The Arthur–Selberg trace formula and some applications to arithmetic statistics

Author: Kenz Kallal (kenzkallal@college.harvard.edu)
Advisor: Prof. Mark Kisin

22 March 2021

Abstract

This undergraduate thesis is broadly about the theme of applying the Arthur–Selberg trace formula in various contexts to gain insight into the problems of arithmetic statistics. To this end, we develop, in essentially full detail, the technical prerequisites in the spectral theory of automorphic forms for $GL(2)$, and we prove two specializations of the trace formula: the version at the infinite place using Selberg's language of point-pair invariants, and the adelic version specifically for traces of classical Hecke operators. After fully setting up the technology of the trace formula, we proceed with the applications to arithmetic statistics. We provide an exposition of one of the results of Sarnak's thesis, in which he applied the trace formula (in the guise of a prime geodesic theorem for hyperbolic surfaces of finite area) to prove an asymptotic averaging law for class numbers of real quadratic fields which made an important stride towards removing the dependence on the regulator of the classical Gauss–Siegel asymptotic formula. Then, we move on to imaginary quadratic fields, where we present a new approach to proving the existence of fields whose class group has trivial ℓ-torsion for a small list of primes ℓ. Our approach (the original content of this thesis) uses congruences between modular eigenforms of different weight (those induced by the holomorphic Eisenstein series $E_{\ell-1}$ and which were used by Deligne–Serre to extend the construction of Galois representations associated to modular forms to the case of weight $k = 1$) in conjunction with the trace formula.

Acknowledgements

I am deeply indebted to my thesis advisor Mark Kisin, for his suggestion to study the trace formula, his countless other invaluable suggestions, and for his enthusiastic and unwavering support over the past two years. I would also like to thank my parents Hédi Kallal and Malika Zeghal, who have selflessly supported me in all aspects of my life for the past two decades and change.

Thanks are also due to all of the professors who taught me during my undergraduate education. It is a pleasure to thank Ahmed Abbes for his advice about learning algebraic geometry; Fabian Gundlach and Melanie Wood for teaching me about arithmetic statistics in their respective topics courses; Barry Mazur, for agreeing to supervise me and Matthew Hase-Liu in our reading course; Peter Sarnak, for generously sending me his thesis and for his detailed response to my questions about the proof of the prime geodesic theorem in his thesis; and Wei Zhang, for teaching me about the trace formula in his course on automorphic forms and representations.

Finally, I would like to thank all of the undergraduate and graduate students who inhabit the mathematics department common room (at Harvard and elsewhere) and make it an excellent environment for learning and doing mathematics. I have particular appreciation for the contributions to my education resulting from many helpful conversations and collaborations with Yaolin Kevin Chang, James Hotchkiss, Brian Lawrence, Vinh-Kha Le, Hao Billy Lee, Alec Leng, Matthew Hase-Liu, Tomoka Kan, Dongryul Kim, Hudson Kirkpatrick, Siyan Daniel Li-Huerta, Samuel Marks, Amal Mattoo, Mikayel Mkrtchyan, Drew Moore, Rohil Prasad, Tarik Rashada, Shyam Narayanan, Vijay Srinivasan, Alec Sun, Eric Wang, David Xiang, Zijian Yao, and Fan Zhou.

Contents

Chapter 1

Introduction

"It is *good* to be confused!"

Glenn Stevens

1.1 | Historical background

Ever since the time of the ancient Greeks, *Diophantine equations* (polynomial equations for which the solutions we seek are integers) have been of fundamental interest to number theorists. Elementary questions about Diophantine equations, though stated in simple language, are responsible for much of the vast modern corpus of algebraic number theory. Now that we have made this claim, we have no choice but to back it up with some of the most famous examples of how the modern abstractions of algebraic number theory have been motivated by questions stated hundreds of years ago in the elementary language of Diophantine equations.

Question 1.1.1 (Fermat's challenge to the English mathematicians)**.** What are the solutions $x, y \in \mathbf{Z}$ to the Diophantine equation

$$x^2 = y^3 - 2?$$

Question 1.1.2 (Fermat's last theorem)**.** The pythagorean triples (solutions to the Diophantine equation $x^2 + y^2 = z^2$) can be explicitly generated by parametrizing rational points on the unit circle, or by Hilbert's Theorem 90. Are there any nontrivial solutions $x, y, z \in \mathbf{Z}$ to

$$x^n + y^n = z^n$$

if $n \geq 3$?

Question 1.1.3 (Gauss' quadratic reciprocity law)**.** Given a prime p, which primes q can divide the values of $x^2 - p$? Note that this is really asking about when the diophantine equation

$$x^2 - p = qy$$

has solutions $x, y \in \mathbf{Z}$.

Question 1.1.4 (General reciprocity law)**.** Given an arbitrary polynomial $f(x) \in \mathbf{Z}[x]$, which primes q can possibly divide the output? More generally, can we predict how the mod-q reduction of f in $\mathbf{F}_q[x]$ factors?

Question 1.1.5. Given a fixed integer n, which primes p are of the form $x^2 + ny^2$? More generally, given a quadratic polynomial

$$f(x, y) = ax^2 + bxy + cy^2 \in \mathbf{Z}[x, y]$$

(or eventually an arbitrary homogeneous polynomial in an arbitrary number of variables), which primes p are in the image of f?

All five questions 1.1.1-1.1.5 are natural questions about numbers that only use the language of prime numbers and polynomial expressions. They are also very old: Fermat studied Question 1.1.1 and Question 1.1.2 in the middle of the 17th century; Question 1.1.3, Question 1.1.4 and Question 1.1.5 were studied starting at least in the 18th century by Euler, Legendre, Lagrange, and Gauss (see [Cox2013] for a reasonably elementary modern treatment, and [Gau1966] for Gauss' famous work on these questions). But at least in the current state of knowledge, none of them can be solved in generality without the modern tools of algebraic number theory.

For Question 1.1.1, the key insight is to move the 2 over to the left hand side and factor it in the ring

$$\mathbf{Z}[\sqrt{-2}] = \{a + b\sqrt{-2} : a, b \in \mathbf{Z}\},$$

at which point one has

$$y^3 = (x - \sqrt{-2})(x + \sqrt{-2}).$$

One proves that every element in $\mathbf{Z}[\sqrt{-2}]$ factorizes uniquely (up to units) into irreducible elements, from which it follows that $x - \sqrt{-2}$ and $x + \sqrt{-2}$ both must be perfect cubes in $\mathbf{Z}[\sqrt{-2}]$. This turns out to be enough information to deduce directly that the only solutions to the Diophantine equation $x^2 = y^3 - 2$ are

$$(x, y) = (5, \pm 3).$$

But here's the rub: if we want to repeat this kind of argument to some slightly more general equations, such as

$$x^2 = y^3 - n,$$

then we can repeat it verbatim only if $\mathbf{Z}[\sqrt{-n}]$ has the unique factorization property. For positive n, this almost never happens. The failure of the unique factorization property for the ring of integers $\mathcal{O}_K$ of an algebraic number field K is quantified by the *ideal class group* Cl_K, which is the quotient of the group of fractional ideals of $\mathcal{O}_K$ by the subgroup of principal ideals (so $\mathrm{Cl}_K = 1$ is equivalent to $\mathcal{O}_K$ having the unique factorization property, by the general theory of Dedekind domains [Neu1999]). Gauss [Gau1966] conjectured that the group $\mathrm{Cl}_{\mathbf{Q}(\sqrt{-n})}$ is trivial only for some finite list of $n > 0$, namely

$$n = 1, 2, 3, 7, 11, 19, 43, 67, 163.$$

Gauss's class number 1 conjecture was finally proved by Heegner–Stark [Hee1952, Sta1969] (and inde-

pendently by Baker [Bak1968]) in the 1950s and 1960s, using the theory of modular forms and elliptic curves with complex multiplication. But algebraic number fields and knowledge of their class groups is not just useful for these concrete questions if the class group is trivial. In particular, one of the most well-known steps towards answering Question 1.1.2 (Fermat's last theorem) in the negative is due to Kummer [Kum1850], and involves a similar strategy as the answer to Question 1.1.1 explained above. In particular, the Fermat equation

$$x^n + y^n = z^n$$

can be factored over $\mathbf{Z}[\zeta_n] = \mathcal{O}_{\mathbf{Q}(\zeta_n)}$, where ζ_n is a primitive n-th root of 1. Using the factorization $T^n - 1 = \prod_{i=0}^{n-1}(T - \zeta_n^i)$, the Fermat equation becomes

$$z^n = \prod_{i=0}^{n-1}(x + \zeta_n^i y).$$

If $\mathrm{Cl}_{\mathbf{Q}(\zeta_n)} = 1$, then one deduces that the quantities $x + \zeta_n^i y$ are n-th powers in $\mathbf{Z}[\zeta_n]$, and it is relatively straightforward (though not trivial) to deduce Fermat's last theorem from this information. After discovering this argument, Lamé famously made the mistake of claiming that he had produced a proof of Fermat's last theorem[1]. Still, Kummer discovered that the argument could still be made to work when n is prime and

$$n \nmid \#\mathrm{Cl}_{\mathbf{Q}(\zeta_n)}.$$

Such primes n are called *regular primes*.

Class groups of number fields are also of key importance for Question 1.1.5. It is a standard fact from algebraic number theory that for a quadratic field $K = \mathbf{Q}(\sqrt{n})$ of discriminant D, there is a bijection between Cl_K and the set of $GL_2(\mathbf{Z})$-equivalence classes of primitive binary quadratic forms[2] of discriminant D (the article of Wood [Woo2011] which also generalizes this much further and the book of Cohen [Coh1993] on computational algebraic number theory are two good references). So it is no surprise that the answer to Question 1.1.5 depends on knowledge of the class number of some quadratic orders[3]. Surprisingly, Question 1.1.5 can only be answered in generality via *class field theory*, a part of algebraic number theory developed over the course of the first half of the 20th century by Kronecker, Hilbert, Takagi, and Artin, among others [Hil1896, Hil1902, Tak2014, Art1929, AT2009, CF1967]. As a basic example (this argument is most of the content of [Cox2013]; I learned it from an exercise in B. Conrad's course [Con2009, Homework 9]), if $\mathbf{Z}[\sqrt{-n}] = \mathcal{O}_{\mathbf{Q}(\sqrt{n})}$, then a prime p is of the form $x^2 - ny^2$ if and only if p splits into two principal primes $(x - \sqrt{-n}y)(x + \sqrt{-n}y)$ in $\mathbf{Q}(\sqrt{-n})$. By class field theory, this is equivalent to p splitting completely in the the Hilbert class field of $\mathbf{Q}(\sqrt{-n})$, which is a degree-$\#\mathrm{Cl}_{\mathbf{Q}(\sqrt{-n})}$ extension of $\mathbf{Q}(\sqrt{-n})$. Therefore, the answer to Question 1.1.5 depends on how some polynomial over $\mathbf{Z}[\sqrt{-n}]$ depending only on n splits modulo p, which reduces Question 1.1.5 to

[1] At least this story is part of number theory folklore, recorded for instance in [Edw1977].

[2] A *primitive binary quadratic form* is just one of the quadratic polynomials of interest in Question 1.1.5, namely a homogeneous quadratic polynomial in 2 variables such that all 3 coefficients are coprime.

[3] Here we have been sweeping a detail under the rug: the ring $\mathbf{Z}[\sqrt{n}]$ is not necessarily equal to the ring of integers $\mathcal{O}_{\mathbf{Q}(\sqrt{n})}$; instead it is a non-maximal order (see [Neu1999, Ch. 1]). But this difficulty is not the central one in the theory, and we lose no important ideas by ignoring it.

Question 1.1.4 with the added information that the degree of this polynomial equals $\#\mathrm{Cl}_{\mathbf{Q}(\sqrt{-n})}$.

Finally, Question 1.1.3 and the generalized version Question 1.1.4 concern the general notion of *reciprocity*. Gauss [Gau1966] famously resolved Question 1.1.3 by elementary methods in many different ways over the course of his life. The case of Question 1.1.4 where f has abelian Galois group is part of the content of class field theory. From one lens, the reciprocity law of class field theory is about a correspondence between abelian *Hecke L-functions* (analytic objects that come from 1-dimensional complex representations of ray class groups, which are just a natural generalization of ideal class groups) and abelian *Artin L-functions* (analytic objects that come from 1-dimensional complex representations of $\mathrm{Gal}(\overline{\mathbf{Q}}/\mathbf{Q})$). The idea that this should generalize to higher-dimensional representations and with coefficients in other fields is a fundamental one fueling the Langlands program, and particularly modern work on the Langlands reciprocity conjecture. In fact, Wiles and Taylor–Wiles' resolution of Question 1.1.2 in the 1990s [Wil1995, TW1995] and the continuing work on modularity [Kis2009a, Kis2009b, KW2009] is a central part of the Langlands program.

A common view is that the theory of automorphic forms and representations (key objects on one side of the Langlands program; the Hecke L-function of class field theory are attached to automorphic representations of $GL(1)$) began with Tate's thesis [Tat1950], which set up the language and proved the local and global functional equation in the case $GL(1)$. This thesis essentially assumes familiarity with that work, and explains the theory for $GL(2)$ in all the necessary detail. The main theme of this thesis is the development of the Arthur–Selberg trace formula in various forms for $GL(2)$, and the application of this to concrete questions about the distribution of ideal class groups of number fields. Though it is not the main focus of this thesis, we remark here that the trace formula is a central tool in the Langlands program in and of itself: for instance, it is used in the proof of the Jacquet–Langlands correspondence [JL1970, GJ1979], and in the proof of the cyclic base change lifting for $GL(2)$, which itself was a key ingredient in the proof of Fermat's last theorem (in the guise of the Langlands–Tunnell theorem [Lan1980, Tun1981, CSS1997]). It is also the key ingredient in the Langlands–Kottwitz method [Lan1973, Kot1984, Sch2011, Sch2013].

1.2 | Brief overview

The purpose of Section 1.1 was to give some motivation for why class numbers of algebraic number fields are interesting for the purposes of attacking concrete questions in number theory, and for why the modern methods of algebraic number theory and particularly the Langlands program should be expected to shed light on them. This section is meant to give more detailed information about the actual contents of this thesis.

Given that Gauss' class number-1 conjecture for imaginary quadratic fields has been resolved, one might hope to eventually be able to have simple explicit formulas for class numbers of quadratic fields. Having such convenient theoretical information about individual class numbers seems to be out of reach thus far (though this does not mean there are not efficient algorithms for computing class numbers of individual number fields [Coh1993]). In fact, it is not known whether infinitely many *real* quadratic fields $\mathbf{Q}(\sqrt{n})$ for $n > 0$ have class number 1, though it is predicted that about 76% of them (ordered by discriminant) do [CL1984].

On the other hand, Gauss [Gau1966] saw that the class numbers have considerable regularity to them *when looked at on average*. In particular, he proved

Theorem 1.2.1 (Gauss, 1801)**.** *Let $\mathscr{D}_{\mathrm{Im}}$ be the set of discriminants of imaginary quadratic fields. Then*

$$\sum_{\substack{D\in\mathscr{D}_{\mathrm{Im}}\\ -D<T}} h_D \sim \frac{\pi}{36}\prod_p (1-p^{-2}-p^{-3}+p^4)T^{3/2}$$

as $T\to\infty$, where the product is over the positive rational primes and h_D denotes the class number of the imaginary quadratic field of discriminant D.

However, Gauss found that no such asymptotic law seemed to hold for averages of class numbers of real quadratic fields. This is due to the fact that the unit groups of real quadratic fields are infinite, and one must account for this by weighting everything by the regulator. Gauss conjectured the following statement, which was eventually proved in [Sie1944].

Theorem 1.2.2 (Siegel, 1944)**.** *Let $\mathscr{D}_{\mathrm{Re}}$ be the set of discriminants of real quadratic fields. Then*

$$\sum_{\substack{D\in\mathscr{D}_{\mathrm{Re}}\\ -D<T}} h_D R_D \sim \frac{\pi^2}{36}\prod_p (1-p^{-2}-p^{-3}+p^4)T^{3/2}$$

as $T\to\infty$, where the product is over the positive rational primes and h_D denotes the class number of the real quadratic field of discriminant D.

Nowadays, the problem of the asymptotic behavior of class numbers, and more generally the asymptotic distribution of class groups of global fields in natural families, is an important part of a field called *arithmetic statistics*. The most famous conjecture of arithmetic statistics in this direction is from [CL1984]:

Conjecture 1.2.3 (Cohen–Lenstra heuristics, 1983)**.** Let ℓ be an odd prime. For any finite abelian ℓ-group G,

$$\Pr[\mathrm{Cl}_{k_D}[\ell^\infty]\cong G | D\in\mathscr{D}_{\mathrm{Im}}] = \frac{1}{\#\mathrm{Aut}G}\prod_{i\geq 1}(1-\ell^{-i})$$

where k_D denotes the imaginary quadratic field of discriminant D. The probability is meant in the sense of natural density, ordered by discriminant.

Conjecture 4.2.1 seems to be very far out of reach, and so are its various generalizations to other families of number fields [CM1987, CM1990]. The question in arithmetic statistics that this thesis will focus on is a consequence of Conjecture 4.2.1 which is still very open:

Conjecture 1.2.4. Let ℓ be an odd prime. Then

$$\Pr[\mathrm{Cl}_{k_D}[\ell]=1 | D\in\mathscr{D}_{\mathrm{Im}}] = \Pr[h_D\not\equiv 0 \pmod{\ell} | D\in\mathscr{D}_{\mathrm{Im}}] = \prod_{i\geq 1}(1-\ell^{-i}) > 0.$$

To give an idea of how open this conjecture is: when $\ell \geq 5$, it is unknown whether a positive proportion[4] of imaginary quadratic fields K have $\mathrm{Cl}_K[\ell] = 1$.

We will explain some special instances of the theme of applying information about automorphic forms together with a powerful tool from the Langlands program called the *Arthur–Selberg trace formula* [Sel1956, DL1971, Lan2001, Art2005] (the research program around various forms of the trace formula and related issues of invariance and stabilization is ongoing and still being led by Arthur [Art1981, Art1983, Lan1983, Art2002, Art2001, Art2003]) to make progress towards both of the questions in arithmetic statistics we have introduced so far: Conjecture 1.2.4, as well as the question of how to generalize Theorem 1.2.1 to the real quadratic case without having to weight the class number by the regulator.

One explicit way to write down the Arthur–Selberg trace formula for $GL(2)$ is called the *Eichler–Selberg trace formula*. The statement (in the case of full level) is as follows.

Theorem 1.2.5. *Let $k > 2$ be an even integer, and S_k the $\mathbf{C}$-vector space of cusp forms of weight k and level 1, equipped with the Hecke operators T_m for all $m \geq 1$. Then*

$$\mathrm{Tr} T_m|_{S_k} = -\frac{1}{2}\sum_{|t|\leq 2\sqrt{m}} P_k(t,m)H(t^2-4m) - \frac{1}{2}\sum_{d_1 d_2 = m} \min(d_1,d_2)^{k-1},$$

where $P_k(t,m)$ is defined to be

$$P_k(t,m) = \frac{\rho^{k-1} - \overline{\rho}^{k-1}}{\rho - \overline{\rho}},$$

where ρ is the quadratic algebraic number with norm m and trace t, and H denotes the Hurwitz class number[5]. Note that $P_k(t,m)$ is a polynomial in t and m, and is equal to the coefficient of x^{k-2} in the formal power series $(1 - tx + mx^2)^{-1} \in \mathbf{Z}[m,t][\![x]\!]$.

We recover some special cases (previously known but by different methods [Har1974, Wil2015]) of Conjecture 1.2.4 by using Theorem 1.2.5 in conjunction with congruences between modular eigenforms of different weight coming from the congruence of holomorphic Eisenstein series $E_{p-1} \equiv 1 \mod p$ (this type of congruence was notably used by Deligne–Serre [DS1974] to show the existence of Galois representations associated to eigenforms of weight $k = 1$). This novel approach is the original content of this thesis, and the present result is stated and proved in Theorem 4.2.27.

We also explain how to apply the Selberg trace formula at the infinite place to obtain a prime geodesic theorem for finite-area hyperbolic surfaces and deduce one of the main results of Sarnak's thesis [Sar1980, Sar1982], namely

Theorem 1.2.6 (Sarnak, 1980)**.** *For $D \in \mathscr{D}_{\mathrm{Re}}$, let h_D^+ denote the narrow class number of the real quadratic field of discriminant D, and R_D the narrow regulator. Then*

$$\sum_{\substack{D\in\mathscr{D}_{\mathrm{Re}} \\ e^{R_D}\leq T}} h_D^+ \sim Li(T^2)$$

[4] From now on we omit the fact that densities are computed by ordering by discriminant, since this is the only way it will be done in this thesis.

[5] The Hurwitz class number is defined just like the usual class number, except the quadratic forms are counted with weights inverse to the size of the stablizer in $SL_2(\mathbf{Z})$

as $T \to \infty$.

This thesis is organized as follows. In Chapter 2, we follow some of the standard references [Bum1997, GGPS1969, Lan1985, Lan1976, Iwa2002] in establishing in full detail the basic spectral theory of automorphic forms on $GL(2)$, albeit only in the context of $GL(2, \mathbf{R})^+$. In the long Chapter 3, we provide detailed proofs of both versions of the Arthur–Selberg trace formula we will use: the purely analytic version for $GL_2(\mathbf{R})$ (essentially following [Hej1976, Iwa2002]) which we will use later to prove the prime geodesic theorem and Weyl's law for hyperbolic surfaces; and the version for $GL_{2/\mathbf{Q}}$ used to deduce Theorem 1.2.5 (essentially following [GJ1979, KL2006]). Finally, in Chapter 4, we present the applications of these tools to the problems of arithmetic statistics introduced in this section: Sarnak's result Theorem 1.2.6 [Sar1980, Sar1982] for asymptotic averages of class numbers of real quadratic fields, and our new approach to Conjecture 1.2.4 (about torsion in class groups of imaginary quadratic fields) using the trace formula in conjunction with additional congruence data (as briefly described above).

Parts of Chapter 2 and Chapter 3 have previously appeared in notes I wrote for the University of Chicago graduate students' seminar on automorphic forms and representations [Kal2020].

Chapter 2

Review of basic spectral theory of automorphic forms

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

This chapter follows the references [Bum1997] and [Iwa2002] closely, with occasional input from [Lan1976], [GGPS1969], and [Lan1985].

2.1 | Automorphic forms and representation theory

Let $\Gamma \subset SL_2(\mathbf{Z})$ be a congruence subgroup and $\mathbf{H} = \{z \in \mathbf{C} : \Im(z) > 0\}$ the complex upper half-plane. Then recall

Definition 2.1.1. Let $k \geq 0$ be an integer and $\chi : \Gamma \to \mathbf{C}^\times$ a unitary character. A *Maass form* of weight k and character χ is a smooth function

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

satisfying

1. A polynomial growth condition at the cusps[1];
2. the transformation law
$$f(\gamma z) = \chi(\gamma) \left(\frac{cz+d}{|cz+d|} \right)^k f(z)$$
for all
$$\gamma = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \gamma;$$

[1] What this means is that for any $\sigma \in \Gamma\backslash SL_2(\mathbf{Z})$ (these take $i\infty$ to the other cusps of $\Gamma\backslash\mathbf{H}$), $f(\sigma(x+iy)) \ll y^N$ for some N.

3. and the differential equation

$$\Delta_k f = \lambda f$$

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

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

For this section and the next two, we assume $\Gamma = SL_2(\mathbf{Z})$. This will make the theory easier to write down (since we only have one cusp and therefore only one family of Eisenstein series), and it is the only thing we will need in the archimedean theory in order to establish Sarnak's theorem on class numbers in Chapter 4. However, we note here that (with the exception of specific information about small Laplace eigenvalues which is only important for the error terms in Chapter 4) all the parts of the theory presented here generalize easily to general Γ.

The reason for the strange-looking transformation law in Definition 2.1.1 is that the definition comes from looking at functions on $\mathbf{H}$ from the perspective of representation theory. This perspective allows one to go from the classical notion of modular forms to the modern language of automorphic forms and representations. In particular, $\mathbf{H}$ is equipped with a transitive smooth action of $GL_2(\mathbf{R})^+$, and

$$\mathrm{Stab}_{GL_2(\mathbf{R})^+}(i) = \left\{\begin{pmatrix} a & b \\ -b & a \end{pmatrix} : a^2 + b^2 > 0\right\} = \mathbf{R}^\times_{>0} SO_2(\mathbf{R}).$$

We write $SO_2(\mathbf{R}) = K^\circ$, because it is the connected component of the maximal compact subgroup $K = O_2(\mathbf{R}) \subset GL_2(\mathbf{R})$. So we may rewrite $\mathbf{H}$ as the homogeneous space

$$\mathbf{H} \cong GL_2(\mathbf{R})^+/Z^\circ K^\circ \cong SL_2(\mathbf{R})/K^\circ.$$

The point of this observation is that the Maass forms of weight[2] 0 which also enjoy the property of being square-integrable with respect to the induced hyperbolic metric on $\Gamma\backslash\mathbf{H}$ can be considered as elements of the complex Hilbert space

$$L^2(\Gamma\backslash GL_2(\mathbf{R})^+/Z^\circ K^\circ, \chi) = L^2(\Gamma\backslash SL_2(\mathbf{R})^+/K^\circ, \chi)$$

consisting of the measurable functions $f : GL_2(\mathbf{R})^+ \to \mathbf{C}$ with the property that[3]

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

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

$$\int_{\Gamma\backslash SL_2(\mathbf{R})} |f(x)|^2\, dx < \infty \tag{2.1}$$

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

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

where dx is the Haar measure on $SL_2(\mathbf{R})$. One checks (for instance via the Iwasawa decomposition for $SL_2(\mathbf{R})$) that the measure on $\Gamma\backslash SL_2(\mathbf{Z})/K^\circ = \Gamma\backslash\mathbf{H}$ coming from the Haar measure on $SL_2(\mathbf{R})$ coincides with the measure induced by the Riemannian metric on $\mathbf{H}$, so since K° is compact and of measure 1 when normalized correctly, Equation (2.1) is equivalent to f being of finite L^2-norm when considered as a function on $\mathbf{H}$. In summary, we have

Lemma 2.1.2. *There is an isomorphism of complex Hilbert spaces*

$$\varphi : L^2(\Gamma\backslash\mathbf{H}, \chi) \to L^2(\Gamma\backslash SL_2(\mathbf{R})/K^\circ, \chi)$$

given by sending f to the complex-valued function

$$\begin{pmatrix} y^{1/2} & xy^{-1/2} \\ 0 & y^{-1/2} \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \mapsto f(x+iy).$$

Proof. This is just the explicit realization of the isomorphism $\Gamma\backslash\mathbf{H} \cong \Gamma\backslash SL_2(\mathbf{R})/K^\circ$, using the Iwasawa decomposition and the fact that this isomorphism is given by

$$g \mapsto g(i),$$

and hence $x+iy$ is the image of the coset of $\begin{pmatrix} y^{1/2} & xy^{-1/2} \\ 0 & y^{-1/2} \end{pmatrix}$. As remarked above, one checks that the left Haar measure on the upper-triangular Borel subgroup is given by $\frac{1}{y^2}\, dxdy$ when using the coordinates above, which shows that φ respects the inner product of the two L^2 spaces. □

This L^2 space is really only useful insofar as we can use representation theory to study it. We would want to define a left action of $GL_2(\mathbf{R})^+$, via right regular representation $(g\cdot f)(x) = f(xg)$. The problem with this is that K° is not in the center of $GL_2(\mathbf{R})^+$, so this action would not take $L^2(\Gamma\backslash SL_2(\mathbf{R})/K^\circ, \chi)$ to itself.

So we remove the requirement of K°-invariance, and consider the larger Hilbert space

$$\mathfrak{H} = L^2(\Gamma\backslash SL_2(\mathbf{R}), \chi),$$

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

$$L^2(\Gamma\backslash SL_2(\mathbf{R})/K^\circ, \chi)$$

where the square-integrable Maass forms live then sits inside of $\mathfrak{H}$ as the set of vectors on which K° operates trivially. Also,the space of smooth vectors for the Lie group action of $GL_2(\mathbf{R})^+$ on $\mathfrak{H}$ is

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

defined in the obvious way[4]. It is a general fact that the smooth vectors in Hilbert space representations

[4]Though the "obvious" way still requires adding the requirement of square-integrability.

of Lie groups are dense:

Lemma 2.1.3. *Let $\pi : G \to \mathrm{End}(\mathfrak{H})$ be a representation of a Lie group G into a Hilbert space $\mathfrak{H}$. The space of smooth vectors for this representation, $\mathfrak{H}^\infty$, is dense in $\mathfrak{H}$.*

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

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

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

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

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

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

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

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

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

The square-integrable Maass forms of weight 0 live in the $GL_2(\mathbf{R})^+$-smooth vectors of

$$L^2(\Gamma\backslash SL_2(\mathbf{R})/K^\circ, \chi) \subset L^2(\Gamma\backslash SL_2(\mathbf{R}), \chi) = \mathfrak{H},$$

that is they are smooth vectors in the K°-isotypic subspace of $\mathfrak{H}$ corresponding to the trivial representation of K°. But $K^\circ = SO_2(\mathbf{R})$ has some other irreducible representations, which are all 1-dimensional (as $SO_2(\mathbf{R})$ is abelian) and given by

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \mapsto e^{ik\theta}$$

for $k \in \mathbf{Z}$. By the Peter–Weyl theorem [Bum2013, Theorem 4.3], we have a decomposition

$$\mathfrak{H} = \bigoplus_{k \in \mathbf{Z}} \mathfrak{H}_k,$$

of representations of K°, where the direct sum is the Hilbert space direct sum[5] and $\mathfrak{H}_k$ is the K°-isotypic subspace corresponding to the irreducible representation $e^{ik\theta}$.

This is a general technique in representation theory: when you have a Hilbert space representation of a group G, restrict it to a maximal compact subgroup K and use the representation theory of compact groups to your advantage. In our case, there are two reasons why it is more convenient to think about the connected component $G = GL_2(\mathbf{R})^+$ rather than $GL_2(\mathbf{R})$:

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

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

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

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

Lemma 2.1.4. *The map*

$$\sigma_k : L^2(\Gamma\backslash\mathbf{H}, \chi, k) \to \mathfrak{H}_k$$

given by

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

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

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

is an isomorphism of Hilbert spaces.

Proof. We checked already in Lemma 2.1.2 that this map respects the inner product. To recover f from $\sigma_k f$, we just take $f(x+iy) = (\sigma_k f)\begin{pmatrix} y & x \\ 0 & 1 \end{pmatrix}$, which works because the K-component of this matrix

[5] So the claim is really that the algebraic direct sum on the right hand side is dense in $\mathfrak{H}$

in the Iwasawa decomposition is zero. It just remains to check that $\sigma_k f$ being a $e^{ik\theta}$-simultaneous eigenvector for the action of K is equivalent to f satisfying the symmetry property for Maass forms of weight k. This is because if $\sigma_k f = F \in L^2(\Gamma\backslash PGL_2(\mathbf{R})^+, \chi, k)$, then

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

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

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

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

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

so

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

and thus

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

as desired. □

So we have provided a natural explanation of the symmetry condition satisfied by the Maass forms of weight k in terms of the representation theory of $SL_2(\mathbf{R})$. Where do the modular forms fit into this picture? Actually the answer is very simple.

Lemma 2.1.5. *Suppose that $f : \mathbf{H} \to \mathbf{C}$ is a modular form of weight k and character χ for Γ. Then*

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

Proof. This is a straightforward observation about the relationship between the symmetry conditions

satisfied by modular forms and Maass forms. In particular,

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

as required. □

So from the perspective of representation theory, the theory of modular forms is subsumed by the theory of Maass forms. Note that we haven't yet accounted for the entirety of Definition 2.1.1: we are still missing

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

These conditions are useful because they guarantee the existence of a Fourier expansion (see [Bum1997, §1.9]). We will see later that despite the growth condition, the Maass forms still provide a full decomposition of $\mathfrak{H}$ into the discrete spectrum (coming from Maass cusp forms) and the continuous spectrum (coming from non-holomorphic Eisenstein series[6]). The point is that we are looking to decompose the spectrum of Δ, and the Maass forms are the basic building blocks of that.

Let $\mathfrak{g} = \mathcal{M}_{2\times 2}(\mathbf{R})$ be the Lie algebra of $GL(2,\mathbf{R})^+$. Since $C^\infty(\Gamma\backslash SL_2(\mathbf{R}), \chi)$ are the $GL(2,\mathbf{R})^+$-smooth vectors in $L^2(\Gamma\backslash SL_2(\mathbf{R}), \chi)$, they admit an action of the universal enveloping algebra $U(\mathfrak{g})$. The weight-0 Laplacian on the upper half-plane, which we might originally justify as being the Laplace-Beltrami operator for the Poincaré upper half-plane, turns out to transfer over (via the map of Lemma 2.1.4) to this setting as the Casimir element of $U(\mathfrak{g})$. In fact, it is generally true that if you choose a bi-invariant metric on a Lie group G, then the Laplace-Beltrami operator with respect to that metric coincides with the Casimir element corresponding to the induced inner product on the Lie algebra $\mathfrak{g}$.

