Contents

1 Introduction **4**
- 1.1 What is an automorphic form? . . . 4
- 1.2 A rough definition of automorphic forms on Lie groups . . . 5
- 1.3 Specializing to $G = \mathrm{SL}(2, \mathbb{R})$ . . . 5
- 1.4 Goals for the course . . . 7
- 1.5 Recommended Reading . . . 7

2 Automorphic forms from elliptic functions **8**
- 2.1 Elliptic Functions . . . 8
- 2.2 Constructing elliptic functions . . . 9
- 2.3 Examples of Automorphic Forms: Eisenstein Series . . . 14
- 2.4 The Fourier expansion of G_{2k} . . . 17
- 2.5 The j-function and elliptic curves . . . 19

3 The geometry of the upper half plane **19**
- 3.1 The topological space $\Gamma\backslash\mathcal{H}$ . . . 20
- 3.2 Discrete subgroups of $\mathrm{SL}(2, \mathbb{R})$ . . . 22
- 3.3 Arithmetic subgroups of $\mathrm{SL}(2, \mathbb{Q})$ . . . 23
- 3.4 Linear fractional transformations . . . 24
- 3.5 Example: the structure of $SL(2, \mathbb{Z})$ . . . 27
- 3.6 Fundamental domains . . . 28
- 3.7 $\Gamma\backslash\mathcal{H}^*$ as a topological space . . . 31
- 3.8 $\Gamma\backslash\mathcal{H}^*$ as a Riemann surface . . . 34
- 3.9 A few basics about compact Riemann surfaces . . . 35
- 3.10 The genus of $X(\Gamma)$ . . . 37

4 Automorphic Forms for Fuchsian Groups **40**
- 4.1 A general definition of classical automorphic forms . . . 40
- 4.2 Dimensions of spaces of modular forms . . . 42
- 4.3 The Riemann-Roch theorem . . . 43
- 4.4 Proof of dimension formulas . . . 44
- 4.5 Modular forms as sections of line bundles . . . 46
- 4.6 Poincaré Series . . . 48
- 4.7 Fourier coefficients of Poincaré series . . . 50
- 4.8 The Hilbert space of cusp forms . . . 54
- 4.9 Basic estimates for Kloosterman sums . . . 56
- 4.10 The size of Fourier coefficients for general cusp forms . . . 60

5 L-functions associated to cusp forms **63**
5.1 The Mellin transform 63
5.2 Weil's converse theorem 69

6 Hecke Operators **73**
6.1 Initial Motivation 73
6.2 Hecke operators on lattices 75
6.3 Explicit formulas for Hecke operators 79
6.4 A geometric definition of Hecke operators 82
6.5 Hecke operators as correspondences 87
6.6 Brief remarks on Hecke operators for Fuchsian groups 88

7 Modularity of elliptic curves **90**
7.1 The modular equation for $\Gamma_0(N)$. 90
7.2 The canonical model for $X_0(N)$ over $\mathbb{Q}$ 93
7.3 The moduli problem and $X_0(N)$ 97
7.4 Hecke Correspondences for $Y_0(N)$ 99
7.5 The Frobenius Map 101
7.6 Eichler-Shimura theory 103
7.7 Zeta functions of curves over finite fields 105
7.8 Hasse-Weil L-functions of a curve over $\mathbb{Q}$ 107
7.9 Eichler-Shimura theory in genus one 108

8 Defining automorphic forms on groups **111**
8.1 Eigenfunctions of Δ on $\mathcal{H}$ – Maass forms 111
8.2 Automorphic forms on groups – rough version 117
8.3 Basic Lie theory for $\mathrm{SL}(2,\mathbb{R})$ 118
8.4 K-finite and $\mathcal{Z}$-finite functions 120
8.5 Distributions and convolution of functions 121
8.6 A convolution identity for $\mathcal{Z}$-finite, K-finite functions 126
8.7 Fundamental domains again - Siegel sets 132
8.8 Growth conditions 135
8.9 Definition and first properties of an automorphic form 136

9 Finite dimensionality of automorphic forms **138**
9.1 Constant term estimates and cuspidality 138
9.2 Finite dimensionality of automorphic forms 142

10 Eisenstein series **147**
10.1 Definition and initial convergence 147
10.2 Constant terms of Eisenstein series 156
10.3 Outline of Proof of Analytic Continuation 160
10.4 The meromorphic continuation principle 167
10.5 Continuation consequences: functional equations, etc. 171
10.6 Spectral decomposition of $L^2(\Gamma\backslash G)$ 173

11 Applications **176**
11.1 Review of representation theory 176
11.2 Representations of $\mathrm{SL}(2,\mathbb{R})$ 177

These are (not terribly original) notes for the lectures. For further reading, see the bibliography at the end of Section 1.

1 Introduction

1.1 What is an automorphic form?

An automorphic form is a generalization of a certain class of periodic functions. To understand their definition, we begin with the simpler example of periodic functions on the real line.

We consider functions $f : \mathbb{R} \to \mathbb{C}$ which are (for simplicity) periodic with period 1. That is,

$$f(x+n) = f(x) \quad \text{for all } x \in \mathbb{R} \text{ and all } n \in \mathbb{Z}$$

We want to use differential calculus to study them, so we use a translation invariant measure. In this case, its the familiar Lebesgue measure for $\mathbb{R}$ which we'll just denote by the usual dx. Generally speaking, such a translation invariant measure on a topological group is referred to as a "Haar measure."

A very powerful tool for studying such functions is harmonic analysis. That is, we want to find an orthogonal system with which we may express any (reasonably nice) periodic functions (e.g. integrable functions on $\mathbb{R}/\mathbb{Z}$) – this is often referred to as a "complete orthogonal system." Since orthogonal systems are, by assumption, countable we may denote its members rather suggestively by $e_n(x)$.

Then (after normalizing to obtain an orthonormal basis) we may expand f as follows:

$$f(x) = \sum_{n \in \mathbb{Z}} a_n \, e_n(x), \quad \text{where} \quad a_n = \int_0^1 f(x) \overline{e_n(x)} \, dx.$$

As most of you probably already know, one such choice for this orthogonal system is to set $e_n(x) = e^{2\pi i n x}$ for each $n \in \mathbb{Z}$. But why this choice?

For inspiration, we can go back to the original problem that motivated Fourier himself. He was seeking a general solution for the heat equation in a thin metal plate. It was known that if the heat source was expressible as a sinusoidal wave, then the solution was similarly expressible as a sinusoidal wave. Fourier's idea was to use a superposition of these waves to attack the problem for an arbitrary function as heat source. In short, a good basis for the solution to his problem came from the eigenfunctions of a differential operator!

Which differential operator? In this case, there's a particularly natural choice – the Laplace operator Δ associated to our metric dx. In general, this can be defined

on any Riemannian manifold by $\Delta(f) = \operatorname{div} \operatorname{grad} f$. In the coordinate x, this is just $\frac{d^2}{dx^2}$ for our example. Indeed, the spectral theorem for compact, self-adjoint operators guarantees that the collection of eigenfunctions will give an orthogonal basis. (The word "spectrum" for an operator T here means the set of elements λ such that $T - \lambda I$ is not invertible. This notion will play an especially important role for us in the course.)

To the extent that this basis is good for solving the problem at hand, Fourier analysis can be an extremely useful tool.

1.2 A rough definition of automorphic forms on Lie groups

To generalize from the setting above, we let X be a locally compact space with a discontinuous group action by a discrete group Γ. Then an *automorphic function* $f : X \to \mathbb{C}$ is just a function invariant with respect to this action:

$$f(\gamma x) = f(x) \quad \text{for all } \gamma \text{ in } \Gamma.$$

Recall that a discontinuous action of a discrete group Γ on a topological space X means that for any point x in X, there exists a neighborhood U_x of x such that $\gamma(U_x) \cap U_x = \emptyset$ for all non-trivial γ in Γ.

An important special case is when $X = G/K$, where G is a locally compact group, and K is a closed subgroup. For example, if G is a Lie group, then this quotient is a Riemannian manifold which admits a differential calculus. In general, a deep theorem from functional analysis asserts that all locally compact groups have a (unique up to constant left- or right-) Haar measure. Then an *automorphic form* is a simultaneous eigenfunction of the algebra $\mathcal{D}$ of invariant differential operators on X. This algebra includes the Laplace-Beltrami operator.

Typically, we allow for a more general transformation law for these functions:

$$f(\gamma x) = j(\gamma, x) f(x)$$

where j may consist of differential factors and multiplier systems. (More on those in later lectures). For now, we content ourselves with an example of such a j. Again with x as a coordinate on the manifold X and suppose $d(\gamma x) = c(\gamma, x) d(x)$. Then setting $j(\gamma, x) = c(\gamma, x)^{-1}$, we obtain all 1-forms on $\Gamma \backslash G/K$ that are eigenfunctions of $\mathcal{D}$.

1.3 Specializing to $G = \mathrm{SL}(2, \mathbb{R})$

For us, our main example will be $G = \mathrm{SL}(2, \mathbb{R})$ and $K = \mathrm{SO}(2, \mathbb{R})$. Remember, that

$$\mathrm{SL}(2, \mathbb{R}) = \{\gamma \in \mathrm{Mat}_{\mathbb{R}}(2 \times 2) \mid \det(\gamma) = 1\},$$

$$\mathrm{SO}(2,\mathbb{R}) = \{\gamma \in \mathrm{Mat}_{\mathbb{R}}(2\times 2) \mid \gamma^T\gamma = I, \det(\gamma) = 1\} = \left\{ \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \middle| \, \theta \in [0, 2\pi) \right\}.$$

Then the resulting space $X = G/K$ is isomorphic to the complex upper half-plane $\mathcal{H} = \{z \in \mathbb{C} \mid \Im(z) > 0\}$. This isomorphism can be seen from the fact that G acts transitively on the point i in the upper half plane by linear fractional transformation:

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} : z \to \frac{az+b}{cz+d},$$

and the stabilizer of i is K.

For the discrete group Γ, we often will take $\mathrm{SL}(2,\mathbb{Z})$ or a finite index subgroup. The most important class of subgroups are called "congruence subgroups." They are subgroups which contain

$$\Gamma(n) = \{\gamma \in \mathrm{SL}(2,\mathbb{Z}) \mid \gamma \equiv I \pmod{n}\}$$

for some positive integer n.

To finish the definition, we need only note that the algebra of invariant differential operators is one dimensional, generated by the Laplacian for $\mathcal{H}$. Here the invariant measure (in terms of Lebesgue measure dx and dy) is $d\mu = y^{-2}dxdy$ with associated Laplacian

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

where, as usual,

$$\frac{\partial}{\partial z} = \frac{1}{2}\left(\frac{\partial}{\partial x} - i\frac{\partial}{\partial y}\right), \quad \frac{\partial}{\partial \overline{z}} = \frac{1}{2}\left(\frac{\partial}{\partial x} + i\frac{\partial}{\partial y}\right).$$

The presence of the $\frac{\partial}{\partial \overline{z}}$ on the right in Δ explains why "classical automorphic forms" are defined as complex analytic functions $f : \mathcal{H} \to \mathbb{C}$ which satisfy a transformation law of the form:

$$f(\gamma z) = (cz+d)^k f(z) \quad \text{for all } \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$$

for some fixed integer k.

In the next lecture, we'll investigate examples of these functions for $\Gamma = \mathrm{SL}(2,\mathbb{Z})$, and then begin addressing aspects of the theory for general discrete subgroups of $SL(2,\mathbb{R})$. An old result of Poincaré states that a subgroup of $\mathrm{SL}(2,\mathbb{R})$ is discrete if and only if it acts discontinuously on $\mathcal{H}$ as a subgroup of $\mathrm{PSL}(2,\mathbb{R})$. Such groups are called *Fuchsian groups* and we'll spend a couple of days making a careful study of them.

1.4 Goals for the course

The goal of this course is to present several applications of automorphic forms to number theory, arithmetic geometry, and functional analysis. Here are several topics we'll cover in these areas:

- (Number Theory) We'll use theta functions to study the problem of representing integers using binary quadratic forms. If time permits, we may address finer questions about the equidistribution of the solutions on the surface $Q(x) = n$ as the integer n tends to infinity.

- (Arithmetic Geometry) We use Fourier coefficients of automorphic forms to build generating functions (called "L-functions") which are (at least conjecturally) connected to generating functions coming from arithmetic geometry. We'll prove a simple case of this correspondence between Hecke L-functions and those coming from elliptic curves with complex multiplication.

- (Geometric and Functional Analysis) We will study the spectrum of the Laplace-Beltrami operator for the quotient space $\Gamma\backslash G/K$ for various choices of Γ. When the group Γ is arithmetic, this leads to interesting consequences in number theory. Our main tool here will be the Selberg Trace Formula, which leads to asymptotics for these eigenvalues in the form of Weyl's law.

Interestingly there are unifying themes to the study of these three problems. One is the need to understand the Fourier coefficients of automorphic forms - sometimes via exact formulas and at other times using estimates. Another is the use of Hecke operators to study the cases when Γ is arithmetic.

If time permits, I'm hoping to discuss the translation between classical automorphic forms and automorphic representations of adele groups. This is somehow slightly cheating, since this is more algebraic than analytic, but it's very important to know both languages when working in automorphic forms.

1.5 Recommended Reading

Here are a few books which cover some of the material we'll discuss.

1. D. Bump, "Automorphic Forms and Representations," Cambridge Studies in Advanced Math. v. 55, (1997).

2. D. Hejhal, "The Selberg Trace Formula for PSL(2,R)," Springer Lecture Notes in Math. 1001 (1983).

3. H. Iwaniec, "Spectral Methods of Automorphic Forms," Second Edition. AMS Graduate Studies v. 53, (2002).

4. H. Iwaniec, "Topics in Classical Automorphic Forms," AMS Graduate Studies v. 17, (1997).

5. P. Sarnak, "Some Applications of Modular Forms," Cambridge Univ. Press, (1990)

6. G. Shimura "Introduction to the Arithmetic Theory of Automorphic Functions," Princeton Univ. Press, (1971).

2 Automorphic forms from elliptic functions

In this lecture, we'll give alternate motivation for the appearance of automorphic forms, which will lead directly to a construction of specific examples. It's also the very first case in the study of Shimura varieties, which seeks choices of G, K and Γ (using the notation of the previous lecture) so that the resulting quotient space $\Gamma\backslash G/K$ has an interpretation as a moduli space of abelian varieties.

For us, $G = \mathrm{SL}(2,\mathbb{R})$, $K = \mathrm{SO}(2,\mathbb{R})$ and $\Gamma = \mathrm{SL}(2,\mathbb{Z})$, as before. We'll give an interpretation of $\Gamma\backslash G/K \simeq \Gamma\backslash\mathcal{H}$ as a moduli space of elliptic curves. The basic idea is to identify an elliptic curve with a lattice in $\mathbb{C}$, and then lattices with a point in the upper half-plane. From this investigation, automorphic forms will appear naturally. To explore the connection between elliptic curves and lattices, we take a step back and consider *elliptic functions* on $\mathbb{C}$.

2.1 Elliptic Functions

An elliptic function is a meromorphic function on $\mathbb{C}$ which is periodic with set of periods a full lattice Λ in $\mathbb{C}$. Any such lattice can be written as the set:

$$\Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2$$

for a pair of generators ω_1 and ω_2 in $\mathbb{C}$ which are linearly independent over $\mathbb{R}$. Note that elliptic functions for a fixed lattice Λ form a field. Any elliptic function f having no poles is both holomorphic and bounded on $\mathbb{C}$. Hence by Liouville's theorem, any such f must be constant.

We now choose a set of representatives for $\mathbb{C}/\Lambda$ corresponding to a point α in $\mathbb{C}$ as follows:

$$\alpha + t_1\omega_1 + t_2\omega_2, \quad 0 \leq t_1, t_2 < 1.$$

We refer to such a set as a *fundamental parallelogram* (with respect to the given basis).

Proposition 2.1.1. *Given an elliptic function f, we may choose a fundamental parallelogram P such that*

$$\sum_{w \in P} \operatorname{res}_f(w) = 0.$$

Proof. Choose a fundamental parallelogram P such that there are no zeros or poles on the boundary ∂P. (This is possible because any non-zero meromorphic function has only finitely many such zeros and poles in any fundamental parallelogram.) Then integrating along the boundary, Cauchy's theorem immediately gives the result since the contributions on opposite sides of the parallelogram cancel by periodicity. □

Corollary 2.1.2. *An elliptic function has at least two poles in its fundamental parallelogram.*

Proposition 2.1.3. *Let f be an elliptic function with fundamental parallelogram P whose boundary avoids zeros and poles. Let a_i be the location of the singular points of f (i.e., the zeros and poles) inside P. Let m_i be the corresponding order (with sign) at each a_i. Then*

$$\sum_{a_i \in P} m_i = 0.$$

Proof. Because f is elliptic, so is its logarithmic derivative f'/f. Integrating f'/f over the boundary ∂P and applying Proposition 2.1.1 gives the result. (Indeed, according to the "argument principle" the orders of the singular set of f are the residues of the poles of f'/f.) □

Proposition 2.1.4. *With all hypotheses as in Proposition 2.1.3,*

$$\sum_{a_i \in P} m_i a_i \equiv 0 \pmod{\Lambda}$$

Proof. We leave this as an exercise to the reader. Hint: it uses a similar contour integration as in the argument principle above. □

2.2 Constructing elliptic functions

One natural way to construct such functions is by averaging over the lattice. A first guess might be:

$$f(z) = \sum_{\lambda \in \Lambda} \frac{1}{(z - \lambda)^2}$$

but the terms in this series are too big, and don't converge. Instead, we have to do something a bit more clever. Define the *Weierstrass $\wp$-function* by

$$\wp(z) = \frac{1}{z^2} + \sum_{\lambda \in \Lambda \setminus \{0\}} \left[\frac{1}{(z-\lambda)^2} - \frac{1}{\lambda^2} \right]$$

We need to address the issue of convergence – we must show that it converges uniformly on compact sets not including lattice points. Note that for z on compact sets not including lattice points, the summands:

$$\frac{1}{(z-\lambda)^2} - \frac{1}{\lambda^2} = \frac{z^2 - 2z\lambda}{\lambda^2 (z-\lambda)^2}$$

have order of magnitude $\frac{1}{|\lambda|^3}$. We thus obtain convergence with the following lemma:

Lemma 2.2.1. *If $s > 2$, the series*

$$\sum_{\lambda \in \Lambda \setminus \{0\}} \frac{1}{|\lambda|^s}$$

converges.

Proof. The proof is left as an exercise to the reader. One approach is to use an estimate for the number of lattice points in the annulus defined by circles of radius $n-1$ and n, centered at the origin. □

More importantly, we haven't even demonstrated that the function $\wp(z)$ has period lattice Λ. Notice that its derivative

$$\wp'(z) = -2 \sum_{\lambda \in \Lambda} \frac{1}{(z-\lambda)^3}$$

is an elliptic function, as it is clearly invariant under the period lattice and converges uniformly on compact sets away from lattice points by the previous lemma.

Proposition 2.2.2. *The function $\wp(z)$ is elliptic.*

Proof. Let ω_1 be one of the generators of the lattice. Then because the function $\wp'(z)$ is elliptic, for any point z not a pole,

$$\wp'(z) = \wp'(z + \omega_1) \Rightarrow \wp(z) = \wp(z + \omega_1) + C,$$

for some constant C. But $\wp(z)$ is clearly even, according to its definition, and so choosing $z = -\omega_1/2$ we see that $C = 0$. The same argument works for the other generator of the lattice ω_2, which ensures the required periodicity for all points in the lattice. □

Theorem 2.2.3. *The field of elliptic functions (with respect to the lattice Λ) is generated by $\wp$ and $\wp'$.*

Proof. Any function can be decomposed into even and odd parts via the familiar identity:

$$f(z) = \frac{f(z) + f(-z)}{2} + \frac{f(z) - f(-z)}{2}$$

and each of these pieces are elliptic. If f is odd, then $\wp' f$ is even, so it suffices to prove that any even elliptic function f is a rational function in $\wp$.

To this end, note that if f is even and has a zero or pole of order m at some point a, then f has a zero or pole of order m at $-a$, simply because

$$f^{(k)}(a) = (-1)^k f^{(k)}(-a).$$

Lemma 2.2.4. *Suppose f has a zero (or pole) at $a \equiv -a \pmod{\Lambda}$, then f has a zero (or pole) of even order at a.*

Proof. We prove this for a zero, as the proof for poles follows similarly. There are exactly four points with the property that $a \equiv -a \pmod{\Lambda}$ in the fundamental parallelogram P, represented by

$$0, \frac{\omega_1}{2}, \frac{\omega_2}{2}, \frac{\omega_1 + \omega_2}{2}.$$

Since f even and elliptic implies f' is odd and elliptic, so for these points a above, $f'(a) = 0$, so f has a zero of order at least 2. When applied to the function

$$g(z) = \wp(z) - \wp(a)$$

for any of the three representatives $a \not\equiv 0 \bmod \Lambda$, we see that $g(z)$ has a zero of order at least 2 at a. Hence the order of the zero must be exactly 2, since Proposition 2.1.3 guarantees the sum of orders in P are 0 and the only pole is of order 2 at the origin. Now consider $f(z)/g(z)$, which is even, elliptic, and holomorphic at a. If $f(a)/g(a)$ is non-zero, we're done. Otherwise, repeat the argument. To handle the remaining case of $u \equiv 0 \bmod \Lambda$, we may use $1/\wp$ instead for g. This completes the proof for zeros of f. □

To finish the theorem, given any even elliptic function f, label the singular points a_i with orders m_i as before. If a_i happens to be one of the points with representative $-a_i$ in P, then choose one of the two representatives. Consider the product:

$$\prod_{a_i} [\wp(z) - \wp(a_i)]^{m_i'}$$

where $m_i' = m_i$ unless $a_i \equiv -a_i$, and then $m_i' = m_i/2$. By previous lemma, for all $z \not\equiv 0 \bmod \Lambda$, this function has the same order at z as f. Applying Proposition 2.1.3 to the product, this is also true at the origin. Hence the quotient of the f and the product above is an elliptic function without zeros or poles, hence constant. $\square$

Finally, we give an algebraic relation between $\wp$ and $\wp'$ by comparing their power series expansions at the origin. Such a relation is to be expected because elliptic functions without poles are constant.

We compute the power series for $\wp(z)$ by expanding the summands in a geometric series:

$$\begin{aligned}
\wp(z) &= \frac{1}{z^2} + \sum_{\lambda \in \Lambda \setminus \{0\}} \left[\frac{1}{(z-\lambda)^2} - \frac{1}{\lambda^2} \right] \\
&= \frac{1}{z^2} + \sum_{\lambda \in \Lambda \setminus \{0\}} \left[\frac{1}{\lambda^2} \left(1 + \frac{z}{\lambda} + \frac{z^2}{\lambda^2} + \cdots \right)^2 - \frac{1}{\lambda^2} \right] \\
&= \frac{1}{z^2} + \sum_{\lambda \in \Lambda \setminus \{0\}} \frac{1}{\lambda^2} \sum_{n=1}^{\infty} (n+1) \left(\frac{z}{\lambda} \right)^n \\
&= \frac{1}{z^2} + \sum_{n=1}^{\infty} (n+1) G_{n+2} z^n
\end{aligned}$$

where

$$G_n = \sum_{\lambda \neq 0} \frac{1}{\lambda^n} \qquad \text{(Note that } G_n = 0 \text{ if } n \text{ is odd).} \tag{1}$$

Differentiating term by term, we obtain:

$$\wp'(z) = \frac{-2}{z^3} + \sum_{n=1}^{\infty} n(n+1) G_{n+2} z^{n-1} = \frac{-2}{z^3} + 6G_4 z + 20 G_6 z^3 + \cdots$$

From these expansions, we obtain:

Proposition 2.2.5. *The elliptic functions $\wp$ and $\wp'$ satisfy the relation:*

$$(\wp'(z))^2 = 4\wp(z)^3 - 60 G_4 \wp(z) - 140 G_6.$$

Proof. Comparing power series expansions at $z = 0$ for both sides, we see their difference is an elliptic function with no poles (because we've removed the pole at 0 and $\wp$ and $\wp'$ are holomorphic elsewhere) hence the difference must be a constant function. Since the power series expansion of the difference has constant term equal to 0, the difference must be identically 0. $\square$

Moreover, we can factor the right-hand side.

Proposition 2.2.6. *We have the identity*

$$(\wp'(z))^2 = 4(\wp(z) - e_1)(\wp(z) - e_2)(\wp(z) - e_3),$$

where we have defined

$$e_1 = \wp\left(\frac{\omega_1}{2}\right), \quad e_2 = \wp\left(\frac{\omega_2}{2}\right), \quad e_3 = \wp\left(\frac{\omega_1 + \omega_2}{2}\right).$$

Proof. As previously discussed, $\wp(z) - \wp(\omega_1/2)$ has a zero at $\omega_1/2$ of order 2, so $\wp'(\omega_1/2) = 0$. Moreover, $\wp(\omega_1), \wp(\omega_2)$, and $\wp(\omega_1 + \omega_2/2)$ must be distinct complex numbers, else one of $\wp(z) - \wp(\omega_i/2)$ would have zeros and poles in contradiction to Proposition 2.1.3. Comparing the zeros and poles of the two sides of the identity to be proved, we see again that their difference must be 0. □

In fact we've also proven the following:

Corollary 2.2.7. *Let G_4 and G_6 be complex numbers defined as above. Then the polynomial*

$$y^2 = 4x^3 - 60G_4x - 140G_6$$

has non-zero discriminant Δ. That is,

$$\Delta = (60G_4)^3 - 27(140G_6)^2 = 16(e_1 - e_2)^2(e_2 - e_3)^2(e_1 - e_3)^2 \neq 0.$$

This condition guarantees that the associated cubic curve is non-singular, i.e. an "elliptic curve."

Theorem 2.2.8. *Let E be the elliptic curve corresponding to the lattice Λ given by*

$$y^2 = 4x^3 - 60G_4x - 140G_6$$

Then the map

$$\begin{array}{rccl} \phi: & \mathbb{C}/\Lambda & \longrightarrow & E \subset \mathbb{P}^2(\mathbb{C}) \\ & z & \longmapsto & [\wp(z), \wp'(z), 1] \end{array}$$

is a complex analytic isomorphism of complex Lie groups. (I.e. it's simultaneously an isomorphism of Riemann surfaces and a group homomorphism.)

Proof. The image of ϕ is contained in E by Proposition 2.2.6. For the remaining details, see Silverman, "The Arithmetic of Elliptic Curves," Proposition VI.3.6(b). □

In fact, much more can be said. It turns out that if E_1 and E_2 are elliptic curves defined over $\mathbb{C}$, the two curves are isomorphic over $\mathbb{C}$ if and only if their corresponding lattices Λ_1 and Λ_2 are *homothetic*. That is, there exists a non-zero complex number α such that $\Lambda_1 = \alpha\Lambda_2$. (See section VI.4 of Silverman for the proof.)

Furthermore, the "Uniformization Theorem" for elliptic curves says that for any A and B with $A^3 - 27B^2 \neq 0$, there is a unique lattice $\Lambda \subset \mathbb{C}$ such that $60G_4 = A$ and $140G_6 = B$. (See, for example, Serre's "Course in Arithmetic," Proposition VII.5.)

But where are the automorphic forms?

2.3 Examples of Automorphic Forms: Eisenstein Series

We now return to the problem of constructing automorphic forms on $\mathrm{SL}(2,\mathbb{Z})\backslash\mathcal{H}$. In our previous discussion of elliptic functions, we regarded the lattice Λ as fixed. Now we consider what happens when we allow the lattice to vary. First, we give a description of the space of lattices.

Proposition 2.3.1. *The set $\mathcal{L}/\mathbb{C}^\times$ of lattices in $\mathbb{C}$ modulo homothety may be identified with $G\backslash\mathcal{H}$ where $G = \mathrm{SL}(2,\mathbb{Z})/\{\pm I\}$.*

Proof. Consider first the set M of pairs (ω_1, ω_2) such that $\Im(\omega_1/\omega_2) > 0$. Then the map:

$$\begin{aligned} M &\longrightarrow \mathcal{L} \\ (\omega_1, \omega_2) &\longmapsto \Lambda = \mathbb{Z}\omega_1 + \mathbb{Z}\omega_2 \end{aligned}$$

which is clearly surjective as the condition on the imaginary part can always be satisfied by reordering ω_1 and ω_2. Furthermore, $\mathrm{SL}(2,\mathbb{Z})$ acts on this set by matrix multiplication (as one checks that it preserves the condition that $\Im(\omega_1/\omega_2) > 0$.)

Moreover, two elements m, m' of M define the same lattice if and only if $\gamma m = m'$ for some γ in $\mathrm{SL}(2,\mathbb{Z})$. This is clearly sufficient, as the transformation is invertible. If two pairs of complex numbers define the same lattice, they are related by a matrix with integer coefficients and determinant ± 1. The sign of the determinant is $+1$ if the sign of $\Im(\omega_1/\omega_2)$ is preserved, so the fact that both pairs are from M implies the matrix is indeed in $\mathrm{SL}(2,\mathbb{Z})$.

To obtain the result, note that the quotient $M/\mathbb{C}^\times$ may be identified with $\mathcal{H}$ via the map $(\omega_1, \omega_2) \mapsto z = \omega_1/\omega_2$. The action of $\mathrm{SL}(2,\mathbb{Z})$ on M transforms into an action of G on $\mathcal{H}$. □

Combining this result with those of the previous section, we conclude:

Corollary 2.3.2. *The quotient space $G\backslash\mathcal{H}$ is in bijection with the set of isomorphism classes of elliptic curves.*

The transformation property for classical automorphic forms arises naturally from homogeneous functions on lattices. First, we recall the definition of these transformations.

Definition 2.1. *A function $f : \mathcal{H} \to \mathbb{C}$ is said to be* **weakly modular** *of weight k if*

$$f(\gamma z) = (cz + d)^k f(z) \quad \text{for all } \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2, \mathbb{Z}).$$

(Note: Some authors differ on whether such f should also be required to be meromorphic (e.g. Serre) or not (e.g. Milne). We will not assume this in order to cleanly state the following result.)

Proposition 2.3.3. *Let $F : \mathcal{L} \to \mathbb{C}$ be a homogeneous function of weight k on the set of lattices. That is,*

$$F(\alpha\Lambda) = \alpha^{-k} F(\Lambda) \quad \text{for all } \alpha \in \mathbb{C}^\times.$$

Then we may define a function $f(z) := F(\Lambda(z, 1))$ on $\mathcal{H}$ according to Proposition 2.3.1. The function f is weakly modular of weight k and this identification defines a bijection between homogeneous functions of weight k on the set of lattices $\mathcal{L}$ and weakly modular functions on $\mathcal{H}$.

Proof. The function f is well-defined because of the homogeneity of F, so that F only depends on the ratio $\omega_1/\omega_2 = z$. Write

$$F(\Lambda(\omega_1, \omega_2)) = \omega_2^{-k} f(\omega_1/\omega_2). \tag{2}$$

Because F is a function on lattices, it is invariant under the action of $\mathrm{SL}(2, \mathbb{Z})$:

$$F(\Lambda(a\omega_1 + b\omega_2, c\omega_1 + d\omega_2)) = F(\Lambda(\omega_1, \omega_2)), \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2, \mathbb{Z}).$$

Rewriting both sides of this equation in terms of (2) gives the weakly modular transformation property for f. Given a weakly modular function f on $\mathcal{H}$, the identification (2) may be used to define a homogeneous F on $\mathcal{L}$ as well. Indeed, all lattices are homothetic to one of form $\Lambda(z, 1)$ for some $z \in \mathcal{H}$ by Proposition 2.3.1 and this together with the homogeneity requirement determines F. □

We've already seen such homogeneous functions on the set of lattices in the previous section. These are the functions G_k of (1) that appeared in the power series expansion for $\wp(z)$ at $z = 0$. Considered as a function on lattices, we may write $G_k(\Lambda(\omega_1, \omega_2)) = \omega_2^{-k} G_k(z)$ where

$$G_k(z) = \sum_{(m,n)\neq(0,0)} (mz + n)^{-k}.$$

This series is known as the weight k Eisenstein series, and by Lemma 2.2.1 the series converges to a holomorphic function on $\mathcal{H}$ for $k > 2$.

Finally, we arrive at the definition of a modular form for $\mathrm{SL}(2, \mathbb{Z})$.

Definition 2.2. *A function $f : \mathcal{H} \to \mathbb{C}$ is called a modular form of weight k with respect to $\mathrm{SL}(2, \mathbb{Z})$ if it satisfies the following three conditions:*

1. *f is a weakly modular function of weight k for $\mathrm{SL}(2, \mathbb{Z})$,*
2. *f is holomorphic on $\mathcal{H}$,*
3. *f is "holomorphic at ∞."*

The last of these conditions requires further explanation. The first condition implies that f is invariant by the translation operator

$$T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} : z \mapsto z + 1.$$

Thus we may write $f(z) = g(e(z))$ where $e(z) = e^{2\pi i z}$. Letting $q = e^{2\pi i z}$, the resulting function $g(q)$ is defined on a punctured disc $0 < |q| < \epsilon$. The condition that f is "holomorphic at ∞" means that g is holomorphic at $q = 0$.

Equivalently, given a translation invariant function f, we may express the associated g as a power series at $q = 0$:

$$f(z) = g(q) = \sum_{n\in\mathbb{Z}} a_n q^n$$

and then this last condition simply means that the coefficients a_n vanish for $n < 0$. The coefficients a_n are the *Fourier coefficients of f*.

Remarks: The proper way to view this definition is that $\mathrm{SL}(2, \mathbb{Z})\backslash\mathcal{H}$ has a one point compactification by adding the point "$i\infty$" making the resulting space into a compact Riemann surface. We'll explore this more generally for any discrete group

Γ of $SL(2,\mathbb{R})$ in the coming lectures. Thus f may be viewed as a holomorphic function on this compact Riemann surface. This matches our earlier definition of an automorphic form from the first lecture, since the holomorphicity of f implies $\Delta(f) = 0$.

Some authors (e.g. Iwaniec) allow the function to be meromorphic everywhere (including "at infinity") but this is non-standard.

Proposition 2.3.4. *The functions G_k for $k > 2$ are modular forms of weight k. (Remember $G_k = 0$ if k odd.)*

Proof. By Lemma 2.2.1 and Proposition 2.3.3, it remains only to check that G_k is holomorphic at ∞. It suffices to show that the limit of G_k as $q \to \infty$, that is as $z \to i\infty$, exists. Since

$$G_k(z) = \sum_{(m,n)\neq(0,0)} (mz+n)^{-k},$$

we see that each summand vanishes in the limit unless $m = 0$ (and we may pass to the limit of summands by uniform convergence). Hence

$$\lim_{z\to i\infty} G_k(z) = 2\sum_{n=1}^{\infty} n^{-k} = 2\zeta(k).$$

□

2.4 The Fourier expansion of G_{2k}

We now compute the Fourier coefficient of G_{2k} using a standard trick based on the product formula for the sine function.

Lemma 2.4.1.

$$\pi\cot(\pi z) = \frac{1}{z} + \sum_{m=1}^{\infty}\left(\frac{1}{z+m} + \frac{1}{z-m}\right)$$

Proof. This follows immediately from logarithmic differentiation of the product formula for the sine function:

$$\sin\pi z = \pi z\prod_{m=1}^{\infty}\left(1-\frac{z}{m}\right)\left(1+\frac{z}{m}\right).$$

The product representation itself may be shown by comparing zeros of the two functions and applying Liouville's theorem. □

Theorem 2.4.2. *For any integer $k \geq 2$,*

$$G_{2k}(z) = 2\zeta(2k) + 2\frac{(2\pi i)^{2k}}{(2k-1)!}\sum_{n=1}^{\infty}\sigma_{2k-1}(n)q^n, \quad q = e^{2\pi i z},$$

where $\sigma_j(n)$ is the divisor function

$$\sigma_j(n) = \sum_{d|n} d^j$$

Proof. The previous lemma gives an expression for G_{2k} in terms of $\cot(z)$ upon taking $2k-1$ derivatives. We seek an alternate expression for $\cot(z)$ in terms of exponentials in order to obtain the Fourier series for G_{2k}. Note:

$$\cos(z) = \frac{e^{iz} + e^{-iz}}{2}, \quad \sin(z) = \frac{e^{iz} - e^{-iz}}{2i},$$

so

$$\cot(z) = i\frac{e^{iz} + e^{-iz}}{e^{iz} - e^{-iz}} = i - \frac{2i}{1 - e^{2iz}}.$$

Hence $\pi\cot(\pi z) = \pi i - 2\pi i\sum_{n=1}^{\infty} q^n$. Equating both expressions for $\pi\cot(\pi z)$ and taking $2k-1$ derivatives, we obtain:

$$\sum_{n\in\mathbb{Z}} \frac{1}{(z+n)^{2k}} = \frac{1}{(2k-1)!}(2\pi i)^{2k}\sum_{a=1}^{\infty} a^{2k-1}q^a. \tag{3}$$

Now

$$\begin{aligned}
G_{2k}(z) &= \sum_{(m,n)\neq(0,0)} (mz+n)^{-2k} \\
&= 2\zeta(2k) + 2\sum_{m=1}^{\infty}\sum_{n\in\mathbb{Z}}(mz+n)^{-2k} \\
&= 2\zeta(2k) + 2\frac{(2\pi i)^{2k}}{(2k-1)!}\sum_{m=1}^{\infty}\sum_{a=1}^{\infty} a^{2k-1}q^{ma} \quad \text{(by applying (3))} \\
&= 2\zeta(2k) + 2\frac{(2\pi i)^{2k}}{(2k-1)!}\sum_{n=1}^{\infty}\sigma_{2k-1}(n)q^n \quad \text{(after reindexing the sum).}
\end{aligned}$$

□

2.5 The j-function and elliptic curves

Throughout this section, we set $\Gamma = \mathrm{SL}(2,\mathbb{Z})$ for brevity. We have seen that $\Gamma\backslash\mathcal{H}$ is a moduli space for isomorphism classes of elliptic curves. Hence, if we can produce a modular function on $\mathcal{H}$ with respect to Γ (i.e. a weight 0 weakly modular function defined for all points in $\mathcal{H}$), this will be an invariant of elliptic curves.

Our smallest weight Eisenstein series are G_4 and G_6. It is further clear that modular forms with respect to Γ form a graded ring over $\mathbb{C}$, graded by the weight. So one way to produce a modular function is to take the quotient of two (linearly independent) modular forms of equal weight. One can check, from the Fourier expansions of the previous section that G_4^3 and G_6^2 are linearly independent modular forms of weight 12.

Writing $g_2 = 60G_4$ and $g_3 = 140G_6$, we set

$$j(z) = \frac{1728 g_2^3}{\Delta}, \quad \text{with} \quad \Delta = g_2^3 - 27 g_3^2,$$

where the factor 1728 has been chosen so that the Fourier coefficients of $j(z)$ are integral, as can be verified from Theorem 2.4.2. The first few q-series coefficients of $j(z)$ are:

$$j(z) = \frac{1}{q} + 744 + 196884q + 21493760q^2 + \cdots$$

Our basic philosophy is that arithmetic data about elliptic curves should correspond to information about special values of modular functions (and modular forms). For example, if E is an elliptic curve defined over a number field K, then E is isomorphic over $\mathbb{C}$ to an elliptic curve with lattice $\Lambda = \Lambda(\tau, 1)$ with $\tau \in \mathcal{H}$. One can conclude that $j(\tau) \in K$ – that is, a transcendental function has an algebraic special value!

If the lattice $\mathbb{Z}\tau + \mathbb{Z}$ is the ring of integers in a quadratic imaginary field K, then $j(\tau)$ generates the Hilbert class field of K, the maximal unramified abelian extension. For a wonderful exposition of these facts, see Cox's "Primes of the form $x^2 + ny^2$."

3 The geometry of the upper half plane

In this section, we explore the geometry of the upper half plane. For a fixed choice of discrete subgroup Γ, we study the quotient $\Gamma\backslash\mathcal{H}$, first as a topological space and then as a Riemann surface. This will allow us to formulate a definition for automorphic forms for arbitrary discrete groups Γ and then calculate dimensions of spaces of these automorphic forms. Much of this information is just a streamlined version of what

appears in Shimura's book "Introduction to the Arithmetic Theory of Automorphic Functions." His book contains an exhaustive account of the theory for discrete subgroups of $\mathrm{SL}(2,\mathbb{R})$.

3.1 The topological space $\Gamma\backslash\mathcal{H}$

Let Γ be a group acting on a topological space X. We seek conditions under which the quotient space $\Gamma\backslash X$ is Hausdorff.

Definition 3.1. *The action of Γ on X is said to be discontinuous if, for every $x \in X$, there exists a neighborhood U_x such that*

$$\{\gamma \in \Gamma \mid \gamma U_x \cap U_x \neq \emptyset\}$$

is finite.

Proposition 3.1.1. *Let G be a locally compact group acting on a topological space X such that for a point $x_0 \in X$, the stabilizer K of x_0 in G is compact and*

$$\begin{array}{rcl} \varphi : G/K & \longrightarrow & X \\ gK & \longmapsto & gx_0 \end{array}$$

is a homeomorphism. Then the following conditions on a subgroup Γ of G are equivalent:

(a) Γ acts discontinuously on X;
(b) For any compact subsets A and B of X, $\{\gamma \in \Gamma \mid \gamma(A) \cap B \neq \emptyset\}$ is finite;
(c) Γ is a discrete subgroup of G.

Proof. The equivalence of (a) and (b) is straightforward. We will show (b) is equivalent to (c). Given compact sets A and B of X, let $C = \pi^{-1}(A)$ be the lift of A to G (not G/K) and similarly $D = \pi^{-1}(B)$. Then $\gamma(A) \cap B \neq \emptyset$ implies $\gamma(C) \cap D \neq \emptyset$. That is, $\gamma \in \Gamma \cap (DC^{-1})$.

We claim that A compact implies $\pi^{-1}(A)$ compact, and hence C and D and thus DC^{-1} are compact. Assuming the claim, then if Γ discrete, $\Gamma \cap DC^{-1}$ is finite (since compact and discrete), giving (c) implies (b). To prove the claim, take an open cover of $G = \cup V_i$ whose closures $\overline{V}_i$ are compact. Then $A \subset \cup\pi(V_i)$ where the union runs over only finitely many i. Thus $\pi^{-1}(A) \subset \cup V_i K \subset \cup \overline{V}_i K$ (again taking the union over this finite set). Each $\overline{V}_i K$ is compact (as the image of $\overline{V}_i \times K$ under the multiplication map). Thus $\pi^{-1}(A)$ is a closed subset of a compact set, so compact.

Finally, to prove (b) implies (c), let V be a compact neighborhood of the identity e in G. Let $x = \pi(e)$. Then

$$\Gamma \cap V \subset \{g \in \Gamma \mid gx \in \pi(V)\}$$

For A and B as in the statement (b), we take $A = \{x\}$ and $B = \pi(V)$. Then by assumption, $\Gamma \cap V$ is a finite set, so Γ is discrete. □

Proposition 3.1.2. *Let Γ be a discrete subgroup of G, with all the hypotheses of the previous result. Then:*

(a) For any x in X, $\{\gamma \in \Gamma \mid \gamma x = x\}$ is finite.

(b) For any x in X, there is a neighborhood U_x of x such that if $\gamma \in \Gamma$ with $U_x \cap \gamma U_x \neq \emptyset$, then $\gamma x = x$.

(c) For any x and y in X not in the same Γ-orbit, there exist neighborhoods U_x and V_y of x and y such that $\gamma U_x \cap V_y = \emptyset$ for every $\gamma \in \Gamma$.

Proof. For part (a), the set in question is expressible as $\pi^{-1}(x) \cap \Gamma$ where again π is the map $g \mapsto gx$. By the previous proposition, inverse images of compact sets under π are again compact, so the intersection is compact and discrete, hence finite.

To prove (b), let V be a compact neighborhood of x. By Proposition 3.1.1(b), there is a finite set $\{\gamma_1, \ldots, \gamma_n\}$ in Γ such that $V \cap \gamma_i V \neq \emptyset$. Reindexing if necessary, let $\gamma_1, \ldots, \gamma_s$ be the subset of γ_i's which fix x. For each $i > s$, choose disjoint neighborhoods V_i of x and W_i of $\gamma_i x$ and let

$$U = V \cap \left(\bigcap_{i>s} V_i \cap \gamma^{-1} W_i \right).$$

Then U has the required property, since for $i > s$, $\gamma_i U \subset W_i$ but W_i is disjoint from V_i which contains U.

To prove (c), we again use Proposition 3.1.1(b). Choose compact neighborhoods A of x and B of y and let $\gamma_1, \ldots, \gamma_n$ be the finite set in Γ such that $\gamma_i A \cap B \neq \emptyset$. Since x and y are assumed to be inequivalent under Γ, we can find disjoint neighborhoods U_i of $\gamma_i x$ and V_i of y. Then setting

$$U = A \cap \gamma_1^{-1} U_1 \cap \cdots \cap \gamma_n^{-1} U_n, \quad V = B \cap V_1 \cap \cdots \cap V_n$$

gives the required pair of neighborhoods. □

Corollary 3.1.3. *With hypotheses as in the previous proposition, the space $\Gamma\backslash X$ is Hausdorff.*

Proof. Given any two points x and y not in the same Γ orbit, we may choose neighborhoods U and V as in Proposition 3.1.2(c). The images of U and V in $\Gamma\backslash X$ are then the required disjoint neighborhoods of Γx and Γy, respectively. □

3.2 Discrete subgroups of $\mathrm{SL}(2,\mathbb{R})$

Discrete subgroups of $\mathrm{SL}(2,\mathbb{R})$ are also known as *Fuchsian groups*. To see that a subgroup Γ of $\mathrm{SL}(2,\mathbb{R})$ is discrete, it suffices to check that the identity element is an isolated point in Γ. In particular, any subgroup of $\mathrm{SL}(2,\mathbb{Z})$ is discrete. We will be focusing on congruence subgroups of $\mathrm{SL}(2,\mathbb{Z})$.

Definition 3.2. *A congruence subgroup of* $\mathrm{SL}(2,\mathbb{Z})$ *is any subgroup containing* $\Gamma(N)$ *for some* N*, where*

$$\Gamma(N) = \left\{ \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z}) \;\middle|\; \begin{pmatrix} a & b \\ c & d \end{pmatrix} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \ (mod\ N) \right\}.$$

The smallest such N *is called the level of the congruence subgroup.*

The subgroups $\Gamma(N)$ are often referred to as "principal congruence subgroups" and $\Gamma(1)$ is common shorthand for $\mathrm{SL}(2,\mathbb{Z})$. The principal congruence subgroups fit into the exact sequence:

$$1 \to \Gamma(N) \to \mathrm{SL}(2,\mathbb{Z}) \to \mathrm{SL}(2,\mathbb{Z}/N\mathbb{Z}) \to 1$$

where the map from $\mathrm{SL}(2,\mathbb{Z}) \to \mathrm{SL}(2,\mathbb{Z}/N\mathbb{Z})$ is just the canonical projection. This shows that $\Gamma(N)$ is normal in $\mathrm{SL}(2,\mathbb{Z})$ and of finite index. (We leave the surjectivity of this projection as an exercise.)

Another important class of congruence subgroups are labeled $\Gamma_0(N)$ and defined as follows:

$$\Gamma_0(N) = \left\{ \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}(2,\mathbb{Z}) \;\middle|\; c \equiv 0 \pmod{N} \right\}.$$

Discrete subgroups of $\mathrm{SL}(2,\mathbb{R})$ may also be made from quaternion algebras. Recall that for any pair of rational numbers a, b such that $ab \neq 0$, we may define the quaternion algebra $B_{a,b} = B$ over $\mathbb{Q}$ as a $\mathbb{Q}$-algebra with basis $\{1, i, j, k\}$ satisfying the relations:

$$i^2 = a, \quad j^2 = b, \quad ij = k = -ji.$$

Given an element $\alpha = w + xi + yj + zk$, define its conjugate to be $\overline{\alpha} = w - xi - yj - zk$ and the norm map

$$\mathrm{Nm}(\alpha) = \alpha\overline{\alpha} = w^2 - ax^2 - by^2 + abz^2.$$

Then $B \otimes \mathbb{R}$ is an algebra over $\mathbb{R}$. There are exactly two quaternion algebras over $\mathbb{R}$, namely $B(1,1) = \mathrm{Mat}(2,\mathbb{R})$ and $B(-1,-1)$, which is the usual Hamiltonian

quaternion algebra and is a division algebra. The algebra B over $\mathbb{Q}$ is called *indefinite* if $B \otimes \mathbb{R} \simeq \mathrm{Mat}(2, \mathbb{R})$ and *definite* if isomorphic to the division algebra.

Suppose $B \otimes \mathbb{R} \simeq \mathrm{Mat}(2, \mathbb{R})$. Then the norm map corresponds to the determinant, and so we have an induced isomorphism between

$$\{\alpha \in B \otimes \mathbb{R} \mid \mathrm{Nm}(\alpha) = 1\} \xrightarrow{\simeq} \mathrm{SL}(2, \mathbb{R})$$

An order in B is a subring $\mathcal{O}$ that is finitely generated over $\mathbb{Z}$, and hence a free $\mathbb{Z}$ module of rank 4. If we set $\Gamma_{a,b}$ to be set of elements of $\mathcal{O}$ of norm 1, then $\Gamma_{a,b}$ is mapped to a discrete group of $\mathrm{SL}(2, \mathbb{R})$ under the above isomorphism.

Note that if a and b are chosen so that $B = \mathrm{Mat}(2, \mathbb{Q})$ and we take $\mathcal{O}$ to be $\mathrm{Mat}(2, \mathbb{Z})$, then we recover the classical theory with $\Gamma_{a,b} = \mathrm{SL}(2, \mathbb{Z})$. But if B is not isomorphic to $\mathrm{Mat}(2, \mathbb{Q})$ then the families of groups are in fact quite different from the theory of congruence subgroups. In particular the quotient space $\Gamma \backslash \mathcal{H}$ will be compact. This simplifies the spectral theory, as the quotient space no longer has continuous spectrum. For more details, see Chapter 5 of Miyake's book "Modular Forms"; in particular Theorem 5.2.13 of his book explains how indefinite quaternion algebras over $\mathbb{Q}$ which are division algebras over $\mathbb{Q}$ have orders with discrete subgroups that are co-compact.

3.3 Arithmetic subgroups of $\mathrm{SL}(2, \mathbb{Q})$

Roughly speaking, arithmetic subgroups are of interest to number theorists because they possess a large family of commuting self-adjoint operators, the so-called "Hecke operators." Before defining the notion of arithmetic subgroups for $\mathrm{SL}(2, \mathbb{Q})$, we require another definition.

Definition 3.3. *Two subgroups Γ and Γ' of a group G are said to be commensurable if $\Gamma \cap \Gamma'$ is of finite index in both Γ and Γ'.*

Proposition 3.3.1.

(a) Commensurability is an equivalence relation.

(b) Given two commensurable subgroups Γ, Γ' of a topological group G, then Γ is discrete if and only if Γ' is discrete.

(c) Given two commensurable closed subgroups Γ, Γ' of a locally compact group G, then $\Gamma \backslash G$ is compact if and only if $\Gamma' \backslash G$ is compact.

Proof. The proof is left as an exercise to the reader. □

Definition 3.4. *A subgroup of $\mathrm{SL}(2, \mathbb{Q})$ is said to be arithmetic if it is commensurable with $\mathrm{SL}(2, \mathbb{Z})$.*

For example, any congruence subgroup is arithmetic, as it has finite index in $\mathrm{SL}(2,\mathbb{Z})$. The congruence subgroups are sparse among all arithmetic subgroups. If $C(m)$ is the number of congruence subgroups of index $< m$ and $A(m)$ is the number of arithmetic subgroups of index $< m$, then $C(m)/A(m) \to 0$ as $m \to \infty$. See Remark 1.5 of Milne's lecture on "Shimura varieties and the work of Langlands" (`www.jmilne.org/math/xnotes/svq.pdf`) for a justification.

The matrix group $SL(2)$ is exceptional in having many non-arithmetic discrete groups. For other algebraic groups, Margulis proved essentially that any discrete subgroup Γ of $G(\mathbb{R})$ such that $\Gamma\backslash G(\mathbb{R})$ has finite volume is arithmetic. Moreover, there are many groups for which all arithmetic subgroups are congruence subgroups. (See Gopal Prasad's 1990 ICM lecture "Semi-simple groups and arithmetic subgroups" for more details.)

3.4 Linear fractional transformations

We begin by considering linear fractional transformations on $\mathbb{C} \cup \{\infty\}$. For any pair

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}(2,\mathbb{C}), \qquad z \in \mathbb{C} \cup \{\infty\},$$

we may consider the linear fractional transformation:

$$\gamma(z) = \frac{az+b}{cz+d} \quad \text{where} \quad \gamma(\infty) = \frac{a}{c}.$$

From the theory of the Jordan canonical form, each matrix γ (not scalar) is conjugate to one of the following forms:

$$\begin{pmatrix} a & 1 \\ 0 & a \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}, \quad \text{with } a \neq b.$$

Hence, each transformation is essentially one of:

$$z \longmapsto z + a^{-1} \quad \text{or} \quad z \longmapsto cz, \quad c \neq 1.$$

Matrices γ conjugate to the first of these cases are called "parabolic." These matrices act by translation, and their lone fixed point is ∞. Those in the second class are divided into three groups. If $|c| = 1$, they are called elliptic. If c is real and positive, they are called "hyperbolic." All other matrices are known as "loxodromic." These elements conjugate to the second class all have two fixed points. If we specialize to matrices with $\det(\gamma) = 1$, then the classification can be reinterpreted in terms of the trace $\mathrm{tr}(\gamma)$.

Proposition 3.4.1. *Given an element $\gamma \in \mathrm{SL}(2,\mathbb{C})$ not equal to $\pm I$, then*

$$\begin{aligned}
\gamma \textit{ is parabolic} &\iff \operatorname{tr}(\gamma) = \pm 2 \\
\gamma \textit{ is elliptic} &\iff \operatorname{tr}(\gamma) \textit{ is real and } |\operatorname{tr}(\gamma)| < 2 \\
\gamma \textit{ is hyperbolic} &\iff \operatorname{tr}(\gamma) \textit{ is real and } |\operatorname{tr}(\gamma)| > 2 \\
\gamma \textit{ is loxodromic} &\iff \operatorname{tr}(\gamma) \textit{ is not real.}
\end{aligned}$$

Proof. Since $\det(\gamma) = 1$, then the Jordan form of γ is either

$$\begin{pmatrix} \pm 1 & 1 \\ 0 & \pm 1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} a & 0 \\ 0 & a^{-1} \end{pmatrix}, \quad \text{with } a \neq \pm 1.$$

This gives the first three equivalences. For the last, suppose γ is conjugate to the diagonal matrix above so that $\operatorname{tr}(\gamma) = a + a^{-1}$. If $\operatorname{tr}(\gamma)$ is real, then either a is real (then γ hyperbolic) or a is imaginary with $a\overline{a} = 1$ (then γ elliptic). The reverse direction is clear. □

Finally, restricting our focus to transformations with real matrices, if $\gamma \in \mathrm{GL}(2,\mathbb{R})$, then set

$$j(\gamma, z) = cz + d \quad \text{for} \quad z \in \mathbb{C},\ \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Then one may check that

$$\det(\gamma)\, \Im(z) = |j(\gamma, z)|^2\, \Im(\gamma(z))$$

so that restricting γ to $\mathrm{GL}^+(2,\mathbb{R})$, invertible matrices with positive determinant, we send $\mathcal{H}$ to itself. Since scalar matrices induce the identity map, we may restrict our attention to $\mathrm{PSL}(2,\mathbb{R}) = \mathrm{SL}(2,\mathbb{R})/\{\pm 1\}$.

Recall that $\mathrm{SL}(2,\mathbb{R})$ acts transitively on $\mathcal{H}$. Indeed, if we consider the point $z = i$ (essentially the "origin" of the hyperbolic plane as it maps to the center of the unit disk under the Cayley transform) then for any $a, b \in \mathbb{R}$ with $a > 0$,

$$\begin{pmatrix} a^{1/2} & a^{-1/2}b \\ 0 & a^{-1/2} \end{pmatrix} (i) = ai + b$$

The isotropy group of $z = i$ is $\mathrm{SO}(2,\mathbb{R})$. Hence the isotropy group of any element $z \in \mathcal{H}$ is the set

$$\tau\, \mathrm{SO}(2,\mathbb{R})\tau^{-1}, \quad \text{where } \tau \text{ in } \mathrm{SL}(2,\mathbb{R}) \text{ maps } \tau(i) = z.$$

This shows that an element of $\mathrm{SL}(2,\mathbb{R})$ with at least one fixed point in $\mathcal{H}$ is either $\pm I$ or elliptic.

The group $\mathrm{SL}(2,\mathbb{R})$ also acts transitively on $\mathbb{R}\cup\{\infty\}$. Moreover, the isotropy subgroup of ∞ is

$$\left\{ \begin{pmatrix} a & b \\ 0 & a^{-1} \end{pmatrix} \middle| \, a \in \mathbb{R}^\times, b \in \mathbb{R} \right\}$$

and the subset of all parabolic elements in this isotropy subgroup are those with $a = \pm 1$ and $b \neq 0$. Thus, any element $\gamma \neq \pm I$ in $\mathrm{SL}(2,\mathbb{R})$ having at least one fixed point in $\mathbb{R}\cup\{\infty\}$ is either parabolic or hyperbolic. Summarizing, we have shown:

Proposition 3.4.2. *Let $\gamma \in \mathrm{SL}(2,\mathbb{R})$ such that $\gamma \neq \pm I$. Then*

$$\begin{aligned} \gamma \text{ is parabolic} &\iff \gamma \text{ has one fixed point on } \mathbb{R}\cup\{\infty\} \\ \gamma \text{ is elliptic} &\iff \gamma \text{ has one fixed point } z \text{ in } \mathcal{H} \text{ and the other fixed point is } \overline{z} \\ \gamma \text{ is hyperbolic} &\iff \gamma \text{ has two fixed points on } \mathbb{R}\cup\{\infty\}. \end{aligned}$$

Now fix a discrete group Γ. The points $x \in \mathbb{R}\cup\{\infty\}$ such that $\tau x = x$ for some parabolic element $\tau \in \Gamma$ will be called a *cusp* of Γ. The points $z \in \mathcal{H}$ such that $\tau(z) = z$ for an elliptic element γ of Γ will be called *elliptic points* of Γ. Both will play a distinguished role in defining the Riemann surface structure from the quotient $\Gamma\backslash\mathcal{H}$.

