BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
S 0273-0979(2013)01404-6
Article electronically published on January 14, 2013

THE ROLE OF THE RAMANUJAN CONJECTURE IN ANALYTIC NUMBER THEORY

VALENTIN BLOMER AND FARRELL BRUMLEY

Dedicated to the 125th birthday of Srinivasa Ramanujan

Abstract. We discuss progress towards the Ramanujan conjecture for the group GL_n and its relation to various other topics in analytic number theory.

Contents

1. Introduction

In a remarkable article [111], published in 1916, Ramanujan considered the function

$$\Delta(z) = (2\pi)^{12} e^{2\pi i z} \prod_{n=1}^{\infty} (1 - e^{2\pi i n z})^{24} = (2\pi)^{12} \sum_{n=1}^{\infty} \tau(n) e^{2\pi i n z},$$

where $z \in \mathbb{H} = \{z \in \mathbb{C} \mid \Im z > 0\}$ is in the upper half-plane. The right hand side is understood as a definition for the arithmetic function $\tau(n)$ that nowadays bears

Received by the editors June 8, 2012.

2010 *Mathematics Subject Classification.* Primary 11F70.

Key words and phrases. Ramanujan conjecture, L-functions, number fields, non-vanishing, functoriality.

The first author was supported by the Volkswagen Foundation and a Starting Grant of the European Research Council. The second author is partially supported by the ANR grant ArShiFo ANR-BLANC-114-2010 and by the Advanced Research Grant 228304 from the European Research Council.

Table 1

n	$\tau(n)$	n	$\tau(n)$	n	$\tau(n)$	n	$\tau(n)$	n	$\tau(n)$
1	1	7	-16744	13	-577738	19	10661420	25	-25499225
2	-24	8	84480	14	401856	20	-7109760	26	13865712
3	252	9	-113646	15	1217160	21	-4219488	27	-73279080
4	-1472	10	-115920	16	987136	22	-12830688	28	24647168
5	4830	11	534612	17	-6905934	23	18643272	29	128406630
6	-6048	12	-370944	18	2727432	24	21288960	30	-29211840

Ramanujan's name. He computed (at least) the first 30 values of $\tau(n)$ and made two fundamental conjectures about $\tau(n)$.

On the one hand, he conjectured that τ is *multiplicative*, that is, $\tau(nm) = \tau(n)\tau(m)$ whenever n and m are relatively prime, and that it satisfies the second order recurrence relation

$$\tau(p^{k+1}) = \tau(p)\tau(p^k) - p^{11}\tau(p^{k-1})$$

for p prime and $k \geqslant 1$. The reader can check this for the first few primes against the above table. On the other hand, he writes

> *There is reason for supposing that $\tau(n)$ is of the form $O(n^{\frac{11}{2}+\varepsilon})$ and not of the form $o(n^{\frac{11}{2}})$.*

His precise conjecture is that $|\tau(n)| \leqslant d(n)n^{11/2}$, where $d(n)$ is the number of positive divisors of n.

The multiplicativity relations and the upper bound can be restated more analytically as follows: the Dirichlet series generated by the renormalized coefficients $\lambda_\Delta(n) := \tau(n)n^{-11/2}$ has a product expansion

$$\sum_{n=1}^{\infty} \frac{\lambda_\Delta(n)}{n^s} = \prod_{p \text{ prime}} \frac{1}{1 - \lambda_\Delta(p)p^{-s} + p^{-2s}},$$

and the denominator $1 - \lambda_\Delta(p)p^{-s} + p^{-2s}$ has, as a consequence of $|\lambda_\Delta(p)| \leqslant 2$, zeros only on the line $\Re s = 0$. The product expansion (and hence the multiplicativity of $\tau(n)$) was immediately proved by Mordell [97]. The vanishing condition, which is a sort of local Riemann hypothesis and is in fact equivalent to the upper bound on $\tau(n)$, revealed itself to be substantially harder.

The function $\Delta(z)$ is called the discriminant function, as it associates to $z \in \mathbb{H}$ the discriminant of the elliptic curve $\mathbb{C}/(\mathbb{Z}.z + \mathbb{Z}.1)$. It is a constant multiple of the 24th power of the Dedekind η-function which plays an important role in the study of the number of representations of an integer as a sum of squares. One of its most important properties is its transformation behavior under certain fractional linear transformations, which we proceed to describe in some generality.

A *modular form* of weight k for $\mathrm{SL}_2(\mathbb{Z})$ is a holomorphic function f on $\mathbb{H}$ that is bounded near $i\infty$ and satisfies

$$f\left(\frac{az+b}{cz+d}\right) = (cz+d)^k f(z), \quad \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}). \tag{1}$$

Applying this transformation rule with the unipotent matrix $\left(\begin{smallmatrix} 1 & 1 \\ & 1 \end{smallmatrix}\right)$, we see that f is 1-periodic, and hence can be expanded into a Fourier series

$$f(z) = \sum_{n=0}^{\infty} a_f(n) e^{2\pi i n z}.$$

We write M_k for the space of such functions and S_k for the subspace of *cusp forms*, the subspace of such functions f with $a_f(0) = 0$. What makes this setting so rich from an arithmetic standpoint is that M_k is endowed with an action of a commutative algebra $\bigoplus_{n \in \mathbb{N}} \mathbb{C}.T_n$ of self-adjoint operators $T_n : M_k \to M_k$, the Hecke algebra, leaving stable the subspace S_k.

One can verify that $\Delta(z)$ satisfies the transformation property (1) for $k = 12$, and hence is a modular form of weight 12. It has a vanishing constant Fourier coefficient so that $\Delta \in S_{12}$. One can show that S_{12} is one dimensional, so each T_n acts as a scalar and Δ is automatically a Hecke eigenform. The eigenvalue of T_n acting on $(2\pi)^{-12}\Delta$ is precisely the coefficient $\tau(n)$. An introduction to this theory can be found in a large number of books, for example [80, 130].

To bring in more arithmetic, one relaxes the transformation property (1) to certain subgroups Γ of finite index in $\mathrm{SL}_2(\mathbb{Z})$ defined by congruence conditions on their entries. The resulting quotient spaces $\Gamma\backslash\mathbb{H}$ can be viewed as moduli spaces of elliptic curves with additional structure [42, Section 1.5]. The corresponding theory is extremely rich: it produces the classical modular forms on which much of modern number theory is based.[1] The generalization of the classical Ramanujan conjecture to this context is known as the Ramanujan–Petersson conjecture, in recognition of Petersson's extension [106] to congruence subgroups and arbitrary weight. The conjecture states that the (appropriately normalized) Hecke eigenvalues satisfy $|\lambda(n)| \leqslant d(n)$. When applied to the discriminant function, this reproduces Ramanujan's original conjecture.

The history of work leading to the resolution of this conjecture highlights some of the finest mathematical achievements of the last century. Eichler [46] was the first to recognize the role of arithmetic geometry in relation to the Ramanujan–Petersson conjecture, by reducing the weight $k = 2$ case to the Weil conjectures for algebraic curves over finite fields. Important contributions by Shimura, Kuga, Ihara, and Deligne similarly reduced the Ramanujan–Petersson conjecture for all weights $k \geqslant 2$ to the full Weil conjectures (see [39]). With these implications established, the Ramanujan–Petersson conjecture became a theorem when Deligne proved the Weil conjectures in full generality [37]. More recently, a succession of works by Clozel, Harris, Taylor, and their coworkers (see Section 4.3 for precise references) has succeeded in establishing the Ramanujan conjecture for certain cuspidal automorphic forms arising through cohomological methods, which in effect extends Deligne's theorem to the most general framework.

Yet the automorphic spectrum of GL_2 contains much more than the holomorphic forms, most notably the Maaß forms, which we will describe in the next section. As important as it is to know that automorphic forms of a prescribed type verify the Ramanujan conjecture—the subset of cohomological forms, for example—analytic applications often require a control on worst-case scenarios, so that no cusp form should violate the Ramanujan conjecture "by too much". The goal of this article

[1] Eichler classified modular forms as the fifth arithmetic operation—besides addition, subtraction, multiplication, and division.

is to describe some recent progress by the authors [9] towards the generalization of the classical Ramanujan conjecture to cuspidal automorphic forms on GL_2 over a number field. We start by stating the general conjecture as it pertains to the group GL_n, then describe through examples the role that *bounds* towards the Ramanujan conjecture (not establishing the full conjecture, but valid for *all* automorphic forms) play in analytic number theory. We spend some time sketching the central ideas in the article [9] and other historical precedents. In the last section, we offer some perspectives on our current understanding of the conjecture.

There are already a host of fine expositions of the Ramanujan conjecture. For example, Cogdell's Park City notes [33] provide an excellent introduction to the conjecture for GL_n. The IAS/Park City Summer School and the Toronto Clay Summer School proceedings contain articles by Clozel [28] and Sarnak [123], respectively, which together provide probably the most detailed source about what is known about the Ramanujan conjecture in the general setting of connected linear reductive groups. Our hope is that this exposition will serve as a reminder of the mysteries surrounding the arithmetic significance of Maaß forms, for which the Ramanujan conjecture is famously still open.

2. Background on Maass forms

While the holomorphic forms of weight $k \geqslant 2$ for $\mathrm{SL}_2(\mathbb{Z})$ can be identified, via the transformation formula (1), with sections of certain line bundles on the modular curve $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$, any weight zero form is properly invariant under the group $\mathrm{SL}_2(\mathbb{Z})$ and hence descends to the quotient $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ as a well-defined function. Now a standard argument [80, p. 117 and p. 129] shows that a holomorphic modular form of weight zero on a space such as $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ must be constant. In order to obtain an interesting class of functions (with which we can ultimately perform harmonic analysis), one must therefore relax the holomorphy condition. In place of solutions to the Cauchy–Riemann differential equations, one considers eigenfunctions of the hyperbolic Laplace operator. These non-holomorphic, but real analytic, $\mathrm{SL}_2(\mathbb{Z})$-invariant functions were called *waveforms* by Maaß, who introduced them, in analogy with a vibrating membrane.

We proceed to give precise definitions in a slightly more general context, as in the original paper [93]. In particular, we allow Maaß forms with weight so that the previously considered holomorphic forms become (after a small renormalization) a special case of Maaß forms. This will complete the description of the automorphic spectrum of GL_2 over $\mathbb{Q}$. For more details the reader is invited to consult, for instance, Sections 1.9 and 2.1 of the book [19], Section 4 of the article [43], or the overview [17].

2.1. **Weight k Maaß forms.** The group $\mathrm{SL}_2(\mathbb{R})$ acts on the upper half-plane $\mathbb{H}$ by fractional linear transformations. For an integer k and $g = \left(\begin{smallmatrix} * & * \\ c & d\end{smallmatrix}\right) \in \mathrm{SL}_2(\mathbb{R})$, we define an operator $R_g^{(k)}$ on functions on $\mathbb{H}$ by

$$R_g^{(k)} f(z) = \left(\frac{cz+d}{|cz+d|}\right)^{-k} f(g.z).$$

For any positive integer N we introduce the Hecke congruence subgroup

$$\Gamma_1(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) \mid c \equiv 0 \,(\mathrm{mod}\ N), a \equiv d \equiv 1 \,(\mathrm{mod}\ N) \right\}.$$

Let $Y_1(N) = \Gamma_1(N)\backslash\mathbb{H}$, and denote by $L^2(Y_1(N),k)$ the L^2-space of functions satisfying

$$R_\gamma^{(k)} f = f \quad \text{for } \gamma \in \Gamma_1(N). \tag{2}$$

The Hilbert space structure is taken with respect to the inner product

$$\int_{Y_1(N)} f(z)\overline{g(z)}\frac{dxdy}{y^2}. \tag{3}$$

We can make the transformation formula (2) more transparent by bringing to the foreground the underlying notions in representation theory. For this it is helpful to lift the function f from $\mathbb{H}$ to a function F on the group $\mathrm{SL}_2(\mathbb{R})$. The recipe for doing so is as follows. First note that since $\mathrm{SL}_2(\mathbb{R})$ acts transitively on $\mathbb{H}$ and the stabilizer of i is $\mathrm{SO}(2)$, we can identify $\mathbb{H}$ with the quotient $\mathrm{SL}_2(\mathbb{R})/\mathrm{SO}(2)$. Then, if f satisfies (2), the lifted function $F(g) := (R_g^{(k)} f)(i)$ satisfies

$$F(\gamma g \kappa) = e^{ik\theta} F(g) \tag{4}$$

for all $g \in \mathrm{SL}_2(\mathbb{R})$, $\gamma \in \Gamma_1(N)$, and

$$\kappa = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \in \mathrm{SO}(2).$$

One should think of (4) as describing the (Hilbert space direct sum) decomposition of the right regular representation of $\mathrm{SL}_2(\mathbb{R})$ on the highly reducible space

$$L^2(\Gamma_1(N)\backslash\mathrm{SL}_2(\mathbb{R})) = \bigoplus_k L^2(Y_1(N),k),$$

according to the action of the Abelian subgroup $\mathrm{SO}_2(\mathbb{R})$.

