperbolic term dominated everything else: the norms of hyperbolic conjugacy classes are all at least 2 so the denominators are bounded below, and the exponential term clearly dominates the $\log N(\gamma_0)$ since $N(\gamma) \geq N(\gamma_0)$ as well as the $\beta^{-1/2}$ (because it is negative exponential in $1/\beta$). So the identity term is the only

one that makes a contribution. For that term, we use the estimate $\tanh(\pi r) = 1 + O(e^{-2\pi r})$. The upper bound on the error integrates to

$$\ll \int_0^\infty r e^{-\beta\left(\frac{1}{4}+r^2\right)-2\pi r}\, dr \ll 1.$$

The rest is

$$\int_0^\infty r e^{-\beta\left(\frac{1}{4}+r^2\right)}\, dr = e^{-\beta/4} \cdot \left[\frac{-1}{2\beta} e^{-\beta r^2}\right]_{r=0}^{\infty} = \frac{1}{2\beta} + O(1)$$

as $\beta \to 0$. This shows the estimate on the heat kernel

$$\sum_n e^{-\beta\lambda_n} = \frac{\mu(\Gamma\backslash\mathbf{H})}{4\pi}\beta^{-1} + O(1)$$

as $\beta \to 0$. This implies the desired asymptotic formula for the λ_i by Karamata's Tauberian theorem [11]. □

Weyl's law generalizes to where there are non-hyperbolic terms as well. This was an application of an estimate of the geometric side to gain fine control over the spectral side. Indeed, the previous bound coming from Bessel's inequality is only enough to say $\{i : \lambda_i \leq T\} \ll T^2$, so this is a substantial improvement.

On the other hand, once one can control the spectral side, it is also possible to use it to deduce things about the geometric side. Intrinsic to the compact Riemannian manifold $\Gamma\backslash\mathbf{H}$ are the lengths of the closed geodesics on it. Of course, given a geodesic γ, we probably only want to know the length of γ, and not the geodesics $\gamma(2t)$, $\gamma(3t), \ldots$, which trace over the image of γ multiple times. In other words, we are interested in

Definition 4.13 (Prime geodesics)**.** Let X be a Riemannian manifold. A *prime geodesic* on X is a closed geodesic that traces out its image exactly once.

Remark 4.14. On the other hand, if $\gamma(t)$ is a prime geodesic on X, then X is also equipped with the time-reversal of γ, namely $t \mapsto \gamma(-t)$. For our purposes, these count as different prime geodesics even though they trace out the same image.

Just as we are interested in the asymptotics of the prime numbers, we are interested in the asymptotics of lengths of prime geodesics on $\Gamma\backslash\mathbf{H}$.

Moreover, we have a bijection

$$\{\text{hyperbolic conjugacy classes of } \Gamma\} \to \{\text{closed geodesics on } \Gamma\backslash\mathbf{H}\}$$

taking a conjugacy class represented by a hyperbolic element $\gamma \in \Gamma$ to the closed geodesic given by the projection to $\Gamma\backslash\mathbf{H}$ of the arc of the geodesic on $\mathbf{H}$ connecting the two fixed points (on the real axis) of γ constituting a fundamental domain of the action of $\langle\gamma\rangle$ on that geodesic. Note that the length of the closed geodesic corresponding to a hyperbolic $\gamma \in \Gamma$ is given by (for any z in the geodesic connecting the two fixed points)

$$d_{\mathbf{H}}(z, \gamma z) = \log N(\gamma),$$

and the length of the underlying prime geodesic is $\log N(\gamma_0)$, where γ_0 is a generator of the centralizer of γ in Γ. This provides an opportunity to apply the trace formula to study these quantities. I learned about this application from Sarnak's thesis [19], where he explains a somewhat more general result.

Theorem 4.15 (Selberg, 1956)**.** *Suppose* $\Gamma\backslash\mathbf{H}$ *is compact. Then*

$$\#\{\textit{prime geodesics } \tau \textit{ on } \Gamma\backslash\mathbf{H} : \mathrm{len}(\tau) \leq \log T\} \sim Li(T)$$

as $T \to \infty$*.*

This proof is reproduced from Sarnak's thesis.