Proposition 3.4.3. *If z is an elliptic point of Γ then $\{\gamma \in \Gamma \mid \gamma(z) = z\}$ is a finite cyclic group.*

Proof. Recall that the set

$$\{\gamma \in \Gamma \mid \gamma(z) = z\} = \tau \,\mathrm{SO}(2,\mathbb{R})\tau^{-1} \cap \Gamma$$

where $\tau(i) = z$. Since Γ is discrete and $\mathrm{SO}(2)$ is compact, the intersection must be a finite group. Moreover, $\mathrm{SO}(2,\mathbb{R})$ is isomorphic to $\mathbb{R}/\mathbb{Z}$, whose finite subgroups are all cyclic (of form $n^{-1}\mathbb{Z}/\mathbb{Z}$ for some integer n). □

Corollary 3.4.4. *The elements of Γ of finite order are the elliptic elements of Γ and $\{\pm I\} \cap \Gamma$.*

Proof. If γ has finite order, then it is conjugate in $\mathrm{SL}(2,\mathbb{C})$ to a diagonal matrix with diagonal $(\zeta, \bar{\zeta})$ where ζ is a root of unity. By definition, such an element is elliptic or equal to $\pm I$. The other direction is clear from the previous proposition. □

3.5 Example: the structure of $SL(2,\mathbb{Z})$

Proposition 3.5.1. *The group $SL(2,\mathbb{Z})$ is generated by the two matrices*

$$S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \textit{and} \quad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

Proof. We take an arbitrary matrix in $\mathrm{SL}(2,\mathbb{Z})$ and show that it may be reduced to the identity by a sequence of multiplications by S and T. First observe that S is an inversion with $S^2 = -I$ and

$$S \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} -c & -d \\ a & b \end{pmatrix}.$$

The matrix T and its iterates $T^n = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}$ act by translation:

$$T^n \begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} a+cn & b+dn \\ c & d \end{pmatrix}.$$

Given any matrix γ with $c \neq 0$, this implies that there exists an n such that $T^n\gamma$ has upper left entry in $[0, |c|)$. Then applying S switches elements in the first column. So we may apply these operations repeatedly to successively reduce the element in the bottom left entry to 0. The resulting matrix must then be of the form

$$\begin{pmatrix} \pm 1 & m \\ 0 & \pm 1 \end{pmatrix}$$

so applying T^{-m} we arrive at $\pm I$. Finally applying S^2 gives I. □

We now determine the parabolic and elliptic elements of $\mathrm{SL}(2,\mathbb{Z})$.

Proposition 3.5.2. *The cusps of $\mathrm{SL}(2,\mathbb{Z})$ are the points of $\mathbb{Q} \cup \{\infty\}$, and they all lie in a single $\mathrm{SL}(2,\mathbb{Z})$-orbit.*

Proof. The matrix T fixes ∞. If $m/n \in \mathbb{Q}$, we may assume $\gcd(m,n) = 1$ so that there exist r, s such that $rm - sn = 1$. Let

$$\gamma = \begin{pmatrix} m & s \\ n & r \end{pmatrix}.$$

Then $\gamma(\infty) = m/n$, so m/n is fixed by the parabolic element $\gamma T \gamma^{-1}$ and in the same $\mathrm{SL}(2,\mathbb{Z})$-orbit as ∞. Conversely, every parabolic element of Γ is conjugate to $\pm T$ and may be written in the form $\alpha = \pm\gamma T \gamma^{-1}$ for some $\gamma \in \mathrm{GL}(2,\mathbb{Q})$. The point fixed by α is $\gamma(\infty) \in \mathbb{Q} \cup \{\infty\}$. □

Proposition 3.5.3. *The elliptic points of* $\mathrm{SL}(2,\mathbb{Z})$ *are all* $\mathrm{SL}(2,\mathbb{Z})$*-equivalent to either* $z = i$ *or* $z = e^{2\pi i/3} = (1 + i\sqrt{3})/2$.

Proof. If γ is an elliptic element of $\mathrm{SL}(2,\mathbb{Z})$, then by Proposition 3.4.1 the trace satisfies $|\operatorname{tr}(\gamma)| < 2$ and must be integral. Since the characteristic polynomial is quadratic, the only possibilities for such a γ are $x^2 + 1$ or $x^2 \pm x + 1$. In any case, γ has finite order and its eigenvalues are roots of a quadratic equation, so must be of order dividing 4 or 6 (but not of order 2 according to the above list). It is not hard to see that if $\gamma^4 = 1$, then γ is conjugate to $\pm \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, so the fixed points are all $\mathrm{SL}(2,\mathbb{Z})$ equivalent to i. Similarly, if $\gamma^3 = 1$, then γ is conjugate to either

$$\begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix},$$

and hence every elliptic point of order 3 is equivalent to $z = e^{2\pi i/3}$. See Section 1.4 of Shimura's book for explicit details. □

3.6 Fundamental domains

Let Γ continue to denote a discrete subgroup of $\mathrm{SL}(2,\mathbb{R})$. When considered as a subgroup of $\mathrm{PSL}(2,\mathbb{R}) = \mathrm{SL}(2,\mathbb{R})/\{\pm 1\}$, these are known as Fuchsian groups. (For an excellent treatment of the subject which includes proofs of theorems we will only state, see Svetlana Katok's book "Fuchsian groups." The first few chapters are available online at `www.math.psu.edu/katok_s/cmi.pdf`) Fuchsian groups can be visualized by means of their *fundamental domains*.

Definition 3.5. *A fundamental domain for* Γ *is a connected open subset* D *of* $\mathcal{H}$ *such that no two points of* D *are equivalent under* Γ *and* $\mathcal{H} = \bigcup_{\gamma\in\Gamma} \gamma(\overline{D})$ *where* $\overline{D}$ *denotes the closure of* D.

We define the distance $\rho(z, w)$ between any two points z, w in $\mathcal{H}$ to be the infimum of all lengths of curves between z and w using the hyperbolic metric. Here we replace the usual Euclidean metric with

$$ds = \frac{1}{y}\sqrt{dx^2 + dy^2}.$$

Geodesics in $\mathcal{H}$ are given by semicircles or half-lines orthogonal to $\mathbb{R}$ (which may be proved using the fact that linear fractional transformations are isometries. See Katok's Theorems I.2.5 and I.3.1 for details.)

One can show that the distance function ρ may be explicitly given by

$$\rho(z,w) = \ln \frac{|z-\overline{w}| + |z-w|}{|z-\overline{w}| - |z-w|}.$$

Various hyperbolic trig functions applied to this distance have simpler expressions.

To any point w in $\mathcal{H}$, we define the Dirichlet region for Γ centered at w:

$$D_w(\Gamma) = \{z \in \mathcal{H} \mid \rho(z,w) \leq \rho(z,\gamma(w)) \text{ for all } \gamma \in \Gamma\}$$

Proposition 3.6.1. *If w is not a fixed point of $\Gamma - \{I\}$, then D_w is a fundamental domain for Γ. Moreover, all fundamental domains have the same positive (but possibly infinite) volume*

$$\int_{D_w} d\mu$$

where μ is the Haar measure on $\mathcal{H}$.

Proof. For the first statement, see Theorem 2.4.2 of Katok. The latter is left as an exercise. □

Note that since Γ is discrete, then the orbits Γz for any point $z \in \mathcal{H}$ has no limit point in $\mathcal{H}$. However, it could have a limit point on the boundary $\mathbb{R} \cup \{\infty\}$. We say that a Fuchsian group Γ is "of the first kind" if every point of $\partial\mathcal{H} = \mathbb{R} \cup \{\infty\}$ is a limit point of Γ. We state a few facts about Fuchsian groups of the first kind and their fundamental domains before quickly specializing to the case of $\mathrm{SL}(2,\mathbb{Z})$.

Theorem 3.6.2. *Let Γ be a Fuchsian group of the first kind. Then we have the following:*

(1) Any Dirichlet region D_w which is a fundamental domain is a (hyperbolic) polygon with an even number of sides (where, if a side contains an elliptic point of order 2, we consider this as two sides).

(2) The sides of D_w can be arranged in pairs of equivalent sides. The elements $\gamma \in \Gamma$ which take one side to its pair generate Γ.

(3) Every fundamental domain has finite volume.

(4) We may choose the fundamental domain so that it is a polygon whose cuspidal vertices are inequivalent under Γ.

(5) Γ is co-compact in $\mathcal{H}$ if and only if it contains no parabolic elements.

Proof. See Katok, or else C.L. Siegel's paper "Discontinuous Groups," Annals of Math. (1943). □

Instead of offering the proof, we are content to see how each of these conditions holds for fundamental domains of $\mathrm{SL}(2,\mathbb{Z})$.

Proposition 3.6.3. *The set of points*

$$D = \{z \in \mathcal{H} \mid |z| > 1, -1/2 < \Re(z) < 1/2\}$$

is a fundamental domain for $\mathrm{SL}(2,\mathbb{Z})$.

Proof. We first show that any point z in $\mathcal{H}$ is equivalent under $\mathrm{SL}(2,\mathbb{Z})$ to a point in $\overline{D}$, the closure of D. Recall that for any

$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} : \quad \Im(\gamma(z)) = \frac{\Im(z)}{|cz+d|^2}.$$

Since the set $\{cz + d \mid c, d \in \mathbb{Z}\}$ is a lattice, then $\min_{(c,d)\neq(0,0)} |cz+d|$ is attained for some pair (c,d). Equivalently, there exists a γ such that $\Im(\gamma(z))$ is maximized. For any such γ, let $\gamma(z) = w$. Applying the inversion matrix $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ to w:

$$\Im(S(w)) = \Im(-1/w) = \Im(w)/|w|^2 \leq \Im(w),$$

so $|w| \geq 1$. Now translating w by an appropriate power of $T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ preserves the imaginary part of w and may be moved to any vertical strip of $\mathcal{H}$ of width 1.

It remains to show that no two points of D are equivalent. Let z and z' be distinct points of D. Suppose that $z' = \gamma z$ with $\Im(z) \leq \Im(z') = \frac{\Im(z)}{|cz+d|^2}$. This implies

$$|c| \cdot \Im(z) \leq |cz + d| \leq 1.$$

This forces $|c| \leq 1$ according to D (since $\Im(z) > \sqrt{3}/2$) and we examine each of these cases. If $c = 0$, then $a, d = \pm 1$, which means γ acts by translation, which is a contradiction. If $|c| = 1$, then $|z \pm d| \leq 1$. Either $|d| \geq 1$ and $z \in D$, which gives $|z + d| > 1$, or else $d = 0$ and then the inequality would force $|z| \leq 1$, contradicting that $z \in D$. □

Note that the shape of this fundamental domain (or any of its translates under Γ) satisfy the conditions of Theorem 3.6.2:

1. D has 4 sides, where the bottom arc formed by the unit circle is divided into two sides to the right and left of i.

2. The transformation T pairs the sides $z = -1/2$ and $z = 1/2$, while S takes $z = e^{i\theta}$ to $z = -e^{-i\theta}$, so pairs the two edges meeting at $z = i$. As we saw before, S and T generate $\mathrm{SL}(2, \mathbb{Z})$.

3. The fundamental domain D is a hyperbolic triangle, so by a special case of the Gauss-Bonnet theorem, has area $\pi - \alpha - \beta - \gamma$ where α, β, γ are the interior angles of the triangle. In our case, these angles are $\pi/3$ at the angles adjacent to the unit circle, and 0 at $i\infty$, so the area of D (or any of its translates under $\mathrm{SL}(2, \mathbb{Z})$) is $\pi/3$.

4. The only cuspidal point in $\overline{D}$ is $\{\infty\}$ (which is indeed a vertex of D), so the condition that D contain only non-equivalent cusps is trivially satisfied.

3.7 $\Gamma\backslash\mathcal{H}^*$ as a topological space

We now show how to compactify $\Gamma\backslash\mathcal{H}$, in order to apply the theory of compact Riemann surfaces. As we will see, if Γ has cusps, then the resulting quotient is not compact. But if Γ is a Fuchsian group of the first kind, we may consider

$$\mathcal{H}^* = \mathcal{H}^*_\Gamma = \mathcal{H} \cup P_\Gamma, \quad \text{where } P_\Gamma \text{ is the set of cusps of } \Gamma.$$

We first explain how to define the topology on $\mathcal{H}^*$:

- For $z \in \mathcal{H}$, the fundamental system of open neighborhoods for $z \in \mathcal{H}^*$ is just that for $z \in \mathcal{H}$.

- For $x \in P_\Gamma$, the fundamental system of open neighborhoods at x is the family

$$\{\sigma^{-1}U_\ell \mid \ell > 0\} \quad \text{where} \quad U_\ell = \{z \in \mathcal{H} \mid \Im(z) > \ell\}, \quad \text{and} \quad \sigma(x) = \infty$$

Note that if our cusp is of the form $x = -d/c$, then the family of neighborhoods

$$\sigma^{-1}U_\ell = \{z \in \mathcal{H} \mid \Im(z)/|cz + d|^2 > \ell\},$$

which is a circle of radius $(2\ell c^2)^{-1}$ tangent to the real axis at x.

Theorem 3.7.1. *Let Γ be a Fuchsian group. The quotient space $\Gamma\backslash\mathcal{H}^*$ is Hausdorff.*

Proof. First note that since we had an action of $\mathrm{SL}(2, \mathbb{Z})$ on $\mathbb{C} \cup \{\infty\}$, the quotient space is defined and we may regard $\Gamma\backslash\mathcal{H}$ as a subspace. Our earlier proof that $\Gamma\backslash\mathcal{H}$ is Hausdorff followed from Proposition 3.1.2(c) which stated:

For any x and y in X not in the same Γ-orbit, there exist neighborhoods U_x and V_y of x and y such that $\gamma U_x \cap V_y = \emptyset$ for every $\gamma \in \Gamma$.

If x and y are in $\mathcal{H}$, then we can use this fact without change. If instead, at least one of x or y is in P_Γ, we must show the same is true in the topological space $X = \mathcal{H}^*$. We label these cases:

Case 1: $x \in P_\Gamma$ while $y \in \mathcal{H}$.

Case 2: Both x and y are in P_Γ.

Both of these cases rely on the following lemma:

Lemma 3.7.2. *Assume that ∞ is a cusp of Γ. Then the stabilizer of ∞ is of the form*

$$\Gamma_\infty = \left\{ \pm \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}^m \;\middle|\; m \in \mathbb{Z} \right\}, \quad \textit{(for some } h > 0.\textit{)}$$

Let $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma$. If $|ch| < 1$ then $c = 0$.

Proof. We define a sequence of matrices $\gamma_n \in \Gamma$ by

$$\gamma_0 = \gamma, \quad \text{and} \quad \gamma_{n+1} = \gamma_n \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix} \gamma_n^{-1}.$$

Denote the entries of $\gamma_n = \begin{pmatrix} a_n & b_n \\ c_n & d_n \end{pmatrix}$. Then by explicit computation one can check that for $|ch| < 1$, the sequence converges to $\begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}$. Since Γ is discrete, there exists an n such that $\gamma_n = \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}$. Noting that $c_n = -c(ch)^{2^n - 1}$, this implies $c = 0$. □

Proof in Case 1: Now we show how this lemma implies the required result when x is a cusp and $y \in \mathcal{H}$. Suppose x is a cusp of Γ and $\sigma \in \mathrm{SL}(2, \mathbb{R})$ satisfies $\sigma(x) = \infty$. It suffices to show that for any compact subset A of $\mathcal{H}$, there exists a positive number ℓ such that

$$A \cap \gamma \sigma^{-1} U_\ell = \emptyset$$

for any $\gamma \in \Gamma$.

First, replacing Γ by $\sigma^{-1}\Gamma\sigma$ and A by $\sigma^{-1}A$, we may assume that $x = \infty$, $\sigma = I$. Now if $\gamma \in \Gamma_\infty$, the stabilizer of ∞, then $\gamma U_\ell = U_\ell$. If $\gamma \notin \Gamma_\infty$, then by Lemma 3.7.2, $|c| \geq 1/|h|$ so that

$$\gamma U_\ell \subset \{ z \in \mathcal{H} \mid \Im(z) < h^2/\ell \}.$$

Thus we may choose ℓ so that our compact set $A \subset \{z \in \mathcal{H} \mid h^2/\ell < \Im(z) < \ell\}$.

Proof in Case 2: Given x and y, cusps of Γ, and $\sigma, \tau \in \mathrm{SL}(2,\mathbb{R})$ such that $\sigma(x) = \tau(y) = \infty$. Then we have

$$\sigma\Gamma_x\sigma^{-1} = \left\{ \begin{pmatrix} 1 & h_1 \\ 0 & 1 \end{pmatrix}^m \;\middle|\; m \in \mathbb{Z} \right\}, \quad \tau\Gamma_y\tau^{-1} = \left\{ \begin{pmatrix} 1 & h_2 \\ 0 & 1 \end{pmatrix}^m \;\middle|\; m \in \mathbb{Z} \right\}$$

for some $h_1, h_2 > 0$. Then we claim that provided ℓ_1, ℓ_2 are chosen so that $\ell_1\ell_2 > |h_1h_2|$, then

$$\gamma\sigma^{-1}U_{\ell_1} \cap \tau^{-1}U_{\ell_2} = \emptyset \quad \text{for all } \gamma \in \Gamma \text{ such that } \gamma x \neq y.$$

Again, by replacing Γ by $\sigma^{-1}\Gamma\sigma$ and τ by $\tau\sigma$, we may assume that $x = \infty$ and $\sigma = I$. Assume that $\gamma\sigma^{-1}U_{\ell_1} \cap \tau^{-1}U_{\ell_2} \neq \emptyset$. Let $\delta = \tau\gamma$ and write $\delta^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Then for any $z \in \delta U_{\ell_1} \cap U_{\ell_2}$, we have

$$\ell_1\ell_2 < \Im(\delta^{-1}z)\Im(z) = \Im(z)^2/|cz+d|^2 \leq c^{-2}.$$

Now using a similar technique to Lemma 3.7.2, consider

$$\delta \begin{pmatrix} 1 & h_1 \\ 0 & 1 \end{pmatrix} \delta^{-1} := \begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix} \quad \text{where } c_1 = -c^2h_1.$$

Applying Lemma 3.7.2 to $\tau\Gamma\tau^{-1}$ with element $\delta \begin{pmatrix} 1 & h_1 \\ 0 & 1 \end{pmatrix} \delta^{-1}$, we see that since

$$|c_1h_2| = |c^2h_1h_2| < c^2\ell_1\ell_2 \leq 1$$

we conclude $c_1 = 0$ and hence $c = 0$. This implies $\gamma\infty = \tau^{-1}\delta\infty = \tau^{-1}\infty = x_2$. □

We call a point in the quotient space $\Gamma\backslash\mathcal{H}^*$ an *elliptic point* (resp. a *cusp*) if its preimage in $\mathcal{H}^*$ under the canonical projection is an elliptic point (resp. a cusp).

Corollary 3.7.3. *If $\Gamma\backslash\mathcal{H}^*$ is compact, then the number of elliptic points and cusps of $\Gamma\backslash\mathcal{H}^*$ is finite.*

Proof. Given any point $z \in \mathcal{H}^*$, we showed in the proof of Theorem 3.7.1 that there exists a neighborhood U of z such that $\gamma U \cap U \neq \emptyset$ if and only if $\gamma z = z$. In the latter case, U may be chosen so that z is the unique elliptic point or cusp. Let π denote the canonical projection from $\mathcal{H}^*$ to $\Gamma\backslash\mathcal{H}^*$. Then $\pi(U) - \{\pi(z)\}$ contains neither elliptic points nor cusps. Since $\Gamma\backslash\mathcal{H}^*$ is assumed compact and the sets $\pi(U)$ are open, it takes only finitely many of them to cover the quotient. The result follows. □

Theorem 3.7.4 (Siegel)**.** *Let Γ be a Fuchsian group. Then $\Gamma\backslash\mathcal{H}^*$ is compact if and only if $\mu(\Gamma\backslash\mathcal{H}^*)$ is finite.*

Proof. The only-if direction is clear. For the proof of the converse, see Theorem 1.9.1 of Miyake's "Modular Forms." □

This theorem is meant to justify our emphasis on Fuchsian groups *of the first kind*. In fact, some authors (including Shimura and Miyake) define these groups to be those for which $\Gamma\backslash\mathcal{H}^*$ is compact. But we can arrive at this using Theorem 3.6.2, part (3), together with Siegel's result.

3.8 $\Gamma\backslash\mathcal{H}^*$ as a Riemann surface

Recall that a Riemann surface is a one-dimensional, connected complex analytic manifold. That is, a Riemann surface is a connected Hausdorff space X with an atlas:

- Every point $x \in X$ has a neighborhood U_x and a homeomorphism ϕ_x of U_x onto an open subset of $\mathbb{C}$.
- If $U_x \cap U_y \neq \emptyset$, the map $\phi_x \circ \phi_y^{-1} : \phi_y(U_x \cap U_y) \to \phi_x(U_x \cap U_y)$ is holomorphic.

To define the complex structure on $\Gamma\backslash\mathcal{H}^*$, recall that to any point $z \in \mathcal{H}^*$, there is an open neighborhood U such that

$$\Gamma_z = \{\gamma \in \Gamma \mid \gamma(U) \cap U \neq \emptyset\} \quad \text{where} \quad \Gamma_z : \text{stabilizer of } z.$$

Then there's a natural injection of $\Gamma_z\backslash U \to \Gamma\backslash\mathcal{H}^*$, with $\Gamma_z\backslash U$ an open neighborhood of $\pi(z)$, the image of z under the canonical projection to $\Gamma\backslash\mathcal{H}^*$. If z is neither an elliptic point nor a cusp, then $\Gamma_z \subset \{\pm I\}$ so that the map $\pi : U \to \Gamma_z\backslash U$ is a homeomorphism. Then we may take the pair $(\Gamma_z\backslash U, \pi^{-1})$ as part of the complex structure.

If instead z is an elliptic point of $\mathcal{H}^*$, then let $\bar{\Gamma}_z = \Gamma_z/(\Gamma \cap \{\pm 1\})$. Let λ be a holomorphic isomorphism of $\mathcal{H}$ onto the unit disc D with $\lambda(z) = 0$. Recall that $\bar{\Gamma}_z$ is cyclic, say of order n. By Schwarz' lemma, $\lambda\bar{\Gamma}_z\lambda^{-1}$ consists of the transformations

$$D \to D : w \mapsto \zeta_n^k w, \quad k \in [0, n-1], \quad \zeta_n = e^{2\pi i/n}.$$

Then we can define the chart $\phi : \Gamma_z\backslash U \to \mathbb{C}$ by $\phi(\pi(z)) = \lambda(z)^n$. The resulting ϕ is a homeomorphism onto an open subset of $\mathbb{C}$.

Example: Elliptic points of $\mathrm{SL}(2,\mathbb{Z})$

Every elliptic point is $\mathrm{SL}(2,\mathbb{Z})$-equivalent to either i or ρ, a cube root of unity. So it suffices to provide charts in these two cases. Suppose $z = i$. Then the Cayley transform $z \mapsto \frac{z-i}{z+i}$ maps $\mathcal{H}$ to D, the open disk, and maps i to 0. The stabilizer (mod $\pm I$) of i is the two element set $\{I, S\}$. The action of S on $\mathcal{H}$ is transformed to the automorphism $z \mapsto -z$ of D. The function

$$z \mapsto \left(\frac{z-i}{z+i}\right)^2$$

is thus a holomorphic function defined in a neighborhood of i and invariant under S. This is our coordinate chart for $\pi(i)$. The chart for ρ works similarly, with ρ replacing i in the Cayley transform, and then cubing the resulting map as the stabilizer has order 3 generated by ST.

Finally, we must explain how to handle the cusps. As we argued earlier, if x is a cusp of Γ then choosing $\sigma \in \mathrm{SL}(2,\mathbb{R})$ so that $\sigma(x) = \infty$,

$$\sigma\Gamma_x\sigma^{-1} = \left\{ \pm \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}^m \;\middle|\; m \in \mathbb{Z} \right\}, \quad \text{for some } h > 0.$$

Then we define a homeomorphism ϕ from $\Gamma_x\backslash U$ to an open subset of $\mathbb{C}$ by $\phi(\pi(z)) = e(\pi(z)/h)$ where $e(x) = e^{2\pi i x}$.

With this complex structure, one may check that the composition of charts satisfy the holomorphicity condition on intersections. Hence we have shown that for any Fuchsian group Γ, the quotient space $\Gamma\backslash\mathcal{H}^*$ can be given the structure of a Riemann surface. Moreover, if Γ is *of the first kind*, the quotient is compact.

3.9 A few basics about compact Riemann surfaces

Let us recall a few facts, without proof, about compact Riemann surfaces. For more information on these basic results, see the following books:

- R. Gunning, "Lectures on Riemann Surfaces" (1966)
- R. Miranda, "Algebraic Curves and Riemann Surfaces" (1995)
- P. Griffiths, "Introduction to Algebraic Curves" (1989)

Let $\mathcal{R}$ be a compact Riemann surface. It's a topological space, so we may define homology groups $H_i(\mathcal{R},\mathbb{Z})$. Then

$$H_0(\mathcal{R},\mathbb{Z}) \simeq \mathbb{Z}; \quad H_1(\mathcal{R},\mathbb{Z}) \simeq \mathbb{Z}^{2g}; \quad H_2(\mathcal{R},\mathbb{Z}) \simeq \mathbb{Z}; \quad H_i(\mathcal{R},\mathbb{Z}) \simeq 0 \quad \text{for all } i > 2,$$

where g is a non-negative integer called the genus of $\mathcal{R}$. It follows from the Riemann-Hurwitz theorem (to be discussed shortly) that this is the same genus as that appearing in the Riemann-Roch theorem.

The Euler-Poincaré characteristic χ of $\mathcal{R}$ is defined by

$$\chi(\mathcal{R}) := \sum_i (-1)^i \dim(H_i(\mathcal{R}, \mathbb{Z})) = 2 - 2g$$

Taking a triangulation of $\mathcal{R}$, then if V, E, F, denote number of vertices, edges, and faces, respectively,

$$2 - 2g = V - E + F.$$

For example, we may triangulate $\mathrm{SL}(2,\mathbb{Z})\backslash\mathcal{H}^*$ using the fundamental domain D presented in Proposition 3.6.3. Recall this has 4 sides, counting the arcs on both sides of $z = i$ as separate sides. But these are identified in pairs in the quotient. There are three vertices i, ∞, and $e^{2\pi i/6} = e^{2\pi i/3}$. If we add any vertex in the interior of D and connect it to all four vertices of the boundary of D, we obtain a triangulation with a total of 4 vertices, 6 edges, and 4 faces. This is just one way to confirm that the genus of $\mathrm{SL}(2,\mathbb{Z})\backslash\mathcal{H}^*$ is 0.

Finally, we require the Riemann-Hurwitz formula, which will allow us to compute the genus of the Riemann surfaces corresponding to families of congruence subgroups.

Theorem 3.9.1 (Riemann-Hurwitz formula)**.** *Let $f : \mathcal{R}' \to \mathcal{R}$ be a holomorphic mapping of compact Riemann surfaces that is m-to-1 except at finitely many points. For each point p in $\mathcal{R}'$, let e_p denote the ramification index of p. Then*

$$2g(\mathcal{R}') - 2 = m(2g(\mathcal{R}) - 2) + \sum_{p \in \mathcal{R}'} (e_p - 1).$$

The integer m is the degree of the covering map f. In the classical language of coordinate charts, the ramification index at $z_0 \in \mathcal{R}'$ above $w_0 = f(z_0) \in \mathcal{R}$ may be understood as the integer e appearing in the expansion

$$\phi'_{w_0}(f(z)) = a_e \phi_{z_0}(z)^e + a_{e+1}\phi_{z_0}(z)^{e+1} + \cdots, \quad a_e \neq 0,$$

in a neighborhood of z_0, where ϕ' is a chart for z_0 and ϕ is a chart for w_0. The integer e can be shown to be independent of chart. One can prove this theorem using a triangulation of $\mathcal{R}$ in which the ramification points are included among the vertices, then lift this triangulation to $\mathcal{R}'$.

We now study the ramification indices, in the general situation where Γ' has finite index in Γ, a Fuchsian group of the first kind. Again to any $z \in \mathcal{H}^*$, let

$$\overline{\Gamma}_z = \{\gamma \in \Gamma \mid \gamma(z) = z\}, \quad \overline{\Gamma}'_z = \overline{\Gamma}_z \cap \overline{\Gamma}'.$$

Here the notation $\overline{\Gamma}$ means we are considering the group Γ as a subgroup of $\mathrm{PSL}(2,\mathbb{R})$.

Let f be a covering map of degree n for these Riemann surfaces with $f^{-1}(p) = \{q_1, \ldots, q_h\}$ where $\pi(z) = p$ under the canonical projection from $\mathcal{H}^*$:

$$\begin{array}{rcl} f : \Gamma'\backslash\mathcal{H}^* & \longrightarrow & \Gamma\backslash\mathcal{H}^* \\ \{q_1, \ldots, q_h\} & \longmapsto & p \end{array}$$

Finally choose points w_k such that $\pi'(w_k) = q_k$ where $\pi' : \mathcal{H}^* \to \Gamma'\backslash\mathcal{H}^*$.

Proposition 3.9.2. *The ramification index e_k of f at q_k is $[\overline{\Gamma}_{w_k} : \overline{\Gamma}'_{w_k}]$. If $w_k = \sigma_k(z)$ for some $\sigma_k \in \overline{\Gamma}$, then $e_k = [\overline{\Gamma}_z : \sigma_k^{-1}\overline{\Gamma}'\sigma_k \cap \overline{\Gamma}_z]$. Moreover, if $\overline{\Gamma}'$ is normal in $\overline{\Gamma}$ then $e_1 = \cdots = e_h$ and $[\overline{\Gamma} : \overline{\Gamma}'] = e_1 h$.*

Proof. The first assertion is clear from the form of our complex structure, and the definition of ramification index. The second claim follows since $\overline{\Gamma}_{w_k} = \sigma_k \overline{\Gamma}_z \sigma_k^{-1}$. If $\overline{\Gamma}'$ is normal, then this last index is independent of σ_k, so the final claim follows. $\square$

3.10 The genus of $X(\Gamma)$

Given a Fuchsian group of the first kind Γ, we often denote the resulting Riemann surface $\Gamma\backslash\mathcal{H}$ by $Y(\Gamma)$ and the *compact* Riemann surface $\Gamma\backslash\mathcal{H}^*$ by $X(\Gamma)$. In fact, because we commonly deal with the congruence subgroups $\Gamma(N)$ and $\Gamma_0(N)$ as defined in Section 3.2, we often further shorten the notation by writing $X(\Gamma(N)) = X(N)$ and $X(\Gamma_0(N)) = X_0(N)$, etc.

In this section, we determine the genus of $X(N)$ and $X_0(N)$. Note that since both these families consist of finite index subgroups Γ of $\mathrm{SL}(2,\mathbb{Z})$, the natural map

$$f : \Gamma\backslash\mathcal{H}^* \longrightarrow \mathrm{SL}(2,\mathbb{Z})\backslash\mathcal{H}^*$$

is a holomorphic map of Riemann surfaces and the degree of the covering map f is precisely $[\mathrm{PSL}(2,\mathbb{Z}) : \overline{\Gamma}]$ where $\overline{\Gamma}$ denotes the image of Γ in $\mathrm{PSL}(2,\mathbb{Z}) = \mathrm{SL}(2,\mathbb{Z})/\{\pm 1\}$. This index, together with the ramification indices at elliptic points and cusps, will give us a formula for the genus of $X(\Gamma)$ according to the Riemann-Hurwitz formula.

Proposition 3.10.1. *For any positive integer N, the index*

$$[\mathrm{SL}(2,\mathbb{Z}) : \Gamma(N)] = N^3 \prod_{p|N} (1 - p^{-2}).$$

Proof. As discussed in Section 3.2, there is a natural exact sequence of groups

$$1 \to \Gamma(N) \to \mathrm{SL}(2,\mathbb{Z}) \to \mathrm{SL}(2,\mathbb{Z}/N\mathbb{Z}) \to 1$$

so the index $[\mathrm{SL}(2,\mathbb{Z}) : \Gamma(N)] = |\mathrm{SL}(2,\mathbb{Z}/N\mathbb{Z})|$. To give an exact formula for the order of this group, we write $N = \prod_{p_i} p_i^{r_i}$. Then we have isomorphisms

$$\mathbb{Z}/N\mathbb{Z} \simeq \prod_{p_i} (\mathbb{Z}/p_i^{r_i}\mathbb{Z}), \quad \text{and} \quad \mathrm{SL}(2,\mathbb{Z}/N\mathbb{Z}) \simeq \prod_{p_i} \mathrm{SL}(2,\mathbb{Z}/p_i^{r_i}\mathbb{Z}),$$

Furthermore, we may determine $|\mathrm{SL}(2,\mathbb{Z}/p^r\mathbb{Z})|$ from the order of $|\mathrm{GL}(2,\mathbb{Z}/p^r\mathbb{Z})|$ by considering it as the kernel of the determinant map, which gives:

$$|\mathrm{GL}(2,\mathbb{Z}/p^r\mathbb{Z})| = \varphi(p^r)|\mathrm{SL}(2,\mathbb{Z}/p^r\mathbb{Z})| \quad \varphi : \text{Euler phi function.} \tag{4}$$

Thus it remains to determine the size of $|\mathrm{GL}(2,\mathbb{Z}/p^r\mathbb{Z})|$. Again, we use an exact sequence to reduce to the case $r = 1$:

$$1 \to \ker(\phi) \to \mathrm{GL}(2,\mathbb{Z}/p^r\mathbb{Z}) \xrightarrow{\phi} \mathrm{GL}(2,\mathbb{Z}/p\mathbb{Z}) \to 1,$$

where $\ker(\phi)$ consists of matrices

$$\{\gamma \in \mathrm{Mat}(2,\mathbb{Z}/p^r\mathbb{Z}) \mid \gamma \equiv I \pmod{p}\} = \left\{ I + p\begin{pmatrix} a & b \\ c & d \end{pmatrix} \;\middle|\; a,b,c,d \in \mathbb{Z}/p^{r-1}\mathbb{Z} \right\},$$

so $|\ker(\phi)| = p^{4(r-1)}$. Finally, $|\mathrm{GL}(2,\mathbb{Z}/p\mathbb{Z})| = (p^2-1)(p^2-p)$, since we may freely choose the top row $(a,b) \neq (0,0)$ and then choose the bottom row to be linearly independent from the top. Thus we conclude that

$$|\mathrm{GL}(2,\mathbb{Z}/p^r\mathbb{Z})| = (p^2-1)(p^2-p)p^{4(r-1)}$$

and using (4) we have

$$|\mathrm{SL}(2,\mathbb{Z}/p^r\mathbb{Z})| = p^{3r}(1-p^{-2})$$

and the result follows by applying the isomorphism for $\mathrm{SL}(2,\mathbb{Z}/N\mathbb{Z})$. □

Corollary 3.10.2. *Let* $\overline{\Gamma(N)} := \Gamma(N)/\{\pm I \cap \Gamma(N)\}$. *The index*

$$\mu_N = [\mathrm{PSL}(2,\mathbb{Z}) : \overline{\Gamma(N)}] = \begin{cases} \frac{N^3}{2}\prod_{p|N}(1-p^{-2}) & \text{if } N > 2 \\ 6 & \text{if } N = 2. \end{cases}$$

Proof. The case $N = 2$ is the only N for which $-I \in \Gamma(N)$. □

We now study the ramification indices using Proposition 3.9.2. Indeed we need only consider elliptic points and cusps, as these are the only points in $\mathcal{H}^*$ with non-trivial stabilizer. First, a simple fact about elliptic points of principle congruence subgroups.

Proposition 3.10.3. *If $N > 1$, then $\Gamma(N)$ has no elliptic points.*

Proof. This follows immediately from the fact that any elliptic element in $\mathrm{SL}(2,\mathbb{Z})$ is conjugate to one of:

$$\pm\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \pm\begin{pmatrix} 0 & -1 \\ 1 & -1 \end{pmatrix}, \quad \pm\begin{pmatrix} -1 & 1 \\ -1 & 0 \end{pmatrix},$$

which fix either i or $\rho = e^{2\pi i/3}$, together with the fact that $\Gamma(N)$ is a normal subgroup. Indeed, none of these is congruent to I modulo N if $N > 1$, and normality implies their conjugates can't be either. □

Using Proposition 3.9.2, we see that for each elliptic point of $\Gamma(1) := \mathrm{SL}(2,\mathbb{Z})$, the ramification index is $[\overline{\Gamma}(1)_z : \overline{\Gamma}(N)_z]$. Since $\overline{\Gamma}(N)_z = \{I\}$ for $N > 1$, the ramification index is just 2 or 3, depending on the order of the elliptic element.

For the cusps, each is $\Gamma(1)$-equivalent to ∞. Moreover

$$\overline{\Gamma(1)}_\infty = \left\{ \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^m \;\middle|\; m \in \mathbb{Z} \right\}, \quad \overline{\Gamma(N)}_\infty = \left\{ \begin{pmatrix} 1 & N \\ 0 & 1 \end{pmatrix}^m \;\middle|\; m \in \mathbb{Z} \right\}$$

so that $[\overline{\Gamma}(1)_\infty : \overline{\Gamma}(N)_\infty] = N$. In particular, this says that $\Gamma(N)$ has μ_N/N inequivalent cusps.

Proposition 3.10.4. *Let $N > 1$. The genus g_N of $\Gamma(N)\backslash\mathcal{H}^*$ is given by*

$$g_N = 1 + \mu_N \cdot \frac{(N-6)}{12N}$$

Proof. We apply the Riemann-Hurwitz formula (Theorem 3.9.1) to the covering $f : \Gamma(N)\backslash\mathcal{H}^* \to \Gamma(1)\backslash\mathcal{H}^*$. Since the genus of $X(1)$ is 0, this gives:

$$g_N = 1 - \mu_N + \frac{1}{2}\sum_p (e_p - 1)$$

As remarked above, there are μ_N/N inequivalent cusps of $\Gamma(N)$. Each of these has $e_p = N$. Over i, there are $\mu_N/2$ points of index 2 and over $e^{2\pi i/3}$ there are $\mu_N/3$ points of index 3. Putting this all together, we have

$$g_N = 1 - \mu_N + \frac{1}{2}\left[\left(\mu_N - \frac{\mu_N}{N}\right) + \frac{\mu_N}{2} + \frac{2\mu_N}{3}\right],$$

and simplifying gives the result. □

One can use similar methods for any finite index subgroup of $\Gamma(1)$.

Proposition 3.10.5. *Given any finite index subgroup Γ whose projection $\overline{\Gamma}$ in $\mathrm{PSL}(2,\mathbb{Z})$ has index μ, then the genus g of the Riemann surface $\Gamma\backslash\mathcal{H}^*$ is*

$$g = 1 + \frac{\mu}{12} - \frac{\nu_2}{4} - \frac{\nu_3}{3} - \frac{\nu_\infty}{2}$$

where ν_2, ν_3 are the numbers of Γ-inequivalent elliptic points of order 2 and 3, respectively, and ν_∞ is the number of Γ-inequivalent cusps.

Proof. This is again an exercise in using the Riemann-Hurwitz formula together with some careful counting of ramification indices. We leave it as an exercise to the reader. □

The Riemann surfaces $X_0(N)$ play an extremely important role in number theory. Indeed, one version of the Shimura-Taniyama-Weil conjecture (now a theorem) is that there exists a surjective map of algebraic curves $X_0(N) \to E$, where N is the conductor of the elliptic curve E. In particular, the number N is divisible only by primes p at which E has bad reduction (i.e. reduces mod p to a singular curve).

4 Automorphic Forms for Fuchsian Groups

4.1 A general definition of classical automorphic forms

Let Γ be a Fuchsian group of the first kind. Recall our earlier notation

$$j(\gamma, z) = (cz + d), \quad \gamma \in \Gamma, z \in \mathcal{H}$$

which arose naturally in studying the action of Γ on $\mathcal{H}$. Following Shimura and many other authors, given a function $f : \mathcal{H} \to \mathbb{C}$ we define an action of Γ on the space of functions by

$$f|[\gamma]_k \stackrel{def}{=} f(\gamma(z))j(\gamma, z)^{-k}.$$

We sometimes simply write $f|[\gamma]$ when $k = 0$ (i.e., the usual action by linear fractional transformation). The verification that this is an action is straightforward and left to the reader. The action is sometimes referred to as the "slash operator" (of weight k for Γ).

Definition 4.1 (Classical automorphic forms)**.** *A function $f : \mathcal{H} \to \mathbb{C}$ is called a (classical) automorphic form of weight k with respect to Γ if f satisfies the following conditions:*

- *f is meromorphic on $\mathcal{H}$,*
- *$f|[\gamma]_k = f$ for all $\gamma \in \Gamma$,*
- *f is meromorphic at every cusp s of Γ.*

To explain this last condition in more detail, we handle each cusp s as before by translating the action to ∞. If $\sigma(s) = \infty$ for some $\sigma \in \mathrm{SL}(2, \mathbb{R})$. As remarked in the last section Γ_s, the stabilizer of s, then satisfies

$$\sigma \Gamma_s \sigma^{-1} \cdot \{\pm 1\} = \left\{ \pm \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}^m \;\middle|\; m \in \mathbb{Z} \right\}, \quad \text{(for some } h > 0\text{)}.$$

Because f is invariant under the slash operator, then $f|[\sigma^{-1}]_k$ is invariant under $z \mapsto z+h$. Thus we may write $f|[\sigma^{-1}]_k$ as a meromorphic function $g(q)$ in the domain $0 < |q| < \epsilon$ for some ϵ, where $q = e^{2\pi i z/h}$. The condition that f is "meromorphic at the cusp s" then means that in the expansion:

$$f|[\sigma^{-1}]_k = g(q) = \sum_n a_n e^{2\pi i n z/h} \tag{5}$$

we have $a_n = 0$ for all $n < n_0$ for some fixed integer n_0. [1]

It is not hard to check that this last condition on being meromorphic at the cusps does not depend on the choice of σ mapping $\sigma(s) = \infty$ and needs to be checked only for the finite number of Γ-inequivalent cusps.

The expansion of $f|[\sigma^{-1}]_k$ of form (13) is often called the *Fourier expansion* for f at the cusp s (where $\sigma(s) = \infty$) and the coefficients a_n appearing in the series are called *Fourier coefficients*.

In light of our discussion of the complex structure of $\Gamma\backslash\mathcal{H}^*$ in Section 3.8, we see that the function $q(z) = e^{2\pi i z/h}$ may be regarded as a chart from the cusp at infinity. So the last condition is merely saying that the function is meromorphic on $\mathcal{H}^*$. In fact if $k = 0$, Definition 4.1 is equivalent to saying f is a meromorphic function on the quotient $\Gamma\backslash\mathcal{H}^*$. If $k = 2$, then since $\frac{d}{dz}(\gamma(z)) = j(\gamma, z)^{-2}$, Definition 4.1 is equivalent to saying f is a meromorphic differential 1-form on the compact Riemann surface $\Gamma\backslash\mathcal{H}^*$.

[1] This must be slightly adjusted if k is odd, and $-I$ is not in Γ. Then $\sigma\Gamma_s\sigma^{-1}$ may be generated by either $\begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}$ or $-\begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}$. In this latter case $f|[\sigma^{-1}]_k$ is multiplied by -1 under $z \mapsto z + h$. This adds the additional requirement that the function g is odd and the coordinate function is $q = e^{\pi i z/h}$.

The forms of weight $2k$ have interpretations in terms of $(dz)^k$-forms on $\Gamma\backslash\mathcal{H}^*$. This k-fold differential form can be locally written in the form

$$\omega = f(z)(dz)^k$$

and acting by an element $\gamma \in \mathrm{SL}(2,\mathbb{R})$, we obtain

$$\gamma \cdot \omega = f(\gamma(z))(d\gamma(z))^k = f(\gamma(z))\frac{d}{dz}(\gamma(z))^k(dz)^k.$$

This is just $f(z)(dz)^k$ for all $\gamma \in \Gamma$ if f is a weight $2k$ automorphic form for Γ. That is, automorphic forms of weight $2k$ for Γ correspond to meromorphic k-fold differential forms on $\Gamma\backslash\mathcal{H}^*$.

Definition 4.2 (Modular forms)**.** *A function $f : \mathcal{H} \to \mathbb{C}$ is called a modular form of weight k with respect to Γ if it satisfies the following conditions:*

- *f is holomorphic on $\mathcal{H}$,*
- *$f|[\gamma]_k = f$ for all $\gamma \in \Gamma$,*
- *f is holomorphic at every cusp s of Γ.*

Again, this last condition is just as for classical automorphic forms in Definition 4.1, except that the Fourier coefficients a_n in the expansion at any cusp s of Γ must now vanish for all negative integers n (i.e. $f|[\sigma^{-1}]_k = g(q)$ is holomorphic at $q = 0$). We conclude with a final definition:

Definition 4.3 (Cusp forms)**.** *A function $f : \mathcal{H} \to \mathbb{C}$ is called a cusp form of weight k with respect to Γ if it is a modular form whose constant Fourier coefficient a_0 vanishes in the expansion at every cusp of Γ.*

4.2 Dimensions of spaces of modular forms

Let $M_{2k}(\Gamma)$ denote the space of modular forms of weight $2k$ for a subgroup Γ of finite index in $\mathrm{SL}(2,\mathbb{Z})$. We have decided to restrict to this case because, although we can handle the case of odd weights for groups Γ that don't contain $-I$, this adds many caveats that make precise statements much more convoluted.

Theorem 4.2.1. *For $k \geq 2$, the dimension of $M_{2k}(\Gamma)$ is given by the formula*

$$\dim(M_{2k}(\Gamma)) = (2k-1)(g-1) + \nu_\infty k + \sum_p \lfloor k(1 - e_p^{-1}) \rfloor$$

where g denotes the genus of $X(\Gamma)$, ν_∞ is the number of Γ-inequivalent cusps, and the sum is taken over a set of representatives of elliptic points p of Γ. Finally, $\lfloor x \rfloor$ is the greatest integer function.

Remark 4.1. For $k = 0$, the space of modular forms $M_0(\Gamma)$ consists of holomorphic functions on $X(\Gamma)$, which is compact. Hence the only such functions are constant and $M_0(\Gamma) = \mathbb{C}$. As will be clear from the proof of our theorem, using the Reimann-Roch formula, the space of modular forms $M_k(\Gamma)$ is empty for k negative.

4.3 The Riemann-Roch theorem

Before diving into the proof, we recall the Riemann-Roch theorem, which counts the number of functions on a compact Riemann surface having a prescribed set of zeros and poles. Recall that we attach a divisor to a differential form as follows. Given a point $p \in X$, our Riemann surface, then let (U, z) be a coordinate neighborhood containing p. Then locally, the differential form $\omega = f(z)dz$ for some meromorphic f. We set $\mathrm{ord}_p(\omega) = \mathrm{ord}_p(f)$ so that, just as for functions,

$$\mathrm{div}(\omega) = \sum \mathrm{ord}_p(\omega)P.$$

Given any non-zero differential form ω, then any other is of the form $f\omega$ for some meromorphic function f. Hence the linear equivalence class of the divisor $\mathrm{div}(\omega)$ is independent of ω; let $K = \mathrm{div}(\omega)$ (as in the German "kanonisch").

Finally, let $\ell(D)$ denote the dimension of the space of functions

$$L(D) = \{f : X \to \mathbb{C},\ \text{meromorphic} \mid\ \mathrm{div}(f) + D \geq 0\},$$

where $D \geq 0$ for any divisor D means that all of the coefficients n_p of points $p \in X$ have $n_p \geq 0$. This depends only on the equivalence class of D.

Theorem 4.3.1. *Let X be a compact Riemann surface. Then there is an integer $g \geq 0$ such that for any divisor D,*

$$\ell(D) = \deg(D) + 1 - g + \ell(K - D). \tag{6}$$

A proof would take us too far afield. We refer the reader to Section 7 of Gunning's book.

Corollary 4.3.2. *A canonical divisor K has degree $2g - 2$, with $\ell(K) = g$.*

Proof. Set $D = 0$ in the Riemann-Roch theorem. Then $L(D)$ consists of functions f with $\mathrm{div}(f) \geq 0$; the only such functions are constant, and hence (6) implies $\ell(K) = g$. If instead we set $D = K$, then we get $\deg(K) = 2g - 2$. □

Generally speaking, the hard term to compute in (6) is $\ell(K - D)$. However, if $\deg(D) > 2g - 2$, then $L(K - D) = 0$. Indeed, under this assumption, $\deg(\mathrm{div}(f) + K - D) < 0$ for any meromorphic f on X. Thus we've shown

Corollary 4.3.3. *If* $\deg(D) > 2g - 2$*, then* $\ell(D) = \deg(D) + 1 - g$*.*

4.4 Proof of dimension formulas

In order to apply the Riemann-Roch theorem, we must understand the relationship between zeros and poles of k-fold differentials on a compact Riemann surface and the zeros and poles of the corresponding modular form.

Lemma 4.4.1. *Let f be an automorphic form of weight $2k$ for Γ corresponding to the k-fold differential ω on $X(\Gamma)$. Let $q \in \mathcal{H}^*$ denote the preimage of $p \in \Gamma\backslash\mathcal{H}^*$ under the canonical projection. We have the following relations between f and ω:*

- *If q is not a cusp nor elliptic point, then* $\mathrm{ord}_q(f) = \mathrm{ord}_p(\omega)$.
- *If q is an elliptic point of ramification index e,* $\mathrm{ord}_q(f) = e\,\mathrm{ord}_p(\omega) + k(e-1)$.
- *If q is a cusp, then* $\mathrm{ord}_q(f) = \mathrm{ord}_p(\omega) + k$.

Proof. We study the case where q is an elliptic point first. Recall that our complex structure at q was defined by taking a neighborhood $U \subset \mathcal{H}$ via isomorphism to the open unit disk D with q mapped to 0. Then acting on the unit disk by $\varphi : z \to z^e$. This may all be summarized in the commutative diagram:

$$\begin{array}{ccccc} q \in & \mathcal{H}^* \supset U & \xrightarrow[\simeq]{\lambda} & D & z \\ \downarrow & \pi\downarrow & & \downarrow\varphi & \\ p \in \Gamma\backslash\mathcal{H}^* \supset \Gamma_q\backslash U & & \xrightarrow[\simeq]{\phi} & D & z^e \end{array}$$

and we set $\phi(\pi(q)) = \varphi(\lambda(q)) = \lambda(z)^e$.

Now if g is a function on the target D with zero of order m, then $g^* = g \circ \varphi$ will have a zero of order me. Similarly, for a k-fold differential form ω on the target D, we set

$$\omega^* = g(\varphi(z))(d\varphi(z))^k = g(z^e)(ez^{e-1}dz)^k,$$

so the automorphic form f on the copy of D that's the domain of φ corresponds to ω^*. Again, since q was mapped via local isomorphism to the origin of the disk D, we indeed verify

$$\mathrm{ord}_0(f) = e\,\mathrm{ord}_0(\omega) + k(e-1).$$

Now if q is cuspidal, we chose the chart $\phi(\pi(q)) = e(\pi(q)/h)$. Let η denote the map from $\mathcal{H}$ to the coordinate on the punctured disk $z' = \eta(z) = e(z/h)$ where $e(x) := e^{2\pi i x}$. Then consider the differential

$$\omega^* = g(z')(dz)^k.$$

Then $dz' = (2\pi i/h)z'dz$. Lifting the differential to ω on $\mathcal{H}$ we get

$$\omega = (2\pi i/h)^k g(\eta(z))\eta(z)^k (dz)^k$$

and so the corresponding automorphic form f is $(2\pi i/h)^k g(\eta(z))\eta(z)^k$, i.e., $f^*(q) = (2\pi i/h)^k g(q)q^k$, which gives the result.

Finally for the points that are neither cuspidal nor elliptic, π is a local isomorphism from the neighborhood U to $\Gamma_q\backslash U$, so we just get immediate equality. □

With this lemma, we are at last ready to find the dimensions of the spaces of automorphic forms of weight $2k$.

Proof of Theorem 4.2.1. Since f is holomorphic, we have $\mathrm{ord}_q(f) \geq 0$ for all points q in $\mathcal{H}^*$. In view of the previous lemma, this forces the following inequalities on the corresponding k-fold differential ω at points $p \in X(\Gamma)$:

$$\begin{aligned} e\,\mathrm{ord}_p(\omega) + k(e-1) &\geq 0, \text{ if } p \text{ is the image of an elliptic point} \\ \mathrm{ord}_p(\omega) + k &\geq 0, \text{ if } p \text{ is the image of a cusp} \\ \mathrm{ord}_p(\omega) &\geq 0, \text{ if } p \text{ is neither the image of a cusp nor elliptic point.} \end{aligned}$$

Now given any k-fold differential ω_0 then $h\omega_0 = \omega$ for some meromorphic function h, with

$$\begin{aligned} \mathrm{ord}_p(h) + \mathrm{ord}_p(\omega_0) + k(1-1/e) &\geq 0, \text{ if } p = \pi(q),\ q\text{: elliptic point} \\ \mathrm{ord}_p(h) + \mathrm{ord}_p(\omega_0) + k &\geq 0, \text{ if } p = \pi(q),\ q\text{: cusp} \\ \mathrm{ord}_p(h) + \mathrm{ord}_p(\omega_0) &\geq 0, \text{ if } p = \pi(q),\ q\text{: neither.} \end{aligned}$$

Combining these, we have

$$\mathrm{div}(h) + D \geq 0$$

where

$$D = \operatorname{div}(\omega_0) + \sum_{i:p_i \leftrightarrow \text{ cusp}} kp_i + \sum_{i:p_i \leftrightarrow \text{ elliptic}} \lfloor k(1-1/e_i) \rfloor p_i$$

By Corollary 4.3.2, the degree of the canonical divisor of a 1-form is $2g-2$. Hence the degree of a k-fold differential is $k(2g-2)$. Thus the degree of D is

$$k(2g-2) + k \cdot \nu_\infty + \sum_{i:p_i \leftrightarrow \text{ elliptic}} \lfloor k(1-1/e_i) \rfloor$$

Applying the Riemann-Roch theorem with this D, we may use Corollary 4.3.3 to conclude the result. □

For example, we may apply Theorem 4.2.1 to $\Gamma(1)$, the full modular group, to obtain

$$\dim(M_{2k}) = 1 - k + \lfloor k/2 \rfloor + \lfloor 2k/3 \rfloor, \quad k > 1.$$

The theorem shows that modular forms are abundant on finite index subgroups of $\mathrm{SL}(2,\mathbb{Z})$, and shortly we will investigate ways of constructing bases for the vector space of weight $2k$ modular forms.

4.5 Modular forms as sections of line bundles

Given a Riemann surface X, a line bundle is a map of complex manifolds $\pi : L \to X$ such that, for some open cover $X = \bigcup_i U_i$, $\pi^{-1}(U_i)$ is locally isomorphic to $U_i \times \mathbb{C}$. For any open set $U \subset X$, let $\Gamma(U, L)$ denote the group of sections of L over U. That is,

$$\Gamma(U, L) = \{ f : U \to L, \text{ holomorphic} \mid \pi \circ f = \mathrm{id} \}$$

For example, if $L = U \times \mathbb{C}$ then $\Gamma(U, L)$ is just the holomorphic functions on U.

Let p be the quotient map $\mathcal{H}$ to $Y(\Gamma) = \Gamma \backslash \mathcal{H}$. Then given a line bundle $\pi : L \to Y$ we may construct a line bundle over $\mathcal{H}$ by

$$p^*(L) = \{(h, l) \in \mathcal{H} \times L \mid p(h) = \pi(l)\}.$$

Moreover, Γ acts on $p^*(L)$ according to its action on the component in $\mathcal{H}$. If we are given an isomorphism $\phi : \mathcal{H} \times \mathbb{C} \to p^*(L)$, then we may translate the action of Γ on $p^*(L)$ into an action on $\mathcal{H} \times \mathbb{C}$ over $\mathcal{H}$.

For $\gamma \in \Gamma$ and $(\tau, z) \in \mathcal{H} \times \mathbb{C}$, we write this action formally as

$$\gamma \cdot (\tau, z) = (\gamma\tau, j_\gamma(\tau) z), \quad j_\gamma(\tau) \in \mathbb{C}^\times$$

Then our formal expression for $\gamma\gamma'(\tau, z)$ takes form

$$\gamma\gamma'(\tau z) = \gamma(\gamma'\tau, j_{\gamma'}(\tau)z) = (\gamma\gamma'\tau, j_\gamma(\gamma'\tau) \cdot j_{\gamma'}(\tau)z).$$

Since this is an action, we must have:

$$j_{\gamma\gamma'}(\tau) = j_\gamma(\gamma'\tau) \cdot j_{\gamma'}(\tau).$$

This reminds us of the multiplicative version of the cocycle condition.

Definition 4.4. *An automorphy factor is a map $j : \Gamma \times \mathcal{H} \to \mathbb{C}^\times$ such that*

- *For each $\gamma \in \Gamma$, the map $\tau \to j_\gamma(\tau)$ is a holomorphic function on $\mathcal{H}$.*
- *j satisfies the cocycle condition $j_{\gamma\gamma'}(\tau) = j_\gamma(\gamma'\tau) \cdot j_{\gamma'}(\tau)$ for all $\gamma, \gamma' \in \Gamma$ and $\tau \in \mathcal{H}$.*

There is a canonical automorphy factor coming from the tangent space of the action by Γ. Indeed, if we consider the map

$$\Gamma \times \mathcal{H} \to \mathbb{C} : \quad (\gamma, \tau) \mapsto (d\gamma)_\tau$$

where $(d\gamma)_\tau$ denotes the map on the tangent space at τ defined by the map $\gamma : \mathcal{H} \to \mathcal{H}$.