One can reduce the space $L^2(Y_1(N),k)$ still further by generalizing the transformation property (2). Let

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

and let ψ be a Dirichlet character modulo N. Let $Y_0(N) = \Gamma_0(N)\backslash\mathbb{H}$, and denote by $L^2(Y_0(N),k,\psi)$ the L^2-space of functions satisfying

$$R_\gamma^{(k)} f = \psi(d) f, \qquad \text{for } \gamma \in \Gamma_0(N). \tag{5}$$

Then

$$L^2(Y_1(N),k) = \bigoplus_{\psi\,(\mathrm{mod}\,N)} L^2(Y_0(N),k,\psi),$$

since $\Gamma_0(N)/\Gamma_1(N) \cong (\mathbb{Z}/N\mathbb{Z})^\times$. These latter spaces are the building blocks of the theory of Maaß forms.

On $\mathrm{SL}_2(\mathbb{R})$ there exists a unique up to scaling left $\mathrm{SL}_2(\mathbb{R})$-invariant second order differential operator, called the Casimir operator. When we restrict the Casimir operator to one of the weight k spaces $L^2(Y_1(N),k)$, we obtain the weight k Laplacian

$$\Delta_k = -y^2(\partial_x^2 + \partial_y^2) + iky\partial_x,$$

a positive operator. When $k = 0$, this recovers the Laplacian $\Delta = \Delta_0$ on $\mathbb{H}$ associated to the hyperbolic metric. We will use the weight k Laplacian as a tool to decompose the space $L^2(Y_1(N),k,\psi)$.

We have finally arrived at the definition of a *weight k Maaß form*, for $k \in \mathbb{Z}$. By this we mean a function on $\mathbb{H}$ verifying (5), satisfying a moderate growth condition, and which is an eigenfunction of Δ_k. As an eigenfunction of an elliptic operator, a Maaß form f is automatically real analytic. The character ψ is called the *nebentypus* of f, and the integer N is called its *level*. Given a weight k, a level N, a character ψ mod N, and an eigenvalue λ of Δ_k, we denote by $M_k(N, \psi, \lambda)$ the space of Maaß forms with this data. The space of cusp forms, consisting of those f which vanish in the cusps, is denoted $S_k(N, \psi, \lambda)$.

The definition of weight k Maaß form encompasses the weight k holomorphic forms f discussed in the Introduction. One sees this by applying the embedding

$$f \mapsto y^{k/2} f \tag{6}$$

of weight k modular forms into the space of weight k Maaß forms. In this case, the Δ_k eigenvalue of $y^{k/2} f(x + iy)$ is the neat expression $\frac{k}{2}(1 - \frac{k}{2})$.

2.2. **Additional symmetries.** In the preceding discussion we did not make use of the integral and congruence properties of the lattices $\Gamma_1(N)$ and $\Gamma_0(N)$ which define the hyperbolic surfaces $Y_1(N)$ and $Y_0(N)$. To obtain the automorphic spectrum of the group GL_2, thought of as reductive group defined over $\mathbb{Q}$, one should take into account the symmetries of the space $Y_1(N)$ arising from the arithmetic of $\Gamma_1(N)$. These symmetries give rise to what are known as the Hecke operators, denoted T_n.

The simplest Hecke operator actually comes from complex conjugation. We describe it only in the weight zero case. Namely, let T_{-1} be the involutive operator on $M_0(N, \psi, \lambda)$ given by precomposition with $z \mapsto -\bar{z}$. We call a weight zero Maaß form f *even* or *odd* according to whether $T_{-1} f = f$ or $-f$.

Once the full theory of Hecke operators is developed, the space $L^2(Y_1(N), k, \psi)$ can be diagonalized into common eigenforms of Δ_k *and* the T_n, and the automorphic spectrum of $\mathrm{GL}_2/\mathbb{Q}$ will be in a natural one-to-one correspondence with weight $k \geqslant 0$ Maaß forms which are *new*. The terminology *newform* has a precise meaning that will be given later, in Section 3.1. But we can think of *new* in an extended sense, as an informal description of those forms which respect all conceivable symmetries.

One way f can be *old* is if it can be obtained from another Maaß form of different weight. Indeed, for any integer k, there are differential operators $L_k : S_k(N, \psi, \lambda) \to S_{k-2}(N, \psi, \lambda)$ and $K_k : S_k(N, \psi, \lambda) \to S_{k+2}(N, \psi, \lambda)$, called the lowering and raising operators, which allow one to pass between spaces of like parity weights. These maps are surjective and their kernels admit a natural description. For example, we have $L_k f = 0$ precisely when $y^{-k/2} f$ is holomorphic, in which case we necessarily have $f \in S_k(N, \psi, \frac{k}{2}(1 - \frac{k}{2}))$. In this way, we can identify $S_0(N, \psi, \lambda)$ with all its isomorphic even weight companion spaces $S_{2k}(N, \psi, \lambda)$. Similarly, $S_1(N, \psi, \lambda)$, where $\lambda > 1/4$, can be identified with any $S_{2k+1}(N, \psi, \lambda)$. The remaining forms can be identified with the classical weight $k \geqslant 1$ holomorphic cusp forms. One therefore obtains the full cuspidal automorphic spectrum on GL_2 over $\mathbb{Q}$ by considering Maaß cusp forms of weight 0 and 1 together with holomorphic cusp forms of weight $k \geqslant 2$.

The weight $k = 0$ and 1 Maaß cusp forms constitute the automorphic spectrum of *low weight* for $\mathrm{GL}_2/\mathbb{Q}$. They behave in many ways quite differently from the classical integral weight $k \geqslant 2$ holomorphic forms. The following paragraphs are intended to highlight some of these differences.

2.3. **Dihedral Maaß forms.** In this subsection we discuss the historically first explicit construction of Maaß forms. In an important 1927 article [58], Hecke introduced a technique, later generalized by Maaß [93], for constructing automorphic forms for GL_2 from the GL_1 theory. The construction associates with a Hecke character χ of a quadratic field extension K of $\mathbb{Q}$ a Maaß form f_χ (not necessarily cuspidal) such that $L(s, f_\chi) = L(s, \chi)$. One calls a Maaß form obtained by this construction a *dihedral form*. In the automorphic language, f_χ is the automorphic induction of χ from GL_1/K to $\mathrm{GL}_2/\mathbb{Q}$. We briefly reprise the work of Hecke and Maaß in this section; see also [19, Section 1.9] or [63, Section 12] for more details.

Let K be a number field. In adelic language a (unitary) Hecke character is a continuous homomorphism $K^\times\backslash\mathbb{A}_K^\times \to S^1$. Classically, a Hecke character is a product of three characters, η, χ_{fin}, and χ_∞, satisfying a compatibility condition. Here η is a class group character, χ_{fin} is a character of $(\mathcal{O}_K/\mathfrak{q})^\times$ for some integral ideal $\mathfrak{q}$ of the ring of integers $\mathcal{O}_K$ of K, and χ_∞ is a character of the Minkowski space $K_\infty^\times$ (the product of the multiplicative groups of all Archimedean completions of K). The compatibility condition, that $\chi_{\mathrm{fin}}\chi_\infty$ should be trivial on units, makes their product a well-defined character on ideals rather than on numbers.

We begin with Hecke's construction which associates with a Hecke character χ of an imaginary quadratic field K the theta series

$$f_\chi(z) = \sum_{\mathfrak{a}\subseteq\mathcal{O}_K} \chi(\mathfrak{a})\mathrm{N}(\mathfrak{a})^{(k-1)/2}e^{2\pi i\mathrm{N}(\mathfrak{a})z} = \sum_{n\geqslant 1} \lambda_{f_\chi}(n)n^{(k-1)/2}e^{2\pi inz}.$$

Here $k-1 \geqslant 0$ is the frequency of the component at infinity $\chi_\infty(z) = (z/|z|)^{k-1}$ and $\mathrm{N}(\mathfrak{a})$ is the absolute norm of $\mathfrak{a}$. That f_χ is a holomorphic form of weight k and level $|d|\,\mathrm{N}(\mathfrak{q})$, where d is the discriminant of K and $\mathfrak{q}$ the conductor of χ, follows from an application of Poisson summation (see e.g. [63, Section 12]). The Δ_k-eigenvalue of $y^{k/2}f_\chi$ is the same as that of the frequencies $y^{k/2}\exp(2\pi iz)$ from which it is built, namely $\frac{k}{2}(1-\frac{k}{2})$.

Maaß's theory begins with a Hecke character of a *real* quadratic field. He used the classical Whittaker functions as more general frequencies than the exponentials from which to build his theta series. In this way, the condition of holomorphy was relaxed, generating a much larger class of automorphic forms. To describe the construction, we observe that the component at infinity χ_∞ of a unitary Hecke character χ of a real quadratic field K over $\mathbb{Q}$ takes the form $\chi_\infty(x,y) = \mathrm{sgn}(x)^a\mathrm{sgn}(y)^b|x/y|^{ir}$, where $a, b \in \{0,1\}$, $r \in (\pi/\log\varepsilon)\mathbb{Z}$, and ε is a fundamental unit in K; see [19, Section 1.7]. Set k to be 0 or 1 according to whether $a+b$ is even or odd. With such χ Maaß associated the series

$$f_\chi(z) = \sum_{\mathfrak{a}\subseteq\mathcal{O}_K} \chi(\mathfrak{a})\mathcal{W}_{k/2,ir}(\mathrm{N}(\mathfrak{a})z) = \sum_{n\geqslant 1} \lambda_{f_\chi}(n)\mathcal{W}_{k/2,ir}(nz)$$

formed from linear combinations of the Whittaker frequencies $\mathcal{W}_{k/2,ir}$. These are Δ_k-eigenfunctions on $\mathbb{H}$ given $\mathcal{W}_{k/2,ir}(x+iy) = e^{2\pi ix}W_{k/2,ir}(4\pi y)$, where $W_{k/2,ir}$ is the classical weight $k = 0,1$ Whittaker function [152]. Once again, Poisson summation can be used[2] to show that f_χ is a weight k Maaß form for some congruence subgroup, its Δ_k-eigenvalue being the same as that of $\mathcal{W}_{k/2,ir}$, namely $1/4 + r^2$. When $k = 1$ and $r = 0$, the Whittaker function $\mathcal{W}_{1/2,0}(z)$ reduces to the exponential $y^{1/2}e^{2\pi iz}$ so that it comes from a weight 1 holomorphic theta series via (6).

We make two observations. The first is that f_χ is a cusp form whenever χ is not of the form $\psi\circ\mathrm{N}$ on ideals coprime to the conductor of χ for some Dirichlet

[2] One can also employ the converse theorem with level [150] to show the modularity of these lifts—many texts follow this strategy—but this tool, developed later by Weil, was not available to either Hecke or Maaß.

character ψ.[3] In this case f_χ is in fact an eigenfunction for all Hecke operators T_n, and its Hecke eigenvalues at a rational prime p are

$$\lambda_{f_\chi}(p) = \begin{cases} 0, & \text{if } p \text{ is inert in } K, \\ \chi(\mathfrak{p}) + \chi(\bar{\mathfrak{p}}), & \text{if } p \text{ splits in } K \text{ as } \mathfrak{p}\bar{\mathfrak{p}}, \text{ and} \\ \chi(\mathfrak{p}), & \text{if } p = \mathfrak{p}^2 \text{ ramifies in } K, \end{cases}$$

so that, trivially, $|\lambda_{f_\chi}(p)| \leqslant 2$. Our second observation is that one can characterize the low weight dihedral Maaß forms of Δ_k-eigenvalue $1/4$ ($k = 0, 1$): they are precisely those coming from finite order (or ray class) characters. Their coefficients, being the sum of two roots of unity, are thus algebraic integers.

2.4. **The Artin conjecture.** The dihedral Maaß forms are just the first instance of a more general correspondence between complex representations of the Weil group of $\mathbb{Q}$ and certain automorphic forms, as enunciated by Artin and Langlands. We now review what is currently known of this more general picture for Maaß forms. Although many open questions remain, our discussion will highlight some extraordinary recent progress. For background on the Weil group and its representations, we refer the reader to [143].

Class field theory identifies a Hecke character χ of a quadratic field extension K of $\mathbb{Q}$ with a character of the Weil group W_K of K. One may then induce χ to an irreducible 2-dimensional complex representation $\rho_\chi = \mathrm{Ind}_{W_K}^{W_{\mathbb{Q}}}(\chi)$ of $W_{\mathbb{Q}}$. If, in addition, χ is of finite order, then one may identify it with a character of the absolute Galois group G_K of K. In this case, the induced representation $\rho_\chi = \mathrm{Ind}_{G_K}^{G_{\mathbb{Q}}}(\chi)$ has finite image in $\mathrm{GL}_2(\mathbb{C})$ and its image in $\mathrm{PGL}_2(\mathbb{C})$ is known to be isomorphic to a dihedral group. We therefore call the dihedral Maaß forms arising from finite order characters dihedral Maaß forms *of Galois type*; by the description in the previous paragraph, these are precisely the dihedral Maaß forms of eigenvalue $1/4$. The action of ρ_χ on complex conjugation $c \in G_{\mathbb{Q}}$ determines the weight of the corresponding f_χ. We say that a Galois representation is even or odd according to whether $\det(\rho(c))$ is 1 or -1. If ρ_χ is even, then f_χ is of weight zero. If ρ_χ is odd, then f_χ is of weight 1 (and being of eigenvalue $1/4$ is therefore holomorphic).