The center of $U(\mathfrak{g} \otimes \mathbf{C})$ is $\mathfrak{Z} = \mathbf{C}[\Delta, Z_\mathfrak{g}]$, where Δ is the Casimir element with respect to the Killing form, and $Z_\mathfrak{g}$ is the identity matrix and one of the two standard basis elements of the Cartan subalgebra $\mathfrak{h}_\mathbf{C} \subset \mathfrak{g} \otimes \mathbf{C}$, given by the diagonal matrices. Recall that the standard basis of $\mathfrak{h}_\mathbf{C}$ is

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

Note that $Z^\circ = \mathbf{R}^\times_{>0}$ the subgroup of $GL_2(\mathbf{R})^+$ is distinguished in our notation from $Z_\mathfrak{g}$, though the fact that they are both involve the letter Z is suggestive of their relation, namely that $\exp(\mathbf{R} \times Z_\mathfrak{g}) = Z^\circ$.

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

[6] Though the individual Eisenstein series are NOT square-integrable.

those elements are

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

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

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

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

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

Despite the fact that the actual matrix $\hat{H} = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}$ is not as nice, we prefer to use this one because a decomposition of a K°-finite representation of $GL_2(\mathbf{R})$ into K°-isotypic subspaces should correspond to a decomposition into eigenspaces of H. By our previous discussion on the irreducible representations of $K^\circ = SO_2(\mathbf{R})$, such a K°-finite representation V of G decomposes as an algebraic direct sum

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

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

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

If V is the space of K°-finite vectors in $\mathfrak{H} = L^2(\Gamma\backslash SL_2(\mathbf{R}), \chi)$, then in the previous section we saw that Maass forms and modular forms of weight k and character χ for Γ can be thought of as elements of $V(k)$. By virtue of the way we changed variables via the Cayley transform, this decomposition is also an eigenspace decomposition for the action of $\hat{H}$, since

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

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

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

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

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

decomposition

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

where

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

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

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

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

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

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

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

The Maass–Shimura operators also provide the key representation–theoretic distinction between the Maass forms that come from (anti-)holomorphic modular forms (see Lemma 2.1.5) and those that don't.

Lemma 2.1.6. *Let $f \in C^\infty(\Gamma \backslash SL_2(\mathbf{R}), \chi, k)$ be nonzero.*

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

Proof. In the coordinates on $GL_2(\mathbf{R})^+$ coming from the Iwasawa decomposition

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

the Maass–Shimura differential operators may be explicitly given by

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

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

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

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

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

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

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

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

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

2.1.1 | Discrete decomposition of the cuspidal subspace

The ultimate goal here is to decompose the right regular representation $(\pi, \mathfrak{H})$ into irreducible components. We do this following the reference of Bump [Bum1997], which uses the strategy of applying the theory of $(\mathfrak{g}, K^\circ)$-modules originally due to Harish-Chandra [HC1953, HC1954a, HC1954b]. In the decomposition, there is a serious distinction between the cuspidal part of $\mathfrak{H}$ and the rest: the cuspidal part will decompose *discretely* as a direct sum of irreducibles. The rest will compose *continuously* as a direct integral of subgrpresentations generated by Eisenstein series. So we begin with a more serious discussion of cuspidality, and with the decomposition of the cuspidal subspace. Let $N \subset SL_2(\mathbf{R})$ be the upper-triangular nilpotent radical in the Levi decomposition of the upper-triangular parabolic (Borel) subgroup of $SL_2(\mathbf{R})$[7]. The definition of cuspidal is obvious when $\chi|_{\Gamma\cap N} = 1$: in that case, f is periodic under horizontal translations, so we say that f is cuspidal at ∞ if its constant Fourier coefficient vanishes, i.e.

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

[7] Some general language is used here to suggest what the appropriate generalization is to automorphic forms on arbitrary reducive Lie groups, but the point is that N is the group of upper-triangular matrices with 1's on the diagonal. You integrate over $(\Gamma \cap N)\backslash N \cong \mathbf{Z}\backslash\mathbf{R}$ to compute Fourier coefficients.

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

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

This will not be relevant to us since we will restrict to the case $\Gamma = SL_2(\mathbf{Z})$, and therefore to the case of exactly one cusp at ∞. When χ has finite image (true of the most important case, when χ is a nebentypus character) and $\chi|_{\Gamma\cap N} \neq 1$, f is periodic with respect to horizontal translations, but not by everything in $\Gamma \cap N$: one must restrict to the kernel, and the constant Fourier coefficient at ∞ is

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

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

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

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

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

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

So we define cuspidality in the following way for forms with character.

Definition 2.1.7. A function $f \in L^2(\Gamma\backslash G, \chi)$ is *cuspidal* at ∞ if $\chi|_{\Gamma\cap N} \neq 1$ or otherwise if

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

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

Now we proceed with the decomposition of the subrepresentation

$$L^2_{\text{cusp}}(\Gamma\backslash SL_2(\mathbf{R})) \subset \mathfrak{H}$$

consisting of cuspidal elements as a direct sum of irreducible components. This is an example of the general technique of obtaining operators by convolving a representation with smooth test functions.

Definition 2.1.8. For the representation $(\pi, \mathfrak{H})$ and a function $\phi \in C_c^\infty(G^\circ)$, we can obtain an operator $\pi(\phi)$ on $\mathfrak{H}$ given by

$$(\pi(\phi)f)(g) = \int_G \phi(h)\pi(h)f(g)\,dh.$$

The following two lemmas were proved in greater generality by Langlands, but the basic ideas are contained in these proofs for $GL_2(\mathbf{R})^+$. The first is an obligatory estimate. The reader should feel free to skip it, but be aware that it is the only place where the cuspidality is an input. So the difficulties of the continuous spectrum are due to the failure of this estimate to hold when not restricted to the cuspidal part.

Lemma 2.1.9. *Suppose $\phi \in C_c^\infty(GL_2(\mathbf{R})^+)$ and $\Gamma \subset SL_2(\mathbf{Z})$ is a congruence subgroup. Then there exists a constant $C_{\phi,\Gamma,\chi}$ depending only on ϕ, χ and Γ such that*

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

for all $f \in L^2_{\text{cusp}}(\Gamma\backslash SL_2(\mathbf{R}), \chi)$.

Proof. The simplest way to carry out an estimate like this is to construct a crude approximation of a fundamental domain for $\Gamma\backslash SL_2(\mathbf{R})$, from the standard knowledge of how $\Gamma\backslash\mathbf{H}$ works. The *Siegel set* defined via the Iwasawa decomposition

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

contains a fundamental domain for $SL_2(\mathbf{Z})\backslash G^\circ$ if $c, d > 0$ are chosen correctly. Choose them correctly, and fix those values. Depending on the choice of Γ, there is a list of finitely many[8] $\xi_i \in SL_2(\mathbf{Z})$ which take ∞ to each of the cusps, and thus

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

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

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

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

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

[8] since there are finitely many cusps

is in $L^2_{\text{cusp}}(\xi\Gamma\xi^{-1}\backslash SL_2(\mathbf{R}), \chi)$. Applying the bound over $\mathcal{G}_{c,d}$ to this function, we have

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

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

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

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

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

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

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

Using the Iwasawa decomposition, there is a fundamental domain for $(\Gamma \cap N)\backslash SL_2(\mathbf{R})$ given by $\mathcal{N}_\Gamma \times A \times K$, where

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

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

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

$$\begin{aligned}
(\pi(\phi)f)(g) &= \int_G \phi(g^{-1}h)f(h)\,dh \\
&= \int_{(\Gamma\cap N)\backslash SL_2(\mathbf{R})} \sum_{\gamma\in\Gamma\cap N} \int_Z f(\gamma hu)\phi(g^{-1}\gamma hu)\,dy = u\,dh \\
&= \int_{(\Gamma\cap N)\backslash SL_2(\mathbf{R})} f(h) \sum_{\gamma\in\Gamma\cap N} \chi(\gamma) \int_Z \phi(ug^{-1}\gamma h)\,du\,dh \\
&= \int_{\mathcal{N}_\Gamma\times A\times K} f(h) \sum_{n\in\mathbf{Z}} \chi\left(\begin{pmatrix} 1 & n_0 n \\ 0 & 1 \end{pmatrix}\right) \int_{Z^\circ} \phi\left(g^{-1}\begin{pmatrix} 1 & n_0 n \\ 0 & 1 \end{pmatrix} h\right) du\,dh.
\end{aligned}$$

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

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

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

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

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

The main task is therefore to estimate

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

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

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

$$\phi_{Z^\circ}(g) = \int_{Z^\circ} \phi(ug)\, du$$

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

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

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

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

hence

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

Fourier transforms of smooth compactly-supported functions are nice, because they decay very fast, in fact faster than any polynomial[9]. So when $n \neq 0$ the terms in this sum can be controlled, which is what we do next. We need our bound to work over arbitrary y-coordinate for h and $y_g \geq c$, though, so we need to take this apart a little more. Writing

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

we have (sine χ is unitary and ϕ_{Z° is Z°-invariant)

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

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

$$= \left| \int_{-\infty}^{\infty} \chi\left(\begin{pmatrix} 1 & n_0 t \\ 0 & 1 \end{pmatrix} \right) \phi_{Z^\circ} \left(\kappa_g^{-1} \begin{pmatrix} y_g^{-1} y_h & y_g^{-1} n_0 t \\ 0 & 1 \end{pmatrix} \kappa_h \right) e^{-2\pi i n t} \, dt \right|$$
$$= |y_g| \left| \int_{-\infty}^{\infty} \chi\left(\begin{pmatrix} 1 & n_0 y_g t \\ 0 & 1 \end{pmatrix} \right) \phi_{Z^\circ} \left(\kappa_g^{-1} \begin{pmatrix} y_g^{-1} y_h & n_0 t \\ 0 & 1 \end{pmatrix} \kappa_h \right) e^{-2\pi i n y_g t} \, dt \right|$$
$$= |y_g| \left| \int_{-\infty}^{\infty} \phi_{Z^\circ} \left(\kappa_g^{-1} \begin{pmatrix} y_g^{-1} y_h & n_0 t \\ 0 & 1 \end{pmatrix} \kappa_h \right) e^{-2\pi i (n - \lambda_\chi n_0) y_g t} \, dt \right|.$$

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

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

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

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

is compact[10], which means that the set of $(y_g^{-1} y_h, t) \in \mathbf{R}_{>0} \times \mathbf{R}$ for which $F_{g,h}(t) \neq 0$ is contained in a compact set, hence the set of possible $y_g^{-1} y_h$ is contained in a compact set[11] as well. We have shown

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

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

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

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

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

for all h, g (we haven't yet used any restriction to fundamental domains). As a result, choosing $N = 2$ so

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

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

that the sum converges, there is a constant $B_{\phi,\Gamma.\chi}$ such that

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

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

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

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

$$y_h \geq ac$$

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

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

The domain of integration here can be covered by finitely many translates of a fundamental domain for $\Gamma\backslash SL_2(\mathbf{R})$ (this is easily seen using the upper half-plane, and then taking products of everything with K which doesn't change the volume). So there is some positive integer N depending only on Γ such that this contribution is bounded by $c^{-1}B_{\phi,\Gamma,\chi}N\|f\|_{L^1}$. This L^1-norm is actually finite and bounded above by $\|f\|_{L^2} < \infty$, because the fundamental domain has finite volume (so it follows from Cauchy–Bunyakovski–Schwarz inquality).

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

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

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

$$\int_{(\Gamma\cap N)\backslash G/Z} f(h)\widehat{\Phi}_{g,h}(0) = \int_{(\Gamma\cap N)\backslash SL_2(\mathbf{R})} f(h) \int_N \phi_Z(g^{-1}nh)\, dn\, dh$$

$$
\begin{aligned}
&= \int_{(\Gamma\cap N)\backslash SL_2(\mathbf{R})} f(h) \int_N \phi_Z(g^{-1}n^{-1}h)\,dn\,dh \\
&= \int_{\mathcal{N}_\Gamma} \int_{(\Gamma\cap N)\backslash SL_2(\mathbf{R})} f(h) \sum_{\gamma\in\Gamma\cap N} \phi_Z(g^{-1}n^{-1}\gamma^{-1}h)\,dh\,dn \\
&= \int_{\mathcal{N}_\Gamma} (\pi(\phi)f)(ng) \\
&= 0
\end{aligned}
$$

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

Proposition 2.1.10. *Let $\phi \in C_c^\infty(G^\circ)$. Then the convolved operator $\pi(\phi)$ is a compact operator on $L^2_{\mathrm{cusp}}(\Gamma\backslash SL_2(\mathbf{R}),\chi)$.*

Proof. First, we consider the case of compact quotient. In that case, for $f \in L^2(\Gamma\backslash SL_2(\mathbf{R}),\chi)$ and $h \in G^\circ$, we have

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

where

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

and $\mathcal{F}$ is a fundamental domain[12] in G° for $\Gamma\backslash G^\circ/Z^\circ = \Gamma\backslash SL_2(\mathbf{R})$. The fact that $\phi \in C_c^\infty(G^\circ)$ means that $K(g,h)$ is smooth in g and h, and $\Gamma\backslash SL_2(\mathbf{R})$ being compact therefore implies that

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

So $\pi(\phi)$ is a Hilbert–Schmidt operator on $L^2(\Gamma\backslash SL_2(\mathbf{R}),\chi) \cong L^2(\mathcal{F})$ [where the isomorphism is as Hilbert spaces], and is therefore compact.

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

In the case of noncompact quotients, to prove the statement, we need to check that $\pi(\phi)$ restricts to a well-defined operator on $L^2_{\text{cusp}}(\Gamma\backslash SL_2(\mathbf{R}), \chi)$. In other words, if

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

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

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

This is not hard to check:

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

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

The argument we have written down so far is not a priori a valid argument for why $\pi(\phi)|_{L^2_{\text{cusp}}}$ is compact (indeed, if it worked without modification, then there would be no need to restrict to the cuspidal part). The reason is that when $\Gamma\backslash SL_2(\mathbf{R})$ is not compact, $K(\cdot,\cdot)$ is not guaranteed to be in $L^2(\mathcal{F}\times\mathcal{F})$. The additional technical observation that must be made is that there is a constant C_ϕ depending only on ϕ such that

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

for all $f \in L^2_{\text{cusp}}(\Gamma\backslash SL_2(\mathbf{R}), \chi)$. This is what we did in Lemma 2.1.9, and it is where the assumption of cuspidality is used.

There are two ways of establishing the compactness of $\pi(\phi)_{L^2_{\text{cusp}}}$ from here. The first, which I learned from Lang, involves more functional analysis. The basic point is that for any $x \in G^\circ$, Lemma 2.1.9 says that the linear functional

$$T_x : L^2_{\text{cusp}}(\Gamma\backslash SL_2(\mathbf{R}), \chi) \to \mathbf{C}$$

given by

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

is bounded. By the Riesz representation theorem, it follows that for all such x, there exists a $q_x \in L^2_{\text{cusp}}(\Gamma\backslash SL_2(\mathbf{R}), \chi)$ such that $T_x(f) = \langle f, q_x\rangle$. The map $x \mapsto q_x$ from G° to L^2_{cusp} has bounded image, because by Lemma 2.1.9

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

so $\|q_x\|_{L_2} \leq C_\phi$ for all x. Also, since $L^2_{\text{cusp}}(\Gamma\backslash G^\circ, \chi) \cong L^2_{\text{cusp}}(\mathcal{F})$ has a countable orthonormal basis[13] $\{u_i\}$, we can write

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

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

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

is a measurable bounded function on G°. Restricting it to a fundamental domain $\mathcal{F}$ for $\Gamma\backslash SL_2(\mathbf{R})$, which has finite volume, and using the Hilbert space isomorphism $L^2_{\text{cusp}}(\Gamma\backslash SL_2(\mathbf{R}), \chi) \cong L^2_{\text{cusp}}(\mathcal{F})$, the function $x \mapsto \langle q_x, q_x \rangle$ is therefore in $L^1(\mathcal{F})$. When $g(x, y)$ is the characteristic function of $U \times V$ the product of measurable sets in X, we have

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

so this iterated integral is well-defined as long as $g(x, y)$ is a step function on $\mathcal{F} \times \mathcal{F}$. By the Cauchy–Bunyakovsky–Schwarz inequality[14] and Lemma 2.1.9, we have (still only as long as g is a step function, which is the only case in which we have established the left hand side is a real thing)

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

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

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

densely defined on the step functions and given by

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

is continuous where it is defined and is therefore extends to all of $L^2(\mathcal{F} \times \mathcal{F})$. By the Riesz representation

[13] $L^2(\mathcal{F})$ is separable by general theory, and L^2_{cusp} is a closed subspace and thus separable too.
[14] technically speaking, one has to repeat the proof to deduce what follows.

theorem, there exists a $Q(\cdot,\cdot) \in L^2(\mathcal{F}\times\mathcal{F})$ such that

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

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

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

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

Note that the only time the assumption of cuspidality was used was to establish the estimate $\|\pi(\phi)f\|_{L^\infty}\leq C_\phi\|f\|_{L^2}$, and this was only used to show that the evaluation-at-x functional was bounded. The rest of the proof is not dependent on the specifics of the situation at all, and is a general technique in functional analysis. □

It is the existence of these compact operators that allows us to decompose the cuspidal subspace discretely. My understanding is that this argument was first written down in [GGPS1969]. It appeared later in [Lan1976] in a more general context but containing essentially the same ideas.

Theorem 2.1.11 (Gelfand–Graev–Piatetski-Shapiro, 1966)**.** *Let $(\pi,\mathfrak{H})$ be the right regular representation of G° on $\mathfrak{H}_{\mathrm{cusp}} = L^2_{\mathrm{cusp}}(\Gamma\backslash G^\circ/Z^\circ,\chi)$. Then we have a discrete decomposition of $\mathfrak{H}_{\mathrm{cusp}}$ as a Hilbert space orthogonal direct sum of irreducible representations of G°*

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

Proof. The basic technique of the proof is the same as usual: let $\mathfrak{H}'$ be a nonzero closed subspace of $\mathfrak{H}_{\mathrm{cusp}}$ which is closed under the action of G°. We will show that $\mathfrak{H}'$ contains a nontrivial irreducible representation of G°, which will show by Zorn's lemma[15] (via the fact that π is unitary) that the desired decomposition exists (though not a priori with finite multiplicity).

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

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

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

$$\langle\pi(\phi)v,w\rangle = \int_{G^\circ}\phi(g)\langle\pi(g)v,w\rangle\,dg$$

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

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

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

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

The fact that the multiplicities are finite does not lie deeper than the rest of the statement: by the spectral theorem for compact self-adjoint operators, $\pi(\phi)$ diagonalizes and has eigenvalues going to zero, so each eigenspace with nonzero eigenvalue is finite-dimensional. Also, from its definition, $\pi(\phi)$ restricts to a well-defined G-intertwining operator on each irreducible component π_i, where it must act as a scalar by Schur's lemma. This scalar only depends on i, so the fact that the nonzero eigenvalues have finite multiplicity implies $m_i < \infty$ as well, as long as $\pi(\phi)$ doesn't act by zero on π_i. This could technically happen, but can be avoided easily, by choosing some nonzero $f \in \pi_i$ and then choosing[16] $\phi \in C_c^\infty(G^\circ)$ such that $|\pi(\phi)f - f|$ is small enough that $\pi(\phi)f$ cannot vanish.

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

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

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

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

Since $0 \neq L_0 \subset \mathfrak{V}$, the definition guarantees that $\mathfrak{V} \neq 0$, and it remains to show that $\mathfrak{V}$ is irreducible. It is irreducible because of the minimal nature of its construction: if it had a proper subrepresentation $\mathfrak{V}_1$,

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

then $\mathfrak{V}_1 \cap V_\lambda$ has to be properly contained in L_0 by the minimality of $\mathfrak{V}$. Unless $\mathfrak{V}_1 = 0$, this contradicts the minimality of L_0. So we just need to show that $\mathfrak{V}_1$ can be chosen so that $\mathfrak{V}_1 \cap V_\lambda \neq 0$. Since π is unitary, we actually have $\mathfrak{V} = \mathfrak{V}_1 \oplus \mathfrak{V}_2$ for closed subrepresentations $\mathfrak{V}_i$. Taking intersections with V_λ, we have

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

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

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

As a consequence of Theorem 2.1.11, after taking the appropriate K°-isotypic subspace, we have

Corollary 2.1.12. *For any element $f \in L^2_{\text{cusp}}(\Gamma\backslash\mathbf{H}, \chi, k)$, we have*

$$f = \sum_{u_j} \langle f, u_j \rangle u_j$$

where u_j runs over the Maass cusp forms of weight k.

2.1.2 | The continuous spectrum and Eisenstein series

Because of the weight-raising and weight-lowering operators $\hat{R}$ and $\hat{L}$, for the Maass forms not coming from modular forms, it suffices to study those of weight 0 and 1. For our purposes, we will stick to those of weight 0. In any event, to finish off the decomposition of $L^2(\Gamma\backslash\mathbf{H})$, we need to construct some non-cuspidal Maass forms. The standard way to do this is via the *Eisenstein series*. This section mostly follows [Iwa2002].

From now on, take $\Gamma = SL_2(\mathbf{Z})$ and $\chi = 1$. This will serve to make things slightly more convenient for us, since normally there is one type of Eisenstein series for every cusp, and one has to consider all of them at the same time. But the main ideas of the theory, as they are relevant to our case[17], will not change.

Definition 2.1.13. For any given $\psi \in C_c^\infty(\mathbf{R}_{>0})$, the corresponding *incomplete Eisenstein series* is the smooth function

$$E(z|\psi) = \sum_{(\Gamma\cap N)\Gamma} \psi(\Im \gamma z).$$

[17] In general there is an issue with the *residual spectrum* coming from residues of Eisenstein series. This issue does not really come up in as serious a way for $\Gamma = SL_2(\mathbf{Z})$.

defined on $\mathbf{H}$

These incomplete Eisenstein series are not exactly what we want to end up with, for example because they are not necessarily eigenfunctions of the Laplacian, but they are a good starting point because of

Lemma 2.1.14. *The span of the $E(z|\psi)$ over all ψ is the orthogonal complement of $L^2_{\text{cusp}}(\Gamma\backslash\mathbf{H})$.*

Proof. First, the fact that $E(z|\psi)$ is invariant under Γ is clear from the definition (that is why we sum over Γ), and the fact that it is square-integrable is clear from the fact that ψ is compactly supported (in fact $E(z|\psi)$ is bounded, which automatically makes it square-integrable thanks to the y^{-2} factor in the hyperbolic measure on $\mathbf{H}$). Let $f \in L^2(\Gamma\backslash\mathbf{H})$. Then

$$\begin{aligned}
\langle f, E(z|\psi)\rangle &= \int_{\mathcal{F}[\Gamma\backslash\mathbf{H}]} f(z)\overline{\sum_{\gamma\in(\Gamma\cap N)\backslash\Gamma}\psi(\Im\gamma z)}\,\frac{dxdy}{y^2} \\
&= \sum_{\gamma\in(\Gamma\cap N)\backslash\Gamma}\int_{\gamma\mathcal{F}[\Gamma\backslash\mathbf{H}]} f(z)\overline{\psi(y)}\,\frac{dxdy}{y^2} \\
&= \int_0^\infty\int_0^1 f(x+iy)\overline{\psi(y)}\frac{dxdy}{y^2} \\
&= \int_0^\infty\left[\int_0^1 f(x+iy)\,dx\right]\overline{\psi(y)}y^{-2}\,dy,
\end{aligned}$$

where $\mathcal{F}[\Gamma\backslash\mathbf{H}]$ denotes the standard fundamental domain in $\mathbf{H}$ for $\Gamma\backslash\mathbf{H}$. If this quantity vanishes for all ψ, then it must have been the case that

$$\int_0^1 f(x+iy)\,dx = 0$$

for almost all y, i.e. that f is cuspidal. The result follows. □

The actual family of Eisenstein series that is useful to us is the one given by Definition 2.1.13 except with the non-compactly supported function $\psi(y) = y^s$ for some $s \in \mathbf{C}$. The reason for this choice is that $x + iy \mapsto y^s$ is the simplest example of a nontrivial eigenfunction for Δ:

Lemma 2.1.15. *The Eisenstein series*

$$E(z,s) = \sum_{\gamma\in(\Gamma\cap N)\backslash\Gamma}(\Im\gamma z)^s$$

is an eigenfunction of Δ with eigenvalue $s(1-s)$.

Proof. Since Δ is Γ-invariant, it suffices to show that $x + iy \mapsto y^s$ is a Δ-eigenfunction with appropriate eigenvalue. And indeed we have

$$\Delta y^s = -y^2\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right)y^s = s(1-s)y^s.$$

□

On the other hand, the Eisenstein series $E(z,s)$ are very much not square-integrable on $\Gamma\backslash\mathbf{H}$. This is because of what their Fourier expansions look like.

Proposition 2.1.16. *For $\Re s > 1$, we have the Fourier expansion*

$$E(x+iy,s) = y^s + \varphi(s)y^{1-s} + \sum_{n\neq 0} \varphi(n,s)\cdot 2y^{1/2}K_{s-1/2}(2\pi|n|y)e^{2\pi inx},$$

where $K_{s-1/2}$ denotes the K-Bessel function,

$$\varphi(s) = \pi^{1/2}\frac{\Gamma(s-1/2)}{\Gamma(s)}\sum_c c^{-2s}\sum_{a\in(\mathbf{Z}/c\mathbf{Z})^\times} 1,$$

and

$$\varphi(n,s) = 2\pi^{1/2}\Gamma(s)^{-1}|n|^{s-1/2}\sum_c c^{-2s}\sum_{a\in(\mathbf{Z}/c\mathbf{Z})^\times} e^{2\pi i\frac{an}{c}}.$$

Proof. First, observe that for

$$\begin{pmatrix} a & b \\ c & d\end{pmatrix} \neq I \in \Gamma = SL_2(\mathbf{Z}),$$

we have

$$\begin{pmatrix} 1 & m \\ 0 & 1\end{pmatrix}\begin{pmatrix} a & b \\ c & d\end{pmatrix}\begin{pmatrix} 1 & n \\ 0 & 1\end{pmatrix} = \begin{pmatrix} a+mc & b+md+na+nmc \\ c & d+nc\end{pmatrix}.$$

This shows that the non-identity double cosets $(\Gamma\cap N)\backslash\Gamma/(\Gamma\cap N)$ are determined by the bottom-left entry c of a representative plus the value of $d \mod c$. Computing explicitly, we can conclude that

$$\begin{aligned} E(z,s) &= \sum_{\gamma\in(\Gamma\cap N)\backslash\Gamma} \Im(\gamma z)^s \\ &= (\Im z)^s + 2\sum_{c\geq 1}\sum_{d\in(\mathbf{Z}/c\mathbf{Z})^\times}\sum_{n\in\mathbf{Z}} \Im\left[\begin{pmatrix} a & b \\ c & d\end{pmatrix}(z+n)\right]^s, \end{aligned}$$

where the matrix $\begin{pmatrix} a & b \\ c & d\end{pmatrix}$ is an arbitrary one in Γ with bottom row (c,d). By Poisson summation, we have

$$\begin{aligned} \sum_{n\in\mathbf{Z}} \Im\left[\begin{pmatrix} a & b \\ c & d\end{pmatrix}(z+n)\right]^s &= \sum_{n\in\mathbf{Z}}\int_{\mathbf{R}}\sum_{n\in\mathbf{Z}} \Im\left[\begin{pmatrix} a & b \\ c & d\end{pmatrix}(z+t)\right]^s e^{-2\pi int}\,dt \\ &= \sum_{n\in\mathbf{Z}}\int_{\mathbf{R}} \Im\left[\frac{a}{c} - \frac{1}{c^2(t+x+d/c+iy)}\right]^s e^{-2\pi int}\,dt \\ &= \sum_{n\in\mathbf{Z}} \exp\left(2\pi i\left(nx+n\frac{d}{c}\right)\right)\int_{\mathbf{R}} \Im\left[\frac{a}{c} - \frac{1}{c^2(t+iy)}\right]^s e^{-2\pi int}\,dt \\ &= \sum_{n\in\mathbf{Z}} \exp\left(2\pi i\left(nx+n\frac{d}{c}\right)\right)\int_{\mathbf{R}} \left[\frac{yc^{-2}}{t^2+y^2}\right]^s e^{-2\pi int}\,dt. \end{aligned}$$

Returning to the original computation, and putting the sum over n on the outside instead of the inside, we see that the $n = 0$ term is

$$2\sum_{c\geq 1} c^{-2s} y^s \left[\int_{\mathbf{R}} \frac{1}{(t^2+y^2)^s}\, dt\right] \sum_{d\in(\mathbf{Z}/c\mathbf{Z})^\times} 1.$$

and the others are

$$\sum_{n\neq 0} \left[e^{2\pi i n x} \int_{\mathbf{R}} \frac{1}{(t^2+y^2)^s} e^{-2\pi i n t}\, dt\right] y^s \cdot 2\sum_{c\geq 1} c^{-2s} \sum_{d\in(\mathbf{Z}/c\mathbf{Z})^\times} e^{2\pi i n d/c}$$

which proves the claimed Fourier expansion thanks to the standard definite integrals

$$\int_{\mathbf{R}} \frac{1}{(t^2+y^2)^s}\, dt = \pi^{1/2} \frac{\Gamma(s-1/2)}{\Gamma(s)} y^{1-2s}$$

and

$$\int_{\mathbf{R}} \frac{1}{(t^2+y^2)^s} e^{-2\pi i n t}\, dt = 2\pi^s \Gamma(s)^{-1} |n|^{s-1/2} y^{-s+1/2} K_{s-1/2}(2\pi|n|y).$$

□

By the standard estimates on K-Bessel functions, for any fixed s, $E(z,s)$ is dominated by the cuspidal terms $y^s + \varphi(s) y^{1-s}$. So the closest that $E(z,s)$ gets to being square-integrable is when $\Re s = 1/2$ (in which case it barely fails to be square-integrable). But to even talk about that, we need to meromorphically extend $E(z,s)$ to the left of $\Re s > 1$. Luckily, at least for our choice of $\Gamma = SL_2(\mathbf{Z})$, this is a direct consequence of the Fourier expansion Lemma 2.1.15, thanks to the fact that the Fourier coefficients theselves have meromorphic continuations. In particular,

$$\begin{aligned}
\varphi(s) &= \pi^{1/2} \frac{\Gamma(s-1/2)}{\Gamma(s)} \sum_{c\geq 1} c^{-2s} \phi_{\text{Euler}}(c) \\
&= \pi^{1/2} \frac{\Gamma(s-1/2)}{\Gamma(s)} \prod_p \sum_{n\geq 0} p^{-2ns} \phi_{\text{Euler}}(p^n) \\
&= \pi^{1/2} \frac{\Gamma(s-1/2)}{\Gamma(s)} \prod_p \left(1 + \sum_{n\geq 1} \frac{p^n - p^{n-1}}{p^{2ns}}\right) \\
&= \pi^{1/2} \frac{\Gamma(s-1/2)}{\Gamma(s)} \prod_p \left(1 + \left(p^{1-2s} - p^{-2s}\right) \frac{1}{1-p^{1-2s}}\right) \\
&= \pi^{1/2} \frac{\Gamma(s-1/2)}{\Gamma(s)} \prod_p \left(\frac{1-p^{-2s}}{1-p^{1-2s}}\right) \\
&= \pi^{1/2} \frac{\Gamma(s-1/2)}{\Gamma(s)} \frac{\zeta(2s-1)}{\zeta(2s)}.
\end{aligned}$$

And

$$\begin{aligned}
\varphi(n,s) &= \pi^s\Gamma(s)^{-1}|n|^{s-1}\sum_{c\geq 1} c^{-2s}\sum_{d\in(\mathbf{Z}/c\mathbf{Z})^\times} e^{2\pi i\frac{dn}{c}} \\
&= \pi^s\Gamma(s)^{-1}|n|^{s-1}\prod_p\left(1+\sum_{m\geq 1}p^{-2ms}\sum_{d\in(\mathbf{Z}/p^m\mathbf{Z})^\times} e^{2\pi i\frac{dn}{p^m}}\right) \\
&= \pi^s\Gamma(s)^{-1}|n|^{s-1}\prod_p\left(1+\sum_{m\geq 1}p^{-2ms}\left(\sum_{d=0}^{p^m-1} e^{2\pi i\frac{dn}{p^m}} - \sum_{d=0}^{p^{m-1}-1} e^{2\pi i\frac{dn}{p^{m-1}}}\right)\right) \\
&= \pi^s\Gamma(s)^{-1}|n|^{s-1}\prod_p\left(1+\sum_{m\geq 1}p^{-2ms}\left(\sum_{d=0}^{p^m-1} e^{2\pi i\frac{dn}{p^m}} - \sum_{d=0}^{p^{m-1}-1} e^{2\pi i\frac{dn}{p^{m-1}}}\right)\right).
\end{aligned}$$

The sum on the inside may be evaluated case-by-case. If $v_p(n)\geq m$, then it is p^m-p^{m-1}. If $v_p(n)=m-1$, then it is $-p^{m-1}$. Otherwise, it vanishes. So we may continue the computation, finding that

$$\begin{aligned}
\varphi(n,s) &= \pi^s\Gamma(s)^{-1}|n|^{s-1}\prod_p\left(1+\sum_{m=1}^{v_p(n)}p^{-2ms}(p^m-p^{m-1}) - p^{-2(v_p(m)+1)s}p^{v_p(m)}\right) \\
&= \pi^s\Gamma(s)^{-1}|n|^{s-1}\prod_p\left(\frac{(1-p^{-2s})(1+p^{(v_p(n)+1)(1-2s)})}{1-p^{1-2s}}\right) \\
&= \pi^s\Gamma(s)^{-1}\zeta(2s)^{-1}|n|^{s-1}\prod_p\left(\sum_{m=0}^{v_p(n)}p^{m(1-2s)}\right) \\
&= \pi^s\Gamma(s)^{-1}\zeta(2s)^{-1}|n|^{-1/2}\sum_{ab=|n|}\left(\frac{a}{b}\right)^{s-\frac{1}{2}}.
\end{aligned}$$

So we can deduce from our newfound explicit knowledge of the Fourier expansion

Corollary 2.1.17. *$E(z,s)$ extends meromorphically in $s\in\mathbf{C}$ to a function with a single simple pole, at $s=1$. The residue at that pole is $3/\pi$ (in particular it doesn't depend on z). Finally, it satisfies the functional equation*

$$E(z,1-s)=\varphi(1-s)E(z,s).$$

Proof. This follows directly from the computations we have done for the Fourier expansion of $E(z,s)$, along with standard facts about the poles, zeros, and residues of the Gamma and Riemann ζ-functions. The only nontrivial fact used is that $\zeta(s)\neq 0$ when $\Re(s)=1$ (which is standard and part of the proof of the prime number theorem). □

This allows us to look at the family of functions on the upper half-plane $E(z,s)$ with $\Re s=1/2$, which still barely fails to be square-integrable because of the cuspidal terms in the Fourier expansion (which are of order $y^{1/2}$).