In general, if M, N, P are complex manifolds with maps

$$M \xrightarrow{\alpha} N \xrightarrow{\beta} P$$

then for any point $m \in M$, we have the identity $(d(\beta \circ \alpha))_m = (d\beta)_{\alpha(m)} \circ (d\alpha)_m$ as maps on tangent spaces. This implies that if we set $j_{\gamma\gamma'}(\tau) = (d\gamma\gamma')_\tau$, the map satisfies the required cocycle condition.

We have already seen this choice of automorphy previously. Since γ acts by linear fractional transformation, we have that $d\gamma$ (as a map $z \mapsto \frac{az+b}{cz+d}$) satisfies

$$d\gamma = \frac{1}{(cz+d)^2} dz$$

so that $j_\gamma(z) = (cz+d)^{-2}$.

Proposition 4.5.1. *There is a one-to-one correspondence between pairs (L, ϕ), where L is a line bundle on $Y(\Gamma)$ and ϕ is an isomorphism $\mathcal{H} \times \mathbb{C} \simeq p^*(L)$, and the set of automorphy factors for $\mathcal{H}$*

Proof. Given (L, ϕ) we have seen how the formal definition of the action on the complex component gives rise to a factor $j_\gamma(\tau)$ with the required properties. For the converse, we may use ϕ and j to define an action of Γ on $\mathcal{H} \times \mathbb{C}$. Then we may define L to be $\Gamma\backslash\mathcal{H} \times \mathbb{C}$ with respect to this action. □

Note that since every line bundle on $\mathcal{H}$ is trivial (that is, there exists an isomorphism ϕ to $\mathcal{H} \times \mathbb{C}$), the previous proposition classifies all line bundles on $\Gamma\backslash\mathcal{H}$.

Given a line bundle L on $Y = Y(\Gamma)$. Then the space of global sections

$$\Gamma(Y, L) = \{F \in \Gamma(H, p^*L) \mid f \text{ commutes with the action of } \Gamma\}.$$

We now explain this latter condition more concretely. Given an isomorphism $\phi : p^*L \to \mathcal{H} \times \mathbb{C}$, then we noted earlier that the action of Γ on $\mathcal{H} \times \mathbb{C}$ may be described in terms of an automorphy factor:

$$\gamma(\tau, z) = (\gamma\tau, j_\gamma(\tau)z).$$

A holomorphic section $F \in \Gamma(\mathcal{H}, L)$ is a map $F(\tau) = (\tau, f(\tau))$ from $\mathcal{H} \to \mathcal{H} \times \mathbb{C}$ where f is a holomorphic map $\mathcal{H} \to \mathbb{C}$. In order for F to commute with the action of γ we must have

$$F(\gamma\tau) = \gamma F(\tau) \iff (\gamma\tau, f(\gamma\tau)) = (\gamma\tau, j_\gamma(\tau)f(\tau)),$$

or more simply

$$f(\gamma\tau) = j_\gamma(\tau)f(\tau).$$

Note that if j is an automorphy factor, so is j^k for any integer k. Hence, taking L_k to be the line bundle on $\Gamma\backslash\mathcal{H}$ corresponding to $j_\gamma(\tau)^{-k} = (d\gamma)_\tau^{-k}$, this condition is precisely the transformation property for weight $2k$ automorphic forms.

Thus the global sections of L_k are in one-to-one correspondence with holomorphic functions on $\mathcal{H}$ satisfying the transformation property. The line bundle L_k may be extended to the compactification $X(\Gamma)$ of $Y(\Gamma)$ and the resulting global sections are thus modular forms of weight $2k$.

4.6 Poincaré Series

We have seen the construction for Eisenstein series arise naturally in the algebraic relation between the Weierstrass function and its derivative, or as a homogeneous function on a lattice. In this section, we explore a more robust way of constructing invariant functions via averaging.

Suppose we want to construct a function f on $\mathcal{H}$ such that $f(\gamma z) = j_\gamma(z)f(z)$ for some automorphy factor j (e.g., $j_\gamma(z) = (cz+d)^k$). Then we may try to define

$$f(z) = \sum_{\gamma \in \overline{\Gamma}} \frac{g(\gamma z)}{j_\gamma(z)},$$

where g is a function on $\mathcal{H}$ with suitably nice growth properties. If this series converges absolutely uniformly on compact sets, then

$$f(\gamma' z) = \sum_{\gamma \in \overline{\Gamma}} \frac{g(\gamma\gamma' z)}{j_\gamma(\gamma' z)} = \sum_{\gamma \in \overline{\Gamma}} \frac{g(\gamma\gamma' z)}{j_{\gamma\gamma'}(z)} j_{\gamma'}(z) = j_{\gamma'}(z) f(z).$$

However, this averaging construction has an obvious flaw in that we're summing over infinitely many matrices with $j_\gamma(z) = 1$, so finding a suitable choice of g is quite difficult.

If $j_\gamma(z) = (cz+d)^k$, then it seems much more reasonable to sum over pairs (c,d) rather than all matrices in Γ. Put another way, $j_\gamma(z) = 1$ if $(c,d) = (0,1)$, so the set of elements in Γ with $j_\gamma(z) = 1$ are precisely the stabilizer of ∞:

$$\Gamma_\infty = \left\{ \pm \begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}^m \;\middle|\; m \in \mathbb{Z} \right\} \quad \text{(for some } h > 0\text{)}.$$

Given any function g on $\mathcal{H}$ with $g(\gamma z) = g(z)$ for all $\gamma \in \Gamma_\infty$, we may consider the average

$$f(z) = \sum_{\gamma \in \Gamma_\infty \backslash \Gamma} \frac{g(\gamma z)}{j_\gamma(z)}.$$

If g is holomorphic and the series converges absolutely uniformly on compact sets, then we obtain a holomorphic function f with $f(\gamma z) = j_\gamma(z) f(z)$.

Definition 4.5. *The Poincaré series of weight $2k$ and character n for Γ is defined by*

$$P_n^{2k}(z) = P_n(z) = \sum_{\gamma \in \Gamma_\infty \backslash \Gamma} \frac{e(n\gamma(z)/h)}{(cz+d)^{2k}},$$

where $e(x) := e^{2\pi i x}$, we're assuming each coset representative γ has form $\begin{pmatrix} a & b \\ c & d \end{pmatrix}$, and that Γ_∞ is generated by translations $z \mapsto z + h$ for some $h > 0$.

More generally, we may define the Poincaré series with respect to any cusp s of Γ by choosing $\sigma \in \mathrm{SL}(2,\mathbb{R})$ such that $\sigma(s) = \infty$. Then $\Gamma_s = \sigma^{-1}\Gamma_\infty\sigma$. Then replacing $g(\gamma z)$ by $g(\sigma\gamma z)$ and $j_\gamma(z)$ by $j_{\sigma\gamma}(z)$, we obtain an invariant function by averaging over $\Gamma_s\backslash\Gamma$.

Proposition 4.6.1. *The Poincaré series $P_n(z)$ of weight $2k > 2$ converges absolutely on compact subsets of $\mathcal{H}$.*

Proof. A set of coset representatives for $\Gamma_\infty\backslash\Gamma$ is given by taking a single representative for each possible bottom row (c, d) of elements in Γ. The convergence may be checked by comparison with the series

$$\sum_{(m,n)\neq(0,0)} \frac{1}{|mz+n|^s}, \quad s > 2$$

which follows from Lemma 2.2.1. □

4.7 Fourier coefficients of Poincaré series

We show that the Poincaré series are modular forms by examining their Fourier expansions at each cusp. Before beginning the proof of the expansion, we require an important decomposition theorem.

Proposition 4.7.1 (Bruhat decomposition)**.** *Let s and t be cusps for Γ with $\sigma_s(s) = \sigma_t(t) = \infty$ for matrices $\sigma_s, \sigma_t \in \mathrm{SL}(2,\mathbb{R})$. Then we have the following decomposition of $\sigma_s\overline{\Gamma}\sigma_t^{-1}$ into disjoint double cosets of B, the set of upper triangular matrices in $\overline{\Gamma}$:*

$$\sigma_s\overline{\Gamma}\sigma_t^{-1} = \delta_{s,t}B \cup \bigcup_{c>0} \bigcup_{d \,(\mathrm{mod}\ c)} B\begin{pmatrix} * & * \\ c & d \end{pmatrix} B$$

where $\delta_{s,t}B$ is empty unless $s = t$ (in which case it equals B), and the union is taken over representatives in $\sigma_s\overline{\Gamma}\sigma_t^{-1}$ having bottom row (c, d) satisfying the subscripted conditions.

Proof. We first examine the elements $\omega \in \sigma_s\overline{\Gamma}\sigma_t^{-1}$ with lower-right entry $c = 0$. These matrices fix ∞, so we set

$$\Omega_{s,t} = \left\{\omega \in \sigma_s\overline{\Gamma}\sigma_t^{-1} \mid \omega(\infty) = \infty\right\}.$$

If this set is non-empty, containing say $\omega = \sigma_s \gamma \sigma_t^{-1}$ for some $\gamma \in \overline{\Gamma}$, then $\gamma(t) = \sigma_s^{-1}\omega\sigma_t(t) = s$ so in fact s and t are equivalent cusps. Then $\gamma \in \Gamma_s$, the stabilizer of the cusp s, and $\omega \in B$, the stabilizer of ∞. Hence,

$$\Omega_{s,t} = \begin{cases} B & \text{if } s = t, \\ \emptyset & \text{otherwise.} \end{cases}$$

If $c > 0$ for an element in $\sigma_s \overline{\Gamma} \sigma_t^{-1}$, the matrix identity

$$\begin{pmatrix} 1 & m \\ 0 & 1 \end{pmatrix} \begin{pmatrix} a & * \\ c & d \end{pmatrix} \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} a + cm & * \\ c & d + cn \end{pmatrix}$$

shows that the double coset $B \begin{pmatrix} * & * \\ c & d \end{pmatrix} B$ determines c uniquely and $d \pmod c$. Moreover, we claim that this double coset does not depend on the top row. Indeed, if ω, ω' are matrices in $\sigma_s \overline{\Gamma} \sigma_t^{-1}$ having the same bottom row, then $\omega' \omega^{-1}$ is in B. This implies $a' = a + cm$ for some m (where a' is the top-right entry of ω' and a is the top-right entry of ω). □

Theorem 4.7.2. *The Poincaré series $P_n(z)$ of weight $2k$ has Fourier expansion at the cusp s given by*

$$P_n(z)|[\sigma^{-1}]_k = \delta_{\infty,s} e(nz/h) + \sum_{m=1}^{\infty} e(mz/h) \sum_{c>0} S_\Gamma(n/h, m/h; c) \mathcal{J}_c(n/h, m/h),$$

where $S_\Gamma(n, m; c)$ is the Kloosterman sum with respect to Γ and cusp s defined by

$$S_{\Gamma,s}(n, m, c) := S_\Gamma(n, m; c) = \sum_{\substack{\begin{pmatrix} a & * \\ c & d \end{pmatrix} \in B \backslash \overline{\Gamma} \sigma^{-1} / B \\ c \text{: fixed}}} e\left(\frac{na + md}{c}\right),$$

and $\mathcal{J}_c(n/h, m/h)$ is a certain Bessel function to be described in the proof.

Proof. Using the double coset decomposition given in Proposition 4.7.1, we write $P_n(z)|[\sigma^{-1}]_{2k}$ in terms of double cosets $B\backslash\overline{\Gamma}\sigma^{-1}/B$.[2] This gives

$$P_n(z)|[\sigma^{-1}]_{2k} = \delta_{\infty,s} e(nz/h) + \sum_{\gamma \in B \backslash \overline{\Gamma} \sigma^{-1} / B} I_\gamma(z) \tag{7}$$

[2]Note that this is slightly less general than our decomposition in Proposition 4.7.1, since we're only finding the Fourier coefficients for the series with respect to ∞. The expansion at a cusp t for the series with respect to s uses the full generality of the Bruhat decomposition.

where, if $\gamma = \begin{pmatrix} a & * \\ c & d \end{pmatrix} \in \overline{\Gamma}\sigma^{-1}$ with $c > 0$,

$$\begin{aligned} I_\gamma(z) &= \sum_{\tau \in B} j_{\gamma\tau}(z)^{-2k} e(n\gamma\tau z/h) \\ &= \sum_{m \in \mathbb{Z}} (c(z+mh)+d)^{-2k} e\left(\frac{n}{h}\left(\frac{a}{c} - \frac{1}{c^2(z+mh)+cd}\right)\right). \end{aligned}$$

By Poisson summation formula, this is

$$I_\gamma(z) = \frac{1}{h}\sum_{m \in \mathbb{Z}} \int_{-\infty}^{\infty} (c(z+u)+d)^{-2k} e\left(\frac{n}{h}\left(\frac{a}{c} - \frac{1}{c^2(z+u)+cd}\right)\right) e(-mu/h)du.$$

Now performing the change of variable $u \mapsto u - z - \frac{d}{c}$, we get

$$\begin{aligned} I_\gamma(z) &= \frac{1}{h}\sum_{m \in \mathbb{Z}} \int_{-\infty}^{\infty} (cu)^{-2k} e\left(\frac{n}{h}\left(\frac{a}{c} - \frac{1}{c^2 u}\right)\right) e\left(m/h\left(-u + z + \frac{d}{c}\right)\right) du \\ &= \frac{1}{h}\sum_{m \in \mathbb{Z}} e\left(mz/h + \frac{an+md}{ch}\right) \int_{-\infty}^{\infty} (cu)^{-2k} e\left(-(m/h)u + \frac{(n/h)}{c^2 u}\right) du. \end{aligned}$$

We denote this latter integral by

$$\mathcal{J}_c(n/h, m/h) = \frac{1}{h}\int_{-\infty}^{\infty} (cu)^{-2k} e\left(-(m/h)u + \frac{(n/h)}{c^2 u}\right) du. \tag{8}$$

Note that we may take this integration over any horizontal line at height y in the complex plane, and result will be independent of y (by Cauchy's theorem). If $m \leq 0$, then taking $y \to \infty$ we see that $\mathcal{J}_c(n/h, m/h) = 0$ for $m \leq 0$. If $m > 0$, then we have

$$\mathcal{J}_c(0, m/h) = \left(\frac{2\pi}{ich}\right)^{2k} \frac{m^{2k-1}}{\Gamma(2k)}$$

and if both $m, n > 0$, we have

$$\mathcal{J}_c(n/h, m/h) = \frac{2\pi}{i^{2k} ch} (m/n)^{\frac{2k-1}{2}} J_{2k-1}\left(\frac{4\pi\sqrt{mn}}{ch}\right),$$

where J_ν is the Bessel function of order ν having power series representation

$$J_\nu(x) = \sum_{\ell=0}^{\infty} \frac{(-1)^\ell}{\ell!\Gamma(\ell+1+\nu)} \left(\frac{x}{2}\right)^{\nu+2\ell}.$$

Both cases of the integration in (8) – $n = 0$ and $n > 0$ – can be found in the tables of Gradshteyn and Rhyzik, "Tables of Integrals, Series, and Products," Tables 8.315.1 and 8.412.2, respectively. Putting this back into the decomposition (7), we have

$$P_n(z)|[\sigma^{-1}]_{2k} = \delta_{\infty,s} e(nz/h) + \sum_{m=1}^{\infty} e(mz/h) \sum_{\gamma \in B\backslash \overline{\Gamma}\sigma^{-1}/B} e\left(\frac{a(n/h) + (m/h)d}{c}\right) \mathcal{J}_c(n/h, m/h).$$

Since the double coset space is parametrized by $c > 0, d (\text{mod } c)$, we may rewrite this latter term over the double cosets as:

$$\sum_{\gamma \in B\backslash \overline{\Gamma}\sigma^{-1}/B} e\left(\frac{a(n/h) + (m/h)d}{c}\right) \mathcal{J}_c(n/h, m/h) = \sum_{c>0} S_\Gamma(n/h, m/h; c) \mathcal{J}_c(n/h, m/h),$$

which gives the result. □

From the shape of the Fourier expansion, and in particular the fact that it includes no negative powers of $e(z/h)$, we immediately conclude the following.

Corollary 4.7.3. *$P_0(z)$ is zero at all cusps except ∞ where $P_0(\infty) = 1$ (hence P_0 is a modular form of weight $2k$ for Γ). If $n \geq 1$, then $P_n(z)$ is a cusp form.*

Note that for $n = 0$ and $\Gamma = \text{SL}(2, \mathbb{Z})$, this Fourier expansion should reduce to the earlier one presented for Eisenstein series. For $\Gamma = \Gamma(1)$, the Kloosterman sum reduces to the one originally defined by Kloosterman in his 1926 paper. (They arose naturally for him in the study of representations of quadratic forms.) The sum has form

$$S(n, m; c) = \sum_{ad \equiv 1 \text{ (mod c)}} e\left(\frac{na + md}{c}\right).$$

Now if $n = 0$, this just degenerates to a Ramanujan sum:

$$S(0, m; c) = \sum_{d \neq 0 \text{ (mod c)}} e\left(\frac{md}{c}\right) = \sum_{\ell | \gcd(c,m)} \mu\left(\frac{c}{\ell}\right) \ell,$$

and hence

$$\sum_{c>0} c^{-k} S(0, m; c) = \zeta(k)^{-1} \sum_{c|m} c^{1-k} = \frac{\sigma_{k-1}(m)}{m^{k-1}\zeta(k)}.$$

4.8 The Hilbert space of cusp forms

We have seen that the differential $y^{-2}dxdy$ is invariant under the action of $\mathrm{SL}(2,\mathbb{R})$, and hence Haar measure for the group is given by

$$\mu(U) = \iint_U \frac{dxdy}{y^2}.$$

Let Γ be a Fuchsian group of the first kind. By the invariance of this measure, we see that the measure of any fundamental domain D for $\Gamma\backslash\mathcal{H}$ is independent of the choice of D and finite. In fact, Shimura explicitly calculates this volume in Theorem 2.20, Section 2.5 of his book, obtaining

$$\mu(D) = \iint_D \frac{dxdy}{y^2} = 2\pi\left(2g - 2 + \nu_\infty + \sum_{p\text{: elliptic}} (1 - \frac{1}{e_p})\right)$$

where as before, g is the genus of $X(\Gamma)$, ν_∞ is the number of inequivalent cusps of ∞ and e_p is the ramification index at a point p, the image of an elliptic point. The proof is somewhat involved, and we won't reproduce it here, but we note that this quantity did arise in the computation of $\deg(\mathrm{div}(f))$ for an automorphic form f with respect to Γ.

Lemma 4.8.1. *Let f and g be modular forms of weight k with respect to Γ. Then the differential*

$$f(z)\,\overline{g(z)}\,y^k\frac{dxdy}{y^2}$$

is invariant with respect to the action of $\mathrm{SL}(2,\mathbb{R})$.

Proof. This follows from the transformation properties of f and g and our familiar identity

$$\Im(\gamma z) = \frac{\Im(z)}{|cz+d|^2},$$

together with the invariance of the differential $y^{-2}dxdy$ under the action of $\mathrm{SL}(2,\mathbb{R})$. □

Lemma 4.8.2. *Let D be a fundamental domain for Γ. Provided at least one of f and g is a cusp form,*

$$\iint_D f(z)\,\overline{g(z)}\,y^k\frac{dxdy}{y^2}$$

converges.

Proof. If we integrate over the subset of D outside a neighborhood of each cusp, then this subset is contained in a compact set and so the integral converges. Thus it suffices to prove that the integral converges on a neighborhood of the cusp at ∞, as all other cusps may be handled similarly by translating them to ∞. Near the cusp at ∞, $f(z)\,\overline{g(z)} = O(e^{-cy})$ for some $c > 0$, since one of f or g is cuspidal. This ensures that (up to constant) the integral is dominated by $\int_\ell^\infty e^{-cy} y^{k-2} dy < \infty$. □

Definition 4.6. *Given two modular forms f and g of weight k for Γ such that at least one of f and g is cuspidal, we define the Petersson inner product of f and g by the integral*

$$\langle f, g\rangle = \iint_D f(z)\,\overline{g(z)}\, y^k \frac{dxdy}{y^2}.$$

In particular, for a cusp form f we set

$$||f||^2 = \langle f, f\rangle = \iint_D |f(z)|^2\, y^k \frac{dxdy}{y^2}.$$

Remark 4.2. By our previous two lemmas, the Petersson inner product defines a positive-definite Hermitian form on $S_k(\Gamma)$, which endows $S_k(\Gamma)$ with the structure of a finite-dimensional Hilbert space.

Theorem 4.8.3. *Let f be a modular form of weight k and $P_n(z)$, $n > 0$, the Poincaré series of weight k for Γ. Then*

$$\langle f, P_n\rangle = \Gamma(k-1)\left(\frac{h^k}{(4\pi n)^{k-1}}\right) a(n)$$

where $a(n)$ is the n-th Fourier coefficient in the expansion of $f(z) = \sum_n a(n)e(nz/h)$.

Proof.

$$\begin{aligned}
\langle f, P_n\rangle &= \int_{\Gamma\backslash\mathcal{H}} y^k f(z) \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} \overline{j_\gamma(z)}^{-k}\, \overline{e(n\gamma z/h)} d\mu \\
&= \int_{\Gamma\backslash\mathcal{H}} \sum_{\gamma\in\Gamma_\infty\backslash\Gamma} \Im(\gamma z)^k f(\gamma z)\, \overline{e(n\gamma z/h)} d\mu \\
&= \int_0^\infty \int_0^h y^{k-2} f(z) \overline{e(nz/h)} dx\, dy \\
&= \int_0^\infty \int_0^h y^{k-2} \left(\sum_{m=0}^\infty a(m) e(mz/h)\right) \overline{e(nz/h)} d\mu
\end{aligned}$$

The only one of these terms in the Fourier series to contribute is $m = n$, by orthogonality. This leaves

$$\langle f, P_n \rangle = h\, a(n) \int_0^\infty y^{k-2} e(-4\pi n y/h) dy = \frac{\Gamma(k-1)}{(4\pi n)^{k-1}} h^k a(n)$$

□

Corollary 4.8.4. *Every cusp form is a linear combination of Poincaré series $P_n(z)$ with $n \geq 1$.*

Proof. The linear space spanned by P_n, $n \geq 1$, is closed because $S_k(\Gamma)$ has finite dimension. A function orthogonal to this subspace must be zero since all of its Fourier coefficients vanish according to Theorem 4.8.3. □

Surprisingly, there are many basic questions about Poincaré series that are still open. For example, there does not exist an explicit description of which Poincaré series form a basis of Γ (for general Fuchsian groups of the first kind). This answer is known for the full modular group $\Gamma(1)$ where we can just take the series $P_n(z)$ with $1 \leq n \leq \dim(S_k(\Gamma))$. Another open problem is to determine conditions for which a Poincaré series is identically 0. For some discussion of work in this direction see:

- Irwin Kra, "On the vanishing of and spanning sets for Poincaré series for cusp forms," *Acta Math.* **vol. 153** 47–116. (1984)

4.9 Basic estimates for Kloosterman sums

We noted that the form of Kloosterman sums appearing in the Fourier coefficients of Poincaré series $P_n(z)$ (Theorem 4.7.2) was not in full generality. We could have averaged over the stabilizer of a cusp other than ∞. This would have resulted in Fourier coefficients which contain the Poincaré series

$$S_{\Gamma,(s,t)}(n,m;c) = \sum_{\substack{\left(\begin{smallmatrix} a & * \\ c & d \end{smallmatrix}\right) \in B\backslash \sigma_s \overline{\Gamma} \sigma_t^{-1}/B \\ c \text{: fixed}}} e\left(\frac{na+md}{c}\right).$$

Thus it is of great interest to try to evaluate these sums, or more practically, to bound their size. Given a pair of cusps s, t in Γ, let

$$C(s,t) = \left\{ c > 0 \;\middle|\; \begin{pmatrix} a & * \\ c & d \end{pmatrix} \in B\backslash \sigma_s \overline{\Gamma} \sigma_t^{-1}/B \right\},$$

and let $c(s,t)$ be the smallest element in $C(s,t)$. One can show that it exists. In particular $c(s,s)^{-1}$ can be shown to be the radius of the largest circle bounding the fundamental domain for $\sigma_s \Gamma \sigma_s^{-1}$.

Proposition 4.9.1. *For any $c \in C(s,t)$,*

$$\left| \left\{ d \ (mod\ c) \ \middle| \ \begin{pmatrix} a & * \\ c & d \end{pmatrix} \in B \backslash \sigma_s \overline{\Gamma} \sigma_t^{-1} / B \right\} \right| \leq c^2 \cdot \max\{c(s,s), c(t,t)\}$$

which gives a bound for $|S_{\Gamma,(s,t)}(n,m;c)| \leq c^2 \cdot \max\{c(s,s), c(t,t)\}$ by trivial estimate.

Proof. See Iwaniec's "Topics in Classical Automorphic Forms," Proposition 2.8. □

Let's attempt to evaluate the classical Kloosterman sum, at least for special choices of the modulus c. Recall, this classical sum has form

$$S(n,m;c) = \sum_{ad \equiv 1} e\left(\frac{na+md}{c}\right).$$

Proposition 4.9.2. *The Kloosterman sum $S(n,m;c)$ satisfies the following properties:*

1. $S(n,m;c) = S(m,n;c)$;
2. $S(an,m;c) = S(n,am;c)$ *if* $\gcd(a,c) = 1$;
3. $S(n,m;c) = \sum_{d \mid \gcd(c,m,n)} dS(mnd^{-2}, 1; cd^{-1})$;
4. *If $c = d_1 d_2$ with $\gcd(d_1,d_2) = 1$, and $\overline{d_1}$, $\overline{d_2}$ denote the multiplicative inverses of d_1 (mod d_2) and d_2 (mod d_1), respectively, then*

$$S(m,n;c) = S(\overline{d_1}m, \overline{d_1}n; d_2) S(\overline{d_2}m, \overline{d_2}n; d_1).$$

Proof. We leave this as an exercise to the reader. The third property is genuinely hard, and was originally proved by Selberg using the Kuznetsov trace formula, though elementary proofs are now known. □

Note that property (4) in the above proposition reduces the determination of Kloosterman sums to prime power moduli.

Lemma 4.9.3. *If the modulus $c = p^{2\alpha}$ with $\alpha \geq 1$ and $\gcd(p, 2n) = 1$ then*

$$S(n, n; p^{2\alpha}) = p^{\alpha}\left(e\left(\frac{2n}{p^{2\alpha}}\right) + e\left(\frac{-2n}{p^{2\alpha}}\right)\right) = 2p^{\alpha}\,\mathrm{Re}\left[e\left(\frac{2n}{p^{2\alpha}}\right)\right]$$

Proof. Consider the Kloosterman sum as running over d and $\overline{d}$ mod $p^{2\alpha}$. We reparametrize the sum by setting $d = a(1 + bp^{\alpha})$ where a takes values mod $p^{2\alpha}$ with $\gcd(a, p) = 1$, and b takes any value mod p^{α}. Then every residue class prime to p mod $p^{2\alpha}$ is attained p^{α} times. Then we have $\overline{d} = \overline{a}(1 - bp^{\alpha}) \pmod{p^{2\alpha}}$ so

$$\begin{aligned} S(n, n; p^{2\alpha}) &= p^{-\alpha} \sum_{a \,(\mathrm{mod}^{\times} p^{2\alpha})} \sum_{b \,(\mathrm{mod}\ p^{\alpha})} e\left(n\frac{a + \overline{a}}{p^{2\alpha}} + n\frac{(a - \overline{a})b}{p^{\alpha}}\right) \\ &= \sum_{\substack{a \,(\mathrm{mod}^{\times} p^{2\alpha}) \\ a \equiv \overline{a} \,(\mathrm{mod}\ p^{\alpha})}} e\left(n\frac{a + \overline{a}}{p^{2\alpha}}\right) \end{aligned}$$

The solutions to $a \equiv \overline{a} \pmod{p^{\alpha}}$ are of the form $a = \pm 1 + tp^{\alpha}$ (with $\overline{a} = \pm 1 - tp^{\alpha}$) for any $t \pmod{p^{\alpha}}$. This gives the result. □

Lemma 4.9.4. *If the modulus $c = p^{2\alpha+1}$ with $\alpha \geq 1$ and $\gcd(p, 2n) = 1$ then*

$$S(n, n; p^{2\alpha+1}) = 2\left(\frac{n}{p}\right) p^{\alpha+1/2}\,\mathrm{Re}\left[\varepsilon_c e\left(\frac{2n}{p^{2\alpha+1}}\right)\right]$$

where $\left(\frac{n}{p}\right)$ is the Legendre symbol, and ε_p is 1 or i according to whether $p \equiv 1$ or -1 (mod 4).

Proof. We proceed just as in the case of even prime powers, setting $d = a(1 + bp^{\alpha+1})$ with a and b as before. Then we reduce to

$$S(n, n; p^{2\alpha+1}) = \sum_{\substack{a \,(\mathrm{mod}^{\times} p^{2\alpha+1}) \\ a \equiv \overline{a} \,(\mathrm{mod} p^{\alpha})}} e\left(n\frac{a + \overline{a}}{p^{2\alpha+1}}\right).$$

Just as before the solutions to $a \equiv \overline{a}$ are $a = \pm 1 + tp^{\alpha}$ where t ranges freely mod $p^{\alpha+1}$, though now $\overline{a} = \pm 1 - tp^{\alpha} \pm t^2 p^{2\alpha}$. Thus

$$\begin{aligned} S(n, n; p^{2\alpha+1}) &= 2\,\mathrm{Re}\left[\sum_{t \,(\mathrm{mod}\ p^{\alpha+1})} e\left(n\frac{2 + t^2 p^{2\alpha}}{p^{2\alpha+1}}\right)\right] \\ &= 2\,\mathrm{Re}\left[p^{\alpha} e\left(\frac{2n}{p^{2\alpha+1}}\right) g(n, p)\right] \end{aligned}$$

where $g(n,p)$ is the Gauss sum

$$g(n,p) = \sum_{t \pmod p} e\left(\frac{nt^2}{p}\right) = \varepsilon_p \left(\frac{n}{p}\right) p^{1/2},$$

where the latter equality is the famous evaluation due to Gauss (see, for example, Ch. 2 Davenport's "Multiplicative Number Theory"). This gives the result. □

In fact we see that both results can be unified into a single statement about odd prime powers p^β with $\beta \geq 2$ and $\gcd(p,n) = 1$. This can be written

$$S(n,n;p^\beta) = 2\left(\frac{n}{p}\right)^\beta p^{\beta/2} \operatorname{Re}\left[\varepsilon_p^\beta e\left(\frac{2n}{p^\beta}\right)\right]. \tag{9}$$

This can be generalized to sums $S(m,n;p^\beta)$ with $\beta \geq 2$ such that $\gcd(p,2mn) = 1$. It's not difficult to show that $S(m,n;p^\beta) = 0$ unless $m \equiv \ell^2 n \pmod c$. If this congruence holds, we may use property (2) of Proposition 4.9.2 to write

$$S(m,n;p^\beta) = S(\ell n, \ell n; p^\beta)$$

and then apply (9). Thus we obtain the following bound.

Corollary 4.9.5. *Let $\beta \geq 2$ and $\gcd(p,2mn) = 1$. Then*

$$|S(m,n;p^\beta)| \leq 2p^{\beta/2}.$$

Similar arguments may be applied to the case $c = 2^\beta$. Note that we have not addressed the case of a single prime modulus (where the above methods of reparametrizing the sum over d break down). This we may obtain from the famous result of Weil – the Riemann hypothesis for curves over finite fields – from which he was able to obtain the identity

$$S(m,n;p) = \alpha_p + \beta_p, \quad \text{where} \quad \alpha_p = \overline{\beta_p},\ \alpha_p\beta_p = p.$$

Thus $|\alpha_p| = |\beta_p| = p^{1/2}$ amd we have $|S(m,n;p)| \leq 2p^{1/2}$. Putting it all together via the multiplicativity property (4) of Proposition 4.9.2, we have

Theorem 4.9.6. *For any positive integer c and any integers m,n, we have*

$$|S(m,n;c)| \leq \gcd(m,n,c)^{1/2} c^{1/2} \sigma_0(c)$$

where $\sigma_0(c)$ denotes the number of divisors of c.

4.10 The size of Fourier coefficients for general cusp forms

In the last section, we bounded Kloosterman sums according to their modulus c. However, the Fourier coefficients of a Poincaré series are indexed by integers n in $S(m, n; c)$. In this section, we explain how to obtain general bounds for $a(n)$, the n-th Fourier coefficient of a cusp form f of weight k, in terms of n. One approach uses the estimates for Kloosterman sums from the previous section.

The key observation is that if f is a cusp form, then

$$F(z) = y^{k/2}|f(z)|$$

is a bounded function on the upper half-plane. This follows because $F(z)$ is both periodic with respect to Γ, our discrete subgroup, and has exponential decay at every cusp. Thus we may write

$$f(z) \ll y^{-k/2} \quad \text{for any } z \text{ in } \mathcal{H}, \tag{10}$$

where the implied constant depends on f. In fact, the converse is also true. Given an automorphic form f with F as above, if F is bounded in $\mathcal{H}$, then for any σ in $\mathrm{SL}(2, \mathbb{R})$,

$$f(z)|[\sigma^{-1}] = y^{-k/2}F(\sigma^{-1}z) \ll y^{-k/2}$$

which implies that $f(z)|[\sigma^{-1}]$ vanishes as $y \to \infty$, hence f is a cusp form. We record this in the following result.

Proposition 4.10.1. *Let f be an automorphic form of weight k with respect to Γ. Then f is a cusp form if and only if $\Im(z)^{k/2}|f(z)|$ is bounded in the upper-half plane.*

Now we make use of (10) to give a first estimate on the growth of Fourier coefficients.

Proposition 4.10.2. *Let f be a cusp form of weight k having For any $N \geq 1$, we have*

$$\sum_{n \leq N} |a(n)|^2 \ll N^k,$$

with the implied constant depending on f.

Proof. We use Parseval's formula in the real variable x of $z = x + iy$ to obtain

$$\sum_n |a(n)|^2 e^{-4\pi n y/h} = \frac{1}{h}\int_0^h |f(z)|^2 \, dx \ll y^{-k}$$

where we've used (10) in the last inequality. This implies

$$\sum_{n\leq N} |a(n)|^2 \ll y^{-k} e^{4\pi N y/h}$$

for any $y > 0$. Choosing $y = N^{-1}$, we obtain the result. □

Corollary 4.10.3. *Let f be a cusp form of weight k with Fourier coefficients $a(n)$. On average, we have $a(n) \ll n^{(k-1)/2}$. For any individual coefficient, we have*

$$a(n) \ll n^{k/2}.$$

Proof. We use Cauchy's theorem, applied to the estimate of Proposition 4.10.2, to obtain

$$\sum_{n\leq N} |a(n)| \leq \left(\sum_{n\leq N} |a_n|^2 \sum_{n\leq N} 1 \right)^{1/2} \ll N^{(k+1)/2}.$$

Dividing by N leads to the first result. The second result is immediate from Proposition 4.10.2. □

Remark 4.3. In order to obtain the bound $a(n) \ll n^{k/2}$ for an individual coefficient, we didn't need to use Parseval's formula. A simple estimate using the Fourier expansion in the real variable x for $f(z)$ would give the result.

Finally, we show how to improve upon this result by using information about Poincaré series.

Theorem 4.10.4. *Suppose that $S_\Gamma(n/h, m/h; c)$, the Kloosterman sums appearing in the Fourier coefficients of a weight k Poincaré series as in Theorem 4.7.2, satisfy*

$$\sum_{c>0} c^{-2\sigma} |S_\Gamma(n/h, n/h; c)| \ll n^{\varepsilon} \tag{11}$$

for some σ with $\frac{1}{2} \leq \sigma < 1$ and any $\varepsilon > 0$ with the implied constant depending on σ, ε, and Γ. Then the Fourier coefficients of any cusp form $f \in S_k(\Gamma)$ of weight $k > 2$ are bounded by

$$a(n) \ll n^{k/2-1+\sigma+\epsilon}.$$

Proof. Since $k > 2$, any cusp form is a linear combination of Poincaré series $P_n(z)$ with $n \geq 1$. Thus it suffices to estimate the Fourier coefficients $p(m, n)$ of

$$P_n(z) = \sum_{m>0} p(m, n) e(mz/h).$$

Let $\mathcal{F} = \{f_i\}$ be an orthonormal basis for the space of cusp forms $S_k(\Gamma)$. By the inner product formula of Theorem 4.8.3, we may write

$$P_n(z) = \Gamma(k-1)\left(\frac{h^k}{(4\pi n)^{k-1}}\right)\sum_{f_i\in\mathcal{F}}\overline{a_{f_i}(n)}f_i(z).$$

Then expanding in a Fourier series, we have

$$p(m,n) = \Gamma(k-1)\left(\frac{h^k}{(4\pi n)^{k-1}}\right)\sum_{f_i\in\mathcal{F}}\overline{a_{f_i}(n)}a_{f_i}(m).$$

Applying the Cauchy-Schwarz inequality, we see that

$$|p(m,n)|^2 \leq \left(\frac{m}{n}\right)^{k-1} p(m,m)p(n,n). \tag{12}$$

Thus it remains to estimate the "diagonal" coefficients $p(n,n)$. Recall these had Fourier coefficients in the expansion at infinity as in Theorem 4.7.2:

$$p(n,n) = 1 + \frac{2\pi}{i^k h}\sum_{c>0} c^{-1}S_\Gamma(n/h, n/h; c)J_{k-1}\left(\frac{4\pi n}{ch}\right).$$

We claim that

$$J_\nu(x) \ll \min\{x^\nu, x^{-1/2}\} \leq x^\delta \quad \text{if } -\frac{1}{2} \leq \delta \leq \nu.$$

Then in view of our assumption on the growth of Kloosterman sums in (11), set $\delta = 2\sigma - 1$ (which satisfies the above inequality since $\nu = k-1$) to give an estimate for $p(n,n)$ independent of c. Then

$$p(n,n) \ll n^{2\sigma-1+\epsilon}.$$

Applying this in (12) gives our desired bound for $p(m,n)$ and hence for the Fourier coefficient $a(n)$. □

Remark 4.4. It's not hard to show that the Kloosterman sums appearing the Fourier coefficients of cusp forms on $\Gamma_0(N)$ can be related to those for classical Kloosterman sums. Then using the Weil bound $|S(n,n;c)| \leq \gcd(n,c)^{1/2}c^{1/2}\sigma_0(c)$, we have the desired estimate on $\sum_{c>0} c^{-2\sigma}S(n,n,c)$ for any $\sigma > 3/4$. This gives $a(n) \ll n^{k/2-1/4+\epsilon}$.

In Section 5.2 of Iwaniec's "Topics in Classical Automorphic Forms," he shows how clever manipulation of inner product formulas allow one to avoid using the Weil bound.

On the other hand, if we use the full strength of Deligne's proof of the Weil conjectures for varieties over finite fields, then we obtain the bound

$$a(n) \ll n^{k/2-1/2}\sigma_0(n) \ll n^{k/2-1/2+\epsilon} \quad \text{for every } \epsilon > 0.$$

5 L-functions associated to cusp forms

In this section, we describe how to associate a generating function (called an "L-function") to a cusp form. Then we present certain conditions on the L-function which guarantee that it may be associated to a cusp form. These are known as "converse theorems" and we present two famous results of Hecke and Weil in this direction.

The case of the full modular group $\Gamma(1)$ gives a good illustration of this principle. As we noted in Section 3.5, $\Gamma(1)$ is generated by the matrices:

$$S = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \text{ (inversion)}, \quad T = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \text{ (translation)}$$

The invariance of a modular form f under the translation T (i.e., $f(z+1) = f(z)$) guarantees that f is expressible as a Fourier series

$$f(z) = \sum_n a_f(n) e(nz).$$

Hecke showed that the transformation under S (i.e. $f(-1/z) = (-z)^k f(z)$) leads to a functional equation for the Dirichlet series

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

The growth properties of cusp forms guarantee that this function is well-behaved as a function of the complex variable s. (As we will see, L-functions for Eisenstein series also make sense, but will result in meromorphic rather than holomorphic functions.) The key ingredient in Hecke's proof is the Mellin transform.

5.1 The Mellin transform

Given any sequence of complex numbers $\{a_n\}$ such that $|a_n| = O(n^M)$ for some M, we may consider its power series

$$f(q) = \sum_{n=1}^{\infty} a_n q^n \quad \text{which is absolutely convergent for } |q| < 1. \tag{13}$$

Alternately, we may consider its Dirichlet series

$$L(s) = \sum_{n=1}^{\infty} a_n n^{-s} \quad \text{which is absolutely convergent for } \mathrm{Re}(s) > M + 1. \tag{14}$$

These two series expressions are related by the Mellin transform and its inversion formula.

Let ϕ be a continuous function on the open interval $(0, \infty)$. Define the Mellin transform of ϕ by

$$\Phi(s) = \int_0^\infty \phi(y)\, y^s\, \frac{dy}{y}, \quad s \in \mathbb{C}$$

wherever this integral is absolutely convergent. If the integral

$$\int_0^1 \phi(y)\, y^s\, \frac{dy}{y}$$

is absolutely convergent for a particular value of s, then it's absolutely convergent for any s with larger real part. Similarly, if the integral

$$\int_1^\infty \phi(y)\, y^s\, \frac{dy}{y}$$

converges for some s, it converges for all s with smaller real part. Thus we see that there exists $\sigma_1, \sigma_2 \in [-\infty, \infty]$ such that the Mellin transform is absolutely convergent for $\mathrm{Re}(s) \in (\sigma_1, \sigma_2)$ and divergent for $\mathrm{Re}(s) < \sigma_1$ or $\mathrm{Re}(s) > \sigma_2$. (We can't conclude anything in general about the convergence at $\mathrm{Re}(s) = \sigma_1$ or σ_2.)

Proposition 5.1.1 (Mellin Inversion Formula)**.** *Given any $\sigma \in (\sigma_1, \sigma_2)$ and $y \in (0, \infty)$, then*

$$\phi(y) = \frac{1}{2\pi i} \int_{\sigma - i\infty}^{\sigma + i\infty} \Phi(s)\, y^{-s}\, ds.$$

where $\Phi(s)$ is the Mellin transform of ϕ.

Proof. We may prove this using the Fourier inversion formula. Let $\phi_\sigma(v) = \phi(e^v)e^{\sigma v}$. By our assumptions on ϕ, we have ϕ_σ continuous and in $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$. Applying the Mellin transform as above with $y = e^v$, we have

$$\Phi(\sigma + it) = \int_{-\infty}^{\infty} \phi_\sigma(v) e^{ivt}\, dv,$$

so applying the Fourier inversion formula, we obtain

$$\phi_\sigma(v) = \frac{1}{2\pi i} \int_{-\infty}^{\infty} \Phi(\sigma + it)\, e^{-ivt}\, dt.$$

Mutliplying both sides by $e^{-\sigma v}$ and setting $y = e^v$ gives the result. □

As an example, note that the Mellin transform of e^{-x} is the Gamma function

$$\Gamma(s) = \int_0^\infty e^{-x} x^s \, \frac{dx}{x}, \quad (\text{for } \operatorname{Re}(s) > 0).$$

Thus the Mellin inversion formula gives, for every $\sigma > 0$,

$$e^{-x} = \frac{1}{2\pi i} \int_{\sigma - i\infty}^{\sigma + i\infty} \Gamma(s) x^{-s} ds.^3$$

The next result shows that the two power series expansions appearing in (13) and (14) are indeed related by the Mellin transform.

Proposition 5.1.2. *Let $\{a_n\}$ be a sequence of complex numbers with $|a_n| = O(n^M)$ for some M. Then let*

$$f(iy) = \sum_{n=1}^\infty a_n e^{-2\pi n y}, \quad L(s) = \sum_{n=1}^\infty a_n n^{-s}.$$

Then for $\operatorname{Re}(s) > \max\{0, M+1\}$,

$$(2\pi)^{-s} \Gamma(s) L(s) = \int_0^\infty f(iy) y^s \frac{dy}{y}.$$

Thus by Mellin inversion, for $\sigma > \max\{0, M+1\}$ and $y > 0$,

$$f(iy) = \frac{1}{2\pi i} \int_{\sigma - i\infty}^{\sigma + i\infty} \Gamma(s) L(s) \, (2\pi y)^{-s} \, ds.$$

Proof. For $\operatorname{Re}(s) > \max\{0, M+1\}$, the Mellin transform

$$\int_0^\infty f(iy) \, y^s \, \frac{dy}{y}$$

is absolutely convergent and so we may substitute its Fourier expansion. Interchanging order of integration and summation, and noting that

$$\int_0^\infty e^{-2\pi n y} \, y^s \, \frac{dy}{y} = (2\pi)^{-s} \Gamma(s) n^{-s}$$

the first statement follows. The second is an immediate application of the Mellin inversion formula. □

[3] This formula may also be proved by residue calculus moving the path of integration to the left past the poles of the Gamma function at negative integers, giving the power series for e^{-y}.

From this result, we see that invariance properties of a function f with Fourier series may be rephrased in terms of functional equations for L-functions (and vice versa). Furthermore, if f is an automorphic form (or better, a cusp form), then the resulting L-function will be meromorphic with known poles (or better, holomorphic in the case of cusp forms).

We say that a function $\phi(s)$ on the complex plane is *bounded in vertical strips* if, for all pairs of real numbers $a < b$, $\phi(s)$ is bounded on the strip $a \leq \mathrm{Re}(s) \leq b$ as $\Im(s) \to \pm\infty$.

Theorem 5.1.3 (Hecke)**.** *Let $\{a_n\}_{n=0}^{\infty}$ be a sequence of complex numbers with $|a_n| = O(n^M)$ for some M and let h be a fixed positive number. Let*

$$L(s) = \sum_{n=1}^{\infty} a_n n^{-s}, \quad f(z) = \sum_{n \geq 0} a_n e(nz/h).$$

Further define the completed L-series by

$$L^*(s) = \left(\frac{2\pi}{h}\right)^{-s} \Gamma(s) L(s).$$

Then (with $c = \pm 1$) the following two conditions are equivalent:

- *The function $L^*(s) + \frac{a_0}{s} + \frac{ca_0}{k-s}$ may be analytically continued to a holomorphic function on the complex plane, which is bounded in vertical strips, and satisfies the functional equation:*
$$L^*(k-s) = cL^*(s).$$
- *As a function of $z \in \mathcal{H}$, f satisfies the transformation property*
$$f(-1/z) = c(z/i)^k f(z)$$

Proof. Suppose the condition on f holds. Then writing

$$\int_0^{\infty} (f(iy) - a_0)\, y^s \, \frac{dy}{y} = \int_0^{\infty} \sum_{n=1}^{\infty} a_n e^{-2\pi n y/h}\, y^s \, \frac{dy}{y}$$

We have already seen that this integral evaluates to $L^*(s)$ for $\mathrm{Re}(s)$ sufficiently large in the course of the proof of Proposition 5.1.2. On the other hand,

$$\int_0^{\infty} (f(iy) - a_0)\, y^s \, \frac{dy}{y} = \int_1^{\infty} (f(iy) - a_0)\, y^s \, \frac{dy}{y} + \int_0^{1} (f(iy) - a_0)\, y^s \, \frac{dy}{y}$$

This choice is inspired by the fact that the first integral is known to be well-behaved, since $f(iy) - a_0 \to 0$ rapidly as $y \to \infty$. In the second of the two integrals, we set $y \to 1/y$ to obtain

$$\int_1^\infty (f(i/y) - a_0)\, y^{-s}\, \frac{dy}{y} = \int_1^\infty (cf(iy)y^k - a_0)\, y^{-s}\, \frac{dy}{y}$$

using the transformation property for f. Then using a bit of algebra, we have

$$c\int_1^\infty (f(iy) - a_0)\, y^{k-s}\, \frac{dy}{y} - \int_1^\infty a_0\, y^{-s}\, \frac{dy}{y} + \int_1^\infty ca_0\, y^{k-s}\, \frac{dy}{y} =$$

$$c\int_1^\infty (f(iy) - a_0)\, y^{k-s}\, \frac{dy}{y} - \frac{a_0}{s} - \frac{ca_0}{k-s}.$$

Putting it all together, this simultaneously gives the functional equation, the fact that $L^*(s) + \frac{a_0}{s} + \frac{ca_0}{k-s}$ defines a holomorphic function on the entire complex plane, and boundedness in vertical strips. These last two properties again follow from the fact that $f(iy) - a_0 \to 0$ rapidly as $y \to \infty$.

Now suppose L^* satisfies all of the stated conditions. From the condition on $|a_n|$ we see that f defines a holomorphic function for $z \in \mathcal{H}$. Hence it suffices to check that the transformation property in z holds for all $z = iy$ with $y > 0$ (as then the difference of the two sides is a holomorphic function on $\mathcal{H}$ which is 0 along the positive imaginary axis, hence identically 0).

Recall that for $s = \sigma + it$ with σ sufficiently large, we have

$$\int_0^\infty (f(iy) - a_0)y^s \frac{dy}{y} = L^*(s, f)$$

So by Mellin inversion formula

$$f(iy) - a_0 = \frac{1}{2\pi i}\int_{\sigma - i\infty}^{\sigma + i\infty} L^*(s)y^{-s}ds = \frac{c}{2\pi i}\int_{\sigma - i\infty}^{\sigma + i\infty} L^*(k-s)y^{-s}ds,$$

Now $L(\sigma + it)$ is of rapid decay as $t \to \infty$ for σ sufficiently large. Indeed, it is the product of an absolutely convergent Dirichlet series together with a Gamma function that decays by Stirling's formula. Recall Stirling's formula gives $|\Gamma(\sigma+it)| \sim \sqrt{2\pi}|t|^{\sigma-1/2}e^{-\pi|t|/2}$ as $t \to \infty$.

Furthermore, if σ is sufficiently smaller than 0, we may use the functional equation for $L^*(s) = cL^*(k-s)$ to conclude that $L^*(s)$ is again of rapid decay as $t \to \infty$. We are assuming that $L^*(s)$ is bounded in vertical strips, which allows us to apply the Phragmén-Lindelöf principle to the values of σ in between.

Lemma 5.1.4 (Phragmén-Lindelöf)**.** *Let $\phi(s)$ be a holomorphic function on the upper part of a vertical strip defined by:*

$$\sigma_1 \leq \mathrm{Re}(s) \leq \sigma_2, \quad \Im(s) > c$$

and such that $\phi(\sigma + it) = O(e^{t^\alpha})$ for some $\alpha > 0$ when $\sigma \in [\sigma_1, \sigma_2]$. If $\phi(\sigma_1 + it) = O(t^M)$ and $\phi(\sigma_2 + it) = O(t^M)$ then $\phi(\sigma + it) = O(t^M)$ for all $\sigma \in [\sigma_1, \sigma_2]$.

For a proof of this result, see Lang's "Algebraic Number Theory," Section XIII.5.

Using the Phragmén-Lindelöf principle, it follows that $L^*(\sigma + it) \to 0$ as $t \to \infty$ uniformly for σ in any compact set. This allows us to move the line of integration leftward by applying Cauchy's theorem (from σ to $k - \sigma$).

$$f(iy) - a_0 = \frac{c}{2\pi i} \int_{k-\sigma-i\infty}^{k-\sigma+i\infty} L^*(k-s) y^{-s} ds + c(a_0 y^{-k} - c a_0)$$

after picking up poles at $s = 0$ and $s = k$. Then making the change of variables $s \mapsto k - s$ gives

$$\frac{c y^{-k}}{2\pi i} \int_{\sigma-i\infty}^{\sigma+i\infty} L^*(s) y^s ds + c(a_0 y^{-k} - c a_0) = c y^{-k}(f(i/y) - a_0) + c a_0 y^{-k} - a_0,$$

which after cancellation, gives the desired result. □

Remark 5.1. In terms of linear fractional transformations, the existence of a Fourier expansion for f guarantees that it is invariant with respect to the translation $T_h : z \mapsto z + h$. The relation $f(-1/z) = c(z/i)^k f(z)$ is a slightly more general version of the action by the inversion $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$. So our theorem may be rephrased as giving an equivalence between L-functions and modular forms of weight k for the group $\Gamma = \langle T_h, S \rangle$. For example, if $h = 1$, $\Gamma = \Gamma(1)$. As a second example, the Riemann zeta function arises in this way as the L-function associated to the Jacobi theta function, which may be viewed as a modular form of weight $1/2$ with respect to $\langle T_2, S \rangle$.

The condition on the Fourier coefficients $|a(n)| = O(n^M)$ for some M follows for cusp forms from our estimates from the previous section, and for Eisenstein series from our explicit description in terms of the divisor function.

We have seen that a single functional equation for an L-function gives the transformation law of its associated function f at a single matrix - an inversion like the matrix S. But discrete groups other than $\Gamma(1)$ will have a longer, more complicated list of generators (and relations). So we need additional conditions on the L-function in order to guarantee that f is a modular form for other choices of Γ.

5.2 Weil's converse theorem

Weil was able to obtain conditions for f to be a modular form on $\Gamma_0(N)$ by requiring functional equations for L-functions twisted by certain primitive characters modulo N. Before discussing his proof, we mention some particulars of these so-called "twisted L-functions."

Throughout this section, we will need to act by matrices in $\mathrm{GL}(2,\mathbb{R})^+$, where the "+" indicates the matrices have positive determinant. For these purposes it is convenient to extend the slash action of weight k as follows:

$$f(z)|[\gamma]_k = (\det\gamma)^{k/2}(cz+d)^{-k} f\left(\frac{az+b}{cz+d}\right).$$

Given a positive integer N, let ψ be a Dirichlet character modulo N. Then denote by $S_k(\Gamma_0(N),\psi)$ the space of cusp forms on $\Gamma_0(N)$, but with the slightly more general transformation property

$$f(z)|[\gamma]_k = \psi(d)f(z) \quad \text{where } \gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \Gamma_0(N).$$

Note that the parity of the weight k determines the value of $\psi(-1)$.

The matrix

$$w_N = \begin{pmatrix} 0 & -1 \\ N & 0 \end{pmatrix}$$

normalizes $\Gamma_0(N)$:

$$w_N \begin{pmatrix} a & b \\ c & d \end{pmatrix} w_N^{-1} = \begin{pmatrix} d & -c/N \\ -Nb & a \end{pmatrix}$$

Hence given any $f \in S_k(\Gamma_0(N),\psi)$ and $\gamma \in \Gamma_0(N)$, we have

$$f|[w_N]_k|[\gamma]_k = f|[w_N\gamma w_N^{-1}]_k|[w_N]_k = \overline{\psi(d)}f|[w_N]_k$$

and thus $g := f|[w_N]_k \in S_k(\Gamma_0(N),\overline{\psi})$.

Proposition 5.2.1. *Let N and D be coprime integers and χ_D a primitive character modulo D. Let $f \in S_k(\Gamma_0(N),\psi)$ and $g = f|[w_N]_k$. If $a(n)$ are Fourier coefficients of f and $b(n)$ are Fourier coefficients of g, define the associated L-functions:*

$$L(s,f,\chi_D) = \sum_{n=1}^{\infty} \chi_D(n)a(n)n^{-s}, \quad L(s,g,\overline{\chi}_D) = \sum_{n=1}^{\infty} \overline{\chi}_D(n)b(n)n^{-s}.$$

Define the completed L-function for f as before with $L^*(s,f,\chi_D) = (2\pi)^{-s}\Gamma(s)L(s,f,\chi_D)$ and similarly for g. Then the completed L-functions satisfy the functional equation:

$$L^*(s,f,\chi_D) = i^k\chi_D(N)\psi(D)\frac{\tau(\chi_D)^2}{D}(D^2N)^{-s+k/2}L^*(k-s,g,\overline{\chi}_D),$$

where $\tau(\chi_D)$ is the Gauss sum

$$\tau(\chi_D) = \sum_{a \ (mod\ D)} \chi_D(a)e(a/D).$$

Proof. Define the power series expansions

$$f_\chi(z) = \sum_{n=1}^{\infty}\chi(n)a(n)e(nz) \quad g_{\overline{\chi}}(z) = \sum_{n=1}^{\infty}\overline{\chi}(n)b(n)e(nz).$$

We'd like to mimic the proof of Theorem 5.1.3 for the "twisted" Fourier expansions f_χ and $g_{\overline{\chi}}$. We derive a relation between them from that of f and g by first relating f to f_χ.