The work of Langlands [86] on solvable base change for GL_2 allows one to partially generalize the above modular correspondence. The finite subgroups of $\mathrm{GL}_2(\mathbb{C})$ are classified by their image in $\mathrm{PGL}_2(\mathbb{C})$. A classical result going back at least to Klein [77] states that the latter fall into five isomorphism classes: cyclic, dihedral, tetrahedral (isomorphic to A_4), octahedral (isomorphic to S_4), and icosahedral (isomorphic to the non-solvable group A_5). The theorem of Langlands [86] and Tunnell [147] then states that if ρ is an irreducible 2-dimensional complex representation of $G_{\mathbb{Q}}$ whose isomorphism type in $\mathrm{PGL}_2(\mathbb{C})$ is not of icosahedral type,[4] then there exists a Maaß cusp form f such that $L(s, f) = L(s, \rho)$, the latter being the Artin L-function (see e.g. [19, Section 1.8] or [100, Lecture 10] for a definition). This result established the solvable case of the Artin conjecture, the statement of which is that to any irreducible 2-dimensional complex Galois representation should correspond a Maaß cusp form with matching L-functions.

[3] The smallest discriminant of a quadratic field allowing for class group characters that do not factor through the norm is $d = -23$.

[4] The cyclic case does not appear for irreducible two dimensional representations, and the dihedral case is just the correspondence of Hecke and Maaß. Langlands treated the tetrahedral and odd octahedral cases while Tunnell's work dealt with the even octahedral case.

This left the non-solvable case of the Artin conjecture open, in which state it stood for many years. It is at this point where the distinction between even and odd Galois representations becomes critical. In 1997, Khare [68] reduced the Artin conjecture for odd representations to the Serre modularity conjecture. In the meantime, Buzzard, Dickinson, Shepherd-Barron, and Taylor in [24] and again Taylor in [146] were able to establish the modularity of a wide class of examples of odd icosahedral representations. Very recently, Khare and Wintenberger [69–71] proved Serre's conjecture, and hence the full strength of the Artin conjecture for odd representations. Remarkably, there has been essentially no progress towards modularity of *even* icosahedral Galois representations.

2.5. **Some mysteries of Maaß forms.** In this section we attempt to convey some of the more mysterious aspects of weight zero Maaß forms.

2.5.1. *Existence.* In the previous two subsections we discussed many important examples of Maaß cusp forms coming from Galois representations (these all have eigenvalue 1/4) as well as those coming from representations of the larger Weil group of $\mathbb{Q}$ (this is the more general class of dihedral Maaß cusp forms, whose eigenvalue is determined by the frequency of the component at infinity of the inducing character). Outside of these examples there are no other explicit constructions of Maaß cusp forms known.

Indeed the very existence of weight zero Maaß cusp forms for $\mathrm{SL}_2(\mathbb{Z})$—the full modular group does not support any of Maaß's lifts, since there are no totally unramified extensions of $\mathbb{Q}$—had to wait until the development of the Selberg trace formula [128]. Selberg's principal motivation for the development of his trace formula was, in fact, to deduce from it the full Weyl law for the asymptotic count of eigenvalues for Maaß forms which in particular implies the existence of Maaß forms; see e.g. [64, Section 15].

The quotients of $\mathbb{H}$ by congruence groups are not compact. The difficulty here as compared with the Weyl law for compact Riemannian manifolds (which had been proved much earlier by Minakshisundaram and Pleijel [95]) is the presence of the continuous spectrum coming from the Eisenstein series. Discrete eigenvalues embedded within the continuous spectrum are known in physics to be highly unstable. While we refer the reader to the excellent discussion in Sarnak's Baltimore lecture [122] for more details on this subject, we do remark that since the continuous spectrum is even, the existence of odd Maaß forms is straightforward to establish. For even Maaß cusp forms the situation is more delicate. Several years ago, Lindenstrauss and Venkatesh [90] came up with an ingenious, yet simple, method to exhibit even weight zero Maaß cusp forms for $\mathrm{SL}_2(\mathbb{Z})$. Their idea is to use the Hecke operator T_p at any fixed prime p to construct an operator on even Maaß forms which, while killing the Eisenstein series, has non-zero image. A summary of the method of Lindenstrauss and Venkatesh in the concrete setting of the upper half-plane (and including a simple proof of the existence of odd Maaß cusp forms) can be found in [53, Chapter 4] or [4, Section 4.3.2].

2.5.2. *Galois correspondence.* Shortly after Deligne proved the Weil conjectures, Deligne and Serre [40] were able to associate with any holomorphic cuspidal weight one Hecke eigenform f an odd irreducible 2-dimensional complex representation ρ of the absolute Galois group $G_{\mathbb{Q}}$ such that $L(s,f) = L(s,\rho)$. Attempts have been made to establish a similar theorem for eigenvalue 1/4 *weight zero* Hecke–Maaß forms, showing that they too are all of Galois type. That they should be so is more or less a folklore conjecture. We point out that Carayol [26] has proposed an approach for proving this Galois correspondence, which reduces to proving a Deligne–Serre type theorem for *degenerate* limits of discrete series on $\mathrm{U}(2,1)$.

As difficult as it may be to show that all weight zero Hecke–Maaß forms of eigenvalue $1/4$ are of Galois type, at least there is a credible conjecture to work with. By contrast, for Maaß forms of Δ_k-eigenvalue ($k = 0,1$) not equal to $1/4$, there are no known links to Galois theoretic objects.[5] Nor does one believe such a correspondence to exist. Indeed, the field generated by the Hecke eigenvalues of Maaß forms of eigenvalue $> 1/4$ is thought to be transcendental, a conjecture that was first put forward in a much wider context by Clozel [29, Conjecture 3.8]. This was proven by Sarnak for weight zero dihedral Maaß forms [121] of eigenvalue $> 1/4$ using classical results in transcendence theory. For the more delicate non-dihedral forms, strong numerical evidence for this conjecture was recently given in a paper by Booker, Strömbergsson, and Venkatesh [14]. Among many other computational results, they show that the first non-zero eigenvalue of the Laplacian on $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$, which numerically is roughly equal to $91.141345\cdots$. is not the solution of any algebraic equation of degree at most 10, all of whose coefficients are of size at most 10^7.

Ultimately, one would like an intrinsic understanding of the arithmetic Maaß forms encode. Quite amazingly, Langlands has conjectured the existence of a group, much larger than the Galois or Weil group of $\mathbb{Q}$, whose unitary irreducible finite-dimensional complex representations (taken up to isomorphism) parametrize the full cuspidal automorphic spectrum of GL_2 over $\mathbb{Q}$, *including* the transcendental Maaß forms. For more on this mysterious, and to this day hypothetical, Langlands group, see Arthur's article [2].

2.5.3. *The Ramanujan conjecture.* One of the most noteworthy consequences of the Deligne–Serre theorem mentioned in the previous subsection is that the Ramanujan–Petersson conjecture is true for holomorphic cuspidal newforms of weight one. From the discussion thus far, one can summarize the situation as follows. Of the cuspidal automorphic spectrum for GL_2 over $\mathbb{Q}$, those forms for which the Ramanujan–Petersson conjecture has been established are the holomorphic modular forms of integer weight $k \geqslant 1$, the dihedral Maaß forms, and the Maaß forms of eigenvalue $1/4$ known to come from Galois representations.

For all other weight $k = 0, 1$ Maaß forms, the Ramanujan–Petersson conjecture is still open. It has often been described as the fundamental unsolved conjecture in the analytic theory of automorphic forms. This is the subject of the next section, where we describe the Ramanujan–Petersson conjecture for Maaß forms at some length, along with its Archimedean (weight zero) counterpart, the Selberg conjecture.

3. The Ramanujan conjecture for Maass forms

We group under the common heading the *Ramanujan conjecture* the two conjectures of Ramanujan–Petersson and Selberg. In fact, the more general set-up in the next section will provide a unifying framework in which to view them. With our

[5] One must be careful to exclude from this statement those Maaß forms admitting a self-twist, for they do in fact correspond to representations of the Weil group, according to the characterization given by Langlands and Labesse [83, Proposition 6.5]. A weight zero Maaß form admitting a self-twist will be of dihedral *Galois* type only when its Laplacian eigenvalue is equal to $1/4$.

attention focused on the classical case, we will be able to highlight connections with other areas that are less transparent in the higher rank situation. For simplicity, we restrict our discussion to the case of weight zero forms. The term Maaß form in this section will implicitly mean weight zero.

3.1. **Ramanujan–Petersson conjecture.** As in the holomorphic setting, the Ramanujan–Petersson conjecture for the weight zero spectrum will apply only to those Maaß cusp forms which are eigenfunctions of the Hecke algebra. With this in mind, we proceed to give the definition of the Hecke operators T_p.

For $p \nmid N$ let $\mathcal{C}_p$ be the $(p+1)$-valent correspondence on $Y_1(N)$ sending a point $z \in Y_1(N)$ to the collection

$$\mathcal{C}_p(z) = \left\{ \frac{az+d}{p} : ad = p, 1 \leqslant b \leqslant d \right\} \subset Y_1(N).$$

If ψ is a Dirichlet character modulo N and $w \in \mathcal{C}_p(z)$ is given by $(az+d)/p$, we write $\psi(w)$ for the value $\psi(a)$. The pth Hecke operator T_p, for $p \nmid N$, acts on the space of Maaß forms for $\Gamma_1(N)$ with nebentypus ψ by the rule

$$T_p(f)(z) = \frac{1}{\sqrt{p}} \sum_{w \in \mathcal{C}_p(z)} \psi(w) f(w).$$

The T_p commute with each other and with Δ_k, and they are normal with respect to the Petersson inner product, defined in (3). In so much as they average a function at a point over $p+1$ of its neighbors, the Hecke operators T_p should be viewed as the appropriate p-adic version of the Laplacian.

A Hecke–Maaß cusp form for $\Gamma_1(N)$ is then a Maaß cusp form which is, in addition, a joint eigenfunction for all the T_p, $p \nmid N$. We call f an oldform if it can be written as $f(z) = g(dz)$ for some g on $Y_1(M)$, where M properly divides N and $d \mid N/M$. Then a cuspidal *newform* of level N is a Hecke–Maaß cusp which is orthogonal to the space of oldforms and whose Fourier–Whittaker expansion

$$f(z) = \sum_{n \neq 0} \rho_f(n) W_{0,ir}(4\pi|n|y) e(nx)$$

is normalized to have $\rho_f(1) = 1$. Letting $\lambda_f(p)$ denote the eigenvalue of T_p on a Hecke–Maaß cuspidal newform f, we have the relation $\lambda_f(p) = \rho_f(p)p^{1/2}$ for $p \nmid N$, and the Ramanujan–Petersson conjecture is that $|\lambda_f(p)| \leqslant 2$.

Alternatively, one can avoid the theory of newforms and Fourier expansions and express the Ramanujan–Petersson conjecture simply as

$$\left\| T_p \big|_{L^2_{\text{cusp}}(Y_1(N))} \right\| \leqslant 2.$$

3.2. **Selberg eigenvalue conjecture.** The hyperbolic Laplacian $\Delta = -y^2(\partial_x^2 + \partial_y^2)$, acting on smooth functions $C^\infty(Y_1(N))$, admits an extension to a self-adjoint operator on the Hilbert space $L^2(Y_1(N))$. The unitary Eisenstein series furnish the continuous spectrum of Δ, which is precisely $[1/4, \infty)$. Now Δ is a positive operator, so its spectrum is bounded from below by 0. Since the eigenvalue 0 is in fact realized by the one-dimensional space of constant functions, a natural question to ask is how small the next smallest eigenvalue is.

In a seminal paper, Selberg [129] conjectured that Maaß cusp forms on congruence hyperbolic surfaces $Y_1(N)$ have Laplacian eigenvalue at least 1/4; in other words, there is a spectral gap of optimal size 1/4 for the surfaces $Y_1(N)$. We will

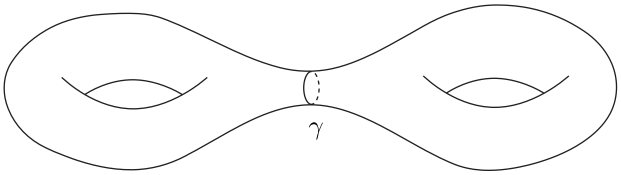


FIGURE 1. A pinched hyperbolic surface

see in Section 4 that the Selberg eigenvalue conjecture is the Archimedean analogue of the Ramanujan conjecture. In the half-century since Selberg stated his conjecture, much progress has been made, yet it remains one of the most important open problems in number theory. The following paragraphs show that the existence of a spectral gap seems to be a reflection of the *arithmetic* nature of surfaces $Y_1(N)$.

To shed light on the geometric and arithmetic nature of Selberg's conjecture, let us first state the variational description

$$\lambda_1(M) = \inf_{\int_M f=0} \frac{\int_M |\nabla f|^2}{\int_M |f|^2} \tag{7}$$

of the first non-zero Laplacian eigenvalue $\lambda_1(M)$ on a compact Riemannian manifold M. If M is a hyperbolic surface, say of genus 2, one can pinch M at the neck, while choosing f to be constantly 1 and -1 on either side, thereby obtaining a continuous deformation of M admitting arbitrarily small λ_1. Figure 1 shows such a surface being pinched at a closed geodesic γ whose length is tending towards 0.