Finally we are equipped to write down $L^2(\Gamma\backslash\mathbf{H})$ as the discrete part plus a direct integral of Eisenstein series with $\Re(s)=1/2$. We already know (Lemma 2.1.14) that the general incomplete Eisenstein series

$E(z|\psi)$ for compactly-supported smooth ψ span the orthogonal complement of $L^2_{\text{cusp}}(\Gamma\backslash\mathbf{H})$. The key manipulation to get from $E(z|\psi)$ to $E(z,s)$ is to use the Mellin inversion formula. In particular, if

$$(\mathcal{M}\psi)(s) = \int_0^\infty x^{-s-1}\psi(x)\,dx$$

denotes the Mellin transform of f (with a slightly different convention than usual), then by the Mellin inversion formula, for $\sigma > 1$, we have

$$\begin{aligned}\frac{1}{2\pi i}\int_{\Re s=\sigma}(\mathcal{M}\psi)(s)E(z,s)\,ds &= \sum_{\gamma\in(\Gamma\cap N)\backslash\Gamma}\frac{1}{2\pi i}\int_{\Re s=\sigma}(\mathcal{M}\psi)(s)\Im(\gamma z)^s\,ds\\ &= \sum_{\gamma\in(\Gamma\cap N)\backslash\Gamma}(\mathcal{M}^{-1}\mathcal{M}\psi)(\Im(\gamma z))\\ &= \sum_{\gamma\in(\Gamma\cap N)\backslash\Gamma}\psi(\Im(\gamma z))\\ &= E(z|\psi).\end{aligned}$$

Therefore, we obtain $E(z|\psi)$ as an integral of Eisenstein series as s traverses a vertical line of real part larger than 1. Shifting the contour of integration to the left until $\Re s = 1/2$ and using the fact that the Eisenstein series all have a pole only at $s = 1$ of residue $3/\pi$, we obtain

$$E(z|\psi) = (\mathcal{M}\psi)(1)\frac{3}{\pi} + \frac{1}{2\pi}\int_{t=-\infty}^{\infty}(\mathcal{M}\psi)\left(\frac{1}{2}+it\right)E\left(z,\frac{1}{2}+it\right)\,dt.$$

From this, at the very least we may conclude that every element of $L^2(\Gamma\backslash\mathbf{H})$ is (arbitrarily close to) a linear combination of (possibly constant) Maass cusp forms of weight 0 plus an integral of Eisenstein series along the line $\Re s = 1/2$. Notice that we have seen that even though $E(z, 1/2+it)$ fails to be square-integrable for any fixed t, it does become square integrable when integrated with respect to t against appropriate functions (for instance the Mellin transform of compactly-supported ψ).

We aren't completely done yet, because the numbers $(\mathcal{M}\psi)(1/2+it)$ aren't necessarily equal to $\langle E(z|\psi), E(z,1/2+it)\rangle$. The problem is that the $E(z,1/2+it)$ are not square-integrable, so their inner products are not defined (which is a problem because we need them to be orthogonal). To fix this, we compute (by the same technique as the proof of Theorem 2.1.21, exploiting the fact that non-cuspidal terms in the Fourier cancel die when integrated horizontall, and the automorphicity of $E(z,s)$)

$$\begin{aligned}\langle E(z|\psi), E(z,1/2+it)\rangle &= \int_{\mathcal{F}[\Gamma\backslash\mathbf{H}]}\left(\sum_{\gamma\in(\Gamma\cap N)\backslash\Gamma}\psi(\Im(\gamma(x+iy)))\right)E(x+iy,1/2-it)\,\frac{dxdy}{y^2}\\ &= \int_0^\infty\psi(y)y^{-2}\int_{-\frac{1}{2}}^{\frac{1}{2}}E(x+iy,1/2-it)\,dxdy\\ &= \int_0^\infty\psi(y)y^{-2}(y^{1/2-it}+\varphi(1/2-it)y^{1/2+it})\,dy\end{aligned}$$

$$= \int_0^\infty \psi(y) y^{-3/2-it}\, dy + \varphi(1/2-it) \int_0^\infty \psi(y) y^{-3/2+it}\, dy$$
$$= (\mathcal{M}\psi)(1/2+it) + \varphi(1/2-it)(\mathcal{M}\psi)(1/2-it)$$

Multiplying by $E(z, 1/2+it)$ and using the functional equation (Corollary 2.1.17), we get

$$\langle E(z|\psi), E(z,1/2+it)\rangle E(z,1/2+it) = (\mathcal{M}\psi)(1/2+it)E(z,1/2+it) + (\mathcal{M}\psi)(1/2-it)E(z,1/2-it),$$

which after integrating t over the real line becomes

$$\int_{-\infty}^\infty \langle E(z|\psi), E(z,1/2+it)\rangle E(z,1/2+it) = 2 \cdot \frac{1}{i} \int_{\Re s = 1/2} (\mathcal{M}\psi)(s) E(z,s)\, ds$$
$$= 4\pi E(z|\psi) - \frac{2}{i}(\mathcal{M}\psi)(1)\frac{3}{\pi}$$

so we may conclude that

$$E(z|\psi) - \frac{1}{4\pi}\int_{-\infty}^\infty \langle E(z|\psi), E(z,1/2+it)\rangle E(z,1/2+it)$$

is a constant function (for arbitrary congruence subgroups we would find that this difference is in the residual spectrum coming from the residues of Eisenstein series), i.e. part of the 0-eigenspace of Δ. Thus, we conclude, as a result of this computation along with

Theorem 2.1.18. *Any $f \in L^2(\Gamma\backslash\mathbf{H})$ decomposes as*

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

where u_j ranges over an orthonormal basis of $L^2(\Gamma\backslash\mathbf{H})$ consisting of normalized eigenvectors of Δ, plus the constant function u_0 which is normalized to have $\|u_0\|_{L^2} = 1$.

Still, the Eisensten series $E(z, 1/2+it)$ are not square-integrable, and, as discussed above, their inner products with each other (over varying $t \in \mathbf{R}$) do not converge. The canonical way to solve this problem is to approximate the Eisenstein series by truncating them. This discussion of truncation is necessary for the development of the trace formula.

Iwaniec does the truncation in the naive way by just deleting the cuspidal term, but this has the disadvantage that the truncated function is not automorphic. So we follow the convention of Arthur's truncation operator, which is just as well since that is the convention typically used when dealing with truncation in the trace formula. In reality, the computations will be identical to those in Iwaniec, because Arthur's truncation operator agrees with Iwaniec's on a fundamental domain where the integral defining an inner product is defined.

Definition 2.1.19. For large T, define the *truncated cuspidal term* of $E(z,s)$ as expected,

$$c_{\text{Eis}}^T(x+iy,s) = \begin{cases} y^s + \varphi(s)y^{1-s} & \text{if } y > T \\ 0 & \text{if } y \le T \end{cases}$$

Definition 2.1.20 (Arthur's truncation operator)**.** For the Eisenstein series $E(z,s)$, define the truncation $\Lambda^T E(z,s)$ to be

$$\Lambda^T E(z,s) = E(z,s) - \sum_{\gamma \in (\Gamma\cap N)\backslash\Gamma} c_{\text{Eis}}^T(\gamma z, s).$$

As promised, for T large enough, we have $\Im(\gamma z) < T$ for all $\gamma \neq I$ in $(\Gamma\cap N)\backslash\Gamma$, which makes Definition 2.1.20 equivalent to Iwaniec's definition when restricted to the standard fundamental domain, namely

$$\Lambda^T E(x+iy,s) = \begin{cases} E(x+iy,s) & \text{if } y \le T \\ E(x+iy,s) - y^s - \varphi(s)y^{1-s} & \text{if } y > T \end{cases}.$$

In any event, the point now is that $\Lambda^T E(x+iy,s)$ decays quickly at the cusp and is therefore square-integrable. Of course it isn't smooth and is not an eigenfunction for Δ. We are ultimately interested in making approximations with truncated Eisenstein series and computing inner products involving them. The basic computation for this is

Theorem 2.1.21 (The Maass–Selberg relations)**.** *For $s_1, s_2 \neq 1$ with $s_1 \neq \overline{s_2}$ and $s_1 + \overline{s_2} \neq 1$, we have*

$$\begin{aligned}\langle \Lambda^T E(z,s_1), \Lambda^T E(z,s_2)\rangle = {} & (s_1 - \overline{s_2})^{-1}\varphi(\overline{s_2})T^{s_1-\overline{s_2}} + (\overline{s_2} - s_1)^{-1}\varphi(s_1)T^{\overline{s_2}-s_1} \\ & + (s_1+\overline{s_2}-1)^{-1}T^{s_1+\overline{s_2}-1} - (s_1+\overline{s_2}-1)^{-1}\varphi(s_1)\varphi(\overline{s_2})T^{1-s_1-\overline{s_2}}\end{aligned}$$

Proof. Iwaniec's proof is by applying Green's formula, but I am not sure if this can be generalized as easily as the standard proof, which is as follows. First, we compute (using the convenient fact that $\Lambda^T(z,s)$ is Γ-invariant in z)

$$\begin{aligned}\left\langle \Lambda^T E(z,s), \sum_{\gamma\in(\Gamma\cap N)\backslash\Gamma} c_{\text{Eis}}^T(\gamma z,s)\right\rangle &= \int_{\mathcal{F}[\Gamma\backslash\mathbf{H}]} \Lambda^T E(x+iy,s) \sum_{\gamma\in(\Gamma\cap N)\backslash\Gamma} \overline{c_{\text{Eis}}^T(\gamma(x+iy),s)}\,\frac{dxdy}{y^2} \\ &= \int_{-\frac12}^{\frac12}\int_0^\infty \Lambda^T E(x+iy,s) c_{\text{Eis}}^T(x+iy,s)\,\frac{dxdy}{y^2} \\ &= \int_{-\frac12}^{\frac12}\int_0^\infty \Lambda^T E(x+iy,s) c_{\text{Eis}}^T(x+iy,s)\,\frac{dydx}{y^2} \\ &= \int_{-\frac12}^{\frac12}\int_T^\infty \Lambda^T E(x+iy,s)(y^{\overline{s_2}} + \varphi(\overline{s_2})y^{1-\overline{s_2}})\,\frac{dydx}{y^2}. \\ &= \int_T^\infty (y^{\overline{s_2}} + \varphi(\overline{s_2})y^{1-\overline{s_2}}) \int_{-\frac12}^{\frac12} \Lambda^T E(x+iy,s)\,dx\,\frac{dy}{y^2}.\end{aligned}$$

Because of the Fourier expansion (Proposition 2.1.16), for any fixed value of y, $\Lambda^T E(x+iy,s)$ is expanded

in Fourier series with no constant term, and hence its integral on a fundamental domain for $\mathbf{Z}\backslash\mathbf{R}$ vanishes. So we have shown that

$$\left\langle \Lambda^T E(z,s), \sum_{\gamma\in(\Gamma\cap N)\backslash\Gamma} c_{\mathrm{Eis}}^T(\gamma z, s) \right\rangle = 0,$$

and thus

$$\langle \Lambda^T E(z,s_1), \Lambda^T E(z,s_2)\rangle = \langle \Lambda^T E(z,s_1), E(z,s_2)\rangle$$

(which is well-defined even though $E(z,s_2)$ is not square-integrable, as we have just shown). Now we may compute (using the fact that the Eisenstein series is automorphic and the fact that the cuspidal term in the Fourier expansion is computed in the usual way)

$$\begin{aligned}
\langle \Lambda^T E(z,s_1), \Lambda^T E(z,s_2)\rangle &= \langle \Lambda^T E(z,s_1), E(z,s_2)\rangle \\
&= \int_{\mathcal{F}[\Gamma\backslash\mathbf{H}]} \left(\sum_{\gamma\in(\Gamma\cap N)\backslash\Gamma} (\Im(\gamma(x+iy))^{s_1} - c_{\mathrm{Eis}}^T(\gamma(x+iy), s_1)) \right) E(z,\overline{s_2})\,\frac{dxdy}{y^2} \\
&= \int_0^\infty \int_{-\frac{1}{2}}^{\frac{1}{2}} (y^{s_1} - c_{\mathrm{Eis}}^T(x+iy,s_1)) E(x+iy,\overline{s_2})\,\frac{dxdy}{y^2} \\
&= \int_0^T \int_{-\frac{1}{2}}^{\frac{1}{2}} y^{s_1} E(x+iy,\overline{s_2})\,\frac{dxdy}{y^2} - \int_T^\infty \int_{-\frac{1}{2}}^{\frac{1}{2}} (\varphi(s_1) y^{1-s_1}) E(x+iy,\overline{s_2})\,\frac{dxdy}{y^2} \\
&= \int_0^T y^{s_1}(y^{\overline{s_2}} + \varphi(\overline{s_2}) y^{1-\overline{s_2}})\,\frac{dy}{y^2} - \varphi(s_1)\int_T^\infty y^{1-s_1}(y^{\overline{s_2}} + \varphi(\overline{s_2}) y^{1-\overline{s_2}})\,\frac{dy}{y^2}
\end{aligned}$$

and the rest follows from direct computation of these definite integrals. □

Recall that we are really interested in what happens on the line $\Re s = \frac{1}{2}$. Let $s = s_1 = s_2 = \sigma + it$. Then for $\sigma \neq \frac{1}{2}$ and $t \neq 0$ (so that $s_1 + \overline{s_2} \neq 1$ and $s_1 \neq \overline{s_2}$ as required by Theorem 2.1.21), we have shown that

$$\langle \Lambda^T E(z,s), \Lambda^T E(z,s)\rangle = \frac{\varphi(\sigma-it)T^{2it} - \varphi(\sigma+it)T^{-2it}}{2it} + \frac{T^{2\sigma-1} - \varphi(\sigma+it)\varphi(\sigma-it)T^{1-2\sigma}}{2\sigma-1}.$$

As $\sigma \to 1/2$, the problem is that we have some division by zero in the last two terms. Luckily, the blowing up that happens in those two terms cancels out, as we have the Taylor expansions near $\sigma = 1/2$

$$T^{2\sigma-1} = 1 + (\log T)(2\sigma-1) + O((\sigma-1/2)^2),$$

$$T^{1-2\sigma} = 1 - (\log T)(2\sigma-1) + O((\sigma-1/2)^2),$$

and

$$\varphi(\sigma+it) = \varphi(1/2+it) + (\sigma-1/2)\varphi'(1/2+it) + O((\sigma-1/2)^2)$$

which implies that $\varphi(\sigma+it)\varphi(\sigma-it)$ is

$$\varphi(1/2+it)\varphi(1/2-it) + (\sigma-1/2)(\varphi'(1/2+it)\varphi(1/2-it) - \varphi'(1/2-it)\varphi(1/2+it)).$$

With the added information that $\varphi(s) = \varphi(1-s)^{-1}$ from the functional equation (Corollary 2.1.17), this simplifies to

$$\varphi(\sigma+it)\varphi(\sigma-it) = 1 + (2\sigma-1)\varphi'(1/2+it)\varphi(1/2+it)^{-1} + O((\sigma-1/2)^2).$$

Plugging this back into our previous expression for $\langle \Lambda^T E(z,s_1), \Lambda^T E(z,s_2)\rangle$ and taking the limit as $\sigma \to 1/2$, we obtain

Corollary 2.1.22. *For nonzero $t \in \mathbf{R}$, we have*

$$\left\langle \Lambda^T E\left(z, \frac{1}{2}+it\right), \Lambda^T E\left(z, \frac{1}{2}+it\right)\right\rangle = \frac{\varphi\left(\frac{1}{2}-it\right)T^{2it} - \varphi\left(\frac{1}{2}-it\right)T^{-2it}}{2it} + 2\log T - \varphi'\left(\frac{1}{2}+it\right)\varphi\left(\frac{1}{2}+it\right)^{-1}.$$

And for $t = 0$, we obtain (by taking the limit)

$$\left\langle \Lambda^T E\left(z, \frac{1}{2}\right), \Lambda^T E\left(z, \frac{1}{2}\right)\right\rangle = 4\log T - 2\varphi'\left(\frac{1}{2}\right).$$

Chapter 3

The Arthur–Selberg trace formula

> "Now witness the firepower of this fully armed and operational battle station!"
>
> Emperor Sheev Palpatine to Luke Skywalker, *Star Wars Episode VI: Return of the Jedi*

In this chapter, we follow the basic references [Hej1976, Iwa2002, KL2006, GJ1979, Art2005] about the various forms of the trace formula that will be necessary for this thesis.

3.1 | The general approach of the trace formula and of its applications

Given a topological group G and a discrete subgroup Γ, we have spent the beginning of this chapter studying the right regular representation π of G acting on $L^2(\Gamma\backslash G)$. Prominently featured in the theory of this representation were the operators

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

for compactly supported functions ϕ on G (see Proposition 2.1.10).

The Arthur–Selberg trace formula is essentially a way of computing the trace of the $\pi(\phi)$ in two different ways: one side is the *spectral side* and comes from decomposing $L^2(\Gamma\backslash G)$ into irreducible representations and looking at the trace on each one of those. The other side is the *geometric side* and comes from writing $\pi(\phi)$ as an integral operator coming from a certain kernel and then computing its trace as a sum of *orbital integrals* corresponding to conjugacy classes of γ.

Thus far, we have been only looking at real groups. In that scenario, we have had $G = SL_2(\mathbf{R})$ and $\Gamma = SL_2(\mathbf{Z})$, though in our more concrete analysis of Eisenstein series we restricted to the $SO_2(\mathbf{R})$-isotypic subspace corresponding to the trivial representation of $SO_2(\mathbf{R})$ (hopefully it was clear how to generalize it to the subspace of arbitrary weight just by using the Iwasawa decomposition).

Manipulating things formally without any regard to convergence, let $\mathcal{F}$ be a fundamental domain for $\Gamma\backslash G$, and for $g \in G$ and $f \in L^2(\Gamma\backslash G)$, observe that

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

so that $\pi(\phi)$ can be viewed as a Hilbert–Schmidt integral operator with kernel

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

In particular, even if $\mathcal{F}$ is noncompact, the fact that ϕ is compactly supported means that the sum defining K_ϕ is finite. If we were to make the additional assumption that $\Gamma\backslash G$ were compact[1], then the L^2 space would decompose discretely (since there are no cusps), and we would have $K_\phi \in L^2(\mathcal{F} \times \mathcal{F})$. The operator $\pi(\phi)$ is compact (as shown in Proposition 2.1.10), so there is an orthonormal basis $\{u_i\}$ of $L^2(\Gamma\backslash G)$ that diagonalizes it. If $\pi(\phi)u_i = a_i u_i$ for each i, then since $\pi(\phi)$ is given by the integral kernel $K_\phi(\cdot,\cdot)$, we have

$$K_\phi(g,h) = \sum_i a_i u_i(g)\overline{u_i(h)},$$

and hence (if $\pi(\phi)$ is actually of trace-class)

$$\begin{aligned} \operatorname{Tr}(\pi(\phi)) &= \sum_i a_i \\ &= \sum_i a_i \langle u_i, u_i \rangle \\ &= \int_{\mathcal{F}} \sum_i a_i u_i(g)\overline{u_i(g)} \, dg \\ &= \int_{\mathcal{F}} K_\phi(g,g) \, dg. \end{aligned}$$

In general, we are therefore interested in computing the integral of K_ϕ along the diagonal, which is supposed to equal the trace of $\pi(\phi)$ in good circumstances. This integral in turn is best written as a sum of *orbital integrals* corresponding to conjugacy classes in Γ.

The basic statement in the case of compact quotient is as follows.

Theorem 3.1.1. *Suppose that $L^2(\Gamma\backslash G, \chi)$ decomposes into irreducibles as*

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

[1]The compactness assumption might omit a lot of spaces of arithmetic interest, such as the space $SL_2(\mathbf{Z})\backslash SL_2(\mathbf{R})$ we have been studying so far, but by the uniformization theorem and Gauss–Bonnet, it still includes the compact Riemann surfaces of genus ≥ 2.

If $\pi(\phi)$ is trace-class, then

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

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

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

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

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

This computation is reproduced from [Art2005, §1]:

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

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

The trace formula Theorem 3.1.1 breaks down on both sides for reasons essentially related to the existence of a proper parabolic subgroup in G. On the left hand side, this is related to the technicalities of the continuous spectrum; on the right hand side, one sees that the divergence of the integral comes from conjugacy classes of Γ which meet a proper parabolic subgroup of G. Arthur [Art1978] has made this principle explicit, showing that both sides converge if and only if G has a proper parabolic subgroup. In this thesis, we are interested mainly in GL_2, where there tend to be nontrivial parabolic subgroups (in particular the upper-triangular Borel subgroup). So we will have no choice but to deal with the convergence issue by Arthur's technique of truncation (which we have already discussed in Section 2.1.2 in the context of truncation of Eisenstein series).

In any particular realization of the Arthur–Selberg trace formula, there are basically three steps:

1. Choose the test function ϕ.

2. Simplify the spectral side.

3. Simplify the geometric side (usually this is done by case-by-case analysis of the different types of conjugacy classes: elliptic, parabolic, hyperbolic, and identity).

One interesting thing about Theorem 3.1.1 is that it is a generalization of Poisson summation. Here is how one obtains Poisson summation from the general perspective of the trace formula.

Example 3.1.2 (Poisson summation)**.** Let $G = \mathbf{R}$ and $\Gamma = \mathbf{Z}$. By Fourier analysis (a.k.a. knowledge of the irreducible representations of $SO_2(\mathbf{R}) = S^1 = \mathbf{Z}\backslash\mathbf{R}$), we know that $L^2(\mathbf{Z}\backslash\mathbf{R})$ is the closure of the span of the functions $x \mapsto e^{2\pi inx}$ for $n \in \mathbf{Z}$. These are eigenfunctions of the Laplacian $\Delta = -\frac{\partial^2}{\partial x^2}$ with eigenvalues $4\pi^2 n^2$. For a test function ϕ smooth on $\mathbf{R}$ with good decay properties, we have a well-defined kernel

$$K_\phi(x, y) = \sum_{n \in \mathbf{Z}} \phi(y - x + n).$$

So the geometric side of the trace formula reads

$$\sum_{n \in \mathbf{Z}} \mu(\mathbf{Z}\backslash\mathbf{R})\phi(n) = \sum_{n \in \mathbf{Z}} \phi(n).$$

On the other hand, $\pi(\phi)$ acts on the eigenfunctions by

$$(\pi(\phi)e^{2\pi in\cdot})(x) = \int_{\mathbf{R}} e^{2\pi in(x+y)}\phi(y)\, dy = e^{2\pi inx} \int_{\mathbf{R}} e^{2\pi iny}\phi(y)\, dy$$

which means the basis we have of eigenfunctions is also a diagonalizing basis for $\pi(\phi)$, with eigenvalues $\hat{\phi}(-n)$. The trace, equal to the geometric side computed above, is the sum of those eigenvalues, and hence we obtain the Poisson summation formula

$$\sum_{n \in \mathbf{Z}} \phi(n) = \sum_{n \in \mathbf{Z}} \hat{\phi}(n).$$

Notice that all of the analytic issues associated with which test functions ϕ can be plugged in are all absorbed into the convergence issues of the trace formula; once this is taken care of, the trace formula becomes a powerful and systematic way to interpret spectral theory in terms of geometry.

There are generally two ways in which the trace formula is applied.

1. Apply it for a single group and exploit the resulting nontrivial identities. This splits into two general strategies:

   (a) Design a test function so that the spectral side is something we understand and the geometric side is something we want. This is the strategy behind proving prime geodesic theorems from the trace formula, and why the error term in the prime geodesic theorem depends on whether there are small eigenvalues present in the spectral side. It is also how our approach to Conjecture 4.2.1 will function: the class numbers will appear on the geometric side as a result of thinking carefully about the orbital integrals, and we will use additional information about modular forms to manage the spectral side.

(b) Design a test function so that the geometric side is something we understand and the spectral side is something we want. This is the strategy behind proving Weyl's law for the distribution of Laplace eigenvalues.

2. Apply the trace formula for two *different* groups, match up orbital integrals on the geometric side, and use this to extract a correspondence between automorphic forms on the two groups. This is the strategy that was intended by Langlands to be the way to prove his functoriality conjecture, which seems so far to indeed be the most promising route [JL1970, GJ1979, Lan1980, Tun1981, Ngô2010].

3.2 | The Selberg trace formula for $SL_2(\mathbf{R})$

In this section, we follow Iwaniec [Iwa2002] and Hejhal [Hej1976] in deriving the trace formula for $L^2(SL_2(\mathbf{Z})\backslash SL_2(\mathbf{R}))$. As discussed above, in this setting there is the added difficulty of the continuous spectrum and the parabolic conjugacy classes in $SL_2(\mathbf{R})$. We deal with these issues by Arthur's method of truncation. For convenience, and because we constructed the Eisenstein series in this setting, we restrict to the weight-0 K°-isotypic subspace, so that we are really looking at $L^2(\Gamma\backslash\mathbf{H})$ where $\Gamma = SL_2(\mathbf{Z})$. Again, we make the claim here that everything here will work for arbitrary congruence subgroup Γ, but our proofs are only complete for $\Gamma = SL_2(\mathbf{Z})$ because that is the context in which we developed the theory of Eisenstein series (for convenience).

Before anything else, a discussion of the test function. In this setting, the choice of test function $\phi \in C_c^\infty(SL_2(\mathbf{R}))$ that actually makes a difference is somewhat limited, for the following reason. The kernel whose corresponding integral operator we want to take the trace of is

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

Since $\phi((\sigma g)^{-1}(\sigma\gamma h)) = \phi(g^{-1}\gamma h)$, the data from ϕ that we care about is really the data of a smooth function

$$\phi : SL_2(\mathbf{R}) \times SL_2(\mathbf{R}) \to \mathbf{C}$$

with the property that $\phi(g,h) = \phi(\sigma g, \sigma h)$ for any $\sigma \in SL_2(\mathbf{R})$. This new perspective is convenient for when we look at the K°-invariant subspace, since we don't have to worry about multiplying the underlying elements of $SL_2(\mathbf{R})$. In particular, the trace formula is meant to compute the trace of the integral operator given by the kernel

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

where

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

is invariant under $SL_2(\mathbf{R})$ acting diagonally. Since $PSL_2(\mathbf{R}) = \mathrm{Isom}(\mathbf{H})$, it follows that $\phi(z,w)$ only

depends on the hyperbolic distance between z and w, which meand we can rewrite it as

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

where Φ is a smooth function on $\mathbf{R}$. In principle Φ is compactly supported or at least has good decay properties. In the literature on the trace formula in this context, such functions ϕ are called *Selberg's point-pair invariants*. For the geometric side, we are interested in computing the quantity

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

where $\mu(z)$ denotes the hyperbolic measure on $\mathbf{H}$ and subscripts denote centralizers. We will compute these orbital integrals by casework on the type of conjugacy classes. For now, we go back to the spectral side.

3.2.1 | The spectral side

Recall from Section 2.1.2 that $L^2(\Gamma\backslash\mathbf{H})$ is a direct sum of Maass cusp forms of weight 0 plus a direct integral of Eisenstein series. To compute the spectral side of the trace, we therefore need to understand the action of $\pi(\phi)$ on the Maass cusp forms and on the Eisenstein series (Example 3.1.2 is a good example for why we should expect this to work). In fact, this computation is insensitive to whether the Maass form is cuspidal.

Lemma 3.2.1. *Suppose $f : \mathbf{H} \to \mathbf{C}$ is such that $\Delta_0 f = \lambda f$ for some $\lambda \in \mathbf{C}$. Then*

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

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

This proof is taken from [Hej1976, Proposition 3.1].

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

$\sigma i = z$, and then we have

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

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

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

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

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

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

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

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

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

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

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

This proves the result, because

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

and the integral on the right hand side only depends on λ and Φ. $\square$

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

It isn't obvious (to me) that Λ is entire from its definition as an integral involving a function satisfying a differential equation depending on λ. Instead, one uses the result of Lemma 3.2.1: to compute Λ, we may choose any test function f we want[2], as long as it has $\Delta_0 f = \lambda f$. This is what allows for

Lemma 3.2.2. *For $r \in \mathbf{C}$, we have*

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

where

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

In fact, Λ is entire.

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

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

So by Lemma 3.2.1,

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

Taking

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

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

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

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

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

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

$$= \int_{-\infty}^{\infty} e^{iru} \int_{e^u+e^{-u}-2}^{\infty} \frac{\Phi(t)}{\sqrt{t-e^u-e^{-u}+2}}\, dt\, du$$

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

Definition 3.2.3. The function h is called the *Harish-Chandra transform* of the test function Φ. From now on, Φ is considered to be fixed, and h is defined to be its Harish-Chandra transform.

Corollary 3.2.4. *The integral kernel K_ϕ decomposes in $L^2(\mathcal{F}\times\mathcal{F})$ as*

$$K_\phi(z,\zeta) = \sum_j h(r_j) u_j(z)\overline{u_j(\zeta)} + \frac{1}{4\pi}\int_{\mathbf{R}} h(r) E\left(z, \frac{1}{2}+ir\right)\overline{E\left(\zeta, \frac{1}{2}+ir\right)}\, dr$$

where the r_j are such that $\frac{1}{4}+r_j^2 = \lambda_j$ the Δ-eigenvalue of u_j.

Proof. Direct consequence of Theorem 2.1.18 and Lemma 3.2.1. N.B.: one has to check that K_ϕ really is square-integrable in both coordinates as long as h satisfies the appropriate decay properties. This whole time, we are implicitly restricting to test functions Φ such that h satisfies those decay properties. □

Regardless, the integral operator induced by the kernel K_ϕ will likely not be of trace class. In particular, according to Corollary 3.2.4 and our previous computations, if it was of trace-class, we would have

$$\begin{aligned}\mathrm{Tr}\pi(\phi) &= \int_{\mathcal{F}[\Gamma\backslash\mathbf{H}]} K_\phi(z,z)\, d\mu(z) \\ &= \sum_j h(r_j)\|u_j\|_{L^2}^2 + \frac{1}{4\pi}\int_{\mathbf{R}} h(r)\left\|E\left(z,\frac{1}{2}+ir\right)\right\|_{L^2}^2 dr.\end{aligned}$$

This does not make sense because $E(z,s)$ is (barely) not square-integrable even when $\Re s = 1/2$. So we are forced to compute the truncated trace, which is done by replacing this integral with an integral over the truncated fundamental domain $\mathcal{F}[\Gamma\backslash\mathbf{H}]_{y\le T}$ for large T. In that case, we are after the quantity we now call

$$\mathrm{Tr}^T\pi(\phi) = \sum_j h(r_j)\int_{\mathcal{F}[\Gamma\backslash\mathbf{H}]_{y\le T}} |u_j(z)|^2\, d\mu(z) + \frac{1}{4\pi}\int_{\mathbf{R}} h(r)\int_{\mathcal{F}[\Gamma\backslash\mathbf{H}]_{y\le T}}\left|E\left(z,\frac{1}{2}+ir\right)\right|^2 d\mu(z)dr \tag{3.1}$$

and since we are looking at the truncated fundamental domain, we might as well replace E with the truncated Eisenstein series $\Lambda^T E$ (see Definition 2.1.20). Since the truncated fundamental domains are compact, there are no issues with convergence here.