Proof. This time, the test function of choice is a little more complicated. Let $T.\epsilon > 0$, and define the Fejér kernel (or its Fourier transform depending on the convention) to be $k_T(x) = 1 - |x|/T$ for $0 \leq |x| \leq T$ and 0 elsewhere. Also, take an even (Schwartz) function $\psi \in C_c^\infty(\mathbf{R})$ supported in $[-1, 1]$ with $\int_{\mathbf{R}} \psi = 1$, and define the dilations in the usual way

$$\psi_\epsilon(x) = \epsilon^{-1}\psi(x/\epsilon).$$

This way, the ψ_ϵ (i.e. the corresponding convolution operators) are supposed to be an approximation to the identity. Since ψ is Schwartz, so is $\hat{\psi}$ and its derivative. For $1 \leq p \leq \infty$, the L^p norms of those are all $O_\psi(1)$ (in particular, they are finite and depend only on ψ).

Also, since these functions are all even, there is no distinction between the Fourier transform and the inverse Fourier transform. So we define

$$g(x) = g_{T,\epsilon}(x) = (k_T * \psi_\epsilon)(x),$$

which is supposed to be a series of smoothed-out approximations to the Fejér kernel, and thus

$$h(x) = \hat{g}_{T,\epsilon} = T\left(\frac{\sin(Tx/2)}{Tx/2}\right)^2 \cdot \hat{\psi}(\epsilon x).$$

First, we estimate the identity term. Since $\tanh(\pi r) \leq 1$ and

$$\left(\frac{\sin(Tr/2)}{Tr/2}\right)^2 \leq 1,$$

we have

$$\begin{aligned}\int_0^1 rh(r)\tanh(\pi r)\,dr &\ll_\psi \int_0^1 T\,d(r^2)\\ &\ll_\psi T.\end{aligned}$$

And since $h(r) \ll \frac{1}{Tr^2}\hat{\psi}(\epsilon r)$, we may also estimate via integration by parts

$$\begin{aligned}\int_1^\infty rh(r)\tanh(\pi r)\,dr &\ll \int_1^\infty h(r)\,d(r^2)\\ &\ll \int_1^\infty \frac{1}{Tr^2}\hat{\psi}(\epsilon r)\,d(r^2)\\ &= \frac{1}{T}\left[\hat{\psi}(\epsilon r)\right]_{r=1}^\infty - \frac{1}{T}\int_1^\infty r^2\frac{d}{dr}\left[\frac{\hat{\psi}(\epsilon r)}{r^2}\right]dr\\ &\ll_\psi \frac{1}{T} + \frac{1}{T}\int_1^\infty \epsilon\hat{\psi}'(\epsilon r)\,dr + \frac{1}{T}\int_1^\infty \hat{\psi}(\epsilon r)\frac{dr}{r}\\ &= \frac{1}{T} + \frac{1}{T}\int_\epsilon^\infty \hat{\psi}'(r)\,dr + \frac{1}{T}\int_\epsilon^\infty \hat{\psi}(r)\frac{dr}{r}\\ &\ll_\psi \frac{1}{T} + \frac{1}{T}\log\left(\frac{1}{\epsilon}\right)\end{aligned}$$

where in the last step we are using the fact that $\hat{\psi}$ is Schwartz. Adding up the two contributions $\int_0^1 + \int_1^\infty$, we have the estimate on the identity term

$$\int_0^\infty rh(r)\tanh(\pi r)\,dr \ll T + \frac{1}{T}\log(1/\epsilon).$$

The test function is engineered to yield essentially a truncated (and weighted) version of a sum involving the lengths of geodesics. In particular, the hyperbolic term is

$$\sum_{I \neq \gamma \in \{\Gamma\}} \frac{\log N(\gamma_0)}{N(\gamma)^{1/2} - N(\gamma)^{-1/2}} g_{T,\epsilon}(\log N(\gamma)).$$