This type of behavior is not unique to continuous deformations of a fixed hyperbolic surface. A similar phenomenon arises when considering cyclic covers of a fixed surface. To construct such a tower, let γ be a closed geodesic on the base surface S_0 whose complement is connected. Any $x_0 \notin \gamma$ then induces a surjective homomorphism $m_\gamma : \pi_1(S_0, x_0) \to \mathbb{Z}$ given by the algebraic intersection multiplicity of a loop with γ. Taking the kernel of the composition of m_γ with the quotient map $\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$, we obtain a cover $p_n : S_n \to S_0$ whose deck transformation group is $\mathbb{Z}/n\mathbb{Z}$. One can cut S_n along two opposing lifts of γ to create two identical halves (see Figure 2 for the case $n = 4$). A test function, which on one half is $+1$ and on the other is -1 with a transition in between, will produce a smaller and smaller Rayleigh quotient (7) as n gets large. For more details on this geometric procedure we refer the reader to [4, Proposition 3.39] or to the original result of Randol [112], obtained by use of the Selberg trace formula.

In the setting of finite volume non-compact surfaces, Selberg [129] constructed a tower of cyclic covers of the modular surface having arbitrarily small λ_1. In Selberg's examples, the eigenfunctions realizing λ_1 arise as residues of Eisenstein series and so are not cuspidal. Later, Zograf [153, Theorem 4] showed that for these same examples, one can actually find cusp forms realizing λ_1. In any case, the existence of cyclic covers can be seen algebraically from the fact that $\mathrm{PSL}_2(\mathbb{Z})$ contains a finite index subgroup Γ^* isomorphic to the free group on two generators.[6] As Γ^* is the fundamental group of $\Sigma = \Gamma^* \backslash \mathbb{H}$ and $\pi_1(\Sigma)$ maps surjectively via its

[6] The notation Γ^* is Selberg's. In fact Γ^* is the image in $\mathrm{PSL}_2(\mathbb{Z})$ of $\Gamma(2)$, the principal congruence subgroup consisting of integer matrices congruent to the identity mod 2.

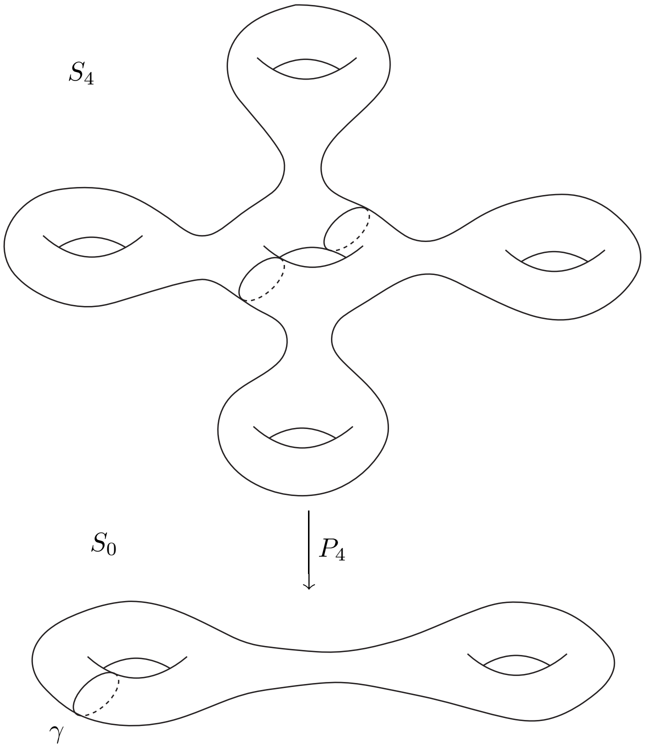


Figure 2. A cyclic cover

Abelianization to $H^1(\Sigma, \mathbb{Z}) \simeq \mathbb{Z}$, one again obtains cyclic covers by taking the kernels of the composition with $\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ (see [129, (2.6)]).

The above examples put into stark relief the uniformly rigid behavior of the tower of congruence surfaces $Y_1(N)$ predicted by the Selberg eigenvalue conjecture. They suggest that the geometry of $Y_1(N)$ should be asymptotically much different from cyclic covers, whose highly radial structure we saw was obtained by a series of regular identifications. What makes the surfaces $Y_1(N)$ extraordinary is that they are, in the words of Robert Brooks [15], *short and fat* and possessed of *highly interesting symmetries*. Consider the closely related surface $Y(N) = \Gamma(N)\backslash\mathbb{H}$, where $\Gamma(N)$ is the kernel of the reduction mod N map on $\mathrm{SL}_2(\mathbb{Z})$. If one were to model $Y(N)$ by the Cayley graph of $\Gamma(N)$ (with respect to a given choice of generators), "short and fat" would be expressed as the graph being sparse and highly connected, which is the colloquial definition of an expander graph [124]. A large group of "interesting symmetries", on the other hand, means that the first non-trivial representation of the group of deck transformations $\mathrm{SL}_2(\mathbb{Z})/\Gamma(N)$ is of high dimension relative to its order. This is clearly not the case for cyclic covers, which, being Abelian, admit only one-dimensional representations. The influential work of Sarnak and Xue [126], based on ideas of Kazhdan, makes use of this latter property to obtain lower bounds on λ_1 for congruence towers of arithmetic lattices.

While the Selberg eigenvalue conjecture is wide open, it has been verified for small levels. Let $\lambda_1^{\mathrm{disc}}$ denote the smallest positive eigenvalue in the discrete spectrum. Using the variational characterization and elementary arguments, Roelcke

[114] showed that for the modular surface $\lambda_1^{\rm disc}({\rm SL}_2(\mathbb{Z})\backslash\mathbb{H}) > 1/4$. In a 1985 article [61], Huxley confirmed the Selberg eigenvalue conjecture for $N \leqslant 18$, showing the strict inequality $\lambda_1^{\rm disc}(Y_1(N)) > 1/4$ for these N. For quite some time, no progress was made towards extending this range of N, for whenever Huxley's technique worked it necessarily gave a strict inequality. In Section 2.3 we showed that for N large enough (e.g. $N = 23$) there exist Maaß forms of eigenvalue $1/4$, so the Selberg conjecture, if true, is sharp. Recently Booker and Strömbergsson [13] confirmed that $\lambda_1^{\rm disc}(Y_1(N)) \geqslant 1/4$ for square-free $N \leqslant 857$, using methods strong enough to detect the presence of eigenvalue $1/4$ Maaß forms.

3.3. **Dynamical reformulation.** It is especially interesting to recast the Selberg and the Ramanujan–Petersson conjectures for Maaß forms in the context of effective equidistribution results. Although there exist quite a few ways to do so (effective equidistribution of the Hecke correspondences $\mathcal{C}_p(z)$ as p gets large [31, 32, 120], effective ergodic theorem for ${\rm SL}_2(\mathbb{R})$ [55]), the most pertinent to our discussion involves the equidistribution of segments of long closed horocycles.

For our purposes, a *horocycle* on $Y_1(N) = \Gamma_1(N)\backslash G/K$ will be the projection from the homogeneous space $\Gamma_1(N)\backslash G$ of an orbit of the upper triangular unipotent subgroup

$$U = \left\{ u_x = \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} : x \in \mathbb{R} \right\}.$$

The action of U on $\Gamma_1(N)\backslash G$ is by right multiplication. A horocycle is *closed* if it comes from a periodic U orbit. If $y \in \Gamma_1(N)\backslash G$ lies on a periodic U orbit, there exists some minimal $\ell > 0$ such that $y.u_\ell = y$. Closed horocycles are associated to the cusps of $Y_1(N)$. For example, if

$$y_t = \Gamma_1(N) \begin{pmatrix} t^{1/2} & 0 \\ 0 & t^{-1/2} \end{pmatrix},$$

then putting $\ell = 1/t$ one has $y_t.u_\ell = y_t$ as elements in $Y_1(N)$. The projection down to $Y_1(N)$ of this periodic U orbit $y_t.U$ is a closed horocycle that we shall denote by H_t. Equivalently, H_t is the projection down to $Y_1(N)$ of $\{x + it \mid x \in \mathbb{R}\} \subset \mathbb{H} = G/K$. This particular collection of closed horocycles $\{H_t : t > 0\}$ is associated to the cusp at infinity of $Y_1(N)$.

We return to a general case of a periodic U orbit $y.U$ of period ℓ. For any subinterval $I \subset [0, \ell]$ we consider the probability measure

$$\nu_I(f) = \frac{1}{|I|} \int_I f(y.u_x)dx$$

integrating a test function $f \in C_c(\Gamma_1(N)\backslash G)$ over a segment of the orbit. We translate this measure orthogonally by the group of diagonal matrices

$$A = \left\{ a_t = \begin{pmatrix} t^{1/2} & 0 \\ 0 & t^{-1/2} \end{pmatrix} : t > 0 \right\},$$

putting

$$\nu_I^t(f) = \frac{1}{|I|} \int_I f(y.u_x a_t)dx.$$

For any fixed interval I, as $t \to 0$, the translated segment of the horocycle grows longer and longer. For instance, the horocycle H_1 is mapped under a_t to H_t. Figure 3 shows the image of $H_{1/100}$ on the standard fundamental domain of the modular surface ${\rm SL}_2(\mathbb{Z})\backslash\mathbb{H}$.

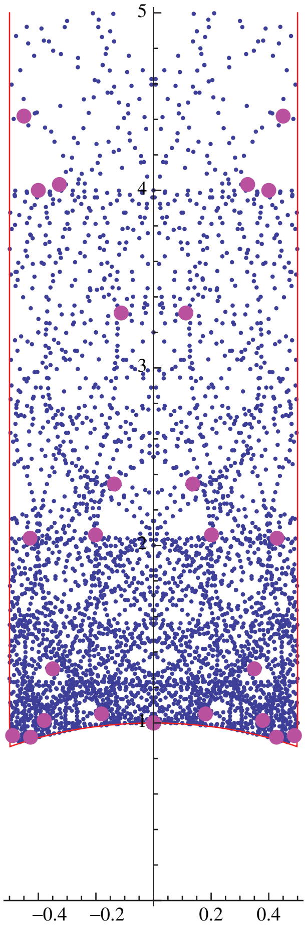


FIGURE 3. 5000 sample points of the horocycle $H_{1/100}$; special points $\pm 2/p + i/100$ for primes $2 < p < 100$

It is reasonable to expect that H_t should become equidistributed on $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ as $t \to 0$, with respect to the hyperbolic measure. More generally, ν_I^t should tend to the uniform measure m. When $I = [0, \ell]$, this is in fact a well-known result of Sarnak [119]. Improving upon this, Strömbergsson [140] allowed the interval I to shrink with t. He showed that if the Selberg eigenvalue conjecture is true then

$$\left|\nu_I^t(f) - \int_{\Gamma_1(N)\backslash G} f dm\right| = O_{f,\varepsilon}(N^A |I|^B t^{1/2-\varepsilon})$$

for every $f \in C_c^\infty(\Gamma_1(N)\backslash G)$ and sufficiently large constants A, B, as long as $|I| > t^{1/2}$. This implication can be reversed; the equivalence of this statement with the Selberg eigenvalue conjecture is reviewed in forthcoming lecture notes by the second author and Akshay Venkatesh [18].

3.4. **The role of Kloosterman sums.** In his famous 1926 paper [78], Kloosterman introduced a certain type of exponential sum which arises naturally in the study of the representation number of integers by quaternary quadratic forms. His method was based upon the Hardy–Littlewood circle method, which breaks up an integral over the unit circle into pieces concentrated around rational numbers according to the size of their denominator. Kloosterman's decisive contribution was to make this decomposition more precise by eliminating overlapping pieces and then collecting those pieces of same size. Kloosterman's version of the circle method, usually referred to as the *Kloosterman refinement*, led to the introduction of what are now called *Kloosterman sums*. For integers m, n and a natural number c, the Kloosterman sum is defined as

$$S(m,n,c) = \sum_{x \in (\mathbb{Z}/c\mathbb{Z})^\times} \exp\left(2\pi i \frac{mx + nx^{-1}}{c}\right). \tag{8}$$

Bounding these sums non-trivially was the key to the solution to Kloosterman's problem of determining the asymptotics of representation numbers of positive quaternary quadratic forms. Clearly, one has $|S(m,n,c)| \leqslant c$; Kloosterman obtained $S(m,n,c) = O(c^{3/4+\varepsilon})$, when $(m,n) = 1$.

This refined circle method approach turned out to be fruitful in the theory of holomorphic cusp forms. Soon after Hecke proved his bound $\lambda_f(n) = O(n^{1/2})$—the simplest approximation towards the Ramanujan conjecture—from quite general principles, Kloosterman [79] was able to improve this to $\lambda_f(n) = O_\varepsilon(n^{3/8+\varepsilon})$. Estermann [47] and Salié [118] then completely clarified this connection between non-trivial bounds for Kloosterman sums and bounds for Fourier coefficients of holomorphic cusp forms of integer weight $k \geqslant 2$. They showed that a bound of the form $c^{1-\delta}$ for (8) implies that $\lambda_f(p) = O(p^{(1-\delta)/2})$.