To this end, recall the elementary property of Gauss sums that

$$\sum_{n \ (\mathrm{mod}\ D)} \chi_D(n)e(nm/D) = \tau(\chi_D)\overline{\chi_D(m)}$$

which implies that (upon replacing χ by $\overline{\chi}$ and noting $|\tau(\chi)|^2 = D$)

$$\chi_D(n) = \frac{\chi_D(-1)\tau(\chi_D)}{D}\sum_{m \bmod D}\overline{\chi_D(m)}e(nm/D).$$

This formula gives an interpolation of the character χ since the right-hand side makes sense for all real numbers n. For us, multiplying both sides by $a_f(n)e(nz)$ and summing over all integers n, we obtain

$$f_{\chi_D}(z) = \frac{\chi_D(-1)\tau(\chi_D)}{D}\sum_{m \ (\mathrm{mod}^\times D)}\overline{\chi_D(m)}f\left|\left[\begin{pmatrix} D & m \\ 0 & D\end{pmatrix}\right]_k\right.$$

To obtain a relation between f_{χ_D} and $g_{\overline{\chi_D}}$, we examine

$$\begin{aligned} f_{\chi_D}\left|\left[\begin{pmatrix} 0 & -1 \\ D^2N & 0\end{pmatrix}\right]_k\right. &= f_{\chi_D}\left|\left[\begin{pmatrix} 0 & -1/ND \\ D & 0\end{pmatrix}\right]_k\right. \\ &= \frac{\chi_D(-1)\tau(\chi_D)}{D}\sum_{m \ (\mathrm{mod}^\times D)}\overline{\chi_D(m)}g\left|\left[\begin{pmatrix} 0 & 1 \\ -N & 0\end{pmatrix}\begin{pmatrix} D & m \\ 0 & D\end{pmatrix}\begin{pmatrix} 0 & -1/ND \\ D & 0\end{pmatrix}\right]_k\right. \\ &= \frac{\chi_D(-1)\tau(\chi_D)}{D}\sum_{m \ (\mathrm{mod}^\times D)}\overline{\chi_D(m)}g\left|\left[\begin{pmatrix} D & -r \\ -Nm & s\end{pmatrix}\begin{pmatrix} D & r \\ 0 & D\end{pmatrix}\right]_k\right. \end{aligned}$$

where $r = r(m)$ and $s = s(m)$ are chosen so that $Ds - rNm = 1$. This implies $\overline{\chi_D(m)} = \chi_D(-N)\chi_D(r)$, so we may reparametrize the sum to give

$$f_{\chi_D}\left|\left[\begin{pmatrix} 0 & -1 \\ D^2N & 0 \end{pmatrix}\right]_k = \frac{\chi_D(N)\tau(\chi_D)}{D}\sum_{r \pmod{D}^\times} \chi_D(r) g\left|\left[\begin{pmatrix} D & -r \\ -Nm & s \end{pmatrix}\begin{pmatrix} D & r \\ 0 & D \end{pmatrix}\right]_k. \tag{15}$$

But

$$g\left|\left[\begin{pmatrix} D & -r \\ -Nm & s \end{pmatrix}\right]_k = \psi(D) g \tag{16}$$

since $g \in S_k(\Gamma_0(N), \overline{\psi})$. Finally, using the same identity as with f_χ and f,

$$g_{\overline{\chi_D}} = \frac{\chi_D(-1)\tau(\overline{\chi}_D)}{D}\sum_{r \pmod{D}^\times} \chi_D(r) g\left|\left[\begin{pmatrix} D & r \\ 0 & D \end{pmatrix}\right]_k \tag{17}$$

Putting the previous three equations (15), (16), and (17) together, we obtain

$$f_{\chi_D}\left|\left[\begin{pmatrix} 0 & -1 \\ D^2N & 0 \end{pmatrix}\right]_k = \chi_D(N)\psi(D)\frac{\tau(\chi_D)^2}{D} g_{\overline{\chi_D}}. \tag{18}$$

We may then apply a Mellin transform to f_χ and manipulate the result as in Theorem 5.1.3 to conclude the relation on L-functions. □

Hecke realized that the reasoning here might be reversible. The rough idea is that functional equations for twisted L-functions are essentially equivalent to the identity (18) via Mellin inversion. Together with the more elementary properties of (15) and (17), these should imply (16). However it was left to Weil to provide the definitive converse theorem, which required only functional equations for L-functions twisted by primitive characters χ_D (as opposed to all characters mod D).

Theorem 5.2.2 (Weil)**.** *Let N be a positive integer and ψ a Dirichlet character modulo N. Let $\{a(n)\}$ and $\{b(n)\}$ be sequences of complex numbers such that $|a(n)|, |b(n)| = O(n^M)$ for some M. For D such that $\gcd(D, N) = 1$, let χ_D be a primitive character modulo D. Define*

$$L_1(s, \chi_D) = \sum \chi(n)a(n)n^{-s} \quad L_2(s, \overline{\chi}_D) = \sum \overline{\chi}(n)b(n)n^{-s}$$

and the associated completed L-functions

$$L_1^*(s, \chi_D) = (2\pi)^{-s}\Gamma(s)L_1(s, \chi_D) \quad L_2^*(s, \overline{\chi}_D) = (2\pi)^{-s}\Gamma(s)L_1(s, \overline{\chi}_D).$$

Suppose that for all but finitely many primes p not dividing N the following conditions hold for every primitive character χ_p of conductor p:

1. *$L_1^*(s,\chi_p)$ and $L_2^*(s,\overline{\chi}_p)$ have analytic continuation to holomorphic functions for all $s \in \mathbb{C}$.*

2. *$L_1^*(s,\chi_p)$ and $L_2^*(s,\overline{\chi}_p)$ are bounded in vertical strips*

3. *We have the functional equation*

$$L_1^*(s,\chi_p) = i^k \chi_p(N)\psi(D)\frac{\tau(\chi)^2}{D}(D^2N)^{-s+k/2}L_2^*(s,\overline{\chi}_p).$$

Then $f(z) = \sum_n a(n)e(nz)$ is a modular form in $M_k(\Gamma_0(N),\psi)$.

Proof. For a complete proof, see Section 1.5 of Bump's "Automorphic Forms and Representations." We only give a very brief sketch. As we noted above, Mellin inversion and the functional equations give the identity (18) along the same lines as in Hecke's theorem. The identities (17) and (15) are elementary and in particular don't require automorphicity of f and g. Thus combining these three equations, we get an identity between $g\left|\begin{pmatrix} D & r \\ Nm & s \end{pmatrix}\right.$ and $\psi(D)g$ with both sides twisted by a primitive character χ_D. Here D and s may be arbitrary primes outside of a finite set S, according to our assumptions in the theorem.

Now the brilliant idea is that we may make use of the following result: given a holomorphic function f on $\mathcal{H}$, then if $f|[\gamma]_k = f$ for some elliptic element $\gamma \in \mathrm{SL}(2,\mathbb{R})$ of infinite order, then $f \equiv 0$. Then if we set

$$g' = g\left|\begin{pmatrix} D & r \\ Nm & s \end{pmatrix}\right. - \psi(D)g$$

and find an elliptic matrix γ as above such that $g'|[\gamma]_k = g'$, then we are almost done. This is precisely what is accomplished with some clever matrix manipulation of the twisted identity for g.

To argue that this identity leads to invariance of g slashed by a generic matrix of $\Gamma_0(N)$ requires one last step. In the above, we emphasize that D and s may only be taken to be arbitrary primes (outside of a finite set). Given a generic matrix in $\Gamma_0(N)$, we may act on the left and right by upper triangular matrices (which leaves g invariant according to the Fourier expansion). This shifts the diagonal entries of our generic matrix, each of which are relatively prime to N. Dirichlet's theorem on primes in an arithmetic progression guarantees that some such shift of each diagonal element is prime. □

6 Hecke Operators

6.1 Initial Motivation

Recall that the space of weight 12 cusp forms for $\Gamma(1)$ is one dimensional, and has generator $\Delta = g_2^3 - 27g_3^2$ where $g_2 = 60G_4$ and $g_3 = 140G_6$. Jacobi showed (as a consequence of his famous triple product formula) that in fact $\Delta(z)$ expressed as a q-series where $q = e(z)$ is of the form

$$\Delta(z) = (2\pi)^{12} q \cdot \prod_{n=1}^{\infty} (1-q^n)^{24}, \quad q = e(z)$$

If instead we express this as a Fourier series of the form $f(z) = \sum_n \tau(n) e(nz)$, then we may ask about properties of the arithmetic function $\tau(n)$. Ramanujan conjectured that

$$\begin{aligned} \tau(mn) &= \tau(m)\tau(n) && \text{if } \gcd(m,n) = 1, \\ \tau(p^{n+1}) &= \tau(p)\tau(p^n) - p^{11}\tau(p^{n-1}) && \text{if } p \text{ is prime and } n \geq 1. \end{aligned} \tag{19}$$

The fact that τ is multiplicative seems quite remarkable. These two properties were first proved by Mordell (1917) just a year after Ramanujan's conjecture in which he introduced the first form of Hecke operators. These were later codified by Hecke in a definitive paper from 1937. Before explaining this in detail, we note a consequence for the L-series associated to Δ.

Let $L(\Delta, s)$ denote the L-function attached to the cusp form Δ. That is,

$$L(\Delta, s) = \sum_{n=1}^{\infty} \tau(n) n^{-s}.$$

Proposition 6.1.1. *The properties of (19) are equivalent to the identity*

$$L(\Delta, s) = \prod_{p:\ prime} (1 - \tau(p)p^{-s} + p^{11-2s})^{-1}.$$

Note that the Riemann zeta function had a similar product expansion, and any such identity of a Dirichlet series given as an infinite product over primes is referred to as an Euler product. Issues of convergence are usually transferred to equivalent conditions about infinite sums using the logarithm. For a brief introduction to infinite products, see Section 5.2.2 of Ahlfors' "Complex Analysis." The above expression for $L(\Delta, s)$ is valid for $\mathrm{Re}(s) > 1$.

Proof. From the multiplicativity of τ, we have

$$L(\Delta, s) = \sum_{n=1}^{\infty} \tau(n) n^{-s} = \prod_{p \text{: prime}} \left(\sum_{r=0}^{\infty} \tau(p^r) p^{-rs} \right).$$

To evaluate this latter series note that the recursion in powers of p implies that

$$(1 - \tau(p)X + p^{11}X^2) \left(\sum_{r=0}^{\infty} \tau(p^r) X^r \right) = 1.$$

Taking $X = p^{-s}$ in the above gives the form of the Euler product. □

Why should the Fourier coefficients of an automorphic form be multiplicative? The insight is that we may regard the Fourier coefficients $a(n)$ as eigenvalues of a certain family of operators $T(n)$ defined for each positive integer n. If this operator is self-adjoint with respect to the Petersson inner product on $S_k(\Gamma)$ and some subset of the $T(n)$ commute, then we may apply the (finite dimensional version of the) spectral theorem.

Theorem 6.1.2. *Given a finite dimensional complex vector space V with positive definite Hermitian form $\langle \cdot, \cdot \rangle$, then if $\{T_i\}$ is a commuting family of self-adjoint operators, then V has a basis of simultaneous eigenvectors for all T_i.*

Proof. Since $\mathbb{C}$ is algebraically closed, we are guaranteed an eigenvector e_1 for T_1. Let $V_1 = (\mathbb{C}e_1)^{\perp}$. Since T_1 is self-adjoint, then V_1 is stable under T_1 and so has an eigenvector. Repeating this process we obtain an eigenbasis for T_1. Now since T_2 commutes with T_1 it preserves each T_1-eigenspace. Each such eigenspace can be further decomposed into an eigenspace of T_2's. Repeating this process, we arrive at a decomposition of V into subspaces for which each T_i acts by a scalar. Choose a basis for each subspace and take the union. □

We will show that there exist operators

$$T(n) : M_k(\Gamma(1)) \to M_k(\Gamma(1))$$

which preserve the space of cusp forms and moreover satisfy

$$T(m) \circ T(n) = T(mn) \quad \text{if } \gcd(m, n) = 1$$

and hence in particular any set $\{T(n)\}$ for which the n are pairwise coprime will give a commuting family. Note that since $S_{12}(\Gamma(1))$ is one-dimensional, then Δ must

be such a simultaneous eigenfunction (provided such operators exist). In the next section, we explain their definition.

Finally, we mention briefly some motivation for expressing the L-function of an automorphic form as an Euler product. Generally speaking, we use the term "L-function" for any Dirichlet series having a certain list of rare analytic properties (often including analytic continuation to a meromorphic function, a functional equation, and an Euler product). There exist other means of defining L-functions attached to algebraic varieties. These are defined by specifying their Euler factors at each prime p and then defining the L-function as the resulting product.

For example, given an elliptic curve E defined over $\mathbb{Q}$, we define

$$L(E,s) = \prod_{\substack{p\text{: prime of} \\ \text{good red.}}} (1 - a(p)p^{-s} + p^{1-2s})^{-1}$$

where

$$a(p) = p + 1 - |E(\mathbb{F}_p)|, \quad |E(\mathbb{F}_p)| = \#\text{ of points on } E \text{ over } \mathbb{F}_p.$$

We'll explain later in the course where the definition of the Euler factor comes from and why $a(p)$ is a natural choice. For now, we merely note the resemblance of this Euler factor to the one presented in Proposition 6.1.1. In fact, we will show that weight 2 cusp forms for $\Gamma_0(N)$ (which are simultaneous eigenfunctions of the Hecke operators $T(p)$) have Euler factors of precisely the same form, now with $a(p)$ equal to the p-th Fourier coefficient. We may now give a restatement of the Taniyama-Shimura-Weil theorem.

Theorem 6.1.3 (Wiles, Taylor-Wiles, BCDT)**.** *Given an elliptic curve E over $\mathbb{Q}$ of conductor N, there exists a weight 2 Hecke eigen-cuspform f on $\Gamma_0(N)$ such that*

$$L(E,s) = L(f,s).$$

We should explain why this is equivalent to the form stated in Section 3.10. But we also postpone this until after completing the discussion of Hecke operators. This theorem is now understood as being part of a much larger framework, the Langlands program, in which all such L-functions attached to algebraic varieties (or more generally motives) are conjectured to be associated to L-functions coming from automorphic forms.

6.2 Hecke operators on lattices

Let $\mathcal{L}$ be the set of full lattices in $\mathbb{C}$. Earlier, we identified homogeneous functions on the set of lattices with functions on the upper half plane transforming by the usual

factor of automorphy. In this section, we define Hecke operators on lattices and then use this identification with modular functions to transfer the definition to $M_k(\Gamma(1)$.

Let $\mathcal{K}$ be the free abelian group generated by elements of $\mathcal{L}$. Thus the elements of this group are $\sum n_i[\Lambda_i]$ for $n_i \in \mathbb{Z}$ and $\Lambda_i \in \mathcal{L}$. Then define the operator $T(n)$ by

$$T(n)[\Lambda] = \sum_{[\Lambda:\Lambda']=n} [\Lambda']$$

where the sum is taken over all sublattices Λ' having index n. Recall that the index of the lattice is just the ratio of its covolumes or, equivalently, the number of lattice points of Λ in a fundamental parallelogram for Λ'. (Then we extend $T(n)$ linearly to all of $\mathcal{K}$.) Note that the sum above is finite because any such sublattice contains $n\Lambda$ and the quotient $\Lambda/n\Lambda$ is finite. Finally, let $R(n)$ be the linear operator on $\mathcal{K}$ determined by

$$R(n)[\Lambda] = [n\Lambda].$$

Proposition 6.2.1. *If m and n are relatively prime, then*

$$T(m) \circ T(n) = T(mn).$$

Further, if p is a prime and $r \geq 1$ then

$$T(p^r) \circ T(p) = T(p^{r+1}) + pR(p) \circ T(p^{r-1}).$$

Proof. We have $T(mn)[\lambda]$ equal to the sum over all sublattices of index mn, while $T(m) \circ T(n)[\Lambda]$ is the sum indexed by pairs (Λ', Λ'') where $[\Lambda : \Lambda'] = n$ and $[\Lambda' : \Lambda''] = m$. But since $\gcd(m, n) = 1$, there is a unique Λ' fitting in the chain

$$\Lambda \supset \Lambda' \supset \Lambda''$$

since $\Lambda/mn\Lambda$ is a direct sum of a group of order n^2 and one of order m^2. This proves the first assertion.

Now with p a prime, $T(p^r) \circ T(p)[\Lambda]$ is a sum of lattices Λ'' of index p^{r+1} indexed by pairs (Λ', Λ'') where $[\Lambda : \Lambda'] = p$ and $[\Lambda' : \Lambda''] = p^r$. While

$$pR(p) \circ T(p^{r-1})[\Lambda] = p \sum_{[\Lambda:\Lambda']=p^{r-1}} R(p)[\Lambda'] = p \sum_{[p\Lambda:\Lambda'']=p^{r-1}} [\Lambda'']$$

where the later sum ranges over all $\Lambda'' \subset p\Lambda$ with $[p\Lambda : \Lambda''] = p^{r-1}$. Note that each such Λ'' is a sublattice of index p^{r+1} in Λ, each of which appears exactly once in the sum for $T(p^{r+1})[\Lambda]$. If we let a be the number of times Λ'' appears in the sum for

$T(p^r) \circ T(p)[\Lambda]$ and b be the number of times it appears in $R(p) \circ T(p^{r-1})[\Lambda]$, then we must show $a = 1 + pb$.

Suppose that $\Lambda'' \not\subset p\Lambda$ (i.e., $b = 0$). Note that a is the number of lattices Λ' containing Λ'' and of index p in Λ. A simple counting argument shows that $a = 1$. Indeed, such a Λ' contains $p\Lambda$ and projects to an element of order p in $\Lambda/p\Lambda$ containing the image of Λ''. Since $\Lambda'' \not\subset p\Lambda$, its image is an element of order p so must equal the image of Λ'. This determines Λ' uniquely according to the one-to-one correspondence between sublattices of index p in Λ and subgroups of index p in $\Lambda/p\Lambda$.

If instead $\Lambda'' \subset p\Lambda$ then $b = 1$. Any lattice Λ' of index p contains $p\Lambda \supset \Lambda''$. So we merely need to count the number of subgroups of $\Lambda/p\Lambda$ of index p. This is $p + 1$ (the number of lines through the origin in $\mathbb{F}_p$ which is $(p^2 - 1)/(p - 1)$). □

Corollary 6.2.2. *Let $\mathcal{H}(\Gamma(1))$ be the $\mathbb{Z}$-subalgebra generated by the operators $T(p)$ and $R(p)$ for all primes p. Then $\mathcal{H}(\Gamma(1))$ is commutative and contains $T(n)$ for all n.*

We may denote $\mathcal{H}(\Gamma(1))$ simply by $\mathcal{H}$ when no confusion may arise about the group in question or the similar notation used for the upper half plane.

Corollary 6.2.3. *For any integers m, n, we have the identity of linear operators on $\mathcal{K}$:*

$$T(m) \circ T(n) = \sum_{0<d|\gcd(m,n)} d \cdot R(d) \circ T(mn/d^2).$$

Remark 6.1. Note the similarity between this relation and the third property of Kloosterman sums appearing in Proposition 4.9.2 due to Selberg.

Proof. Using Proposition 6.2.1, we may reduce this to a statement about prime powers:

$$T(p^r) \circ T(p^s) = \sum_{i \le \min\{r,s\}} p^i \cdot R(p^i) \circ T(p^{r+s-2i}),$$

which follows easily from the previous proposition by induction. □

Given a function F on the set of lattices $\mathcal{L}$, we may extend linearly to obtain a function on $\mathcal{K}$, the free abelian group generated by lattices. That is,

$$F\left(\sum_i n_i[\Lambda_i]\right) = \sum_i n_i F(\Lambda_i).$$

Then we may define $(T(n) \cdot F)$ by on $\mathcal{L}$ by the formula

$$(T(n) \cdot F)([\Lambda]) = F(T(n)[\Lambda]) = \sum_{[\Lambda:\Lambda']=n} F([\Lambda'])$$

Note that if F is homogeneous of weight k, then by definition $F(n\Lambda) = n^{-k}F(\Lambda)$. In terms of our previously defined operator $R(n)$ we see that this assumption implies

$$R(n) \cdot F = n^{-k}F.$$

Proposition 6.2.4. *Given a function $F : \mathcal{L} \to \mathbb{C}$ that is homogeneous of weight k, then $T(n) \cdot F$ is again of weight k. Moreover, for any positive integers m and n,*

$$T(m) \cdot (T(n) \cdot F) = \sum_{0<d|\gcd(m,n)} d^{1-k}T(mn/d^2) \cdot F.$$

Proof. This follow immediately from the fact the definitions of the operator $T(n) \cdot F$ and Corollary 6.2.3. □

Recall further that there is a one-to-one correspondence between functions F on $\mathcal{L}$ of weight k and functions f on the upper half plane $\mathcal{H}$ that are weakly modular of weight k according to

$$F(\Lambda(\omega_1, \omega_2)) = \omega_2^{-k} f(\omega_1/\omega_2) \quad \text{and} \quad f(z) = F(\Lambda(z, 1))$$

Thus we may define the action of the "Hecke operator" $T(n)$ on the function $f(z)$ by

$$T(n) \cdot f(z) = n^{k-1}(T(n) \cdot F)(\Lambda(z, 1)).$$

The factor n^{k-1} has been included to make some formulas a bit nicer (in particular, making coefficients integral).

Proposition 6.2.5. *If f is a weakly modular form of weight k, then $T(n) \cdot f$ is also a weakly modular form of weight k. Further*

1. *$T(m) \cdot T(n) \cdot f = T(mn) \cdot f$ if $\gcd(m, n) = 1$.*
2. *$T(p) \cdot T(p^r) \cdot f = T(p^{r+1}) \cdot f + p^{k-1}T(p^{r-1}) \cdot f$ for p prime and $r \geq 1$.*

Proof. Both properties are immediate from Proposition 6.2.4, taking into account the normalization appearing in $T(n) \cdot f(z) = n^{k-1}(T(n) \cdot F)(\Lambda(z, 1))$. □

6.3 Explicit formulas for Hecke operators

We would like to give a more explicit description of the action of Hecke operators $T(n)$ for $\Gamma(1)$, by describing the sublattices of index n.

Let $M(n)$ denote the set of 2×2 matrices with integer coefficients and determinant n. Given any lattice Λ in $\mathbb{C}$, we choose a basis ω_1, ω_2 for Λ. Then for any matrix $M \in M(n)$ define

$$M \cdot \Lambda = \Lambda(a\omega_1 + b\omega_2, c\omega_1 + d\omega_2), \quad M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Then $M \cdot \Lambda$ is a sublattice of Λ of index n, and every such lattice is of this form for some $M \in M(n)$. Note further that $M\Lambda = M'\Lambda$ if and only if $M' = UM$ for some $U \in \mathrm{SL}(2, \mathbb{Z})$.

Thus we require a decomposition of $M(n)$ into $\mathrm{SL}(2, \mathbb{Z})$ cosets. This is provided by the following short lemma.

Lemma 6.3.1. *Given a 2×2 matrix A with integer coefficients and determinant n, there is an invertible matrix in $\mathrm{Mat}_2(\mathbb{Z})$ such that*

$$U \cdot A = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \quad ad = n, a \geq 1, 0 \leq b < d,$$

where the integers a, b, d are uniquely determined.

Proof. We leave the proof of this fact to the reader, as it is an exercise in applying invertible row operations to get A to an upper triangular matrix. The uniqueness follows since a is seen to be the g.c.d. of the first column of A, d is then determined by $ad = n$ and b is uniquely determined modulo d. □

Thus we have

$$M(n) = \bigcup \mathrm{SL}(2, \mathbb{Z}) \cdot \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$$

where the union is disjoint and the set of a, b, and d are as in the lemma.

In short, given a lattice $\Lambda = \Lambda(\omega_1, \omega_2)$ the sublattices of index n are precisely those of the form

$$\Lambda(a\omega_1 + b\omega_2, d\omega_2), \quad a, b, d \in \mathbb{Z} : \ ad = n, a \geq 1, 0 \leq b < d.$$

This finally allows us to write $T(n) \cdot f(z)$ more explicitly in the form

$$T(n) \cdot f(z) = n^{k-1} \sum_{\substack{a,b,d:\ ad=n \\ a\geq 1,\ 0\leq b<d}} d^{-k} f\left(\frac{az+b}{d}\right). \tag{20}$$

With this, we can now give formulas for the Fourier coefficients of $T(n) \cdot f(z)$.

Proposition 6.3.2. *Let f be a modular form of weight k for $\Gamma(1)$, with Fourier expansion*

$$f(z) = \sum_{m \geq 0} c_f(m) e(mz).$$

Then $T(n) \cdot f(z)$ is also a modular form and has Fourier expansion

$$T(n) \cdot f(z) = \sum_{m \geq 0} b(m) e(mz) \quad \text{where} \quad b(m) = \sum_{a \mid \gcd(m,n), a \geq 1} a^{k-1} c_f\left(\frac{mn}{a^2}\right).$$

In particular $T(n)$ preserves the space of cusp forms.

Proof. From (20), we see that $T(n) \cdot f$ is holomorphic on the upper half-plane, since f is holomorphic. Expanding (20) using the Fourier series for f we have

$$T(n) \cdot f(z) = n^{k-1} \sum_{\substack{a,b,d:\ ad=n \\ a \geq 1,\ 0 \leq b < d}} d^{-k} \sum_{m \geq 0} c_f(m) e\left(\frac{m(az+b)}{d}\right).$$

Interchanging the order of summation and summing over b, we make use of the identity

$$\sum_{0 \leq b < d} e\left(\frac{mb}{d}\right) = \begin{cases} d & \text{if } d \mid m. \\ 0 & \text{otherwise.} \end{cases}$$

So we may assume $d|m$ and reparametrize the sum over integers $m' = m/d$:

$$T(n) \cdot f(z) = n^{k-1} \sum_{\substack{a,d,m':\ ad=n \\ m' \geq 0,\ a \geq 1}} d^{1-k} c_f(dm') e\left(\frac{m'az}{d}\right).$$

Now letting $m = am'$ and replacing d by n/a, we obtain the m-th Fourier coefficient $b(m)$ for $T(n) \cdot f$ of form

$$b(m) = \sum_{a \mid \gcd(n,m), a \geq 1} a^{k-1} c_f\left(\frac{nm}{a^2}\right),$$

for $m \geq 0$. If $m < 0$, then $b(m) = 0$ since the coefficients of f vanish at negative integers as well. Finally, the last identity for $b(m)$ shows that $T(n) \cdot f$ preserves cusp forms. □

This explicit formula leads immediately to the following result.

Corollary 6.3.3. *The Fourier coefficients $b(m)$ of $T(n) \cdot f(z)$ satisfy:*

1. $b(0) = \sigma_{k-1}(n)c_f(0)$;
2. $b(1) = c_f(n)$;
3. *In the special case $n = p$,*

$$b(m) = \begin{cases} c_f(pm) & p \nmid m \\ c_f(pm) + p^{k-1}c_f(m/p) & p|m. \end{cases}$$

Proposition 6.3.4. *Let $f = \sum c_f(m)e(mz)$ be a non-zero modular form of weight k for $\Gamma(1)$ that is a simultaneous eigenform for all the $T(n)$:*

$$T(n) \cdot f = \lambda(n) \cdot f, \quad \lambda(n) \in \mathbb{C}.$$

Then $c_f(1) \neq 0$ and if we normalize f so that $c_f(1) = 1$, then

$$c_f(n) = \lambda(n) \quad \text{for all } n \geq 1.$$

Proof. Item 2 of Corollary 6.3.3 shows that $T(n) \cdot f$ has $b(1) = c_f(n)$. Since f is an eigenform, we also have $b(1) = \lambda(n)c_f(1)$, and hence $c_f(n) = \lambda(n)c_f(1)$. If $c_f(1) = 0$, then the relation implies that f is identically 0 since all its Fourier coefficients vanish. Hence, normalizing gives the result. □

Corollary 6.3.5. *Let $f = \sum_n c_f(n)e(nz)$ be a normalized eigenform of weight k. Then*

1. $c_f(m)c_f(n) = c_f(mn)$ *if* $\gcd(m, n) = 1$,
2. $c_f(p)c_f(p^r) = c_f(p^{r+1}) + p^{k-1}c(p^{r-1})$ *if p is prime and $r \geq 1$.*

Proof. Since the eigenvalues satisfy these relations, the claim follows. □

Finally, returning to the Dirichlet series associated to the normalized eigenform f, we may express $L(s, f)$ as an Euler product using the above relations, just as in the proof for $L(s, \Delta)$. We record this result in the following proposition, remembering that we should take $\mathrm{Re}(s) \gg 0$ (depending on our control of the growth of $|c_f(n)|$ as a function of n) in order to guarantee convergence of the series and hence, the resulting Euler product.

Proposition 6.3.6. *For any normalized eigenform f on $\Gamma(1)$,*

$$L(s,f) = \prod_p (1 - c_f(p)p^{-s} + p^{k-1-2s})^{-1} \quad \textit{for } \mathrm{Re}(s) \gg 0.$$

All of the results in the previous section require the existence of a Hecke eigenform. In order to apply the spectral theorem, which *guarantees* a basis of Hecke eigenforms for $S_k(\Gamma(1))$, we must demonstrate that the operators $T(n)$ are Hermitian. We do this in the next section after giving an alternate definition of the Hecke operators.

6.4 A geometric definition of Hecke operators

In the last section, we presented a definition of Hecke operators as a sum over sublattices Λ of index n. Upon a choosing a basis for $\Lambda = (\omega_1, \omega_2)$, this is equivalent to specifying a 2×2 matrix with integer coefficients and determinant n. The set of matrices $M(n)$ having determinant n was partitioned into $\mathrm{SL}(2,\mathbb{Z})$ orbits having upper triangular representatives. They key to showing self-adjointness is that these right coset representatives may be chosen so that they are also *left* coset representatives. This will follow easily from a double coset decomposition of $GL(2,\mathbb{Q})^+$. We'll proceed to these theorems now, and then later explain why double coset decompositions are natural from a geometric point of view, following discussion in Milne's notes "Modular functions and modular forms."

Proposition 6.4.1. *Let $\alpha \in \mathrm{GL}(2,\mathbb{Q})^+$. Then we have a decomposition into disjoint right cosets of*

$$\Gamma(1)\alpha\Gamma(1) = \bigcup_{i=1}^{\ell} \Gamma(1)\alpha_i$$

for some $\{\alpha_i\}$ in $\mathrm{GL}(2,\mathbb{Q})^+$.

Proof. We will show that

$$|\Gamma(1)\backslash\Gamma(1)\alpha\Gamma(1)| = [\Gamma(1) : \alpha^{-1}\Gamma(1)\alpha \cap \Gamma(1)].$$

This latter index is finite because $\alpha^{-1}\Gamma(1)\alpha$ is a congruence subgroup. Indeed, given $\alpha \in \mathrm{GL}(2,\mathbb{Q})^+$, then pick M_1, M_2 such that $M_1\alpha, M_2\alpha^{-1} \in \mathrm{Mat}(2,\mathbb{Z})$. Let $M = M_1M_2$. Given any $\gamma \in \Gamma(M)$, write $\gamma = I + Mg$ for some $g \in \mathrm{Mat}(2,\mathbb{Z})$. Then $\alpha\gamma\alpha^{-1} = I + (M_1\alpha)g(M_2\alpha^{-1}) \in \Gamma(1)$, hence $\Gamma(M) \subset \alpha^{-1}\Gamma(1)\alpha$.

To show that this is indeed the cardinality, note that the quotient is in bijection with

$$\Gamma(1)\backslash\Gamma(1)\alpha\Gamma(1)\alpha^{-1} \simeq (\Gamma(1) \cap \alpha\Gamma(1)\alpha^{-1})\backslash\alpha\Gamma(1)\alpha^{-1}$$

After conjugating by α, this has cardinality equal to $[\Gamma(1) : \alpha^{-1}\Gamma(1)\alpha \cap \Gamma(1)]$. $\square$

Now given $\alpha \in \mathrm{GL}(2,\mathbb{Q})^+$, define the Hecke operator T_α on $M_k(\Gamma(1))$ by

$$f|T_\alpha = \sum_{i=1}^{\ell} f|[\alpha_i]_k$$

with α_i in Proposition 6.4.1. We remind the reader that the slash action for matrices $\gamma \in \mathrm{GL}(2,\mathbb{Q})^+$ is given by

$$f(z)|[\gamma]_k = (\det\gamma)^{k/2}(cz+d)^{-k} f\left(\frac{az+b}{cz+d}\right).$$

The Hecke operator T_α is well-defined since f is modular, so independent of the choice of α_i. Further $f|T_\alpha$ is modular since, for any $\gamma \in \Gamma(1)$, Proposition 6.4.1 implies that $\Gamma(1)\alpha_i\gamma$ is a permutation of the right cosets $\Gamma(1)\alpha_i$. Hence, there exist $\gamma_i \in \Gamma(1)$ such that $\alpha_i\gamma$ are a permutation of the $\gamma_i\alpha_i$. Thus we obtain

$$(f|T_\alpha)|[\gamma]_k = \sum_i f|[\alpha_i\gamma]_k = \sum_i f|[\gamma_i\alpha_i]_k = \sum_i f|[\alpha_i]_k = f|T_\alpha.$$

Finally, we explain how to put a ring structure on the collection of Hecke operators. Given $\alpha, \beta \in \mathrm{GL}(2,\mathbb{Q})^+$, consider

$$f|T_\alpha|T_\beta = \sum_{i,j} f|[\alpha_i\beta_j]_k = \sum_{\sigma \in \Gamma(1)\backslash \mathrm{GL}(2,\mathbb{Q})^+} m(\alpha,\beta;\sigma) f|[\sigma]_k$$

where the α_i and β_j are right coset representatives as in Proposition 6.4.1, and $m(\alpha,\beta;\sigma)$ record the multiplicity of indices (i,j) such that $\sigma \in \Gamma(1)\alpha_i\beta_j$. This multiplicity depends only on the double coset $\Gamma(1)\sigma\Gamma(1)$ (again owing to the fact that acting on the right by γ permutes right cosets in the double coset decomposition), so we may write more simply

$$f|T_\alpha|T_\beta = \sum_{\sigma \in \Gamma(1)\backslash \mathrm{GL}(2,\mathbb{Q})^+/\Gamma(1)} m(\alpha,\beta;\sigma) f|T_\sigma.$$

Thus we may consider the free abelian group $\mathcal{R}$ generated by the symbols T_α where α runs over a complete set of double coset representatives for $\Gamma(1)\backslash \mathrm{GL}(2,\mathbb{Q})^+/\Gamma(1)$ and multiplication defined by

$$T_\alpha \cdot T_\beta = \sum_{\sigma \in \Gamma(1)\backslash \mathrm{GL}(2,\mathbb{Q})^+/\Gamma(1)} m(\alpha,\beta;\sigma) T_\sigma.$$

One can check that this multiplication is indeed associative, so this gives a ring structure on $\mathcal{R}$ acting on the space of modular forms $M_k(\Gamma(1)$.

Thus it is of interest to determine an explicit set of coset representatives for $\Gamma(1)\backslash \mathrm{GL}(2,\mathbb{Q})^+/\Gamma(1)$, which is accomplished by the elementary divisors theorem.

Proposition 6.4.2. *The set of diagonal matrices*

$$\mathrm{diag}(d_1, d_2) = \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} \quad d_1, d_2 \in \mathbb{Q} \text{ and } d_1/d_2 \text{ a positive integer}$$

gives a complete set of coset representatives for $\Gamma(1)\backslash \mathrm{GL}(2,\mathbb{Q})^+/\Gamma(1)$.

Proof. One can prove this directly using a sequence of invertible elementary row and column operations on 2×2 matrices. However, it is better viewed as a consequence of the very general elementary divisors theorem (Theorem III.7.8 in Lang). See Bump, Proposition 1.4.2 for details of the proof. □

We may now connect this new definition of Hecke operators to the one previously given in terms of lattices. Further, we may show directly that the Hecke algebra is commutative and the operators are self-adjoint.

We gave an earlier definition of Hecke operators using a right coset decomposition of $M(n)$, the set of 2×2 matrices with determinant n. We see now that

$$M(n) = \bigcup_{d_1,d_2} \Gamma(1) \begin{pmatrix} d_1 & 0 \\ 0 & d_2 \end{pmatrix} \Gamma(1)$$

where the union is over pairs (d_1, d_2) such that $d_1 d_2 = n$ and $d_2 | d_1$. Thus if we define

$$T(n) = \sum_{(d_1,d_2)} T_{\mathrm{diag}(d_1,d_2)},$$

then we recover our earlier definition of the Hecke operator $T(n)$ as in Section 6.3.

Proposition 6.4.3. *Given any* $\alpha \in \mathrm{GL}(2,\mathbb{Q})^+$*, we have the equality*

$$\Gamma(1)\alpha\Gamma(1) = \Gamma(1)({}^{\top}\alpha)\Gamma(1).$$

Proof. This is immediate from Proposition 6.4.2, since each double cosets contains a diagonal representative. □

Corollary 6.4.4. *Given* $\alpha \in \mathrm{GL}(2,\mathbb{Q})^+$*, the right coset representatives* α_i *in Proposition 6.4.1 may be chosen so that they are simultaneously left coset representatives:*

$$\Gamma(1)\alpha\Gamma(1) = \bigcup_{i=1}^{\ell} \alpha_i \Gamma(1).$$

Proof. From Proposition 6.4.2, the double coset $\Gamma(1)\alpha\Gamma(1)$ contains a diagonal representative, hence equals $\Gamma(1)^{\top}\alpha\Gamma(1) = \cup^{\top}\alpha_i\Gamma(1)$. Since α_i and ${}^{\top}\alpha_i$ generate the same double coset, we have

$$\Gamma(1)\alpha_i \cap {}^{\top}\alpha_i\Gamma(1) \neq \emptyset.$$

Let β_i be an element of this intersection. Then replacing each α_i with the corresponding β_i gives the desired set of representatives. □

Theorem 6.4.5. *The Hecke algebra $\mathcal{R}$ is commutative.*

Proof. The action of transposition is an antiautomorphism of $\mathrm{GL}(2,\mathbb{Q})^+$:

$${}^{\top}(AB) = {}^{\top}B\ {}^{\top}A,$$

and this induces an anti-automorphism of $\mathcal{R}$. To prove this latter assertion, we write the definition of the structure constants $m(\alpha,\beta;\sigma)$ more symmetrically in terms of double cosets. Let $\deg(\alpha) \stackrel{def}{=} |\Gamma(1)\backslash\Gamma(1)\alpha\Gamma(1)|$.

Claim: Let α_i, β_j be coset representatives for the double cosets of α, β, resp. Then

$$m(\alpha,\beta;\sigma) = \frac{1}{\deg(\sigma)}|\{(i,j)\ |\sigma \in \Gamma(1)\alpha_i\beta_j\Gamma(1)\}|.$$

To prove the claim, let σ_k be right coset representatives for σ:

$$\Gamma(1)\sigma\Gamma(1) = \bigcup \Gamma(1)\sigma_k.$$

Then $\sigma \in \Gamma(1)\alpha_i\beta_j\Gamma(1)$ if and only if $\sigma_k \in \Gamma(1)\alpha_i\beta_j$, and the number of such σ_k is $\deg(\sigma)$. Thus the claim follows because the number of pairs (i,j) such that $\sigma \in \Gamma(1)\alpha_i\beta_j\Gamma(1)$ is equal to

$$\sum_k m(\alpha,\beta;\sigma_k) = \sum_k m(\alpha,\beta;\sigma) = \deg(\sigma)m(\alpha,\beta;\sigma).$$

To show that transposition gives rise to an antiautomorphism of $\mathcal{R}$, we must show that $m(\alpha,\beta;\sigma) = m(\beta,\alpha;{}^{\top}\sigma)$. Indeed, extending linearly the action $* : T_\alpha \mapsto T_{{}^{\top}\alpha}$, we have

$$(T_\alpha \cdot T_\beta)^* = \sum m(\alpha,\beta;\sigma)T_{{}^{\top}\sigma}, \quad T_\beta^* \cdot T_\alpha^* = T_\beta \cdot T_\alpha = \sum m(\beta,\alpha;\sigma)T_\sigma.$$

We may choose right coset representatives that are simultaneously left coset representatives, as in Corollary 6.4.4. Thus using Proposition 6.4.3,

$$\Gamma(1)\alpha\Gamma(1) = \bigcup \Gamma(1)\alpha_i = \bigcup \Gamma(1)^{\top}\alpha_i,$$

and similarly for β with representatives β_j. Thus we have

$$\begin{aligned} m(\alpha,\beta;\sigma) = &\ \tfrac{1}{\deg(\sigma)}|\{(i,j)\ |\sigma \in \Gamma(1)^\top\alpha_i{}^\top\beta_j\Gamma(1)\}| \\ = &\ \tfrac{1}{\deg(\sigma)}|\{(i,j)\ |^\top\sigma \in \Gamma(1)\beta_j\alpha_i\Gamma(1)\}| \quad = m(\beta,\alpha;{}^\top\sigma). \end{aligned}$$

But by Proposition 6.4.3, transposition of double cosets acts as the identity in $\mathcal{R}$. Hence $\mathcal{R}$ must be commutative. This idea is often referred to as the "Gelfand trick" and is ubiquitous in the theory of Hecke algebras. □

Theorem 6.4.6. *The operators T_α on $S_k(\Gamma(1))$ are self-adjoint with respect to the Petersson inner product.*

Proof. One may check by change of variables $z \mapsto \alpha^{-1}z$ that

$$\langle f|[\alpha]_k, g\rangle = \langle f, g|[\alpha^{-1}]_k\rangle.$$

The left-hand side of the equality shows that the expression is invariant under left-translation $\alpha \mapsto \gamma\alpha$. The right-hand side shows that the expression is invariant under right translation $\alpha \mapsto \alpha\gamma$. Hence the inner product formula above depends only on the double coset of α in $\Gamma(1)\backslash \mathrm{GL}(2,\mathbb{Q})^+/\Gamma(1)$. Now to determine the self-adjointness, we examine

$$\langle f|T_\alpha, g\rangle = \sum_i \langle f|[\alpha_i]_k, g\rangle = \deg(\alpha)\langle f|[\alpha]_k, g\rangle = \deg(\alpha)\langle f, g|[\alpha^{-1}]_k\rangle,$$

where $\deg(\alpha) \stackrel{def}{=} |\Gamma(1)\backslash\Gamma(1)\alpha\Gamma(1)|$, as above. The second equality follows because these α_i are all defined to be in the same double coset of α. By Proposition 6.4.3, we have

$$\deg(\alpha)\langle f, g|[\alpha^{-1}]_k\rangle = \deg(\alpha)\langle f, g|[{}^\top\alpha^{-1}]_k\rangle.$$

Since scalar matrices act trivially, we may express this last quantity in terms of $\beta = \det(\alpha)^\top\alpha^{-1}$. Since $\beta = S({}^\top\beta^{-1})S^{-1}$ where S is the inversion matrix in $\mathrm{SL}(2,\mathbb{Z})$, then α and β lie in the same double coset (up to the trivial action of a diagonal matrix) and hence

$$\langle f|T_\alpha, g\rangle = \deg(\alpha)\langle f, g|[{}^\top\alpha^{-1}]_k\rangle = \deg(\alpha)\langle f, g|[\beta]_k\rangle = \deg(\alpha)\langle f, g|[\alpha]_k\rangle = \langle f, g|T_\alpha\rangle.$$

□

6.5 Hecke operators as correspondences

We have defined Hecke operators more generally via coset representatives for double cosets $\Gamma(1)\alpha\Gamma(1)$. Previously, we had chosen $\alpha \in \mathrm{GL}(2,\mathbb{Q})^+$. More generally, we may choose any $\alpha \in \mathrm{GL}(2,\mathbb{R})^+$ such that there exists a scalar multiple of α with integer coefficients. Indeed, this was the key ingredient in demonstrating that $\alpha^{-1}\Gamma(1)\alpha$ has finite index in $\Gamma(1)$ in Proposition 6.4.1, which works equally well for any finite index subgroup $\Gamma \subset \Gamma(1)$.

Using a similar argument to the proof of this earlier proposition, we may show that if $\alpha \in \mathrm{GL}(2,\mathbb{R})^+$, then if

$$\Gamma = \bigcup (\Gamma \cap \alpha^{-1}\Gamma\alpha)\beta_i \quad \text{then} \quad \Gamma\alpha\Gamma = \bigcup \Gamma\alpha_i \quad \text{where } \alpha_i = \alpha \cdot \beta_i,$$

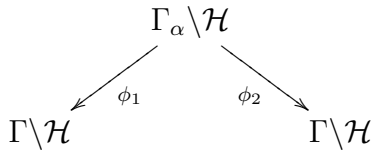
where both unions are disjoint. But why is double coset decomposition natural from the point of view of trying to construct operators on our Riemann surface $\Gamma\backslash\mathcal{H}$?

Given a point Γz in this space, we might try to act by α via group multiplication or linear fractional transformation, but we see both are doomed to fail. The map $\Gamma z \mapsto \alpha\Gamma z$ only gives a right coset if the group is normal in $\mathrm{GL}(2,\mathbb{R})^+$, which is not true. The map $\Gamma z \mapsto \Gamma\alpha z$ is similarly bad, because it depends on the coset representative z in Γz.

We rectify this situation by considering double cosets $\Gamma\alpha\Gamma = \bigcup \Gamma\alpha_i$ and then the map

$$\Gamma\backslash\mathcal{H} \to \Gamma\backslash\mathcal{H} : \ \Gamma z \mapsto \{\Gamma\alpha_i z\}$$

is independent of coset representatives for z. Of course, this isn't a function, but rather a one-to-many map. The right way to understand this map is as a correspondence. Let $\Gamma_\alpha = \Gamma \cap \alpha^{-1}\Gamma\alpha$. Consider the diagram:



Here the map on the right $\phi_2 : \Gamma_\alpha z \mapsto \Gamma\alpha z$. This is now a well-defined map on this restricted quotient. The map on the left is just $\phi_1 : \Gamma_\alpha z \mapsto \Gamma z = \bigcup \Gamma_\alpha \beta_i z$. Now given a function f on $\Gamma\backslash\mathcal{H}$, then the pullback $f \circ \phi_2$ is a function on $\Gamma_\alpha\backslash\mathcal{H}$. To obtain a function on $\Gamma\backslash\mathcal{H}$ we would like to take the inverse image of Γz under ϕ_1. This is not well-defined as Γz has many inverse images. However, the trace (i.e., summing over the inverse images) is well-defined. Thus $f \circ \phi_2 \circ \mathrm{tr}(\phi_1^{-1})(z) = \sum_i f|[\beta_i\alpha]_0 = \sum_i f|[\alpha_i]_0 = f|T_\alpha$. Here we assumed f is an automorphic function,

but more generally we may take f to be a k-fold differential on $\Gamma\backslash\mathcal{H}$ to obtain the definition for arbitrary weight modular forms in a similar way.

Thus we've made sense of the Hecke operator, initially a "one-to-many" map, as a correspondence. In general, a correspondence is any diagram of finite-to-one maps as in the above triple. Each element x in the lower left space may be thought of as mapped to elements $\{z\}$ in the lower right space that are in the image under ϕ_2 of the inverse image of x under ϕ_1.

6.6 Brief remarks on Hecke operators for Fuchsian groups

We have gradually broadened our perspective, first using the point of view taken in Serre's "Course in Arithmetic" and then generalizing to Bump's point of view from "Automorphic Forms and Representations." The most comprehensive discussion of Hecke operators remains Shimura's book "Introduction to the arithmetic theory of automorphic functions," and the reader familiar with our discussion of Hecke operators thus far should be well equipped to handle its level of abstraction. We content ourselves with a few brief comments about the general case of Fuchsian groups.

Given a discrete subgroup $\Gamma \in \mathrm{GL}(2,\mathbb{R})^+$, we may consider its "commensurator" $\widetilde{\Gamma}$ defined by

$$\widetilde{\Gamma} = \{\alpha \in \mathrm{GL}(2,\mathbb{R})^+ \mid \Gamma \sim \alpha^{-1}\Gamma\alpha\} \quad \text{where '}\sim\text{' means commensurable.}$$

Then for each $\alpha \in \widetilde{\Gamma}$, this definition ensures that we may decompose its double coset $\Gamma\alpha\Gamma$ as a finite sum of left or right cosets of Γ. We may define a Hecke algebra $\mathcal{R}$ with multiplication given by structure constants $m(\alpha,\beta;\sigma)$ just as before and since $\mathrm{GL}(2,\mathbb{R})^+$ possesses an anti-involution, then $\mathcal{R}$ is commutative. As before, we take

$$f|T_\alpha \stackrel{def}{=} \det(\alpha)^{k/2-1}\sum_{\alpha_i} f|[\alpha_i]_k, \quad \text{where} \quad \Gamma\alpha\Gamma = \bigcup_i \Gamma\alpha_i.$$

Let's specialize to the important examples $\Gamma = \Gamma_0(N)$ for some positive integer N. In this case, the algebra may be shown to be generated by operators

$$T_p = \Gamma\begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix}\Gamma; \quad R_p = \Gamma\begin{pmatrix} p & 0 \\ 0 & p \end{pmatrix}\Gamma, \quad p\text{: prime.}$$

One may show that $T(n)$ is self-adjoint provided that $\gcd(n,N) = 1$. This gives a simultaneous eigenbasis for the set of all operators $T(n)$ with $\gcd(n,N) = 1$ and from this we may establish a partial Euler product for a restricted Dirichlet series.

(Remember that the corresponding relations for the Fourier coefficients $c_f(n)$ of f only came from being a simultaneous eigenform, and we can no longer guarantee this for all n.) Given a cuspidal eigenform f for $\Gamma_0(N)$ with $c_f(1) = 1$, then

$$L^{(N)}(f,s) = \sum_{n:\gcd(n,N)=1} c_f(n)n^{-s} = \prod_{p\nmid N}(1 - c_f(p)p^{-s} + p^{k-1-2s})^{-1}.$$

Two problems arise here. First, it may not be true that $c_f(1) \neq 0$ as we only have the relation

$$c_f(n) = \lambda_f(n)c_f(1) \quad \text{for } n \text{ with } \gcd(n,N) = 1,$$

where $\lambda_f(n)$ denotes the eigenvalue of f with respect to $T(n)$. Second, we would prefer to define our Euler product over all primes p rather than this restricted class.

One could hope to obtain an eigenfunction for all $T(n)$ using a "multiplicity one" theorem. If one could show that the space of modular forms having system of eigenvalues $\{\lambda(n)\}$ for $\gcd(n,N) = 1$ is at most one dimensional, then since all the Hecke operators $T(p)$ with $p|N$ act on this eigenspace (as they commute with all other $T(n)$) then we would obtain a simultaneous eigenform.

Unfortunately, multiplicity one fails in general. It was Atkin and Lehner (1970) who identified the obstruction. Just as Dirichlet L-functions for primitive characters are the appropriate class satisfying a nice functional equation, there is an analogue of primitivity for modular forms. Given any proper divisor M of N, write $Md = N$. Then given a modular form $f \in M_k(\Gamma_0(M),\psi)$, then

$$g(z) = f(dz) \in M_k(\Gamma_0(N),\psi')$$

where ψ' is the induced character modulo N from ψ. More simply, f itself is a modular form on $M_k(\Gamma_0(N),\psi')$ as it trivially satisfies all of the conditions for modularity. If we let $S_k(\Gamma_0(N),\psi)^{\text{old}}$ be the image of these forms under these two types of maps coming from non-trivial divisors of N, then we may take its orthogonal complement $S_k(\Gamma_0(N),\psi)^{\text{new}}$, and this space is stable under $T(p)$ with $p \nmid N$ and satisfies the desired multiplicity one condition. See Theorem 1.4.5 of Bump for a statement of this result and for detailed references about the proof.

For example, if $N = q$, a prime, then

$$S_k(\Gamma_0(q)) = S_k(\Gamma(1)) \oplus q^*S_k(\Gamma(1)) \oplus S_k(\Gamma_0(q))^{\text{new}},$$

where $q^*S_k(\Gamma(1))$ consists of forms $f(qz)$ with $f \in S_k(\Gamma(1))$.

7 Modularity of elliptic curves

We will attempt to say as much as possible (before Spring Break) about the Taniyama-Shimura-Weil conjecture and its proof in special cases.

7.1 The modular equation for $\Gamma_0(N)$.

We have seen that $X_0(N) = \Gamma_0(N)\backslash\mathcal{H}^*$ is a compact Riemann surface. Now we show that $X_0(N)$ can in fact be defined as an algebraic curve (i.e., algebraic variety of dimension one) over a number field. As a first step, we show that $X_0(N)$ is birationally equivalent to the curve $F(X, Y) = 0$ for some canonically defined polynomial F with coefficients in $\mathbb{Q}$. This polynomial is referred to as the "modular equation" for $\Gamma_0(N)$.

We begin with a few elementary facts about divisors on compact Riemann surfaces. Let $\mathbb{C}(X)$ denote the field of meromorphic functions on X.

Proposition 7.1.1. *The field $\mathbb{C}(X)$ is an algebraic function field of dimension 1 over $\mathbb{C}$. That is, given a non-constant $f \in \mathbb{C}(X)$, then $\mathbb{C}(X)$ is a finite algebraic extension of the rational function field $\mathbb{C}(f)$. Moreover,*

$$[\mathbb{C}(X) : \mathbb{C}(f)] = \deg(f)_0 = \deg(f)_\infty.$$

See any standard text on Riemann surfaces for a proof of this fact.

Proposition 7.1.2. *The field $\mathbb{C}(X(1))$ is $\mathbb{C}(j)$ where j is the modular function defined previously (refer to Section 2.5) by*

$$j(z) = \frac{(12g_2)^3}{\Delta} = \frac{1}{q} + 744 + 196884q + \cdots, \quad \text{where } \Delta = g_2^3 - 27g_3^2 \text{ and } q = e(z)$$

Proof. In Corollary 2.2.7, we showed $\Delta(z) \neq 0$ for all $z \in \mathcal{H}$. Hence, $j(z)$ is holomorphic on $\mathcal{H}$, and viewed as a function on $\Gamma(1)\backslash\mathcal{H}^* = X(1)$, has only a simple pole at $z = \infty$ according to the Fourier expansion at ∞. Now applying Proposition 7.1.1, we see that $[\mathbb{C}(X) : \mathbb{C}(j)] = \deg(j)_\infty = 1$. □

Let $\mathbb{C}(X_0(N))$ denote the field of modular functions for $\Gamma_0(N)$ (i.e. the field of meromorphic modular forms of weight 0.)

Theorem 7.1.3. *The field $\mathbb{C}(X_0(N))$ is generated over $\mathbb{C}$ by $j(z)$ and $j(Nz)$. The minimal polynomial $F(j, Y) \in \mathbb{C}(j)[Y]$ for $j(Nz)$ over $\mathbb{C}(j)$ has degree*

$$\mu = [\Gamma(1) : \Gamma_0(N)] = N \cdot \prod_{p|N} \left(1 + \frac{1}{p}\right).$$

Moreover, $F(j,Y)$ is a polynomial in j with integer coefficients – that is, $F(X,Y) \in \mathbb{Z}[X,Y]$. For $N > 1$, $F(X,Y)$ is symmetric in X and Y. In the special case of $N = p$ prime,

$$F(X,Y) \equiv X^{p+1} + Y^{p+1} - X^pY^p - XY \text{ (mod } p\text{)}.$$

For example, when $N = 2$:

$$\begin{aligned} F(X,Y) = X^3 + Y^3 - X^2Y^2 + 1488XY(X+Y) - 162000(X^2+Y^2) \\ + 40773375XY + 8748000000(X+Y) - 157464000000000 \end{aligned}$$

This constant term is $2^{12} \cdot 3^9 \cdot 5^9$, while the coefficient of the linear terms is $2^8 \cdot 3^7 \cdot 5^6$.

Proof. Certainly $j(z)$ is a modular function for $\Gamma_0(N)$, as it's modular for $\Gamma(1)$. To see that $j(Nz)$ is a modular function for $\Gamma_0(N)$, write a typical element in the form $\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ with $c = c'N$ for some integer c' and consider

$$j(N\gamma z) = j\left(\frac{Naz+Nb}{cz+d}\right) = j\left(\frac{Naz+Nb}{Nc'z+d}\right) = j\left(\frac{a(Nz)+Nb}{c'(Nz)+d}\right) = j(Nz),$$

where the last step follows simply because $\begin{pmatrix} a & Nb \\ c' & d \end{pmatrix}$ is in $\Gamma(1)$.

We now investigate the minimal polynomial for $j(NZ)$ over $\mathbb{C}(j)$. Given any modular function $f(z)$ on $\Gamma_0(N)$, consider the finite set $f(\gamma_i z)$ where

$$\Gamma(1) = \bigcup_{i=1}^{\mu} \Gamma_0(N)\gamma_i.$$

Then $f(\gamma_i z)$ depends only on the coset $\Gamma_0(N)\gamma_i$. Because right translation by any element $\gamma \in \Gamma(1)$ permutes the right cosets $\Gamma_0(N)\gamma_i$, then the collection of functions $\{f(\gamma_i\gamma z)\}$ is, as a set, equal to the collection of functions $\{f(\gamma_i z)\}$. Thus we see that any symmetric polynomial in $f(\gamma_i z)$ will be invariant under $\Gamma(1)$. Because $f(z)$ is a modular function on $\Gamma_0(N)$, any such symmetric polynomial will be meromorphic. Hence the symmetric polynomial, as a modular function on $\Gamma(1)$, is expressible as a rational function of $j(z)$ according to Proposition 7.1.2. In short, we have demonstrated that

$$\prod_{i=1}^{\mu}(Y - f(\gamma_i z)) \tag{21}$$

is a polynomial in Y of degree μ with coefficients in $\mathbb{C}(j)$ having $f(z)$ as a root (using the factor in the product corresponding to the identity coset of $\Gamma_0(N)$). Note that since this construction works for any $f \in \mathbb{C}$, the degree of $\mathbb{C}(X_0(N))$ over $\mathbb{C}(j)$ is at most μ.

We now show that the polynomial in (21) with $f(z) = j(Nz)$ is a minimal polynomial for $j(Nz)$ over $\mathbb{C}(j)$. First, we claim that for any $f(z) \in \mathbb{C}(X_0(N))$, the $f(\gamma_i z)$ are also roots of the minimal polynomial $F(j, Y)$ for $f(z)$ over $\mathbb{C}(j)$. Indeed,

$$F(j(z), f(z)) = 0 \iff F(j(\gamma_i z), f(\gamma_i z)) = 0 \iff F(j(z), f(\gamma_i z)) = 0.$$

where the first equivalence is the change of variables $z \mapsto \gamma_i z$ and the second follows from the modularity of j on $\Gamma(1)$. Hence, to show (21) with $f(z) = j(Nz)$ is minimal for $j(Nz)$, it remains only to show that $j(N\gamma_i z)$ are distinct.

Suppose $j(N\gamma_i z) = j(N\gamma_{i'} z)$ for some $i \neq i'$. Since j defines an isomorphism between $X(1)$ and $\mathbb{P}^1(\mathbb{C})$, the Riemann sphere.[4] Thus if $j(N\gamma_i z) = j(N\gamma_{i'} z)$ for all z, then there exists a $\gamma \in \Gamma(1)$ such that $N\gamma_i z = \gamma N\gamma_{i'} z$ for all z. This implies

$$\begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix} \gamma_i = \pm\gamma \begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix} \gamma_{i'}$$

so that

$$\gamma_i \gamma_{i'}^{-1} \in \Gamma(1) \cap \begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix}^{-1} \Gamma(1) \begin{pmatrix} N & 0 \\ 0 & 1 \end{pmatrix} = \Gamma_0(N),$$

which contradicts the assumption that γ_i and $\gamma_{i'}$ were in distinct cosets of $\Gamma_0(N)$.

Thus the minimal polynomial for $j(Nz)$ over $\mathbb{C}(j)$ is $F(j, Y) = \prod_i (Y - j(N\gamma_i z))$. From this explicit form, we see that $F(j, Y)$ is holomorphic on $\mathcal{H}$, and hence its coefficients must be polynomial in j. So in fact $F(X, Y) \in \mathbb{C}[X, Y]$ rather than just $\mathbb{C}(X)[Y]$.