All the terms with $\log N(\gamma) \geq T + \epsilon$ vanish straightaway (because $g_{T,\epsilon} = k_T * \psi_\epsilon$ vanishes for those values by definition of the convolution), so this is a sum over the geodesics we are actually interested in, namely those with $\log N(\gamma) < T + \epsilon$ (the difference between T and $T + \epsilon$ won't really matter). Recall that convolution by ψ_ϵ approximates the identity, in the send that $\epsilon \to 0$, $\|g_{T,\epsilon} - k_T\|_{L^\infty(\mathbf{R})} \leq \epsilon$ independently of T. So the trace formula reads

$$\sum_\tau \frac{\tau_0}{e^{\tau/2} - e^{-\tau/2}} k_T(\tau) + O\left(\sum_{\tau \leq T+\epsilon} \frac{\tau_0}{e^{\tau/2} - e^{-\tau/2}}\epsilon + T + \frac{1}{T}\log(1/\epsilon)\right),$$

where τ ranges over the lengths, with multiplicity, of closed geodesics on $\Gamma\backslash\mathbf{H}$, and τ_0 is the length of the underlying prime geodesic. The spectral side is

$$\sum_n T\left(\frac{\sin(Tr_n/2)}{Tr_n/2}\right)^2 \hat{\psi}(\epsilon r_n).$$

Since the sequence of $\lambda_n \geq 0$ is discrete and tends to infinity, all but finitely many of the r_n are real. Moreover, the contribution of the terms where r_n is real to the spectral side is

$$\sum_{\substack{n \geq 0 \\ \lambda_n \geq \frac{1}{4}}} T\left(\frac{\sin(Tr_n/2)}{Tr_n/2}\right)^2 \hat{\psi}(\epsilon r_n) = \int_0^\infty T\left(\frac{\sin(Tr/2)}{Tr/2}\right)^2 \hat{\psi}(\epsilon r)\, d(\#\{n : r_n < r\})$$

$$\ll_{\psi,\Gamma} T + \frac{1}{T}\log\left(\frac{1}{\epsilon}\right),$$

where this estimate is obtained using the fact that $\#\{n : r_n < r\} \ll_\Gamma r^2$ (Theorem 4.12) and the same technique we used to estimate the identity term. So the terms where $\lambda_n \geq 1/4$ are absorbed into the $O(T + T^{-1}\log(\epsilon^{-1})$ error. Writing $r_n = it_n$ for the finitely many n with $\lambda_n < 1/4$, the analysis of the spectral side is now reduced to[22]

$$-4\sum_{\substack{n \geq 0 \\ \lambda_n < \frac{1}{4}}} \frac{\sin^2(Tit_n/2)}{Tt_n^2}\hat{\psi}(\epsilon i t_n) = -4\sum_{\substack{n \geq 0 \\ \lambda_n < \frac{1}{4}}} \frac{\sin^2(Tit_n/2)}{Tt_n^2}\int_{\mathbf{R}} \psi(r)e^{-\epsilon t_n r}\,dr$$

$$= -4\sum_{\substack{n \geq 0 \\ \lambda_n < \frac{1}{4}}} \frac{\sin^2(Tit_n/2)}{Tt_n^2}\left[\|\psi\|_{L^1} + \int_{-1}^1 \psi(r)(e^{-\epsilon t_n r} - 1)\,dr\right]$$

[22]Using the assumption that $\|\psi\|_{L^1} = 1$ and $\text{supp}\psi \subset [-1, 1]$, plus the fact that $t_n \leq 1/2$ for each n and $e^x - 1 \ll x$ for x bounded above.

$$
\begin{aligned}
&= -4 \sum_{\substack{n\geq 0 \\ \lambda_n < \frac{1}{4}}} \frac{\sin^2(Tit_n/2)}{Tt_n^2} \left[1 + O(\epsilon)\right] \\
&= \sum_{\substack{n\geq 0 \\ \lambda_n < \frac{1}{4}}} \frac{e^{-Tt_n} + e^{Tt_n} - 2}{Tt_n^2} \left[1 + O(\epsilon)\right] \\
&= \sum_{\substack{n\geq 0 \\ \lambda_n < \frac{1}{4}}} \left[\frac{e^{Tt_n}}{Tt_n^2} + O\left(\frac{1}{T}\right)\right] \left[1 + O(\epsilon)\right] \\
&= \sum_{\substack{n\geq 0 \\ \lambda_n < \frac{1}{4}}} \frac{e^{Tt_n}}{Tt_n^2} + O(\epsilon e^{T/2} + T^{-1}).
\end{aligned}
$$