The terms coming from truncated Eisenstein series must now be estimated. By Corollary 2.1.22, we have (for $r \neq 0$ which doesn't contribute anything to the integral anyway)

$$\int_{\mathcal{F}[\Gamma\backslash\mathbf{H}]_{y\le T}}\left|E^T\left(z,\frac{1}{2}+ir\right)\right|^2 d\mu(z) \le \int_{\mathcal{F}[\Gamma\backslash\mathbf{H}]}\left|E^T\left(z,\frac{1}{2}+ir\right)\right|^2 d\mu(z)$$

$$= \frac{\varphi\left(\frac{1}{2}-ir\right)T^{2ir} - \varphi\left(\frac{1}{2}-ir\right)T^{-2ir}}{2ir} + 2\log T - \frac{\varphi'}{\varphi}\left(\frac{1}{2}+ir\right).$$

So the whole term in Equation (3.1) coming from Eisenstein series is bounded above by

$$\frac{1}{4\pi}\int_{\mathbf{R}} h(r)\left[\frac{\varphi\left(\frac{1}{2}-ir\right)T^{2ir} - \varphi\left(\frac{1}{2}-ir\right)T^{-2ir}}{2ir} + 2\log T - \frac{\varphi'}{\varphi}\left(\frac{1}{2}+ir\right)\right] dr.$$

Only the first term is something we are interested in simplifying further. Indeed, we have (by the evenness of h)

$$\frac{1}{8\pi i}\int_{\mathbf{R}} \frac{h(r)}{r}\left[\varphi\left(\frac{1}{2}-ir\right)T^{2ir} - \varphi\left(\frac{1}{2}-ir\right)T^{-2ir}\right] dr$$
$$= \frac{1}{4\pi i}\int_{\mathbf{R}} \frac{h(r)}{r}\left[\varphi\left(\frac{1}{2}-ir\right)T^{2ir} - \varphi\left(\frac{1}{2}\right)\right] dr,$$

where the $\varphi(1/2) = 1$ term is subtracted from $\varphi(1/2-ir)T^{2ir}$ to make $r^{-1}h(r)(\varphi(1/2-ir)T^{2ir} - \varphi(1/2))$ integrable for r ranging over $\mathbf{R}$ (and in particular the integrand does not have a pole at $r=0$). Since

$$\varphi(s) = \pi^{1/2}\frac{\Gamma(s-1/2)}{\Gamma(s)}\frac{\zeta(2s-1)}{2s}$$

is bounded on small neighborhoods of the line $\Re s = 1/2$, if h decays fast enough, we can move the contour of integration up from $\mathbf{R}$ to $\mathbf{R}+\epsilon i$, which then allows us to integrate both terms separately (since $h(r)/r$ no longer has any poles on the line of integration). There are no poles in the way, so we are left with

$$\frac{1}{4\pi i}\int_{\Im r=\epsilon} \frac{h(r)}{r}\left[\varphi\left(\frac{1}{2}-ir\right)T^{2ir} - 1\right] dr$$
$$= \frac{1}{4\pi i}\int_{\Im r=\epsilon} \frac{h(r)}{r}\varphi\left(\frac{1}{2}-ir\right)T^{2ir}\,dr - \frac{1}{4\pi i}\int_{\Im r=\epsilon}\frac{h(r)}{r}\,dr.$$

Again, since φ is actually bounded on small neighborhoods of $\Re s = 1/2$, for small ϵ, the first term is

$$\begin{aligned}\frac{1}{4\pi i}\int_{\Im r=\epsilon} \frac{h(r)}{r}\varphi\left(\frac{1}{2}-ir\right)T^{2ir}\,dr &\ll \int_{\Im r=\epsilon}\frac{h(r)}{r}T^{2ir}\\ &\ll \int_{\mathbf{R}}\frac{h(r+i\epsilon)}{r+i\epsilon}T^{2ir-2\epsilon}\\ &\ll T^{-2\epsilon}\end{aligned}$$

and the second term is

$$-\frac{1}{4\pi i}\int_{\Im r=\epsilon}\frac{h(r)}{r}\,dr = \frac{1}{4}h(0),$$

by moving the contour of integration down to $\Im r = -\epsilon$, using the symmetry of h and the fact that $h(r)/r$

has only a simple pole at $r = 0$ of residue $h(0)$. Anyway, we conclude from this the upper bound

$$\mathrm{Tr}^T \pi(\phi) \leq \sum_j h(r_j) + g(0) \log T - \frac{1}{4\pi} \int_{\mathbf{R}} h(r) \frac{\varphi'}{\varphi}\left(\frac{1}{2} + ir\right) dr + \frac{1}{4} h(0) + O(T^{-2\epsilon}),$$

where

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

is the (suitably normalized) Fourier transform of h.

In fact, this upper bound is supposed to be an equality, as we show now.

Proposition 3.2.5 (Spectral side of the trace formula)**.** *For large T,*

$$\mathrm{Tr}^T \pi(\phi) = \sum_j h(r_j) + g(0) \log T - \frac{1}{4\pi} \int_{\mathbf{R}} h(r) \frac{\varphi'}{\varphi}\left(\frac{1}{2} + ir\right) dr + \frac{1}{4} h(0) + O(T^{-2\epsilon}).$$

Proof. The upper bound in this equality was proven by integrating $K_\phi(z, z)$ (except with Eisenstein series replaced with truncated Eisenstein series) over the entire fundamental domain rather than the truncated one. So it suffices to show that the contribution of the remainder of the fundamental domain is absorbed into the error term. For the constant function u_0 (which makes up the residual spectrum in our case $\Gamma = SL_2(\mathbf{Z})$), we have

$$\begin{aligned} \int_{\mathcal{F}[\Gamma\backslash\mathbf{H}]_{y \geq T}} |u_0(z)|^2 \, d\mu(z) &\ll \mu(\mathcal{F}[\Gamma\backslash\mathbf{H}]_{y \geq T}) \\ &\ll \frac{1}{T}. \end{aligned}$$

And since Maass cusp forms have exponential decay at the cusp (thanks to the Fourier expansion), the same bound holds for all the u_j's (except the implied constant might depend on s_j; this doesn't end up mattering — we just end up with an infinite sum of these implied constants times $h(s_j)$ which converges thanks to the decay properties of h). As for the Eisenstein series term, the same bound

$$\int_{\mathbf{R}} h(r) \int_{\mathcal{F}[\Gamma\backslash\mathbf{H}]_{y \geq T}} \left| \Lambda^T E\left(z, \frac{1}{2} + ir\right) \right|^2 d\mu(z) dr \ll_h \frac{1}{T}$$

holds, again thanks to the fact that the truncated Eisenstein series decay exponentially at the cusp. The extra T^{-1} are absorbed in the $O(T^{-2\epsilon})$ error, as long as ϵ was chosen to be sufficiently small. □

3.2.2 | The geometric side

We are now meant to evaluate the geometric side of the trace formula, namely

$$\mathrm{Tr}^T \pi(\phi) = \int_{\mathcal{F}[\Gamma\backslash\mathbf{H}]_{y \leq T}} K_\phi(z, z) \, d\mu(z)$$

$$= \sum_{\gamma \in \{\Gamma\}} \int_{\mathcal{F}[\Gamma_\gamma \backslash \mathbf{H}]_{y \leq T}} \phi(z, \gamma z)\, d\mu(z).$$

It remains to compute the orbital integrals $\int_{\mathcal{F}[\Gamma_\gamma \backslash \mathbf{H}]_{y \leq T}} \phi(z, \gamma z)\, d\mu(z)$ for each type of conjugacy class γ. As a warm-up, we begin with the identity term. To do this, we need to make an observation about the integral transform involved in the definition of $g(u)$.

Lemma 3.2.6. *If $\Phi \in C_c^\infty(\mathbf{R})$, then for any $t \in \mathbf{R}$,*

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

Proof. The proof of this formula is the main part of [Hej1976, Proposition 4.1]. I don't know why it is not a coincidence, but it seems related to the theory of the Abel transform . The fact that Φ is compactly supported makes all of our manipulations below kosher. First, observe that

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

So we may compute

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

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

Proposition 3.2.7. *The identity term on the geometric side is*

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

where h is the holomorphic function defined as above.

Proof. By definition, we have

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

and from Lemma 3.2.6 (substituting $e^u + e^{-u} - 2$ for x),

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

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

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

so differentiating under the integral sign yields

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

hence

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

as desired. □

Since $\Gamma\backslash\mathbf{H}$ has finite volume, the truncation really makes no difference to us for the identity term: as $T \to \infty$, the identity term will converge to

$$\frac{\mu(\Gamma\backslash\mathbf{H})}{2\pi}\int_0^\infty rh(r)\tanh(\pi r)\, dr$$

anyway. This is part of the general theme that the divergence in the orbital integrals is caused only by the proper parabolic subgroups, so the only thing truncation will be necessary for is the parabolic conjugacy classes. For this reason, we do those last.

For now, we do the hyperbolic classes. For a hyperbolic $\gamma \in \Gamma$, we know that γ is conjugate in $PSL_2(\mathbf{R})$ to a transformation of the form $z \mapsto N(\gamma)z$ for some positive constant $N(\gamma)$. That quantity is also characterized by the fact that

$$\begin{aligned}\inf_{z\in\mathbf{H}} d_{\mathbf{H}}(z,\gamma z) &= \inf_{z\in\mathbf{H}} d_{\mathbf{H}}(z,N(\gamma)z)\\ &= \int_{x+iy\in\mathbf{H}} \int_y^{N(\gamma)y} \frac{dy}{y}\\ &= \log N(\gamma).\end{aligned}$$

Lemma 3.2.8. *If $\gamma \in SL_2(\mathbf{Z})$ is hyperbolic, then*

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

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

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

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

This simplifies to the actual claim, using the fact that

$$g(u) = \int_{e^u+e^{-u}} \frac{\Phi(t)}{\sqrt{t - \frac{(N(\gamma)-1)^2}{N(\gamma)}}}$$

□

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

Now, time for the elliptic term. This term basically never has any impact on anything, because there are only finitely many elliptic elements of $SL_2(\mathbf{Z})$ or any other congruence subgroup – most elements have no fixed points in $\mathbf{H}$, which is why it is easy to define the Riemann surface $\Gamma\backslash\mathbf{H}$. For an elliptic element $\gamma \in \Gamma$, its centralizer is finite cyclic (it is the stabilizer of the fixed point of γ) and generated by some γ_0.

Proposition 3.2.9. *The elliptic orbital integrals evaluate to*

$$\int_{\mathcal{F}[\Gamma_\gamma\backslash\mathbf{H}]} \phi(z,\gamma z)\, d\mu(z) = \frac{\tau}{2\pi\sin\tau} \int_{\mathbf{R}} h(r) \frac{\cosh(\pi r - 2\tau r)}{\cosh \pi r}\, dr,$$

where $\tau \in [0,\pi)$ is the angle such that γ is conjugate in $SL_2(\mathbf{R})$ to a rotation by angle $\pm\tau$.

Proof. Every elliptic element is conjugate in $SL_2(\mathbf{R})$ to some

$$\kappa_\tau = \begin{pmatrix} \cos\tau & -\sin\tau \\ \sin\tau & \cos\tau \end{pmatrix}.$$

So we have

$$\gamma = \eta^{-1}\kappa_\tau\eta,$$

and hence (by the same argument as in Lemma 3.2.8)

$$\int_{\mathcal{F}[\Gamma_\gamma\backslash\mathbf{H}]} \phi(z,\gamma z)\, d\mu(z) = \int_{\mathcal{F}[\Gamma_{\kappa_\tau}\backslash\mathbf{H}]} \phi(z,\kappa_\tau z)\, d\mu(z).$$

The fundamental domain for Γ_{κ_τ} might not be as easy to describe as the fundamental domain for hyperbolic elements, but we are saved by the fact that κ_τ has finite order (as the stabilizer of i in Γ is finite), namely π/τ. So all of $\mathbf{H}$ is tiled by translates of $\mathcal{F}[\Gamma_{\kappa_\tau}\backslash\mathbf{H}]$ by κ_τ, and the orbital integral over all of these regions are the same thanks to the $SL_2(\mathbf{R})$-invariance of ϕ. So in fact the value we are interested in computing does not require us to understand this fundamental domain, and we just need to compute

$$\frac{\tau}{\pi} \int_{\mathbf{H}} \phi(z,\kappa_\tau z)\, d\mu(z).$$

This is easiest to compute in the coordinates that come from the Cartan decomposition of $SL_2(\mathbf{R})$, rather than the Iwasawa decomposition (makes sense, since we are interested in the action of K°). In

particular, we have

$$SL_2(\mathbf{R}) = K^\circ A K^\circ,$$

where A is the subgroup of diagonal matrices. This gives us new coordinates

$$z = \kappa_\tau e^{-r} i$$

for $\theta \in [0, \pi)$ and $r \in [0, \infty)$. One checks that

$$d\mu(z) = (2 \sinh r)\, dr d\theta,$$

from which it follows that

$$\begin{aligned}
\int_{\mathcal{F}[\Gamma_\gamma \backslash \mathbf{H}]} \phi(z, \gamma z)\, d\mu(z) &= \frac{\tau}{\pi} \int_{\mathbf{H}} \phi(z, \kappa_\tau z)\, d\mu(z) \\
&= \frac{\tau}{\pi} \int_0^\pi \int_0^\infty \phi(\kappa_\theta e^{-r} i, \kappa_\tau \kappa_\theta e^{-r} i)(2 \sinh r)\, dr d\theta \\
&= \frac{\tau}{\pi} \int_0^\pi \int_0^\infty \phi(\kappa_\theta e^{-r} i, \kappa_\theta \kappa_\tau e^{-r} i)(2 \sinh r)\, dr d\theta \\
&= \tau \int_0^\infty \phi(e^{-r} i, \kappa_\tau e^{-r} i)(2 \sinh r)\, dr \\
&= \tau \int_0^\infty \Phi((\sinh r \sin \tau)^2)(2 \sinh r)\, dr \\
&= \frac{\tau}{\sin \tau} \int_0^\infty \frac{\Phi(u)}{\sqrt{u + \sin^2 \tau}}\, du \\
&= \frac{\tau}{\pi} \int_0^\infty g(r) \frac{\cosh(r/2)}{\cosh r - \cos(2\tau)}\, dr \\
&= \frac{\tau}{2\pi \sin \tau} \int_{\mathbf{R}} h(r) \frac{\cosh(\pi r - 2\tau r)}{\cosh \pi r}\, dr
\end{aligned}$$

as desired. □

It is finally time to deal with the parabolic elements. We are saved in this case by the fact that the parabolic conjugacy classes are very easy to describe: they have representatives which are in $\Gamma \cap N$, i.e. are just horizontal translations. The technical difficulty here comes from the fact that the truncation is actually necessary: as we saw in Proposition 3.2.7, Lemma 3.2.8, and Proposition 3.2.10, all the other orbital integrals converge as $T \to \infty$, which implies (by Proposition 3.2.5) that the parabolic orbital integrals have to be of order $\log T$. So in this situation we need to keep track of the truncation.

Proposition 3.2.10. *The sum of the parabolic orbital integrals is*

$$\sum_{\substack{\gamma \in \{\Gamma\} \\ \text{parabolic}}} \int_{\mathcal{F}[\Gamma_\gamma \backslash \mathbf{H}]_{y \le T}} \phi(z, z+n) = g(0) \log T - g(0) \log 2 + \frac{1}{4} h(0) - \frac{1}{2\pi} \int_{\mathbf{R}} h(r) \frac{\Gamma'}{\Gamma}(1 + ir)\, dr.$$

Proof. Every parabolic conjugacy class of Γ has a representative in Γ of the form

$$\gamma_n = \begin{pmatrix} 1 & n \\ 0 & 1 \end{pmatrix}.$$

for $n \neq 0$. The centralizer of γ_n is exactly

$$\Gamma \cap N = \left\{ \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}^n : n \in \mathbf{Z} \right\},$$

so the sum we wish to compute is

$$\begin{aligned}
\sum_{n\neq 0} \int_{\mathcal{F}[(\Gamma\cap N)\backslash \mathbf{H}]} \phi(z, z+n) &= \sum_{n\neq 0} \int_0^T \int_{-\frac{1}{2}}^{\frac{1}{2}} \Phi\left(\frac{n^2}{y^2}\right) \frac{dxdy}{y^2} \\
&= \sum_{n\neq 0} \int_0^T \Phi\left(\frac{n^2}{y^2}\right) y^{-2} dy \\
&= \sum_{n\neq 0} \frac{1}{2n} \int_{n^2/T^2}^{\infty} \frac{\Phi(u)}{\sqrt{u}}\, du \\
&= \sum_{n\geq 1} \frac{1}{n} \int_{n^2/T^2}^{\infty} \frac{\Phi(u)}{\sqrt{u}}\, du \\
&= \int_{T^{-2}}^{\infty} \frac{\Phi(u)}{\sqrt{u}} \left(\sum_{1\leq n\leq T\sqrt{u}} \frac{1}{n} \right) du \\
&= \int_{T^{-2}}^{\infty} \frac{\Phi(u)}{\sqrt{u}} \left(\log(T\sqrt{u}) + \gamma_{\text{Euler}} + O\left(\frac{1}{T\sqrt{u}}\right) \right) du
\end{aligned}$$

where γ_{Euler} is the Euler–Mascheroni constant. The contribution of the error inside the integral is of the order

$$\frac{1}{T} \int_{T^{-2}}^{\infty} \frac{\Phi(u)}{u}\, du \ll_{\Phi} \frac{1}{T} \log T,$$

and the rest is

$$\int_0^{\infty} \frac{\Phi(u)}{\sqrt{u}} \left(\log(T\sqrt{u}) + \gamma_{\text{Euler}} \right) du$$

which is

$$g(0)(\log T + \gamma_{\text{Euler}}) + \frac{1}{2} \int_0^{\infty} \frac{\Phi(u)}{\sqrt{u}} \log(u)\, du.$$

Computing further, we have

$$\frac{1}{2} \int_0^{\infty} \frac{\Phi(u)}{\sqrt{u}} \log(u)\, du = -g(0)\gamma_{\text{Euler}} + \frac{1}{4} h(0) - \frac{1}{2\pi} \int_{\mathbf{R}} h(r) \frac{\Gamma'}{\Gamma}(1+ir)\, dr,$$

from which the desired result follows. □

Now that we have computed the spectral side and all the necessary orbital integrals, we can write

down the final version of the trace formula in this context.

Theorem 3.2.11 (Selberg trace formula for $SL_2(\mathbf{R})$). *Let $h(s)$ be an even holomorphic function defined on a neighborhood of the strip $|\Im s| \leq 1/2$ satisfying the appropriate growth conditions such that the spectral side below converges absolutely, and let $g(u) = \frac{1}{2\pi}\int_{\mathbf{R}} e^{-iru}h(r)\,dr$ be its Fourier transform. Then*

$$\begin{aligned}\sum_j h(r_j) - \frac{1}{4\pi}\int_{\mathbf{R}} h(r)\frac{\varphi'}{\varphi}\left(\frac{1}{2}+ir\right)\,dr \\ = -g(0)\log 2 - \frac{1}{2\pi}\int_{\mathbf{R}} h(r)\frac{\Gamma'}{\Gamma}(1+ir)\,dr \\ + \sum_{\gamma\in\{\Gamma\}_{\mathrm{Ell}}} \frac{\tau_\gamma}{2\pi\sin\tau_\gamma}\int_{\mathbf{R}} h(r)\frac{\cosh(\pi r - 2\tau_\gamma r)}{\cosh \pi r}\,dr \\ + \sum_{\gamma\in\{\Gamma\}_{\mathrm{Hyp}}} \frac{\log N(\gamma_0)}{N(\gamma)^{1/2}-N(\gamma)^{-1/2}} g(\log N(\gamma)) \\ + \frac{\mu(\mathcal{F}[\Gamma\backslash\mathbf{H}])}{2\pi}\int_0^\infty rh(r)\tanh(\pi r)\,dr.\end{aligned}$$

Proof. Combine Proposition 3.2.5, Proposition 3.2.7, Proposition 3.2.10, Lemma 3.2.8, and Proposition 3.2.9. Subtract the term proportional to $\log T$ from both sides, and then take $T \to \infty$. □

Now is a good time to mention all the technical analytic conditions we have swept under the rug in the proofs. In particular, we have been assuming this whole time (as made explicit in the hypothesis to Theorem 3.2.11) that the truncated operators are actually of trace-class. This is true when Φ is compactly supported, but as we see in Theorem 3.2.11, an attractive feature in this explicit version of the trace formula is that it doesn't depend on Φ itself at all — only on h and its Fourier transform[3]. So it makes more sense to ask what kind of growth condition we need to require of h to make the spectral side converge absolutely (in which case it is a formal consequence that the geometric side will also). It turns out that if h is defined on the neighborhood $|\Im s| \leq 1/2 + \delta$, then it is enough to have h satisfy the growth condition

$$h(r) \leq (|r|+1)^{-2-\delta}. \tag{3.2}$$

In the case where $\Gamma\backslash\mathbf{H}$ was compact, one could show convergence under the weaker growth condition $h(r) \leq (|r|+1)^{-4-\delta}$ by constructing a Green's function, that is, a Hilbert–Schmidt integral kernel for the resolvent of the Laplace–Beltrami operator. The existence of a Green's function in that context (which itself is a consequence of basic ODE theory) then shows that

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

where the λ_j are the Laplace eigenvalues of the Maass (cusp) forms on $\Gamma\backslash\mathbf{H}$ (see [Bum1997, §2.3]). It is harder to get this argument to work in the presence of the continuous spectrum, but luckily the following

[3] Of course, one can recover Φ and thus ϕ from h (see the definitions) — otherwise the hypotheses to Theorem 3.2.11 wouldn't make sense.

weak version of Weyl's law for the finite-volume case is fairly easy to establish given what we have developed so far.

Lemma 3.2.12.

$$\sum_{|r_j|<T} 1 + \int_{-T}^{T} \frac{\varphi'}{\varphi}(1+ir)\, dr \ll T^2.$$

Proof. By Bessel's inequality, we have, for $\zeta \in \mathbf{H}$ in the standard fundamental domain,

$$\sum_j |h(r_j)u_j(\zeta)|^2 + \frac{1}{4\pi}\int_{\mathbf{R}} |h(r)E(\zeta, 1/2+ir)|^2\, dr \leq \int_{\mathcal{F}[\Gamma\backslash\mathbf{H}]} |K_\phi(z,\zeta)|^2\, d\mu(z).$$

Taking Φ to be a smooth approximation to the characteristic function of an interval $[0, R]$, we find that this yields

$$\sum_{|r_j|<T} |u_j(\zeta)|^2 + \int_{-T}^{T} |E(\zeta, 1/2+ir)|^2\, dr \ll T^2 + Ty$$

where $\zeta = x + iy$ is chosen to be in the standard fundamental domain (this choice is convenient because it guarantees that $y = \sup_{\gamma\in\Gamma} \Im(\gamma\cdot\zeta)$). Integrating this inequality over the truncated fundamental domain $\mathcal{F}[\Gamma\backslash\mathbf{H}]_{y\leq T}$ and using Corollary 2.1.22 along with the decay properties at the cusp guaranteed by Lemma 2.1.15, we get the desired result. □

The crude bound Lemma 3.2.12 (which we will improve to a precise asymptotic in the standard application of the trace formula) is enough to establish the validity of Theorem 3.2.11 for test functions h satisfying the condition Equation (3.2).

3.3 | The Eichler–Selberg trace formula

The goal of this section is to prove the analog of Theorem 3.2.11 in the global setting, that is in the setting of automorphic forms on $GL_2(\mathbf{A_Q})$. The adelic setting is where the Hecke operators are most naturally defined, and the point of this section is to derive the Eichler–Selberg trace formula for traces of Hecke operators acting on the classical holomorphic cusp forms of level 1 and weight k. In doing this, we will follow the references of Knightly–Li [KL2006] and Gelbart–Jacquet [GJ1979].

In the language of Section 3.1, we are looking at the group $G = GL_2(\mathbf{A_Q})$ and the discrete subgroup $\Gamma = GL_2(\mathbf{Q})$. In reality, we are restricting to the case of trivial central character (since we are interested in level 1), so the actual representation we are looking at is the right regular representation π of $GL_2(\mathbf{A_Q})$ acting on

$$L^2(PGL_2(\mathbf{Q})\backslash PGL_2(\mathbf{A_Q})).$$

The automorphic forms in $L^2(PSL_2(\mathbf{Z})\backslash PSL_2(\mathbf{R}))$ embed in the obvious way in here. The trace formula is a computation of

$$\mathrm{Tr}\pi(\phi)$$

where $\phi : GL_2(\mathbf{A_Q}) \to \mathbf{C}$ is smooth (in the usual adelic sense) and is nice enough to that $\pi(\Phi)$ is of trace class.

The standard notation for this setup (which will allow us to save a little space is as follows). Let $G = GL_{2/\mathbf{Q}}$, Z its center (the scalar matrices), and $\overline{G} = G/Z = PGL_{2/\mathbf{Q}}$. As we have already seen, the key feature of G that brings all the technical difficulties is the existence of the proper parabolic subgroup P (which we take to be the upper-triangular Borel subgroup), which admits the Levi decomposition

$$P = MN,$$

where

$$M = \left\{ \begin{pmatrix} * & \\ & * \end{pmatrix} \right\}$$

and

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

3.3.1 | Choosing the test function

In real life, there are (like in the previous section) complications relating to the continuous spectrum, and an analogous theory of Eisenstein series to account for those complications. But the test function we are interested in will make sure that $\pi(\phi)$ kills the continuous spectrum, and in fact only acts on the space of weight-k holomorphic cusp forms. Define

$$\phi = \prod_{v \in M_{\mathbf{Q}}} \phi_v,$$

where each ϕ_v is smooth on $G_v = GL_2(\mathbf{Q}_v)$. It is our job to define the ϕ_v so that $\pi(\phi)$ gives us exactly the Hecke operator we want (it needs to kill the orthogonal complement of $S_k(\Gamma(1), \mathbf{C})$ in $L^2(\overline{G}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q}))$ and act as the Hecke operator T_n on $S_k(\Gamma(1), \mathbf{C})$). Once we have done this, our life will be much easier than in the general case: we will not have to worry about the continuous spectrum, and we will be guaranteed that $\pi(\phi)$ is of trace class (as it is a finite-rank operator).

Lemma 3.3.1. *Suppose $k > 2$. Let*

$$\phi_\infty(g) = d_k \langle \pi(g) f_k, f_k \rangle,$$

where d_k is the formal dimension of the weight-k discrete series representation of $GL_2(\mathbf{R})$ and f_k is any nonzero lowest-weight vector in that representation. For each finite prime p, let ϕ_p be the characteristic function of the compact subset $Z(\mathbf{Q}_p)T(n)_p$, where

$$T(n)_p = \{g \in M_2(\mathbf{Z}_p) | \det g \in n\mathbf{Z}_p^\times\} \subset GL_2(\mathbf{Q}_p).$$

Then $\pi(\phi)$ kills the orthogonal complement of $S_k(\Gamma(1), \mathbf{C})$ and acts on it by the Hecke operator $n^{1-\frac{k}{2}}T_n$.

Proof. The set $T(n)_p$ is just the set of 2×2 matrices with coefficients in $\mathbf{Z}_p$ such that the p-adic valuation

of the determinant equals that of n. By the Cartan decomposition, we have

$$GL_2(\mathbf{Q}_p) = \bigcup_{i \geq j} GL_2(\mathbf{Z}_p) \begin{pmatrix} p^i & 0 \\ 0 & p^j \end{pmatrix} GL_2(\mathbf{Z}_p),$$

which means that we can write $T(n)_p$ as a union of double cosets

$$T(n)_p = \bigcup_{\substack{i \geq j \\ i+j=v_p(n)}} GL_2(\mathbf{Z}_p) \begin{pmatrix} p^i & 0 \\ 0 & p^j \end{pmatrix} GL_2(\mathbf{Z}).$$

This means that (ignoring the archimedean place for now) $\pi(\phi_{\mathrm{fin}})$ agrees with the global definition of T_n as a double coset operator (see e.g. [Bum1997, §3.6]).

The test function at the infinite place is meant to restrict this to $S_k(\Gamma(1), \mathbf{C})$. This is thanks to the general theory of matrix coefficients. First of all, ϕ_∞ is absolutely integrable modulo center, as can be checked from the standard explicit realization of the discrete series representation (see [Bum1997, Theorem 2.6.5]; note that this is where we use the fact that $k > 2$). So (despite the fact that ϕ_∞ is not compactly supported modulo center), the operator $\pi(\phi)$ is well-defined.

Again by [Bum1997, Theorem 2.6.5], one checks that

$$\int_{N(\mathbf{R})} \phi_\infty(gnh)\, dn = 0$$

for any $g, h \in G(\mathbf{R})$, and hence that

$$\int_{N(\mathbf{A_Q})} \phi\left(gnh\right)\, dn = \left(\int_{N(\mathbf{R})} \phi_\infty(g_\infty n h_\infty)\, dn\right) \left(\int_{N(\mathbf{A_{Q,\mathrm{fin}}})} \prod_{p<\infty} \phi_p(g_p n_p h_p)\, dn\right) = 0 \tag{3.3}$$

and as a result

$$\int_{N(\mathbf{A_Q})} \phi(gn^{-1}mnh)\, dn = 0 \tag{3.4}$$

for any $m \in M(\mathbf{A_Q})$.

Therefore, for any $f \in L^2(\overline{G}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q}))$, we have

$$\begin{aligned}
\int_{N(\mathbf{Q})\backslash N(\mathbf{A_Q})} (\pi(\phi)f)(ng)\, dn &:= \int_{N(\mathbf{Q})\backslash N(\mathbf{A_Q})} \int_{\overline{G}(A)} \phi(g^{-1}n^{-1}x) f(x)\, dx dn \\
&= \int_{N(\mathbf{Q})\backslash N(\mathbf{A_Q})} \int_{N(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})} \sum_{\delta \in N(\mathbf{Q})} \phi(g^{-1}n^{-1}\delta x)\, dx dn \\
&= \int_{N(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})} f(x) \int_{N(\mathbf{Q})\backslash N(\mathbf{A_Q})} \sum_{\delta \in N(\mathbf{Q})} \phi(g^{-1}n^{-1}\delta x)\, dn dx \\
&= \int_{N(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})} f(x) \int_{N(\mathbf{A_Q})} \phi(g^{-1}n^{-1}\delta x)\, dn dx \\
&= 0
\end{aligned}$$

thanks to Equation (3.3) (here we have used the integrability of ϕ modulo center and approximated f by bounded functions to apply Fubini's theorem). In other words, $\pi(\phi)$ sends everything to the cuspidal subspace. Recall (from above or from [Bum1997, Proposition 2.3.1]) moreover that the adjoint of $\pi(\phi)$ is given by $\pi(\phi^*)$, where

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

Since ϕ_∞ is given by a matrix coefficient, we know immediately that $\phi_\infty^* = \phi_\infty$. The same is true for ϕ_p at the finite places p, since complex conjugation doesn't do anything to a characteristic function, and

$$\phi_p(g^{-1}) = \phi_p((\det g) \cdot g^{-1}) = \phi_p(g),$$

where $\det g$ is considered as an element of $Z(\mathbf{Q}_p)$, since

$$\det((\det g) \cdot g^{-1}) = \frac{\det^2 g}{\det g} = \det g.$$

Therefore, we conclude that $\pi(\phi)$ is self-adjoint and that

$$\pi(\phi) L^2(\overline{G}(\mathbf{Q})\backslash \overline{G}(\mathbf{A_Q})) \subset L^2_{\text{cusp}}(\overline{G}(\mathbf{Q})\backslash \overline{G}(\mathbf{A_Q})),$$

and thus

$$\pi(\phi) L^2_{\text{cusp}}(\overline{G}(\mathbf{Q})\backslash \overline{G}(\mathbf{A_Q}))^\perp = 0.$$

Hence, $\pi(\phi)$ acts only on the cuspidal subspace. Knowing this is convenient because of the fact that $L^2_{\text{cusp}}(\overline{G}(\mathbf{Q})\backslash \overline{G}(\mathbf{A_Q}))$ decomposes discretely into irreducible representations of $G(\mathbf{A_Q})$ with finite multiplicities. We proved this fact in the archimedean setting (Theorem 2.1.11), which is most of the content of the general proof (see [Bum1997, §3.3] for more details). In particular,

$$\pi = L^2_{\text{cusp}}(\overline{G}(\mathbf{Q})\backslash \overline{G}(\mathbf{A_Q})) = \bigoplus \pi_i$$

where the π_i's are allowed to repeat with finite multiplicity, and by the tensor product theorem (see Flath or Bump) we have

$$\pi_i = \bigotimes_{v \in M_{\mathbf{Q}}} \pi_{i,v},$$

where $\pi_{i,v}$ is an automorphic representation of $G(\mathbf{Q}_v)$. One checks directly that via this isomorphism

$$\pi(\phi)|_{\pi_i} = \pi_{i,\infty}(\phi_\infty) \otimes \pi_{i,\text{fin}}(\phi_{\text{fin}}).$$

By Schur's lemma for matrix coefficients (see [Bum2013, Theorem 2.4]), we conclude that $\pi(\phi)$ kills all of the irreducible representations π_i whose infinity-type is not the discrete series of lowest-weight k, and otherwise acts trivially on the infinite part of π_i. It follows (since the action of $\pi(\phi)$ on π_i is then exactly the same as the Hecke action of T_n, up to the usual scaling factor) that

$$\text{Tr} T_n|_{S_k(\Gamma(1),\mathbf{C})} = n^{\frac{k}{2}-1} \text{Tr} \pi(\phi),$$

as claimed. □

3.3.2 | The identity term

For our purposes, Lemma 3.3.1 is the spectral side of the trace formula. In this section and the ones that follow, we compute the geometric side, which will ultimately give a formula for $\mathrm{Tr} T_n$ in terms of class numbers of imaginary quadratic fields.