Next we show that the coefficients of $F(X, Y)$ of $\mathbb{C}[X, Y]$ are integral. Recall that the Fourier coefficients of j are integral, which follows from a straightforward computation using Theorem 2.4.2. To understand the Fourier expansion of $j(N\gamma_i z)$, we regard $N\gamma_i$ as a matrix in $M(N)$ – 2×2 integer matrices with determinant N. Since $j(\gamma N\gamma_i z) = j(N\gamma_i z)$ for any $\gamma \in \Gamma(1)$ then Lemma 6.3.1 guarantees the

[4] We haven't stated this explicitly before. It follows from the fact that $\infty \mapsto \infty$ according to the Fourier expansion of j and the isomorphism of $\Gamma(1)\backslash\mathcal{H}$ and $\mathbb{C}$ follows from the discussion at the end of Section 2.2 using the theory of elliptic curves. Alternately, any non-constant meromorphic function f on X defines a map to $\mathbb{P}^1(\mathbb{C})$ and then f takes each value on the Riemann sphere n times, where n is the number of poles (with multiplicity).

existence of an upper triangular matrix $\begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$ in this right coset with $ad = N$. Thus

$$j(N\gamma_i z) = j\left(\frac{az+b}{d}\right).$$

In terms of the Fourier expansions, this relation shows that $j(N\gamma_i z)$ has a Fourier expansion in powers of $q^{1/N}$ whose coefficients are in $\mathbb{Z}[e^{2\pi i/N}]$ since the Fourier coefficients of j are in $\mathbb{Z}$. This implies that the symmetric polynomials in $\{j(N\gamma_i z)\}$ have Fourier coefficients which are algebraic integers. We claim that this implies the symmetric polynomials (which are known thus far to be in $\mathbb{C}[j(z)]$) are expressible as polynomials P in j with algebraic integer coefficients.

Suppose not and write $P = \sum c_n j(z)^n$ and suppose m is the largest subscript among coefficients which are not algebraic integers. Then the coefficient of q^{-m} in the q expansion of P (obtained by substituting the q-expansion for j) is not an algebraic integer, so P can't be equal to a symmetric polynomial in the $\{j(N\gamma_i z)\}$.

So far, we have shown that the modular equation is of the form

$$F(X,Y) = \sum_{m,n} c_{m,n} X^m Y^n, \quad c_{m,n} \text{ algebraic integers, } c_{0,\mu} = 1.$$

Since $F(j(z), j(Nz)) = 0$, we may substitute the q-expansion for $j(z)$ and $j(Nz)$ and to obtain linear equations for the $c_{m,n}$ in each power of q. Since the q-series for $j(z)$ and $j(Nz)$ are integral, these involve only rational coefficients.

Because the minimal polynomial of $j(Nz)$ over $\mathbb{C}(j)$ is unique, this system has a unique solution over $\mathbb{C}$ and by the form of the equations, this has a solution over $\mathbb{Q}$. But since the $c_{m,n}$ were also algebraic integers, they must in fact be integers.

We leave the symmetry and congruence property when $N = p$ as exercises to the reader. □

7.2 The canonical model for $X_0(N)$ over $\mathbb{Q}$

We want to explain how the modular equation $F(X,Y)$ for $X_0(N)$ presented in the previous section defines a curve over the rational numbers. Defining an algebraic variety over a non-algebraically closed field is not as straightforward as in the algebraically closed case. There we may realize affine varieties as the solution sets to polynomials with coefficients in the field. Instead, we rephrase properties of curves in terms of their function fields, as we now explain.

Given a field k of characteristic 0, and a set of polynomials $\phi_1, \ldots, \phi_m \in k[x_1, \ldots, x_n]$, let I be the ideal generated by the ϕ_i in an algebraic closure $\overline{k}$ of k. That is, consider

$$I = \langle \phi_1, \ldots, \phi_m \rangle \subset \overline{k}[x_1, \ldots, x_n].$$

Then let C be the set of simultaneous solutions of the polynomials in I. Provided I is a prime ideal of $\overline{k}[x_1, \ldots, x_n]$, then we may define the integral domain

$$\overline{k}[C] \stackrel{def}{=} \overline{k}[x_1, \ldots, x_n]/I$$

and its field of fractions $\overline{k}(C)$ is called the *function field* of C over $\overline{k}$.

Definition 7.1. *If $\overline{k}(C)$ is a finite extension of $\overline{k}(t)$ where t is transcendental over $\overline{k}$ then C is said to be an affine algebraic curve over k. If, for each point $P \in C$, the $m \times n$ derivative matrix $[D_j\phi_i(P)]$ has rank $n-1$ then the curve is said to be non-singular.*

We study rational functions locally, via the maximal ideal m_P at each point $P \in C$:

$$m_P \stackrel{def}{=} \{f \in \overline{k}(C) \mid f(P) = 0\}.$$

Proposition 7.2.1. *Given a non-singular algebraic curve C and any point $P \in C$, then m_P/m_P^2 is a one-dimensional vector space over $\overline{k}$.*

Proof Sketch. Consider the perfect pairing

$$m_P/m_P^2 \times T_P(C) \longrightarrow \overline{k} \ : \ \langle f, v \rangle = \nabla f(P) \cdot v,$$

where $T_P(C)$ denotes the (1-dimensional) tangent space at P:

$$T_P(C) = \{v \in \overline{k}^n \mid [D_j\phi_i(P)]v = 0\}.$$

□

Then passing to the local ring $\overline{k}[C]_P$ of C over $\overline{k}$ at P (i.e. the localization with respect to all functions g vanishing at P), we have its unique maximal ideal

$$M_p \stackrel{def}{=} m_P\overline{k}[C]_P = \{f/g \in \overline{k}[C]_P \mid f(P) = 0\}.$$

Using the isomorphism $m_P/m_P^2 \to M_P/M_P^2$ together with the previous proposition, we conclude that M_P is principal.[5] This allows us to define a valuation at each point P on the curve.

$$v_P : \overline{k}[C] \to \mathbb{N} \cup \{\infty\} \quad \text{where} \quad f \mapsto \begin{cases} \infty & \text{if } f = 0, \\ e & \text{if } f = t^e u \text{ in } \overline{k}[C]_P,\ t\text{: uniformizer of } M_P. \end{cases}$$

[5]This follows from a standard application of Nakayama's Lemma.

This is then extended to the function field by defining the valuation of f/g to be $v_P(f) - v_P(g)$. It's the uniformizer and its valuation that allow us to make sense of rational functions on the curve. In particular, any rational function f/g on C defines a map to $\mathbb{P}^1(\overline{k})$ where f/g maps to 0 or ∞ according to whether $v_P(f/g) > 0$ or $v_P(f/g) < 0$, respectively.

Finally, we note that since all of our methods are local, they work equally well for projective varieties, and dehomogenizing a projective variety with respect to any variable gives affine pieces with isomorphic function fields. Hence it makes sense to consider the function field $\overline{k}(C)$ associated to a projective curve C by defining it on any (non-empty) affine piece.

Recall that a function field K of k (as a general term, not associated to a variety) is a finite extension of $k(t)$, where t is a transcendental element, such that $K \cap \overline{k} = k$.

Theorem 7.2.2. *There is an equivalence of categories between non-singular, projective algebraic curves over k (with equivalence given by isomorphism over k) and function fields over k (with equivalence given by isomorphism fixing k point-wise). The equivalence of categories is induced by the map*

$$C \mapsto k(C),$$

the subfield of $\overline{k}(C)$ defined as the field of fractions of $k[x_1, \ldots, x_n]/(I \cap k[x_1, \ldots, x_n])$.

Proof Sketch. We say only how to construct the map in the opposite direction. Given K, a finite extension of $k(t)$, the primitive element theorem guarantees the existence of some u such that $K = k(t, u)$ where u satisfies some polynomial relation with coefficients in $k(t)$. After clearing denominators if necessary, we obtain a polynomial $\varphi(t, u) = 0$ with $\varphi \in k[x, y]$. Since $K \cap \overline{k} = k$, one may show that the polynomial φ is irreducible over $\overline{k}$ (though this takes some argument). Thus

$$\{(x, y) \in (\overline{k})^2 \mid \varphi(x, y) = 0\}$$

defines a plane curve C' which can have finitely many singular points. Then we desingularize (see Chapter 7 of Fulton's "Algebraic Curves") to obtain a non-singular C having function field K. Note that C and C' are not necessarily isomorphic, but only birationally equivalent (recalling that a rational map is a weaker condition than that of morphism, being defined not at all points but at all but finitely many points in C). □

Note that this equivalence also suggests that we could have made our initial definition of an affine variety in terms of curves rather than function fields. We

wanted to stress the utility of this latter perspective in working locally, but Milne takes the former approach, which we very quickly review:

Equivalently, we could start with a finitely-generated algebra A over k:

$$A = k[X_1, \ldots, X_n]/\langle \phi_1, \ldots, \phi_m \rangle$$

for which $A \otimes_k \overline{k}$ is an integral domain, and then define the affine variety over k as the ringed space

$$\mathrm{Specm}(A) \stackrel{def}{=} (\mathrm{specm}(A), \mathcal{O}),$$

where $\mathrm{specm}(A)$ is the set of maximal ideals in A with a sheaf $\mathcal{O}$ compatibly defined according to its topology.

Given an affine variety X_k over k defined in this way, there is a canonical way of constructing a variety $X_{\overline{k}}$ over $\overline{k}$:

$$X_k = \mathrm{Specm}(A) \mapsto X_{\overline{k}} = \mathrm{Specm}(A \otimes_k \overline{k}).$$

In this case, we say X_k is a model for $X_{\overline{k}}$ over k. From our dictionary for affine varieties, we see that describing a model for $X_{\overline{k}}$ over k amounts to giving an ideal in $k[X_1, \ldots, X_n]$ which generates I as an ideal in $\overline{k}$.

We need one last important result from the theory of compact Riemann surfaces:

Theorem 7.2.3. *Every compact Riemann surface X has a unique structure of a non-singular projective, algebraic curve over $\mathbb{C}$. Under this correspondence, meromorphic functions are rational functions on the curve.*

Proof. For a statement of this result, and references for the proof, see Section I.2 of Griffiths' "Introduction to Algebraic Curves," especially Theorem 2.2. □

We may now finally explain how to obtain a model for $X_0(N)$ over $\mathbb{Q}$. According to Theorem 7.2.3, $X_0(N)$ has a unique structure as a projective algebraic curve over $\mathbb{C}$, which we denote by $X_0(N)_{\mathbb{C}}$ for emphasis. On the other hand, we have shown that the field of meromorphic functions

$$\mathbb{C}(X_0(N)) \simeq \mathbb{C}(j(z), j(Nz)) \simeq \mathbb{C}[X, Y]/\langle F_N(X, Y) \rangle$$

Using the equivalence of categories in Theorem 7.2.2, we may thus produce a projective curve, appropriately desingularized, call it $\overline{C}$ which is isomorphic to $X_0(N)_{\mathbb{C}}$.

But the same construction of $\overline{C}$ from Theorem 7.2.2 works equally well for $\mathbb{Q}$ as we've shown the coefficients of $F(X, Y)$ are rational. This, too, results in a projective non-singular curve C and we may thus regard C to be a model for the algebraic curve $X_0(N)_{\mathbb{C}}$. Note that in the course of this argument, we explained why the main theorem of the previous section provided only a birational equivalence.

7.3 The moduli problem and $X_0(N)$

We noted earlier that the quotient space $\Gamma(1)\backslash\mathcal{H} = Y(1)$ parametrizes the set of all lattices up to homothety, and hence each point corresponds to an isomorphism class of elliptic curves over $\mathbb{C}$. At the time, we loosely described $Y(1)$ as a "moduli space for elliptic curves over $\mathbb{C}$." In this section, we make this notion somewhat more precise and give a corresponding characterization of $Y_0(N)$ for $N > 1$.

Moreover, our method for finding the space of meromorphic functions on $X_0(N)$ may have appeared ad hoc. At the end of the section, we connect the problems of finding a generating set for this space and the moduli problem. We begin by following Milne in giving a brief summary of Mumford's precise definition of a moduli variety.

Definition 7.2. *A* **moduli problem** *over $k = \overline{k}$ (algebraically closed) is a contravariant functor $\mathcal{F}$ from the category of algebraic varieties over k to the category of sets. Typically, $\mathcal{F}(V)$ will be the set of isomorphism classes of objects over V.*

For example, let V be a variety over k and consider the family of elliptic curves E over V. That is, there is a map of algebraic varieties $E \to V$ where E is the subvariety of $V \times \mathbb{P}^2$ defined by a Weierstrass equation, say

$$Y^2Z + a_1XYZ + a_3YZ^2 = X^3 + a_2X^2Z + a_4XZ^2 + a_6Z^3$$

with coefficients a_i regular functions on V such that the regular function $\Delta(\{a_i\}) \neq 0$ on V. If we let $\mathcal{E}$ be the set of isomorphism classes of E over V, then $\mathcal{E}$ is contravariant, so defines a "moduli problem" over k. For example, we could take $V = \mathbb{A}^1(\mathbb{C})$.

Definition 7.3. *A* **solution to the moduli problem** *$\mathcal{F}$ is a pair (V, α) consisting of a variety V over k and a bijection $\alpha : \mathcal{F}(k) \to V(k)$ satisfying the following conditions:*

1. *Given another variety T over k, there exists a regular map $T(k) \to V(k)$ defined as follows. Let $f \in \mathcal{F}(T)$. Then since a point $t \in T(k)$ can be regarded as a map of varieties $\mathrm{Specm}(k) \to T$, then by contravariance of $\mathcal{F}$ we have a map $\mathcal{F}(T) \to \mathcal{F}(k) : f \mapsto f_t$. Composing with α we obtain a map*

$$T(k) \to V(k) : t \mapsto \alpha(f_t),$$

which we require to be a regular map (i.e. defined by a morphism of algebraic varieties).

2. *The bijection α is universal with respect to the above property: Given another map β such that maps $t \mapsto \beta(f_t)$ defines morphisms of algebraic varieties, then $\beta \circ \alpha^{-1}$ is a morphism of varieties.*

It is not hard to see that any such solution is unique up to isomorphism, using the universality property. On the other hand, such a solution may not exist in all cases and Mumford received the Fields medal largely for his construction of a solution for moduli varieties of curves and abelian varieties.

Returning to our example for elliptic curves, we can now see that the pair $(\mathbb{A}^1(\mathbb{C}), j)$ is a solution to the moduli problem $\mathcal{E}$ over $\mathbb{C}$. Indeed, if $E \to T$ is a family of curves over T, then we consider the map $T(\mathbb{C}) \to \mathbb{A}^1(\mathbb{C}) : t \mapsto j(E_t)$. This is a morphism since $j = c_4^3/\Delta$ where c_4 is a polynomial in the coefficients a_i (see p. 46 of Silverman) and $\Delta \neq 0$.

For universality, we must show that $j \mapsto \beta(E_j) : \mathbb{A}^1(\mathbb{C}) \to Z(\mathbb{C})$ is a morphism of varieties for any variety Z over $\mathbb{C}$. Here E_j denotes an elliptic curve with j-invariant equal to $j \in \mathbb{C}$. Let U be the open set of $\mathbb{A}^1$ omitting 0 and 1728. There is a one-parameter family of elliptic curves E_u given by

$$E_u : Y^2Z + XYZ = X^3 - \frac{36}{u - 1728}XZ^2 - \frac{1}{u - 1728}Z^3, \quad u \in U$$

with the property that $j(E_u) = u$. By assumption on the properties of (Z, β), E/U defines a morphism $U \to Z : u \mapsto \beta(E_u)$. Thus we are done, since this map is just restriction of $j \mapsto \beta(E_j)$ on $\mathbb{A}^1$, hence this map is a morphism as well. For this reason, the one-parameter family E_u is often referred to as the "universal elliptic curve."

We now turn to the moduli problem related to $Y_0(N)$. Given an elliptic curve E over $\mathbb{C}$, let S be a cyclic subgroup of order N in $E(\mathbb{C})$. Define an equivalence on all pairs (E, S) by setting $(E, S) \sim (E', S')$ if there exists an isomorphism $E \to E'$ mapping S to S'. Then given any complex variety V we may consider $\mathcal{E}_N(V)$ to be the set of equivalence classes of pairs (E, S) over V. Again, this defines a contravariant functor and hence a moduli problem.

Proposition 7.3.1. *The map*

$$\begin{array}{ccc} \mathcal{H} & \longrightarrow & \mathcal{E}_N(\mathbb{C}) \\ z & \longmapsto & \left(\mathbb{C}/\Lambda(z,1), \Lambda(z, \frac{1}{N})/\Lambda(z,1)\right) \end{array}$$

induces a bijection $Y_0(N) = \Gamma_0(N)\backslash\mathcal{H} \to \mathcal{E}_N(\mathbb{C})$.

Proof. Just show that the pair (E, S) corresponding to z is isomorphic to the pair (E', S') for z' if and only if z and z' are related by an element of $\Gamma_0(N)$. $\square$

Further let $\mathcal{E}_N^{\hom}(\mathbb{C})$ denote the set of homomorphisms $\alpha : E \to E'$ over $\mathbb{C}$ whose kernel is a cyclic group of order N. Then there is a bijection

$$\mathcal{E}_N^{\hom}(\mathbb{C}) \to \mathcal{E}_N(\mathbb{C}) : \ \alpha \mapsto (E, \ker(\alpha))$$

with inverse $(E, S) \mapsto (E \mapsto E/S)$. For example, the pair $\big(\mathbb{C}/\Lambda(z,1), \Lambda(z, \frac{1}{N})/\Lambda(z,1)\big)$ has $E/S = \mathbb{C}/\Lambda(z, \frac{1}{N})$ and so after changing by homothety maps to the element $(\mathbb{C}/\Lambda(z,1) \xrightarrow{N} \mathbb{C}/\Lambda(Nz,1))$. Then consider the map

$$\mathcal{E}_N^{\hom}(\mathbb{C}) \to \mathbb{A}^2(\mathbb{C}) : \ (E, E') \mapsto (j(E), j(E')).$$

By our above discussion, the image of this map is contained in the (singular) algebraic curve $C(\mathbb{C})$ defined by the modular equation $F_N(X, Y)$. Putting it all together, we have shown the following.

Theorem 7.3.2. *The moduli problem $\mathcal{E}_N$ has a solution $(Y_0(N), \phi)$ over $\mathbb{C}$. Here ϕ is the map*

$$\mathcal{E}_N(\mathbb{C}) \xrightarrow{\phi} Y_0(N) \xrightarrow{(j, j_N)} C(\mathbb{C})$$

where the composition is given by $(E, S) \mapsto (j(E), j(E/S))$. This remains a solution to the moduli problem over any field k with characteristic not dividing N, defining ϕ implicitly by this relation of maps and using the model for $Y_0(N)$ over $\mathbb{Q}$.

The moduli problem over $\mathbb{C}$ does not require the additional map (j, j_N), of course. It's just given by the bijection in Proposition 7.3.1 above. However, for other fields, one only defines ϕ implicitly via the composition of maps. The theorem can be proven for $\mathbb{Q}$ similarly to the earlier result for elliptic curves.

7.4 Hecke Correspondences for $Y_0(N)$

So far, we have two ways of viewing $Y_0(N) \subset X_0(N)$ – either as an algebraic curve defined by the modular equation (at least up to desingularization) or as the solution to the moduli problem for pairs (E, S) of an elliptic curve E and a cyclic subgroup S of order N. We used both to solve the moduli problem over arbitrary fields k with $N \nmid \mathrm{char}(k)$. Here we use both to define the Hecke correspondence $T(p)$ on the reduced curve $\widetilde{X}_0(N)$, the reduction of $X_0(N)$ modulo p. The theorem of Eichler and Shimura expresses $T(p)$ in terms of the Frobenius morphism.

In Section 6.5 at the end of our unit on Hecke theory, we described the operators T_α in terms of correspondences on curves $\Gamma\backslash\mathcal{H} = Y(\Gamma)$ according to the map

$$T_\alpha : \Gamma z \longmapsto \Gamma\alpha_i z, \quad \text{where } \Gamma\alpha\Gamma = \bigcup \Gamma\alpha_i$$

and this may be formally extended to $X(\Gamma)$.

Thinking of $X(\Gamma)$ as a projective algebraic curve, this induces a map on divisors which preserves divisors of degree 0, $\mathrm{Div}^0(X(\Gamma))$, as well as principal divisors. Thus we obtain a map on the Picard group:

$$\mathrm{Pic}^0(X) \stackrel{def}{=} \mathrm{Div}^0(X)/\{ \text{ principal divisors } \},$$

given by

$$\mathrm{Pic}^0(X) \longrightarrow \mathrm{Pic}^0(X) : \ [z] \longmapsto \sum_i [\alpha_i z].$$

In fact, given any correspondence of algebraic curves, we obtain such an induced map between their Picard groups. These correspondences form an abelian group under addition. In the case where the correspondence takes a curve X to itself, then composition defines a multiplication. Here composition means that we think of these correspondences as multivalued functions and compose them as in the usual composition of functions. This gives a ring structure to the space of correspondences.

Returning to the special example of Hecke operators, we now examine the case of $\Gamma = \Gamma_0(N)$ and α corresponding to the double coset $\Gamma_0(N)\begin{pmatrix} 1 & 0 \\ 0 & p \end{pmatrix}\Gamma_0(N)$. In this case, we previously referred to T_α as $T(p)$. Viewing $Y_0(N)$ as a curve over $\mathbb{C}$, we have identified points of $Y_0(N)$ with homomorphisms $E \to E'$ having kernel S, a cyclic subgroup of order N. As we will mention in the next section, the subgroup $E[p]$ of points of order dividing p is isomorphic to $(\mathbb{Z}/p\mathbb{Z})^2$ (See Corollary 6.4(b) of Silverman). Hence there are $p+1$ cyclic subgroups of $E[p]$ of order p. Label them $S_0, \ldots, S_p$ – geometrically they correspond to lines through the origin in $\mathbb{F}_p$. Then $T(p)$ sends the homomorphism $\alpha : \ E \mapsto E'$ to the collection of homomorphisms

$$\{E_i \mapsto E_i' \mid i = 0, \ldots, p\} \quad \text{where} \quad E_i = E/S_i, \quad E_i' = E'/\alpha(S_i).$$

Going a step further, we may consider $Y_0(N)$ as the curve C defined by the modular equation $F_N(X, Y)$ via the map $z \mapsto (j(z), j(Nz))$. If we represent a point of C by a pair (j, j_N) then given two elliptic curves E, E' with $j(E) = j$ and $j(E') = j_N$, our earlier identifications imply that there exists a homomorphism $\alpha : E \to E'$ with

kernel a cyclic subgroup of order N. Hence we may rephrase the correspondence $T(p)$ in terms of points on C as follows:

$$T(p) : (j, j_N) \longmapsto \{(j_i, j_{N,i}) \mid i = 0, \ldots, p\} \quad \text{where } j_i = j(E/S_i), j_{N,i} = j(E'/\alpha(S_i)).$$

These last two descriptions of the action of $T(p)$ remain valid over any field of characteristic not equal to p. In the next section, we review basics of elliptic curves and the Frobenius morphism that allow us to study this last very interesting case.

7.5 The Frobenius Map

Now let C be an algebraic curve defined over a field k of characteristic $p \neq 0$. Then we may define a new curve $C^{(q)}$ where q is a power of p by taking each polynomial ϕ_j in its defining ideal to $\phi_j^{(q)}$, the polynomial obtained by raising each coefficient of ϕ_j to the q-th power. Thus we obtain a natural map, the q-th power Frobenius morphism, defined by

$$\phi : C \longrightarrow C^{(q)} : \ [x_0, \ldots, x_n] \longmapsto [x_0^q, \ldots, x_n^q]$$

We now describe the basic properties of the Frobenius map.

Given a non-constant rational map of curves $\phi : C_1 \to C_2$ defined over k, then composition with ϕ induces an inclusion of function fields

$$\phi^* : k(C_2) \to k(C_1) : \ \phi^*(f) = f \circ \phi.$$

Lemma 7.5.1. *For a non-constant map ϕ, $k(C_1)$ is a finite extension of $\phi^* k(C_2)$.*

Proof. See Hartshorne, II.6.8. □

Thus it makes sense to define the degree of a map $\phi : C_1 \to C_2$ by

$$\deg(\phi) \stackrel{def}{=} [k(C_1) : \phi^* k(C_2)]$$

unless ϕ is constant, in which case we set $\deg(\phi) = 0$. Moreover, we say that the map ϕ is separable (resp. inseparable, purely inseparable) if the corresponding extension of function fields $k(C_1)$ over $\phi^* k(C_2)$ has this property.

For example, the p-th Frobenius map on $\mathbb{P}^1(\overline{\mathbb{F}}_p)$ is $\phi(t) = t^p$ on the affine part, so the induced map on function fields gives $k(C) = \mathbb{F}_p(t)$ over $\phi^* k(C^{(p)} = \mathbb{F}_p(s)$ with $s = t^p$. The minimal polynomial of t over $\mathbb{F}_p(s)$ is just $x^p - s$ so the degree of the extension is p, despite the fact that the Frobenius map is a bijection. Moreover, we see that the extension is purely inseparable. These facts hold more generally:

Proposition 7.5.2. *Let k be a perfect field of characteristic p. Let ϕ be the q-th power Frobenius map from $C \to C^{(q)}$. Then*

1. $\phi^* k(C^{(q)}) = k(C)^q = \{f^q \mid f \in k(C)\}$
2. *ϕ is purely inseparable.*
3. $\deg(\phi) = q$.

Proof. See Silverman, Proposition II.2.11. □

Corollary 7.5.3. *Every map $\psi : C_1 \to C_2$ of smooth curves over k factors into a separable map λ and the q-th power Frobenius map where $q = \deg_i(\psi)$, the inseparable degree. That is:*

$$\psi : \ C_1 \xrightarrow{\phi} C_1^{(q)} \xrightarrow{\lambda} C_2.$$

In particular, if ψ is purely inseparable, then λ is an isomorphism.

Proof. See Silverman, Corollary II.2.12. □

We now apply these results to the multiplication by p map as an endomorphism of an elliptic curve. First recall the following lemma:

Lemma 7.5.4. *Given any integer m, the multiplication-by-m map on E, an elliptic curve, has degree m^2. Moreover, in the special case $m = p$, the map is separable if $\mathrm{char}(k) \neq p$. If $\mathrm{char}(k) = p$, then either the map is purely inseparable and $E[p] = \{0\}$ or its separable degree is p and $E[p] \simeq \mathbb{Z}/p\mathbb{Z}$.*

Proof. See Silverman, Corollary III.6.4. □

If we begin with an elliptic curve E over $\mathbb{Q}$ and we reduce modulo p and obtain a non-singular curve $\tilde{E}$ over $\mathbb{F}_p$, then we say that E has "ordinary reduction" if $\tilde{E}[p] \simeq \mathbb{Z}/p\mathbb{Z}$ and "supersingular reduction" if $\tilde{E}[p] = \{0\}$.

Thus, if the multiplication-by-p map is purely inseparable, then Corollary 7.5.3 implies that

$$E \xrightarrow{\phi_{p^2}} E^{(p^2)} \xrightarrow{\simeq} E.$$

So we must have $E \simeq E^{(p^2)}$ in this case, which in turn implies that $j(E) = j(E^{(p^2)}) = j(E)^{p^2}$, where the last equality follows easily from the description of j in terms of Weierstrass coefficients. Hence if E has no points of order p, then $j(E) \in \mathbb{F}_{p^2}$. This shows that there are only finitely many isomorphism classes of supersingular elliptic curves. See Silverman, Section V.4, for more precise criterion determining supersingularity.

7.6 Eichler-Shimura theory

Given the curve $X_0(N)$ as a variety over $\mathbb{Q}$, we may consider its reduction modulo p and call the resulting curve $\tilde{X}_0(N)$. For almost all primes $p \nmid N$ this curve will have good reduction (i.e. descends to a non-singular curve over $\mathbb{F}_p$). For these primes p, the Hecke correspondence $T(p)$ (viewed as a curve in $X_0(N) \times X_0(N)$ – essentially the graph of the many valued function) similarly descends to a Hecke correspondence $\tilde{T}(p)$ on the reduced curve.

In Section 6.5, we gave a diagram for the action of T_α in terms of $\Gamma_\alpha = \Gamma \cap \alpha^{-1}\Gamma\alpha$. In the special case of $\Gamma = \Gamma_0(N)$ and $\alpha = \begin{pmatrix} 1 & 0 \\ 0 & p \end{pmatrix}$, we have $\Gamma_\alpha = \Gamma_0(Np)$ and the correspondence:

$$\begin{array}{ccc} & \Gamma_0(Np)\backslash\mathcal{H} & \\ \swarrow^{\phi_1} & & \searrow^{\phi_2} \\ \Gamma_0(N)\backslash\mathcal{H} & & \Gamma_0(N)\backslash\mathcal{H} \end{array}$$

In more modern language, we seek a solution to the (coarse) moduli problem for pairs (E, C) which is simultaneously a model for $Y_0(N)$ defined over $\mathbb{Z}[1/N]$. This permits a definition of $\tilde{T}(p)$ over fields of characteristic p.

Further, note that any morphism of curves $\phi : C_1 \to C_2$ may be viewed as a correspondence by considering the projections to C_1 and C_2 of the graph of ϕ as a subset of $C_1 \times C_2$. The transpose of a correspondence is obtained by reversing the roles of C_1 and C_2 (i.e., take preimages of C_2 and then their resulting images in C_1), and thus every morphism has a well-defined transpose.

Theorem 7.6.1 (Eichler, Shimura)**.** *Let p be a prime for which $X_0(N)$ has good reduction. Then*

$$\tilde{T}(p) = \phi_p + \phi_p'$$

where ϕ_p denotes the p-th power Frobenius map over $\overline{\mathbb{F}}_p$, ϕ_p' denotes its transpose, and equality is understood as occurring in the ring of correspondences of $\tilde{X}_0(N)/\mathbb{F}_p$ defined over $\overline{\mathbb{F}}_p$.

Most accounts of this theorem in the literature are either too long or too short for our purposes. A very brief account can be found in either:

- "Rational points on modular elliptic curves," H. Darmon, CBMS no. 101. (2004) See in particular Section 2.5 and the sketch of the proof in Theorem 2.10 on pp. 18–19.

- "Fermat's Last Theorem," H. Darmon, F. Diamond, and R. Taylor, in "Elliptic curves, modular forms, and Fermat's last theorem." (1997) (See in particular the discussion titled "Hecke operators:" on pp. 31–32.) Their result is a slight variant of ours as they work with Hecke operators on $\Gamma_1(N)$.

Alternately, Knapp's book "Elliptic Curves" has an entire chapter (Ch. 11, pp. 302-385) devoted to Eichler-Shimura theory, but does not take up the proof of the theorem above, only discussing the statement (Equation 11.118) and some of its implications, which we pursue in the next section.

Proof Sketch. It suffices to show that they agree on a dense subset of the $\overline{\mathbb{F}}_p$ points of $\tilde{X}_0(N)$. Thus we may make two simplifying assumptions: first, suppose $\tilde{P}$ is a point of $\tilde{Y}_0(N)$ so that each point can be given the form $(j(\tilde{E}), j(\tilde{E}'))$ for a map $\tilde{\alpha} : \tilde{E} \to \tilde{E}'$. Second, we may assume that $\tilde{E}$ has p-torsion of order p (as otherwise, the j invariant must lie in $\mathbb{F}_{p^2}$, as discussed at the end of the previous section).

Over $\overline{\mathbb{Q}}_p$, which has characteristic 0, we have a description of the Hecke operator $T(p)$ in terms of homomorphisms between elliptic curves E, E' having cyclic kernel of order N. Let α be a lifting of $\tilde{\alpha}$ to $\overline{\mathbb{Q}}_p$. Consider the reduction map

$$E[p](\overline{\mathbb{Q}}_p) \to \tilde{E}[p](\overline{\mathbb{F}}_p)$$

(defined by passing to the residue class field) which has kernel of order p. Let S_0 be the kernel of this map so that the other subgroups of order p in $E[p](\overline{\mathbb{Q}}_p)$ are labeled $S_1, \ldots, S_p$.

Consider the map $[p] : \tilde{E} \to \tilde{E}$. It factors, for any $i = 0, \ldots, p$, as

$$\tilde{E} \xrightarrow{\varphi} \tilde{E}/S_i \xrightarrow{\psi} \tilde{E}.$$

If $i = 0$, then φ is purely inseparable of degree p since it is the reduction of the map $E \mapsto E/S_0$ which has degree p and zero kernel. This forces ψ to be separable of degree p as we are assuming that the p-torsion (the kernel of $[p]$) has order p. By Corollary 7.5.3, this implies $\tilde{E}^{(p)} \simeq \tilde{E}/S_0$. The same holds for $\tilde{E}'$. Thus,

$$(j(\tilde{E}_0), j(\tilde{E}'_0)) = (j(\tilde{E}^{(p)}), j(\tilde{E}'^{(p)})) = (j(\tilde{E})^p, j(\tilde{E}')^p) = \phi_p(j(\tilde{E}), j(\tilde{E}')).$$

For $i \neq 0$, the map φ has non-trivial kernel of order p (the reduction of S_i) so is separable (and hence ψ must be purely inseparable). This implies $\tilde{E} \simeq \tilde{E}_i^{(p)}$ and likewise for $\tilde{E}'$. Thus

$$(j(\tilde{E}_i^{(p)}), j(\tilde{E}_i'^{(p)})) = (j(\tilde{E}), j(\tilde{E}')),$$

or equivalently that each of the pairs $(j(\tilde{E}_i), j(\tilde{E}'_i))$ is in the inverse image of ϕ_p and hence in the image of $\phi'_p(j(\tilde{E}), j(\tilde{E}'))$. □

7.7 Zeta functions of curves over finite fields

We begin by recalling Weil's conjectures for varieties over finite fields and then examine them in more detail for curves. Our discussion will closely follow Silverman's in V.2 of "Arithmetic of Elliptic Curves."

Let V be a projective variety over a finite field $\mathbb{F}_q$ having $q = p^r$ elements. Suppose that V is defined by polynomials $f_1, \ldots, f_m$. Then we form a generating function using the number of solutions to V over $\mathbb{F}_{q^n}$ for positive integers n:

$$Z(V;T) = \exp\left(\sum_{n=1}^{\infty} |V(\mathbb{F}_{q^n})| \frac{T^n}{n}\right),$$

where exp is defined according to the usual power series expansion. This is known as the zeta function of V over $\mathbb{F}_q$.

Theorem 7.7.1 (Weil Conjectures)**.** *Given V, a projective variety of dimension n over $\mathbb{F}_q$, the zeta function $Z(V,T)$ satisfies the following properties:*

- *Rationality: $Z(V;T) \in \mathbb{Q}(T)$*
- *Functional Equation: There exists an integer ε such that*

$$Z(V;1/q^nT) = \pm q^{n\varepsilon/2} T^{\varepsilon} Z(V;T)$$

- *Riemann Hypothesis: There exists a factorization*

$$Z(V;T) = \frac{P_1(T)\cdots P_{2n-1}(T)}{P_0(T)P_2(T)\cdots P_{2n}(T)}$$

with each $P_i(T) \in \mathbb{Z}[T]$, $P_0(T) = 1 - T$, and $P_{2n}(T) = 1 - q^nT$. Each of the remaining P_i with $1 \le i \le 2n-1$ factor over $\mathbb{C}$ as

$$P_i(T) = \prod_j (1 - \alpha_{ij}T) \quad \textit{with} \quad |\alpha_{ij}| = q^{i/2}.$$

The rationality was proved by Dwork (1960), the functional equation via ℓ-adic cohomology by the Artin/Grothendieck school, and then the Riemann hypothesis by Deligne (1973).

In the special case where V is an elliptic curve E, the theorem states that we have

$$Z(E;T) = \frac{(1-\alpha T)(1-\beta T)}{(1-T)(1-qT)}.$$

We now explain the meaning of the complex numbers α, β in algebraic terms.

First recall that any isogeny $\phi : E_1 \to E_2$ induces a map on ℓ-torsion points (or more generally ℓ^n-torsion points $E_1[\ell^n] \to E_2[\ell^n]$ for any n) and thus induces a map ϕ_ℓ on their Tate modules where we define

$$T_\ell(E) = \varprojlim_n E[\ell^n],$$

which carries the natural structure of a $\mathbb{Z}_\ell$-module. In particular if $E_1 = E_2$ we obtain an endomorphism of $T_\ell(E)$. Let ρ_ℓ be this induced homomorphism

$$\rho_\ell : \ \mathrm{End}(E) \longrightarrow \mathrm{End}(T_\ell(E)). \tag{22}$$

Silverman proves the much stronger result, in Thereom III.7.4, that the map

$$\mathrm{End}(E) \otimes \mathbb{Z}_\ell \longrightarrow \mathrm{End}(T_\ell(E))$$

is an isomorphism.

If ℓ is a prime different from $\mathrm{char}(k)$ then we may choose a basis for $T_\ell(E)$ and write the endomorphism ϕ_ℓ as a 2×2 matrix.

Lemma 7.7.2. *Let $\phi \in \mathrm{End}(E)$. Then*

$$\det(\phi_\ell) = \deg(\phi), \quad \mathrm{tr}(\phi_\ell) = 1 + \deg(\phi) - \deg(1 - \phi).$$

Proof. See Silverman, Proposition V.2.3. □

In the special case that ϕ is the q-th power Frobenius endomorphism (for the field $\mathbb{F}_q$), then $(1 - \phi)$ is a separable map (see Corollary III.5.5 of Silverman) and hence

$$|E(\mathbb{F}_q)| = |\ker(1 - \phi)| = \deg(1 - \phi).$$

Similarly, using the q^n-th power Frobenius, we have

$$|E(\mathbb{F}_{q^n})| = \deg(1 - \phi^n) = \det(\phi_\ell^n),$$

where we've used the previous lemma in the last equality. If we let α and β be the roots of the characteristic polynomial for ϕ_ℓ, then we have

$$|E(\mathbb{F}_{q^n})| = 1 - \alpha^n - \beta^n + q^n$$

by considering the characteristic polynomial of ϕ_ℓ^n after changing basis to put ϕ_ℓ in Jordan canonical form, and using the fact that $\alpha\beta = \det(\phi_\ell) = \deg \phi = q$. The Weil

conjectures for elliptic curves follow easily from this last identity. In particular, we obtain the interesting polynomial $P_1(T)$ in the numerator is given by

$$P_1(T) = 1 - aT + qT^2 = (1 - \alpha T)(1 - \beta T)$$

Using the Lemma we obtain the more familiar characterization of a in terms of counting points over $\mathbb{F}_q$:

$$a = \alpha + \beta = \mathrm{tr}(\phi_\ell) = 1 + q - \deg(1 - \phi) = 1 + q - |E(\mathbb{F}_q)|.$$

To build an L-function over a number field K, we will use these zeta functions at each residue field corresponding to primes p of K.

7.8 Hasse-Weil L-functions of a curve over $\mathbb{Q}$

Given a complete non-singular curve over $\mathbb{Q}$, we may consider its reduction modulo p. For almost all primes p, this will be a complete non-singular curve over $\mathbb{F}_p$. Then we consider the Hasse-Weil L-function

$$L(C, s) = \prod_p L_p(C, p^{-s})^{-1}$$

where the product is taken over all primes p and if p is a prime of good reduction, then $L_p(C, T)$ is defined to be the lone polynomial $P_1(T)$ appearing in the numerator of $Z(C/\mathbb{F}_p; T)$, the zeta function of a curve over a finite field. At primes with bad reduction, one needs a slightly adjusted definition which is best given in terms of representation theory.

However, for elliptic curves, we can give a precise definition of the bad factors according to their type of reduction. Suppose p is a prime of bad reduction for E. Recall that the bad reduction is classified according to the type of singularity appearing in the reduced curve. If it is a node, then E is said to have multiplicative reduction (which is "split" if the slopes of the tangent lines at the node are in $\mathbb{F}_q$, and otherwise "non-split"). If the singularity is a cusp, then we say E has additive reduction. These geometric qualities can be easily obtained from the Weierstrass equation for the curve. (See Proposition VII.5.1 of Silverman for the details.)

Then to each of these types of bad reduction at p, we assign the local L-factor

$$L_p(E, T) = \begin{cases} (1 - T) & \text{split, multiplicative,} \\ (1 + T) & \text{non-split, multiplicative,} \\ 1 & \text{additive.} \end{cases}$$

With these definitions, we commonly write

$$L(E,s) = \prod_p L_p(E, p^{-s})^{-1},$$

where the p in p^{-s} should be regarded as the order of the residue class field at the prime p. The same definition works the L-function of an elliptic curve over any number field. We may use the Riemann hypothesis for elliptic curves to conclude that the L-series converges to an analytic function for $\mathrm{Re}(s) > 3/2$.

Thanks to the modularity of elliptic curves (which, in one form, states that there exists a cusp form f such that $L(E,s) = L(f,s)$, where the latter L-function is defined via Mellin transform as usual), we know that $L(E,s)$ possesses analytic continuation to the entire complex plane and satisfies a functional equation relating the values at s and $2-s$. All known proofs of these analytic properties follow from matching the L-function with one coming from automorphic forms.

It was Weil who refined the statement of Taniyama's conjecture, predicting that the level of the corresponding modular form f should be equal to the conductor N of the elliptic curve. This conductor is an integral ideal defined as follows. To each place v of the number field K, define the exponent

$$f_v = \begin{cases} 0 & \text{if } E \text{ has good reduction at } v, \\ 1 & \text{if } E \text{ has multiplicative reduction at } v, \\ 2 & \text{if } E \text{ has additive reduction, and } v \text{ is not above } 2,3, \\ \geq 2 & \text{if } E \text{ has additive reduction, and } v \text{ is above } 2,3. \end{cases}$$

Ogg has given a formula for computing f_v exactly. See his article "Elliptic curves and wild ramification" in Amer. J. Math (1967). Then the conductor of E is defined to be (the integral ideal) $N = \prod_{v \in K} p_v^{f_v}$.

For example, if $K = \mathbb{Q}$, then the precise form of the functional equation is

$$L^*(E, 2-s) = \pm L^*(E,s), \quad \text{where } L^*(E,s) \stackrel{def}{=} N^{s/2}(2\pi)^{-s}\Gamma(s)L(E,s).$$

7.9 Eichler-Shimura theory in genus one

Suppose that N is chosen such that $X_0(N)$ has genus one. Recalling our formulas proved in the problem sets for the genus, this implies that N is one of twelve integers in the interval from 11 to 49. Taking the cusp $\{\infty\}$ to be the distinguished point on the curve, these cases of $X_0(N)$ are then elliptic curves defined over $\mathbb{Q}$.

Further, we've computed dimension formulas for the space of cusp forms of a given weight on $X_0(N)$. In particular, the dimension of the space of weight 2 cusp forms is g, the genus of the curve. So there is a unique (up to constant) cusp form f of weight 2 associated to $X_0(N)$, corresponding to the one-dimensional space of holomorphic 1-forms on the Riemann surface.

Theorem 7.9.1. *Let N be chosen so that $X_0(N)$ has genus one, and let f be the normalized cusp form (i.e. $a_f(1) = 1$) of weight 2 on $X_0(N)$. Then*

$$L(X_0(N), s) = L(f, s),$$

where the left-hand side is the Hasse-Weil L-function of an elliptic curve and the right-hand side is the Mellin transform of f.

Remark 7.1. One potential cause of confusion in these modularity results is that there are TWO sources of connection to elliptic curves. The first was used to give a moduli interpretation of the curve, which was useful in formulating versions of the Hecke correspondence in the proof of the Eichler-Shimura relation. The second has only been hinted at so far – for N with $X_0(N)$ of genus one, the curve $X_0(N)$ is the elliptic curve appearing in the modularity theorem. More generally, one needs an appropriate quotient of the Jacobian associated to $X_0(N)$.

Proof. We prove that the L-functions agree for all but finitely many places v – those corresponding to the primes of good reduction for $X_0(N)$. Because the space of cusp forms has dimension one, and f is assumed normalized, then it is automatically an eigenfunction of the Hecke operators $T(p)$ with $T(p)f = a_f(p)f$ for all primes p. Since the Petersson inner product is self-adjoint, these eigenvalues (or equivalently Fourier coefficients) are guaranteed to be real. (A bit more work shows they can in fact be taken to be integral, but we won't need that here.) We must show $a_f(p) = a_E(p)$, where we recall that $a_E(p)$ is given by the trace of Frobenius acting on the Tate module.

Now consider $\tilde{X}_0(N)$, the reduction of $X_0(N)$ modulo p. The endomorphisms ϕ_p, ϕ_p' (Frobenius and its transpose) act on $\tilde{X}_0(N)$ and satisfy

$$\phi_p \circ \phi_p' = [\deg(\phi_p)] = [p] \tag{23}$$

(as the transpose of ϕ_p may be viewed as a dual isogeny. See Section III.6 of Silverman for details).

If we let ρ_ℓ denote the map from $\mathrm{End}(E)$ to $\mathrm{End}(T_\ell(E))$ as in (22), then the relation (23) implies that

$$(I_2 - \rho_\ell(\phi_p)T)(I_2 - \rho_\ell(\phi_p')T) = I_2 - (\rho_\ell(\phi_p + \phi_p'))T + pT^2. \tag{24}$$

By the Eichler-Shimura relation (Theorem 7.6.1) we can replace $\phi_p + \phi'_p$ by $\tilde{T}(p)$. Milne claims that since the ℓ-adic representation is unchanged by reduction modulo $p \neq \ell$, then we may replace $\tilde{T}(p)$ by $T(p)$. Thus the right-hand side of (24) is just

$$I_2 - \begin{pmatrix} a_p & 0 \\ 0 & a_p \end{pmatrix} T + pT^2.$$

Upon taking determinants and using Lemma 7.7.2, the claim follows. □

8 Defining automorphic forms on groups

In the second half of our course, we will study a wider class of automorphic forms for discrete subgroups of $\mathrm{SL}(2,\mathbb{R})$ – those that are eigenfuctions of the Laplace-Beltrami operator Δ. In fact, it is better to work with functions on the group $\mathrm{SL}(2,\mathbb{R})$ rather than on $\mathcal{H}$ so we will make this transition quickly.

8.1 Eigenfunctions of Δ on $\mathcal{H}$ – Maass forms

Working in rectangular coordinates, we write $z = x + iy$ and the corresponding Laplacian $\Delta = -y^2\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)$ on $\mathcal{H}$. We seek automorphic functions that are eigenfunctions of Δ.

Definition 8.1 (Maass form)**.** *Given a discrete subgroup Γ of $\mathrm{SL}(2,\mathbb{R})$, a smooth function $f : \mathcal{H} \to \mathbb{C}$ is called a* **Maass form** *with respect to Γ if it satisfies the following three conditions:*

- *$f(\gamma z) = f(z)$ for all $\gamma \in \Gamma$,*
- *$\Delta f(z) = \lambda f(z)$ for some $\lambda \in \mathbb{C}$,*
- *For any x, $f(x+iy) = O(y^N)$ as $y \to \infty$ for some N.*

Note in particular that holomorphic functions are eigenfunctions of the Laplacian with eigenvalue 0. However, this definition does not make immediately clear what the relationship between Maass forms and modular forms might be, since we require Maass forms to be "automorphic functions" (i.e. automorphic forms of weight 0). One can consider Maass forms of higher weight, hence the reason for the use of "form" not "function."

Previously, we studied modular forms using their Fourier series. Each such modular form was expressible as a function of $e(z/h)$ for some h because it's a holomorphic function satisfying $f(z) = f(z+h)$. But we're no longer assuming holomorphicity, so we need to study how we can represent Maass forms in some canonical way or via a good basis. We do still have a Fourier expansion in the real variable x from the invariance under translation, but need to understand what happens in the y component of z. To answer this question, we study eigenfunctions of Δ.

If we want f to be a function in y alone (i.e. constant in x) then the second-order differential equation is easily seen to produce two linearly independent solutions:

$$\frac{1}{2}(y^s + y^{1-s}) \quad \text{and} \quad \frac{1}{2s-1}(y^s - y^{1-s}), \tag{25}$$

where $\lambda = s(1-s)$. In particular, the map $s \mapsto \lambda$ is 2-to-1 on $\mathbb{C}$, except at $s = 1/2 \mapsto \lambda = 1/4$. At $s = 1/2$, the pair of eigenfunctions are $y^{1/2}$ and $y^{1/2} \log y$, respectively. If $s \neq 1/2$, it is often more convenient to consider the simpler pair of solutions y^s, y^{1-s}.

Building up to slightly more complicated solutions, we could seek functions $f(z)$ which are periodic in x of period 1. For example, a natural guess is $f(z) = e(x)F(2\pi y)$ for some function F. The change of variables $y \mapsto 2\pi y$ is just for convenience, as we find that F must satisfy the ordinary differential equation

$$F''(y) + (\lambda y^{-2} - 1)F(y) = 0.$$

We see that as $y \to \infty$, this differential equation is essentially $F''(y) = F(y)$ and so we expect two linearly independent solutions whose asymptotic behavior in y is given by e^y or e^{-y}. Indeed, this differential equation is well-studied and has solutions

$$(2\pi^{-1}y)^{1/2} K_{s-1/2}(y) \sim e^{-y} \quad \text{and} \quad (2\pi y)^{1/2} I_{s-1/2}(y) \sim e^y.$$

Here K and I are standard Bessel functions. For example, for $y > 0$, we have

$$K_s(y) = \frac{1}{2} \int_0^\infty e^{-y(t+t^{-1})/2} t^s \frac{dt}{t}.$$

For a proof of this fact, we refer the reader to Whittaker and Watson (1927), though their notation is slightly different than ours. If our initial guess $f(z) = e(x)F(2\pi y) = o(e^{2\pi y})$, this rules out the solution using the I Bessel function. In short, $f(z)$ is a multiple of

$$W_s(z) = 2y^{1/2} K_{s-1/2}(2\pi y) e(x),$$

the "Whittaker function." Similarly, if we let $f(z) = e(rx)F(2\pi y)$, we obtain as a solution

$$W_s(r; z) = 2y^{1/2} K_{s-1/2}(2\pi |r| y) e(rx).$$

An alternative method for finding these solutions of Δ is via averaging. If we want $f(z)$ to be periodic in x of period 1, we must verify that

$$f\left(\begin{pmatrix} 1 & u \\ 0 & 1 \end{pmatrix} z\right) = e(u) f(z) \quad \text{for all } u \in \mathbb{R}.$$

To arrange this property for an eigenfunction of Δ, we take the simplest eigenfunction y^s and average over the subgroup of such matrices:

$$\int_{-\infty}^{+\infty} \overline{e(u)} \operatorname{Im}\left(\begin{pmatrix} 1 & u \\ 0 & 1 \end{pmatrix} z\right)^s du,$$

except that we see immediately that these translation matrices leave $\mathrm{Im}(z)$ invariant, so we rectify this by setting

$$f(z) = \int_{-\infty}^{+\infty} \overline{e(u)}\, \mathrm{Im}\left(w_0 \begin{pmatrix} 1 & u \\ 0 & 1 \end{pmatrix} z \right)^s du, \quad \text{where} \quad w_0 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

After some elementary manipulation, we obtain

$$\begin{aligned} f(z) &= \int_{-\infty}^{+\infty} e(u)\, \mathrm{Im}\left(\frac{-1}{z-u} \right)^s du \\ &= e(x) y^{1-s} \int_{-\infty}^{+\infty} (1+t^2)^{-s} e(ty) dt = \pi^s \Gamma(s)^{-1} W_s(z). \end{aligned}$$

Though this is only initially absolutely convergent for $\mathrm{Re}(s) > 1/2$, the resulting identity gives an analytic continuation to the complex s-plane to an eigenfunction of Δ periodic in x. Putting together our observations so far, we have:

Proposition 8.1.1. *Let $f(z)$ be an eigenfunction of Δ with eigenvalue $\lambda = s(1-s)$ which satisfies*

$$f(z+m) = f(z) \quad \textit{for all } m \in \mathbb{Z},$$

and the growth condition $f(z) = o(e^{2\pi y})$ as $y \to \infty$. Then

$$f(z) = f_0(y) + \sum_{n \neq 0} a(n) W_s(n; z) \quad \textit{for } a(n) \in \mathbb{C},$$

where the constant term $f_0(y)$ is a linear combination of the two solutions appearing in (25). The series converges absolutely and uniformly on compact sets.

We say that a Maass form f is a "cusp form" with respect to Γ if these constant terms ($n = 0$) of $f(\sigma_i z)$ are 0 for each element σ_i mapping the cusp x_i of Γ to ∞.

Does the Maass form possess similar analytic properties to those of classical automorphic forms? Yes. But in order to build an L-function from a Maass form f, we note that its expansion in terms of Whittaker functions has entries at all integers, not just non-negative ones. To remedy this, consider the involution

$$\iota(x+iy) = -x + iy.$$

According to the form of Δ, if $f(z)$ is an eigenfunction, so is $f \circ \iota(z)$. Moreover, the eigenvalues of ι acting on this space of functions must be ± 1 as $\iota^2 = 1$ and the space of Maass forms having fixed eigenvalue may be diagonalized with respect to

this involution. We say that the resulting forms are "even" if $f \circ \iota = f$ and "odd" if $f \circ \iota = -f$. Even Maass forms have Whittaker expansions in which $a(n) = a(-n)$ for all $n \in \mathbb{Z}$. (Similarly, odd Maass forms have $a(n) = -a(-n)$.) In the following, we restrict our attention to even and odd Maass forms.

Given a Maass form f, even or odd, then similar to our construction with modular forms we may form its L-function. Let $a_f(n) := a(n)$ be the coefficients of the Whittaker expansion of f and define

$$L(w, f) = \sum_{n=1}^{\infty} a(n) n^{-w}.$$

We use w as the complex variable to avoid serious confusion with the spectral parameter s.

Proposition 8.1.2. *Let f be a Maass cusp form with respect to Γ containing the inversion S. Write the eigenvalue $\lambda = s(1-s)$. The function $L(w, f)$ is initially absolutely convergent for $\mathrm{Re}(w) > 3/2$. Further, if we define the completed L-function:*

$$L^*(w, f) = \pi^{-w} \Gamma\left(\frac{w + (s - 1/2) + \epsilon}{2}\right) \Gamma\left(\frac{w - (s - 1/2) + \epsilon}{2}\right) L(w, f), \text{ where } \epsilon = \begin{cases} 0 & f \text{ even,} \\ 1 & f \text{ odd.} \end{cases}$$

Then $L^*(w, f)$ has analytic continuation to all $w \in \mathbb{C}$ and satisfies the functional equation

$$L^*(w, f) = (-1)^{\epsilon} L^*(1 - w, f).$$

Proof. We first show that $a(n) = O(n^{1/2})$ from which the initial domain of absolute convergence follows.

$$\left|a(n) 2\sqrt{y} K_{s-1/2}(2\pi |n| y)\right| = \left| \int_0^1 f(x + iy) e^{-2\pi i r x} \, dx \right| \leq \int_0^1 |f(x + iy)| \, dx.$$

But this latter expression is bounded by an absolute constant C since f is cuspidal (hence of rapid decay as z approaches any cusp) and so bounded on the fundamental domain, hence bounded by automorphicity. This expression is valid for any n independent of y, so we may choose $y = 1/n$ and $a(n) = O(n^{1/2})$ follows.

Now suppose f is an even Maass form. Consider the (shifted) Mellin transform

$$\int_0^{\infty} f(iy) y^{w - 1/2} \frac{dy}{y}.$$

This defines an absolutely convergent function for all w, since the K-Bessel function guarantees the integrand is small as $y \to \infty$ and the transformation property $f(iy) = f(i/y)$ gives the convergence for $y \to 0$. Substituting the Fourier expansion for $f(iy)$ into the integrand, we use the identity

$$\int_0^\infty K_{s-1/2}(y) y^w \frac{dy}{y} = 2^{w-2} \Gamma\left(\frac{w+s-1/2}{2}\right) \Gamma\left(\frac{w-s+1/2}{2}\right),$$

to show that the integrand equals $L^*(w, f)$. The reader should keep careful track of powers of 2 here, as one comes from using $a(n) = a(-n)$. The integral identity follows from substituting the integral definition of the K-Bessel function into the left-hand side to produce:

$$\frac{1}{2} \int_0^\infty \int_0^\infty e^{-(t+t^{-1})y/2} t^{s-1/2} y^w \frac{dy}{y} \frac{dt}{t}.$$

Then perform the change of variables $(u, v) = (ty/2, t^{-1}y/2)$. Noting that $\frac{du}{u} \wedge \frac{dv}{v} = 2\,\frac{dy}{y} \wedge \frac{dt}{t}$, we obtain

$$2^{w-2} \int_0^\infty \int_0^\infty e^{-u-v} u^{(w+s-1/2)/2} y^{(w-s+1/2)/2} \frac{du}{u} \frac{dv}{v},$$

from which we can read off the equality of Gamma functions. Finally, the functional equation then follows from the transformation property $f(iy) = f(i/y)$. (Here we've assumed that the group contains the inversion $S = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$ to guarantee invariance of this form. In general, we require some inversion for this to be true. For example, on $\Gamma_0(N)$ we would have $f(iy) = \pm f(i/Ny)$.)

To handle the case of f odd, the same Mellin transform doesn't produce a Dirichlet series indexed by positive integers as we can no longer use the identity $a(n) = a(-n)$ in the Whittaker expansion. Instead, we take a Mellin transform in $w + 1/2$ of $g(iy)$ where $g(z) = \frac{1}{4\pi i} \frac{\partial f}{\partial x}(z)$. We leave the details to the reader. □

Two immediate questions arise. Do Maass forms exist? And if so, do Maass cusp forms exist? The first question is easily answered by our familiar procedure of averaging, this time over a discrete subgroup of $\mathrm{SL}(2, \mathbb{R})$ rather than via a continuous integral. Starting with the simplest eigenfunction y^s, we define the (spectral) Eisenstein series as an average over all translates of Γ:

$$E(z, s) = \sum_{\gamma \in \Gamma_\infty \backslash \Gamma} \mathrm{Im}(\gamma(z))^s.$$

Note that just as with classical modular forms, we sum over a quotient of Γ after modding out by the (infinite) stabilizer of the cusp at ∞. As usual, there is an Eisenstein series defined for each cusp as in Section 4.6, and by writing the above we're assuming that ∞ is one such cusp. Using the same estimates as in the holomorphic case (e.g., Problem 2 of PSet 1), we see that $E(z,s)$ converges absolutely for $\mathrm{Re}(s) > 1$. The averaging makes clear that $E(z,s)$ is automorphic with respect to Γ.