So the trace formula reads

$$
\sum_{\substack{n\geq 0 \\ \lambda_n < \frac{1}{4}}} \frac{e^{Tt_n}}{Tt_n^2} = \sum_{\tau \leq T+\epsilon} \frac{\tau_0}{e^{\tau/2} - e^{-\tau/2}} (k_T(\tau) + O(\epsilon)) + O\left(T + \frac{1}{T}\log\left(\frac{1}{\epsilon}\right) + \epsilon e^{T/2}\right).
$$

Using the trivial bound[23]

$$
\#\{\tau : \tau \leq x\} \ll_\Gamma e^x
$$

(which also implies that there is a well-defined positive smallest length of a closed geodesic), we may estimate

$$
\sum_{\tau \leq T+\epsilon} \frac{\tau_0}{e^{\tau/2} - e^{-\tau/2}} \ll_\Gamma (T+\epsilon) e^{T+\epsilon} \ll e^{1.1T},
$$

where the dependence of the implied constant on Γ comes from both the implied constant from the trivial bound and from the length of the shortest geodesic on $\Gamma\backslash\mathbf{H}$ (also in the last bound we have used that $\epsilon \to 0$). Setting $\epsilon = e^{-1.1T}$, the trace formula now reads

$$
\sum_{\substack{n\geq 0 \\ \lambda_n < \frac{1}{4}}} \frac{e^{Tt_n}}{Tt_n^2} = \sum_{\tau \leq T} \frac{\tau_0}{e^{\tau/2} - e^{-\tau/2}} \left(1 - \frac{\tau}{T}\right) + O\left(T\right)
$$

as $T \to \infty$. Note that (again using the trivial bound)

$$
\begin{aligned}
\sum_{\tau \leq T} \frac{\tau_0}{e^{\tau/2} - e^{-\tau/2}} - \frac{\tau_0}{e^{\tau/2}} &\leq \sum_{\tau \leq T} \frac{\tau}{e^{\tau/2} - e^{-\tau/2}} - \frac{\tau}{e^{\tau/2}} \\
&= \sum_{\tau \leq T} \frac{\tau}{e^{3\tau/2} - e^{\tau/2}}
\end{aligned}
$$

[23]See [10, Proposition 2.5]. The point is that every hyperbolic conjugacy class has a representative γ whose underlying geodesic on $\mathbf{H}$ meets the canonical fundamental domain $\mathcal{F}[\Gamma\backslash\mathbf{H}]$, and we know that $d_{\mathbf{H}}(z, \gamma z) = \log N(\gamma)$ for $z \in \gamma$. The conjugacy classes of log-norm at most x therefore all have the property that they have a representative γ such that $d_{\mathbf{H}}(z_0, \gamma\mathcal{F}) \leq x + \mathrm{diam}\mathcal{F}$ where z_0 is some point in $\mathcal{F}$ fixed beforehand. In other words, $\gamma\mathcal{F} \cap B_{x+\mathrm{diam}\mathcal{F}}(z_0) \neq \emptyset$, and thus $\gamma\mathcal{F} \subset B_{x+2\mathrm{diam}\mathcal{F}}(z_0)$. The number of $\gamma \in \Gamma$ that satisfy this last inequality (which we have shown is an upper bound for the number we are interested in) is (by covering a subset of $B_{x+2\mathrm{diam}\mathcal{F}}(z_0)$ with disjoint translates of $\mathcal{F}$ and looking at areas) at most $\mu(B_{x+2\mathrm{diam}\mathcal{F}}(z_0))/\mu(\mathcal{F})$, so the trivial bound follows from the fact that the area of a hyperbolic disc of radius r is asymptotic to πe^r as $r \to \infty$.

$$\begin{aligned}
&\ll_\Gamma \int_0^\infty xe^{-3x/2}\, d(\#\{\tau < x\}) \\
&\ll_\Gamma -\int_0^\infty e^x \frac{d}{dx}\left[xe^{-3x/2}\right]\, dx \\
&\ll \int_0^\infty (1+x)e^{-x/2}\, dx \\
&\ll 1
\end{aligned}$$

so that difference is absorbed in the error and we have (after multiplying by T)

$$\sum_{\substack{n\geq 0\\ \lambda_n<\frac{1}{4}}} \frac{e^{Tt_n}}{t_n^2} = \sum_{\tau\leq T} \frac{\tau_0}{e^{\tau/2}}\left(T-\tau\right) + O\left(T^2\right).$$