Finally, using algebro-geometric techniques, Weil [149] established the optimal bound $|S(m,n,c)| \leqslant d(c)\sqrt{c}$, for $(m,n) = 1$, from which it follows that $\lambda_f(p) = O(p^{1/2-1/4})$. These historical developments were sketched in great clarity already by Selberg in [129]. There he introduced new techniques as a way to bound the low weight automorphic spectrum, and, more critically, applied them to bound from below the Laplacian eigenvalue of weight zero Maaß forms. His celebrated result[7] is that $\lambda_f(\infty) \geqslant 3/16$; writing $\lambda_f(\infty) = s(1-s)$, this is equivalent to $|\mathrm{Re}(s)| \leqslant 1/2 - 1/4$, which visibly puts Selberg's result in parallel with that of Kloosterman and Weil.

Petersson [105] found a fruitful approach to this connection using the theory of Poincaré series. His investigations would eventually lead to the spectral relation [63, 107] that is now referred to as the Petersson trace formula. A generalization of this formula due to Bruggeman [16] and Kuznetsov [82] can be used to derive all of the above bounds in a unified way; see e.g. [94, Remark 2.3]. The trace formula approach is essentially equivalent to Selberg's: both rely on the spectral theory of the Laplacian on $L^2(Y_1(N))$, the key point being that $\lambda_f(v)$, for v an arbitrary place of $\mathbb{Q}$, can be written as a sum of Kloosterman sums. It bears mentioning, however, that the original derivation of bounds on Fourier coefficients of cusp forms by Kloosterman and Estermann, directly from non-trivial bounds on Kloosterman sums, continues to be of relevance. Indeed, it is this approach which most clearly reveals the geometric significance of Kloosterman sums. We explain briefly this point of view, which is elaborated in the notes [18].

[7] Selberg's 3/16 theorem is probably the most typographically hidden of any mathematical result of comparable renown. In the original typesetting, it is buried within an expository paragraph and appears directly after a page turn.

The geometric link between the Kloosterman refinement and bounds towards the Ramanujan conjecture for $\mathrm{GL}_2/\mathbb{Q}$ is made through the identification of the circle S^1 with a low closed horocycle H_t in $Y_1(N)$. We have already hinted at this in the dynamical reformulation of the Selberg conjecture in the previous section, where the notation H_t was introduced. At the finite places, this link is even clearer, for the circle method extracts the nth Fourier coefficient as an integral against a fixed additive character over $H_{1/n}$. For example, if f is a weight k holomorphic form and $\tilde{f}$ is the lift of f to $\Gamma_1(N)\backslash G$ given by

$$\tilde{f} : \Gamma_1(N) \begin{pmatrix} a & b \\ c & d \end{pmatrix} \mapsto f\left(\frac{ai+b}{ci+d}\right)(ci+d)^{-k},$$

then

$$a_f(n) = n^{-1} \int_0^n \tilde{f}(y_{1/n}.u_x)e(-x)dx.$$

Consider a point $z = p/q + in^{-1}$ mapping to the horocycle $H_{1/n}$, whose x-coordinate is rational of denominator $q \leqslant n$. It is mapped into the standard fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$ to the point $\bar{p}/q + in/q$, where $\bar{p}$ is the multiplicative inverse of p mod q. Thus, when one restricts to rational points on the circle of bounded height, one obtains a finite collection of points in the fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$ whose y-coordinate is bounded and whose x-coordinate varies rather erratically. It is this *arithmetic randomness* that bounds on Kloosterman sums encode (see Figure 3). Now, if the rational points along the long closed horocycle appear randomly in the quotient $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$, then the horocycle itself can be shown to equidistribute, as in the previous section. As cusp forms are orthogonal to constants, this is enough to prove bounds towards the Ramanujan conjecture at either finite or infinite places.

3.5. **The Rankin–Selberg method.** In Selberg's article [129], a great deal of attention is paid to the relative merits of two different methods for the estimation of Fourier coefficients. The first method was sketched above, in both its spectral and geometric incarnations, and it will be revisited in Section 9.1. The second, which we shall discuss in detail in Section 7, now goes under the name of the Rankin–Selberg method. It is a method which uses the machinery of L-functions; Kloosterman sums, at least on the surface, do not intervene.

At its heart, the Ramanujan conjecture is a statement about the purity of certain critical exponents. While geometric methods are enlightening, to reach these conjectural exponents one must have recourse to an iterative process, so that weak bounds can be bootstrapped to better ones. This is provided by automorphic forms on higher rank groups, the theory of functoriality, and L-functions such as those introduced by Rankin and Selberg. The next few sections are intended to widen the discussion to include these more powerful tools. With that said, we will eventually reveal (in Section 7) the role that certain generalized Kloosterman sums play in the method of Rankin and Selberg: there is more unity to the various analytic approaches to Ramanujan than once thought!

4. The Ramanujan conjecture for GL_n

We now pass to the general setting of the group GL_n, defined over a number field F. We refer to [54] for an introduction to this theory and some relevant notation. A dictionary between classical language and adelic language can be found in [49].

Let $\mathbb{A}_F$ be the ring of adeles of F. Fix a unitary character ω of $Z(F)\backslash Z(\mathbb{A}_F)$, where Z is the center of GL_n. Let $L^2(\mathrm{GL}_n(F)\backslash\mathrm{GL}_n(\mathbb{A}_F),\omega)$ denote the space of left $\mathrm{GL}_n(F)$-invariant, square integrable functions transforming under the center by ω. The group of adelic points $\mathrm{GL}_n(\mathbb{A}_F)$ acts on this space by right translation, preserving the natural Haar measure. The resulting unitary representation is highly reducible, and as a first decomposition we can write it as a direct sum of the continuous and discrete spectrum. A *cuspidal automorphic representation* on $\mathrm{GL}_n(\mathbb{A}_F)$ is an irreducible subrepresentation (π, V_π) of the discrete spectrum such that the period integral

$$\int_{U(F)\backslash U(\mathbb{A}_F)} \pi(u).\xi \, du$$

vanishes for almost every $\xi \in V_\pi$, where U is the unipotent radical of any proper parabolic subgroup of GL_n.

A cuspidal automorphic representation π factorizes [48] as a restricted tensor product $\pi \simeq \bigotimes \pi_v$, where π_v—an irreducible unitary representation of $\mathrm{GL}_n(F_v)$—is unramified for all but a finite number of places v. The Ramanujan conjecture for GL_n, first formulated by Langlands [85, §8], states that the local components π_v appearing in a cuspidal automorphic representation π satisfy a certain analytic condition known as *temperedness*.[8] We review the definition of temperedness below and record several “first results” towards the Ramanujan conjecture in this degree of generality.

4.1. **Tempered representations.** By definition, if K_v is a maximal compact subgroup of $\mathrm{GL}_n(F_v)$, an irreducible unitary representation (π_v, V_{π_v}) of $\mathrm{GL}_n(F_v)$ is *tempered* if for every K_v-finite vector $\xi_v \in V_{\pi_v}$ the associated matrix coefficient $\langle \pi_v(g).\xi_v, \xi_v\rangle$ is in $L^{2+\varepsilon}(\mathrm{PGL}_n(F_v))$ for any $\varepsilon > 0$. In other words, π_v should be weakly contained in the regular representation.

Temperedness is a concept of immense importance in the subject of automorphic forms, especially as concerns the Ramanujan conjecture. In light of this, we take some time to review below several equivalent formulations of the definition.

Denote by Ξ_v the Harish-Chandra function on $\mathrm{PGL}_n(F_v)$. It is defined as the unique bi-K_v-invariant function in $I(1)$ whose value at the identity is 1. (Here, $I(\chi)$ is unitary induction from the upper triangular Borel subgroup to $\mathrm{PGL}_n(F_v)$.) A result of Harish-Chandra shows that $\Xi_v \in L^{2+\varepsilon}(\mathrm{PGL}_n(F_v))$ for every $\varepsilon > 0$. Let $p \geqslant 1$. Then it is known that (cf. [36], [104])

(1) $\langle \pi_v(g).\xi_v, \xi_v\rangle \in L^{2p+\varepsilon}(\mathrm{PGL}_n(F_v))$ for all $\varepsilon > 0$ and all K_v-finite vectors $\xi_v \in V_{\pi_v}$
if and only if

(2) $|\langle \pi_v(g).\xi_v, \xi_v\rangle| \leqslant \dim\langle K_v.\xi_v\rangle \Xi_v(g)^{1/p}$ for all K_v-finite unit vectors $\xi_v \in V_{\pi_v}$ and all $g \in \mathrm{PGL}_n(F_v)$.

This gives an equivalent formulation of temperedness when we put $p = 1$. We say that a unitary representation of $\mathrm{GL}_n(F_v)$, which is not necessarily irreducible but whose center nevertheless acts by a character, possesses a *spectral gap* if there exists a finite $p \geqslant 1$ such that one of the above equivalent conditions holds.

For generic representations π_v, we can replace the matrix coefficients $\langle \pi_v(g).\xi_v, \xi_v\rangle$ with Whittaker functions. If ψ_v is a non-degenerate character of the

[8] Sarnak [123] has an eloquent colloquial formulation of the conjecture: *only tempered representations are excited arithmetically.*

standard unipotent upper triangular subgroup $N(F_v)$, then π_v is called *generic* if there exists a non-zero intertwining operator between π_v and a subspace $\mathcal{W}(\pi_v, \psi_v)$ of the space $\mathcal{W}(\psi_v)$ of functions on $\mathrm{GL}_n(F_v)$ transforming under left multiplication by $N(F_v)$ by ψ_v. If $\pi = \bigotimes_v \pi_v$ is a cuspidal automorphic representation of $\mathrm{GL}_n(\mathbb{A}_F)$, then π_v is generic for every v. In this case, the above properties can be reformulated with the functions $W \in \mathcal{W}(\pi_v, \psi_v)$ replacing matrix coefficients.[9]

4.2. **Local parameters.** We now describe a more structural way to measure how far a given generic irreducible unitary representation π_v of $\mathrm{GL}_n(F_v)$ is from being tempered.

4.2.1. *Langlands quotients.* The idea here is to realize π_v as an induced representation from "twisted" tempered representations. This is commonly referred to as a Langlands quotient, although in this setting, due to our genericity assumption, no quotienting procedure is necessary. More precisely, there exists

(1) a parabolic subgroup P of $\mathrm{GL}_n(F_v)$, with Levi decomposition $P = MU$ where
$$M \simeq \mathrm{GL}_{n_1} \times \cdots \times \mathrm{GL}_{n_r},$$
(2) tempered representations τ_i on each block GL_{n_i}, and
(3) real numbers $\sigma_\pi(v, i)$,

such that π_v is isomorphic to the induced representation from M to $\mathrm{GL}_n(F_v)$ of $\bigotimes \tau_i[\sigma_\pi(v, i)]$, where $\tau[\sigma] = \tau|\det|^\sigma$. It follows from the unitarity of π_v that $\{\sigma_\pi(v, i)\} = \{-\sigma_\pi(v, i)\}$. As will become apparent later, this inductive description is all that is necessary to understand the region of holomorphy of local L-functions.

As an example, let $F_v = \mathbb{R}$, $n = 2$, and take π_v to be a discrete series representation. Then π_v is tempered (since its matrix coefficients are in fact square integrable), and the corresponding induction data is trivial: the parabolic P is the entire group $\mathrm{GL}_2(F_v)$, $i = 1$, $n_1 = n = 2$, $\tau_1 = \pi_v$ and $\sigma_\pi(v, 1) = 0$. The same description holds for tempered principal series representations, as well as the special and supercuspidal representations of $\mathrm{GL}_2(\mathbb{Q}_p)$, both of the latter being square-integrable.

On the other hand, if π_v is a complementary series representation of $\mathrm{GL}_2(F_v)$, then the parabolic P is the standard Borel upper triangular subgroup and $\sigma_\pi(v, 1)$, $\sigma_\pi(v, 2) \in (-1/2, 1/2)$ are the (non-zero) real parts of the inducing characters. The closer the $\sigma_\pi(v, i)$'s are to 0, the closer π_v is to being tempered. In fact, $\max_{i=1,2} |\sigma_\pi(v, i)|$ is simply $(1/2)(1 - 1/p)$, for the real number $p \geqslant 1$ appearing in the equivalence of Section 4.1. In particular, the right-regular representation of $\mathrm{PGL}_2(\mathbb{R})$ on $L^2_0(\Gamma\backslash\mathrm{PGL}_2(\mathbb{R}))$, for Γ a lattice in $\mathrm{PGL}_2(\mathbb{R})$, has a spectral gap if and only if there exists a $\delta > 0$ such that no complementary series representations with $\max_{i=1,2} |\sigma_\pi(v, i)| > 1/2 - \delta$ are weakly contained in it.

[9] We mention this reformulation only for perspective, especially as concerns the remarks in Section §8.3. While completely natural, it is a highly non-trivial statement. It was first conjectured by Lapid and Mao [87, Conjecture 3.5], and very recently proved (over non-Archimedean fields) independently by Delorme [41, Théorème 8] and Sakellaridis and Venkatesh [117, Corollary 6.3.4]. The Archimedean case was established earlier in the book of Wallach [148, Chapter 15].