Because Δ commutes with the action of $\mathrm{SL}(2,\mathbb{R})$, each of the summands $\mathrm{Im}(\gamma(z))^s$ is an eigenfunction of Δ as well. So we have

$$\Delta E(z,s) = s(1-s)E(z,s) \quad \text{for } \mathrm{Re}(s) > 1.$$

One of Maass' great achievements was a proof (1949) of the analytic continuation of Eisenstein series. One way of proceeding is to analyze the Fourier expansion in x (or rather, Whittaker expansion). The exponential decay of the K-Bessel functions ensures that the non-constant terms are well behaved. Thus it remains to analyze the constant term.

For example, if $\Gamma = \mathrm{SL}(2,\mathbb{Z})$, then

$$E(z,s) = \sum_{\gamma \in \Gamma_\infty \backslash \mathrm{SL}(2,\mathbb{Z})} \Im(\gamma(z))^s = y^s \sum_{\substack{(c,d)\in\mathbb{Z}^2 \\ \gcd(c,d)=1}} \frac{1}{|cz+d|^{2s}}$$

and we may compute

$$a_0(y,s) = \int_0^1 E(x+iy,s)\,dx = 2y^s + 2\pi^{2s-1}\frac{\Gamma(1-s)\zeta(2-2s)}{\Gamma(s)\zeta(2s)}y^{1-s}.$$

The term $2y^s$ comes from the pairs $(c,d) = (0,\pm 1)$, and the latter term requires some careful manipulation after change of variables. The form suggests that we can remove many poles (coming from zeros of the zeta function) by considering the "normalized" Eisenstein series

$$E^*(z,s) = \frac{1}{2}\pi^{-s}\Gamma(s)\zeta(2s)E(z,s).$$

Then the analytic continuation of the constant term follows from the analytic continuation of the Riemann zeta function. We also note the striking fact that the constant term is invariant under $s \mapsto 1-s$ according to the functional equation for the zeta function. This functional equation persists in the non-constant terms. Indeed, one may show that for the normalized Eisenstein series,

$$a(n;y,s) = 2|n|^{s-1/2}\sigma_{1-2s}(|n|)\sqrt{y}K_{s-1/2}(2\pi|n|y),$$

where σ_{1-2s} is the divisor function with exponents $1-2s$. From the integral definition of the K-Bessel function, we see that $K_\nu = K_{-\nu}$. Together with the simple functional equation for the divisor function, this implies

$$E^*(z,s) = E^*(z,1-s).$$

For proofs of these facts, see Section 1.6 of Bump's "Automorphic Forms and Representations."

8.2 Automorphic forms on groups – rough version

We've been working with automorphic forms as functions on $\mathcal{H} \simeq G/K$ where $G = \mathrm{SL}(2,\mathbb{R})$ and $K = \mathrm{SO}(2,\mathbb{R})$, the maximal compact subgroup of G. This identification was realized by considering the G-orbit of $z = i$ in $\mathcal{H}$, which had stabilizer K. Using this, we can lift an automorphic form f of weight m defined as a function on $\mathcal{H}$ to a function $\tilde{f} : G \to \mathbb{C}$ defined by

$$\tilde{f}(g) = j(g,i)^{-m} f(g(i)).$$

The point i is fixed by K and so the co-cycle condition on $j(g,z) = cz + d$ reduces for $k, k' \in K$ to

$$j(kk',i) = j(k,i)j(k',i).$$

Thus the map $\chi : k \mapsto j(k,i)$ is a character of K. Then $\tilde{f}$ inherits the following properties from f, an automorphic form of weight m with respect to a discrete subgroup Γ in G:

- $\tilde{f}(\gamma g) = \tilde{f}(g)$ for all $\gamma \in \Gamma$.
- $\tilde{f}(gk) = \chi(k)^{-m}\tilde{f}(g)$ for all $k \in K$.
- There exists a polynomial P in $\mathcal{C}$ – the Casimir operator on G – such that

$$P(\mathcal{C})\tilde{f} = \left(\frac{m^2}{2} - m\right)\tilde{f}.$$

- A suitable growth condition.

The first two conditions are clear from the definition of the lifting. The latter two require further explanation, which will only be completed after reviewing a few basics of Lie theory. For now, we only note that $\mathcal{C}$ is a second-order differential operator

closely related to Δ on $\mathcal{H}$. Further, the second and third conditions are very natural from the point of view of representation theory, and correspond to K-finite and $\mathcal{Z}$-finite vectors in certain representation spaces. (Here $\mathcal{Z}$ is the center of the universal enveloping algebra $\mathcal{U}(\mathrm{Lie}(G))$.)

In what follows, G will always denote $\mathrm{SL}(2,\mathbb{R})$ and $K = \mathrm{SO}(2)$. Our approach is to give proofs which generalize (with very little change) to the case of arbitrary Lie groups and beyond. We want the notation to suggestively reflect this. Our proofs will closely follow Borel's book "Automorphic forms on $\mathrm{SL}(2,\mathbb{R})$," (Cambridge Tracts in Math., v. 130).

8.3 Basic Lie theory for $\mathrm{SL}(2,\mathbb{R})$

In this section, we explain that the universal enveloping algebra of $\mathfrak{g} = \mathrm{Lie}(G)$, denoted $\mathcal{U}(\mathfrak{g})$, may be identified with the ring of left-invariant differential operators on G. Then we describe a distinguished element $\mathcal{C}$ of $\mathcal{U}(\mathfrak{g})$ and relate it to Δ on $\mathcal{H}$. A standard reference for this material is Chapter 3 of Warner's book "Foundations of Differentiable Manifolds and Lie Groups" or Bump's book "Lie Groups."

Recall that for $G = \mathrm{SL}(2,\mathbb{R})$, its Lie algebra $\mathfrak{g}$ is given by

$$\mathfrak{g} = \{M \in \mathrm{Mat}(2,\mathbb{R}) \mid \mathrm{tr}(M) = 0\}.$$

It has standard basis given by the triple H, E, F, where

$$H = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad E = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad F = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

satisfying the commutator relations

$$[H, E] = 2E, \quad [H, F] = -2F, \quad [E, F] = H.$$

To every $Y \in \mathfrak{g}$, we may associate a one-parameter subgroup of G defined by the exponential

$$t \mapsto e^{tY} = \sum_{n=0}^{\infty} \frac{t^n Y^n}{n!} \quad \text{for } t \in \mathbb{R}.$$

The group G acts on functions of G by right and left translation as follows:

$$l_g(f(x)) = f(g^{-1}x) \quad \text{and} \quad r_g(f(x)) = f(xg).$$

Each element $Y \in \mathfrak{g}$ may be identified with a left-invariant derivation on G defined by

$$Yf(x) = \frac{d}{dt} f(xe^{tY})\Big|_{t=0} \quad \text{where } f \in C^\infty(G), x \in G, Y \in \mathfrak{g}.^6$$

As an algebra over $\mathbb{C}$, the Y generate the left-invariant differential operators on G (including higher order operators). It is isomorphic to the universal enveloping algebra $\mathcal{U}(\mathfrak{g})$ of $\mathfrak{g}$. The Poincaré-Birkhoff-Witt theorem states that for any basis A, B, C for $\mathfrak{g}$, the monomials $A^m B^n C^p$ (for $m, n, p \in \mathbb{N}$) form a vector space basis of $\mathcal{U}(\mathfrak{g})$. Thus any left-invariant differential operator is a finite linear combination of monomials $X_1 \cdots X_s$ with $X_i \in \mathfrak{g}$. The action of such a monomial operator on f is given by

$$X_1 \cdots X_s f(x) = \frac{d^s}{dt_1 \cdots dt_s} f(xe^{t_1 X_1} \cdots e^{t_s X_s})\Big|_{t_1 = \cdots = t_s = 0}.$$

Given a representation $\pi : \mathfrak{g} \to \mathrm{End}(V)$, we say that a bilinear form B on V is invariant if

$$B(\pi(Y)v, w) + B(v, \pi(Y)w) = 0, \quad \text{for all } Y \in \mathfrak{g} \text{ and } v, w \in V.$$

Recall that the adjoint representation $\mathrm{ad} : \mathfrak{g} \to \mathrm{End}(\mathfrak{g})$ is defined by $\mathrm{ad}(X)Y = [X, Y]$. The Killing form $B(X, Y) = \mathrm{tr}(\mathrm{ad}(X) \circ \mathrm{ad}(Y))$ is symmetric and invariant with respect to ad (see Proposition 10.3 of Bump).

Theorem 8.3.1 (Theorem 10.2 of Bump)**.** *Suppose that $\mathfrak{g}$ admits a non-degenerate symmetric invariant bilinear form B. Given a basis $X_1, \ldots, X_d$ of $\mathfrak{g}$, let $X_1', \ldots, X_d'$ be a dual basis with respect to B – that is, $B(X_i, X_j') = \delta_{i,j}$. Then the element*

$$\mathcal{C} \stackrel{def}{=} \sum_{i=1}^{d} X_i X_i' \in \mathcal{U}(\mathfrak{g})$$

is in the center of the universal enveloping algebra.

Taking B to be the Killing form, we may represent $\mathcal{C}$, called the "Casimir element" of $\mathcal{U}(\mathfrak{g})$, in terms of the standard basis H, E, F as follows:

$$\mathcal{C} = \frac{1}{2}H^2 + EF + FE = \frac{1}{2}H^2 + H + 2FE = \frac{1}{2}H^2 - H + 2EF.$$

[6]The usual definition of $\mathfrak{g} = \mathrm{Lie}(G)$ is the set of left-invariant vector fields on G. Since vector fields and global derivations are in 1-to-1 correspondence on smooth manifolds, the left-invariant derivations on G are given by elements of $\mathfrak{g}$. (See Proposition 6.3 of Bump's "Lie Groups" and the discussion on p. 42 of the book for the equivalence with the matrix definition.)

One can check explicitly that it lies in the center of $\mathcal{U}(\mathfrak{g})$ by showing it commutes with the generators H, E, F via commutation relations. The center $\mathcal{Z}$ of the universal enveloping algebra $\mathfrak{g}$ can be shown to be equal to the polynomial algebra $\mathbb{C}[\mathcal{C}]$ in the Casimir operator $\mathcal{C}$.

Just as the Lie algebra $\mathfrak{g}$ defines differential operators on G, it also defines differential operators on any homogeneous space $\mathcal{H}$ for G by

$$\frac{d}{dt} f(e^{tY} \cdot z)\Big|_{t=0} \quad \text{for } f \in C^\infty(\mathcal{H}), z \in \mathcal{H}, Y \in \mathfrak{g}.$$

In particular, if $\mathcal{H}$ is the upper half-plane and we express $e^{tY} \cdot z = x(t) + iy(t)$, then in operator notation, we act on functions f by

$$\frac{dx(t)}{dt}\Big|_{t=0} \frac{\partial}{\partial x} + i \, \frac{dy(t)}{dt}\Big|_{t=0} \frac{\partial}{\partial y}.$$

Note that since we're evaluating derivatives at $t = 0$, it suffices to compute $(I+tY)\cdot z$, using the first two terms in the power series expansion for e^{tY}.

To calculate the action of the Casimir element $\mathcal{C}$ on functions on $\mathcal{H}$, we compute the action of the generators H, E, F. For example,

$$E = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \; : \quad (I + tE)z = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} z = z + t.$$

Thus if $z = x + iy$, then $x(t) = x + t$ and $y(t) = y$. Hence, $dx/dt = 1$ and $dy/dt = 0$ for all t and therefore E acts by $\partial/\partial x$. Repeating this for F and H (which are slightly harder but follow similarly), we obtain

$$E = \frac{\partial}{\partial x}, \quad F = (y^2 - x^2)\frac{\partial}{\partial x} - 2xy\frac{\partial}{\partial y}, \quad H = 2\left(x\frac{\partial}{\partial x} + y\frac{\partial}{\partial y}\right).$$

Using any of the equivalent expressions for the Casimir element, we find

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

8.4 K-finite and $\mathcal{Z}$-finite functions

The group $K = \mathrm{SO}(2, \mathbb{R})$ may be identified with the circle group $S^1 = \{e^{i\theta}\}$. So we let k_θ be the element of K corresponding to θ under this isomorphism. The characters

of K are its one-dimensional representations, which form a group $\hat{K}$ isomorphic to $\mathbb{Z}$. To each $m \in \mathbb{Z}$, we have the corresponding character

$$\chi_m : k_\theta \mapsto e^{im\theta}.$$

Any continuous finite-dimensional representation of K is a direct sum of 1-dimensional ones. We say that a function f on G has (right) K-type m if, for all $x \in G, k \in K$, we have

$$f(xk) = \chi_m(k) f(x).$$

Definition 8.2 (K-finiteness)**.** *A function f is said to be (right) K-finite if the set of all (right) translations $r_k f$ by elements $k \in K$ spans a finite-dimensional vector space.*

There are, of course, analogous definitions for left translations. Any such K-finite function may then be expressed as a finite sum:

$$f = \sum_i f_i, \quad \text{where each } f_i \text{ has } K\text{-type } m_i, \text{ for some integer } m_i.$$

Definition 8.3 ($\mathcal{Z}$-finiteness)**.** *A locally integrable function f on G is said to be $\mathcal{Z}$-finite if it is annihilated by an ideal J of finite codimension in $\mathcal{Z}$.*

We explain the extension to locally integrable functions (or even distributions) in the next section (See Remark 8.1). For now, the reader may consider f to be smooth on G.

Since $\mathcal{Z}$ is generated as a polynomial algebra by $\mathcal{C}$, this just means that the sequence $\{\mathcal{C}^n f\}$ is contained in a finite-dimensional vector space. Equivalently, such an f is $\mathcal{Z}$-finite if there exists a non-constant polynomial in one-variable P such that $P(\mathcal{C})f = 0$.

8.5 Distributions and convolution of functions

We follow Harish-Chandra in making systematic use of convolution with smooth functions of compact support. These arguments may be easily carried over to the case of a general reductive group.

Given two locally integrable functions f, g on G (i.e., integrable on any compact set in their domain), their convolution is defined by

$$(f * g)(x) = \int_G f(xy) g(y^{-1})\, dy = \int_G f(y) g(y^{-1}x)\, dy,$$

whenever the integral exists. This is guaranteed if one of the functions has compact support. Recall that convolution acts like multiplication on spaces of functions. However, there is generally not a function in the space that acts as the multiplicative identity.

For example, consider $L^1(\mathbb{R})$, the space of integrable functions on the real line. An elementary exercise in Fubini's theorem shows that if $f, g \in L^1(\mathbb{R})$, then $f*g \in L^1(\mathbb{R})$ and its Fourier transform $(f * g)\hat{} = \hat{f}\hat{g}$. However, $L^1(\mathbb{R})$ does not have an identity element under convolution.[7]

However, if we restrict to a special class of functions, then there are convolution operators that act as the identity. For example, if we take 1_B to be the characteristic function of the unit "ball" in $\mathbb{R}$ centered at the origin (divided by its volume 2), then convolution gives

$$(f * 1_B)(x) = \int_{\mathbb{R}} f(y) 1_B(x-y)\, dy = \frac{1}{2}\int_{x-1}^{x+1} f(y)\, dy.$$

Thus 1_B acts as the identity if f is equal to its average value over the integral. This is true if f is harmonic – that is, an eigenfunction of the Laplacian! We will seek to prove a similar result for $G = \mathrm{SL}(2,\mathbb{R})$ in the case that f is $\mathcal{Z}$-finite (and K-finite, which will further restrict the choice of identity element.)

To carry this out, it is useful to have the language of distributions, sometimes referred to as "generalized functions." The basic idea is to reinterpret a function f as a linear functional T_f on a collection of test functions, often $C_c^\infty(G)$ – smooth functions on G having compact support. In particular, given an locally integrable function f, we have the very natural functional (convolution evaluated at the identity of G):

$$T_f(\phi) = \int_G f(y)\phi(y)\, dy \quad \text{for any } \phi \in C_c^\infty(G).$$

But we don't want to restrict just to functionals arising this way. In general, we allow for any continuous linear functional on a collection of test functions (appropriately topologized). See p. 136 of Rudin's "Functional Analysis" for details in the case $G = \mathbb{R}^n$ and Warner, p. 41 for the generalization to manifolds. The classic example not of the above form is the Dirac δ distribution, for which $\delta(f) = f(0)$.

[7] If u were such a unit element, then we could take a function f such that $\hat{f}$ is non-vanishing, e.g., $f(x) = e^{-2\pi|x|}$. Then $||\hat{f} - \hat{f}\hat{u}||_{L^\infty(\hat{\mathbb{R}})} \leq ||f - f*u||_{L^1(\mathbb{R})} = 0$, which implies $\hat{u} \equiv 1$ on $\hat{\mathbb{R}}$. This contradicts the Riemann-Lebesgue lemma: For any $u \in L^1(\mathbb{R})$, $\hat{u}(x) \to 0$ as $x \to \infty$. See Theorem 1.4.1 of Benedetto's "Harmonic Analysis and Applications."

We may also define the convolution of a distribution with a test function. We rewrite the definition of convolution of two functions suggestively as

$$(f * g)(x) = \int_G f(y)g(y^{-1}x)\,dy = \int_G f(y)(r_x g^-)(y)\,dy, \quad \text{where } g^-(y) = g(y^{-1}).$$

This is just $T_f(r_x g^-)$, using our earlier notation. This suggests that given any distribution D, we define

$$(D * g)(x) = D(r_x g^-)$$

Thus by identifying elements $Y \in \mathcal{U}(\mathfrak{g})$ with distributions supported at the origin[8], we have Y_r as a right-invariant operator and Y as a left-invariant differential operator acting as

$$Y_r f(x) = (Y * f)(x), \quad Yf(x) = (f * (-Y))(x).$$

The reader allergic to distributions can simply take this to be the definition of the convolution with an element of $\mathcal{U}(\mathfrak{g})$. Note, however, that in the simpler case of $L^1(\mathbb{R})$, our definition of convolution gives $\delta * f(x) = f(x)$, the desired identity in convolution (though of course δ is a functional, not an $L^1(\mathbb{R})$ function).

Here are three nice properties of the convolution operator, and its relation to elements $D \in \mathcal{U}(\mathfrak{g})$ viewed as a distribution.

1. (**Convolution is a smoothing operator**) If f is continuous and $g \in C_c^\infty(G)$, then $f * g \in C^\infty(G)$. Further, given $D \in \mathcal{U}(\mathfrak{g})$, then $D(f * g) = f * (Dg)$.

2. (**Convolution is associative**) If f, g, h are locally integrable and g, h have compact support, then

$$(f * g) * h = f * (g * h).$$

3. (**Convolution relation with $\mathcal{Z}$**) If f is smooth, $g \in C_c^\infty(G)$, $D \in \mathcal{Z}$, then

$$D(f * g) = (Df) * g.$$

The first two properties are straightforward from the definition. We give a proof of the third.

[8]See 2.2.7 of Benedetto for a definition of support of a distribution. It's essentially the complement of points where D acts as the 0 functional.

Proof of Property (3). Since $D \in \mathcal{Z}$, we have $((Df) * g)(x)$ is equal to

$$\int_G (Df)(xy)g(y^{-1})\,dy = \int_G r_y(Df)(x)g(y^{-1})\,dy = \int_G D(r_y f)(x)g(y^{-1})\,dy.$$

Since g was assumed to have compact support, we may take the integration be over a compact set of y such that $y^{-1} \in \mathrm{Supp}(g)$. Hence we may interchange the order of integration and convolution and the left-hand side of the above equation becomes

$$D\left(\int_G (r_y f)(x)g(y^{-1})\,dy\right) = D(f * g)(x)$$

□

Definition 8.4 (Dirac sequence)**.** *A sequence of functions $\{\alpha_n\}$ with $\alpha_n \in C_c^\infty(G)$ is called a Dirac sequence if $\alpha_n(x) \geq 0$ for all $x \in G, n \in \mathbb{N}$,*

$$\mathrm{Supp}(\alpha_n) \to 1 \text{ as } n \to \infty, \quad \text{and} \quad \int_G \alpha_n(x)\,dx = 1 \quad \text{for all } n \in \mathbb{N}$$

Proposition 8.5.1. *For any function $f \in C(G)$ and $D \in \mathcal{U}(\mathfrak{g})$, the sequence $\{Df * \alpha_n\}$ converges absolutely and uniformly on compact sets to Df.*

Proof. We must show that given any $D \in \mathcal{U}(\mathfrak{g})$, any compact set C in G and $\epsilon > 0$, there exists a j such that

$$|(Df * \alpha_j)(x) - Df(x)| \leq \epsilon \quad \text{for all } x \in C.$$

There exists an open, relatively compact neighborhood U of the identity (which we may assume to be symmetric under $y \mapsto y^{-1}$) such that

$$|Df(xy) - Df(x)| \leq \epsilon \quad \text{for all } x \in C, y \in U.$$

Now using the fact that the total integral of each α_j is equal to 1, we have for any j:

$$(Df * \alpha_j)(x) - Df(x) = \int_G Df(xy^{-1})\alpha_j(y)\,dy - \int_G Df(x)\alpha_j(y)\,dy,$$

and thus (since $\alpha_j \geq 0$)

$$|(Df * \alpha_j)(x) - Df(x)| \leq \int_G |Df(xy^{-1}) - Df(x)|\alpha_j(y)\,dy. \tag{26}$$

Now using the shrinking support of α_n as $n \to \infty$, we choose j large enough so that $\mathrm{Supp}(\alpha_j) \subset U$. Thus the left-hand side of (26) is

$$\int_U |Df(xy^{-1}) - Df(x)|\alpha_j(y)\,dy \leq \epsilon \int_U \alpha_j(y)\,dy = \epsilon,$$

where we've used the symmetry of U to conclude the inequality. □

Theorem 8.5.2. *Let f be $\mathcal{Z}$-finite and K-finite on one side. Then f is smooth.*

Remark 8.1. Generally, we take f to be smooth in the first place, but it is quite interesting to note that these properties imply smoothness. Recall that K-finiteness makes sense for any function on G, and that $\mathcal{Z}$-finiteness is defined for locally integrable functions if meant in the sense of distributions. That is, if f is a locally integrable distribution, then $\mathcal{C}^n f$ in the sense of distributions is the functional

$$\varphi \mapsto \int_G f(x)\mathcal{C}^n\varphi(x)\,dx \quad \text{where } \varphi \in C_c^\infty(G), n \in \mathbb{N}.$$

We say f is $\mathcal{C}$ finite if the distributions ranging over all n span a finite dimensional complex vector space of distributions.

Proof. The $\mathcal{Z}$-finiteness guarantees the existence of a (monic) polynomial P in $\mathcal{C}$ such that $P(\mathcal{C})(f) = 0$. We want to show that this, together with K-finiteness, imply that f is annihilated by $Q(\Omega)$ for some polynomial Q in an elliptic operator Ω. Then the result will follow from the elliptic regularity theorem.[9]

We manufacture such an elliptic operator Ω by rewriting $\mathfrak{g}$ in terms of the natural basis coming from the Cartan decomposition. Let $\mathfrak{s}_0$ be the space of 2×2 symmetric real matrices of trace 0. Let $S_0 = \exp(\mathfrak{s}_0)$ – positive, non-degenerate symmetric matrices of determinant 1. Then (as a special case of the Cartan decomposition for semisimple groups) we may write

$$G = S_0K: \quad g = sk \quad \text{where } s = (g^t g)^{1/2}.$$

The Lie algebra $\mathrm{Lie}(K) = \mathfrak{k}$ is spanned by the matrix

$$W = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} = E - F,$$

[9]Recall that a (ℓth-order) differential operator is said to be elliptic if the top-degree terms $\sum_{[\alpha]=\ell} a_\alpha D^\alpha$ are non-zero upon substituting any point $\mathbf{x}^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$. Remember that the coefficients a_α are themselves complex-valued functions on $\mathbb{R}^n$, say $\mathbf{y} = (y_1, \ldots, y_n)$, and ellipticity means that we check this non-singularity at all $\mathbf{x}$ for each $\mathbf{y}$. See page 240 of Warner.

while the Lie algebra $\mathfrak{s}_0$ is spanned by

$$H \quad \text{and} \quad Z = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = E + F.$$

With respect to the basis H, Z, W, we see that

$$\mathcal{C} = \frac{1}{2}(H^2 + Z^2 - W^2),$$

which makes it clear that though $\mathcal{C}$ is not an elliptic differential operator,

$$\Omega = \mathcal{C} + W^2 = \frac{1}{2}(H^2 + Z^2 + W^2),$$

is elliptic.

It suffices to assume that f is of K-type m for some integer m (as the general K-finite function is just a finite sum of these). In this case, we see that $W(f) = i \cdot mf$:

$$Wf(x) = \frac{d}{dt}(f(xe^{tW}))\Big|_{t=0} = \frac{d}{dt}(f(x)e^{imt})\Big|_{t=0} = imf(x).$$

Thus $W^2 f = -m^2 f$. If f is of type m, then factoring $P(\mathcal{C}) = \prod(C - \lambda_i)$, we have

$$P(\mathcal{C})f = 0 \Longleftrightarrow \prod_{i=1}^{n}(\mathcal{C} + W^2 - \lambda_i + m^2)f = 0.$$

But this latter operator is just a polynomial in $\Omega = \mathcal{C} + W^2$, hence f is smooth (because 0 is smooth). For a proof of the regularity theorem, see pp. 227–249 in Warner. □

Using a stronger form of the regularity theorem, we can prove that such an f is in fact analytic (remembering that f is a function on the real manifold G), and we use this stronger version in the proof of the main theorem of the next section. For this, one needs Theorem 7.5.1 (p. 178) of Hörmander's "Linear Partial Differential Operators."

8.6 A convolution identity for $\mathcal{Z}$-finite, K-finite functions

Let

$$I_c^\infty(G) = \{\alpha \in C_c^\infty(G) \mid \alpha(xk) = \alpha(kx), k \in K, x \in G\}.$$

The significance of this set is that if f is of K-type m on the right, so is $f * \alpha$ for any $\alpha \in I_c^\infty(G)$. We refer to such functions α as K-invariant.

Theorem 8.6.1 (Harish-Chandra)**.** *Let f be $\mathcal{Z}$-finite and K-finite. Given a neighborhood U of the identity, there exists a function $\alpha \in I_c^\infty(G)$ with* $\mathrm{Supp}(\alpha)$ *in U such that*

$$f * \alpha = f.$$

Proof. First suppose that f has K-type m. Let V be the smallest G-invariant closed subspace of $C^\infty(G)$ containing $\mathcal{U}(\mathfrak{g}) \cdot f$. Then we will show that:

1. $P(\mathcal{C})$, the polynomial that annihilates f, annihilates all $g \in V$.
2. For each character $\chi = \chi_n$, the subspace

$$V_\chi = \{v \in V \mid v \text{ is of } K\text{-type } n\}$$

   is finite dimensional for every $\chi \in \hat{K}$.

Momentarily assuming these facts, we now complete the proof of the theorem. In general, f will be a finite sum of functions with K-types m_i. The above facts imply that there exists a finite-dimensional subspace L of $C^\infty(G)$ containing f and stable under convolutions $*\alpha$ for all $\alpha \in I_c^\infty(G)$. In particular, L contains $f * \alpha$ for all α and so the convolutions $*\alpha$ (restricted to L) form a finite dimensional subspace W of $\mathrm{End}(L)$.

Let $\{\alpha_i\} \in I_c^\infty(G)$ be a Dirac sequence. Then $g * \alpha_i \to g$ for any $g \in V$. Hence $*\alpha_i$ tends to the identity endomorphism. Thus, the identity endomorphism is in the closure of W. But since W is finite-dimensional, the space is closed and contains the identity endomorphism. That is, it is realized by convolution with some $\alpha \in I_c^\infty(G)$. □

Before proving the two required properties, we need the following result.

Lemma 8.6.2. *Let $f \in C^\infty(G)$ be of right K-type $\chi = \chi_m$. Then for any $D \in \mathcal{U}(\mathfrak{g})$, Df is of finite K-type on the right.*

Proof. It suffices to show this for $D = X_1 \cdots X_s$ with $X_i \in \mathfrak{g}$, as any element of $\mathcal{U}(\mathfrak{g})$ is a finite sum of such elements. Recall that

$$(r_k Df)(x) = (Df)(xk) = \frac{d^s}{dt_1 \cdots dt_s} f(xke^{t_1 X_1} \cdots e^{t_s X_s})\Big|_{t_1 = \cdots = t_s = 0}. \tag{27}$$

But

$$f(xke^{t_1 X_1} \cdots e^{t_s X_s}) = f(xe^{t_1 {}^k X_1} \cdots e^{t_s {}^k X_s} k) = \chi_m(k) f(xe^{t_1 {}^k X_1} \cdots e^{t_s {}^k X_s}),$$

where we've used that $ge^{tX}g^{-1} = e^{t\,{}^gX}$ for any $g \in G$. (Here ${}^gX$ denotes $gXg^{-1} = \mathrm{Ad}\, g(X)$.) Thus we may rewrite (27) in the form

$$(r_k Df)(x) = \chi_m(k)({}^kX_1 \cdots {}^kX_s)f(x).$$

Now if A, B, C is a basis for $\mathfrak{g}$, then ${}^kX_1 \cdots {}^kX_s$ is a linear combination of products of s factors, each equal to one of A, B, or C. (The coefficients in the linear combination depend on k.) Thus there exist finitely many elements $D_1, \ldots, D_N \in \mathcal{U}(\mathfrak{g})$ such that $r_k Df$ is a linear combination (with constant coefficients) of $D_i f$, $i = 1, \ldots, N$ for each $k \in K$. □

Theorem 8.6.3. *Let $f \in C_c^\infty(G)$ be $\mathcal{Z}$-finite such that $P(\mathcal{C})$ is a monic polynomial in $\mathcal{C}$ annihilating f. Further, let f be of K-type m. Then*

1. *The closure of $\mathcal{U}(\mathfrak{g}) \cdot f$ in $C^\infty(G)$ (with respect to the C^∞ topology) is a G-invariant subspace V.*

2. *For every $\chi = \chi_n \in \hat{K}$, the space*

$$\{V_n = \{v \in V \mid r_k v = \chi_n(k)v, k \in K\}$$

is finite dimensional.

The theorem gives the two ingredients necessary to complete the proof of Theorem 8.6.1. In particular, the subspace V is the smallest closed subspace of $C^\infty(G)$ containing $\mathcal{U}(\mathfrak{g}) \cdot f$. Moreover, since $P(\mathcal{C})$ commutes with $\mathcal{U}(\mathfrak{g})$, then every element of $\mathcal{U}(\mathfrak{g}) \cdot f$ is annihilated by $P(\mathcal{C})$ and hence, so is every element of V.

Proof of 1. Let U be the smallest G-invariant closed subspace of $C^\infty(G)$ containing $\mathcal{U}(\mathfrak{g}) \cdot f$. It contains V and of course we must show $U = V$. By the Hahn-Banach theorem, it suffices to show that for any continuous linear functional b on U that vanishes on V and any $v \in V$, b vanishes on $r_g v$ for any $g \in G$. (Indeed, there are two possibilities. Either the space V is G-invariant or there exists a $g \in G$ and $v \in V$ such that $r_g v$ is in $U - V$. But then the Hahn-Banach theorem guarantees that there exists a functional extending b on V that has value, say, equal to 1 at an element of $U - V$, so this contradicts that all such b vanish. See Corollary to Theorem 3 in Chapter IV of Yosida's "Functional Analysis" for the appropriate corollary to the Hahn-Banach theorem.) Equivalently, given any such b and $v \in V$, we must show that the map

$$\varphi : g \mapsto b(r_g v)$$

is identically 0. We do this by showing that φ is $\mathcal{Z}$-finite and K-finite, hence analytic, and has all derivatives vanishing at the identity.

The $\mathcal{Z}$-finiteness is clear, since $\mathcal{C}$ commutes with right translations, so $P(\mathcal{C})$ annihilates φ because it annihilates v. For K-finiteness, we use the previous lemma to assert that given any $v \in V$, there exists $v_1, \dots v_s \in \mathcal{U}(\mathfrak{g}) \cdot f$ such that $r_k v$ is a linear combination of the v_i. Hence, φ is a linear combination of $g \mapsto b(r_g v_i)$ and thus also K-finite. Using the strong form of the elliptic regularity theorem, this implies φ is analytic.

Further, for any $D \in \mathcal{U}(\mathfrak{g})$, we have $(D\varphi)(1) = b(Dv) = 0$, since b was assumed to vanish on V. Since $\mathcal{U}(\mathfrak{g})$ consists of all higher derivatives at the identity, it follows that φ vanishes for any $g \in G$, i.e., $b(r_g v)$ vanishes for all g. □

The proof of part 2 requires some additional discussion. In particular, we give a brief reminder about the theory of complex $\mathcal{U}(\mathfrak{g})$-modules. Let $G_{\mathbb{C}} = \mathrm{SL}(2, \mathbb{C})$ with corresponding Lie algebra $\mathfrak{g}_{\mathbb{C}}$, the set of 2×2 complex matrices of trace 0. We take, as a natural basis of $\mathfrak{g}_{\mathbb{C}}$ the matrices gHg^{-1}, gEg^{-1} and gFg^{-1}, where H, E, F are as before and g is defined by the equation $gHg^{-1} = iW$. Let $gEg^{-1} =: Y$ and $gFg^{-1} =: Z$. Since Ad g is an automorphism of $\mathfrak{g}_{\mathbb{C}}$, these matrices Y, Z, iW continue to satisfy the same commutation relations as their counterparts E, F, H:

$$[iW, Y] = 2Y, \quad [iW, Z] = -2Z, \quad [Y, Z] = iW.$$

Since $\mathcal{U}(\mathfrak{g})$ is an algebra over $\mathbb{C}$, we may identify it with $\mathcal{U}(\mathfrak{g}_{\mathbb{C}})$. Just as before, via the Poincaré-Birkhoff-Witt theorem, it is spanned by products of the form $Y^a Z^b (iW)^c$ for $a, b, c \geq 0$.

Let M be a complex vector space that is a $\mathcal{U}(\mathfrak{g})$-module. We say that an element $v \in M$ is of weight λ (with respect to iW) such that $iW \cdot v = \lambda v$ with $\lambda \in \mathbb{C}$. The subspace of such elements is called the weight space, which we denote by M_λ. The effect of Y and Z on v is simple to describe:

$$\begin{aligned} (iW)Y = Y(iW) + 2Y \quad &\Longrightarrow \quad (iW)(Y \cdot v) = (\lambda + 2) Y \cdot v \\ (iW)Z = Z(iW) - 2Z \quad &\Longrightarrow \quad (iW)(Z \cdot v) = (\lambda - 2) Z \cdot v \end{aligned}$$

The following result is true for any semisimple Lie algebra $\mathfrak{g}$, though we continue to give proofs in the rank one case. For the general proof, see Theorem 1 of Harish-Chandra's "Representations of a semisimple Lie group on a Banach space, I."

Theorem 8.6.4. *Let $M = \mathcal{U}(\mathfrak{g}) \cdot v$ with v both $\mathcal{Z}$-finite and an eigenvector of iW of weight λ. Then M is a countable direct sum of finite-dimensional weight spaces M_μ for iW with $\mu \in \{\lambda + 2\mathbb{Z}\}$. In particular, if v is an eigenfunction of $\mathcal{C}$, then* $\dim(M_\mu) \leq 1$.

Proof. The line $\mathbb{C}\cdot v$ is stable under powers of (iW), so $M = \mathcal{U}(\mathfrak{g})\cdot v$ is spanned over $\mathbb{C}$ by elements of the form $Y^aZ^b\cdot v$ for $a, b \geq 0$. But from the action of Y and Z above, we see that M is a direct sum of weight spaces M_μ with $\mu \in \{\lambda + 2\mathbb{Z}\}$.

Now suppose that v is an eigenvector of $\mathcal{C}$, corresponding to the special case $P(\mathcal{C}) = \mathcal{C} - c$ for some constant c. We will show that the M_μ are at most one-dimensional. We begin by showing that if $x \in M_\mu$ then $Y^aZ^ax \in \mathbb{C}\cdot x$ for all $a \geq 0$. The proof is by induction.

If $a = 1$, then YZ is expressible as a polynomial Q in iW and $\mathcal{C}$. Indeed, $\mathcal{C} = -\frac{1}{2}W^2 - iW + 2YZ$ so

$$YZ = Q(iW, \mathcal{C}) = \frac{1}{2}\mathcal{C} + \frac{i}{2}W + \frac{1}{4}W^2 \Longrightarrow YZ\cdot x = Q(\mu, c)x.$$

Using the similar identity $\mathcal{C} = -\frac{1}{2}W^2 + iW + 2ZY$, we may show x is an eigenvector for ZY. Now suppose we have the result for a and consider $Y^{a+1}Z^{a+1} = Y(Y^aZ^a)Z$. First, $Z\cdot x \in M_{\mu-2}$ and by induction $Y^aZ^a(Z\cdot x)$ is a multiple of $Z\cdot x$. Thus $Y(Y^aZ^a)Z\cdot x$ is a multiple of $YZ\cdot x$ which is a multiple of x.

Now given any $a \geq b$, we have

$$Y^aZ^b\cdot v = Y^{a-b}Y^bZ^b\cdot v \in \mathbb{C}Y^{a-b}v \in M_{\lambda+2(a-b)}.$$

Similarly for $a \leq b$,

$$Y^aZ^b\cdot v = Y^aZ^aZ^{b-a}\cdot v \in \mathbb{C}Z^{b-a}v \in M_{\lambda+2(a-b)}.$$

Thus M, initially spanned by $Y^aZ^b\cdot v$ with $a, b \geq 0$, is in fact spanned by Y^mv and Z^mv for $m \geq 0$ with respective weights $\lambda + 2m$ and $\lambda - 2m$. Since each of these vectors has a distinct weight, the associated weight space is one-dimensional.

To finish the general case of a $\mathcal{Z}$-finite function, we give similar argument by induction. We have treated the case $P(\mathcal{C}) = \mathcal{C} - c$. First suppose the theorem is true for $P(\mathcal{C}) = (\mathcal{C} - c)^{a-1}$ and then we'll show it for $P(\mathcal{C}) = (\mathcal{C} - c)^a$.

Suppose that our vector v is annihilated by $(\mathcal{C} - c)^a$. Consider the short exact sequence of cyclic $\mathcal{U}(\mathfrak{g})$ modules (i.e., those generated by $\mathcal{U}(\mathfrak{g})$ translates of a single vector):

$$0 \longrightarrow \mathcal{U}(\mathfrak{g})\cdot(\mathcal{C} - c)v \longrightarrow M = \mathcal{U}(\mathfrak{g})\cdot v \xrightarrow{\phi} M/\mathcal{U}(\mathfrak{g})\cdot(\mathcal{C} - c)v.$$

The last of these modules on the right is cyclic since it's generated by the image of v under the last map ϕ with $\mathcal{C} - c$ annihilating $\phi(v)$. Similarly, $\mathcal{U}(\mathfrak{g})\cdot(\mathcal{C} - c)v$ is generated by $(\mathcal{C} - c)v$ which is annihilated by $(\mathcal{C} - c)^{a-1}$. Thus, by induction hypothesis, both cyclic modules have finite dimensional weight spaces.

We showed earlier that any cyclic module M may be decomposed into a direct sum of eigenspaces M_μ of iW. Our exact sequence clearly descends to an exact sequence on eigenspaces:

$$0 \longrightarrow (\mathcal{U}(\mathfrak{g}) \cdot (\mathcal{C} - c)v)_\mu \longrightarrow M_\mu \longrightarrow (M/\mathcal{U}(\mathfrak{g}) \cdot (\mathcal{C} - c)v)_\mu.$$

Since the outer spaces are finite dimensional, the space M_μ must be finite dimensional. This completes the induction.

For an arbitrary $P(\mathcal{C}) = \prod_{i=1}^{s} (C - c_i)^{a_i}$, let

$$P^{(i)}(\mathcal{C}) = P(\mathcal{C})(C - c_i)^{-a_i},$$

the polynomial with i-th factor removed. Since the $P^{(i)}(\mathcal{C})$ have no common factor, we may find $Q_i(\mathcal{C})$ such that

$$\sum_{i=1}^{s} Q_i(\mathcal{C}) P^{(i)}(\mathcal{C}) = 1.$$

Then setting $v_i = Q_i(\mathcal{C})P^{(i)}(\mathcal{C})v$ we have $\sum_i v_i = v$ and M is the sum of the submodules $M_i = \mathcal{U}(\mathfrak{g})v_i$. Thus it suffices to prove finite-dimensionality for the M_i's. But v_i is annihilated by $(C - c_i)^{a_i}$ (since $(C - c_i)^{a_i} Q_i(\mathcal{C}) P^{(i)}(\mathcal{C})v = Q_i(\mathcal{C})P(\mathcal{C})v = 0$). Thus M_i is finite dimensional by the previous induction argument. □

We are now almost ready to prove Part 2 of Theorem 8.6.3. There we were considering the closure V of $M = \mathcal{U}(\mathfrak{g}) \cdot f$, where f was $\mathcal{Z}$-finite and K-finite of type m. Considering the weights of iW instead of W, we see that $Wf = (im)f$ and the weights of W on $\mathcal{U}(\mathfrak{g}) \cdot f$ are in $i\mathbb{Z}$. In view of the previous theorem, it suffices to show that V_n as defined in Theorem 8.6.3 is equal to M_{in}.

Let dk denote the Haar measure on K, normalized to have total volume 1. In the coordinate θ introduced in Section 8.4 identifying k and $e^{i\theta}$, we have $dk = d\theta/2\pi$. This identification also makes clear the identity

$$\int_K \chi_m(k)\chi_{-n}(k)dk = \delta_{m,n}$$

Given any function $h \in C^\infty(G)$, for each $x \in G$ we consider the function

$$h_n : x \mapsto \int_K h(xk)\chi_n(k^{-1})\, dk$$

which is essentially $(h * \chi_n)(x)$ on K though we want to allow for $x \in G$. It may be viewed as a convolution on G if we view $\chi_n\, dk$ as a measure of G supported on K

and is referred to as the "n-th Fourier component of h" (with respect to K). If h is of type m, then this simplifies further to

$$h_n(x) = \int_K h(x)\chi_m(k)\overline{\chi_n(k)}\,dk = h(x)\delta_{m,n}.$$

Thus $*\chi_n$ projects functions in $C^\infty(G)$ to elements of K-type n (on the right).

Proof of Theorem 8.6.3, Part 2. Let $M = \mathcal{U}(\mathfrak{g}) \cdot f$ and V be the closure of M in $C^\infty(G)$. Theorem 8.6.4 guarantees that M_{in} is finite dimensional (and hence closed in V). Thus we must show that any $v \in V_n$ is a limit of elements in M_{in}. We have

$$M = M_{in} \oplus \bigoplus_{\ell \neq n} M_{i\ell} \Longrightarrow V = M_{in} \oplus \overline{\bigoplus_{\ell \neq n} M_{i\ell}}.$$

But $*\chi_n$ annihilates the space $\oplus_\ell M_{i\ell}$ and hence annihilates its closure since the map $*\chi_n$ is continuous. Thus given $v \in V_n$ there exists $a_j \in M_{in}$ and b_j in the complement of M_{in} in V such that $v = \lim_j (a_j + b_j)$. Taking projectors $*\chi_n$ on both sides,

$$v = v * \chi_n = \lim_j (a_j * \chi_n + b_j * \chi_n) = \lim_j a_j.$$

This completes the proof. □

8.7 Fundamental domains again - Siegel sets

Before (finally) getting to the general definition of an automorphic form on G, we need a precise measure of "moderate growth." This will be phrased in terms of Siegel sets, which provide an alternative to Dirichlet regions for describing fundamental domains for $\Gamma\backslash G$.

First we recall our earlier approach to fundamental domains in Section 3.6. There we considered the Dirichlet region

$$D_w(\Gamma) = \bigcap_{\gamma \in \Gamma - \{\pm 1\}} E(\gamma, w), \quad \text{where} \quad E(\gamma, w) = \{z \in \mathcal{H} \mid \rho(z, w) \leq \rho(z, \gamma w)\}$$

where ρ denotes the distance in the hyperbolic metric. Its interior $D_w^\circ(\Gamma)$ is given by replacing $\leq$ with $<$ in the definition of $E(\gamma, w)$. Let $\overline{\mathcal{F}}$ denote the closure of $\mathcal{F}$ in $\overline{\mathcal{H}} = \mathcal{H} \cup \mathbb{R} \cup \{\infty\}$.

In defining fundamental domains $\mathcal{F}$ in Section 3.6, we simply required the interior of $\mathcal{F}$ to be a connected subset of Γ-inequivalent points whose closure has transitive

action by Γ in $\mathcal{H}$. We can be a bit more precise about this by requiring the following pair of conditions on $\mathcal{F}$ which Borel calls the "Siegel property":

$$\Gamma \cdot \mathcal{F} = \mathcal{H} \quad \text{and} \quad \{\gamma \in \Gamma \,|\, \gamma\mathcal{F} \cap \mathcal{F} \neq 0\} \text{ is finite.}$$

This definition may be applied to any group Γ acting on a set $\mathcal{H}$.

Poincaré showed that, provided we choose w such that its isotropy group (as a subgroup of $\mathrm{PSL}(2,\mathbb{R})$) is trivial, $D_w(\Gamma)$ satisfies the first of the two conditions of the Siegel property, sometimes referred to as a "Poincaré fundamental set."

Theorem 8.7.1 (Siegel)**.** *Let $D = D_w(\Gamma)$ be a Poincaré fundamental set for Γ of finite area. Then ∂D is the union of finitely many geodesic segments, $\partial\overline{D} \cap \overline{\mathcal{H}}$ is finite, and $\Gamma\{\partial\overline{D} \cap \overline{\mathcal{H}}\}$ is the set of cusps for Γ. Further, $\Gamma\backslash\mathcal{H}^*$ is compact and D has the Siegel property.*

We now use an alternate construction of sets having the Siegel property which is more amenable to computation. This requires the notion of a parabolic pair, or p-pair for short. For us, a parabolic subgroup P of G will be the stabilizer of a line in $\mathbb{R}^2$. Since K acts transitively on these subgroups, any P is conjugate to

$$P_0 = \left\{ \begin{pmatrix} a & b \\ 0 & a^{-1} \end{pmatrix} \right\} < \mathrm{SL}(2,\mathbb{R}),$$

and

$$P_0 = \pm N_0 A_0, \quad N_0 = \left\{ \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \,\middle|\, x \in \mathbb{R} \right\}, \quad A_0 = \left\{ \begin{pmatrix} t & 0 \\ 0 & t^{-1} \end{pmatrix} \,\middle|\, t > 0 \right\}.$$

Similarly $P^\circ = N_P A$, where P° is the connected component of 1 in P, N_P is the derived subgroup of P, and A is any conjugate of A_0 in P. Here A_0 is the unique Cartan subgroup (that is, conjugate of A_0) of P_0 with Lie algebra orthogonal to that of K. Thus P has a unique Cartan subgroup with the same property. A p-pair is a pair (P, A) with P parabolic and A a Cartan subgroup. The p-pair is "normal" if A is the unique Cartan in P orthogonal to K. (We will always assume our p-pairs are normal.) Given any such p-pair (P, A) with $N_P = N$, we have the corresponding Iwasawa decomposition $G = NAK$. K acts transitively on the set of normal p-pairs and so many arguments can be reduced to the single p-pair (P_0, A_0).

Definition 8.5 (Siegel set)**.** *Given a normal p-pair (P, A), let N be the unipotent radical of P. A Siegel set with respect to (P, A) is a subset of the form*

$$\mathfrak{S} = \mathfrak{S}_{\omega,t} = \omega \cdot A_t \cdot K \quad \omega \subset N, \ \textit{compact},$$

and

$$A_t = \{a \in A \mid a^\rho > t\} \quad a^\rho \stackrel{def}{=} ||ae_P||,$$

where e_P is the unit vector in the direction of the line through the origin in $\mathbb{R}^2$ fixed by P.

For example, if $A = A_0$ corresponding to P_0, then the fixed line is the horizontal axis $e_1\mathbb{R}$, so

$$A_{0,t} = \left\{ \begin{pmatrix} t_a & 0 \\ 0 & t_a^{-1} \end{pmatrix} \;\middle|\; t_a > t \right\}.$$

The reader familiar with root systems for Lie groups will note that the definition of ρ defines an element of the character group $X(A)$ – continuous homormorphisms from A to $\mathbb{C}^\times$.

This is a Siegel set with respect to the group G and we can translate this to a Siegel set on $\mathcal{H}$ via the map $\mathfrak{S} \mapsto \mathfrak{S}' = \mathfrak{S} \cdot i$. If we identify the parabolic subgroup P with a fixed point u on $\partial\overline{\mathcal{H}}$, then we say $\mathfrak{S}$ is the Siegel set at u. Again returning to the case of $(P, A) = (P_0, A_0)$ corresponding to $u = \infty$, we could take for some $h > 0$:

$$\omega = \left\{ \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix} \;\middle|\; |n| \leq h \right\}.$$

Then

$$\mathfrak{S}'_{t,\omega} = \{z = x + iy \in \mathcal{H} \mid |x| \leq h, y > t^2\}.$$

A Siegel set has finite invariant measure. Since K acts transitively on the set of all p-pairs, it suffices to check this for the Siegel sets of (P_0, A_0). In this case,

$$\mathrm{Vol}(\mathfrak{S}'_{t,\omega}) = \int_{\mathfrak{S}'} y^{-2} dx\, dy = \int_\omega dx \int_{y>t^2} y^{-2} dy = t^{-2} \int_\omega dx < \infty$$

since ω is assumed compact.

Theorem 8.7.2. *Suppose Γ has cofinite volume in G. Let $u_1, \ldots, u_l$ be a set of representatives for Γ-equivalence classes of cusps. Then Γ has a fundamental set D in $\mathcal{H}$ of the form $C \cup (\bigcup_i \mathfrak{S}'_i)$ where $\mathfrak{S}'_i$ is a Siegel set at the cusp u_i, the $\mathfrak{S}'_i$ have disjoint images in $\Gamma\backslash\mathcal{H}$ and C is compact. The set D has finite volume and satisfies the Siegel property.*

For a proof of this theorem, see 3.17 of Borel's "Automorphic Forms on $\mathrm{SL}(2, \mathbb{R})$."

Sketch of proof. Because $X(\Gamma)$ is Hausdorff, we may choose disjoint neighborhoods V_i of the cuspidal representatives u_i. Then $V_i - \{u_i\}$ may be taken to be the image of a Siegel set $\mathfrak{S}'_i$ (Choose ω big enough to contain an interval including the fundamental domain for Γ_{u_i} in N. Then varying $\mathfrak{S}'_i \cup \{u_i\}$ over t is seen to give a fundamental system of open neighborhoods of u_i in $X(\Gamma)$.) The complement of $\bigcup V_i$ is relatively compact, and is the image of a compact set C under the projection. Checking the Siegel property is an exercise in the properties of a discontinuous group action. $\square$

8.8 Growth conditions

Suppose for simplicity that ∞ is a cusp of Γ and (as before) we let $\Gamma_\infty = \Gamma \cap N_0$, the stabilizer of ∞. Let f be a function on G such that

$$f(\gamma g k) = f(g)\chi(k) \quad \text{for } \gamma \in \Gamma_\infty, g \in G, k \in K. \tag{28}$$

Then given any (normal) p-pair (P, A), let $\mathfrak{S} = \omega A_t K$ be a corresponding Siegel set. We say that f is of "moderate growth" on $\mathfrak{S}$ if there exists a $\lambda \in \mathbb{R}$ such that

$$|f(g)| = O(a(g)^\lambda), \quad \text{where } g = n(g)a(g)k(g) \tag{29}$$

in the Iwasawa decomposition. Here we mean $a(g)^\lambda = ||a(g) \cdot e_P||^\lambda$. We say that f is "rapidly decreasing" if the condition in (29) holds for all λ. In either case, we say the growth in a smooth function f is "uniform" if the bound holds for Df with any $D \in \mathcal{U}(\mathfrak{g})$. Finally, if f is of moderate growth on a Siegel set $\mathfrak{S}$ corresponding to the parabolic P with associated cusp τ, we say f is of moderate growth "at the cusp τ" for short.

For example, if f is of the form (28), then $|f|$ can be viewed as a function on $\mathcal{H} = G/K$. If $(P, A) = (P_0, A_0)$ and $\mathfrak{S}'_t$ is the standard Siegel set in $\mathcal{H}$ consisting of points $z = x + iy$ with $|x| \leq h$ and $y \geq t$, then the moderate growth condition is that there exists a λ such that

$$|f(z)| = O(y^\lambda) \quad \text{for all } z \in \mathfrak{S}'.$$

Further, let F be a function on $\overline{\mathcal{H}} = \mathcal{H} \cup \mathbb{R} \cup \{\infty\}$ which is Γ_∞-invariant and has convergent power series expansion of form

$$F(z) = \sum_{n \geq n_0} a_n e^{2\pi i n z/h}, \quad \text{where } \Gamma_\infty = \langle T_h \rangle.$$

Then since $|e^{2\pi i n z/h}| = e^{-2\pi n y/h}$, we see that

$$F \text{ has moderate growth in a neighborhood of } z \iff a_n = 0 \text{ for all } n < 0.$$
$$F \text{ is rapidly decreasing in a neighborhood of } z \iff a_n = 0 \text{ for all } n \leq 0.$$

In particular, we see that classical automorphic forms satisfy the moderate growth conditions at the cusps as their Fourier expansions are required to be holomorphic at the cusps. Further, cusp forms are rapidly decreasing at the cusps.

8.9 Definition and first properties of an automorphic form

We may at last give the general definition of an automorphic form. Though this is stated for $G = \mathrm{SL}(2, \mathbb{R})$, the form of the definition is correct for any Lie group.

Definition 8.6 (Automorphic form)**.** *A smooth function $f : G \to \mathbb{C}$ is said to be an automorphic form with respect to a discrete subgroup Γ of G if the following conditions are satisfied:*

1. *$f(\gamma g) = f(g)$ for all $\gamma \in \Gamma, g \in G$.*
2. *f is K-finite.*
3. *f is $\mathcal{Z}$-finite.*
4. *f is of moderate growth at the cusps of Γ.*

In Section 8.2, we outlined a proof that by defining

$$\tilde{f}(g) = j(g, i)^{-m} f(g(i)), \tag{30}$$

for f, a classical automorphic form on $\mathcal{H}$ with respect to Γ, then $\tilde{f}$ is an automorphic form on G according to the definition above. That is, we confirmed the first two properties, and stated the last two. In Section 8.3, it was shown that the Casimir element $\mathcal{C} = -2\Delta$. Holomorphic functions on $\mathcal{H}$ have eigenvalue 0 under Δ, but the presence of $j(g, i)^{-m}$ in (30) shifts the eigenvalue to $m^2/2 - m$. Finally, about moderate growth, we already noted that f has moderate growth at the cusps according to the holomorphicity condition. Thus it suffices to show that $j(g, i)^{-m}$ is of moderate growth. Using the Iwasawa decomposition and cocycle relation:

$$j(g, i) = j(nak, i) = j(n, a(i))j(a, i)j(k, i) = a(g)^{-\rho}$$

since $j(n, z) \equiv 1$ for any $n \in N_0$, $j(k, i)$ has modulus 1 for all k, and $j(a, i) = t_a^{-1} = a^{-\rho}$ if $a = \mathrm{diag}(t_a, t_a^{-1})$.

We may also use (30) to provide a simple interpretation for the Petersson inner product:

$$\langle f_1, f_2 \rangle = \int_{\Gamma \backslash \mathcal{H}} y^{m-2} f_1(z) \overline{f_2(z)} \, dx \, dy,$$

where f_1, f_2 are automorphic forms on $\mathcal{H}$ of weight m. Indeed, if $\tilde{f}_1$ and $\tilde{f}_2$ are related to f_1 and f_2 as in (30), then we claim that the Petersson inner product is just the usual scalar product

$$\langle \tilde{f}_1, \tilde{f}_2 \rangle = \int_{\Gamma\backslash G} \tilde{f}_1(g)\overline{\tilde{f}_2(g)}\, dg.$$

To see this, notice that

$$\tilde{f}_1(g)\overline{\tilde{f}_2(g)} = f_1(g(i))\overline{f_2(g(i))}|j(g,i)|^{-2m} = f_1(g(i))\overline{f_2(g(i))}\,\mathrm{Im}(g(i))^m.$$

The function on the right-hand side above is K-invariant, and so we have the desired matching of integrals by setting $z = g(i)$ with $z = x + iy$ and noting that $dg = dk\, d\mu$ where μ is Haar measure on $\mathcal{H}$.

We record several properties of automorphic forms f which follow from this definition.

Proposition 8.9.1. *An automorphic form f with respect to $\Gamma < G$ further satisfies the following:*

1. *f is real analytic.*
2. *Given a neighborhood U of 1 in G, there exists a function $\alpha \in I_c^\infty(G)$ with $\mathrm{Supp}(\alpha) \subset U$ such that $f * \alpha = f$.*
3. *f has uniform moderate growth at every cusp of Γ.*

Proof. The first two statements are Theorem 8.5.2, using the strong form of the elliptic regularity theorem, and Theorem 8.6.1 since (to apply both results) we may use that f is both K-finite and $\mathcal{Z}$-finite. The third fact requires proof, and is a nice illustration of the utility of Theorem 8.6.1.

We claim that if f is of moderate growth on a Siegel set $\mathfrak{S}_t$ with exponent λ, then so is $f * \alpha$ for any $\alpha \in C_c^\infty(G)$. Assuming the claim for the moment, then

$$D(f * \alpha)(g) = (f * D\alpha)(g) \quad \text{for all } D \in \mathcal{U}(\mathfrak{g}),$$

and since $D\alpha$ is again in $C_c^\infty(G)$, the claim implies

$$|D(f * \alpha)(g)| = O(a(g)^\lambda). \tag{31}$$

If we choose α to be the convolution identity guaranteed above, then (31) holds with f in place of $f * \alpha$, which proves property 3.