For small $h>0$ (going to 0 as $T\to\infty$), we can take the difference quotient of both sides as a function of T. On the left hand side, that is

$$\begin{aligned}
\sum_{\substack{n\geq 0\\ \lambda_n<\frac{1}{4}}} \frac{e^{t_n(T+h)}-e^{t_nT}}{ht_n^2} &= \sum_{\substack{n\geq 0\\ \lambda_n<\frac{1}{4}}} \frac{e^{t_nT}(t_nh+O(h^2))}{ht_n^2} \\
&= \sum_{\substack{n\geq 0\\ \lambda_n<\frac{1}{4}}} \frac{e^{t_nT}}{t_n} + O(he^{T/2})
\end{aligned}$$

(where we have used the fact that $t_0 = 1/2$ is the largest of the t_n's). And the right hand side becomes

$$\sum_{\tau\leq T} \frac{\tau_0}{e^{\tau/2}} + \sum_{T<\tau\leq T+h} \frac{\tau_0}{e^{\tau/2}}\left(\frac{T+h-\tau}{h}\right) + O((T+h)^2/h)$$

which means that (since the terms in the sum $\sum_{T<\tau\leq T+h}$ are all positive)

$$\sum_{\tau\leq T} \frac{\tau_0}{e^{\tau/2}} \leq \sum_{\substack{n\geq 0\\ \lambda_n<\frac{1}{4}}} \frac{e^{t_nT}}{t_n} + O\left(he^{T/2}+T+h+\frac{T^2}{h}\right).$$

Taking the difference quotient from the left, we get (by the same arguments) the same thing on the left hand side and on the right hand side except for the $\sum_{T<\tau\leq T+h}$ term is negated, so in fact

$$\sum_{\tau\leq T} \frac{\tau_0}{e^{\tau/2}} \geq \sum_{\substack{n\geq 0\\ \lambda_n<\frac{1}{4}}} \frac{e^{t_nT}}{t_n} + O\left(he^{T/2}+T+h+\frac{T^2}{h}\right).$$

Setting $h = Te^{-T/4}$, we obtain

$$\sum_{\tau\leq T} \frac{\tau_0}{e^{\tau/2}} = \sum_{\substack{n\geq 0\\ \lambda_n<\frac{1}{4}}} \frac{e^{t_nT}}{t_n} + O\left(Te^{T/4}\right).$$

The contribution of the non-prime geodesics here is bounded by

$$\begin{aligned}\sum_{\tau_0 \leq T} \tau_0 \sum_{k=2}^{\infty} e^{-k\tau_0/2} &\ll_\Gamma \sum_{\tau_0 \leq T} \tau_0 e^{-\tau_0} \\ &= \int_0^T x e^{-x} d(\#\{\tau_0 < x\}) \\ &\ll_\Gamma T + \int_0^T e^x \frac{d}{dx}\left[x e^{-x}\right] dx \\ &\ll_\Gamma T^2\end{aligned}$$

(using the trivial bound again) which is absorbed into the error term, and hence we can rewrite the expression from the trace formula with the geometric side purely in terms of lengths of prime geodesics, namely

$$\sum_{\tau_0 \leq T} \frac{\tau_0}{e^{\tau_0/2}} = \sum_{\substack{n \geq 0 \\ \lambda_n < \frac{1}{4}}} \frac{e^{t_n T}}{t_n} + O\left(T e^{T/4}\right). \tag{4.16}$$

This lets us conclude via the usual technique of integration by parts. Let $F(T)$ be the quantity equal to both sides of Equation (4.16). Then the thing we are interested in is (using both the left and right hand sides of Equation (4.16) and the fact that F vanishes for small enough inputs)

$$\begin{aligned}\#\{\tau_0 < T\} &= \int_0^T x^{-1} e^{x/2} \, dF(x) \\ &= T^{-1} e^{T/2} F(T) - \int_\alpha^T F(x) \frac{d}{dx}\left[x^{-1} e^{x/2}\right] dx \\ &= T^{-1} e^{T/2} F(T) - \int_\alpha^T \left(\sum_{\substack{n \geq 0 \\ \lambda_n < \frac{1}{4}}} \frac{e^{t_n x}}{t_n} + O\left(x e^{x/4}\right) \right) \left(-x^{-2} e^{x/2} + \frac{1}{2} x^{-1} e^{x/2} \right) dx.\end{aligned}$$