4.2.2. *Satake parameters.* For all but a finite number of places v, the local component π_v is unramified. In this case, π_v can be realized as an induced representation $I(\chi)$, where χ is a character of the diagonal torus, viewed as a character of the Borel subgroup of upper triangular matrices by trivially extending it over the unipotent. Obviously, χ can be regarded as an ordered collection of n unramified characters of $F_v^\times$: $\chi_i = |\cdot|_v^{\mu_\pi(i,v)}$, $i = 1, \ldots, n$. When v is finite, and ϖ_v is a uniformizing element, the n complex numbers $\chi_i(\varpi_v)$, often written $\alpha_\pi(v, i)$, are called the Satake parameters of π_v. One can alternatively think of the complex numbers $\alpha_\pi(v, i)$ as the eigenvalues of a semi-simple conjugacy class within $\mathrm{GL}_n(\mathbb{C})$. To work uniformly with all places v, whether finite or infinite, it is preferable to use the parameters $\mu_\pi(v, i)$ appearing in the exponents. In any case, the unramified representation π_v is tempered if and only if all of the $\mu_\pi(v, i)$ are purely imaginary, an equivalence originally observed by Satake for finite places. (For finite places $\mathfrak{p}$, that $\Re\mu_\pi(\mathfrak{p}, i) = 0$ is the same as to say $|\alpha_\pi(\mathfrak{p}, i)| = 1$.) In fact, when π_v is unramified yet not tempered, the real parts $\mu_\pi(v, i)$ are precisely the parameters $\sigma_\pi(v, i)$ encountered above.

Let us see how the above definitions recover more familiar objects in the classical case of $F = \mathbb{Q}$ and $n = 2$. Let π be an everywhere unramified cuspidal automorphic representation on $\mathrm{GL}_2(\mathbb{A}_\mathbb{Q})$ with trivial central character. Then π is generated by a weight 0 cuspidal Hecke–Maaß newform f for $\mathrm{SL}_2(\mathbb{Z})$. Denote by $\lambda_f(n)$ its Hecke eigenvalues and by $\lambda_f(\infty)$ its Laplacian eigenvalue. The connection between these eigenvalues and the parameters $\mu_\pi(v, i)$ is given by

$$\begin{cases} \lambda_f(p) = p^{\mu_\pi(p,1)} + p^{\mu_\pi(p,2)} = p^{\mu_\pi(p,1)} + p^{-\mu_\pi(p,1)}, & v = p \text{ a prime,} \\ \lambda_f(\infty) = \frac{1}{4} - \mu_\pi(\infty, 1)^2 = \frac{1}{4} - \mu_\pi(\infty, 2)^2, & v = \infty. \end{cases}$$

In both cases we see directly that $\Re\mu_\pi(v, i) = 0$ implies the classical Ramanujan conjecture and the Selberg eigenvalue conjecture.

4.3. **Local bounds.** The first general result towards the Ramanujan conjecture for GL_n is due to Jacquet and Shalika [67]. For any place v of F they established the bound

$$\max_i |\sigma_\pi(v, i)| < \frac{1}{2}. \tag{9}$$

The proof uses only local methods: they exploit the genericity of π_v to show that the local Rankin–Selberg L-function[10] $L(s, \pi_v \times \tilde{\pi}_v)$ is holomorphic on $\Re s > 1$ and regular on $\Re s = 1$. Since for $\Re s$ large enough one has the factorization

$$L(s, \pi_v \times \tilde{\pi}_v) = \prod_{i,j} L(s + \sigma_i - \sigma_j, \tau_i \times \tilde{\tau}_j), \tag{10}$$

the product running over the tempered representations appearing in the induced description of Section 4.2.1, the bounds (9) then follow from the fact that for τ_v, τ_v' tempered the local L-factor $L(s, \tau_v \times \tau_v')$ is holomorphic on $\Re s > 0$.

When (9) is applied to cuspidal automorphic representations of $\mathrm{GL}_2(\mathbb{A}_\mathbb{Q})$, one recovers the trivial Hecke bound $|\lambda_\pi(p)| < p^{1/2} + p^{-1/2}$ (cf. Section 3.4) at finite

[10] One can in fact bypass the appeal to local L-functions and simply read off the above bound by the classification of the generic dual of $\mathrm{GL}_n(F_v)$ inside the unitary dual.

places and the trivial positivity bound $\lambda_\pi(\infty) > 0$ at infinity. We note, however, that although the proof only takes into consideration local information,[11] the bounds of Jacquet and Shalika are nonetheless non-trivial with respect to the *unitary* dual for $n \geqslant 3$. To see this it suffices to observe that the trivial representation of $\mathrm{GL}_n(F_v)$ is induced by a character having maximal parameter equal to $(n-1)/2$. This reflects the well-known phenomenon of automatic spectral gap in higher rank. Indeed, when the equivalence of Section 4.1 is applied to the unitary representation of $\mathrm{PGL}_n(\mathbb{A}_F)$ acting on the *cuspidal* spectrum $L^2_{\mathrm{cusp}}(\mathrm{PGL}_n(F)\backslash\mathrm{PGL}_n(\mathbb{A}_F))$, the Jacquet–Shalika bounds give the value $p = 1 + 1/(n-2)$ at every place v.

A striking application of the bounds (9) can be found in the book by Harris and Taylor [57]. Their Corollary VI.2.7 employs the Jacquet–Shalika bounds together with a purity statement (their Proposition III.2.1) to deduce that certain discrete series representations cannot appear in the cohomology of their unitary Shimura varieties outside of the middle dimension. This idea dates back to Drinfeld's proof of the Ramanujan conjecture for GL_2 over a function field [56, §2.2]. The point is that between the exponents $-1/2$ and $1/2$ given by (9) the only integer exponent that can appear is 0. This is just one of the ingredients in the proof by Harris and Taylor, building upon the work of many authors and completed recently by Shin [137] and Caraiani [25], that if

(1) F is a CM field,
(2) π is a cuspidal automorphic representation of $\mathrm{GL}_n(\mathbb{A}_F)$, for $n \geqslant 3$,
(3) $\tilde{\pi} \simeq \pi \circ c$, where c is complex conjugation, and
(4) π_∞ is regular and algebraic,

then π corresponds to a compatible system of continuous ℓ-adic n-dimensional Galois representations, unramified at all but finitely many places, and de Rham at ℓ. It can be shown that such π satisfy the Ramanujan conjecture. Moreover, from this theorem one can deduce an analogous theorem [144, Theorem 3.6] for totally real number fields F. This body of work represents the most complete generalization of Deligne's work to GL_n.

4.4. **Global bounds.** Despite the appearance of the bounds (9) in the work of Harris and Taylor, more often than not these bounds fall just short of what is needed in a given problem. For instance, in the case $n = 2$ it just fails to give any non-trivial information such as spectral gap for the right regular representation of $\mathrm{PGL}_2(\mathbb{A}_F)$ on $L^2_0(\mathrm{PGL}_2(F)\backslash\mathrm{PGL}_2(\mathbb{A}_F))$, the L^2-subspace orthogonal to constants. In certain situations, any better bound, even a marginally better one, serves as a key ingredient for deep and sometimes unexpected applications.

The breakthrough which led to an improvement upon (9) was the development of the global theory of the Rankin–Selberg L-function on $\mathrm{GL}_n \times \mathrm{GL}_n$ by Jacquet, Piatetski-Shapiro, and Shalika, in the early 1980s. When combined with classical techniques for arithmetic functions whose associated Dirichlet series satisfy certain functional equations [27, 84], this allowed for an improvement over (9) in many cases. This was first observed by Serre [131] and can be viewed as a far-reaching generalization of Rankin's method [113].

[11] To apply these bounds, one does make use of the critical (global) assumption that π is a cusp form. Residual Eisenstein series (such as the trivial representation) are not globally generic, and so their local components cannot all be generic.

While a striking application of the Rankin–Selberg theory, the above improvement suffered two drawbacks. The first is that the method only seemed to work at the finite places. Secondly, the resulting bound depended on the number of Γ-factors in the functional equation and hence deteriorated in quality in the degree of the number field (see (21) below). These issues were later resolved by Luo, Rudnick, and Sarnak [91, 92] who gave what is still the most comprehensive result. They found a new way to apply Rankin–Selberg L-functions which allowed them to prove the same bounds at *all* (unramified) places—including the elusive Archimedean places, producing the first ever improvement on Selberg's 3/16 bound—and over a general number field, without loss of quality in the degree of the number field. Their techniques were extended to ramified places independently by Bergeron and Clozel [5, Théorème 7.0.2] and Müller and Speh [98]. Taken together, the complete result is that for every place v of F one has

$$\max_i |\sigma_\pi(i,v)| \leqslant 1/2 - 1/(n^2+1). \tag{11}$$

We will describe the methods of Rankin and Serre, and Luo, Rudnick and Sarnak, as well as other refinements, in Section 7. We also provide some further discussion in Section 8.2.

We describe a recent application which uses the full strength of the bounds (11). Let F/F' be a cyclic extension of number fields of prime degree, and suppose that π and π' are cusp forms on GL_n over F whose local components coincide for almost all primes of degree 1 over F'. Such situations arise when considering the Artin conjecture over solvable extensions. Then Ramakrishnan [110] has shown π and π' are twist equivalent, that is $\pi' = \pi \otimes \chi$ for some finite order character of χ of F. The proof of this result uses the bounds (11), in particular the fact that the gain over 1/2 is independent of v *and the number field* F.

4.5. **The Ramanujan conjecture for classical groups.** Although this article focuses mainly on the group GL_n, the following section provides a short discussion on the Ramanujan conjecture for other classical groups.

The Ramanujan conjecture for quasi-split classical groups (symplectic, unitary, and orthogonal) must be stated with more care. Here one might first restrict one's attention to *globally generic* cuspidal automorphic representations. By definition, a cuspidal automorphic representation π of a connected reductive algebraic group G over F is globally generic if there exists a vector $\xi \in V_\pi$ such that

$$\int_{U(F)\backslash U(\mathbb{A}_F)} (\pi(u).\xi)\overline{\psi(u)}du \not\equiv 0,$$

for some maximal unipotent subgroup U (defined over F) and non-degenerate character ψ of $U(F)\backslash U(\mathbb{A}_F)$. As has been mentioned, all cusp forms on GL_n are globally generic (and hence everywhere locally generic), a fact which follows from the adelic Fourier–Whittaker expansion of Piatetski-Shapiro [108] and Shalika [135]. For cusp forms on other groups, this is not necessarily the case. Examples of non-generic cusp forms, such as genus 2 holomorphic Siegel modular forms on Sp_4 which are not tempered at any finite place, were discovered some time ago in [81] and [60]. Restricting to globally generic cuspidal automorphic representations $\pi = \bigotimes_v \pi_v$, the generalized Ramanujan conjecture again states that each π_v should be tempered.

The work of Cogdell, Kim, Piatetski-Shapiro, and Shahidi [35] establishes the functorial transfer of globally generic cuspidal automorphic forms from quasi-split classical groups to GL_n. Their method, a culmination of work spanning many years, couples the Langlands–Shahidi method of constructing "nice" L-functions with the converse theorem. This transfer, along with the work on descent by Ginzberg, Rallis, and Soudry [52], sets up an implication whereby one deduces the Ramanujan conjecture for generic forms on classical groups from that for GL_n. This is a concrete example of the general principle that to prove the full Ramanujan conjecture for a certain wide class of groups[12] should reduce, via functoriality and explicit cuspidality conditions, to proving it for the "mother group" GL_n. See the nice discussion of these ideas in Section 10 of [35], and in particular Corollary 10.2 there.

Recent work of Arthur [3] establishes the functorial transfer of automorphic forms from quasi-split special orthogonal and symplectic groups[13] to GL_n. Arthur's approach goes by the stable trace formula, and in particular makes critical use of the Fundamental Lemma proved by Ngô [103]. Compared to the results of [35], Arthur's work is more general and gives more precise results. In particular, he makes no assumption of global genericity: the trace formula is in a sense blind to whether or not a cusp form has a non-zero Fourier coefficient. As a consequence of Arthur's work, in the implication of the previous paragraph, one can replace the condition of being globally generic with the seemingly weaker condition of being everywhere locally generic [134]. Moreover, Arthur's work provides a formulation of the Ramanujan conjecture for the entire cuspidal spectrum of classical groups as well as a reduction of its proof to that for GL_n.

As an example of the last statement, consider the spectrum of the Laplacian Δ on real hyperbolic manifolds of congruence type. Such M are of the form $\Gamma\backslash\mathbb{H}^n$, where Γ is a congruence lattice in $\mathrm{SO}(n,1)^0$ and $\mathbb{H}^n = \mathrm{SO}(n,1)^0/\mathrm{SO}(n)$ is a model for hyperbolic n-space (the unique, up to isometry, connected simply-connected Riemannian manifold of constant negative sectional curvature -1). A conjecture of Burger, Li, and Sarnak [22] states that

$$\mathrm{Spec}_\Delta(M) \subset \bigcup_{0\leqslant j<\frac{n-1}{2}} \left\{ \left(\frac{n-1}{2}\right)^2 - \left(\frac{n-1}{2} - j\right)^2 \right\} \cup \left[\left(\frac{n-1}{2}\right)^2, +\infty \right[.$$

When the Laplacian eigenvalues are written as $s(\rho - s)$, where $s \in \mathbb{C}$ and $\rho = (n-1)/2$, then the above singletons correspond to half integer s. These are the low-energy quantum states of M; when $j \neq 0$ (so $n \geqslant 4$) they correspond to non-constant non-tempered eigenfunctions which are, by necessity, non-generic. When M is compact, as is the case when Γ comes from an F-form of $\mathrm{SO}(n,1)$ where F is a totally real number field of degree at least 2 over $\mathbb{Q}$, all non-zero eigenvalues lie in the cuspidal spectrum. The remaining continuous interval is the tempered spectrum. Using Arthur's recent work, in combination with the Burger–Li–Sarnak functorial principles, Bergeron and Clozel [6] were able to make dramatic progress on the above conjecture. They show, among other things, that $\mathrm{Spec}_\Delta(M)$ is contained in the above set of quantum states union the slightly larger continuous interval

$$\left[\left(\frac{n-1}{2}\right)^2 - \left(\frac{1}{2} - \frac{1}{N^2+1}\right)^2, +\infty \right[,$$

[12] The precise class in question here consists of the *twisted endoscopic* groups of GL_n.