Thus it suffices to prove the claim that f of moderate growth implies $f * \alpha$ of moderate growth. Given any $\alpha \in C_c^\infty(G)$, then $\mathrm{Supp}(\alpha)$ is contained in a relatively compact, symmetric neighborhood U. We consider

$$|f * \alpha(x)| = \left| \int_G f(y)\alpha(y^{-1}x)\, dy \right| \leq ||\alpha||_\infty \int_{xU} |f(y)|\, dy.$$

Thus it suffices to show that if $x \in \mathfrak{S}_t$, then there exists a corresponding $\mathfrak{S}_{t'}$ containing xU such that $f(y) = O(a(y)^\lambda)$ on $\mathfrak{S}_{t'}$.

Since U is relatively compact, KU is relatively compact, so via the Iwasawa decomposition, there exists compact subsets $C_A \subseteq A$ and $C_N \subseteq N$ such that $KU \subset C_N C_A K$. Now

$$xU = n(x)a(x)k(x)U \subset Na(x)C_N C_A K = Na(x)C_A K,$$

where in the last step we used that N is normal in NA, so $NaC_N \subseteq Na$ for any $a \in A$. Since $x \in \mathfrak{S}_t$, we know that $a(x)^\rho > t$ and so we may find a t' with $(a(x) \cdot a')^\rho = a(x)^\rho (a')^\rho > t'$ for $a' \in C_A$ since this latter set is compact. Hence, $|f(y)| = O(a(x)^\lambda)$ for $y \in xU \subset \mathfrak{S}'_t$. This gives the bound

$$||\alpha||_\infty \int_{xU} |f(y)|\, dy = ||\alpha||_\infty \cdot \mathrm{vol}(U) O(a(x)^\lambda) = O(a(x)^\lambda),$$

as desired. □

9 Finite dimensionality of automorphic forms

The key ingredient in the proof is the fact that cusp forms are rapidly decreasing on Siegel sets. (This will also be used in the analytic continuation of Eisenstein series in the next section.)

9.1 Constant term estimates and cuspidality

Let $P = NA$ be a cuspidal parabolic subgroup for Γ. Set $\Gamma_N = \Gamma \cap N$ and let f be a continuous Γ_N invariant function on G that is locally L^1. The constant term f_P of f along P is defined as follows:

$$f_P(g) = \int_{\Gamma_N \backslash N} f(ng)\, dn \quad \text{for all } g \in G,$$

where we've normalized the measure dn so that the total volume of $\Gamma_N\backslash\Gamma$ is equal to 1. As the integration is performed on the left, we have the following relations. First

$$D(f_P) = (Df)_P \quad \text{for all } D \in \mathcal{U}(\mathfrak{g}), f \in C^\infty(\Gamma_N\backslash G),$$

and

$$(f * \varphi)_P = f_P * \varphi \quad \text{for } \varphi \in C_c^\infty(G).$$

Note that our terminology is apt here. If F is a classical automorphic form with cusp ∞ and stabilizer $\Gamma_N = \langle T_h \rangle$, set $f(g) = j(g,i)^{-m}F(g(i))$ as before with $g(i) = z = x + iy$. Then the constant term along $P = P_0$ is

$$f_P(g) = \int_0^h j(g,i)^{-m}F(g(i))\frac{dx}{h} \quad \text{where } N_0 = \left\{ \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \right\}.$$

Substituting in the Fourier expansion of F we have

$$f_P(g) = j(g,i)^{-m} \sum_{n \geq 0} a_n e^{-2\pi n y/h} \int_0^h e^{2\pi i n x/h} \frac{dx}{h} = j(g,i)^{-m} a_0.$$

We now prepare for the main lemma on constant terms. First, choose an element $Y \in \mathrm{Lie}(N)$ such that e^Y generates the (infinite cyclic) group Γ_N. Then for $a \in A$, we have

$$\mathrm{Ad}\, a(Y) = aYa^{-1} = a^{2\rho} \cdot Y.$$

(Indeed, we can check this for $P = P_0$ then it will hold for all P by conjugation in K. This amounts to the matrix multiplication $\left(\begin{smallmatrix} t_a & 0 \\ 0 & t_a^{-1} \end{smallmatrix}\right)\left(\begin{smallmatrix} 0 & c \\ 0 & 0 \end{smallmatrix}\right)\left(\begin{smallmatrix} t_a^{-1} & 0 \\ 0 & t_a \end{smallmatrix}\right) = t_a^2\left(\begin{smallmatrix} 0 & c \\ 0 & 0 \end{smallmatrix}\right) = a^{2\rho}Y$.) In other words, the character 2ρ of A is the (unique) simple root of $\mathfrak{g}$ with respect to the Lie algebra $\mathrm{Lie}(A)$. As such, we call $2\rho = \alpha$, which is more common in the literature (and suggestive of the general case).

Lemma 9.1.1. *Let X_1, X_2, X_3 be a basis for the Lie algebra $\mathfrak{g}$ of G. Let $f \in C^1(\Gamma_N\backslash G)$. Then there exists a constant $c > 0$, independent of f, such that*

$$|(f - f_P)(g)| \leq c \cdot a(g)^{-\alpha} \left(\sum_{i=1}^3 |X_i f|_P(g) \right) \quad \textit{for all } g \in G.$$

Proof. First note that $|X_i f|_P$ is defined because X_i is a left-invariant differential operator, so $X_i f$ remains Γ_N invariant on the left. Now with Y again chosen so that e^Y generates Γ_N, we have

$$(f_P - f)(g) = \int_0^1 (f(e^{tY}g) - f(g))\, dt.$$

But the integrand may be rewritten:

$$f(e^{tY}g) - f(g) = \int_0^t \frac{d}{ds} f(e^{(u+s)Y} \cdot g)\Big|_{s=0} du.$$

In our earlier convolution notation, we could write

$$f(e^{tY}g) - f(g) = \int_0^t ((-Y) * f)(e^{uY}g)\, du.$$

Unfortunately, this is a right-invariant differential operator on G. We need to express the integrand in terms of left-invariant differential operators acting on f. First, we have

$$\frac{d}{dt} f(e^{tY}g)\Big|_{t=0} = \frac{d}{dt} f(gg^{-1}e^{tY}g)\Big|_{t=0} = \frac{d}{dt} f(ge^{t\,\mathrm{Ad}\, g^{-1}(Y)})\Big|_{t=0}.$$

Now using the Iwasawa decomposition $g = n(g)a(g)k(g)$, we have

$$\mathrm{Ad}\, n(Y) = Y, \quad \mathrm{Ad}\, a(g)^{-1}(Y) = a(g)^{-\alpha}Y,$$

as we checked in the remark above the lemma. For $k(g)$ we can only say that, since $\mathrm{Ad}\, g$ acts on the Lie algebra $\mathfrak{g}$,

$$\mathrm{Ad}\, k(g)^{-1}(Y) = \sum_{i=1}^{3} c_i(k(g)^{-1})X_i$$

for smooth functions c_i. Thus

$$(-Y * f)(g) = a(g)^{-\alpha} \sum_{i=1}^{3} c_i(k(g)^{-1}) X_i f(g).$$

Since K is compact, the functions c_i are bounded. Choose $c \geq \max_{i,k(g)} |c_i(k)|$. Then

$$|(-Y * f)(e^{uY}g)| \leq c \cdot a(g)^{-\alpha} \sum |X_i f(e^{uY} \cdot g)|$$

since $a(g) = a(e^{uY}g)$ for all g in G, as e^Y was taken to generate N. Now considering the inequality

$$|f(e^{tY}g) - f(g)| \leq \int_0^t |(-Y * f)(e^{uY}g)|\, du \leq c \cdot a(g)^{-\alpha} \sum \int_0^1 |X_i f|(e^{uY}g)\, du.$$

But these latter integrals are just $|X_i f|_P(g)$. Integrating this constant in t, and noting the total volume of $\Gamma_N \backslash N$ is 1, gives the result. □

Theorem 9.1.2. *Let P be a cuspidal parabolic subgroup with respect to Γ. Let ψ be a smooth function on G, left-invariant under Γ_N, and of uniform moderate growth on a Siegel set $\mathfrak{S}_t$ corresponding to P. Then $\psi - \psi_P$ is rapidly decreasing on $\mathfrak{S}_t$.*

Proof. By assumption, there exists $\lambda \in \mathbb{R}$ (which may be considered as an element of the character group $X(A)_{\mathbb{R}}$) such that

$$|D\psi(g)| = O(a(g)^{\lambda}) \quad \text{for all } g \in \mathfrak{S}_t \text{ and all } D \in \mathcal{U}(\mathfrak{g}).$$

In particular, letting $D = X_1, X_2, X_3$ and applying the previous lemma, we have

$$|(\psi - \psi_P)(g)| = O(a(g)^{\lambda - \alpha}) \quad \text{for } g \in \mathfrak{S}_t.$$

But in fact, we can do better still. Since $(f_P)_P = f_P$ for any locally L^1 function that is left Γ_N invariant, we have $(\psi - \psi_P)_P = 0$. Applying the lemma repeatedly, e.g. to $(\psi - \psi_P) - (\psi - \psi_P)_P$, we obtain the stronger result for any positive integer m

$$|(\psi - \psi_P)(g)| = O(a(g)^{\lambda - m\alpha}) \quad \text{for } g \in \mathfrak{S}_t.$$

We may identify $X(A)$ with $\mathbb{C}$ (and $X(A)_{\mathbb{R}}$ with $\mathbb{R}$) via the map

$$s \in \mathbb{C} \longleftrightarrow \chi_s : a \mapsto a^{s\rho} = ||a \cdot e_P||^s.$$

Thus given any $\mu \in X(A)_{\mathbb{R}}$, there exists an m such that

$$\mu = \lambda - m\alpha + r\rho \quad \text{for some } r > 0.$$

Hence $\psi - \psi_P$ is rapidly decreasing. □

Corollary 9.1.3. *Let f be a locally L^1 function on G, left-invariant under Γ_N, of moderate growth on $\mathfrak{S}_t$. Given $\varphi \in C_c^{\infty}(G)$, then $f * \varphi - (f * \varphi)_P$ is rapidly decreasing on $\mathfrak{S}_t$.*

Proof. As noted earlier, convolution on the right remains left-invariant with respect to Γ_N. Convolution with a function in $C_c^{\infty}(G)$ is a smoothing operator, so $f * \varphi$ is smooth and (according to Proposition 8.9.1) of uniform moderate growth on $\mathfrak{S}_t$. Thus $f * \varphi$ satisfies all the conditions of the previous theorem and the result follows. □

Corollary 9.1.4. *Let f be an automorphic form on G with respect to Γ. Let P be a cuspidal parabolic subgroup for Γ. Then $f - f_P$ is rapidly decreasing on a Siegel set with respect to P.*

Proof. According to the definition of automorphic form, f is of moderate growth on a Siegel set $\mathfrak{S}_t$. Further, by Theorem 8.6.1, there exists a φ in $C_c^\infty(G)$ such that $f = f * \varphi$. Thus we may apply the previous corollary. □

Definition 9.1 (Cuspidality)**.** *Let f be a locally L^1 function on $\Gamma\backslash G$. Then f is said to be cuspidal for Γ if the constant term with respect to every cuspidal parabolic subgroup is zero.*

Given $\gamma \in \Gamma$, let $P' = {}^\gamma P$. Then we have $N' = {}^\gamma N$, so the automorphism given by conjugation by γ induces an isomorphism of $\Gamma_N\backslash N$ onto $\Gamma_{N'}\backslash N'$ taking dn to dn'. Then $f_P(g) = f_{P'}(\gamma g)$ for all $g \in G$:

$$\int_{\Gamma_N\backslash N} f(ng)\,dn = \int_{\Gamma_N\backslash N} f(\gamma n\gamma^{-1}\gamma g)\,dn = \int_{\Gamma_{N'}\backslash N'} f(n'\gamma g)\,dn'.$$

Thus it suffices to check this condition on a set of representatives of Γ-conjugacy classes of cuspidal parabolic subgroups.

From the above definition, Corollary 9.1.4 implies the following.

Corollary 9.1.5. *A cuspidal automorphic form is rapidly decreasing on any Siegel set with respect to a cuspidal parabolic subgroup.*

9.2 Finite dimensionality of automorphic forms

Given a non-zero ideal J in $\mathcal{Z}$ and a character χ of K, we use the notation $\mathcal{A} = \mathcal{A}(\Gamma, J, \chi)$ to denote the space of automorphic forms with respect to Γ that are annihilated by J and have K-type χ on the right. The main theorem of this section will be the finite dimensionality of these spaces for any choice of data.

The basic outline of the argument is to consider the map ψ which maps f to each of its constant terms along each cuspidal parabolic P_i:

$$\psi : \ f \mapsto (f_{P_1}, \ldots, f_{P_l}) \quad \text{for any } f \in \mathcal{A}.$$

We show that both the kernel and image of ψ are finite dimensional. Note that the kernel is precisely the set of cuspidal automorphic forms. The finite dimensionality for it follows from the growth estimates of the previous section together with a lemma due to Godement. The fact that the image is finite dimensional follows from showing that the constant terms f_P are all solutions to a rather simple ordinary differential equation. We begin with the lemma.

Lemma 9.2.1 (Godement)**.** *Let Z be a locally compact space with positive measure μ such that $\mu(Z)$ is finite. Let V be a closed subspace of $L^2(Z,\mu)$ consisting of essentially bounded functions. Then V is finite dimensional.*

Recall that a function is "essentially bounded" if $\inf(S) < \infty$ where

$$S = \{x \in \mathbb{R}_{\geq 0} \,|\, \mu(|f|^{-1}(x,\infty)) = 0\}.$$

Let $||\cdot||_\infty$ denote this essential supremum.

Proof. For every $f \in L^2(Z,\mu)$ we have

$$||f||_2 \leq \mu(Z)||f||_\infty$$

The map $(V, ||\cdot||_\infty) \to (V, ||\cdot||_2)$ given by the identity map on V is therefore continuous (and naturally, a bijection). It is a consequence of the open mapping theorem that its inverse is also continuous. (See for example Yosida's "Functional Analysis," II.5, p. 77) Hence there exists a $c > 0$ such that

$$||f||_\infty \leq c\,||f||_2 \quad \text{for all } f \in V.$$

Let $v_1, \ldots, v_n$ be an orthonormal subset of V. For any complex coefficients $a_i \in \mathbb{C}$, we then have

$$\left|\sum a_i v_i(z)\right| \leq c\left|\left|\sum a_i v_i\right|\right|_2 = c\left(\int_Z \sum a_i v_i \cdot \overline{\sum a_i v_i}\, d\mu\right)^{1/2} = c\left(\sum |a_i|^2\right)^{1/2}$$

for almost all z. Setting $a_i = \overline{v_i(z)}$ for $i = 1, \ldots, n$, this becomes

$$\sum |v_i(z)|^2 \leq c\left(\sum |v_i(z)|^2\right)^{1/2} \Longrightarrow \sum |v_i(z)|^2 \leq c^2.$$

Integrating over the total space Z, we find

$$n \leq c^2\mu(Z)$$

which implies that $\dim(V) \leq c^2\mu(Z)$. □

In order to prove that $\ker(\psi)$ is finite dimensional, we need two further lemmas. By choosing Γ to be Fuchsian of the first kind, we have that $\Gamma\backslash G$ is of finite volume, say C. This implies that for $p \geq 1$, the space $L^p(\Gamma\backslash G)$ is contained in $L^1(\Gamma\backslash G)$. In fact, by Hölder's inequality, we have

$$\int_{\Gamma\backslash G} |f(x)|\, dx \leq \left(\int_{\Gamma\backslash G} |f(x)|^p\, dx\right)^{1/p} \cdot \left(\int_{\Gamma\backslash G} dx\right)^{1/q}, \quad \text{where } \frac{1}{p} + \frac{1}{q} = 1,$$

so that $||f||_1 \leq C^{1/q}||f||_p$ for $f \in L^p(\Gamma\backslash G)$. Writing ${}^\circ L^p(\Gamma\backslash G)$ for the cuspidal elements of $L^p(\Gamma\backslash G)$ then we have

$$ {}^\circ L^p(\Gamma\backslash G) = L^p(\Gamma\backslash G) \cap {}^\circ L^1(\Gamma\backslash G). $$

Lemma 9.2.2. *The subspace ${}^\circ L^p(\Gamma\backslash G)$ is closed in $L^p(\Gamma\backslash G)$ for all $p \geq 1$.*

Proof. According to the above discussion, it suffices to prove this for $p = 1$. (We really just need the case $p = 2$ in what follows.)

Given a Γ-cuspidal parabolic subgroup P and a function $\varphi \in C_c^\infty(N\backslash G)$, let

$$ \lambda_{P,\varphi}(f) \stackrel{def}{=} \int_{\Gamma_N\backslash G} f(x)\varphi(x)dx. $$

Then

$$ \lambda_{P,\varphi}(f) = \int_{N\backslash G} \varphi(\overline{x})\, d\overline{x} \int_{\Gamma_N\backslash N} f(nx)dn = \int_{N\backslash G} \varphi(\overline{x}) f_P(\overline{x})\, d\overline{x}, $$

where $\overline{x}$ denotes the projection of x into $N\backslash G$ and $d\overline{x}$ is the quotient measure, the unique N-invariant measure on G such that $dx = d\overline{x}\, dn$. Thus $f_P = 0$ if and only if $\lambda_{P,\varphi}(f) = 0$ for all $\varphi \in C_c^\infty(G)$. This allows us to characterize ${}^\circ L^1(\Gamma\backslash G)$ as the intersection of the kernels of $\lambda_{P,\varphi}$ as we range over all cuspidal parabolics P and all $\varphi \in C_c^\infty(N_P\backslash G)$. Hence it suffices to show that $\lambda_{P,\varphi}$ is continuous (see for example Rudin's "Functional Analysis," Theorem 1.18).

Since $\mathrm{Supp}(\varphi)$ is compact in $N\backslash G$ we may write its support, viewing φ as a function on G, as $N \cdot D$ where D is compact. But Γ_N is cocompact in N, so we could also express the support as $\Gamma_N \cdot E$ with E compact. Then

$$ |\lambda_{P,\varphi}(f)| \leq \int_E |f(x)\varphi(x)|dx \leq ||\varphi||_\infty \int_E |f(x)|\, dx. $$

There exist finitely many compact neighborhoods $U_1, \ldots, U_m$ of points in E such that E is contained in the union of the U_i and $G \to \Gamma\backslash G$ maps U_i homeomorphically onto its image for all i. This implies

$$ \int_E |f(x)|\, dx \leq m \int_{\Gamma\backslash G} |f(x)|\, dx \Longrightarrow |\lambda_{P,\varphi}| \leq ||\varphi||_\infty m||f||_1. $$

Thus $\lambda_{P,\varphi}$ is continuous, completing the proof. □

We give one more simple lemma before stating the main theorem.

Lemma 9.2.3. *Let $f \in L^2(\Gamma\backslash G)$. Then f is of type $\chi \in \hat{K}$ on the right if and only if*

$$\int_K f(xk)\overline{\chi(k)}\,dk = f(x) \quad \textit{for all } x \in \Gamma\backslash G.$$

Proof. First, by assumption, $f(xk) = f(x)\chi(k)$ so that the integrand $f(xk)\overline{\chi(k)} = f(x)$ and the total volume of K is normalized to be 1. In the other direction, suppose the integral identity holds for all such $x \in \Gamma\backslash G$. Using the identity for $f(x\ell)$ and making the variable change $k \mapsto k' = \ell k$, we obtain

$$f(x\ell) = \int_K f(x\ell k)\overline{\chi(k)}\,dk = \int_K f(xk')\chi(\ell)\overline{\chi(k')}\,dk' = \chi(\ell)f(x).$$

□

Theorem 9.2.4. *The space $\mathcal{A} = \mathcal{A}(\Gamma, J, \chi)$ of automorphic forms of Γ annihilated by the non-zero ideal J of $\mathcal{Z}(\mathfrak{g})$ and of type $\chi \in \hat{K}$ on the right is finite dimensional.*

Proof. As we noted above, let $P_1, \ldots, P_l$ be representatives of the Γ-conjugacy classes of cuspidal parabolic subgroups. Let ψ continue to denote the map

$$\psi : \ f \mapsto (f_{P_1}, \ldots, f_{P_l}) \quad \text{for } f \in \mathcal{A}.$$

We will show that $\ker(\psi)$ and $\text{image}(\psi)$ are finite dimensional.

The kernel of ψ is the set of cusp forms ${}^\circ\mathcal{A}$ contained in $\mathcal{A}$. Any cusp form is rapidly decreasing at the cusps and is therefore bounded on $\Gamma\backslash G$ and belongs to $L^2(\Gamma\backslash G)$. We would like to apply Godement's Lemma 9.2.1, which requires that ${}^\circ\mathcal{A}$ is closed in $L^2(\Gamma\backslash G)$. To do this, we check that each of the properties of $\mathcal{Z}$-finiteness, K-finiteness, and cuspidality are preserved by convergent sequences in L^2.

Let $f_n \to f$ be a convergent sequence in $L^2(\Gamma\backslash G)$ with $f_n \in {}^\circ\mathcal{A}$ for all n. This implies convergence in $L^1(\Gamma\backslash G)$ by our earlier discussion and hence convergence in the distribution sense. Thus if $Z \in \mathcal{Z}(\mathfrak{g})$ then $Zf_n \to Zf$ in L^2. This shows that f is annihilated by J (at least as a distribution initially) since the f_n are.

By Lemma 9.2.3,

$$\int_K f_n(xk)\overline{\chi(k)}\,dk = f_n(x)$$

so this condition continues to hold for f and hence f has K-type χ as well (at least up to a set of measure 0). But now that we know f is $\mathcal{Z}$-finite and K-finite, we conclude that it is analytic by elliptic regularity. Thus $f \in \mathcal{A}$. Finally, since ${}^\circ L^2(\Gamma\backslash G)$ is closed in $L^2(\Gamma\backslash G)$ by Lemma 9.2.2, this implies ${}^\circ\mathcal{A}$ is closed in $L^2(\Gamma\backslash G)$. Now we may apply Lemma 9.2.1 to conclude the kernel of ψ is finite dimensional.

To finish, we show that $\{f_P \,|\, f \in \mathcal{A}\}$ is a finite-dimensional vector space of functions. Let p be the fixed point of P in $\partial\overline{X}$ and $g \in K$ such that $g(\infty) = p$. Then, as usual, replacing Γ by ${}^g\Gamma$ we may assume that P is the group of upper triangular matrices and $p = \infty$. Since f_P is N-invariant on the left (corresponding to a change of variables in the integration over N) and of K-type χ on the right, then f_P is completely determined by its restriction to A (according to the Iwasawa decomposition $G = NAK$.) Thus it suffices to show that these restrictions generate a finite dimensional space.

Since the function f_P is N-invariant on the left, it is annihilated by the *right-invariant* differential operator E_r corresponding to the usual generator E in the Lie algebra. (Remember "right-invariant" operators act on the left.) We claim that in fact

$$Ef_P(a) \stackrel{def}{=} \frac{d}{dt} f_P(ae^{tE})\Big|_{t=0} = 0 \quad \text{for all } a \in A,$$

where E is the usual left-invariant operator as indicated in the definition.

Indeed we can perform our usual manipulations with respect to the adjoint action:

$$f_P(ae^{tE}) = f_P(ae^{tE}a^{-1}a) = f_P(e^{t\,\mathrm{Ad}\,a(E)}a) = f_P(e^{ta^\alpha E}a)$$

where we're setting $\alpha = 2\rho$ as usual. Thus,

$$\frac{d}{dt} f_P(ae^{tE})\Big|_{t=0} = a^\alpha \frac{d}{dt} f_P(e^{tE}a)\Big|_{t=0} = a^\alpha \frac{d}{dt} f_P(a)\Big|_{t=0} = 0.$$

One expression for the Casimir element $\mathcal{C}$ in terms of H, E, F has form

$$\mathcal{C} = \frac{1}{2}H^2 - H + EF.$$

Remembering that EF acts by first taking derivatives in the one-parameter subgroup with respect to E, we have $EF(f_P) = 0$ since $E(f_P) = 0$. If $P(\mathcal{C})$ is the monic polynomial annihilating f, then

$$P(\mathcal{C})f_P = Q(H)f_P \quad \text{where } Q(H) = P(H^2/2 - H),$$

annihilates f_P. Identifying $\mathbb{R}$ with A via the map $t \mapsto \mathrm{diag}(e^t, e^{-t})$. Then H just acts by d/dt and f_P may be viewed as a function of $\mathbb{R}$ which is then a solution to the differential equation

$$Q(\frac{d}{dt})f_P = 0.$$

The space of such solutions is finite dimensional. □

10 Eisenstein series

10.1 Definition and initial convergence

Recall that to any p-pair (P, A), we have an associated Iwasawa decomposition $G = NAK$ and write

$$g = n\,a(g)k \quad \text{with } n \in N, a(g) \in A, k \in K.$$

With respect to the Euclidean norm $||\cdot||$, we have

$$||g^{-1} \cdot e_P|| = ||k^{-1}a(g)^{-1}n^{-1} \cdot e_P|| = a(g)^{-\rho},$$

since k leaves the Euclidean norm invariant and $ne_P = e_P$. Set

$$h_P(g) = ||g^{-1}e_P||^{-1} = a(g)^{\rho}. \tag{32}$$

We have the following properties for any $n \in N, a \in A, k \in K, g \in G$, which both follow immediately from the definition:

$$h_P(nagk) = h_P(a) \cdot h_P(g), \qquad h_{P'}({}^k g) = h_P(g) \quad \text{where } P' = {}^k P. \tag{33}$$

This function h_P plays a critical role in the definition of the Eisenstein series. We record several facts that we'll need later in proving that the Eisenstein series is an automorphic form. Here we use Borel's notation: For strictly positive functions we write

$$f(x) \prec g(x)$$

if there exists a positive constant c such that $f(x) \leq cg(x)$ for all x. If $f \prec g$ and $f \succ g$ then we write $f \asymp g$.

Lemma 10.1.1. *Let (P, A) be a cuspidal p-pair with corresponding Siegel set $\mathfrak{S}$. The function h_P in (32) satisfies the following properties.*

1. *Given a compact set $C \subset G$, then $||gv|| \asymp ||v||$ for all $g \in C, v \in \mathbb{R}^2$.*
2. *For any $x \in \mathfrak{S}, v \in \mathbb{R}^2$, $||x^{-1}v|| \asymp ||a(x)^{-1}v||$.*
3. *For any $x \in G$ and $y \in C$ compact, $h_P(xy) \asymp h_P(x)$.*
4. *For any $y \in G$ and $x \in \mathfrak{S}$, $h_P(yx) \prec h_P(y)h_P(x)$.*
5. *For all $\gamma \in \Gamma$, $h_P(\gamma) \prec 1$.*

6. *For $\gamma \in \Gamma, c, d \in C, x \in \mathfrak{S}$, $h_P(\gamma cxd) \prec h_P(x)$.*

7. *If (P', A') is another cuspidal pair, then $h_P(yx) \prec h_P(y)h_{P'}(x)$ for any $y \in G$ and $x \in \mathfrak{S}'$ corresponding to P'.*

Proof. We take these in order.

Proof of 1. By rescaling, it suffices to show that $||gv|| \asymp 1$ when $g \in C$ and $v \in S_1$. This is clear because $C \cdot S^1$ is compact and does not contain the origin. □

Proof of 2. Using the Iwasawa decomposition, $x = n_x a(x) k_x$ so that

$$||x^{-1}v|| = ||k_x^{-1}a(x)^{-1}n_x^{-1}v|| = ||a(x)^{-1}n_x^{-1}v|| = ||a(x)^{-1}n_x^{-1}a(x)a(x)^{-1}v||.$$

But $a(x)^{-1}n_x a(x)$ is contained in a compact set if $x \in \mathfrak{S}$ (since $a^{-1}n_x a = a^{-2\rho}n_x$, $n_x \in \omega$ and $a^{-2\rho}$ is bounded on A_t in the Siegel set $\mathfrak{S} = \mathfrak{S}_{\omega,t}$). Hence the claim follows from part (1). □

Proof of 3. By definition, this means

$$||y^{-1}x^{-1} \cdot e_P|| \asymp ||x^{-1}e_P||$$

and this again follows from part (1). □

Proof of 4. This is equivalent to showing for all $x \in \mathfrak{S}, y \in G$:

$$||x^{-1}y^{-1}e_P|| \succ ||y^{-1}e_P|| \cdot ||x^{-1}e_P||.$$

Write $y^{-1}e_P = c_1(y)e_P + c_2(y)\tilde{e}_P$, where $\tilde{e}_P$ is a unit vector orthogonal to e_P. Thus $||y^{-1}e_P||^2 = c_1(y)^2 + c_2(y)^2$. Now using property (2),

$$||x^{-1}y^{-1}e_P|| \asymp ||a(x)^{-1}y^{-1}e_P||. \tag{34}$$

On the other hand,

$$a(x)^{-1}y^{-1}e_P = h_P(x)^{-1}(c_1(y)e_P + h_P^2(x)c_2(y)\tilde{e}_P),$$

so that

$$||a(x)^{-1}y^{-1}e_P||^2 = h_P(x)^{-2}(c_1^2(y) + h_P^4(x)c_2^2(y)) \succ h_P(x)^{-2}(c_1^2(y) + c_2^2(y)),$$

which is equal to $h_P(x)^{-2}h_P(y)^{-2}$, and comparing with (34) completes the proof of (4). □

Proof of 5. We may assume that $(P, A) = (P_0, A_0)$. Then if $\gamma \in \Gamma_\infty$ we have $a(\gamma) = \pm \mathrm{Id}$ and $h_P(\gamma) = 1$. In the case $P = P_0$,

$$h_P(x) = ||x^{-1}e_1||^{-1} = (c^2 + d^2)^{-1/2} \quad \text{where } x = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Now using the contrapositive of Lemma 3.7.2, we are guaranteed that c is bounded below if $c \in \Gamma - \Gamma_\infty$ (in fact, $|c| \geq 1/|h|$ where $\Gamma_\infty = \langle T_h \rangle$.) □

Proof of 6. Using parts (3), (4), (3) again and (5) in that order:

$$h_P(\gamma cxd) \asymp h_P(\gamma cx) \prec h_P(\gamma c)h_P(x) \asymp h_P(\gamma)h_P(x) \prec h_P(x).$$

□

Proof of 7. Choose $k \in K$ such that ${}^kP' = P$, which implies $ke_{P'} = \pm e_P$. Then for $x \in \mathfrak{S}'$ and $y \in G$ we have

$$||x^{-1}y^{-1}e_P||^{-1} = ||x^{-1}y^{-1}ke_{P'}||^{-1} \prec h_{P'}(y^{-1}k)h_{P'}(x^{-1}) = ||y^{-1}e_P||^{-1}||x^{-1}e_{P'}||^{-1},$$

where in the last step, we used that $ke_{P'} = \pm e_P$. □

This completes the proof of the lemma. □

For any cuspidal P with $P^\circ = NA$, let

$$\Gamma_N := \Gamma \cap N, \quad \Gamma_P := \Gamma \cap P,$$

noting that $\Gamma_P \subset Z(G)\Gamma_N$ where $Z(G)$ is the center of G, and thus $[\Gamma_P : \Gamma_N] \leq 2$.

Further let

$$F(P, m) = \{\varphi \in C^\infty(\Gamma_P NA \backslash G) \mid \varphi \text{ of right } K\text{-type } m\}.$$

These are quite restrictive conditions on the functions φ. In particular,

$$\varphi(nak) = \varphi(1)\chi_m(k) \quad \text{where } n \in N, a \in A, k \in K,$$

and thus $\dim F(P, m) \leq 1$. Suppose that $\Gamma_P \neq \Gamma_N$ so that there exists an element ϵn_0 where $\epsilon = -\mathrm{Id}$ in $Z(G)$ such that

$$\Gamma_P = \Gamma_N \cup (\epsilon n_0)\Gamma_N.$$

(If $n_0 = 1$, then $\Gamma_P = \Gamma_N \times Z(G)$ and if $n_0 \neq 1$ then Γ_P is infinite cyclic generated by ϵn_0.) Since ϵ is central and in K, we have

$$\varphi(g) = \varphi(\epsilon g) = \varphi(g\epsilon) = \varphi(g)\chi_m(\epsilon).$$

Thus if $\chi_m(\epsilon) = -1$, then $\varphi = 0$ and $\dim F(P, m) = 0$. Hence we may assume that either m is even or else $\Gamma_P = \Gamma_N$. In this case $\dim F(P, m) = 1$ and we let $\varphi_P = \varphi_{P,m}$ be the function in $F(P, m)$ such that its value at the identity is 1.

Recall that $P = NA \cup \epsilon NA$. Thus

$$\varphi_P(pg) = \varphi_P(p)\varphi_P(g) \quad \text{with} \quad \varphi_P(p) = \begin{cases} 1 & p \in NA \\ (-1)^m & p \in \epsilon NA. \end{cases} \tag{35}$$

For notational compactness, we write

$$\varphi_{P,m,s}(g) = \varphi_{P,s}(g) = \varphi_P(g) \cdot h(g)^{1+s}.$$

From (33) and (35) we have that

$$\varphi_{P,s}(pagk) = \varphi_{P,s}(p)h_P(a)^{1+s}\varphi_{P,s}(g)\chi_m(k) \quad \text{for any } g \in G, \tag{36}$$

and

$$\varphi_{P',s}({}^k g) = \varphi_{P,s}(g) \quad \text{where } P' = {}^k P,\ k \in K.$$

Definition 10.1 (Eisenstein series)**.** *Given a cuspidal parabolic P and $s \in \mathbb{C}$, we define the (spectral) Eisenstein series by*

$$E_{P,s}(g) = \sum_{\gamma \in \Gamma_P \backslash \Gamma} \varphi_{P,s}(\gamma g).$$

In the definition, we continue to suppress the dependence on the K-type m. For example, in the special case $P = P_0$,

$$h_P(g) = ||g^{-1}e_1||^{-1} = (c^2 + d^2)^{-1/2} = \mathrm{Im}(g(i))^{1/2}, \quad \text{with } g = \begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

and so

$$h_P(\gamma g) = \mathrm{Im}(\gamma g(i))^{1/2} = \frac{\mathrm{Im}(g(i))^{1/2}}{|j(\gamma, g(i))|}.$$

Thus letting $g(i) = z = x + iy$ we have

$$E_{P_0,s}(g) = \sum_{\gamma \in \Gamma_{P_0} \backslash \Gamma} \frac{y^{(s+1)/2}\varphi_P(\gamma x)}{|c_\gamma z + d_\gamma|^{s+1}} \quad \text{where } \gamma = \begin{pmatrix} a_\gamma & b_\gamma \\ c_\gamma & d_\gamma \end{pmatrix}.$$

In particular, if $m = 0$ then this is

$$E_{P_0,s}(g) = \sum_{\gamma \in \Gamma_{P_0} \backslash \Gamma} \frac{y^{(s+1)/2}}{|c_\gamma z + d_\gamma|^{s+1}}.$$

Back to the general case, let $s = \sigma + it$ and denote by $E_{P,0}(g) = E_0(g)$ the term-wise majorant

$$E_0(g) = \sum_{\gamma \in \Gamma_P \backslash \Gamma} |\varphi(g) h_P^{s+1}(\gamma g)| = \sum_{\gamma \in \Gamma_P \backslash \Gamma} h_P(\gamma g)^{\sigma+1}.$$

Theorem 10.1.2.

1. *$E = E_{P,s}$ converges absolutely and locally uniformly for $\sigma > 1$.*
2. *If (P', A') is a cuspidal pair and a corresponding Siegel set $\mathfrak{S}' = \mathfrak{S}_{P',t}$ then there exists a $c > 1$ such that*

$$E_0(g) \prec \frac{c^\sigma}{\sigma - 1} h_{P'}^{\sigma+1}(g) \quad g \in \mathfrak{S}', \sigma > 1.$$

3. *$\mathcal{C}(E) = \frac{1}{2}(s^2 - 1)E$, for $\sigma > 1$ and $\mathcal{C}$ denotes the Casimir operator.*
4. *E is an automorphic form for $\sigma > 1$.*
5. *If P' is a cuspidal subgroup for Γ not conjugate to P by an element of Γ, and $\mathfrak{S}'$ is a Siegel set with respect to P' then for $\sigma > 1$,*

$$E_0(g) \cdot h_{P'}(g)^{-(\sigma+1)} \longrightarrow 0 \quad \text{on } \mathfrak{S}',$$

when $h_{P'}(g) \to \infty$.

Proof. Note that (2) is stronger than (1), since any compact set is contained in a Siegel set. (2) also guarantees that E has moderate growth at the cusps. Hence (4) is implied by (2) and (3), since $\varphi_{P,s}$ is of type m (hence so is E) and E is clearly left Γ-invariant. Thus it suffices to prove (2), (3), and (5) and we split the proof into three pieces accordingly, beginning with Godement's proof of (2):

Proof of 2. If (P', A') is another cuspidal pair, then by part (7) of Lemma 10.1.1, we have

$$h_P(g_1 g_2) \prec h_P(g_1) h_{P'}(g_2) \quad \text{where } g_1 \in G \text{ and } g_2 \in \mathfrak{S}_{t'} =: \mathfrak{S}',$$

a Siegel set with respect to P'. This implies that

$$E_0(g) = \sum h_P(\gamma g)^{1+\sigma} \prec h_{P'}(g)^{1+\sigma} \sum h_P(\gamma)^{1+\sigma} \quad \text{for } g \in \mathfrak{S}' \text{ and } \gamma \in \Gamma.$$

Thus it suffices to show that there exists a constant $c > 1$ such that

$$\sum h_P(\gamma)^{\sigma+1} \prec c^\sigma (\sigma - 1)^{-1} \quad \text{for } \sigma > 1.$$

Fix a symmetric compact neighborhood C of the identity such that $\Gamma \cap C \cdot C = \{1\}$. These conditions are chosen so that

$$\gamma C \cap \gamma' C \neq \emptyset \Rightarrow \gamma = \gamma' \quad \text{for all } \gamma, \gamma' \in \Gamma.$$

By properties (3) and (5) of Lemma 10.1.1, we can find constants $b, d > 0$ such that

$$h_P(\gamma) \leq d h_P(\gamma y) \leq db \quad \text{for } \gamma \in \Gamma, y \in C, \tag{37}$$

where we've used both directions of the $\asymp$ in property (3). Raising both sides to the power $\sigma + 1$ and integrating over C:

$$h_P(\gamma)^{\sigma+1} \leq \mu(C)^{-1} d^{\sigma+1} \int_C h_P(\gamma y)^{\sigma+1}\, dy,$$

or in terms of the absolute majorant E_0,

$$E_0(x) \leq \mu(C)^{-1} d^{\sigma+1} \left(\sum_{\gamma \in \Gamma_P \backslash \Gamma} \int_{\gamma C} h_P(\gamma y)^{\sigma+1}\, dy \right) h_{P'}(x)^{\sigma+1}. \tag{38}$$

Define

$$A(0, b) = \{a \in A \mid h_P(a) \leq b\} \quad (b > 0).$$

Then according to (37), $\Gamma C \subset N A(0, b) K$. Hence

$$\Gamma_P \backslash \Gamma C \subset \Gamma_N \backslash N \times A(0, b) \times (Z(G) \cap \Gamma) \backslash K.$$

A simple exercise shows that Haar measure on G can be written as

$$dg = a(g)^{-2\rho} dn\, da\, dk,$$ [10]

[10] We choose Haar measure on K to have total volume 1, we identify N with $\mathbb{R}$ and take Lebesgue measure, and if $t = a^\rho$ is our coordinate on A, then in terms of Lebesgue measure dt we have $da = dt/t$. Because P° is not unimodular, so that $dp = dn\, da$ is right-invariant but the left-invariant measure is $a^{-2\rho}\, dn\, da$. The usual Haar measure $y^{-2} dx dy$ on $\mathcal{H} = G/K$ is twice this measure.

and if we take $h = h_P(a) = a^\rho$ as our coordinate on A, then the restriction of Haar measure from G to A gives $h^{-3}dh$. Returning to (38), there is thus a constant $\delta > 0$ such that

$$E_0(x) \leq \delta \left(\int_{\Gamma_P \backslash NK} dn\, dk \right) \left(\int_0^b h^{\sigma-2}\, dh \right) d^{\sigma+1} h_{P'}^{\sigma+1}(x).$$

The first integral is finite and strictly positive. The second is convergent for $\sigma > 1$ and then equal to $b^{\sigma-1}(\sigma-1)^{-1}$. Thus the claim follows by setting $c = bd$. □

Proof of 3. It suffices to prove that

$$\mathcal{C}\varphi_{P,s} = \frac{s^2-1}{2}\varphi_{P,s}.$$

The function h_P is N-invariant on the left and K-invariant on the right according to the definition. According to (36) we may write

$$\varphi_{P,s}(nak) = h_P^{s+1}(a)\chi_m(k) \quad \text{for } n \in N, a \in A, k \in K.$$

As $\mathcal{C}$ is bi-invariant, then $\mathcal{C}\varphi_{P,s}$ is also left-invariant (and also right K-invariant). Thus it suffices to check that $\varphi_{P,s}$ is an eigenfunction of $\mathcal{C}$ as a function on A. Again, conjugating if necessary, we may assume $P = P_0$. Write

$$\mathcal{C} = \frac{1}{2}H^2 - H + 2EF.$$

Since A normalizes N, we may write

$$\varphi_{P,s}(ae^{t_1 E}e^{t_2 F}) = \varphi_{P,s}(ae^{t_2 F}) \quad \text{for all } a \in A,$$

and thus $EF(\varphi_{P,s}) = 0$. This gives $\mathcal{C}\varphi_{P,s}(a) = (\frac{1}{2}H^2 - H)\varphi_{P,s}(a)$. On A, $\varphi_{P,s} = h_P^{1+s}$. Similar to an earlier argument, we take $h = h_P$ as a coordinate on A and identify A with its Lie algebra which is isomorphic to $\mathbb{R}$ via $a \mapsto \log h_P(a)$. Then $h_P^{s+1}(a) = e^{t(s+1)}$ where $e^t = a_t$ and H becomes d/dt. Thus we need only verify that

$$\left(\frac{1}{2}\frac{d^2}{dt^2} - \frac{d}{dt} \right) e^{t(s+1)} = \left(\frac{(s+1)^2}{2} - (s+1) \right) e^{t(s+1)} = \left(\frac{s^2-1}{2} \right) e^{t(s+1)}.$$

□

Proof of 5. By property 7 of Lemma 10.1.1, there exists a constant $c_1 > 0$ such that

$$h_P(\gamma g)^{\sigma+1} \cdot h_{P'}(g)^{-(\sigma+1)} \leq c_1 h_P(\gamma)^{\sigma+1}$$

for $\gamma \in \Gamma$ and $g \in \mathfrak{S}'$. By part 2 of this theorem, the series

$$\sum_{\Gamma_P \backslash \Gamma} h_P(\gamma)^{\sigma+1}$$

is uniformly convergent on compact sets for $\sigma > 1$. Then summing both sides over $\Gamma_P \backslash \Gamma$ we at least have that $E_0(g) h_{P'}(g)^{-(\sigma+1)}$ is bounded by an absolute constant independent of g. We claim that it suffices to show that for any given $\gamma \in \Gamma$,

$$h_P(\gamma g) \cdot h_{P'}(g)^{-1} \to 0 \quad \text{as} \quad h_{P'}(g) \to \infty \tag{39}$$

for $g \in \mathfrak{S}'$. The statement (39) follows from showing that $h_P(\gamma g) \asymp h_{P'}(g)^{-1}$ since then

$$h_P(\gamma g) h_{P'}(g)^{-1} \asymp h_{P'}(g)^{-2} \to 0 \quad \text{as} \quad h_{P'}(g) \to \infty.$$

This asymptotic is proved similar to property 4 in Lemma 10.1.1, separating the action of $\gamma^{-1} e_P$ into its component along e_P and its orthogonal unit vector $\tilde{e}_P$. Then use the Iwasawa decomposition. (To see that the claim implies part 5, note that (39) (raised to the $\sigma + 1$ power) implies that the partial sums for $E_0(g) h_{P'}(g)^{-(\sigma+1)}$ are all 0 in the limit, and the uniform convergence implies the same must be true for the limit of the sequence of partial sums. That is, we may interchange the two limits.) □

This completes the proof, according to the initial discussion of implications. □

Finally, we demonstrate that Eisenstein series are orthogonal to cusp forms with respect to the inner product:

$$\langle f, g \rangle = \int_{\Gamma \backslash G} f(x) \overline{g(x)} \, d\mu_{\Gamma \backslash G}(x).$$

Here $d\mu_{\Gamma \backslash G}$ is the unique G-invariant measure on the quotient such that $d\mu_G = d\mu_{\Gamma \backslash G} \cdot d\mu_\Gamma$ (as an integration identity for all $f \in C_c(G)$.) Similarly, we may define an inner product on any quotient with respect to a unimodular subgroup (e.g. N).

In fact, we can make similar claims for relative quotients. That is, given two unimodular subgroups $L \supset M$ of G, then $d\mu_{L \backslash G}$ is a quotient of $d\mu_{M \backslash G}$ by $d\mu_{M \backslash L}$ in the following sense. If f is continuous and integrable on $M \backslash G$ then the function

$$g \mapsto \int_{M \backslash L} |f(xg)| d\mu_{M \backslash L}(x)$$

is integrable on $L\backslash G$ and

$$\int_{M\backslash G} f(x)\, d\mu_{M\backslash G}(x) = \int_{L\backslash G} d\mu_{L\backslash G}(\dot{x}) \int_{M\backslash L} f(yx)\, d\mu_{M\backslash L}(y). \tag{40}$$

Conversely, if f is continuous on $M\backslash G$ such that

$$x \mapsto \int_{M\backslash L} |f(yx)|\, d\mu_{M\backslash L}(y) \quad \text{is integrable on } L\backslash G,$$

then f is integrable on $M\backslash G$ with integral given by (40).

The orthogonality follows immediately from the next result.

Proposition 10.1.3. *Let f be a continuous, rapidly decreasing function on $\Gamma\backslash G$. Then for $\mathrm{Re}(s) = \sigma > 1$,*

$$\langle E_{P,s}, f\rangle_{\Gamma\backslash G} = \langle \varphi_{P,s}, f_P\rangle_{N\backslash G}.$$

Proof. The function $\varphi_{P,s}$ is left-invariant under Γ_P. We first claim that $\varphi_{P,s}\cdot\overline{f}$ is integrable on $\Gamma_P\backslash G$ and thus applying (40) with $L = \Gamma$ and $M = \Gamma_P$,

$$\langle E_{P,s}, f\rangle_{\Gamma\backslash G} = \langle \varphi_{P,s}, f\rangle_{\Gamma_p\backslash G}.$$

To see this, note that

$$\begin{aligned}\int_{\Gamma_P\backslash\Gamma} \overline{f}(\gamma x)\cdot\varphi_{P,s}(\gamma x)\, d\mu_{\Gamma_P\backslash\Gamma} &= \sum_{\gamma\in\Gamma_P\backslash\Gamma} \overline{f}(\gamma x)\cdot\varphi_{P,s}(\gamma x)\\ &= \overline{f}(x)\cdot E_{P,s}(x).\end{aligned}$$

To show that $\varphi_{P,s}\cdot\overline{f}$ is integrable, note that in Theorem 10.1.2 we showed that the term-wise majorant $E_0(g)$ is of moderate growth and hence so is $|E_{P,s}|$. Since $\overline{f}$ is rapidly decreasing on $\Gamma\backslash G$, then so is $E_{P,s}\cdot\overline{f}$ (and hence this product is integrable). By the converse described just before the proposition, this implies $\varphi_{P,s}\cdot\overline{f}$ is integrable on $\Gamma_P\backslash G$ as desired.

To finish the proposition, we are now reduced to proving that

$$\langle \varphi_{P,s}, f\rangle_{\Gamma_P\backslash G} = \langle \varphi_{P,s}, f_P\rangle_{N\backslash G}.$$

Again we use (40), but now with $L = \Gamma_P N$ and $M = N$, which gives (since $\varphi_{P,s}$ is left-Γ_P invariant):

$$\langle \varphi_{P,s}, f\rangle_{\Gamma_P\backslash G} = \int_{\Gamma_P N\backslash G} \varphi_{P,s}(\dot{x})\, d\dot{x} \int_{\Gamma_P\backslash\Gamma_P N} \overline{f}(n\cdot\dot{x})\, dn.$$

Then we are finished provided that

$$\int_{\Gamma_P N\backslash G} \varphi_{P,s}(\dot{x})\, d\dot{x} \int_{\Gamma_P\backslash\Gamma_P N} \overline{f}(n\cdot\dot{x})\, dn = \int_{N\backslash G} \varphi_{P,s}(\dot{x})\, d\dot{x} \int_{\Gamma_N\backslash N} \overline{f}(n\cdot\dot{x})\, dn.$$

This is almost trivial – remember that Γ_N is a subgroup of index at most 2 in Γ_P – and is immediate in the case that $\Gamma_P = \Gamma_N$. Suppose not, then $N\backslash G$ is a two-fold cover of $\Gamma_P N\backslash G$ but there is a compensating factor of 2 from the integral over $\Gamma_P\backslash\Gamma_P N$, which gives $2\overline{f}_P$. □

Corollary 10.1.4. *For* $\mathrm{Re}(s) > 1$*, the Eisenstein series* $E_{P,s}$ *is orthogonal to all cusp forms.*

10.2 Constant terms of Eisenstein series

Lemma 10.2.1. *Let* f *be an automorphic form for* Γ *of right* K*-type* m*. Assume that* f *is an eigenfunction of* $\mathcal{C}$ *with eigenvalue* $(s^2-1)/2$ *for* $s \neq 0$*. Let* P *be a cuspidal parabolic subgroup for* Γ*. Then there exist functions constants* $d, c \in \mathbb{C}$ *such that*

$$f_P = d\,\varphi_{P,s} + c\,\varphi_{P,-s}$$

Proof. Recall that the constant term is given by

$$f_P(g) = \int_{\Gamma_N\backslash N} f(ng)\, dn,$$

and in particular is left N-invariant and of K-type m on the right, so as usual, is determined by its restriction to A. Apply the same argument as in part 3 of Theorem 10.1.2 (that is, use that A normalizes N to determine that the action of the generator E of $\mathfrak{g}$ is trivial). Then we find that f_P (as a function on A) satisfies the ODE

$$\left(\frac{1}{2}H^2 - H\right) f_P(a) = \frac{1}{2}(s^2-1) f_P(a)$$

Thinking of the parameter λ as an eigenvalue for H, then for $s \neq 0$ we have:

$$(\lambda^2/2) - \lambda = (s^2-1)/2 \iff \lambda = 1 \pm s.$$

This shows that any $f_P(a)$ satisfying the ODE is a linear combination of h_P^{1+s} and h_P^{1-s} with constant coefficients. (Again, this can be seen as in part 3 of Theorem 10.1.2 by identifying $\mathrm{Lie}(A)$ with $\mathbb{R}$.) The functions $h_P^{1\pm s}$ are left N-invariant and right K-invariant as functions on G, while f_P is of K-type m on the right. Thus, in extending

the identity $f_P = \mu h_P^{1+s} + \nu h_P^{1-s}$ on A to an identity on G, we see that μ and ν are in $F(P, m)$:

$$F(P, m) = \{\varphi \in C^\infty(\Gamma_P NA\backslash G) \mid \varphi \text{ of right } K\text{-type } m\}.$$

Recall that this space is at most one-dimensional with generators $\varphi_{P,m}$, so in fact we could write

$$f_P = d\,\varphi_{P,s} + c\,\varphi_{P,-s} \quad \text{where } d, c \in \mathbb{C}.$$

□

We now apply this result to Eisenstein series, remembering that we have two parabolic subgroups to keep track of – the group $P = NA$ from which we form the quotient $\Gamma_P\backslash\Gamma$ defining the series $E_{P,s}$ and a second parabolic P' at which we perform the Fourier-Whittaker expansion. We will write $E(P, s)_{P'}$ for the constant term of of $E_{P,s}$ at P'. From the previous lemma, we know that for $\mathrm{Re}(s) > 1$, there exist functions which we denote $d_P^{(P')}(s)$ and $c_P^{(P')}(s)$ such that

$$E(P, s)_{P'} = d_P^{(P')}(s)\,\varphi_{P',s} + c_P^{(P')}(s)\,\varphi_{P',-s}. \tag{41}$$

Proposition 10.2.2.

1. $d_P^{(P)}(s) \equiv 1$.
2. *If P' is not conjugate to P (by an element of Γ), then* $d_P^{(P')}(s) = 0$.
3. *There exists a constant $C > 1$ such that*

$$|c_P^{(P')}(s)| \prec C^\sigma \cdot (1 - \sigma)^{-1} \quad \textit{where } \sigma = \mathrm{Re}(s).$$

Proof. Part 2 follows from Theorem 10.1.2, part 5. Recall that for P' not Γ-conjugate to P, this gave

$$E(P, s)(x)h_{P'}^{-(s+1)}(x) \to 0 \quad \text{as} \quad h_{P'}(x) \to \infty, x \in \mathfrak{S}'$$

where the convergence is uniform on compact sets of $\Gamma\backslash G$. Thus the same is true with $E(P, s)_{P'}$ replacing $E(P, s)$ in the displayed equation above. (After all, their difference is rapidly decreasing on $\mathfrak{S}'$.) So dividing both sides of (41) by $h_{P'}^{s+1}(x)$ and taking the limit as $h_{P'}(x) \to \infty$ gives

$$0 = \left(d_P^{(P')}(s)\,\varphi_{P'} h_{P'}^{1+s}(x) + c_P^{(P')}(s)\,\varphi_{P'} h_{P'}^{1-s}(x)\right) h_{P'}^{-(s+1)}(x) = d_P^{(P')}(s).$$

The proof of Part 1 is surprisingly thorny. The basic idea follows just as in the classical case – we use a Bruhat decomposition for G to write

$$\Gamma = \Gamma_P \dot\cup \ (\Gamma \cap PwN)\,, \quad w = \begin{pmatrix} & -1 \\ 1 & \end{pmatrix}.$$

Letting $\Gamma_w := \Gamma \cap PwN$ for convenience, we have the identity of quotients

$$\Gamma_P\backslash\Gamma = \{1\} \dot\cup \Gamma_P\backslash\Gamma_w.$$

(PwN is often referred to as the "big cell" in the Bruhat decomposition.) Thus we may decompose the Eisenstein series into partial sums E_1 and E_w accordingly, with $E_1 = \varphi_{P,s}$. This function is already left-N invariant, so taking its constant term (with respect to P) has no effect: $(E_1)_P = \varphi_{P,s}$. Thus it remains to show that $(E_w)_P$ has coefficient of $\varphi_{P,s}$ equal to 0, according to (41).

If we can show that

$$(E_w)_P(ax) = h_P(a)^{1-s}(E_w)_P(x), \quad \text{for } a \in A, x \in G, \tag{42}$$

then part 1 follows. Indeed, then $(E_w)_P h_P^{s-1}$ is left-invariant under A (and also left-invariant by N and of right K-type m by definition). Thus is belongs to $F(P,m)$ and so

$$(E_w)_P = c_P^{(P')}(s)\varphi_{P,-s} \quad \text{for } \sigma > 1$$

for some well-defined function $c_P^{(P')}(s)$. Thus it suffices to show (42), which requires a careful analysis of the integral defining $(E_w)_P$. Recall from our definitions

$$E_w(ax) = \sum_{\gamma\in\Gamma_P\backslash\Gamma_w} \varphi_{P,s}(\gamma ax) \Longrightarrow (E_w)_P(ax) = \sum_{\gamma\in\Gamma_P\backslash\Gamma_w} \int_{\Gamma_N\backslash N} \varphi_{P,s}(\gamma nax)\, dn.$$

Our plan is to rewrite $\Gamma_P\backslash\Gamma_w = (\Gamma_P\backslash\Gamma_w/\Gamma_N)\cdot\Gamma_N$ and then unfold the integral over $\Gamma_N\backslash N$ to one over N.

Lemma 10.2.3. *If $\gamma \in \Gamma_w$ and $\delta, \delta' \in \Gamma_N$, then*

$$\Gamma_P\gamma\delta = \Gamma_P\gamma\delta' \Longrightarrow \delta = \delta'.$$

Proof of lemma. Express $\gamma = p_\gamma w n_\gamma$ and then note that the uniqueness in the Bruhat decomposition for PwN implies $n_\gamma\delta = n_\gamma\delta'$. □

Using the lemma,

$$(E_w)_P(ax) = \sum_{\gamma\in\Gamma_P\backslash\Gamma_w/\Gamma_N}\sum_{\delta\in\Gamma_N}\int_{\Gamma_N\backslash N}\varphi_{P,s}(\gamma\delta nax)\,dn = \sum_{\gamma\in\Gamma_P\backslash\Gamma_w/\Gamma_N}\int_N \varphi_{P,s}(\gamma nax)\,dn.$$

To demonstrate (42), it suffices to show that

$$\int_N \varphi_{P,s}(\gamma nax)\,dn = h_P^{1-s}(a)\int_N \varphi_{P,s}(\gamma nx)\,dn.$$

To show this, we prove that for $\gamma\in\Gamma_w$,

$$\varphi_{P,s}(\gamma nax) = h_P^{1+s}(a^{-1})\varphi_{P,s}(\gamma n_\gamma^{-1}\,({}^{a^{-1}}n_\gamma)\,({}^{a^{-1}}n)x)$$

then integrate over N, making the change of variables $n_\gamma^{-1}\,({}^{a^{-1}}n_\gamma)\,({}^{a^{-1}}n)\mapsto n'$. The first two elements (given in terms of n_γ) are constant multiples that do not change the measure. The last conjugation by a has the effect of $dn = h_P(a)^2\,dn'$, and this gives the integral identity over N, and with it, Part 1 at last.