As usual, the easiest one is the identity term. That term is

$$\begin{aligned}\int_{\overline{G}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})} \phi(g^{-1}Ig) &= \phi(1)\mathrm{vol}(\overline{G}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})) \\ &= \phi(1)\frac{\pi}{3}\end{aligned}$$

(thanks to the standard computation of the Tamagawa number of PGL_2). At infinity, we have

$$\phi_\infty(1) = d_k = \frac{k-1}{4\pi}$$

(the formal degree of the discrete series representations of $GL_2(\mathbf{R})$ is a straightforward computation from the explicit description [Bum1997, Theorem 2.6.5]). As for the finite places, recall that for a rational prime $p < \infty$,

$$\phi_p(1) = \begin{cases} 1, & \text{if } 1 \in Z(\mathbf{Q}_p)T(n)_p \\ 0, & \text{otherwise} \end{cases}.$$

Since $\det(xm) = x^2 \det(m)$, we know that $\phi_p(1) = 0$ whenever $v_p(n)$ is odd. So if n is not a perfect square, $\phi_p(1) = 0$. On the other hand, if n is a perfect square, we have

$$1 = \frac{1}{\sqrt{n}} \cdot \sqrt{n},$$

where the $\sqrt{n} \in \mathbf{Q}_p^\times = Z(\mathbf{Q}_p)$ and the n is also considered as an element of $Z(\mathbf{Q}_p) \subset T(n)_p$. Hence, we can deduce

Proposition 3.3.2. *For ϕ defined as in Lemma 3.3.1, the identity term of the trace formula is*

$$\int_{\overline{G}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})} \phi(g^{-1}Ig) = \begin{cases} \frac{k-1}{12}, & \textit{if } n \textit{ is a perfect square} \\ 0 & \textit{otherwise} \end{cases}.$$

3.3.3 | The elliptic term

Now we consider the elliptic conjugacy classes of $\overline{G}(\mathbf{Q}) = PGL_2(\mathbf{Q})$. These are the conjugacy classes whose elements have zero eigenvalues in $\mathbf{Q}$. The elliptic term is particularly nice because the elliptic conjugacy classes have no intersection with the upper-triangular parabolic subgroup $P(\mathbf{Q})$. As we know from Arthur, there are never any problems with convergence on the geometric side for conjugacy classes

that do not meet $P(\mathbf{Q})$. Lifting to $GL_2(\mathbf{Q})$ (by definition of ϕ, each elliptic element of $PGL_2(\mathbf{Q})$ such that $\phi(g^{-1}\gamma g) \neq 0$ for any g has exactly two lifts to $GL_2(\mathbf{Q})$ of determinant n), the elliptic term is

$$\begin{aligned}
\int_{\overline{G}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})} \sum_{\substack{\gamma\in\overline{G}(\mathbf{Q})\\ \text{elliptic}}} \phi(g^{-1}\gamma g)\,dg &= \frac{1}{2}\int_{\overline{G}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})} \sum_{\substack{\gamma\in G(\mathbf{Q})\text{elliptic}\\ \det\gamma=n}} \phi(g^{-1}\gamma g) \\
&= \frac{1}{2}\sum_{\substack{[\gamma]\text{ elliptic conj. class}\\ \det\gamma=n}} \int_{\overline{G}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})} \sum_{\delta\in G_\gamma(\mathbf{Q})\backslash G(\mathbf{Q})} \phi(g^{-1}\delta^{-1}\gamma\delta g)\,dg \\
&= \frac{1}{2}\sum_{\substack{[\gamma]\text{ elliptic conj. class}\\ \det\gamma=n}} \int_{\overline{G}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})} \sum_{\delta\in\overline{G_\gamma(\mathbf{Q})}\backslash\overline{G}(\mathbf{Q})} \phi(g^{-1}\delta^{-1}\gamma\delta g)\,dg \\
&= \frac{1}{2}\sum_{\substack{[\gamma]\text{ elliptic conj. class}\\ \det\gamma=n}} \int_{\overline{G_\gamma(\mathbf{Q})}\backslash\overline{G}(\mathbf{A_Q})} \phi(g^{-1}\delta^{-1}\gamma\delta g)\,dg.
\end{aligned}$$

The computation of the orbital integral on the inside depends on whether γ remains elliptic in $G(\mathbf{R})$, i.e. on whether the complex roots of the characteristic polynomial of γ are not real. Since $\det\gamma = n$, the characteristic polynomial is of the form

$$p_\gamma(X) = X^2 - tX + n,$$

where $t = \mathrm{Tr}\gamma$.

Lemma 3.3.3. *If γ is diagonalizable over $\mathbf{R}$ (i.e. the roots of p_γ are real), then*

$$\int_{\overline{G_\gamma(\mathbf{Q})}\backslash\overline{G}(\mathbf{A_Q})} \phi(g^{-1}\gamma g)\,dg = 0.$$

Proof. As one can check by linear algebra, the centralizer $G_\gamma(\mathbf{Q})$ is equal to $\mathbf{Q}[\gamma]^\times = \mathbf{Q}(\sqrt{t^2-4n})^\times$ (this is true when γ is elliptic or hyperbolic in $G(\mathbf{Q})$). The same thing holds if you replace $\mathbf{Q}$ with $\mathbf{A_Q}$. In particular, $G_\gamma(\mathbf{Q})$ and $G_\gamma(\mathbf{A_Q})$ are abelian, hence unimodular, so there are no issues with writing

$$\begin{aligned}
\int_{\overline{G_\gamma(\mathbf{Q})}\backslash\overline{G}(\mathbf{A_Q})} \phi(g^{-1}\gamma g)\,dg &= \int_{\overline{G_\gamma(\mathbf{A_Q})}\backslash\overline{G}(\mathbf{A_Q})} \int_{\overline{G_\gamma(\mathbf{Q})}\backslash\overline{G_\gamma(\mathbf{A_Q})}} \phi(g^{-1}\delta^{-1}\gamma\delta g)\,d\delta dg \\
&= \mathrm{vol}(\overline{G_\gamma(\mathbf{Q})}\backslash\overline{G_\gamma(\mathbf{A_Q})}) \int_{\overline{G_\gamma(\mathbf{A_Q})}\backslash\overline{G}(\mathbf{A_Q})} \phi(g^{-1}\gamma g)\,dg,
\end{aligned}$$

since elements of $G_\gamma(\mathbf{A_Q})$ commute with γ by definition. Moreover,

$$\int_{\overline{G_\gamma(\mathbf{A_Q})}\backslash\overline{G}(\mathbf{A_Q})} \phi(g^{-1}\gamma g)\,dg = \left(\int_{\overline{G_\gamma(\mathbf{R})}\backslash\overline{G}(\mathbf{R})} \phi_\infty(g^{-1}\gamma g)\,dg\right)\left(\int_{\overline{G_\gamma(\mathbf{A}_{\mathrm{fin}})}\backslash\overline{G}(\mathbf{A}_{\mathrm{fin}})} \phi_\infty(g^{-1}\gamma g)\,dg\right)$$

and the archimedean integral vanishes already, thanks to Equation (3.4) (since γ is assumed to be diagonalizable over $\mathbf{R}$ and the archimedean orbital integral is only sensitive to the conjugacy class of

γ). $\square$

Now we suppose that $\gamma \in \overline{G}(\mathbf{Q})$ remains elliptic over $\mathbf{Q}$, i.e. that p_γ's roots are two distinct complex numbers. By the proof of Lemma 3.3.3, we still have

$$\int_{\overline{G_\gamma(\mathbf{A_Q})}\backslash\overline{G}(\mathbf{A_Q})} \phi(g^{-1}\gamma g)\,dg = \left(\int_{\overline{G_\gamma(\mathbf{R})}\backslash\overline{G}(\mathbf{R})} \phi_\infty(g^{-1}\gamma g)\,dg\right)\left(\int_{\overline{G_\gamma(\mathbf{A}_{\mathrm{fin}})}\backslash\overline{G}(\mathbf{A}_{\mathrm{fin}})} \phi_{\mathrm{fin}}(g^{-1}\gamma g)\,dg\right),$$

except the archimedean orbital integral has no reason to vanish. Instead, γ is conjugate to one of the standard elliptic elements of $PGL_2(\mathbf{R})$, i.e. elements of $K^\circ = SO_2(\mathbf{R})$, so since the orbital integral is insensitive to the conjugacy class, and $\mathrm{vol}(K^\circ) = 1$, so we are reduced to computing

$$\int_{\overline{G}(\mathbf{R})} \phi_\infty(g^{-1}\gamma g)\,dg,$$

where we can assume that $\gamma \in \sqrt{n}\cdot SO_2(\mathbf{R})$, i.e.

$$\gamma = \sqrt{n}\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}$$

where θ depends on γ. Since this has such an explicit form, we can use the Cartan decomposition $SL_2(\mathbf{R}) = K^\circ M(\mathbf{R}) K^\circ$ plus the explicit description of the discrete series of which ϕ_∞ is the matrix coefficient to explicitly compute

$$\int_{\overline{G}(\mathbf{R})} \phi_\infty(g^{-1}\gamma g)\,dg = -\frac{e^{i(k-1)\theta} - e^{-i(k-1)\theta}}{e^{i\theta} - e^{-i\theta}}.$$

The complex eigenvalues of γ are $\gamma_1 = \sqrt{n}e^{i\theta}$ and $\gamma_2 = \sqrt{n}e^{-i\theta}$, so we obtain

Lemma 3.3.4. *If $\gamma \in \overline{G}(\mathbf{Q})$ is elliptic, then*

$$\left(\int_{\overline{G_\gamma(\mathbf{R})}\backslash\overline{G}(\mathbf{R})} \phi_\infty(g^{-1}\gamma g)\,dg\right) = \begin{cases} -n^{1-\frac{k}{2}}\frac{\gamma_1^{k-1}-\gamma_2^{k-1}}{\gamma_1-\gamma_2}, & \text{if } \gamma \text{ remains elliptic in } G(\mathbf{R}) \\ 0 & \text{otherwise} \end{cases}.$$

Since γ remains elliptic over $\mathbf{R}$ if and only if its characteristic polynomial $X^2 - tX + n$ has two distinct complex roots, it follows that the elliptic term is

$$-\frac{1}{2}n^{1-\frac{k}{2}} \sum_{\substack{\gamma \\ t^2<4n}} \frac{\gamma_1 - \gamma_2^{k-1}}{\gamma_1 - \gamma_2}\left(\int_{\overline{G_\gamma(\mathbf{A}_{\mathrm{fin}})}\backslash\overline{G}(\mathbf{A}_{\mathrm{fin}})} \phi_{\mathrm{fin}}(g^{-1}\gamma g)\,dg\right).$$

Note that γ_1 and γ_2 only depend on t and $n$since they are the complex roots of the characteristic polynomial of γ.

The nonarchimedean orbital integrals are where the class numbers of imaginary quadratic fields come in.

Lemma 3.3.5. *The nonarchimedean orbital integral corresponding to a given γ which is elliptic in $G(\mathbf{Q})$ and remains so in $G(\mathbf{R})$ with characteristic polynomial $X^2 - tX + n$ is*

$$\int_{\overline{G_\gamma(\mathbf{A}_{\text{fin}})}\backslash\overline{G}(\mathbf{A}_{\text{fin}})} \phi_{\text{fin}}(g^{-1}\gamma g)\,dg = \sum_{\mathcal{O}\supset\mathbf{Z}+\mathbf{Z}\gamma} h(\mathcal{O})\frac{2}{|\mathcal{O}^\times|}$$

where the sum is over orders $\mathcal{O}$ of the imaginary quadratic field $\mathbf{Q}[\gamma] = \mathbf{Q}(\sqrt{t^2-4n})$.

Proof. Consider the maximal compact

$$\overline{K} = \prod_{p<\infty} \overline{G}(\mathbf{Z}_p).$$

By definition,

$$\phi_p(g^{-1}\gamma g)$$

is $\overline{G}(\mathbf{Z}_p)$-invariant on both sides as a function of g (since the determinant of the input doesn't change), so we can split the orbital integral into double cosets on which ϕ_{fin} is constant and then use the fact that all the measures used are Haar measures to compute

$$\begin{aligned}
\int_{\overline{G_\gamma(\mathbf{A}_{\text{fin}})}\backslash\overline{G}(\mathbf{A}_{\text{fin}})} \phi_{\text{fin}}(g^{-1}\gamma g)\,dg &= \sum_{x\in\overline{G_\gamma(\mathbf{Q})}\backslash\overline{G}(\mathbf{A}_{\text{fin}})/\overline{K}} \int_{\overline{G_\gamma(\mathbf{Q})}x\overline{K}\subset\overline{G_\gamma(\mathbf{A}_{\text{fin}})}\backslash\overline{G}(\mathbf{A}_{\text{fin}})} \phi_{\text{fin}}(g^{-1}\gamma g)\,dg \\
&= \sum_{x\in\overline{G_\gamma(\mathbf{Q})}\backslash\overline{G}(\mathbf{A}_{\text{fin}})/\overline{K}} \text{vol}(\overline{G_\gamma(\mathbf{Q})}x\overline{K})\phi_{\text{fin}}(x^{-1}\gamma x)\,dx \\
&= \sum_{x\in\overline{G_\gamma(\mathbf{Q})}\backslash\overline{G}(\mathbf{A}_{\text{fin}})/\overline{K}} \text{vol}(\overline{G_\gamma(\mathbf{Q})}x\overline{K}x^{-1})\phi_{\text{fin}}(x^{-1}\gamma x)\,dx
\end{aligned}$$

where $\overline{G_\gamma(\mathbf{Q})}x\overline{K}x^{-1}$ and $\overline{G_\gamma(\mathbf{Q})}x\overline{K}$ are considered as subsets of $\overline{G_\gamma(\mathbf{A}_{\text{fin}})}\backslash\overline{G}(\mathbf{A}_{\text{fin}})$. Conjugating the whole thing by x, we see that $\text{vol}(\overline{G_\gamma(\mathbf{Q})}x\overline{K}x^{-1})$ is the same as $\text{vol}(x^{-1}\overline{G_\gamma(\mathbf{Q})}x\overline{K})$ using the Haar measure on $x\overline{G_\gamma(\mathbf{A}_{\text{fin}})}x^{-1}\backslash\overline{G}(\mathbf{A}_{\text{fin}})$. Since $x^{-1}\overline{G_\gamma(\mathbf{Q})}x$ is discrete in $\overline{G}(\mathbf{A}_{\text{fin}})$, and $\overline{K}$ is a compact open subgroup, this measure is equal to

$$\frac{\text{vol}(\overline{K})}{|\overline{K}\cap x^{-1}\overline{G_\gamma(\mathbf{Q})}x|} = \frac{2}{|K\cap x^{-1}G_\gamma(\mathbf{Q})x|}.$$

The set $K\cap x^{-1}G_\gamma(\mathbf{Q})x$ has a natural interpretation as the set of elements $x^{-1}gx$ for $g\in G_\gamma(\mathbf{Q})$ such that $x^{-1}gx$ restricts to a linear automorphism of $\mathbf{Z}_p^2\subset\mathbf{Q}_p^2$ for each $p<\infty$. In other words, after undoing the conjugation, it is in bijection with the set of $g\in G_\gamma(\mathbf{Q}) = \mathbf{Q}[\gamma]^\times = \mathbf{Q}(\sqrt{t^2-4n})$ that act as a linear automorphism on the lattice $\Lambda_x\subset\mathbf{Q}(\sqrt{t^2-4n})$ corresponding to[4]

$$x\in\overline{G_\gamma(\mathbf{Q})}\backslash\overline{G}(\mathbf{A}_{\text{fin}})/\overline{K} = G_\gamma(\mathbf{Q})\backslash G(\mathbf{A}_{\text{fin}}/K).$$

[4]Recall that the double coset space $G_\gamma(\mathbf{Q})\backslash G(\mathbf{A}_{\text{fin}}/K$ is in bijection with the lattices in $\mathbf{Q}[\gamma] = \mathbf{Q}(\sqrt{t^2-4n}) = \mathbf{Q}^2$ up to multiplication by $G_\gamma(\mathbf{Q}) = \mathbf{Q}(\sqrt{t^2-4n})^\times$

In other words, it is exactly the set of units in the order

$$\mathcal{O}_{\Lambda_x} = \{\alpha \in \mathbf{Q}(\sqrt{t^2 - 4n}) : \alpha\Lambda_x \subset \Lambda_x\}.$$

Putting this back into our previous computations, we have found that the nonarchimedean orbital integral is

$$\int_{\overline{G_\gamma(\mathbf{A}_{\text{fin}})}\backslash\overline{G}(\mathbf{A}_{\text{fin}})} \phi_{\text{fin}}(g^{-1}\gamma g)\, dg = \sum_{x \in G_\gamma(\mathbf{Q})\backslash G(\mathbf{A}_{\text{fin}})/K} \frac{2}{|\mathcal{O}_{\Lambda_x}|}\phi_{\text{fin}}(x^{-1}\gamma x).$$

Since we are already assuming that $\det\gamma = n$, the definition of $\phi_p(x_p^{-1}\gamma_p x_p)$ is that it is 1 if and only if $x_p^{-1}\gamma_p x_p$ has coefficients in $\mathbf{Z}_p$, i.e. if and only if γ_p sends the lattice represented by x_p inside itself. Putting this local fact together, we get

$$\phi_{\text{fin}}(x^{-1}\gamma x) = \begin{cases} 1, & \text{if } \gamma \in \mathcal{O}_{\Lambda_x} \\ 0, & \text{otherwise} \end{cases},$$

hence the nonarchimedean orbital integral is

$$\sum_{\substack{\Lambda \subset \mathbf{Q}[\gamma] \\ \gamma \in \mathcal{O}_\Lambda \\ \text{up to } \mathbf{Q}[\gamma]^\times}} \frac{2}{|\mathcal{O}_\Lambda^\times|} = \sum_{\mathcal{O} \supset \mathbf{Z}+\mathbf{Z}\cdot\gamma} \frac{2}{|\mathcal{O}^\times|} h(\mathcal{O}),$$

as desired. □

Let the maximal order in $\mathbf{Q}[\gamma] = \mathbf{Q}(\sqrt{t^2 - 4n})$ be $\mathcal{O}_{\mathbf{Q}[\gamma]} = \mathbf{Z} + \mathbf{Z}\cdot\tau$, so that the non-maximal orders are all given by $\mathbf{Z} + \mathbf{Z}\cdot f\tau$ and have discriminant

$$\text{disc}(\mathbf{Z} + \mathbf{Z}\cdot f\tau) = f^2\Delta_{\mathbf{Q}[\gamma]}.$$

The discriminant of the non-maximal order $\mathbf{Z}[\gamma] = \mathbf{Z} + \mathbf{Z}\cdot\gamma$ is $t^2 - 4n$. So the orders being summed over in the formula for the nonarchimedean elliptic orbital integral in Lemma 3.3.5 are precisely the imaginary quadratic orders of discriminants

$$\frac{t^2 - 4n}{m^2}$$

for all positive integers m such that

$$m^2 | \frac{t^2 - 4n}{\Delta_{\mathbf{Q}[\gamma]}}.$$

By the usual correspondence between equivalence classes of quadratic forms of given discriminant and fractional ideal classes for the quadratic order of that discriminant, and the fact that unit groups of orders are in bijection with stabilizers of quadratic forms, the quantity

$$\sum_{\mathcal{O} \supset \mathbf{Z}+\mathbf{Z}\cdot\gamma} \frac{2}{|\mathcal{O}^\times|} h(\mathcal{O})$$

is the same thing as the *Kronecker–Hurwitz class number* of discriminant $t^2 - 4n$, that is the number of binary quadratic forms of discriminant $t^2 - 4n$, weighted by the sizes of the stabilizers in $PSL_2(\mathbf{Z})$. It doesn't really matter, but the viewpoint of quadratic forms is the one we will take in the next chapter when we apply this to quadratic forms.

In any event, since an elliptic conjugacy class in $PGL_2(\mathbf{Q})$ is determined by its characteristic polynomial, we have proved

Corollary 3.3.6. *The elliptic orbital integral is*

$$-\frac{n^{1-\frac{k}{2}}}{2}\sum_{t^2<4n}\frac{\rho^{k-1}-\overline{\rho}^{k-1}}{\rho-\overline{\rho}}H(t^2-4n),$$

where $H(t^2-4n)$ is the Kronecker–Hurwitz class number, and ρ denotes a complex root of the polynomial $X^2 - tX + n$.

Remark 3.3.7. The sum in Corollary 3.3.6 is finite and is guaranteed to converge. For the same reason (the finiteness of the set of $t \in \mathbf{Z}$ such that $t^2 < 4n$), the elliptic orbital integral we started out only had a finite sum in it to begin with (there are finitely many elliptic conjugacy classes for which the corresponding term doesn't vanish). Of course, one must still technically check the absolute convergence of the orbital integral, which we didn't bother to do (it is straightforward to check thanks to the fact that the elliptic conjugacy classes do not intersect the subgroup $P(\mathbf{Q})$).

3.3.4 | The hyperbolic orbital integral

Knightly–Li [KL2006] make the claim that for the actual test function ϕ chosen in Lemma 3.3.1, truncation is not necessary and only done for pedagogical reasons. I do not think that the correct claim is quite as strong as this. Truncation might seem unnecessary because the operator $\pi(\phi)$ kills the Eisenstein series anyway, but this does not change the fact that conjugacy classes intersecting the parabolic subgroup $P(\mathbf{Q})$ cause serious convergence problems on the geometric side.

Consider a nontrivial hyperbolic element $\gamma \in \overline{M}(\mathbf{Q})$. Its centralizer is again $\overline{M}(\mathbf{Q})$ (that is how diagonal matrices work), so the corresponding orbital integral is

$$\int_{\overline{M}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})} f(g^{-1}\gamma g)\,dg.$$

If this converged absolutely, we would be able to use Fubini to get

$$\int_{\overline{M}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})} f(g^{-1}\gamma g)\,dg = \int_{\overline{M}(\mathbf{A_Q})\backslash\overline{G}(\mathbf{A_Q})}\int_{\overline{M}(\mathbf{Q})\backslash\overline{M}(\mathbf{A_Q})} f(g^{-1}m^{-1}\gamma m g)\,dm dg.$$

Since $\overline{M}(\mathbf{A_Q})$ centralizes $\overline{M}(\mathbf{Q})$, the inner integral is the integral of a constant function, and thus equals

$$\mathrm{vol}(\overline{M}(\mathbf{Q})\backslash\overline{M}(\mathbf{A})) = \mathrm{vol}(\mathbf{Q}\backslash\mathbf{A}^\times),$$

which is infinite. So this is ABSOLUTELY NOT absolutely integrable, and we cannot actually apply the

tricks that we would like to apply. To remedy this, we recall that in the classical situation (Section 3.2) there was a natural truncation operator that cut off the cuspidal part of a function once it got near the cusps. Motivated by this (see in particular Definition 2.1.20), we repeat essentially the same process and then compute without asking too many questions.

Definition 3.3.8 (Logarithmic height function)**.** For $g \in G(\mathbf{Q}_v)$, define its height via the Iwasawa decomposition by just looking at the $M(\mathbf{Q}_v)$-coordinate, and then defining

$$H\left(\begin{pmatrix} a & \\ & b \end{pmatrix}\right) = \log\left|\frac{a}{b}\right|_v.$$

Then define the global height on $G(\mathbf{A_Q})$ as the sum of local heights at all $v \in M_{\mathbf{Q}}$.

Note that when $v = \infty$, we get the logarithm of the absolute value of the imaginary part of $g(i)$.

Define the truncation of the kernel in the following way:

$$\tilde{\Lambda}^T K_\phi(g,g) = \sum_{\gamma \in \overline{G}(F)} \phi(g^{-1}\gamma g) - \sum_{\xi \in P(\mathbf{Q})\backslash G(\mathbf{Q})} \sum_{\gamma \in \overline{P}(\mathbf{Q})} \phi(g^{-1}\xi^{-1}\gamma\xi g)\chi_{[T,\infty)}(H(\xi g)).$$

This method of truncation makes the integral over $g \in \overline{G}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})$ absolutely convergent without changing its actual value, as long as T is sufficiently large (this will become clear once we actually compute it). This particular technique of truncation is due to Arthur and the resulting integral kernel $\tilde{\Lambda}^T K_\phi$ is called *Arthur's modified kernel.*

Rstricting to some hyperbolic conjugacy class $\gamma \in [\gamma_0]$ in $\overline{G}(\mathbf{Q})$ where $\gamma_0 \in \overline{M}(\mathbf{Q})$ (such a representative is guaranteed to exist by definition of hyperbolic — there must be two distict rational eigenvalues), it is useful to note that

$$[\gamma_0] \cap \overline{M}(\mathbf{Q}) = \{\gamma_0, w\gamma_0 w^{-1}\}$$

where w is the matrix that switches the two canonical basis vectors. Splitting up the γ (in both terms) based on their conjugacy class, we see that the truncation term in the definition of $\tilde{\Lambda}_2^T K_\phi(g,g)$ equals

$$\begin{aligned}
&\sum_{[\gamma_0]} \left(- \sum_{\xi \in P(\mathbf{Q})\backslash G(\mathbf{Q})} \sum_{\gamma \in [\gamma_0] \cap \overline{P}(\mathbf{Q})} \phi(g^{-1}\xi^{-1}\gamma\xi g)\chi_{[T,\infty)}(H(\xi g)) \right) \\
&\quad = \sum_{[\gamma_0]} \left(- \sum_{\xi \in P(\mathbf{Q})\backslash G(\mathbf{Q})} \sum_{n \in N(\mathbf{Q})} (\phi(g^{-1}\xi^{-1}\gamma_0 n\xi g) + \phi(g^{-1}\xi^{-1}w\gamma_0 w^{-1} n\xi g))\chi_{[T,\infty)}(H(\xi g)) \right) \\
&\quad = \sum_{[\gamma_0]} \left(- \sum_{\xi \in P(\mathbf{Q})\backslash G(\mathbf{Q})} \sum_{n \in N(\mathbf{Q})} (\phi(g^{-1}\xi^{-1}n^{-1}\gamma_0 n\xi g) + \phi(g^{-1}\xi^{-1}n^{-1}w\gamma_0 w^{-1} n\xi g))\chi_{[T,\infty)}(H(\xi g)) \right) \\
&\quad = \sum_{[\gamma_0]} \left(- \sum_{\xi \in M(\mathbf{Q})\backslash G(\mathbf{Q})} (\phi(g^{-1}\xi^{-1}\gamma_0\xi g) + \phi(g^{-1}\xi^{-1}w\gamma_0 w^{-1}\xi g))\chi_{[T,\infty)}(H(\xi g)) \right)
\end{aligned}$$

$$= \sum_{[\gamma_0]} \left(- \sum_{\xi \in M(\mathbf{Q})\backslash G(\mathbf{Q})} \phi(g^{-1}\xi^{-1}\gamma_0\xi g)(\chi_{[T,\infty)}(H(\xi g)) + \chi_{[T,\infty)}(H(w\xi g))) \right)$$

thanks again to the Levi decomposition $P = MN$.

Since the centralizer of γ_0 is $M(\mathbf{Q})$, we can split up $K_\phi(g,g)$ in the usual way to finally obtain

$$\tilde{\Lambda}^T K_\phi(g,g) = \sum_{[\gamma_0]} \sum_{\xi \in M(\mathbf{Q})\backslash G(\mathbf{Q})} \phi(g^{-1}\xi^{-1}\gamma_0\xi g)(1 - \chi_{[T,\infty)}(H(\xi g)) - \chi_{[T,\infty)}(H(w\xi g))).$$

The truncated hyperbolic term on the geometric side is the integral of this sum over $g \in \overline{G}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})$, which we may commute with the sum over hyperbolic classes $[\gamma_0]$ and combine with the sum over $M(\mathbf{Q})\backslash G(\mathbf{Q})$ to get (by Fubini, crucially using the fact that the modified kernel is absolutely integrable)

$$\begin{aligned}
&\sum_{[\gamma_0]} \int_{\overline{M}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})} \phi(g^{-1}\gamma_0 g)(1 - \chi_{[T,\infty)}(H(g)) - \chi_{[T,\infty)}(H(wg)))\, dg \\
&\quad = \sum_{[\gamma_0]} \int_{\overline{M}(\mathbf{A})\backslash\overline{G}(\mathbf{A_Q})} \phi(g^{-1}\gamma_0 g) \int_{\overline{M}(\mathbf{Q})\backslash\overline{M}(\mathbf{A_Q})} (1 - \chi_{[T,\infty)}(H(mg)) - \chi_{[T,\infty)}(H(wmg)))\, dm dg
\end{aligned}$$

(using again the fact that $M(\mathbf{A_Q})$ centralizes γ_0).

Now we are lucky that the truncated terms prevent the inner integral (over a set of infinite measure) from diverging. In fact, we can simplify it further: using the Iwasawa decomposition $g = m_g n_g k_g$, we see that

$$H(wg) = H(wn_g) - H(g),$$

hence (as a result of the fact that $\mathbf{Q}^\times\backslash\mathbf{A}_\mathbf{Q}^\times = \mathbf{R}_{>0}^\times \times \hat{\mathbf{Z}}^\times$ so the finite places have no effect on heights)

$$\begin{aligned}
&\int_{\overline{M}(\mathbf{Q})\backslash\overline{M}(\mathbf{A_Q})} (1 - \chi_{[T,\infty)}(H(mg)) - \chi_{[T,\infty)}(H(wmg)))\, dm \\
&= \mathrm{vol}(\mathbf{Q}^\times\backslash\mathbf{A}_\mathbf{Q}^1) \int_{-\infty}^{\infty} \left(1 - \chi_{[T,\infty)}\left(H\left(\begin{pmatrix} e^r & \\ & 1 \end{pmatrix} g\right)\right) - \chi_{[T,\infty)}\left(H\left(w\begin{pmatrix} e^r & \\ & 1 \end{pmatrix} g\right)\right)\right) dr \\
&= \mathrm{vol}(\mathbf{Q}^\times\backslash\mathbf{A}_\mathbf{Q}^1) \int_{-\infty}^{\infty} 1 - \chi_{[T,\infty)}(r + H(g)) - \chi_{[T,\infty)}(H(wn_g) - H(g) - r))\, dr \\
&= \mathrm{vol}(\mathbf{Q}^\times\backslash\mathbf{A}_\mathbf{Q}^1) \int_{-\infty}^{\infty} 1 - \chi_{[T,\infty)}(r) - \chi_{[T,\infty)}(H(wn_g) - r))\, dr \\
&= \mathrm{vol}(\mathbf{Q}^\times\backslash\mathbf{A}_\mathbf{Q}^1)(2T - H(wg) - H(g)),
\end{aligned}$$

since the function being integrated is the characteristic function of the interval $[H(wn_g) - T, T]$ and $H(wn_g) \leq 0$. So we have essentially proved

Lemma 3.3.9. *The hyperbolic orbital integral for the test function introduced in Lemma 3.3.1 is*

$$\int_{\overline{G}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})} \sum_{\gamma \text{ hyp.}} \phi(g^{-1}\gamma g) = -\sum_{\alpha\in\mathbf{Q}^\times-\{1\}} \int_{\overline{M}(\mathbf{A_Q})\backslash\overline{G}(\mathbf{A_Q})} \phi\left(g^{-1}\begin{pmatrix}\alpha & \\ & 1\end{pmatrix} g\right)(H(wg)+H(g))\,dg.$$

Proof. Since the hyperbolic conjugacy classes in $PGL_2(\mathbf{Q})$ are precisely those represented by some

$$\gamma_0 = \begin{pmatrix}\alpha & \\ & 1\end{pmatrix}$$

for $\alpha \neq 1$, our computations above tell us that the hyperbolic orbital integral is

$$\sum_{\alpha\in\mathbf{Q}^\times-\{1\}} \mathrm{vol}(\mathbf{Q}^\times\backslash\mathbf{A}_\mathbf{Q}^1) \int_{\overline{M}(\mathbf{A_Q})\backslash\overline{G}(\mathbf{A_Q})} \phi\left(g^{-1}\begin{pmatrix}\alpha & \\ & 1\end{pmatrix} g\right)(2T - H(wg) - H(g))\,dg.$$

But $\mathrm{vol}(\mathbf{Q}^\times\backslash\mathbf{A}_\mathbf{Q}^1) = 1$ (as computed in [Tat1950]), and the $2T$ term integrates to 0 thanks to Equation (3.4). So if we take $T \to \infty$, in the limit we get the orbital integral we are interested in. □

Now we specialize Lemma 3.3.9 more explicitly.