[13] The transfer from special unitary groups is work in progress by members of the Paris Book Project; see [96] and [151].

where $N = n$ if n is even and $N = n+1$ if n is odd. The quantity $1/2 - 1/(N^2+1)$ comes from the bounds (11) toward the Ramanujan conjecture for the cuspidal spectrum on GL_n. The Ramanujan conjecture for GL_n is therefore shown to imply the conjecture of Burger, Li, and Sarnak.

5. Numerical improvements towards the Ramanujan conjecture and applications

It is important to observe that Langlands' functoriality conjecture for all symmetric powers implies almost trivially the Ramanujan conjecture. Let π be a cuspidal automorphic representation of $\mathrm{GL}_n(\mathbb{A}_F)$ with unitary central character. Let $r \geqslant 2$, and assume the representation $\Pi = \mathrm{sym}^r \pi$ on $\mathrm{GL}_m(\mathbb{A}_F)$ with $m = \binom{n+r-1}{r}$ is automorphic. If v is a place where π is unramified, then applying (9) on $\mathrm{GL}_{m'}$, for appropriate $m' \leqslant m$, we have

$$\max_{1 \leqslant j \leqslant n} |\Re \mu_\pi(v,j)| \leqslant \frac{1}{r} \max_{1 \leqslant j \leqslant m'} |\Re \mu_\Pi(v,j)| < \frac{1}{2r}.$$

The right-hand side can be made arbitrarily small for large r. A similar argument works for places at which π is ramified.

5.1. **Known functorial lifts and first applications.** Unfortunately, non-endoscopic functorial transfers, such as the symmetric power lifts from GL_2, are shrouded in mystery at this time. Only a few precious (and hard-earned) cases have been established. For example, one can lift forms from GL_2 using

1. the symmetric square lift from GL_2 to GL_3, established by Gelbart and Jacquet [50], building on the groundbreaking work of Shimura [136];
2. the symmetric cube lift from GL_2 to GL_4, by Kim and Shahidi [76];
3. the symmetric fourth power lift from GL_2 to GL_5, by Kim [73].

In each of these cases there is a characterization of when the lift is cuspidal. Lifts from other low rank groups include the exterior square lift from GL_4 to GL_6 by Kim [73] and Rankin-Selberg convolutions on $\mathrm{GL}_2 \times \mathrm{GL}_2$ by Ramakrishnan [109] and on $\mathrm{GL}_2 \times \mathrm{GL}_3$ by Kim and Shahidi [76].

It is expected to be quite difficult to expand the above list to include more cases. The work of Kim and Shahidi relies crucially on the Langlands–Shahidi method of constructing L-functions through the constant term of Eisenstein series. For example, for sym^3 they use the Eisenstein series on the exceptional groups E_6 and E_7. The list of exceptional Lie groups being finite, completely new ideas will be needed to go further.

As an immediate consequence of these functorial lifts, and the corresponding cuspidality conditions, one can dramatically improve the local bounds of (9). For example, assume π is not a dihedral form, since otherwise (cf. §2.3) the Ramanujan conjecture is already known to hold. If we use no more than the Gelbart–Jacquet lift, then $\mathrm{sym}^2(\pi)$ is cuspidal and, applying the local bound (9) to $\mathrm{sym}^2(\pi)$ on GL_3, we gain by a factor 2 over the same bound for GL_2. Indeed, one recovers $\max_{i=1,2} \Re\mu_\pi(v,i) < 1/4$, which strengthens Selberg's bound $\lambda_\pi(\infty) \geqslant 3/16$ to a strict inequality. Now if we use all of the above functorial lifts, one can do much better, for assuming that $\mathrm{sym}^4(\pi)$ is cuspidal, one immediately gains four-fold over the local bound. Indeed, applying (9) to $\mathrm{sym}^4(\pi)$ on GL_5, one obtains

$$\max_{i=1,2} |\Re \mu_\pi(v,i)| < 1/8. \tag{12}$$

This strengthens Selberg's bound to $\lambda_\pi(\infty) > 15/64$, a substantial improvement.

5.2. **More refined results.** The bounds (12) illustrate clearly the role of functoriality: one bootstraps strong bounds in lower rank from modest bounds in higher rank. As good as these bounds are, the theory of (global) L-functions allows us to squeeze a bit more from the functorial lifts, and to slightly improve the above exponents. Indeed, instead of applying the local bound (9) in the above arguments, we could have of course appealed to the stronger global bound of Luo, Rudnick, and Sarnak (11), which itself is a consequence of the global theory of Rankin–Selberg L-functions. Doing so for the symmetric fourth power lift of a GL_2 form yields, for example, an improvement of the above 1/8 exponent to 3/26.

There is perhaps no better illustration of the fundamental role of L-functions in this subject than the observation (due to Langlands) that the absolute convergence of $L(s, \pi, \mathrm{sym}^k)$ on $\Re s > 1$ for all $k \geq 2$ implies the Ramanujan conjecture for GL_2. See [99, p. 524] for a short proof of this derivation. Given this, it should be expected that the more one knows of the analytic properties of any given symmetric power L-function of a cusp form, the better the bounds one can hope to prove on its local parameters. As a case in point, Kim and Shahidi [75] showed that for a GL_2 cusp form π the (partial) Langlands Euler product $L^S(s, \pi, \mathrm{sym}^9)$, where S is a finite set of places containing all Archimedean places and all ramified primes for π, has a meromorphic continuation to the entire complex plane and obeys a functional equation. From this they are able to deduce that for finite v such that π_v is unramified the preceding bound of 3/26 can be replaced by 1/9. Separately, Kim [74] showed that the same bounds hold for unramified Archimedean places.

Finally, Kim and Sarnak [73] obtained yet another improvement upon the above bounds, seeming to push this circle of ideas to their limit. Let $2 \leqslant n \leqslant 4$ be an integer and π a cuspidal automorphic representation on $\mathrm{GL}_n(\mathbb{A}_\mathbb{Q})$ with unitary central character. They show that

$$\max_i |\sigma_\pi(v, i)| \leqslant \begin{cases} 7/64, & n = 2; \\ 5/14, & n = 3; \\ 9/22, & n = 4. \end{cases} \tag{13}$$

In the years since this breakthrough, it remained a mystery how to extend the methods of Kim and Sarnak to general number fields. One could prove the bound of 1/9 for all number fields, but the slightly better 7/64 was available only over $\mathbb{Q}$ (and later, imaginary quadratic fields, due to Nakasuji [102]). It is often remarked that to develop results for GL_2 over a general number field brings into play certain phenomena that appear in the setting of higher rank groups, such as GL_3. This has a formal sense, in as much as the *real* (semi-simple) rank of GL_2 over a number field F of degree $d = r_1 + 2r_2$ is $r_1 + r_2$, which is at least 2, provided F has at least two Archimedean places. We will return to this discussion in Section 8. In a recent paper [9], we were able to extend the results of Kim and Sarnak to arbitrary number fields. One of our aims in these notes is to convey some of the obstacles present in higher (global or local) rank analysis, and how we were able to overcome them for this particular problem. The heart of our discussion is contained within Section 8.

5.3. **Applications.** The progression of successively better bounds that we outlined above is a hallmark of analytic number theory. Roughly speaking, each improvement comes from a finer understanding of functorial lifts and their analytic manifestations in the realm of L-functions. Despite this alignment, many classical problems, *a priori* having nothing to do with functoriality, can be solved once a threshold exponent towards the Ramanujan conjecture has been beaten. For example, when the symmetric cube functorial lift was proven and the resulting bound of 5/24 was derived as a consequence, a whole host of applications to analytic number theory became available (see Section 8 of [76]) from the simple fact that 5/24 beat the critical exponent of 1/6. Though open to interpretation, the next critical exponent could be said to be 1/12. For example, the article [125] by Sarnak and Tsimerman, on uniform bounds of sums of Kloosterman sums, sets up an application in waiting, once 1/12 is attained. The existence of these critical exponents in applications is sometimes related to issues surrounding sharp cut-offs of counting functions and transitional behavior of special functions.

While the 7/64 bound for GL_2 in (13) cannot be said to beat any particular threshold, we mention below a few applications which rely on the fact this bound is now available over arbitrary number fields. For an exposition of the various other number theoretic applications that followed on the heels of the new instances of functoriality proven by Kim and Shahidi, see the fine lecture notes of Murty [100] and Shahidi [133].

1. Let B be a quaternion division algebra over the number field F, and put $\mathbf{G} = \mathrm{GL}_1(B)$. Let $\sigma = \bigotimes_v \sigma_v$ be an automorphic representation of $\mathbf{G}(\mathbb{A}_F) = B^\times(\mathbb{A}_F)$ of dimension greater than 1. Then the (global) Jacquet–Langlands correspondence [65] associates with σ a cuspidal automorphic representation $\pi = \mathrm{JL}(\sigma) = \bigotimes_v \pi_v$ of $\mathrm{GL}_2(\mathbb{A}_F)$ such that $\pi_v \simeq \mathrm{JL}_v(\sigma_v)$ as irreducible admissible representations, where JL_v denotes the local Jacquet–Langlands map at v. In particular $\pi_v \simeq \sigma_v$ for every v such that B_v is split and π_v is square-integrable whenever B_v is ramified. See [4, Section 8] for a highly readable account, in the classical language, of the Jacquet–Langlands correspondence.

One can apply what is known towards the Ramanujan conjecture on the GL_2 form $\pi = \mathrm{JL}(\sigma)$ to deduce bounds on the Hecke eigenvalues of σ. This is another instance, as in Section 4.5, where functorial transfer to GL_n is used to obtain bounds towards the Ramanujan conjecture for the originating group. In the present case, if, for example, F is totally real and B_v is ramified at all Archimedean places, then the resulting $\pi = \mathrm{JL}(\sigma) \simeq \bigotimes \mathrm{JL}_v(\sigma_v)$ is a holomorphic Hilbert modular form, and one can apply (the extension to F by Carayol and Blasius of) Deligne's theorem to deduce that each σ_v is tempered. If, however, there is some Archimedean place where $B_v = M_2(F_v)$, then Deligne's theorem does not apply and one can do no better than quote the 7/64 bounds (13) of Kim and Sarnak (for $K = \mathbb{Q}$) and the authors (for general F). Likewise, one can apply the Jacquet–Langlands transfer (proved by Arthur and Clozel) to GL_n from the inner form $\mathrm{GL}_r(D)$, where D is a division algebra of dimension ℓ^2 over F and $n = r\ell$, using the best known bounds towards the Ramanujan conjecture, be it (13) for $n \leqslant 4$ or (11) for $n \geqslant 5$.

Let us give a concrete illustration of the above principle. Assume that F is totally real of degree d and that the quaternion algebra B is split at exactly one Archimedean place $v_{\mathbb{R}}$. Let $\mathbf{G} = PB^\times$, the quotient of $B^\times$ by its center. Then

$\mathbf{G}(F_\infty) = (P\mathsf{H}^\times)^{d-1} \times \mathrm{PGL}_2(\mathbb{R})$, where F_∞ is the product of Archimedean completions of F and H denotes Hamilton's quaternions. An order $\mathfrak{o}$ of $B(F)$ then defines a cocompact lattice $\Gamma_\mathfrak{o}$ in $\mathrm{PGL}_2(\mathbb{R})$ by projecting $\mathfrak{o}^\times$ onto the non-compact factor $\mathrm{PGL}_2(\mathbb{R})$. From the identification $\mathbb{H} = \mathrm{PGL}_2(\mathbb{R})/\mathrm{PO}(2)$ it follows that $X_\mathfrak{o} = \Gamma_\mathfrak{o}\backslash\mathbb{H}$ is an arithmetic compact Riemann surface, otherwise known as a Shimura curve.

When interpreted classically, the Jacquet–Langlands correspondence associates with a Hecke–Laplace eigenfunction φ on $X_\mathfrak{o}$ another Hecke eigenfunction f on a non-compact congruence quotient of $\mathbb{H}^d$ whose Laplacian eigenvalue at $v_\mathbb{R}$ agrees with that of φ. In particular, one deduces from (13) applied at the place $v_\mathbb{R}$ of F that $\lambda_1(X_\mathfrak{o})$, the smallest non-zero Laplacian eigenvalue on $X_\mathfrak{o}$, is at least $1/4-(7/64)^2$.