where $\alpha > 0$ is smaller than the length of the shortest prime geodesic. The part of the integral that gets multiplied by $O(xe^{x/4})$ is

$$\begin{aligned}\ll \int_\alpha^T x e^{x/4} \left(-x^{-2} e^{x/2} + \frac{1}{2} x^{-1} e^{x/2} \right) dx &\ll \int_\alpha^T (x^{-1} + 1) e^{3x/4} \, dx \\ &\ll_\alpha e^{3T/4}\end{aligned}$$

and the rest is

$$\#\{\tau_0 < T\} = T^{-1} e^{T/2} F(T) - \int_\alpha^T \sum_{\substack{n \geq 0 \\ \lambda_n < \frac{1}{4}}} \frac{e^{t_n x}}{t_n} \, d(x^{-1} e^{x/2}) + O(e^{\frac{3}{4}T})$$

$$= T^{-1}e^{T/2}\left(\sum_{\tau_0 \leq T} \frac{\tau_0}{e^{\tau_0/2}} - \sum_{\substack{n\geq 0 \\ \lambda_n < \frac{1}{4}}} \frac{e^{t_n T}}{t_n}\right) + \int_\alpha^T x^{-1}e^{x/2} \sum_{\substack{n\geq 0 \\ \lambda_n < \frac{1}{4}}} e^{t_n x}\, dx + O(e^{\frac{3}{4}T})$$

$$= \sum_{\substack{n\geq 0 \\ \lambda_n < \frac{1}{4}}} \int_\alpha^T \frac{e^{\left(t_n + \frac{1}{2}\right)x}}{x}\, dx + O(e^{\frac{3}{4}T}).$$

Plugging in $\log T$ instead of T and changing variables ($u = e^{(t_n+1/2)x}$) in the integral, we get the desired

$$\#\{\tau_0 < \log T\} = \sum_{\lambda_n < \frac{1}{4}} Li\left(T^{t_n + \frac{1}{2}}\right) + O(T^{3/4})$$

(which gives us what we want because $t_0 = 1/2$ and all the other t_n's are smaller). □

Remark 4.17. The error term in Sarnak's thesis is actually $O(T^{3/4}(\log T)^2)$, which originates from the fact that his error term after the Tauberian differentiation argument is $O(T^2 e^{T/4})$ as compared to our $O(Te^{T/4})$. The most likely explanation for this is that there is a mistake in my own replication of his argument. Either way, the asymptotics are the same. In any event, the refinement of the final error term is mostly about the exponent on T, and is the subject of a lot of important and more recent work (see e.g. [17] where the key point is to use the Weil bounds on Kloosterman sums to gain information about the cancellation in the error term) outside the scope of this paper.

Remark 4.18. By the uniformization theorem and Gauss–Bonnet, the surfaces $\Gamma\backslash\mathbf{H}$ account for all compact Riemann surfaces of genus $g \geq 2$. In fact, they account for the compact 2-dimensional Riemannian manifolds with constant negative curvature. So even this most basic form of a prime geodesic theorem is about something concrete and interesting. Note, too, that for positive curvature there are usually too many geodesics for this to make sense: take S^2 with the usual metric, for instance.

4.2. **Remarks on noncompact finite-volume quotients.** Without the assumptions we made in the previous section, the trace formula must be modified to account for the following two issues:

(1) Γ might have elements which are not hyperbolic or the identity. So the full trace formula has an elliptic term and a parabolic term.
(2) $\Gamma\backslash\mathbf{H}$ might not be compact. This brings in issues relating to the continuous spectrum and therefore Eisenstein series.

These extra terms (from elliptic, parabolic, and Eisenstein series) are not a big deal, in the sense that Weyl's law and the prime geodesic theorem above can still be proved in the same way: those extra terms end up being absorbed in the error.