Proposition 3.3.10. *The hyperbolic orbital integral for the specific test function ϕ from Lemma 3.3.1 is actually*

$$-n^{1-\frac{k}{2}} \sum_{\substack{d|n \\ d<\sqrt{n}}} d^{k-1}.$$

Proof. For $\alpha \in \mathbf{Q}^\times - \{1\}$,

$$\phi\left(g^{-1}\begin{pmatrix}\alpha & \\ & 1\end{pmatrix} g\right)$$

is nonzero for some g if and only if $\begin{pmatrix}\alpha & \\ & 1\end{pmatrix}$ has a representative in $G(\mathbf{Q})$ with integer coefficients and determinant n. So we can rewrite the orbital integral as

$$-\sum_{\substack{d|n \\ d<\sqrt{n}}} \int_{\overline{M}(\mathbf{A_Q})\backslash\overline{G}(\mathbf{A_Q})} \phi\left(g^{-1}\begin{pmatrix}n/d & \\ & d\end{pmatrix} g\right)(H(wg)+H(g))\,dg.$$

The restriction that $d < \sqrt{n}$ (and thus the divisor sum is not quite $\sigma_{k-1}(n)$) comes from the restriction that $\alpha \neq 1$. We already computed the term where $\alpha = 1$, and (as expected) saw that it vanishes whenever n is not a perfect square.

Again, the strategy is to split this into archimedean and nonarchimedean orbital integrals. In particular, if $\gamma \in \overline{M}(\mathbf{Q})$ with eigenvalues γ_1, γ_2, then

$$\int_{\overline{M}(\mathbf{A_Q})\backslash\overline{G}(\mathbf{A_Q})} \phi(g^{-1}\gamma g)(H(wg)+H(g))\,dg$$

$$= \int_{\overline{M}(\mathbf{A_Q})\backslash\overline{G}(\mathbf{A_Q})} \phi\left(g^{-1}\gamma g\right)\left(H(wg_\infty)+H(g_\infty)+H(wg_{\mathrm{fin}})+H(g_{\mathrm{fin}})\right) dg$$
$$= \left(\int_{\overline{M}(\mathbf{R})\backslash\overline{G}(\mathbf{R})} \phi_\infty(g^{-1}\gamma g)(H(wg)+H(g))\, dg\right)\left(\int_{\overline{M}(\mathbf{A}_{\mathrm{fin}})\backslash\overline{G}(\mathbf{A}_{\mathrm{fin}})} \phi_{\mathrm{fin}}(g^{-1}\gamma g))\, dg\right)$$
$$+ \left(\int_{\overline{M}(\mathbf{R})\backslash\overline{G}(\mathbf{R})} \phi_\infty(g^{-1}\gamma g)\, dg\right)\left(\int_{\overline{M}(\mathbf{A}_{\mathrm{fin}})\backslash\overline{G}(\mathbf{A}_{\mathrm{fin}})} \phi_{\mathrm{fin}}(g^{-1}\gamma g))(H(wg)+H(g))\, dg\right).$$

Luckily, the second term here vanishes thanks to Equation (3.4), and we can focus our attention instead on the archimedean weighted orbital integral and the nonarchimedean non-weighted orbital integrals.

One computes (using the usual explicit description of the discrete series and the fact that γ is diagonal with entries $\gamma_1, \gamma_2 \in \mathbf{Z}$ with $\gamma_1\gamma_2 = n$) without difficulty that the archimedean weighted orbital integral is

$$\int_{\overline{M}(\mathbf{R})\backslash\overline{G}(\mathbf{R})} \phi_\infty(g^{-1}\gamma g)(H(wg)+H(g))\, dg = n^{1-\frac{k}{2}}\frac{\gamma_2^{k-1}}{\gamma_1-\gamma_2}.$$

As for the non-archimedean non-weighted orbital integral, we have (thanks to the Iwasawa decomposition for $G(\mathbf{Q}_p)$ and the fact that $K_p = GL_2(\mathbf{Z}_p)$ is supposed to have measure 1 and ϕ_p is invariant on both sides by K_p)

$$\begin{aligned}\int_{\overline{M}(\mathbf{Q}_p)\backslash\overline{G}(\mathbf{Q}_p)} \phi_p(g^{-1}\gamma g)\, dg &= \int_{\overline{N}(\mathbf{Q}_p)} \phi_p(n^{-1}\gamma n)\, dn \\ &= \int_{\mathbf{Q}_p} f_p\left(\begin{pmatrix}\gamma_1 & t(\gamma_1-\gamma_2)\\ & \gamma_2\end{pmatrix}\right) dt \\ &= \mathrm{vol}\left(\frac{1}{\gamma_1-\gamma_2}\mathbf{Z}_p\right) \\ &= \frac{1}{|\gamma_1-\gamma_2|_p},\end{aligned}$$

where the last step is because γ_1 and γ_2 are supposed to be integers such that $\gamma_1\gamma_2 = n$, so f_p is 1 if and only if $t(\gamma_1-\gamma_2) \in \mathbf{Z}_p$.

So we conclude that the full orbital integral is

$$-n^{1-\frac{k}{2}} \sum_{\substack{d|n\\ d<\sqrt{n}}} \frac{d^{k-1}}{\gamma_1-\gamma_2} \prod_{p<\infty} \frac{1}{|\gamma_1-\gamma_2|_p} = -n^{1-\frac{k}{2}} \sum_{\substack{d|n\\ d<\sqrt{n}}} d^{k-1}$$

by the product formula. □

3.3.5 | The unipotent orbital integral

The final type of conjugacy class we have yet to account for is the (non-identity) unipotent ones. These are matrices conjugate over $\mathbf{Q}$ to something in $\overline{N}(\mathbf{Q})$ not equal to the identity.

Definition 3.3.11. Let $F : \mathbf{A_Q} \to \mathbf{C}$ be given by

$$F(t) = \int_K \phi\left(k^{-1}\begin{pmatrix} 1 & t \\ & 1 \end{pmatrix} k\right) dk.$$

The corresponding zeta function is

$$Z_F(s) = \int_{\mathbf{A_Q^\times}} F\left(\begin{pmatrix} 1 & t \\ & 1 \end{pmatrix}\right) |t|^s \, d^\times t,$$

which by the basic theory has a pole at $s = 1$. Let

$$f.p._{s=1} Z_F$$

be the constant term of the Laurent series development of Z_F.

Even though the unipotent class manifestly has nontrivial intersection with $\overline{P}(\mathbf{Q})$, there is no need for truncation to do the computation in this case. Everything in sight is actually already absolutely convergent, and truncation would do nothing to the kernel anyway.

Proposition 3.3.12. *The unipotent orbital integral is precisely*

$$\int_{\overline{G}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})} \sum_{\gamma \text{ uni.}} \tilde{\Lambda}^T_{\textit{unipotent}} K_\phi(g,g)\, dg = f.p._{s=1} Z_F.$$

Proof. For rational numbers $t_1, t_2 \neq 0$, one checks directly that

$$\begin{pmatrix} 1 & t_1 \\ & 1 \end{pmatrix}, \begin{pmatrix} 1 & t_2 \\ & 1 \end{pmatrix}$$

are always conjugate to each other by an element of $P(\mathbf{Q})$. So there is in fact just one conjugacy class involved here, and in fact conjugating an element of $N(\mathbf{Q})$ by some $g \in G(\mathbf{Q})$ lands in $N(\mathbf{Q})$ if and only if $g \in P(\mathbf{Q})$. From this we may conclude that the unipotent orbital integral is

$$\sum_{\xi \in P(\mathbf{Q})\backslash G(\mathbf{Q})} \sum_{n \in N(\mathbf{Q}) - \{1\}} \phi(g^{-1}\xi^{-1} n \xi g).$$

When integrated over $\overline{G}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})$, this becomes

$$\int_{\overline{P}(\mathbf{Q})\backslash\overline{G}(\mathbf{A_Q})} \sum_{n \in N(\mathbf{Q}) - \{1\}} \phi(g^{-1} n g)\, dg,$$

and applying the Iwasawa decomposition plus the fact that the integrand is not sensitive to translating g

by anything in $N(\mathbf{A}_{\mathbf{Q}})$, we see that the orbital integral equals

$$\begin{aligned}
&\int_{\mathbf{Q}^\times\backslash\mathbf{A}_{\mathbf{Q}}^\times}\int_K \sum_{t\in\mathbf{Q}^\times}\phi\left(k^{-1}\begin{pmatrix}a^{-1} & \\ & 1\end{pmatrix}\begin{pmatrix}1 & t\\ & 1\end{pmatrix}\begin{pmatrix}a & \\ & 1\end{pmatrix}\right)|a|^{-1}\,dkd^\times a\\
&\quad=\int_{\mathbf{Q}^\times\backslash\mathbf{A}_{\mathbf{Q}}^\times}\sum_{t\in\mathbf{Q}^\times}F\left(at\right)|a|\,d^\times a\\
&\quad=\int_{\mathbf{Q}^\times\backslash\mathbf{A}_{\mathbf{Q}}^\times}\left[\left(\sum_{t\in\mathbf{Q}^\times}\widehat{F}\left(\frac{t}{a}\right)\right)+\widehat{F}(0)-F(0)|a|\right]d^\times a\\
&\quad=\int_{\mathbf{Q}^\times\backslash\mathbf{A}_{\mathbf{Q}}^\times}\left[\left(\sum_{t\in\mathbf{Q}^\times}\widehat{F}\left(\frac{t}{a}\right)\right)-F(0)|a|\right]d^\times a\\
&\quad=f.p._{s=1}Z_F,
\end{aligned}$$

as desired. Here we have used Poisson summation for F and the fact that $\widehat{F}(0)=0$ (by Equation (3.3)). □

It remains to compute the constant term of this zeta integral. For $v\in M_{\mathbf{Q}}$, let

$$F_v(t):=\int_{K_v}\phi_v\left(k^{-1}\begin{pmatrix}1 & t\\ & 1\end{pmatrix}k\right)dk,$$

so that $F(t)=\prod_v F_v(t)$ and

$$Z_F=\prod_v Z_{F_v}$$

as usual. So we just need to understand each of the local zeta integrals well enough. First of all,

$$Z_{F_\infty}(1)=\int_{\mathbf{R}^\times}\phi_\infty\left(\begin{pmatrix}1 & t\\ & 1\end{pmatrix}\right)|t|\,d^\times t=Z_{F_\infty}(1)=\int_{\mathbf{R}}\phi_\infty\left(\begin{pmatrix}1 & t\\ & 1\end{pmatrix}\right)dt=0$$

thanks to the construction of ϕ_∞ (see Equation (3.3)). It is also possible to painfully compute

$$Z'_{F_\infty}(1)=-\frac{1}{2}$$

using the explicit description of the discrete series representation and the matrix coefficient ϕ_∞.

Now we consider the nonarchimedean zeta integrals. For $p\nmid n$, we have

$$\begin{aligned}
F_p(t)&=\int_{K_p}\phi_p\left(k^{-1}\begin{pmatrix}1 & t\\ & 1\end{pmatrix}k\right)dk\\
&=\chi_{\mathbf{Z}_p},
\end{aligned}$$

by definition of f_p. Therefore,

$$Z_{F_p}(s)=\int_{\mathbf{Z}_p}|t|^s\,d^\times t$$

which is the same as the nonarchimedean local factor in [Tat1950], namely

$$Z_{F_p}(s) = \frac{1}{1-p^{-s}}$$

On the other hand, if $p|n$, then F_p and Z_p are identically zero if $v_p(n)$ is odd. Otherwise, ϕ_p just checks whether $\sqrt{n}t \in \mathbf{Z}_p$, hence

$$Z_{F_p}(s) = \int_{n^{-1/2}\mathbf{Z}_p} |t|^s \, d^\times t = p^{-v_p(n)/2} \int_{\mathbf{Z}_p} |t|^s \, d^\times t = p^{-v_p(n)/2} \frac{1}{1-p^{-s}}$$

so the finite part of Z_F is

$$\frac{1}{\sqrt{n}} \zeta_{\mathbf{Q}}(s),$$

which is meromorphic with a simple pole of residue $n^{-1/2}$ at $s = 0$. Since Z_{F_∞} has a simple root at $s = 1$, it follows that

$$f.p._{s=1} Z_F = \frac{1}{\sqrt{n}} \cdot Z'_{F_\infty}(1) = -\frac{1}{2\sqrt{n}}.$$

So we have proved

Proposition 3.3.13. *The unipotent orbital integral vanishes if n is not a perfect square, and otherwise is equal to*

$$-\frac{1}{2\sqrt{n}} = -\frac{1}{2} n^{1-\frac{k}{2}} (\sqrt{n})^{k-1}.$$

3.3.6 | The final statement

We conclude by putting together all the work of the previous sections into the final formula for the trace of the Hecke operator T_n.

Theorem 3.3.14 (Eichler–Selberg trace formula)**.** *Let $n \geq 1$ and $k \geq 4$. Then*

$$\begin{aligned} \operatorname{Tr} T_n|_{S_k(\Gamma(1),\mathbf{C})} = &-\frac{1}{2} \sum_{d|n} \min\left(d, \frac{n}{d}\right)^{k-1} && \text{(unipotent and hyperbolic)} \\ &- \frac{1}{2} \sum_{t^2<4n} \frac{\rho^{k-1} - \overline{\rho}^{k-1}}{\rho - \overline{\rho}} H(t^2 - 4n) && \text{(elliptic and identity)} \end{aligned}$$

where ρ denotes the imaginary quadratic irrational with trace t and norm n.

Proof. Apply Lemma 3.3.1, Proposition 3.3.2, Corollary 3.3.6, Proposition 3.3.10, and Proposition 3.3.13, and divide by $n^{1-\frac{k}{2}}$. Note that the convention

$$H(0) = -\frac{1}{12}$$

allows us to naturally combine the elliptic and identity terms by including the possibility where $t^2 = 4n$ (which is excluded from the elliptic case at first since the polynomial $X^2 - tX + n$ would then have two

rational roots), since that possibility happens exactly when n is a perfect square. Of course this way of stating the trace formula is made less attractive by the fact that ρ is rational, so the quantity $\frac{\rho^{k-1}-\overline{\rho}^{k-1}}{\rho-\overline{\rho}}$ must be interpreted as a limit as the imaginary part goes to zero. □

Chapter 4

Applications to arithmetic statistics

"I cannot do't without [computers]."

Clown, Shakespeare's *A Winter's Tale*

4.1 | The archimedean case: Weyl's law, prime geodesic theorems, and real quadratic fields

This section concerns the application of the trace formula as stated in Theorem 3.2.11 to more concrete problems.

One application is to further pinning down the asymptotic growth of the Laplace eigenvalues of Maass cusp forms, say for $\Gamma = SL_2(\mathbf{Z})$. This proof is taken from [Mar2012, Proposition 10], except I have observed that the same proof works with the appropriate modifications in the finite-volume case (as opposed to the compact case).

Theorem 4.1.1 (Weyl's law for finite-volume hyperbolic surfaces)**.** *Let $\{\lambda_i = s_i(s_i - 1)\}$ be the Laplace–Beltrami eigenvalues of Maass cusp forms of weight 0 for $\Gamma = SL_2(\mathbf{Z})$, arranged in increasing order. Then*

$$\#\{i : s_i \leq T\} + \frac{1}{4\pi}\int_{-T}^{T} -\frac{\varphi'}{\varphi}\left(\frac{1}{2} + it\right) dt \sim \frac{\mu(\Gamma\backslash\mathbf{H})}{4\pi}T^2$$

as $T \to \infty$.

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

$$\sum_n e^{-\beta r_n^2} + \frac{1}{4\pi}\int_{-\infty}^{\infty} -\frac{\varphi}{\varphi}\left(\frac{1}{2} + ir\right) e^{-\beta r^2}\, dr.$$

If we can understand the asymptotic behavior of this quantity as $\beta \to 0$, then we can expect to understand the s_i better (which is what we will do via the standard Tauberian argument).

The Fourier transform of h (using the nonstandard normalization convention we have been using

thus far) is

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

Plugging this into the trace formula (Theorem 3.2.11), we get

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

plus a bounded contribution from the elliptic terms, plus $-\frac{\log 2}{2\sqrt{\pi\beta}}$, plus the parabolic contribution, which is

$$\begin{aligned} -\frac{1}{2\pi} \int_{-\infty}^\infty e^{-\beta t^2} \frac{\Gamma'}{\Gamma}(1+it)\, dt &= O(1) - \frac{1}{\pi} \int_0^\infty e^{-\beta t^2} \log t\, dt \\ &= O(1) + \frac{-\gamma_{\text{Euler}} - \log 4\beta}{4\sqrt{\beta\pi}}. \end{aligned}$$

As $\beta \to 0$, the negative exponential in the hyperbolic term dominates it: the norms of hyperbolic conjugacy classes are all at least 2 so the denominators are bounded below, and the exponential term clearly dominates the $\log N(\gamma_0)$ since $N(\gamma) \geq N(\gamma_0)$ as well as the $\beta^{-1/2}$ (because it is negative exponential in $1/\beta$). So the identity term is the only one that makes a contribution. For that term, we use the estimate $\tanh(\pi r) = 1 + O(e^{-2\pi r})$. The upper bound on the error integrates to

$$\ll \int_0^\infty r e^{-\beta r^2 - 2\pi r}\, dr \ll 1.$$

The rest is

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

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

$$\sum_n e^{-\beta r_n^2} + \frac{1}{4\pi} \int_\infty^\infty \frac{-\varphi}{\varphi}\left(\frac{1}{2} + ir\right) e^{-\beta r^2}\, dr = \frac{\mu(\Gamma\backslash\mathbf{H})}{4\pi}\beta^{-1} + \frac{-\gamma_{\text{Euler}}}{2\sqrt{\beta\pi}} + \frac{-\log 4\beta}{2\sqrt{\beta\pi}} + O(1)$$

as $\beta \to 0$. The right hand side is dominated by the first term. This implies the desired asymptotic formula for the λ_i by Karamata's Tauberian theorem [Kar1931]. □

This was an application of an estimate of the geometric side to gain fine control over the spectral side. Indeed, the previous bounds we had from the basic spectral theory (either Bessel's inequality in the compact case or otherwise Lemma 3.2.12) were nowhere near as strong as this.

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

Definition 4.1.2 (Prime geodesics)**.** Let X be a Riemannian manifold. A *prime geodesic* on X is a closed

geodesic that traces out its image exactly once.

Remark 4.1.3. On the other hand, if $\gamma(t)$ is a prime geodesic on X, then X is also equipped with the time-reversal of γ, namely $t \mapsto \gamma(-t)$. For our purposes, these count as different prime geodesics even though they trace out the same image.

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

Moreover, we have a bijection

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

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

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

and the length of the underlying prime geodesic is $\log N(\gamma_0)$, where γ_0 is a generator of the centralizer of γ in Γ. This provides an opportunity to apply the trace formula to study these quantities. The following prime geodesic theorem for finite-volume hyperbolic surfaces appears (with this proof) in Sarnak's thesis [Sar1980], though Sarnak told me that the result with this error term was known to Selberg.

Theorem 4.1.4 (Selberg, 1956)**.** *Suppose* $\Gamma = SL_2(\mathbf{Z})$[1]*. Then*

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

as $T \to \infty$.

This proof is reproduced from [Sar1980], and it is slightly different (in its choice of test function) from the other proofs in the more readily-available literature.

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

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

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

[1]The result is true for arbitrary discrete subgroups of $SL_2(\mathbf{R})$ such that $\mu(\Gamma\backslash\mathbf{H}) < \infty$, but we can only prove it in this case because we cut corners and only fully developed the theory of the continuous spectrum in the case $\Gamma = SL_2(\mathbf{Z})$. Of course the proof here works perfectly fine also if Γ is such that $\Gamma\backslash\mathbf{H}$ is compact.

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

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

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

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

One checks that $h(x)$ satisfied the required decay condition Equation (3.2) (this is why we have to use a tent function as opposed to a characteristic function). First, we estimate the identity term. Since $\tanh(\pi r) \leq 1$ and

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

we have

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

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

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

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

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

By the same integration by parts argument, combined with the classical fact that

$$\int_{-x}^{x} \frac{\Gamma'}{\Gamma}(1+it)\,dt \ll x^2$$

and Lemma 3.2.12 (or Theorem 4.1.1 if we feel like using the trace formula twice) which says that

$$\int_{-x}^{x} \frac{\varphi'}{\varphi}\left(\frac{1}{2}+ir\right)\,dr \ll x^2,$$

both of the extra terms from the continuous spectrum are absorbed in the $O(T^{-1}+T^{-1}\log\epsilon^{-1})$ error.

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

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

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

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

where τ ranges over the lengths, with multiplicity, of closed geodesics on $\Gamma\backslash\mathbf{H}$, and τ_0 is the length of the underlying prime geodesic. The spectral side is (after moving the term that is absorbed into the error on the geometric side anyway)

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

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

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

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

where this estimate is obtained using the fact that $\#\{n:r_n<r\}\ll_\Gamma r^2$ (Theorem 4.1.1) and the same technique we used to estimate the identity term. So the terms where $\lambda_n\geq 1/4$ are absorbed into the

$O(T + T^{-1}\log(\epsilon^{-1})$ error. Writing $r_n = it_n$ for the finitely many n with $\lambda_n < 1/4$, the analysis of the spectral side is now reduced to[2]

$$\begin{aligned}
-4\sum_{\substack{n\geq 0\\ \lambda_n<\frac{1}{4}}}\frac{\sin^2(Tit_n/2)}{Tt_n^2}\hat{\psi}(\epsilon it_n) &= -4\sum_{\substack{n\geq 0\\ \lambda_n<\frac{1}{4}}}\frac{\sin^2(Tit_n/2)}{Tt_n^2}\int_{\mathbf{R}}\psi(r)e^{-\epsilon t_n r}\,dr\\
&= -4\sum_{\substack{n\geq 0\\ \lambda_n<\frac{1}{4}}}\frac{\sin^2(Tit_n/2)}{Tt_n^2}\left[\|\psi\|_{L^1}+\int_{-1}^{1}\psi(r)(e^{-\epsilon t_n r}-1)\,dr\right]\\
&= -4\sum_{\substack{n\geq 0\\ \lambda_n<\frac{1}{4}}}\frac{\sin^2(Tit_n/2)}{Tt_n^2}\left[1+O(\epsilon)\right]\\
&= \sum_{\substack{n\geq 0\\ \lambda_n<\frac{1}{4}}}\frac{e^{-Tt_n}+e^{Tt_n}-2}{Tt_n^2}\left[1+O(\epsilon)\right]\\
&= \sum_{\substack{n\geq 0\\ \lambda_n<\frac{1}{4}}}\left[\frac{e^{Tt_n}}{Tt_n^2}+O\left(\frac{1}{T}\right)\right]\left[1+O(\epsilon)\right]\\
&= \sum_{\substack{n\geq 0\\ \lambda_n<\frac{1}{4}}}\frac{e^{Tt_n}}{Tt_n^2}+O(\epsilon e^{T/2}+T^{-1}).
\end{aligned}$$

So the trace formula reads

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

Using the trivial bound[3]

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

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

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

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

[3]See [Hej1976, Proposition 2.5]. The point is that every hyperbolic conjugacy class has a representative γ whose underlying geodesic on $\mathbf{H}$ meets the canonical fundamental domain $\mathcal{F}[\Gamma\backslash\mathbf{H}]$, and we know that $d_{\mathbf{H}}(z,\gamma z)=\log N(\gamma)$ for $z\in\gamma$. The conjugacy classes of log-norm at most x therefore all have the property that they have a representative γ such that $d_{\mathbf{H}}(z_0,\gamma\mathcal{F})\leq x+\mathrm{diam}\mathcal{F}$ where z_0 is some point in $\mathcal{F}$ fixed beforehand. In other words, $\gamma\mathcal{F}\cap B_{x+\mathrm{diam}\mathcal{F}}(z_0)\neq\emptyset$, and thus $\gamma\mathcal{F}\subset B_{x+2\mathrm{diam}\mathcal{F}}(z_0)$. The number of $\gamma\in\Gamma$ that satisfy this last inequality (which we have shown is an upper bound for the number we are interested in) is (by covering a subset of $B_{x+2\mathrm{diam}\mathcal{F}}(z_0)$ with disjoint translates of $\mathcal{F}$ and looking at areas) at most $\mu(B_{x+2\mathrm{diam}\mathcal{F}}(z_0))/\mu(\mathcal{F})$, so the trivial bound follows from the fact that the area of a hyperbolic disc of radius r is asymptotic to πe^r as $r\to\infty$.

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

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

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

$$\begin{aligned}
\sum_{\tau\leq T} \frac{\tau_0}{e^{\tau/2}-e^{-\tau/2}} - \frac{\tau_0}{e^{\tau/2}} &\leq \sum_{\tau\leq T} \frac{\tau}{e^{\tau/2}-e^{-\tau/2}} - \frac{\tau}{e^{\tau/2}} \\
&= \sum_{\tau\leq T} \frac{\tau}{e^{3\tau/2}-e^{\tau/2}} \\
&\ll_\Gamma \int_0^\infty x e^{-3x/2}\, d(\#\{\tau<x\}) \\
&\ll_\Gamma -\int_0^\infty e^x \frac{d}{dx}\left[xe^{-3x/2}\right] dx \\
&\ll \int_0^\infty (1+x)e^{-x/2}\, dx \\
&\ll 1
\end{aligned}$$

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\begin{aligned}\sum_{\tau_0\leq T}\tau_0\sum_{k=2}^{\infty}e^{-k\tau_0/2}&\ll_\Gamma\sum_{\tau_0\leq T}\tau_0e^{-\tau_0}\\&=\int_0^T xe^{-x}d(\#\{\tau_0<x\})\\&\ll_\Gamma T+\int_0^T e^x\frac{d}{dx}\left[xe^{-x}\right]dx\\&\ll_\Gamma T^2\end{aligned}$$

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

$$\sum_{\tau_0\leq T}\frac{\tau_0}{e^{\tau_0/2}}=\sum_{\substack{n\geq 0\\ \lambda_n<\frac{1}{4}}}\frac{e^{t_nT}}{t_n}+O\left(Te^{T/4}\right).\tag{4.1}$$

This lets us conclude via the usual technique of integration by parts. Let $F(T)$ be the quantity equal to both sides of Equation (4.1). Then the thing we are interested in is (using both the left and right hand sides of Equation (4.1) and the fact that F vanishes for small enough inputs)

$$\#\{\tau_0<T\}=\int_0^T x^{-1}e^{x/2}\,dF(x)$$

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

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

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

and the rest is

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

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

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

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

Remark 4.1.5. The error term in Sarnak's thesis is actually $O(T^{3/4}(\log T)^2)$, which originates from the fact that his error term after the Tauberian differentiation argument is $O(T^2e^{T/4})$ as compared to our $O(Te^{T/4})$. The most likely explanation for this is that there is a mistake in my own replication of his argument. Either way, the asymptotics are the same. In any event, the refinement of the final error term is mostly about the exponent on T, and is the subject of a lot of important and more recent work (see e.g. [LS1995] where the key point is to use the Weil bounds on Kloosterman sums to gain information about the cancellation in the error term) outside the scope of this paper.

Remark 4.1.6. By the uniformization theorem and Gauss–Bonnet, even when Γ is not a congruence

subgroup and is instead a cocompact group, the surfaces $\Gamma\backslash\mathbf{H}$ account for all compact Riemann surfaces of genus $g \geq 2$. So even this most basic form of a prime geodesic theorem is about something concrete and interesting. Note, too, that for positive curvature there are usually too many geodesics for this to make sense: take S^2 with the usual metric, for instance.

4.1.1 | Class numbers of real quadratic fields

The Maass cusp forms are supposed to have something to do with arithmetic, at least those of Laplace eigenvalue $1/4$ (which are supposed to be attached to even 2-dimensional Galois representations [Lan2004, BSV2006]). However, Sarnak [Sar1982] found another interesting application to arithmetic, via the prime geodesic theorem for finite-volume quotients of $\mathbf{H}$. He applied it to the uncompactified modular curve $Y(\Gamma)$, where $\Gamma = \Gamma(p)$. The lengths of geodesics on this modular curve are the same as regulators of quadratic fields with discriminant divisible by p. Since the argument is essentially the same conceptually but slightly less complicated, we will restrict our attention to $Y(1)$.

Theorem 4.1.7 (Sarnak, 1980)**.**

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

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

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

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

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

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

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

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

$$\sim_{x\to\infty} Li(x^2)$$

as a consequence of Theorem 4.1.4 (which we have remarked is also valid for finite-volume quotients). □

This result was a big step towards the long-open question of decoupling the regulator from the class number in the Gauss–Siegel asymptotic formula Theorem 1.2.2 for $\sum_{d<x} h(d)\log\epsilon_d$ [Sie1944] and its consequences and refinements (e.g. for $\Gamma=\Gamma(p)$) are elaborated on further in [Sar1982].

4.2 | ℓ-torsion in class groups of imaginary quadratic fields

For any fundamental discriminant $D<0$, let $h(D)$ be the class number of the imaginary quadratic field of discriminant D. The Cohen-Lenstra heuristics [CL1984] predict that $h(D)$ is not evenly distributed amongst congruence classes modulo ℓ, i.e. that the ℓ-power part of Cl_K is trivial with probability more than $1/\ell$. In fact, they predict (repeating Conjecture 1.2.4 to refresh our memory)

Conjecture 4.2.1 (Cohen–Lenstra, 1983)**.** Fix an odd prime ℓ, and let $\mathcal{F}_{\mathrm{im}}$ be the set of all imaginary quadratic fields. Then

$$\lim_{T\to\infty}\frac{\#\{K\in\mathcal{F}_{\mathrm{im}}:|\Delta_K|\le T,|\mathrm{Cl}_K|\not\equiv 0 \pmod{\ell}\}}{\#\{K\in\mathcal{F}_{\mathrm{im}}:|\Delta_K|\le T\}}=\prod_{i\ge 1}(1-\ell^{-i}).$$

The issue of the 2-power part of the class group is typically dealt with separately, thanks to genus theory for the 2-torsion (and recent work of A.D. Smith [Smi2017] for the entire 2^∞-torsion). This is why there is no harm in assuming $\ell>2$. As far as I know, it is unknown for $\ell>2$ whether the limit in Conjecture 4.2.1 even exists. One famous result towards this in the case $\ell=3$ is the main result of [DH1971].

Theorem 4.2.2 (Davenport–Heilbronn, 1971)**.**

$$\liminf_{T\to\infty}\frac{\#\{K\in\mathcal{F}_{\mathrm{im}}:|\Delta_K|\le T|\mathrm{Cl}_K|\not\equiv 0 \pmod{3}\}}{\#\{K\in\mathcal{F}_{\mathrm{im}}:|\Delta_K|\le T\}}\ge\frac{1}{2}.$$

See also [NH1988, Theorem 2], a similar positive-density result that further asks the discriminant to satisfy some very specific congruence conditions. Some other important work on the Davenport–Heilbronn theorem, though probably irrelevant to this situation, is Barghava–Shankar–Tsimerman's paper [BST2013] on the second-order terms in the actual Davenport–Heilbronn theorem, which is an asymptotic formula for averages of sizes of 3-power parts of class groups. Note that the lower bound on density resulting from Theorem 4.2.2 is still less than the expected

$$\prod_{i\ge 1}(1-3^{-i})\approx 56.01\%.$$

Setting our sights lower: can we show, analogously to Theorem 4.2.2, that for primes $\ell\ge 5$, a positive proportion of imaginary quadratic fields have class number not divisible by ℓ? As far as I know, this too is unkown.

This section is devoted the technique of using the trace formula (in the form of Theorem 3.3.14) to try and prove the infinitude of sets of imaginary quadratic fields K with $\mathrm{Cl}_K[\ell] = 1$, possibly with certain ramification conditions. This is much weaker than Conjecture 4.2.1, but still essentially represents the cutting edge in what is currently known (especially with ramification conditions). Without the ramification conditions, the infinitude result is already known [Har1974] using a method related to the trace formula applied in the quaternionic setting.