To prove Part 3 of the proposition, we use Theorem 10.1.2, part (2), to conclude that

$$|E(P,s)_{P'}(x)| \le \int_{\Gamma_{N'}\backslash N'} E_0(nx)\,dx \prec \frac{c^\sigma}{\sigma-1}h_{P'}(x)^{\sigma+1}\quad(\sigma>1, x\in\mathfrak{S}').$$

If $P=P'$, then by Part 1,

$$|c_P^{(P)}(s)h_{P'}^{1-\sigma}(x)| \le h_{P'}(x)^{1+\sigma} + \frac{c^\sigma}{\sigma-1}h_{P'}(x)^{\sigma+1}.$$

If P is not Γ-conjugate to P', then

$$|c_P^{(P)}(s)h_{P'}^{1-\sigma}(x)| \le \frac{c^\sigma}{\sigma-1}h_{P'}(x)^{\sigma+1}.$$

Choose $x\in\mathfrak{S}'$ such that $h_{P'}(x)>1$. Then since $c^\sigma(\sigma-1)^{-1}$ has strictly positive lower bound (as $c>1$), then the claim follows by combining these two cases. (Remember that $h_P(x) = h_{P'}({}^k x)$ if ${}^kP' = P$. But in any case, we're often considering only a list of inequivalent cusps for Γ.) □

10.3 Outline of Proof of Analytic Continuation

The basic goal is to prove a meromorphic continuation for Eisenstein series to all of $\mathbb{C}$, together with a functional equation. The proof method is usually attributed to Joseph Bernstein, patterned off of the third proof of Selberg (who gave the first proof in 1956).

Recall from the previous section that Γ_N has index ≤ 2 in Γ_P. The case that $\Gamma_N = \Gamma_P$ is equivalent to assuming that Γ_P is *neat*. Borel defines a subgroup L of $\mathrm{GL}(n, \mathbb{R})$ to be *neat* if, for each $x \in L$, the subgroup $A(x)$ of $\mathbb{C}^\times$ generated by the eigenvalues of x is torsion-free.

By abuse, we may refer to the cusps of the corresponding parabolics as neat. For ease of discussion, we'll number the cuspidal parabolics P_i of Γ for $i = 1, \ldots, \ell$ as follows. Let $\delta \in \{0, 1\}$ such that $\delta \equiv m \ (2)$. Then set

$$\ell(\delta) = \begin{cases} \ell & \text{if } \delta = 0. \\ \#\text{ of neat cusps} & \text{if } \delta = 1. \end{cases}$$

Then number the cuspidal parabolics so that the first $\ell(1)$ of them are neat. This notational device is convenient since $E(P_j, s) = 0$ for $j > \ell(\delta)$ as we have shown that the vector space $F(P_j, m) = 0$ in this case. (Similarly for $f \in C(\Gamma\backslash G)$ with f of right K-type m, then $f_{P_i} = 0$ for $i > \ell(\delta)$.) For further convenience, we drop the P for the parabolic and just use its index, so will write $E_{j,s}$ or $E_j(s)$ for $E_{P_j,s}$, etc.

By Proposition 10.2.2, the constant term $E_i(s)_{P_j}$ is of the form

$$E_j(s)_{P_i} = \delta_{i,j}\varphi_{i,s} + c_{j,i}(s)\varphi_{i,-s} \quad (i, j \leq \ell(\delta)).$$

Then form the vector

$$\mathbf{E}(s) = (E_1(s), \ldots, E_{\ell(\delta)}(s)),$$

and let $\mathbf{C}(s) = (c_{i,j}(s))_{i,j}$, the transpose of the matrix of constant term contributions. These are both holomorphic if $\mathrm{Re}(s) = \sigma > 1$. We will show that both admit a meromorphic continuation to all of $\mathbb{C}$ and satisfy the functional equations

$$\mathbf{E}(-s) = \mathbf{C}(-s)\mathbf{E}(s), \qquad \mathbf{C}(s)\mathbf{C}(-s) = 1.$$

In brief, we use what's often referred to as "the continuation principle." Roughly stated, let X_s be a system of linear equations depending holomorphically on a parameter s in a connected, complex manifold. Suppose that for all s in some non-empty open subset U, the system has a unique solution v_s. The "continuation principle"

asserts that the solution v_s extends to a meromorphic function $s \mapsto v_s$ on the entire complex manifold. Moreover, outside of a proper analytic subset (thus almost everywhere), v_s is the unique solution of X_s.

In order to concoct our system of linear equations varying holomorphically, we study a certain compact operator. Very roughly, compact operators have very nice spectral properties which allow us to build holomorphic functions out of their resolvent sets. The operator we will use is built from several different operators, starting with Arthur's truncation operator.

Given a locally L^1 function on $\Gamma\backslash G$, of right K-type m, the truncation operator Λ^t is defined by

$$\Lambda^t(f) = f - \sum_i \psi_i f_{P_i}, \quad \text{where } \psi_i \text{ is the characteristic function of } \mathfrak{S}_{i,t}.$$

We will focus on the operator on the set

$$H_D = \{f \in L^2(\Gamma\backslash G) \mid f \text{ of } K\text{-type } m;\ f_{P_i}|_{\mathfrak{S}_i} = 0,\ i = 1, \ldots, \ell\},$$

where D in the notation H_D is the complement of $\cup\mathfrak{S}_i$ in a fundamental domain for $\Gamma\backslash G$. Its clearly a subset of the cuspidal part of $L^2(\Gamma\backslash G)$ and should be viewed as a kind of relative cuspidality condition – more precisely, functions in H_D are cuspidal away from D.

When we proved the finite dimensionality of automorphic forms, we showed that ${}^\circ L^2(\Gamma\backslash G)$ is closed in $L^2(\Gamma\backslash G)$ (see Lemma 9.2.2). This was shown by constructing a family of continuous linear functionals (integration against functions in $C_c^\infty(N\backslash G)$) whose intersections of kernels gave the cuspidal part. The same proof can be replicated to show that H_D is closed in $L^2(\Gamma\backslash G)$.

The truncation operator Λ^t is idempotent (since $(f_P)_P = f_P$). That is, as a map from $L^2(\Gamma\backslash G)$ to itself, it acts as a projection onto H_D. In fact, we claim this is an orthogonal projection (meaning that $\ker(\Lambda^t)$ is orthogonal to its image). A short exercise in manipulating inner products shows that a projection is orthogonal if and only if the operator is self-adjoint. Since

$$\langle \Lambda^t f, g\rangle = \langle f, g\rangle - \sum_i \langle \psi_{i,t} g_{P_i}\rangle$$

then to prove self-adjointness, it suffices to show that

$$\langle \psi_{i,t} f_{P_i}, g\rangle = \langle f, \psi_{i,t} g_{P_i}\rangle, \quad i = 1, \ldots, \ell.$$

This, too, is a simple integration exercise in which we interchange the order of integration after making a natural change of variables.

Previously, we defined

$$I_c^\infty(G) = \{f \in C_c^\infty(G) \mid f(gk) = f(kg) \text{ for all } k \in K\}.$$

Recall that a bounded operator A is compact if it transforms bounded sets to relatively compact sets.

Lemma 10.3.1. *Let $\alpha \in I_c^\infty(G)$. Then convolution on the right by α, denoted $*\alpha$, is a compact operator from H_D to $L^2(\Gamma\backslash G)$.*

Proof. See Theorem 9.7 of Borel. One proves a bound on derivatives of convolution with cuspidal functions f in terms of a constant (independent of f) times $||f||_2$, then applies Ascoli's theorem. □

Corollary 10.3.2. *Given any $\alpha \in I_c^\infty(G)$, then $\Lambda^t \circ (*\alpha)$ is a compact operator from H_D to H_D.*

Proof. Since $*\alpha$ is compact, and the projection operator Λ^t is continuous, then the composition is compact. □

We briefly record the additional facts we need (with references) about compact operators. Let H be a separable Hilbert space and denote by $\mathcal{L}(H)$ the algebra of bounded operators on H. This carries the structure of a Banach algebra with unit (with respect to the operator norm topology). See for example, Theorem III.2 and Section VI.1 of Reed and Simon.

Given a linear operator A, consider the operator $A_\lambda := \lambda I - A$ for a complex number $\lambda \in \mathbb{C}$ (where I denotes the identity transform). By *spectral theory* of A, we generally mean the study of λ for which $\lambda I - A$ has an inverse. The set of $\lambda \in \mathbb{C}$ such that $\lambda I - A$ has dense range and bounded (left) inverse is called the resolvent set of A, denoted $\rho(A)$. Let $R(\lambda, A) = (\lambda I - A)^{-1}$, which is called the resolvent. The numbers in $\mathbb{C} - \rho(A)$ are called the spectrum, denoted $\sigma(A)$, in analogy with the finite-dimensional case. The spectrum of A is divided into three disjoint sets –

- λ such that A_λ has no inverse (discrete spectrum $\sigma_p(A)$),
- λ such that A_λ has discontinuous, dense inverse (continuous spectrum), and
- λ such that A_λ has non-dense inverse (residual spectrum).

The discrete spectrum σ_p can also be characterized as the λ for which A has a non-zero eigenvector.

Theorem 10.3.3. *Let A be a closed linear operator on $\mathcal{L}(H)$. Then the resolvent set $\rho(A)$ is an open set in the complex plane and the map $\lambda \mapsto R(\lambda, A)$ is holomorphic on each connected component.*

Proof. See Yosida's "Functional Analysis," Theorem VIII.2.1, p. 211. Let $\lambda_0 \in \rho(A)$ and consider the power series

$$S(\lambda) = R(\lambda_0; A)\left(I + \sum_{n=1}^{\infty} (\lambda_0 - \lambda)^n R(\lambda_0; A)^n\right).$$

The series is convergent in the operator norm whenever $|\lambda_0 - \lambda| \cdot ||R(\lambda_0; A)|| < 1$ and defines a holomorphic function of λ on this circle. To see that $S(\lambda)$ represents $R(\lambda; A)$, note that multiplication by $\lambda I - A = (\lambda - \lambda_0)I + (\lambda_0 I - A)$ on the left and right gives the identity. □

In the case of compact operators, we can say something much stronger.

Theorem 10.3.4 (Riesz-Schauder)**.** *Let A be a compact operator on $\mathcal{L}(H)$. Then $\sigma(A) \subset \sigma_p(A) \cup \{0\}$.*

Proof. See Yosida, Theorem X.5.1, p. 283. □

If moreover A is self-adjoint, then the eigenvalues of A are real, so combining the two theorems gives that $\lambda \mapsto R(\lambda, A)$ is holomorphic on $\mathbb{C}^\times - \mathbb{R}^\times$. We will take $A = \Lambda^t \circ (*\alpha)$ with $\alpha \in I_c^\infty(G)$. For λ, we will use the function

$$\lambda_\alpha(s) = \int_G \varphi_{P,s}(g)\alpha(g^{-1})\, dg,$$

which is entire in s since $\varphi_{P,s}$ is entire. It's also of right K-type m and independent of P since $\alpha \in I_c^\infty(G)$. Not surprisingly, $\lambda_\alpha(s)$ also satisfies nice relations with respect to $\varphi_{P,s}$ and $E(P, s)$.

Lemma 10.3.5. *Let $\alpha \in I_c^\infty(G)$ with λ_α defined as above. Then:*

$$\varphi_{P,s} * \alpha = \lambda_\alpha(s)\varphi_{P,s} \quad (s \in \mathbb{C}), \quad E(P, s) * \alpha = \lambda_\alpha(s)E(P, s) \quad (\mathrm{Re}(s) > 1).$$

Proof. The second equation follows from the first, since the summation for E in terms of $\varphi_{P,s}$ converges absolutely and locally uniformly. Thus, we may switch the order of integration and summation. To prove the first equality, note that both sides are

of right K-type m, so it suffices to prove the identity for $g = na$ with $n \in N, a \in A$. Writing out the definitions:

$$(\varphi_{P,s} * \alpha)(na) = \int_G \varphi_P(nag) h_P(nag)^{s+1} \alpha(g^{-1})\, dg.$$

We may simplify this expression, since φ_P is left invariant under NA and $h_P(nag) = h_P(a)h_P(g)$. This gives:

$$h_P(a)^{s+1} \int_G \varphi_P(g) h_P(g)^{s+1} \alpha(g^{-1})\, dg = h_P(a)^{s+1} \lambda_\alpha(s) = \varphi_{P,s}(na)\lambda_\alpha(s).$$

The last equality follows because h_P is left N-invariant and $\varphi_P(na) \equiv 1$. □

Then, by our discussion above, the composition

$$s \longmapsto \lambda_\alpha(s) \longmapsto (\Lambda^t \circ (*\alpha) - \lambda_\alpha(s))^{-1}$$

will define a meromorphic function on some neighborhood U of $s_0 \in \mathbb{C}$, provided we can show that $\lambda_\alpha(s)$ is bounded away from 0 on U.

Lemma 10.3.6. *Given any $s_0 \in \mathbb{C}$, there exists a neighborhood U of s_0 and a function $\alpha \in C_c^\infty(G)$ such that*

$$|\lambda_\alpha(s)| \geq 1/2 \quad (s \in U).$$

(Such a pair (U, λ_α) is said to be *compatible*.)

Proof. Choose a Dirac sequence in $I_c^\infty(G)$. Then by Proposition 8.5.1,

$$\varphi_{P,s} * \alpha_n \to \varphi_{P,s} \quad \text{uniformly on compact sets.}$$

According to Lemma 10.3.5, we have

$$\varphi_{P,s} * \alpha_n = \lambda_{\alpha_n}(s)\varphi_{P,s}.$$

Since $\varphi_{P,s}$ is nowhere 0, then there exists an $\alpha \in I_c^\infty(G)$ such that $\lambda_\alpha(s_0) \neq 0$. Hence we may find an α such that $\lambda_\alpha(s_0) = 1$, and the existence of a neighborhood U is clear. □

Choose such a compatible pair (U, λ_α) and let $\mu = (\mu_{\pm 1}, \ldots, \mu_{\pm \ell(\delta)})$ be a vector of meromorphic functions restricted to U (i.e. each is meromorphic on a domain containing the closure of U). Then define

$$\Psi_\mu(s) = \sum_{i=1}^{\ell(\delta)} (\mu_i \psi_i \varphi_{i,s} + \mu_{-i} \psi_i \varphi_{i,-s}).$$

For fixed s, this defines a function of moderate growth on $\Gamma\backslash G$. In practice, we'll want to consider $\mu_i = \delta_{i,j}$ and $\mu_{-i} = c_{i,j}(s)$.

Lemma 10.3.7. *Given a choice of μ as above, there exists a unique meromorphic function $F_\mu(s) := F_\mu(s, g)$ from U to functions of moderate growth on $\Gamma\backslash G$ of right K-type m such that:*

1. *$s \mapsto F_\mu(s) - \Psi_\mu(s)$ is a meromorphic function from U to H_D.*
2. *For $\alpha \in I_c^\infty(G)$ compatible with U,*

$$\Lambda^t(F_\mu(s) * \alpha) = \lambda_\alpha(s) \Lambda^t(F_\mu(s)).$$

Strictly speaking, we don't use compatibility in the proof. We just need a choice of $\alpha \in C_c^\infty(G)$, but in practice we will take α compatible with U as above.

Proof. Suppose that there exists a function $F(s)$ and a $g(s) \in H_D$ such that

$$F(s) = g(s) + \Psi_\mu(s), \quad \text{and } F \text{ satisfies condition 2.}$$

Because $\varphi_{\pm s}$ is left N-invariant, $\Lambda^t(\Psi_\mu(s)) = 0$. Thus we have

$$\Lambda^t(F(s) * \alpha) = \Lambda^t(g(s) * \alpha) + \Lambda^t(\Psi_\mu(s) * \alpha) = \lambda_\alpha(s) \Lambda^t(g_s) = \lambda_\alpha(s) g(s),$$

where the second equality follows from assuming F satisfies condition 2, and the last equality follows because Λ^t acts as projection onto H_D and we assumed $g(s) \in H_D$. Rearranging, we may write

$$(\Lambda^t \circ (*\alpha) - \lambda_\alpha(s)) g(s) = -\Lambda^t(\Psi_\mu(s) * \alpha). \tag{43}$$

The right-hand side is in H_D because we need only check that $\Psi_\mu(s) * \alpha$ is in $L^2(\Gamma\backslash G)$ (the truncation operator ensures that it will vanish at the cusps). On Siegel sets $\mathfrak{S}$, $\Lambda^t(f * \alpha) = f * \alpha - (f * \alpha)_P$ is rapidly decreasing, hence in $L^2(\mathfrak{S})$. Since convolution is a smoothing operator, $f * \alpha$ is smooth and hence L^2 on the compact set D. Thus,

the right-hand side of (43) is in H_D and so we may apply $(\Lambda^t \circ (*\alpha) - \lambda_\alpha(s))^{-1}$ to both sides of (43) to obtain:

$$g(s) = -(\Lambda^t \circ (*\alpha) - \lambda_\alpha(s))^{-1}(\Lambda^t(\Psi_\mu(s) * \alpha)). \tag{44}$$

This shows that any such g is unique, being determined by data independent of F. Moreover, taking the above identity as the definition of g then $F = g + \Psi_\mu$ is uniquely determined as well, and satisfies both conditions. □

Notice that for fixed s, g in (44) depends linearly on $\mu_{\pm i}$ and of right K-type m. Thus we may express $g(s)$ in the form

$$g(s) = v_i(s)\mu_i(s) + v_{-i}(s)\mu_{-i}(s) \quad \text{where } s \mapsto v_{\pm i}(s) \text{ is meromorphic in } H_D.$$

We further impose the condition that F is annihilated by $(\mathcal{C} - (s^2-1)/2)$. This is meant in the distribution sense:

$$\int_G F_\mu(s,g) \cdot (\mathcal{C} - (s^2-1)/2)\varphi(g)\, dg = 0 \quad (\varphi \in C_c^\infty(G)). \tag{45}$$

Substituting in $F_\mu(s) = g(s) + \Psi_\mu(s)$, the above integral is a homogeneous linear equation in the $\mu_{\pm i}$ having coefficients in the meromorphic functions of U.

Now for each $j = 1, \ldots, \ell(\delta)$, let $\mathcal{L}_j$ denote the system of linear equations in the $\mu_{\pm i}$ coming from:

- The integral (45) for any $\varphi \in C_c^\infty(G)$,
- Condition 1 from the Lemma 10.3.7, and
- the conditions $\mu_i = \delta_{i,j}$ for all $i = 1, \ldots, \ell(\delta)$.

Adding further Condition 2 from Lemma 10.3.7 for some function $\alpha \in C_c^\infty(G)$, we refer to the resulting linear system as $\mathcal{L}_{j,\alpha}$.

Lemma 10.3.8. *Suppose that U is contained in the right-half plane $\mathrm{Re}(s) > 1$. Then $E_j(s)$ is the only function of moderate growth on $\Gamma\backslash G$ and right K-type m satisfying $\mathcal{L}_{j,\alpha}$ for any $\alpha \in I_c^\infty(G)$ compatible with U. Moreover, it is holomorphic from U to $C^\infty(G)$.*

Proof. $E_j(s)$ satisfies all of the linear equations in $\mathcal{L}_{j,\alpha}$ on U from our previous discussion – it's a smooth eigenfunction of $\mathcal{C}$ with eigenvalue $(s^2-1)/2$, $E_j(s) - \Psi_\mu(s)$ is in H_D according to our analysis of the constant term, and $E_j(s) * \alpha = \lambda_\alpha(s)E_j(s)$ by

Lemma 10.3.5. If $F(s)$ is another such function satisfying $\mathcal{L}_{j,\alpha}$ then F has moderate growth, K-type m, and is $\mathcal{Z}$-finite; in short, F is automorphic.

Consider the function $R(s) = E_j(s) - F(s)$. For fixed s in U, its constant term $R(s)_{P_i}$ is a constant multiple of $\varphi_{i,-s}$ since both E_j and F had constant terms of form $\delta_{i,j}\varphi_{i,s} + c(s)\varphi_{-i,s}$, with possibly different $c(s)$. By assumption $\mathrm{Re}(s) > 1$, so $\varphi_{i,-s}$ is square integrable on $\mathfrak{S}_i$. Because $R(s) - R(s)_{P_i}$ is rapidly decreasing on $\mathfrak{S}_i$, then $R(s)$ is also square integrable on $\mathfrak{S}_i$. As this holds for all $i = 1, \ldots, \ell$ and $R(s)$ is an eigenfunction of $\mathcal{C}$, then $R(s)$ is in $L^2(\Gamma\backslash G)$. We claim (but postpone the proof that) any automorphic form in $L^2(\Gamma\backslash G)$ which is an eigenfunction of $\mathcal{C}$ is either 0 or has real eigenvalue. This implies $R(s) = 0$ for $s \in U, s \notin \mathbb{R}$ and hence is identically 0. □

We will apply the continuation principle (once we've appropriately formulated it) to the system $\mathcal{L}_{j,\alpha}$.

10.4 The meromorphic continuation principle

Let V be a topological vector space and consider a system of linear equations given by the triple

$$X_0 = \{(V_i, T_i, v_i)\}_{i \in I},$$

where the V_i are topological vector spaces, $T_i : V \to V_i$ are continuous linear maps, and the v_i are the column vector of desired equalities. The set I need not be countable. By a solution to X_0, we mean a vector $v \in V$ such that $T_i(v) = v_i$ for all i. Further let $\mathrm{Soln}(X_0)$ be the set of solution vectors $v \in V$.

More generally, we may allow the system to vary in a complex parameter space s (or in any connected, complex manifold). That is, for each $s \in \mathbb{C}$ consider

$$X_s = \{(V_i, T_{i,s}, v_{i,s})\}.$$

Our continuation principle will be phrased in terms of "locally finite" systems of linear equations, whose definition we now begin to explain. First we remind the reader of the notion of "weakly holomorphic."

Definition 10.2 (weakly holomorphic)**.** *Given a topological vector space V, a V-valued function $s \mapsto f(s)$ is said to be "weakly holomorphic" if $s \mapsto \lambda(f(s))$ is holomorphic for every continuous linear functional λ on V.*

Similarly, a family of operators $T_s : V \to W$ between topological vector spaces is "weakly holomorphic" in a parameter s if, for every vector $v \in V$ and every continuous functional $\mu \in W^*$, the function $\mu(T_s(v))$ is holomorphic in s.

We say the system $\{X_s\}$ is holomorphic if both $T_{i,s}$ and $v_{i,s}$ are "weakly holomorphic" in s.

Definition 10.3 (finite type)**.** *Let $\{X_s\}$ be a parametrized system of linear equations in a space V which is holomorphic in s. Suppose there exists a finite-dimensional subspace F and a weakly holomorphic family of continuous linear functionals $f_s : F \to V$ such that for each s, $\mathrm{Im}(f_s) \supset \mathrm{Soln}(X_s)$. Then we say that $\{f_s\}$ is a "finite holomorphic envelope" for the system X and that X is of "finite type."*

Note that if we are given a meromorphic continuation to a unique solution v_s, then taking $F = \mathbb{C}$ and $f_s : \mathbb{C} \to V$ defined by $f_s(z) = z \cdot v_s$ then we trivially obtain a finite holomorphic envelope for parameter values s away from the poles of v_s.

Definition 10.4 (locally finite type)**.** *Let U_α for $\alpha \in A$ be an open cover of the parameter space. If, for each $\alpha \in A$, we have a finite envelope $\{f_s^{(\alpha)} : s \in U_\alpha\}$ for the system $X^{(\alpha)} = \{X_s : s \in U_\alpha\}$. Then we say that $\{f_s^{(\alpha)}, U_\alpha\}_{\alpha \in A}$ is a locally finite holomorphic envelope and X is of locally finite type.*

With all of these definitions in hand, we can at last state the continuation principle precisely.

Theorem 10.4.1 (Continuation Principle)**.** *Let $\{X_s\}$ be a locally finite system of linear equations*

$$T_{i,s} : V \longrightarrow V_i,$$

with s varying in a complex, connected manifold. Suppose that each V_i is locally convex and quasi-complete[11] *If for some $s_0 \in U$, a non-empty open subset, there exists a unique solution v_{s_0} then the solution depends meromorphically on $s \in U$, has a meromorphic continuation to all s in the manifold, and for fixed s not in a locally finite set of analytic hypersurfaces of the manifold, the solution v_s is the unique solution to the system X_s.*

The idea of the proof is to translate the locally finite condition into a statement of linear algebra, from which elementary considerations (e.g., Cramer's rule) produce uniqueness and existence of a weakly holomorphic v_s. Then we use weak-to-strong theorems to finish the proof.

[11] The locally convex assumption guarantees, by Hahn-Banach theorem, a supply of continuous linear functionals that separate points. (cf. Chapter IV of Yosida.) The space is said to be *quasi-complete* if every bounded (in weak operator topology) Cauchy net is convergent.

Proof. The locally finite condition implies there exists a family $f_s : F \to V$ of morphisms with $\mathrm{Im}(f_s) \supset \mathrm{Soln}(X_s)$ with F finite dimensional. Set

$$F_s^{\mathrm{Soln}} = \{v \in F \mid f_s(v) \in \mathrm{Soln}(X_s)\} = \{\text{inverse images in } F \text{ of solutions}\}.$$

Then X_s has a unique solution if and only if $\dim(F_s^{\mathrm{Soln}}) = \dim(\ker(f_s))$.

The weak holomorphy of $T_{i,s}$ and f_s implies the weak holomorphy of their composition $T_{i,s} \circ f_s$ from $F \to V_i$. This is a corollary of Hartogs' theorem (which says that separately analytic functions of several variables are jointly analytic). Indeed consider

$$(s,t) \mapsto \lambda(T_{i,s}(f_t(v)))$$

is separately holomorphic in s and t. Apply Hartogs' theorem and set $s = t$.

Now consider the space $\mathrm{Hom}^\circ(F, V_i)$ of continuous maps $T : F \to V_i$ equipped with the weak operator topology as follows: given $x \in F$ and $\mu \in V_i^*$ define the seminorm $p_{x,\mu}$ by

$$p_{x,\mu}(T) = |\mu(T(x))|.$$

Since V_i is quasi-complete and F is finite-dimensional, then $\mathrm{Hom}^\circ(F, V_i)$ is quasi-complete. In this case, the weakly holomorphic family $s \mapsto T_{i,s} \circ f_s$ is in fact holomorphic. This last statement requires the theory of Gelfand-Pettis integrals. See Rudin's "Functional Analysis" (in particular, the section on Holomorphic Functions on p. 78) for some discussion of this.

Identify F with $\mathbb{C}^n$ for some n. Using linear functionals on V and on V_i which separate points, we can describe $\ker(f_s)$ and F_s^{Soln} using systems of linear equations as follows:

$$\ker(f_s) = \left\{ (x_1, \ldots, x_n) \in F \;\middle|\; \sum_j a_{\alpha_j} x_j = 0, \alpha \in A \right\},$$

$$F_s^{\mathrm{Soln}} = \left\{ (x_1, \ldots, x_n) \in F \;\middle|\; \sum_j b_{\beta_j} x_j = c_\beta, \beta \in B \right\},$$

for some $a_{\alpha_j}, b_{\beta_j}, c_\beta$ all holomorphic complex-valued functions of s (and A and B may be index sets of arbitrary cardinality.) We examine the dimensions of the sets above by placing the coefficients into matrices:

$$M_s(\alpha, j) = a_{\alpha,j}, \quad N_s(\beta, j) = b_{\beta,j}, \quad Q_s(\beta, j) = \begin{cases} b_{\beta,j} & \text{for } j \in [1, n] \\ c_\beta & \text{for } j = n+1. \end{cases}$$

Then we have $\dim(\ker(f_s)) = n - \mathrm{rank}(M_s)$, and $\mathrm{rank}(N_s) \leq \mathrm{rank}(Q_s)$ with solutions only if equality holds. Hence our condition that $\dim(F_s^{\mathrm{Soln}}) = \dim(\ker(f_s))$ may be rewritten $\mathrm{rank}(M_s) = \mathrm{rank}(N_s) = \mathrm{rank}(Q_s)$.

By assumption, these ranks are then all equal on U, where the solution v_s is unique. Let S be the dense subset of the parameter space (the complement of an analytic subset) for which all these ranks are maximal. Then $S \cap U$ is non-empty. But since the ranks are fixed (say, equal to r) for all points $s \in S$, then in fact X_s has a unique solution for all $s \in S$.

It remains to find a meromorphic solution v_s for $\{X_s\}$. As this system has a finite envelope, it suffices to find a meromorphic solution to the parametrized system $Y = \{Y_s\}$ where

$$Y_s = \left\{ \sum b_{\beta,i} x_i = c_\beta \;\middle|\; \beta \in B \right\},$$

where the $b_{\beta,i}$ and c_β again implicitly depend on s. Again, letting r be the maximal rank as above, choose an $s_0 \in S$ and an $r \times r$ minor of full rank

$$D_{s_0} = \{b_{\beta,j} \mid \beta \in \{\beta_1, \ldots, \beta_r\}, j \in \{j_1, \ldots, j_r\}\} \subset N_{s_0}.$$

Let $S' \subset S$ be the set of points s such that D_s has full rank (i.e., non-vanishing determinant). Then the system of equations

$$\left\{ \sum_{j \in \{j_1, \ldots, j_r\}} b_{\beta,j} x_j = c_\beta \;\middle|\; \beta \in \{\beta_1, \ldots, \beta_r\} \right\}$$

has a unique solution $(x_{1,s}, \ldots, x_{r,s})$ for $s \in S'$ by Cramer's rule. Since the coefficients are holomorphic in s, the expression for the solution obtained via Cramer's rule (as a quotient of determinants) shows that the solution is meromorphic in s. Then extending the solution by setting $x_j = 0$ for $j \notin \{j_1, \ldots, j_r\}$ then the r equations are rows of the system Y_s. For $s \in S'$ the equality of the ranks of N_s, Q_s with r imply that the solution will automatically satisfy the rest of the equations in the system Y_s. Weak-to-strong principles ensure that the resulting solution is holomorphic, not just weakly holomorphic. □

This guarantees the meromorphic continuation of Eisenstein series. Indeed, we have already verified that the system $\mathcal{L}_{j,\alpha}$ has a unique solution $E_j(s)$ on any neighborhood U contained in the right half plane $\sigma > 1$. Moreover, the system is of finite type. (This is easy since the existence of a meromorphic continuation on any set U guarantees that the system $\mathcal{L}_{j,\alpha}$ is of locally finite type. Note that Lemma 10.3.7 allowed for any open subset U of $\mathbb{C}$.)

Strictly speaking, our analysis has one minor flaw at $s = 0$. This is just a technicality, due to the fact that the constant terms, which have independent solutions to the Casimir operator h_P^{1+s} and h_P^{1-s} for $s \neq 0$, now coincide for $s = 0$ where the

characteristic equation has a double solution. The resulting solutions are thus h_P and $h_P \log h_P$. (We already saw a version of this phenomenon on p. 112 of the notes, in Section 8.1 where eigenfunctions of Δ on $\mathcal{H}$ were studied.) In this case, we may rewrite the constant term in a way that remains valid for $s = 0$ using the pair of functions

$$\beta_{P,s} = (\varphi_{P,s} + \varphi_{P,-s})/2, \quad \gamma_{P,s} = (\varphi_{P,s} - \varphi_{P,-s})/2s,$$

the analog of (25). Building a system of linear equations out of these constant terms and following as above gives the continuation to $s = 0$.

10.5 Continuation consequences: functional equations, etc.

To prove functional equations, consider the column vector $\mathbf{Q}(s) = \mathbf{C}(-s)^{-1}\mathbf{E}(-s)$. The components of this vector are thus

$$Q_i(s) = \sum_m \mathbf{C}(-s)^{-1}_{i,m} E_m(-s)$$

so that its constant term satisfies

$$Q_i(s)_{P_j} = \sum_m \mathbf{C}(-s)^{-1}_{i,m}(\delta_{m,j}\varphi_{j,-s} + c_{m,j}(-s)\varphi_{j,s}) = \mathbf{C}(-s)^{-1}_{i,j}\varphi_{j,-s} + \delta_{i,j}\varphi_{j,s}.$$

Thus, for any neighborhood U of a point at which $\mathbf{C}$, $\mathbf{C}^{-1}$, and $\mathbf{E}$ are holomorphic, the $Q_i(s)$ satisfies the conditions $\mathcal{L}_{i,\alpha}$ for all α compatible with U. Hence $Q_i(s) = E_i(s)$, and the functional equation is proved. Furthermore, $E_i(s)$ and $Q_i(s)$ have the same constant terms, so $\mathbf{C}(-s)^{-1}_{i,j} = c_{i,j}(s)$ which yields the functional equation for the matrix of constant terms: $\mathbf{C}(-s)\mathbf{C}(s) = 1$.

The only gap remaining in our argument is to show that $\det(\mathbf{C})$ is not identically 0, so that we may invert it. Suppose it was identically 0. Then we can find meromorphic functions m_j on $\mathbb{C}$ with $j = 1, \ldots, \ell(\delta)$ not all 0 such that

$$\sum_j m_j c_{j,i} = 0 \quad (i = 1, \ldots, \ell(\delta)).$$

Set $R = \sum_j m_j E_j$, so that

$$R(s)_{P_i} = \sum_j m_j(s) E_j(s)_{P_i} = \sum_j m_j(s)\,(\delta_{j,i}\varphi_{i,s} + c_{j,i}(s)\varphi_{i,-s}) = m_i(s)\varphi_{i,s}.$$

Let U be a domain in the left half-plane $\sigma < 0$ on which the m_j and E_j are holomorphic in s. For $s \in U$, the function $R(s)$ is square-integrable and an eigenfunction of

$\mathcal{C}$ with eigenvalue $(s^2-1)/2$. Therefore, according to one of the questions on this week's homework, either s is real or $R(s)=0$. Thus $R(s)$ must be identically 0, so we have the relation

$$\sum_j m_j(s)E_j(s) = 0 \quad \text{for all } s \in \mathbb{C} \text{ such that } m_j, E_j \text{ holomorphic.}$$

This contradicts our assumption that all of the m_j's were not 0 (as these are functions of G whose sum is identically 0, where the $m_j(s)$ are independent of G).

Proposition 10.5.1. *Suppose $E_j(s)$ has a pole of order $m(c)$ at a point $c \in \mathbb{C}$. Then $\lim_{s\to c}(s-c)^{m(c)} \cdot E_j(s)$ is an automorphic form having right K-type m and an eigenfunction for $\mathcal{C}$ with eigenvalue $(c^2-1)/2$.*

Proof. It is clear that the function $\lim_{s\to c}(s-c)^{m(c)} \cdot E_j(s)$ has the correct K-type and eigenvalue under the Laplacian. We need only show that the function is of moderate growth. Let $D(c,R)$ be a small disc of radius R around c. We may prove the stronger result that there exists positive constants C, N such that

$$(s-c)^{m(c)}E_j(s)(g)| \leq C||g||^N \quad \text{for } s \in D(c,R) \text{ and } g \in \Gamma\backslash G.$$

Since $(s-c)^{m(c)}E_j(s)$ is holomorphic for $s \in D(c,R)$, its constant terms at each P_i are linear combinations with bounded coefficients of the functions $\varphi_{P,s}, \varphi_{P,-s}$. On a Siegel set, these have growth bounded by a function $h_P(g)^d$ where $d > \max(1+\sigma, 1-\sigma)$ ranging over $s \in D(c,R)$. The same is true for $\lim_{s\to c}(s-c)^{m(c)} \cdot E_j(s)_{P_i}$. Write

$$F_j(s) = E_j(s) - \Psi_j(s)$$

where $\Psi_j(s)$ contains the constant terms at each P_i multiplied by a characteristic function on the associated Siegel set. We claim that $(s-c)^{m(c)}(F_j * \alpha)$ is a bounded holomorphic function on $D(c,R)$ (where we've chosen our small disc D to lie inside a compatible neighborhood U for α). This may be proved along the same lines that $*\alpha$ is a compact operator – see for example, Theorem 9.7(ii) of Borel's book. Finally, expressing

$$(s-c)^{m(c)}E_j(s) = \lambda_\alpha(s)^{-1}(s-c)^{m(c)}(F_j(s)*\alpha + \Psi_j(s)*\alpha)$$

the claim follows noting that $||\lambda_\alpha(s)|| \geq 1/2$ on a compatible neighborhood U. $\square$

Proposition 10.5.2. *Let $c \in \mathbb{C}^\times$ and let $E(s) = \sum_j \mu_j E_j(s)$ where $\mu_j \in C_c(\Gamma\backslash G)$. If $E(s)$ has a pole of order n at c then the automorphic form $\lim_{s\to c}(s-c)^n E(s)$ is orthogonal to cusp forms. (In particular, we allow $n=0$.)*

Proof. Choose a domain $D \subset \mathbb{C}$ that contains a non-empty open subset of the half-plane $\sigma > 1$ and for which the $E_j(s)$ are holomorphic except possibly for poles at $s = c$. Given any cusp form f, consider the function

$$s \mapsto \langle (s-c)^n E(s), f \rangle_{\Gamma\backslash G}.$$

By the previous proposition, this is a holomorphic function on D. By our earlier result, it is 0 for $\sigma > 1$ since cusp forms were orthogonal to those Eisenstein series. Hence, the inner product must be identically 0. □

So far, we've said nothing about the polar behavior of the meromorphic functions $E(P_i, s)$ and $c_{i,j}(s)$. To do so, we need to prove certain inner product formulas for Eisenstein series and their constant terms, known as Maass-Selberg relations. In the interest of time, we omit these somewhat lengthy calculations and merely state the desired result.

Theorem 10.5.3. *The functions $E(P_i, s)$ and $c_{i,j}(s)$ are holomorphic in the half-plane $\mathrm{Re}(s) \geq 0$ except for at most finitely many simple poles in the interval $(0, 1]$. At each pole, the residue of $E(P_i, s)$ is a square-integrable automorphic form.*

10.6 Spectral decomposition of $L^2(\Gamma\backslash G)$

We study the spectral decomposition of $L^2(\Gamma\backslash G)$ with respect to the operator $\mathcal{C}$ – that is, a canonical decomposition of our vector space according to eigenspaces for $\mathcal{C}$. Note first that we have the direct sum decomposition

$$L^2(\Gamma\backslash G) = \bigoplus_{m \in \mathbb{Z}} L^2(\Gamma\backslash G)_m \quad L^2(\Gamma\backslash G)_m = \{f \in L^2(\Gamma\backslash G) \mid f(gk) = \chi_m(k) f(g)\},$$

the set of functions having right K-type m. It is easy to see these spaces are orthogonal. To prove that they span, see the hints following Exercise 2.1.5 on p. 144 in Bump's "Automorphic Forms and Representations."

The operator $\mathcal{C}$ is unbounded as an operator on $L^2(\Gamma\backslash G)$ with domain of definition the subspace of smooth vectors. In the homework for this week, you showed that $\mathcal{C}$ is symmetric. That is, for any smooth functions ϕ, ψ we have

$$\langle \mathcal{C}\phi, \psi \rangle = \langle \phi, \mathcal{C}\psi \rangle.$$

The smooth vectors are dense in $L^2(\Gamma\backslash G)$ using convolution with a Dirac sequence of compactly supported, smooth functions. Thus an adjoint is defined and is an extension of $\mathcal{C}$ (i.e. its restriction to smooth vectors coincides with $\mathcal{C}$ and its domain

of definition contains the smooth vectors.) Since adjoints are always closed, then $\mathcal{C}$ has a minimal closure $\overline{\mathcal{C}}$ which is self-adjoint (and turns out to be equal to $\mathcal{C}^*$. We say that $\mathcal{C}$ is *essentially self-adjoint*.) Such essentially self-adjoint operators have a spectral decomposition. We begin by analyzing the cuspidal part ${}^\circ L^2(\Gamma\backslash G)_m$.

Theorem 10.6.1. *The spectrum of $\mathcal{C}$ in ${}^\circ L^2(\Gamma\backslash G)_m$ is discrete with finite multiplicities and ${}^\circ L^2(\Gamma\backslash G)$ has a basis (as a Hilbert space) consisting of countably many eigenfunctions of $\mathcal{C}$. In particular, cusp forms are dense in ${}^\circ L^2(\Gamma\backslash G)_m$.*

Proof. $(*\alpha)$ is a compact operator on ${}^\circ L^2(\Gamma\backslash G)$, as shown in the homework. Choose a Dirac sequence $\{\alpha_n\}$ such that α_n are symmetric. Then the operators $*\alpha_n$ are self-adjoint. Thus we may apply the Riesz-Schauder theorem (see Yosida, X.5, Theorem 2, p. 284) implies that for each n, ${}^\circ L^2(\Gamma\backslash G)_m$ is a countable direct sum of eigenspaces of $*\alpha_n$ and the eigenspaces of non-zero eigenvalues are finite dimensional. Note that these finite dimensional eigenspaces ranging over all n must span ${}^\circ L^2(\Gamma\backslash G)_m$ – otherwise, there would exist a non-zero f in ${}^\circ L^2(\Gamma\backslash G)_m$ having $f * \alpha_n = 0$ for all n but converge to f.

Further, any such eigenspace E is stable under $\mathcal{C}$, which commutes with the $(*\alpha_n)$. This implies that E consists of K-finite and $\mathcal{Z}$-finite cuspidal functions (i.e. cusp forms). The restriction of $\mathcal{C}$ to E is self-adjoint, so we may diagonalize. This is our desired basis of eigenfunctions of $\mathcal{C}$. Recall that the space of automorphic forms with fixed data (in particular, eigenvalue and K-type) is finite dimensional, so the multiplicities are indeed finite. □

Let $V_m := {}^\circ L^2(\Gamma\backslash G)^\perp$. It remains to investigate the spectral theory of this subspace. From the Maass-Selberg relations, we have that the residues of Eisenstein series are square-integrable. We proved earlier that they are also orthogonal to all cusp forms. So they belong to V_m. Another important family of functions appearing in this space are called "incomplete theta series" by Godement. These are defined by

$$ {}_P E_f(x) := E_f(x) = \sum_{\gamma \in \Gamma_P \backslash \Gamma} f(\gamma x), \quad f \in C_c(\Gamma_P N \backslash G). $$

It is not hard to see that E_f converges absolutely and locally uniformly to a function of $C_c(\Gamma\backslash G)$. In fact, one can show that there are only finitely many elements mod Γ_P for which $f(\gamma x)$ is not identically 0 (a consequence of our discontinuous group action and properties of Siegel sets).

If we assume that f has right K-type m (and thus E_f has right K-type m), then we may write

$$ f = u\varphi_{P,m} \quad u \in C_c(A),\ \varphi_{P,m} \in C(\Gamma_P N A \backslash G), $$

where we regard u as a function of G with is both left N and right K invariant. Thus, rather than writing E_f, we may instead write $E_{u,m}$ with $u \in C_c(A)$ when f is of right K-type m.

Proposition 10.6.2. *The orthogonal complement V_m of the cuspidal part ${}^\circ L^2(\Gamma\backslash G)$ contains the residues of Eisenstein series at poles in $(0,1]$ and the incomplete theta series $E_{u,m}$ where (P,A) range over all cuspidal pairs and $u \in C_c^\infty(A)$. V_m is spanned by the latter series when P runs through a complete set of inequivalent parabolics $P_1, \ldots, P_{\ell(\delta)}$.*

Proof. We have already argued that both families belong to $L^2(\Gamma\backslash G)$ and that residues of Eisenstein series are orthogonal to cusp forms. To see that theta series are orthogonal to cusp forms, we use an inner product identity for automorphic forms ψ of right K-type m:

$$\langle \psi,\ {}_P E_u \rangle_{\Gamma\backslash G} = \langle \psi_P, u \rangle_{N\backslash G}$$

where the latter inner product is clearly 0 for cusp forms ψ. It is proved analogous to our earlier inner product identity using quotients of Haar measures.

Conversely, if ψ is orthogonal to all $E_{u,m}$ for a given P, then its constant term is orthogonal to all $u \in C_c^\infty$ from the above inner product identity. Hence, the constant term is identically 0. But if this holds for all P_i then ψ is a cusp form. □

To say more about the decomposition of the space V_m, we need to know inner product formulas for the incomplete theta series. Given $u_j \in C_c^\infty(A_j)$ and $v_k \in C_c^\infty(A_k)$ we write ψ_j and η_k for the corresponding incomplete theta series, in order to try to streamline some of the notation.

Theorem 10.6.3. *Using the above notation,*

$$\langle \psi_j, \eta_k \rangle = \frac{1}{2\pi} \sum_{n=1}^{\ell(\delta)} \int_0^\infty \langle \psi_j, E_n(i\tau) \rangle \cdot \langle E_n(i\tau), \eta_k \rangle \, d\tau + \sum_{z \in J_m} \langle \psi_j, F_j(z) \rangle \cdot \langle F_k(z), \eta_k \rangle \cdot \langle F_j(z), F_k(z) \rangle$$

where F_j denotes the unit vector $E_j'(z)/||E_j(z)||$ and J_m is the finite set of simple poles for Eisenstein series in $(0,1]$.

11 Applications

We now explain the connections between the spectral decomposition of $L^2(\Gamma\backslash G)$ with respect to $\mathcal{C}$ (which made critical use of the theory of automorphic forms) and the decomposition of $L^2(\Gamma\backslash G)$ with respect to the right action by G via irreducible representations. Along the way, we'll see that many of the objects we've considered in constructions of automorphic forms (particularly Eisenstein series) have natural interpretations in terms of representation theory. We fix the group $G = \mathrm{SL}(2,\mathbb{R})$ throughout.

11.1 Review of representation theory

We give just enough background about representation theory of $\mathrm{SL}(2,\mathbb{R})$ needed to draw parallels between our decompositions of $L^2(\Gamma\backslash G)$. Sections 2.4–2.6 of Bump's "Automorphic Forms and Representations" are an excellent reference for the material covered here and in the next section, though the normalizations involving eigenvalues of $\mathcal{C}$ are somewhat different than what we present (which follows Borel).

Given a topological vector space V, a *continuous representation* (π, V) is a representation such that

$$G \times V \longrightarrow V \ : \ (g, v) \longmapsto \pi(g) \cdot v$$

is continuous. If V is a Hilbert space, then (π, V) is *unitary* if the G action leaves the scalar product on V invariant. (Note that in this case, the operator norm $||\pi(g)||$ is uniformly bounded (by 1), so it suffices to check $g \mapsto \pi(g) \cdot v$ is continuous to verify the continuity condition above.)

Assume V is further *locally complete*.[12] This condition is imposed to guarantee that if $\alpha \in C_c(G)$ then $\int_G \alpha(x)\pi(x) \cdot v \, dx$ converges. More generally, we may replace $\alpha(x)\, dx$ by any compactly supported measure $\mu(x)$ on G. We use this to extend the action of π to functions in $C_c^\infty(G)$.

Further, we say $v \in V$ is *differentiable* (or smooth) if $g \mapsto \pi(g) \cdot v$ is a smooth map. Let V^∞ be the set of smooth vectors, a G-stable subspace of V. We may extend the representation restricted to V^∞ to a representation of $\mathcal{U}(\mathfrak{g})$ on $\mathrm{End}(V^\infty)$. If $\alpha \in C_c^\infty(G)$ and $D \in \mathcal{U}(\mathfrak{g})$ then one can check

$$D(\pi(\alpha) \cdot v) = \pi(D(\alpha)) \cdot v$$

[12] This is a very weak form of completeness – every closed, bounded, absolutely convex subset has norm space as a Banach space. This is weaker than even sequential completeness.

so that $V^\infty \supset \pi(C_c^\infty(G)) \cdot v$. We've seen these concepts before using the right regular representation, and now we see that they can be extended to arbitrary continuous representations. This will be a recurring theme.

As with the regular representation, let V_m denote the space of vectors having K-type m: $\pi(k) \cdot v = \chi_m(k)v$. There exists a natural projector from V onto the space V_m defined by

$$\pi_m : v \mapsto \int_K \chi_{-m}(k)\pi(k) \cdot v \, dk,$$

where we may think of $\chi_{-m}(k)\, dk$ as a measure e_m and then this projector is just $\pi(e_m)$ in our earlier notation. If $\alpha \in I_c^\infty(G)$, then $\pi(\alpha)$ commutes with $\pi(k)$ so $\pi(\alpha)$ leaves V_m invariant, and this implies $\pi_m(V^\infty)$ is dense in V_m.

The space of K-finite vectors, those $v \in V$ such that $\pi(K)\cdot v$ is a finite dimensional subspace of V, is expressible as the algebraic direct sum

$$V_K = \bigoplus_{m \in \mathbb{Z}} V_m.$$

If V_m is finite dimensional, then $V_m \subset V^\infty$ since $V_m \cap C_c^\infty(G) \cdot V$ is dense in V_m. Further, $\mathcal{Z}(\mathfrak{g})$ commutes with G (and hence with $\pi(e_m)$) so V_m (and hence also V_K) consists of K-finite, $\mathcal{Z}$-finite vectors. Representations (π, V) such that V_m is finite-dimensional are called *admissible*. This is the right category of representations to study, as their classification is known (Langlands, 1973, for an arbitrary Lie group). We conclude by explaining the notion of equivalence for admissible representations.

– NEED TO INSERT DEFINITION OF (g,K) MODULE –

11.2 Representations of $\mathrm{SL}(2, \mathbb{R})$

Fix a p-pair (P, A). As ever, some care needs to be taken with $\pm I = Z(G)$ in the group G. Write

$$P = Z(G) \cdot P^\circ, \quad P^\circ = NA$$

and let χ_δ denote the character on P, trivial on P° that takes the value $(-1)^\delta$ on $-I$ for $\delta \in \{0, 1\}$.

The function h_P used to define the Eisenstein series, when restricted to P, is a character trivial on $Z(G) \cdot N$. Given $s \in \mathbb{C}$, set

$$\psi_{P,\delta,s} := \psi_{P,s} = \chi_\delta h_P^{1+s},$$

a character of P trivial on N. Now consider the smoothly induced representation of G from $\psi_{P,s}$, called the *principal series* representation. That is, the space:

$$I(s) := \mathrm{Ind}_P^G(\psi_{\delta,s}) = \{f : C^\infty(G) \to \mathbb{C} \mid f(pg) = \psi_{\delta,s}(p)f(g) \text{ for all } p \in P, g \in G\}$$

acted on by right translation by G. One can complete $I(s)$ with respect to the norm

$$\langle u, v\rangle = \int_K u(k)\overline{v(k)}\, dk \quad (u, v \in I(s))$$

to obtain a Hilbert space $H(s)$ – functions satisfying the condition of $I(s)$ and square-integrable when restricted to K.

The representation $I(s)$ is independent of p-pair, as all parabolics are conjugate. Indeed the map

$$\ell(k^{-1}) : u(g) \longrightarrow u(k^{-1}g)$$

is an intertwining operator between $I(P', s)$ and $I(P, s)$ if $P' = {}^kP$.

Proposition 11.2.1.

i) On $I(s)^\infty$, the Casimir operator acts as multiplication by $(s^2 - 1)/2$.

ii) The space $H(s)_m$ of vectors of K-type m is 1-dimensional and spanned by $\varphi_{P,m,s}$ if $m \equiv \delta\ (2)$ and is 0 otherwise.

iii) If $s \notin \mathbb{Z}$ or if $s \in \mathbb{Z}$ and $s \equiv \delta\ (2)$ then the representations $I(\delta, s)$ are irreducible and the corresponding $(\mathfrak{g}, K)$-modules $I(\delta, s)_K$ and $I(\delta, -s)_K$ are isomorphic.

Proof. See Section 2.5 of Bump's "Automorphic Forms and Representations" for careful proofs. For *i*), it suffices to show h_P^{1+s} is an eigenfunction of $\mathcal{C}$ since $\mathcal{C}$ commutes with right translations. This was done before by showing that EF acts by 0 on h_P.

For *ii*), note that $\varphi_{P,m,s}$ is clearly in $H(s)_m$. For the reverse direction, given $f \in I(s)_m$, note that

$$f(nak) = \psi_{\delta,s}(na)\chi_m(k)f(1) = \varphi_{P,m,s}(nak)f(1).$$

Since $I(s)$ is dense and admissible in $H(s)$ then the $(\mathfrak{g}, K)$-modules $I(s)_K$ and $H(s)_K$ are equal and $I(s) = H(s)^\infty$.

Finally, for *iii*), given the conditions on s, we want to show that

$$\mathcal{U}(\mathfrak{g}) \cdot \varphi_{P,m,s} = I(s)_K.$$

This proves irreducibility as then $I(s)_K$ is a Harish-Chandra module (a $(\mathfrak{g}, K)$ module which is finitely-generated over $\mathcal{U}(\mathfrak{g})$) for which infinitesimal irreducibility is equivalent to irreducibility. This is a calculation – use the Lie algebra basis W, Y, Z as in Theorem 8.6.3 and determine its action on $\varphi_{P,m,s}$. The isomorphism of modules falls out from the calculation. □

Proposition 11.2.2. *The principal series $H(\delta, s)$ is contragredient to $H(\delta, -s)$.*[13] *If $s \in i\mathbb{R}$ then $H(s)$ is unitary and equivalent to $H(-s)$ and the functions $\varphi_{P,m,s}$ with $m \equiv \delta$ (2) form an orthonormal basis.*

Proof. Show that the form

$$\langle u, v \rangle = \int_K u(k)v(k)\, dk \quad u \in H(s), v \in H(-s)$$

is a non-degenerate, continuous bilinear form that is G-invariant. Do this by writing $gk = p\ell_k$ for unique choices of $p \in P^\circ$ and $\ell_k \in K$ for each $g \in G, k \in K$. Then use this scalar product to show $\langle \varphi_{P,m,s}, \varphi_{P,n,s} \rangle = \delta_{m,n}$. □

According to our results so far, the only principal series $H(\delta, s)$ which may be reducible are those with $s \in \mathbb{Z}$ such that $s \not\equiv \delta$ (2). These do turn out to be reducible, and understanding their decomposition essentially completes the classification of (admissible) $(\mathfrak{g}, K)$-modules. In any case, the representations can be understood in terms of the K-types that occur within them.

Recall from the finite dimensional representation theory of $\mathrm{SL}(2, \mathbb{C})$ that there is exactly one irreducible representation (modulo equivalence) for each positive integer n. The action is on the space of homogeneous polynomials in two variables of degree n. The Casimir operator has eigenvalue $n^2/2 + n$ and the K-types are integers $m \in [-n, n]$ such that $m \equiv n$ (2). Call these finite dimensional representations F_n.

For $n \in \mathbb{Z} - \{0\}$, G has irreducible representations D_n – the so called *discrete series* representations. They are square integrable, meaning that they can be realized as a G-invariant subspace of $L^2(G)$. The eigenvalue of $\mathcal{C}$ on D_n is $(n^2 - 1)/2$ and the corresponding K-types are integers $m \in \mathbb{Z}$ that rays to $\pm\infty$ from n. That is, $n \equiv m \pmod 2$, $\mathrm{sgn}(n) = \mathrm{sgn}(m)$ and $|m| > |n|$.

It remains to analyze the reducible principal series. Consider $H(\delta, s)$ for $s \in \mathbb{Z}$ with $s \not\equiv \delta \pmod 2$ and $s \neq 0$. For n a positive integer, $H(\delta, n)$ contains the

[13] Recall that the contragredient representation is defined on $\widehat{V}$, the space of smooth linear functionals on V. The action of G is defined implicitly by the inner product on $V \times \widehat{V}$ using the natural left action $\pi(g)^{-1}v$ of G on V.

representation $D_n \oplus D_{-n}$ with quotient F_{n-1}. On the other hand, $H(\delta, -n)$ contains F_{n-1} with quotient $D_n \oplus D_{-n}$.

If $s = 0$, then $I(1,0)$ is unitary and decomposes as the sum of two irreducible representations $D_{+,0}$ and $D_{-,0}$. Finally, we consider $I(0,1)$. The $(\mathfrak{g}, K)$-module quotient of $I(0,1)_K$ by $D_1 \oplus D_{-1}$ is F_0, the trivial representation.

With this discussion in mind, we now present the Langlands classification of irreducible admissible $(\mathfrak{g}, K)$-modules:

Theorem 11.2.3. *The irreducible, admissible $(\mathfrak{g}, K)$-modules are (up to equivalence) the underlying K-finite vectors in the following list of representations:*

a) the discrete series D_n for $n \in \mathbb{Z} - \{0\}$

b) the unitary principal series $H(\delta, s)$ with $s \in i\mathbb{R}$ and $(\delta, s) \neq (0,0)$

c) the irreducible constituents $D_{+,0}$ and $D_{-,0}$ of $H(1,0)$, the limits of discrete series,

d) the representations $I(0,s)$ for $s \in (0,1)$, called complementary series,

e) the trivial representation F_0.

f) the finite dimensional representations F_n, $n \geq 1$

g) the irreducible principal series $I(\delta, s)$ with $s \notin \mathbb{Z}$, or $s \in \mathbb{Z}$ with $s \equiv \delta \pmod 2$.

The first five in the list classify the irreducible unitary representations of G. These will be the ones we're most interested in, since any representation realized as a G-invariant subspace of $L^2(\Gamma\backslash G)$ will inherit a G-invariant inner product from the L^2 norm.

Theorem 11.2.4. *The space ${}^\circ L^2(\Gamma\backslash G)$ decomposes into a Hilbert direct sum of closed irreducible G-invariant subspaces with finite multiplicity.*

Proof. The right regular representation on ${}^\circ L^2(\Gamma\backslash G)$ has endomorphism $\pi(\alpha)$ for $\alpha \in C_c^\infty(G)$ which is expressible as convolution $*\check{\alpha}$, which is a compact operator. Then there is a general principle that if a Dirac sequence $\{\alpha_n\}$ exists such that $\pi(\alpha_n)$ is compact, there is such a decomposition with finite multiplicities. See Lemma 16.1 of Borel's book. □

These G-invariant subspaces correspond to discrete series representations in the following way. There is a realization of discrete series D_n in $L^1(G)$ if $|n| \geq 3$. To basis vectors in this representation, associate a Poincare series P_n. This will be an intertwining operator from the G translates of D_n to ${}^\circ L^2(\Gamma\backslash G)$.

The other types of irreducible unitary representations also appear in the decomposition of $L^2(\Gamma\backslash G)$. Briefly, the limits of discrete series correspond to holomorphic modular forms. The complementary series $I(0,z)$ have $(\mathfrak{g},K)$-modules isomorphic to the module spanned by derivatives of Eisenstein series on $\Gamma\backslash G$ having a pole at $z \in (0,1)$. Finally, the continuous spectrum contained the remaining incomplete theta functions (expressed in terms of a direct integral against Eisenstein series) may be reinterpreted as a direct integral of unitary principal series.