That being said, the proofs of the generalizations of these facts are essentially the same, once one proves the necessary facts about Eisenstein series. Those facts are outside the scope of these notes, which are already far too long. They are proved in [10, Ch. 2] and [19], and in fact Theorem 4.15 holds for finite-volume quotients of $\mathbf{H}$, including the affine modular curve $Y(1)$.

4.3. **Class numbers of real quadratic fields.** The Maass forms are supposed to have something to do with arithmetic, at least those of Laplace eigenvalue 1/4. However, Sarnak [20] found another interesting application to arithmetic, via the prime geodesic theorem for finite-volume quotients of $\mathbf{H}$. He applied it to the uncompactified modular curve $Y(\Gamma)$, where $\Gamma = \Gamma(p)$. The lengths of geodesics on this modular curve are the same as regulators of quadratic fields with discriminant divisible by p. Since the argument is essentially the same conceptually but slightly less complicated, we will restrict our attention to $Y(1)$

Theorem 4.19 (Sarnak, 1981)**.**

$$\sum_{e^{R_d} \leq x} h(d) \sim Li(x^2)$$

as $x \to \infty$, where the sum is over discriminants d of orders of quadratic fields, R_d denotes the narrow regulator and $h(d)$ denotes the narrow class number.

Proof. For any $\ell > 0$, there is a natural bijection

$$\left\{ \begin{matrix} SL_2(\mathbf{Z})\text{-equivalence classes of primitive binary} \\ \text{quadratic forms } f \text{ such that } 2R_{\mathrm{disc}(f)} = \ell \end{matrix} \right\} \to \{\text{prime geodesics on } Y(1) \text{ of length } \ell\}$$

formed in the following way. Given a primitive binary quadratic form $f = aX^2+bXY+cY^2 \in \mathbf{Z}[X,Y]$ of positive discriminant, the two roots of $f(X,1)$ have a canonical ordering as

$$\left(\frac{-b+\sqrt{\mathrm{disc}(f)}}{a}, \frac{-b-\sqrt{\mathrm{disc}(f)}}{a} \right).$$

So you can take the geodesic γ on $\mathbf{H}$ going from the first root to the second root, then obtain a prime geodesic on $Y(1)$ by looking at a fundamental domain of the action of $\mathrm{Stab}_{SL_2(\mathbf{Z})}(\gamma)$ on γ. One computes directly that the length of the resulting prime geodesic is $2R_{\mathrm{disc}(f)}$ and that this is bijective. Note that we are still following the convention that the time-reversal of a prime geodesic is not necessarily the same one: taking the time-reversal on the right hand side of the bijection corresponds to negating all the coefficients on the left hand side (since then the ordered pair of roots will be reversed).

Now that the bijection with prime geodesics is established, we may compute

$$\begin{aligned} \sum_{R_d \leq \log x} h(d) &= \# \left\{ \begin{matrix} SL_2(\mathbf{Z})\text{-equivalence classes of primitive binary} \\ \text{quadratic forms } f \text{ such that } 2R_{\mathrm{disc}(f)} \leq \log x^2 \end{matrix} \right\} \\ &= \#\{\text{Prime geodesics on } Y(1) \text{ of length } \leq \log x^2\} \\ &\sim_{x\to\infty} Li(x^2) \end{aligned}$$

as a consequence of Theorem 4.15 (which we have remarked is also valid for finite-volume quotients). □

This result was a big step towards the long-open question of decoupling the regulator from the class number in the Gauss–Siegel asymptotic formula for $\sum_{d<x} h(d) \log \epsilon_d$ [22] and its consequences and refinements (e.g. for $\Gamma = \Gamma(p)$) are elaborated on further in [20]

Remark 4.20. Here is a vague question. The class numbers of real quadratic fields come up when applying the trace formula to Maass forms, as seen in this application. When applying the trace formula to spaces of holomorphic modular forms to compute traces of

Hecke operators, it is the class numbers of imaginary quadratic fields that show up. Is this a coincidence? Does it generalize?

Acknowledgements

Thanks are due to Mark Kisin, for suggesting that I study the trace formula and its applications; Fabian Gundlach, whose arithmetic statistics topics course introduced me to the asymptotic averaging problem for class numbers and made me aware of the existence of Sarnak's work on the topic; Wei Zhang, for teaching me about the trace formula in his course on automorphic forms; and Peter Sarnak, for generously sending me his thesis, for his advice on learning about the Selberg trace formula, and his detailed response to my question about the Tauberian differentiation argument in the proof of Theorem 4.15.