Theorem 4.2.3 (Hartung, 1974)**.** *Let $\ell > 2$ be a rational prime. Then there are infinitely many imaginary quadratic fields K such that $\#\mathrm{Cl}_K \not\equiv 0 \pmod{\ell}$.*

Though we do not do better than Theorem 4.2.3, the method presented here (and particularly the idea of using congruences between modular eigenforms of different weight in this context) is completely new, as far as we are aware.

Wiles [Wil2015], also using the trace formula for modular forms on quaternion algebras (directly and also in the guise of the Jacquet–Langlands correspondence [JL1970, GJ1979]), has also made substantial progress on the generalization to arbitrary ramification conditions:

Theorem 4.2.4 (Wiles, 2015)**.** *Let $\ell > 2$ be prime, and let $P_-, P_0, , P_+$ be disjoint finite sets of odd primes, with the property that P_- contains no prime $\equiv 1 \pmod{\ell}$, P_+ contains no prime $\equiv -1 \pmod{\ell}$, and P_0 contains no prime $\equiv 1 \pmod{\ell}$ and $\equiv -1 \pmod{4}$. Then there are infinitely many imaginary quadratic fields K such that $|\mathrm{Cl}_K| \not\equiv 0 \pmod{\ell}$ and L is ramified at each place in P_0, inert at each place in P_-, and split at each place in P_+.*

In this section, we present a new technique that (for now, in its most naïve form) manages to recover the special case of [Har1974] where $\ell \in \{5, 7, 11\}$. Our proof is conditional on Bunyakovsky's standard conjecture on prime values of polynomials, though it seems promising to fix this gap using the standard facts about squarefree values of polynomials.

The method of [Wil2015] is more precisely to combine the trace formula applied to Hecke operators acting on weight-2 modular forms on Shimura curves associated to indefinite quaternion algebras with additional congruence information about those modular forms (and hence about the trace) coming from the existence of Galois representations associated to those modular forms. Ours can be viewed analogously, though the automorphic information we use is strikingly different: we use classical holomorphic modular forms instead of quaternionic modular forms, and we apply the additional information of congruences between modular eigenforms of different weight induced by Eisenstein series (those which were used by [DS1974] to prove the existence of Galois representations associated to modular forms of weight 1).

Remark 4.2.5. The reason why Wiles' technique gives better results is that ours uses relations of the form $\sum_x m_x h_x \equiv 0 \pmod{\ell}$, which makes it tricky to deduce that at least one of the h_x is nonvanishing modulo ℓ. Wiles is able to get relations of the form $\sum_x m_x h_x \not\equiv 0$, which is why he can immediately deduce the existence of some x such that the class number h_x is not divisible by x. The cases that Theorem 4.2.4 misses are precisely where the relation Wiles has is actually $\sum_x m_x h_x \equiv 0$. So even if our technique using modular forms of different weight does not eventually prove to be useful, the fact that we have dealt with the technical exercise of deducing the infinitude result from the more difficult

information already means that Theorem 4.2.4 can be generalized to some of the exceptional cases, albeit only one at a time (a computation must be done for each desired ramification type) and still conditional on Bunyakovsky's conjecture.

This final section is organized as follows. In Section 4.2.1, we get the basic prerequisites involving Hurwitz class numbers out of the way. In Section 4.2.2, we carry out the promised proof of Theorem 4.2.3 in the case $\ell \in \{5, 7, 11\}$, and in Section 4.2.3 we very briefly propose an algorithm resulting from our proof for computing the class numbers of imaginary quadratic fields modulo ℓ.

4.2.1 | Hurwitz class numbers and the Selberg trace formula

In line with the Cohen–Lenstra $1/|\mathrm{Aut}|$ philosophy, a natural way to count quadratic forms is by weighting them by the reciprocal of the stabilizer under the action of $PSL_2(\mathbf{Z})$. We saw in Theorem 3.3.14 that the resulting weighted class numbers came up naturally in the explicit development of the Eichler–Selberg trace formula.

Remark 4.2.6. The reader who is educated in the ways of the force will recognize that we are looking at the $PSL_2(\mathbf{Z})$-action instead of the twisted $GL_2(\mathbf{Z})$-action (see [Woo2011]). So we unfortunately expect the class numbers that show up to be narrow class numbers rather than bona fide class numbers. In our situation, this makes no difference. The biggest reason is because we are looking at imaginary quadratic fields, which have no real embeddings: so the narrow class group is the same as the class group. Even if we were looking at real quadratic fields, we only care about whether the class number is divisible by ℓ, and ℓ is an odd prime (the narrow class number and the class number can only differ by a factor of 2), so it wouldn't make a difference then either. We use the old convention of using $SL_2(\mathbf{Z})$ instead of $GL_2(\mathbf{Z})$ in order to agree with the vast majority of the literature.

Definition 4.2.7. Let $D < 0$ be an integer with $-D \equiv 0, 3 \pmod 4$. The *Hurwitz–Kronecker class number* of the quadratic forms of discriminant D, denoted $H(D)$, is defined to be

$$H(D) = \sum_{f \in \mathcal{V}_{\mathrm{disc}=D}(\mathbf{Z})} \frac{1}{|\mathrm{Stab}_{PSL_2(\mathbf{Z})}(f)|},$$

where $\mathcal{V}_{\mathrm{disc}=D}(\mathbf{Z})$ denotes the set of binary quadratic forms over $\mathbf{Z}$ with discriminant D.

Note that these stabilizers are mostly trivial, because of the fact that $PSL_2(\mathbf{Z})$-stabilizer of a quadratic form $f(X, Y)$ of *negative* discriminant D is equal to the $PSL_2(\mathbf{Z})$-stabilizer in the upper half-plane of one of the roots of $f(X, 1)$. That stabilizer for most points in $PSL_2(\mathbf{Z})\backslash\mathbf{H} = X(1)$ is trivial. The only two exceptions are the equivalence classes of i (stabilizer of size 2) and $e^{2\pi i/3}$ (stabilizer of size 3). So $H(D)$ simply counts quadratic forms f of discriminant D, with multiplicity $1/2$ if f is a $\mathbf{Z}$-multiple of $X^2 + Y^2$, and multiplicity $1/3$ if f is a $\mathbf{Z}$-multiple of $X^2 + XY + Y^2$. The Hurwitz–Kronecker class number may not be an integer, but from this we see that it is guaranteed to be in $\mathbf{Z}[\frac{1}{6}]$. This means that it makes sense to talk about the quantities $H(N)$ modulo any prime $\ell \neq 2, 3$. This is one of the reasons why we assume $\ell \geq 5$.

Definition 4.2.8. An integer $D < 0$ is a *fundamental discriminant* if it is the discriminant of an imaginary quadratic field, i.e. if $-D = 4d$ where $d > 0$ is squarefree and $d \equiv 2, 1 \pmod 4$ or $-D$ is squarefree and congruent to $3 \pmod 4$. For an arbitrary $N < 0$ which is the discriminant of a binary quadratic form over $\mathbf{Z}$ (i.e. $-N \equiv 0, 3 \pmod 4$), the *fundamental part* of N is the largest (in absolute value) negative fundamental discriminant that divides it. In other words, if the squarefree part of $-N$ is congruent to $3 \pmod 4$, then the fundamental part of N is the squarefree part of N. Otherwise (in which case the squarefree part of $-N$ is congruent to 1 or $2 \pmod 4$), the fundamental part of N is four times the squarefree part. We adopt the nonstandard notation $\mathrm{f.p.}(N)$ for the fundamental part of N.

When D is a fundamental discriminant, another way to interpret the stabilizer of $f(X, Y) \in \mathcal{V}_{\mathrm{disc}=D}$ in $PSL_2(\mathbf{Z})$ is via the fact that it is isomorphic to $\{\pm 1\} \backslash \mathcal{O}^\times_{\mathbf{Q}(\sqrt{D})}$. Since $D < 0$, this group is trivial except for when $D = -4, -3$. Since it is always finite, it is moreover isomorphic to the group of roots of unity in $\mathbf{Q}(\sqrt{D})$. As a result, when $D < 0$ is a fundamental discriminant,

$$H(D) = 2\frac{h(D)}{|\mu_{\mathbf{Q}(\sqrt{D})}|},$$

where $\mu_{\mathbf{Q}(\sqrt{D})}$ denotes, as usual, the group of roots of unity in $\mathbf{Q}(\sqrt{D})$.

Even when N is not a fundamental discriminant, the Hurwitz–Kronecker class number has real meaning in terms of class numbers of real quadratic fields:

Lemma 4.2.9. *Suppose $N < 0$ with $-N \equiv 0, 3 \pmod 4$, and write*

$$N = \mathrm{f.p.}(N) \cdot f_N^2.$$

Then

$$H(N) = 2\frac{h(\mathrm{f.p.}(N))}{|\mu_{\mathbf{Q}(\sqrt{\mathrm{f.p.}(N)})}|} \sum_{d | f_N} \mu(d) \left(\frac{\mathrm{f.p.}(N)}{d}\right) \sigma_1(f_N/d),$$

where $\left(\frac{a}{d}\right)$ denotes the Kronecker symbol.

Proof. One begins with the identity

$$H(N) = \sum_{d | f_N} 2\frac{h(N/d^2)}{|\mu_{\mathcal{O}_{N/d^2}}|}, \tag{4.2}$$

where $\mathcal{O}_{N/d^2}$ is the quadratic order of discriminant $N/d^2 = \mathrm{f.p}(N) \cdot (f_N/d)^2$ and $h(N/d^2)$ is the class number of that order. The d term here comes from taking the primitive binary quadratic forms of discriminant $\mathrm{f.p.}(N)$ and multiplying by f_N/d. The number $|\mu_{\mathcal{O}_D}|$ is equal to 2 unless $D = -4, -3$ (the only new content here is that it equals 2 whenever $\mathcal{O}_D$ is a non-maximal order in $\mathbf{Q}(\sqrt{-1})$ or $\mathbf{Q}(\sqrt{-3})$).

As a straightforward consequence of the analytic class number formula, we further have for any discriminant $D < 0$, with $D = \mathrm{f.p.}(D) \cdot f_D^2$,

$$\frac{h(N)}{|\mu_{\mathcal{O}_N}|} = \frac{h(\mathrm{f.p.}(D))}{|\mu_{\mathcal{O}_{\mathbf{Q}(\sqrt{\mathrm{f.p.}(D)})}}|} f_D \prod_{p | f_D} \left(1 - \frac{\left(\frac{\mathrm{f.p.}(D)}{p}\right)}{p}\right),$$

so applying this to each term in Equation (4.2), using the fact that $\mathrm{f.p.}(N/d^2) = \mathrm{f.p.}(N)$, we have

$$\begin{aligned}
H(N) &= \sum_{d \mid f_N} 2\frac{h(N/d^2)}{|\mu_{\mathcal{O}_{N/d^2}}|} \\
&= \sum_{d \mid f_N} 2\frac{h(\mathrm{f.p.}(N/d^2))}{|\mu_{\mathcal{O}_{\mathbf{Q}(\sqrt{\mathrm{f.p.}(N/d^2)})}}|} f_{N/d^2} \prod_{p \mid f_{N/d^2}} \left(1 - \frac{\left(\frac{\mathrm{f.p.}(N/d^2)}{p}\right)}{p}\right) \\
&= 2\frac{h(\mathrm{f.p.}(N))}{|\mu_{\mathcal{O}_{\mathbf{Q}(\sqrt{\mathrm{f.p.}(N/d^2)})}}|} \sum_{d \mid f_N} \frac{f_N}{d} \prod_{p \mid \frac{f_N}{d}} \left(1 - \frac{\left(\frac{\mathrm{f.p.}(N)}{p}\right)}{p}\right) \\
&= 2\frac{h(\mathrm{f.p.}(N))}{|\mu_{\mathcal{O}_{\mathbf{Q}(\sqrt{\mathrm{f.p.}(N/d^2)})}}|} \sum_{d_1 \mid f_N} \frac{f_N}{d_1} \sum_{d_2 \mid \frac{f_N}{d_1}} \mu(d_2) \frac{\left(\frac{\mathrm{f.p.}(N)}{d_2}\right)}{d_2} \\
&= 2\frac{h(\mathrm{f.p.}(N))}{|\mu_{\mathcal{O}_{\mathbf{Q}(\sqrt{\mathrm{f.p.}(N/d^2)})}}|} \sum_{d_1 d_2 \mid f_N} \frac{f_N}{d_1 d_2} \mu(d_2) \left(\frac{\mathrm{f.p.}(N)}{d_2}\right) \\
&= 2\frac{h(\mathrm{f.p.}(N))}{|\mu_{\mathcal{O}_{\mathbf{Q}(\sqrt{\mathrm{f.p.}(N/d^2)})}}|} \sum_{d_2 \mid f_N} \mu(d_2) \left(\frac{\mathrm{f.p.}(N)}{d_2}\right) \sum_{d_1 \mid \frac{f_N}{d_2}} \frac{f_N}{d_1 d_2}
\end{aligned}$$

as desired. □

Corollary 4.2.10. *If $\ell \geq 5$ is a rational prime, and $H(N) \not\equiv 0 \pmod{\ell}$, then $h(\mathrm{f.p.}(N)) = |\mathrm{Cl}_{\mathbf{Q}(N)}| \not\equiv 0 \mod \ell$.*

Corollary 4.2.11. *If $\ell \in \{5, 7, 11\}$, and $d < 0$ a fundamental discriminant, then $H(d) \equiv 0 \pmod{\ell}$ if and only if $H(4d) \equiv 0 \pmod{\ell}$.*

Our strategy for recovering Theorem 4.2.3 will be to use the Eichler–Selberg trace formula to prove the existence of some N (such that we haven't yet proved that $h(\mathrm{f.p.}(N)) \not\equiv 0 \pmod{\ell}$) such that $H(N) \not\equiv 0 \pmod{\ell}$, and conclude from this some new imaginary quadratic field with class number not divisible by ℓ. Reproducing the statement from when we proved it in Chapter 3, we have

Theorem 4.2.12 (Eichler–Selberg trace formula)**.** *Let $k > 2$ be an even integer, and S_k the $\mathbf{C}$-vector space of cusp forms of weight k and level 1, equipped with the Hecke operators T_m for all $m \geq 1$. Then*

$$\mathrm{Tr} T_m|_{S_k} = -\frac{1}{2} \sum_{|t| \leq 2\sqrt{m}} P_k(t, m) H(t^2 - 4m) - \frac{1}{2} \sum_{d_1 d_2 = m} \min(d_1, d_2)^{k-1},$$

where $P_k(t, m)$ is defined to be

$$P_k(t, m) = \frac{\rho^{k-1} - \overline{\rho}^{k-1}}{\rho - \overline{\rho}},$$

where ρ is the quadratic algebraic number with norm m and trace t. Note that $P_k(t, m)$ is a polynomial in t and m, and is equal to the coefficient of x^{k-2} in the formal power series $(1 - tx + mx^2)^{-1} \in \mathbf{Z}[m, t][\![x]\!]$.

In the next section, we will exploit Theorem 4.2.12 together with congruences between modular forms (and thus, in the right circumstances, traces of Hecke operators) of different weight in order to get nontrivial relations of the form

$$\sum_{0 \le t \le 2\sqrt{m}} A(t,m) H(t^2 - 4m) \equiv 0 \pmod{\ell}$$

where $A(t,m) \in \mathbf{Z}[\frac{1}{2}][t,m]$. From the previous discussion, this makes sense when $\ell \geq 5$ since $H(t^2 - 4m) \in \mathbf{Z}[\frac{1}{6}]$.

4.2.2 | Congruences and proof of infinitude

Our result is based on the congruences mod ℓ between cusp eigenforms of weight k and weight $k + \ell - 1$ induced by multiplication by the Eisenstein series $E_{\ell-1}$. This Eisenstein series is very useful in the theory of p-adic variation of modular forms (where $p = \ell$), because its q-expansion is congruent to 1:

Lemma 4.2.13. *If $(\ell-1)|k$, then $E_k \equiv 0 \pmod{p}$, i.e. the q-expansion of $E_k = 1 + \sum_{n=1}^{\infty} a_n q^n \in \mathbf{Q}_\ell[\![q]\!]$ as the property that $v_\ell(a_n) \geq 1$ for all $n \geq 1$.*

Proof. This is a consequence of the standard explicit form for the coefficients of Eisenstein series, combined with the Clausen–von Staudt theorem [Sta1840] on Bernoulli numbers mod ℓ (see [Ser1973, 1.1(d)]). □

Other than the fact that we want to reduce elements of $\mathbf{Z}[\frac{1}{6}]$ modulo ℓ, another reason we want $\ell \geq 5$ is that we need $E_{\ell-1}$ to be a bona fide modular form.

Corollary 4.2.14. *If $f \in S_k$, then $f \cdot E_{\ell-1} \equiv f \pmod{\ell}$.*

This is a nice congruence between modular forms of different weight, but note that it is *not* true that if f is eigenform, then $f \cdot E_{\ell-1} \in S_{k+\ell-1}$ is an eigenform. That being said, thanks to a general lemma [DS1974, Lemme 6.11], it is still congruent mod ℓ to an eigenform in weight $k + \ell - 1$.

Lemma 4.2.15 (Deligne–Serre, 1968)**.** *Let M be a finite-rank free module over a DVR $\mathcal{O}$; denote by $\mathfrak{m}$ the maximal ideal of $\mathcal{O}$, k its residue field, K its fraction field. Let $\mathscr{T}$ be a set of endomorphisms of M which commute with each other. Let $f \in M/\mathfrak{m}M$ be a nonzero simultaneous eigenvector of the operators $T \in \mathscr{T}$, and let a_T be the corresponding eigenvalues. There then exists a DVR $\mathcal{O}' \supset \mathcal{O}$ with maximal ideal $\mathfrak{m}'$ such that $\mathfrak{m}' \cap \mathcal{O} = \mathfrak{m}$, and with field of fractions K' finite over K; and a nonzero element*

$$f' \in M' = M \otimes_{\mathcal{O}} \mathcal{O}'$$

which is a simultaneous eigenvector of the $T \in \mathscr{T}$ with eigenvalues $a'_T \equiv a_T \pmod{)m'}$.

Corollary 4.2.16. *Let f be a modular cusp eigenform of weight k. Then $f \cdot E_{\ell-1}$ is congruent to a cusp eigenform in weight $k + \ell - 1$.*

Proof. Here, M is the $\mathbf{Z}_{(\ell)}$-module of cusp forms of weight $k + \ell - 1$ over $\mathbf{Z}_{(\ell)}$, and $\mathscr{T}$ is the usual Hecke algebra or at least a set of generators for it. We do this instead of $\mathbf{Z}_\ell$ because we want the resulting cusp

form to have rational coefficients rather than ℓ-adic ones. The mod ℓ-reduction of f in $M/\ell M$ is a mod ℓ modular form of weight $k + (p-1)\mathbf{Z}$ (see [Ser1973, §1.2]), and thus lives in $M/\ell M$. Since f is an eigenform, Lemma 4.2.15 guarantees that it lifts to an eigenform of weight $k + \ell - 1$, whose coefficients are in some finite extension of $\mathbf{Q}$, as desired. □

Corollary 4.2.17. *If k and ℓ are such that*

$$\dim S_k = \dim S_{k+\ell-1},$$

then for all $m \geq 1$,

$$\mathrm{Tr} T_m|_{S_k} \equiv \mathrm{Tr} T_m|_{S_{k+\ell-1}} \pmod{\ell}.$$

Proof. Both spaces of cusp forms have bases of the same length consisting of eigenforms with coefficients in finite extensions of $\mathbf{Q}$, say

$$S_k = \bigoplus_{i=1}^{n} \mathbf{C} \cdot f_i, \qquad S_{k+\ell-1} = \bigoplus_{i=1}^{n} \mathbf{C} \cdot g_i.$$

From Corollary 4.2.16, the f_i's are each congruent to some g_i modulo a prime lying over ℓ. But the multiplicity of a mod ℓ eigenform can only increase when increasing the weight, since multiplication by $E_{\ell-1}$ induces an inclusion $M_k(\overline{\mathbf{F}}_\ell) \subset M_{k+\ell-1}(\overline{\mathbf{F}}_\ell)$. Hence, there is a bijection $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ such that $f_i \equiv g_{\sigma(i)} \pmod{\ell}$. This is just different language for a congruence of systems of Hecke eigenvalues

$$\lambda_T^{(f)} \equiv \lambda_T^{(g)} \pmod{\ell}$$

(or really modulo a prime lying over ℓ) for all Hecke operators T, which implies that

$$\mathrm{Tr} T_m|_{S_k} = \sum_{i=1}^{n} \lambda_{T_m}^{(f_i)} \equiv \sum_{i=1}^{n} \lambda_{T_m}^{(g_i)} = \mathrm{Tr} T_m|_{S_{k+\ell-1}} \pmod{\ell}.$$

This time the congruence really is modulo ℓ since the traces of these Hecke operators are guaranteed to be rational integers. □

Combining Corollary 4.2.17 with Theorem 4.2.12, we have

Corollary 4.2.18. *Let $k \geq 4$ be an even integer, and $\ell \geq 5$ a prime such that $\dim S_k = \dim S_{k+\ell-1}$. Then there is a function $G : \mathbf{F}_\ell \times \mathbf{F}_\ell \to \mathbf{F}_\ell$ such that for all $m \geq 1$,*

$$\sum_{0 \leq t \leq 2\sqrt{m}} G(t, m) H(t^2 - 4m) = 0 \in \mathbf{F}_\ell. \tag{4.3}$$

Remark 4.2.19. The fact that $\ell \geq 5$ is already used to justify inverting 2 in the reduction of the trace formula Theorem 4.2.12 mod ℓ, and in applying Corollary 4.2.17, which depends on the existence of the modular form $E_{\ell-1}$.

The condition that $\dim S_k = \dim S_{k+\ell-1}$ is what prevents Corollary 4.2.18 from being useful when $\ell \geq 13$, because of the roughly linear growth of the dimension with respect to the weight.

The function $G(\cdot,\cdot)$ is given by

$$G(t,m) = \begin{cases} \frac{P_k(t,m)-P_{k+\ell-1}(t,m)}{2}, & \text{if } t = 0 \\ P_k(t,m) - P_{k+\ell-1}(t,m), & \text{if } t \neq 0 \end{cases}.$$

For fixed ℓ, it is straightforward (on a computer) to write down G explicitly as a lookup table, or by writing down polynomials whose values coincide with those of $P_k(t,m) - P_{k+\ell-1}(t,m) \pmod{\ell}$. These tables and polynomials are written down in Section 4.2.4 for $k = 12$.

Remark 4.2.20. One might observe at first that Equation (4.3) gives information about the class numbers mod ℓ for each admissible value of k. However, this is not the case: one can check that Equation (4.3) is always a scalar multiple of the same equation if you vary k. This simplifies things, since it means that we can simply choose $k = 12$, which happens to work for all ℓ that have any admissible k at all: for $\dim S_{12} = \dim S_{16} = \dim S_{18} = \dim S_{22} = 1$. This is interesting because it means that the congruences we are looking at aren't just congruences between traces of Hecke operators, but actually congruences between q-expansions of cusp forms, and the trace formula has provided a link between those q-expansions and class numbers of imaginary quadratic fields. From now on, it is assumed that $k = 12$.

Remark 4.2.21. The original question that Professor Kisin asked me was whether the congruence between the traces would be evident directly from the trace formula. I found that when $\ell | m$, it is obvious (in fact the $G(t,m)$ are all zero). This is easy to read off of the formulas in Section 4.2.4 for $\ell \in \{5,7,11\}$, and in fact it is true in general (either by the general version of those formulas or by using the other description of the P_k in Theorem 4.2.12). I recently learned from Google that Koike [Koi1975] proved this in 1975.

However, when $m \not\equiv 0 \pmod{\ell}$, the coefficients are in general nonvanishing, which means that the class numbers conspire to make Corollary 4.2.18 true. It seems difficult to prove nontrivial identities like this in the other direction, since the coefficients are not very easy to work with.

Our strategy for proving Theorem 4.2.3 is simply to start with an $N < 0$ such that $H(N) \not\equiv 0 \pmod{\ell}$, with the property that $N = t^2 - 4m$ and $G(t,m) \neq 0$. Then the (t,m) term in Equation (4.3) is nonzero in $\mathbf{F}_\ell$, so we deduce that there is an $N' \neq N$ of the form $(t')^2 - 4m$ for some other t. One must also ensure that N' actually has fundamental part that we haven't yet deduced nonvanishing for (otherwise this approach could just be giving us many redundant facts), and that the method can actually be used inductively. That induction rests on

Lemma 4.2.22. *Let $(t_1,m_1),(t_2,m_2) \in \mathbf{Z} \times \mathbf{Z}$ with $t_1^2 - 4m_1 = t_2^2 - 4m^2 < 0$, and $m_i \not\equiv 0 \pmod{\ell}$.*

1. *If $\ell = 5$, then with the above hypotheses,*

$$G(t_2,m_2) = 0 \iff G(t_2,m_2) = 0.$$

2. *If $\ell \in \{7,11\}$, then with the above hypotheses plus $m_1 \equiv m_2 \pmod{\ell}$,*

$$G(t_2,m_2) = 0 \iff G(t_2,m_2) = 0.$$

3. *For any $\ell \geq 5$, and $0 \neq f \in \mathbf{F}_\ell$, we have*

$$G(ft, f^2 m) = 0 \iff G(t, m) = 0.$$

Proof. As far as I understand it, what happens in case (1) is a coincidence. It can be checked by looking directly at the values of G, listed in Table 4.1 in Section 4.2.4: one sees that if $m \not\equiv 0 \pmod 5$, $G(t, m) = 0$ implies that $4m - t^2 \in \{1, 4\} \subset \mathbf{F}_5$, while $G(t, m) \neq 0$ implies that $4m - t^2 \in \{0, 2, 3\} \subset \mathbf{F}_5$.

Case (2) is really true for all $\ell \geq 5$, and it is a consequence of the fact that only even powers of t show up in the polynomials $P_k(t, m) \in \mathbf{Z}[t, m]$. For these special values of ℓ, we wrote down these polynomials explicitly in Section 4.2.4, but also it is true in general because (from the alternate description of P_k in the statement of Theorem 4.2.12)

$$\begin{aligned} P_k(t, m) &= \frac{\rho^{k-1} - \overline{\rho}^{k-1}}{\rho - \overline{\rho}} \\ &= \sum_{i=0}^{k-2} \rho^{k-2-i} \overline{\rho}^i \\ &= \left(\mathrm{Tr} \sum_{i=0}^{\frac{k-2}{2}-1} \rho^{k-2-i} \overline{\rho}^i \right) + (\rho\overline{\rho})^{\frac{k-2}{2}} \\ &= \left(\mathrm{Tr} \sum_{i=0}^{\frac{k-2}{2}-1} m^i \rho^{k-2-2i} \right) + m^{\frac{k-2}{2}} \\ &= \left(\sum_{i=0}^{\frac{k-2}{2}-1} m^i \mathrm{Tr}(\rho^{k-2-2i}) \right) + m^{\frac{k-2}{2}}. \end{aligned}$$

So all the monomial terms with a nontrivial power of t come from the terms $\mathrm{Tr}(\rho^{k-2-2i})$. Using the fact that ρ is a root of $X^2 - tX + m$, one checks by induction that the fact that $k - 2 - 2i$ is even implies that as a polynomial in t, m, ρ with ρ-degree 1, each monomial term in ρ^{k-2-2i} has even combined (ρ, t)-degree. Since $\mathrm{Tr}(\rho) = t$, it follows that as a polynomial in m and t, every monomial in ρ^{k-2-2i} has an even power of t.

If $m_1 \equiv m_2 \pmod \ell$ and $t_1^2 - 4m_1 \equiv t_2^2 - 4m_2 \pmod \ell$, then $t_1 = \pm t_2$. But we just showed that $G(t, m) = G(\pm t, m)$ for all $t, m \in \mathbf{F}_\ell$. Note that (2) is much weaker than (1), as we needed $m_1 \equiv m_2 \pmod \ell$.

Now (3) is a consequence of the fact that $P_k(t, m)$ is homogeneous in the variables t^2, m. □

We will also need

Lemma 4.2.23. *Let $d \leq -7$ be a fundamental discriminant, and p an odd prime such that there exist $x, y \in \mathbf{Z}$ such that*

$$4p = x^2 - dy^2.$$

Then this representation of $4p$ as $x^2 - dy^2$ is unique up to changing the signs of x and y.

Proof. This is a variant of the usual standard things in algebraic number theory. If $d \equiv 0 \pmod 4$, then x must be even, so we can divide x and y by 2 to see that it suffices to show that representations $p = x^2 - dy^2$ are unique, where $d < 0$ is squarefree with $-d \equiv 1, 2 \pmod 4$ and is either -2 or ≤ -5. To do that, note that

$$\mathcal{O}_{\mathbf{Q}(\sqrt{d})} = \mathbf{Z}[\sqrt{d}],$$

so this is equivalent to asking, up to negation and conjugation, how many $\alpha \in \mathcal{O}_{\mathbf{Q}(\sqrt{d})}$ have norm equal to p. Any such α generates a prime ideal in $\mathcal{O}_{\mathbf{Q}(\sqrt{d})}$, by the multiplicativity of the norm, so by unique factorization of ideals in Dedekind domains, if α, β both have norm p, then either $(\alpha) = (\beta)$ or $(\alpha) = (\overline{\beta})$. Since conjugating α has the effect of changing the sign of y, we can assume that $(\alpha) = (\beta)$. By the assumption on d, namely that it is ≤ -5 or equal to -2, we have $\mathcal{O}^\times_{\mathbf{Q}(\sqrt{D})} = \{\pm 1\}$, which shows that indeed α is unique up to change of sign and conjugation.

In the case where $d \not\equiv 0 \pmod 4$, we have $-d \equiv 3 \pmod 4$ squarefree. We have

$$\mathcal{O}_{\mathbf{Q}(\sqrt{d})} = \mathbf{Z}\left[\frac{1+\sqrt{d}}{2}\right].$$

If $4p = x^2 - dy^2$, then x and y must have the same parity, so right away this is equivalent to finding, up to conjugation and negation, elements of $\mathcal{O}_{\mathbf{Q}(\sqrt{d})}$ of norm p. The bound $d \leq -7$ ensures that $\mathcal{O}^\times_{\mathbf{Q}(\sqrt{d})} = \{\pm 1\}$, so the same argument as before goes through. □

Finally, our proof of Theorem 4.2.3 for $\ell \in \{5, 7, 11\}$ is conditional on the standard conjecture [Bun1857],

Conjecture 4.2.24 (Bunyakovsky, 1857)**.** Suppose $f \in \mathbf{Z}[X]$ has positive leading coefficient, is irreducible, and has the property that there is no fixed prime p dividing $f(n)$ for all $n \in \mathbf{N}$. Then $f(n)$ is prime for infinitely many $n \in \mathbf{N}$.

Example 4.2.25. The actual case of Conjecture 4.2.24 we will use is $f(X) = d + X^2$, where d is a positive integer. This clearly satisfies the three conditions.

For technical reasons, we will need

Definition 4.2.26. An integer $d < 0$ is an *almost fundamental discriminant* if it is a fundamental discriminant or 4 times a fundamental discriminant.

Theorem 4.2.27. *Suppose Conjecture 4.2.24 is true. Let $\ell \in \{5, 7, 11\}$. Then there are infinitely many $K \in \mathcal{F}_{\mathrm{im}}$ such that $|\mathrm{Cl}_K| \not\equiv 0 \pmod \ell$.*

Proof. We proceed by induction. Suppose we have constructed, for $1 \leq i \leq N$, tuples of integers (d_i, m_i, t_i) with $m_i > 0$ and $d_i < 0$ almost fundamental with $\mathrm{f.p.}(d_i)$ pairwise distinct, satisfying the

properties

$$\begin{cases} H(d_i) \not\equiv 0 \pmod{\ell} \\ G(t_i, m_i) \neq 0 \in \mathbf{F}_\ell \\ d_i = t_i^2 - 4m_i \\ d_i \neq -4 \\ d_i \not\equiv 0 \pmod{3} \\ d_i \not\equiv 0 \pmod{\ell} \\ d_i :=\equiv -3, -8, -11, -12 \pmod{16} \\ \left(\frac{m_i}{\ell}\right) = -1 \end{cases} . \tag{4.4}$$

Let $\mathscr{P}_N$ be the set of odd primes dividing at least one d_i for $i < N$ but not d_N. We will show that there exists a tuple $(d_{N+1}, m_{N+1}, t_{N+1})$ satisfying Equation (4.4) such that d_{N+1} is not divisible by any primes in $\mathscr{P}_N$ and $\text{f.p.}(d_{N+1}) \neq \text{f.p.}(d_N)$. This statement is enough to deduce the inductive step (i.e. produce such a list of length $N+1$; here the only problem is that d_{N+1} needs to have fundamental part different from all the previous d_i, not just d_N): if d_{N+1} is divisible by some odd prime p not dividing d_N, then we are done because then p doesn't divide d_i for $i < N$ either (else it would be in $\mathscr{P}_N$), so $\text{f.p.}(d_{N+1}) \neq \text{f.p.}(d_i)$ for all $i \leq N$; otherwise, the set of odd primes dividing d_{N+1} is a proper (since $\text{f.p.}(d_{N+1}) \neq \text{f.p.}(d_N)$) subset of the set of odd primes dividing d_N, so $\text{f.p.}(d_{N+1})$ properly divides $\text{f.p.}(d_N)$, since the odd parts of the fundamental parts are squarefree. Repeating this argument sufficiently many times, the only problem is if d_{N+k} constructed at each step always has fundamental part properly dividing d_{N+k-1}. But in this case, we eventually reach the scenario where $\text{f.p.}(d_{N+k}) = -p, -4p$ where p is an odd prime (Equation (4.4) means that it must be divisible by at least one odd prime since it is not -4). Now $\text{f.p.}(d_{N+k+1})$ cannot be a proper divisor of this, so we conclude that it is distinct from d_i for all $1 \leq i \leq N+k$. The list

$$\{(d_i, m_i, t_i)\}_{i=1}^{N} \cup \{(d_{N+k+1}, m_{N+k+1}, t_{N+k+1})\}$$

is now a list of size $N+1$ where all the elements satisfy Equation (4.4) and with the $\text{f.p.}(d_i)$'s pairwise distinct. By induction, this means there are arbitrarily large lists of pairwise distinct negative fundamental discriminants with class number $\not\equiv 0 \pmod{\ell}$, using Corollary 4.2.10 to deduce $h(\text{f.p.}(d_i)) \not\equiv 0 \pmod{\ell}$.

To complete the proof, it suffices to do the base case (construct at least one (d, m, t) satisfying Equation (4.4)), and prove the above claim (which we showed suffices to prove the inductive step), namely that there exists a tuple $(d_{N+1}, m_{N+1}, t_{N+1})$ satisfying Equation (4.4) such that d_{N+1} is not divisible by any primes in $\mathscr{P}_N$ and $\text{f.p.}(d_{N+1}) \neq \text{f.p.}(d_N)$.

We will do the base case at the end (it is an explicit construction). The inductive step is where the trace formula comes in. By Corollary 4.2.18,

$$\sum_{0 \leq t \leq 2\sqrt{m}} G(t, m) H(t^2 - 4m) = 0 \in \mathbf{F}_\ell, \tag{4.5}$$

for any integer $m > 0$. For this to give us useful information to bootstrap off of, we need d_N to be one of

the $t^2 - 4m$ showing up here. First we show that conditional on Conjecture 4.2.24, that there exists a $t \in \mathbf{N}$ such that

1. $-d_N + t^2 = 4q$, where q is prime.
2. $4q$ is not a sum of two squares (i.e. $q \equiv 3 \pmod 4$)
3. $q \equiv m_N \pmod \ell$.
4. q is not a quadratic residue or zero modulo any $p \in \mathscr{P}_N$, mod, ℓ, or mod 3.

Since $-d_N \equiv 3, 8, 11, 12 \pmod{16}$, there is a congruence condition on t mod 4 (depending on d_N) which is sufficient for $-d_N + t^2 \equiv 12 \pmod{16}$. Similarly, we already have $d_N = t_N^2 - 4m_N$, so $-d_N \mod \ell$ differs from $4m_N \mod \ell$ by a quadratic residue or zero in $\mathbf{F}_\ell$, which means that $-d_N + t^2 \equiv 4m_N \pmod \ell$ is also a congruence condition on t mod ℓ. Furthermore, there exists a congruence class $\bar{t} \in \mathbf{F}_p$ such that $-d_N + t^2$ is not a quadratic residue or zero modulo $p \in \mathscr{P}_N$, since if there wasn't, then since $-d_N$ is invertible mod p (this is the reason for the less-than-ideal definition of $\mathscr{P}_N$), every element of $\mathbf{F}_p$ would be a quadratic residue (start at one quadratic residue and keep adding $-d_N$ to it). The same works at the primes 3 and ℓ, though at ℓ we are okay because $\left(\frac{m_N}{\ell}\right) = -1$ so this is already encoded in (3). So proving (1)-(4) amounts to showing that subject to the congruence condition

$$t \equiv a \mod 12\ell \prod_{p \in \mathscr{P}_N} p$$

that forces $-d_N + t^2 \equiv 12 \pmod{16}$ and $-d_N + t^2 \equiv 4m_N \pmod \ell$ and $-d_N + t^2$ to not be a quadratic residue or zero mod any $p \in \mathscr{P}_N$, t can be chosen so that $\frac{-d_N + t^2}{4}$ is prime. This is the content of Conjecture 4.2.24, applied to the polynomial

$$g(X) = \frac{1}{4}\left(-d_N + \left(\left(12\ell \prod_{p \in \mathscr{P}_N} p\right) X + a\right)^2\right) \in \mathbf{Z}[X],$$

which has integer coefficients because $-d_N + a^2 \equiv 12 \pmod{16}$ and is thus divisible by 4. The values of $g(x)$ for $x \in \mathbf{N}$ do not have a common prime divisor, because $\frac{-d_N + t^2}{4} \equiv 3 \pmod 4$, so that divisor can't be 2; since t has a condition mod $p \in \mathscr{P}_N$ such that $-d_N + t^2$ is not a quadratic residue or zero mod p, and same with ℓ, the only primes that could divide all the outputs are outside of $\{2, 3\ell\} \cup \mathscr{P}_N$. But t doesn't satisfy any conditions modulo those primes, so $-d_N + t^2$ can take on $\frac{p+1}{2} > 1$ distinct values modulo p, hence it is not divisible by p for some t.

Armed with $t \in \mathbf{N}$ satisfying (1)-(4), whose existence we have conditional on Conjecture 4.2.24, we may complete the inductive step. First, since $t^2 - 4q = t_N^2 - 4m_N = d_N$, $q \equiv m_N \pmod \ell$, and $G(t_N, m_N) \neq 0 \in \mathbf{F}_\ell$ by Equation (4.4), we also know that $G(t, q) \neq 0 \in \mathbf{F}_\ell$ by Lemma 4.2.22. Setting $m = q$ in Equation (4.5), we have

$$\sum_{0 \leq x \leq 2\sqrt{q}} G(x, q) H(x^2 - 4q) = 0 \in \mathbf{F}_\ell.$$

Since

$$G(t,q)H(t^2-4q) = G(t,q)H(d_N) \neq 0 \in \mathbf{F}_\ell,$$

this implies that there is some $0 \leq x \leq 2\sqrt{q}$ with $x \neq t$ such that $G(x,q)H(x^2-4q) \neq 0 \in \mathbf{F}_\ell$. Let $D = x^2 - 4q$, and $d = \text{f.p.}(D)$ so that

$$D = df^2 = x^2 - 4q.$$

The fact that $4q \equiv 12 \pmod{16}$ implies that $-df^2 = -D \equiv 3, 8, 11, 12 \pmod{16}$. Therefore, using the explicit knowledge of perfect squares mod 16, we know $-d \equiv 3, 8, 11, 12$ (if $f^2 \equiv 1, 9$), or $-d \equiv 3 \pmod 4$, in which case $-4d \equiv 12 \mod 16$. Let d_{N+1} equal either d or $4d$, so that $-d_{N+1} \equiv 3, 8, 11, 12 \pmod{16}$. We will construct m_{N+1}, t_{N+1} so that Equation (4.4) is satisfied for $i = N+1$. First, we prove the parts that only rely on d_{N+1}. By definition, d_{N+1} is an almost fundamental discriminant, and by Corollary 4.2.10 and Corollary 4.2.11, we have $H(d_{N+1}) \neq 0 \pmod \ell$. Also, the fact that $d_{N+1} | D = x^2 - 4q$ and $4q$ is not a quadratic residue or zero modulo any $p \in \mathscr{P}_N$ or ℓ or 3 implies that d_{N+1} is not divisible by $3, \ell$, or any $p \in \mathscr{P}_N$. Also, $d_{N+1} \neq -4$, since otherwise we would have $4q = 4g^2 + x^2$ for some g, which would imply that $q \not\equiv 3 \pmod 4$. This plus $d_{N+1} \not\equiv 0 \pmod 3$ implies that the fundamental discriminant $d \leq -7$, which lets us deduce by Lemma 4.2.23 that $\text{f.p.}(d_{N+1}) \neq \text{f.p.}(d_N)$ because

$$-d_N + t^2 = 4q = -df^2 + x^2,$$

since $x \neq t$ (the relaxation of "fundamental" to "almost fundamental" is not a problem since the factor of 4 can be absorbed in the square multiplied by d or d_N).

The only thing left to prove about d_{N+1} is that there are m_{N+1}, t_{N+1} such that $d_{N+1} = t_{N+1}^2 - 4m_{N+1}$, with $G(t_{N+1}, m_{N+1}) \neq 0 \in \mathbf{F}_\ell$ and $\left(\frac{m_{N+1}}{\ell}\right) = -1$. All we know starting out is that

$$D = d_{N+1}g^2 = x^2 - 4q.$$

where $q \equiv m_N \pmod \ell$ and thus $\left(\frac{q}{\ell}\right) = \left(\frac{m_N}{\ell}\right) = -1$. Define $m'_{N+1}, t'_{N+1} \in \mathbf{Z}$ such that in $\mathbf{F}_\ell$,

$$m'_{N+1} \mod \ell = \frac{q \mod \ell}{g^2}, \qquad t'_{N+1} \mod \ell = \frac{x \mod \ell}{g}.$$

Note that g is invertible in $\mathbf{F}_\ell$ because $D = d_{N+1}g^2 \not\equiv 0 \pmod \ell$. Then

$$d_{N+1} \equiv (t'_{N+1})^2 - 4m'_{N+1} \pmod \ell.$$

We can change m'_{N+1} by a multiple of ℓ to make this an quality, as long as we further specify the appropriate congruence mod 2 for t'_{N+1} to make this a congruence mod 4ℓ. Making that change, we have constructed m_{N+1}, t_{N+1} so that

$$d_{N+1} = t_{N+1}^2 - 4m_{N+1}.$$

Division by g^2 in $\mathbf{F}_\ell$ doesn't change the Legendre symbol, so

$$\left(\frac{m_{N+1}}{\ell}\right) = \left(\frac{q}{\ell}\right) = \left(\frac{m_N}{\ell}\right) = -1.$$

Finally, Lemma 4.2.22(3) guarantees that $G(t_{N+1}, m_{N+1}) \neq 0 \in \mathbf{F}_\ell$. This proves the inductive step.

For the base case, we just need to find an almost fundamental discriminant $d < 0$ with $d = t^2 - 4m$ satisfying Equation (4.4). For $\ell = 5$, take e.g. $d = -107$, $t = 1$, $m = 27$. For $\ell = 7$, take e.g. $d = -235$, $t = 1$, $m = 59$. For $\ell = 11$, take e.g. $d = -76$, $t = 1$, $m = 19$. I did this by eyeballing the tables in Section 4.2.4 and a few lines of PARI/GP for computing the Hurwitz–Kronecker class numbers. □

4.2.3 | Computing class numbers mod ℓ

The relations between class numbers given by Corollary 4.2.18 suggests that we might be able to use it to compute tables of Hurwitz class numbers $H(N)$ modulo ℓ. The method of Hartung [Har1974] for proving Theorem 4.2.3 also leads naturally to such an algorithm: see [Coh1993, Ch. 5]. There are two main issues: first, the coefficient of 4 means that by solving systems of linear equations of the form Equation (4.3), we can only get expressions for linear combinations $aH(4N) + bH(4N-1)$. This is solved in [Coh1993, Ch. 5] by introducing another set of linear equations. We could do this too (but note that once we do this, we are using the same input to [Har1974] that proves Theorem 4.2.3 for all ℓ). The other problem, which would only apply to this approach, is that in our case we sometimes have $G(0, m) = G(1, m) = 0$, in which case our equations do not yield much extra information (assuming we have already computed $H(N)$ for $N < 4m - 4$). For example, this always happens when $m = 0$.

For this reason, I am not sure whether my version of the trace formula approach can be used to improve on the existing algorithms, or even match their performance. It may be the case that using Equation (4.3) instead of [Coh1993, Corollary 5.3.9] whenever $G(0, m) \neq G(1, m)$ speeds up the existing algorithms in practice, since one does not need to compute a divisor sum and can look up the coefficients in a hard-coded version of the tables in Section 4.2.4.

4.2.4 | Explicit tables and formulae

This appendix contains explicit descriptions of the functions $G(t, m)$ in Equation (4.3). This is not logically necessary for any of the proofs in this paper, but is maybe useful for implementing the proposed algorithms or checking things in general. Using the description of $P_k(t, m)$ as the coefficient of x^{k-2} in

$$(1 - tx + mx^2)^{-1} \in \mathbf{Z}[t, m][\![x]\!],$$

one uses a computer algebra system (this was 3 lines of SAGE) to obtain

$$P_{12}(t, m) = t^{10} - 9t^8 m + 28t^6 m^2 - 35t^4 m^3 + 15t^2 m^4 - m^5$$
$$P_{16}(t, m) = t^{14} - 13t^{12} m + 66t^{10} m^2 - 165t^8 m^3 + 210t^6 m^4 - 126t^4 m^5 + 28t^2 m^6 - m^7$$
$$P_{18}(t, m) = t^{16} - 15t^{14} m + 91t^{12} m^2 - 286t^{10} m^3 + 495t^8 m^4 - 462t^6 m^5 + 210t^4 m^6 - 36t^2 m^7 + m^8$$
$$P_{22}(t, m) = t^{20} - 19t^{18} m + 153t^{16} m^2 - 680t^{14} m^3 + 1820t^{12} m^4 - 3003t^{10} m^5 + 3003t^8 m^6$$

$$- 1716t^6m^7 + 495t^4m^8 - 55t^2m^9 + m^{10}$$

Reducing modulo the appropriate primes, and considering these as functions rather than formal polynomials (i.e. applying Fermat's little theorem), we get

$$\begin{aligned}
P_{16}(t,m) &\equiv t^2 + 6t^4m + 4t^2m^2 + 4m^3 \pmod 5 \\
P_{18}(t,m) &\equiv t^4 + 5t^2m + t^4m^3 + 5t^2m^4 + m^2 \pmod 7 \\
P_{22}(t,m) &\equiv t^{10} + 3t^8m + 10t^6m^2 + 2t^4m^3 + 5t^2m^4 + m^{10} \pmod{11}.
\end{aligned}$$

It then follows that

$$\begin{aligned}
P_{12}(t,m) - P_{16}(t,m) &\equiv 4t^2m^2 - m + m^3 \pmod 5 \\
P_{12}(t,m) - P_{18}(t,m) &\equiv 6t^4m^3 + 3t^2m^4 + 6m^2 + 6m^5 \pmod 7 \\
P_{12}(t,m) - P_{22}(t,m) &\equiv 10t^8m + 7t^6m^2 + 9t^4m^3 - t^2m^4 + 10m^5 - 2t^4m^3 - m^{10} \pmod{11}
\end{aligned}$$

which provides explicit descriptions of the functions $G : \mathbf{F}_\ell \times \mathbf{F}_\ell \to \mathbf{F}_\ell$ from Corollary 4.2.18. The values of the functions are summarized in the following lookup tables.

		$t \bmod 5$				
		0	1	2	3	4
$m \bmod 5$	0	0	0	0	0	0
	1	0	4	1	1	4
	2	3	2	0	0	2
	3	2	0	3	3	0
	4	0	4	1	1	4

Table 4.1: The values of the function $G : \mathbf{F}_\ell \times \mathbf{F}_\ell \to \mathbf{F}_\ell$ for $\ell = 5$

		$t \bmod 7$						
		0	1	2	3	4	5	6
$m \bmod 7$	0	0	0	0	0	0	0	0
	1	6	0	1	0	0	1	0
	2	3	4	0	0	0	0	4
	3	0	6	1	0	0	1	6
	4	5	0	0	2	2	0	0
	5	0	0	5	2	2	5	0
	6	0	4	0	3	3	0	4

Table 4.2: The values of the function $G : \mathbf{F}_\ell \times \mathbf{F}_\ell \to \mathbf{F}_\ell$ for $\ell = 7$

$m \bmod 11$	$t \bmod 11$										
	0	1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0	0	0	0	0	0
1	10	10	1	0	0	1	1	0	0	1	10
2	0	0	0	0	6	5	5	6	0	0	0
3	10	1	0	1	0	10	10	0	1	0	1
4	10	1	10	0	1	0	0	1	0	10	1
5	10	0	1	1	10	0	0	10	1	1	0
6	0	0	6	5	0	0	0	0	5	6	0
7	0	6	0	0	5	0	0	5	0	0	6
8	0	5	0	6	0	0	0	0	6	0	5
9	10	0	0	10	1	1	1	1	10	0	0
10	0	0	5	0	0	6	6	0	0	5	0

Table 4.3: The values of the function $G : \mathbf{F}_\ell \times \mathbf{F}_\ell \to \mathbf{F}_\ell$ for $\ell = 11$

Bibliography

[Art1929] Emil Artin, *Idealklassen in oberkörpern und allgemeines reziprozitätsgesetz*, Abh. Math. Sem. Univ. Hamburg **7** (1929), no. 1, 46–51. MR3069515

[Art1978] James Arthur, *A trace formula for reductive groups. I. Terms associated to classes in $G(\mathbf{Q})$*, Duke Math. J. **45** (1978), no. 4, 911–952. MR518111

[Art1981] ———, *The trace formula in invariant form*, Ann. of Math. (2) **114** (1981), no. 1, 1–74. MR625344

[Art1983] ———, *The trace formula for reductive groups*, Conference on automorphic theory (Dijon, 1981), 1983, pp. 1–41. MR723181

[Art2001] ———, *A stable trace formula. II. Global descent*, Invent. Math. **143** (2001), no. 1, 157–220. MR1802795

[Art2002] ———, *A stable trace formula. I. General expansions*, J. Inst. Math. Jussieu **1** (2002), no. 2, 175–277. MR1954821

[Art2003] ———, *A stable trace formula. III. Proof of the main theorems*, Ann. of Math. (2) **158** (2003), no. 3, 769–873. MR2031854

[Art2005] ———, *An introduction to the trace formula*, Harmonic analysis, the trace formula, and Shimura varieties, 2005, pp. 1–263. MR2192011

[AT2009] Emil Artin and John Tate, *Class field theory*, AMS Chelsea Publishing, Providence, RI, 2009. Reprinted with corrections from the 1967 original. MR2467155

[Bak1968] A. Baker, *Linear forms in the logarithms of algebraic numbers. IV*, Mathematika **15** (1968), 204–216. MR258756

[BST2013] Manjul Bhargava, Arul Shankar, and Jacob Tsimerman, *On the Davenport-Heilbronn theorems and second order terms*, Invent. Math. **193** (2013), no. 2, 439–499. MR3090184

[BSV2006] Andrew R. Booker, Andreas Strömbergsson, and Akshay Venkatesh, *Effective computation of Maass cusp forms*, Int. Math. Res. Not. (2006), Art. ID 71281, 34. MR2249995

[Bum1997] Daniel Bump, *Automorphic forms and representations*, Cambridge Studies in Advanced Mathematics, vol. 55, Cambridge University Press, Cambridge, 1997. MR1431508

[Bum2013] ———, *Lie groups*, 2nd edition, Graduate Texts in Mathematics, vol. 225, Springer, New York, 2013. MR3136522

[Bun1857] Viktor Bunjakovskii, *Nouveaux théorèmes relatifs à la distinction des nombres premiers et à la décomposition des entiers en facteurs*, Sc. Math. Phys. **6** (1857), 305–329.

[CF1967] John Cassels and Albert Fröhlich, *Algebraic number theory*, Proceedings of an instructional conference organized by the London Mathematical Society (a NATO Advanced Study Institute) with the support of the International Mathematical Union., Academic Press, London; Thompson Book Co., Inc., Washington, D.C., 1967. MR0215665

[CL1984] Henri Cohen and Hendrik Lenstra, *Heuristics on class groups of number fields*, Number theory, Noordwijkerhout 1983 (Noordwijkerhout, 1983), 1984, pp. 33–62. MR756082

[CM1987] Henri Cohen and Jacques Martinet, *Class groups of number fields: numerical heuristics*, Math. Comp. **48** (1987), no. 177, 123–137. MR866103

[CM1990] ———, *Étude heuristique des groupes de classes des corps de nombres*, J. Reine Angew. Math. **404** (1990), 39–76. MR1037430

[Coh1993] Henri Cohen, *A course in computational algebraic number theory*, Graduate Texts in Mathematics, vol. 138, Springer-Verlag, Berlin, 1993. MR1228206

[Con2009] Brian Conrad, *Math 249b: Class field theory* (2009). http://virtualmath1.stanford.edu/ conrad/249BW09Page/.

[Cox2013] David A. Cox, *Primes of the form $x^2 + ny^2$*, 2nd edition, Pure and Applied Mathematics (Hoboken), John Wiley & Sons, Inc., Hoboken, NJ, 2013. Fermat, class field theory, and complex multiplication. MR3236783

[CSS1997] Gary Cornell, Joseph H. Silverman, and Glenn Stevens (eds.), *Modular forms and Fermat's last theorem*, Springer-Verlag, New York, 1997. Papers from the Instructional Conference on Number Theory and Arithmetic Geometry held at Boston University, Boston, MA, August 9–18, 1995. MR1638473

[DH1971] H. Davenport and H. Heilbronn, *On the density of discriminants of cubic fields. II*, Proc. Roy. Soc. London Ser. A **322** (1971), no. 1551, 405–420. MR491593

[DL1971] Michel Duflo and Jean-Pierre Labesse, *Sur la formule des traces de Selberg*, Ann. Sci. École Norm. Sup. (4) **4** (1971), 193–284. MR437462

[DS1974] Pierre Deligne and Jean-Pierre Serre, *Formes modulaires de poids* 1, Ann. Sci. École Norm. Sup. (4) **7** (1974), 507–530 (1975). MR379379

[Edw1977] Harold M. Edwards, *Fermat's last theorem*, Graduate Texts in Mathematics, vol. 50, Springer-Verlag, New York-Berlin, 1977. A genetic introduction to algebraic number theory. MR616635

[Gau1966] Carl Friedrich Gauss, *Disquisitiones arithmeticae*, Translated into English by Arthur A. Clarke, S. J, Yale University Press, New Haven, Conn.-London, 1966. MR0197380

[GGPS1969] Israel Moiseevich Gelfand, Mark Iosifovich Graev, and Ilya Piatetski-Shapiro, *Representation theory and automorphic functions*, Translated from the Russian by K. A. Hirsch, W. B. Saunders Co., Philadelphia, Pa.-London-Toronto, Ont., 1969. MR0233772

[GJ1979] Stephen Gelbart and Hervé Jacquet, *Forms of* $\mathrm{GL}(2)$ *from the analytic point of view*, Automorphic forms, representations and L-functions (Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977), Part 1, 1979, pp. 213–251. MR546600

[Har1974] P. Hartung, *Proof of the existence of infinitely many imaginary quadratic fields whose class number is not divisible by* 3, J. Number Theory **6** (1974), 276–278. MR352040

[HC1953] Harish-Chandra, *Representations of a semisimple Lie group on a Banach space. I*, Trans. Amer. Math. Soc. **75** (1953), 185–243. MR56610

[HC1954a] ———, *Representations of semisimple Lie groups. II*, Trans. Amer. Math. Soc. **76** (1954), 26–65. MR58604

[HC1954b] ———, *Representations of semisimple Lie groups. III*, Trans. Amer. Math. Soc. **76** (1954), 234–253. MR62747

[Hee1952] Kurt Heegner, *Diophantische Analysis und Modulfunktionen*, Math. Z. **56** (1952), 227–253. MR53135

[Hej1976] Dennis A. Hejhal, *The Selberg trace formula for* $\mathrm{PSL}(2, R)$*. Vol. I*, Lecture Notes in Mathematics, Vol. 548, Springer-Verlag, Berlin-New York, 1976. MR0439755

[Hil1896] David Hilbert, *Ein neuer beweis des kroneckerschen fundamentalsatzes über abelsche zahlkörper*, Nachrichten der gesellschaft der wissenschaften zu göttingen, 1896, pp. 29–39.

[Hil1902] ———, *Über die Theorie der relativ-Abel'schen Zahlkörper*, Acta Math. **26** (1902), no. 1, 99–131. MR1554953

[Iwa2002] Henryk Iwaniec, *Spectral methods of automorphic forms*, 2nd edition, Graduate Studies in Mathematics, vol. 53, American Mathematical Society, Providence, RI; Revista Matemática Iberoamericana, Madrid, 2002. MR1942691

[JL1970] Hervé Jacquet and Robert P Langlands, *Automorphic forms on* $\mathrm{GL}(2)$, Lecture Notes in Mathematics, Vol. 114, Springer-Verlag, Berlin-New York, 1970. MR0401654

[Kal2020] Kenz Kallal, *Representation theory at infinity* (2020). http://math.uchicago.edu/ may/REU2020/REUPapers/Kallal.pdf.

[Kar1931] Jovan Karamata, *Neuer Beweis und Verallgemeinerung der Tauberschen Sätze, welche die Laplacesche und Stieltjessche Transformation betreffen*, J. Reine Angew. Math. **164** (1931), 27–39. MR1581248

[Kis2009a] Mark Kisin, *Modularity of 2-adic Barsotti-Tate representations*, Invent. Math. **178** (2009), no. 3, 587–634. MR2551765

[Kis2009b] ———, *Moduli of finite flat group schemes, and modularity*, Ann. of Math. (2) **170** (2009), no. 3, 1085–1180. MR2600871

[KL2006] Andrew Knightly and Charles Li, *Traces of Hecke operators*, Mathematical Surveys and Monographs, vol. 133, American Mathematical Society, Providence, RI, 2006. MR2273356

[Koi1975] Masao Koike, *On some p-adic properties of the Eichler-Selberg trace formula*, Nagoya Math. J. **56** (1975), 45–52. MR382170

[Kot1984] Robert E. Kottwitz, *Shimura varieties and twisted orbital integrals*, Math. Ann. **269** (1984), no. 3, 287–300. MR761308

[Kum1850] E. E. Kummer, *Allgemeiner Beweis des Fermatschen Satzes, daßdie Gleichung $x^\lambda + y^\lambda = z^\lambda$ durch ganze Zahlen unlösbar ist, für alle diejenigen Potenz-Exponenten λ welche ungerade Primzahlen sind und in den Zählern der ersten $1/2(\lambda)$ Bernoullischen zahlen als Factoren nicht vorkommen*, J. Reine Angew. Math. **40** (1850), 130–138. MR1578681

[KW2009] Chandrashekhar Khare and Jean-Pierre Wintenberger, *Serre's modularity conjecture. II*, Invent. Math. **178** (2009), no. 3, 505–586. MR2551764

[Lan1973] Robert P. Langlands, *Modular forms and ℓ-adic representations*, Modular functions of one variable, II (Proc. Internat. Summer School, Univ. Antwerp, Antwerp, 1972), 1973, pp. 361–500. Lecture Notes in Math., Vol. 349. MR0354617

[Lan1976] ———, *On the functional equations satisfied by Eisenstein series*, Lecture Notes in Mathematics, Vol. 544, Springer-Verlag, Berlin-New York, 1976. MR0579181

[Lan1980] ———, *Base change for* $\mathrm{GL}(2)$, Annals of Mathematics Studies, vol. 96, Princeton University Press, Princeton, N.J.; University of Tokyo Press, Tokyo, 1980. MR574808

[Lan1983] ———, *Les débuts d'une formule des traces stable*, Publications Mathématiques de l'Université Paris VII [Mathematical Publications of the University of Paris VII], vol. 13, Université de Paris VII, U.E.R. de Mathématiques, Paris, 1983. MR697567

[Lan1985] Serge Lang, $\mathrm{SL}_2(\mathbf{R})$, Graduate Texts in Mathematics, vol. 105, Springer-Verlag, New York, 1985. Reprint of the 1975 edition. MR803508

[Lan2001] Robert P. Langlands, *The trace formula and its applications: an introduction to the work of James Arthur*, Canad. Math. Bull. **44** (2001), no. 2, 160–209. MR1827854

[Lan2004] ———, *Beyond endoscopy*, Contributions to automorphic forms, geometry, and number theory, 2004, pp. 611–697. MR2058622

[LS1995] Wen Zhi Luo and Peter Sarnak, *Quantum ergodicity of eigenfunctions on* $\mathrm{PSL}_2(\mathbf{Z})\backslash\mathbf{H}^2$, Inst. Hautes Études Sci. Publ. Math. **81** (1995), 207–237. MR1361757

[Mar2012] Jens Marklof, *Selberg's trace formula: an introduction*, Hyperbolic geometry and applications in quantum chaos and cosmology, 2012, pp. 83–119. MR2885182

[Neu1999] Jürgen Neukirch, *Algebraic number theory*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 322, Springer-Verlag, Berlin, 1999. Translated from the 1992 German original and with a note by Norbert Schappacher, With a foreword by G. Harder. MR1697859

[Ngô2010] Bao Châu Ngô, *Le lemme fondamental pour les algèbres de Lie*, Publ. Math. Inst. Hautes Études Sci. **111** (2010), 1–169. MR2653248

[NH1988] Jin Nakagawa and Kuniaki Horie, *Elliptic curves with no rational points*, Proc. Amer. Math. Soc. **104** (1988), no. 1, 20–24. MR958035

[Sar1980] Peter Sarnak, *Prime geodesic theorems*, ProQuest LLC, Ann Arbor, MI, 1980. Thesis (Ph.D.)–Stanford University. MR2630950

[Sar1982] ———, *Class numbers of indefinite binary quadratic forms*, J. Number Theory **15** (1982), no. 2, 229–247. MR675187

[Sch2011] Peter Scholze, *The Langlands-Kottwitz approach for the modular curve*, Int. Math. Res. Not. IMRN **15** (2011), 3368–3425. MR2822177

[Sch2013] ———, *The Langlands-Kottwitz approach for some simple Shimura varieties*, Invent. Math. **192** (2013), no. 3, 627–661. MR3049931

[Sel1956] Atle Selberg, *Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series*, J. Indian Math. Soc. (N.S.) **20** (1956), 47–87. MR88511

[Ser1973] Jean-Pierre Serre, *Formes modulaires et fonctions zêta p-adiques*, Modular functions of one variable, III (Proc. Internat. Summer School, Univ. Antwerp, 1972), 1973, pp. 191–268. Lecture Notes in Math., Vol. 350. MR0404145

[Sie1944] Carl Ludwig Siegel, *The average measure of quadratic forms with given determinant and signature*, Ann. of Math. (2) **45** (1944), 667–685. MR12642

[Smi2017] Alexander Smith, *2^∞-Selmer groups, 2^∞-class groups, and Goldfeld's conjecture*, arXiv preprint arXiv:1702.02325 (2017).

[Sta1840] K. G. C. Staudt, *Beweis eines Lehrsatzes, die Bernoullischen Zahlen betreffen*, J. Reine Angew. Math. **21** (1840), 372–374. MR1578267

[Sta1969] Harold M. Stark, *On the "gap" in a theorem of Heegner*, J. Number Theory **1** (1969), 16–27. MR241384

[Tak2014] Teiji Takagi, *Collected papers*, Springer Collected Works in Mathematics, Springer, Heidelberg, 2014. With a preface to the 1st edition by Shokichi Iyanaga, Edited by Iyanaga, Kenkichi Iwasawa, Kunihiko Kodaira and Kôsaku Yosida, With appendices by Iwasawa, Yosida and Iyanaga, Reprint of the 1990 edition. MR3309918

[Tat1950] John Tate, *Fourier analysis in number fields and Hecke's zeta-functions*, ProQuest LLC, Ann Arbor, MI, 1950. Thesis (Ph.D.)–Princeton University. MR2612222

[Tun1981] Jerrold Tunnell, *Artin's conjecture for representations of octahedral type*, Bull. Amer. Math. Soc. (N.S.) **5** (1981), no. 2, 173–175. MR621884

[TW1995] Richard Taylor and Andrew Wiles, *Ring-theoretic properties of certain Hecke algebras*, Ann. of Math. (2) **141** (1995), no. 3, 553–572. MR1333036

[Wil1995] Andrew Wiles, *Modular elliptic curves and Fermat's last theorem*, Ann. of Math. (2) **141** (1995), no. 3, 443–551. MR1333035

[Wil2015] _____, *On class groups of imaginary quadratic fields*, J. Lond. Math. Soc. (2) **92** (2015), no. 2, 411–426. MR3404031

[Woo2011] Melanie Matchett Wood, *Gauss composition over an arbitrary base*, Adv. Math. **226** (2011), no. 2, 1756–1771. MR2737799