These notes are about material I learned about and presented at the University of Chicago number theory graduate students' automorphic forms learning seminar, organized by Hao Billy Lee. So, I also thank Billy for organizing the seminar, and the other people who have explained things to me in and outside the seminar: Gal Porat, Brian Lawrence, Bingjin Liu, Jason Kountouridis, Noah Taylor, Aaron Slipper, Yulia Kotelnikova, and Chengyang Bao.

Finally, I thank J.P. May for organizing the 2020 University of Chicago REU, without which I would not have met Billy and joined the learning seminar.

References

[1] J. Arthur. An introduction to the trace formula. *Clay Mathematics Proceedings*, 4:1–263, 2005.
[2] A.R Booker, A. Strömbergsson, and A. Venkatesh. Effective computation of maass cusp forms. *International mathematics research notices*, 2006, 2006.
[3] D. Bump. *Automorphic forms and representations*, volume 55 of *Cambridge Studies in Advanced Mathematics*. Cambridge university press, 1998.
[4] D. Bump. *Lie groups*. Springer, 2004.
[5] M. Duflo and J.-P. Labesse. Sur la formule des traces de selberg. In *Annales scientifiques de l'École Normale Supérieure*, volume 4, pages 193–284, 1971.
[6] I.M. Gelfand, M.I. Graev, and I. Piatetski-Shapiro. *Representation theory and automorphic functions*, volume 6. Saunders, 1968.
[7] Harish-Chandra. Representations of a semisimple lie group on a banach space. i. *Transactions of the American Mathematical Society*, 75(2):185–243, 1953.
[8] Harish-Chandra. Representations of semisimple lie groups. ii. *Transactions of the American Mathematical Society*, 76(1):26–65, 1954.
[9] Harish-Chandra. Representations of semisimple lie groups. iii. *Transactions of the American Mathematical Society*, 76(2):234–253, 1954.
[10] D.A. Hejhal. *The Selberg trace formula for $PSL(2, \mathbf{R})$*, volume 2. Springer, 2006.
[11] J. Karamata. Neuer beweis und verallgemeinerung der tauberschen satze welche die laplaceshe und stieltjessche transformation betreffen. *J. reine angew. Math.*, 164:27–39, 1931.
[12] A.W. Knapp. *Representation theory of semisimple groups: an overview based on examples*, volume 36. Princeton university press, 2001.
[13] A. Knightly and C. Li. *Traces of Hecke operators*. Number 133. American Mathematical Soc., 2006.
[14] S. Lang. $SL_2(\mathbf{R})$, volume 105 of *Graduate Texts in Mathematics*. Springer Science & Business Media, 1985.
[15] R.P. Langlands. Irreducible representations of real algebraic groups. *Representation theory and harmonic analysis on semisimple Lie groups*, 31:101–170, 1989.
[16] R.P. Langlands. *On the functional equations satisfied by Eisenstein series*, volume 544. Springer, 2006.
[17] W. Luo and P. Sarnak. Quantum ergodicity of eigenfunctions on $PSL_2(\mathbf{Z})/H^2$. *Publications Mathématiques de l'Institut des Hautes Études Scientifiques*, 81(1):207–237, 1995.

[18] J. Marklof. Selberg's trace formula: An introduction. *Hyperbolic geometry and applications in quantum chaos and cosmology*, 397:83, 2012.
[19] P.C. Sarnak. *Prime geodesic theorems*. PhD thesis, Stanford University, 1981.
[20] P.C. Sarnak. Class numbers of indefinite binary quadratic forms. *Journal of Number Theory*, 15(2):229–247, 1982.
[21] A. Selberg. Harmonic analysis and discontinuous groups in weakly symmetric spaces with applications to dirichlet series. *J. Indian Math. Soc.*, 20:47–87, 1956.
[22] C.L. Siegel. The average measure of quadratic forms with given determinant and signature. *Annals of Mathematics*, pages 667–685, 1944.
[23] J.H. Silverman. *Advanced topics in the arithmetic of elliptic curves*, volume 151 of *Graduate Texts in Mathematics*. Springer Science & Business Media, 1994.