As a secondary application, one can deduce from the above non-trivial bounds on $\lambda_1(X_\mathfrak{o})$ a non-trivial lower bound on the genus g and the gonality d of the Shimura curve $X_\mathfrak{o}$. (The gonality of a proper algebraic curve is the minimal degree of a rational map to $\mathbb{P}^1$.) Indeed, Zograf [153] shows that

$$8\pi(g+1) \geqslant \lambda_1(X_\mathfrak{o})\mathrm{vol}(X_\mathfrak{o}), \qquad 8\pi d \geqslant \lambda_1(X_\mathfrak{o})\mathrm{vol}(X_\mathfrak{o}),$$

and similar inequalities for non-compact congruence quotients. For more results along these lines see the paper by Abramovich [1].

2. Let Q be a non-degenerate quadratic form in three variables defined over a totally real number field F. Let $a \in F$, and let X denote the quasi-affine variety given by $Q(x) = a$. Let S be a finite set of places of F, containing all Archimedean places, and write $\mathcal{O}_S$ for the S-integers of F. We suppose that X is isotropic over S and that $X(\mathcal{O}_S) \neq \emptyset$. (For example, when $F = \mathbb{Q}$ the real points $X(\mathbb{R})$ define a hyperboloid in $\mathbb{R}^3$.) Define the height on its set of rational points $Q(F)$ by

$$H(x) = \prod_v \max_{1\leqslant i\leqslant 3}(1, |x_i|_v).$$

Let $x \in X(\mathbb{R})$ with $||x|| \leqslant r$, and let $\delta > 0$. Using the bounds (13) when $n = 2$, Ghosh, Gorodnik, and Nevo [51] showed that there exists $\epsilon_0 > 0$, depending only on r and δ, such that for every $\epsilon \in (0, \epsilon_0)$ one can find an S-integer point $z \in X(\mathcal{O}_S)$ satisfying

$$||x - z||_\infty \leqslant \epsilon \quad \text{and} \quad H(z) \leqslant \epsilon^{-\frac{128}{25}-\delta}.$$

Here $||\cdot||_\infty$ is the maximum norm at the Archimedean place.

We see then that the Ramanujan conjecture for GL_2 has applications to Diophantine approximation on affine homogeneous varieties. Moreover, one needs only to insert a stronger bound than the trivial 1/2 bound to obtain such a quantitative statement.

3. Most classical subconvexity results for L-functions depend, among other things, on bounds towards the Ramanujan conjecture. As an example, let F denote a totally real number field and consider the family $\pi \otimes \chi$, where π is a fixed cuspidal automorphic representation on $\mathrm{GL}_2(\mathbb{A}_F)$ and χ is a Hecke character of large conductor $\mathfrak{q}$. The Phragmén–Lindelöf convexity principle shows that the central L-value satisfies

$$L(1/2, \pi \otimes \chi) \ll_\varepsilon \mathrm{N}(\mathfrak{q})^{\frac{1}{2}+\varepsilon}$$

for any $\varepsilon > 0$. (Here the implied constant depends additionally on π, F, and the Archimedean component χ_∞, since these quantities are viewed as being fixed.) Any bound in which the exponent of 1/2 is replaced by $1/2 - \delta$, for a fixed $\delta > 0$, is called a subconvexity bound. The strongest currently known[14] result [12] states that

$$L(1/2, \pi \otimes \chi) \ll_\varepsilon \mathrm{N}(\mathfrak{q})^{\frac{1}{2} - \frac{1}{8}(1-2\theta)+\varepsilon}$$

for any $\varepsilon > 0$, where $\theta \in [0, 1/2)$ is any constant such that $\max_{i=1,2} |\Re \mu_\pi(v, i)| \leqslant \theta$ for π_v unramified. Clearly a value of $\theta = 1/2 - \delta$ for some $\delta > 0$ yields a subconvexity estimate, while the Jacquet–Shalika bound, providing $\theta = 1/2$, just fails in this regard. The full Ramanujan conjecture would give an exponent of 3/8, which is the GL_2 analogue of the classical 3/16 bound of Burgess [23] for Dirichlet L-functions.

The first subconvexity bound for the above family of L-functions was in fact obtained by Cogdell, Piatetski-Shapiro, and Sarnak [34], in the case of π holomorphic. They initiated the systematic study of subconvexity for L-functions over general number fields, motivated by the following striking application. Hilbert's eleventh problem asks which integers are integrally represented by a given n-ary quadratic form Q over a number field F. See [7] for an overview of this topic over $\mathbb{Q}$. If the form Q is binary, then it corresponds to some element in the class group of an order in a quadratic extension of F. If $n \geqslant 4$ or $n = 3$ and Q is indefinite at some Archimedean place, then work of Siegel, Andrianov, Kneser, Hsia, Schulze-Pillot, and others showed that local considerations suffice to decide which integers are represented by Q. This left positive definite quadratic forms in three variables as the sole remaining case, until it was solved over the rationals by Duke and Schulze-Pillot [45] and over totally real F in [34]. A crucial input is a subconvexity bound as above for which non-trivial progress towards the Ramanujan conjecture over number fields is essential. In spite of the fact that the automorphic representation in question is holomorphic (of weight 3/2), the argument requires information about the entire spectrum, in particular bounds towards the Ramanujan conjecture for Maaß forms.

6. L-functions

Techniques coming from the theory of L-functions are at the heart of the proofs of the best known bounds towards the Ramanujan conjecture. In this section we review some of the basic properties of automorphic L-functions attached to cusp forms and describe the properties that allow us to link their analytic properties to the bounds we seek.

6.1. Definitions and first properties.

Arguably the most important analytic object attached to an automorphic representation π is its standard L-function. This is the same L-function appearing in the discussion of subconvexity bounds in the previous section. Similarly to the Riemann zeta function, it is given, for $\Re s$ large enough, by an absolutely convergent Euler product

$$\Lambda(s, \pi) = \prod_v L(s, \pi_v)$$

[14] For $F = \mathbb{Q}$ one has the "Ramanujan-independent" bound $L(1/2, \pi \otimes \chi) \ll_\pi q^{3/8+\varepsilon}$ [11] for a Dirichlet character χ modulo q. There does not seem to be a theoretical obstacle to generalize this stronger bound to number fields, but this has not yet been worked out.

over all places v of F. The local factors $L(s,\pi_v)$ are simplest to describe at places v for which π_v is unramified. In this case, when $v=\mathfrak{p}$ is finite one has

$$L(s,\pi_{\mathfrak{p}})=\prod_{i=1}^{n}\left(1-\frac{1}{\mathrm{N}(\mathfrak{p})^{s-\mu_\pi(\mathfrak{p},i)}}\right)^{-1}=\prod_{i=1}^{n}\left(1-\frac{\alpha_\pi(\mathfrak{p},i)}{\mathrm{N}(\mathfrak{p})^{s}}\right)^{-1},$$

where N denotes the norm map on ideals and the local parameters were defined in Section 4.2, while for v infinite,

$$L(s,\pi_v)=\prod_{i=1}^{n}\Gamma_{F_v}(s-\mu_\pi(v,i)),$$

where $\Gamma_{\mathbb{R}}(s)=\pi^{-s/2}\Gamma(s/2)$ and $\Gamma_{\mathbb{C}}(s)=2(2\pi)^{-s}\Gamma(s)$. More generally, we can use the induced description by tempered representations to reduce the definition of the local L-factor to that for tempered representations, as we did earlier when describing the Rankin–Selberg local factor (10). In the present case we write

$$L(s,\pi_v)=\prod_{i}L(s+\sigma_i,\tau_i)$$

over the tempered components in the induced description of π_v in Section 4.2.1.

The analytic behavior of the L-function reflects in a very precise sense properties of the automorphic representation π. One of the first analytic properties of $\Lambda(s,\pi)$ is that it is *nice*, meaning that it admits a meromorphic (in fact, entire in this case) extension to all of $\mathbb{C}$, is bounded in vertical strips, and admits a functional equation relating s and $1-s$. Following standard practice, we write

$$L(s,\pi)=\prod_{\mathfrak{p}}L(s,\pi_{\mathfrak{p}}),$$

for the *finite part* L-function—the Euler product restricted to finite places. This can be written as an absolutely convergent Dirichlet series in the right half-plane $\Re s>1$. In some sense the L-function carries the same information as the automorphic representation π, since by an identity theorem for Dirichlet series one can recover the Satake parameters $\alpha_\pi(\mathfrak{p},i)$ at unramified primes from the analytic properties of $L(s,\pi)$, and by Strong Multiplicity One these determine[15] π. For instance, if $n=2$ and π belongs to an elliptic curve E, then the rank of E is conjectured to be equal to the order of vanishing of $L(s,\pi)$ at $s=1/2$, which is perhaps one of the most spectacular instances of what automorphic L-functions know about the underlying automorphic representation.

Although the standard L-function has pride of place in the hierarchy of L-functions, we shall be more interested in two particular higher degree L-functions: the Rankin–Selberg and symmetric square L-functions. When $\pi=\bigotimes_v\pi_v$, and $\pi'=\bigotimes_v\pi'_v$ are two cuspidal automorphic representations of $\mathrm{GL}_n(\mathbb{A}_F)$ and $\mathrm{GL}_{n'}(\mathbb{A}_F)$, respectively, the Rankin–Selberg L-function is given by the Euler product

$$\Lambda(s,\pi\times\pi')=\prod_{v}L(s,\pi_v\times\pi'_v)$$

for $\Re s$ large enough, with the reduction to tempered local components given by

$$L(s,\pi_v\times\pi'_v)=\prod_{i,j}L(s+\sigma_i+\sigma'_j,\tau_i\times\tau'_j).$$

[15] This is not true on a local level. For instance, all supercuspidal representations of $\mathrm{GL}_2(\mathbb{Q}_p)$ have the same local L-functions, but the representations themselves can be quite different.

Similar factorization identities hold for the symmetric square L-function $\Lambda(s,\pi,\mathrm{sym}^2)$.

Except when $nn' \leqslant 6$, we do not know if the Rankin–Selberg L-function $\Lambda(s,\pi\times\pi')$ defines an *automorphic* L-function, coming from a functorial automorphic lift from $\mathrm{GL}_n\times\mathrm{GL}_{n'}$ to $\mathrm{GL}_{nn'}$. Despite this, the monumental work of Jacquet, Piatetski-Shapiro and Shalika [66] shows that the Rankin–Selberg L-functions are nice, in the sense alluded to in the discussion of the standard L-function. This recovers some shadow of functoriality which may then be used to make progress towards the Ramanujan conjecture. For their part, the symmetric square L-functions proved more resistant to the explication of their analytic properties. From the work of many authors, most notably Bump and Ginzberg [21], Shahidi [132], Kim [72], and Takeda [141], one knows that $L(s,\pi,\mathrm{sym}^2)$ is nice, except possibly for some exceptional poles within the critical strip. In contrast to the Rankin–Selberg L-functions, the absolute convergence of $L(s,\pi,\mathrm{sym}^2)$ in $\Re s>1$ is only known for $n\leqslant 4$.

6.2. **Dirichlet series coefficients and Whittaker functions.** Denote the Dirichlet series expansions of the finite part L-functions as

$$L(s,\pi\times\tilde{\pi})=\sum_{\mathfrak{m}}\lambda_{\pi\times\tilde{\pi}}(\mathfrak{m})\mathrm{N}(\mathfrak{m})^{-s}$$

and

$$L(s,\pi,\mathrm{sym}^2)=\sum_{\mathfrak{m}}\lambda_{\mathrm{sym}^2\pi}(\mathfrak{m})\mathrm{N}(\mathfrak{m})^{-s}.$$

These sums run over all non-zero integral ideals $\mathfrak{m}$ of the ring of integers $\mathcal{O}_F$ of F. At finite places $\mathfrak{p}$, methods of analytic number theory attempt to bound the local parameters $\sigma_\pi(\mathfrak{p},i)$ by means of the coefficients $\lambda_{\pi\times\tilde{\pi}}(\mathfrak{p}^r)$ or $\lambda_{\mathrm{sym}^2\pi}(\mathfrak{p}^r)$. These coefficients are combinatorial expressions in the parameters $\sigma_\pi(\mathfrak{p},i)$. For instance, for $\pi_\mathfrak{p}$ unramified, one has

$$\lambda_{\pi\times\tilde{\pi}}(\mathfrak{p})=\Big|\sum_i\alpha_\pi(\mathfrak{p},i)\Big|^2$$

and

$$\lambda_{\mathrm{sym}^2\pi}(\mathfrak{p})=\sum_{i\leqslant j}\alpha_\pi(\mathfrak{p},i)\alpha_\pi(\mathfrak{p},j).$$

These last two identities reveal the essence of the analytic utility of their corresponding L-functions. In both the Rankin–Selberg and symmetric square L-functions the local parameters appear with multiplicity two within each coefficient (which is not true of, say, the standard or exterior square L-function). While only the Rankin–Selberg construction yields positivity, the symmetric square needs fewer terms ($n(n-1)/2$ rather than n^2) to define it.

One of the drawbacks of the above presentation is that it only pertains to finite places. To treat all places uniformly, one must follow a slightly more general approach. We describe this approach for Rankin–Selberg L-functions, but it applies word for word for the symmetric square. It is known that for tempered representations τ,τ' of $\mathrm{GL}_n(F_v)$ the local L-factor $L(s,\tau\times\tau')$ is holomorphic on the right half-plane $\Re s>0$. From this and the factorization of $L(s,\pi_v\times\tilde{\pi}_v)$ along tempered pairs, the following chain of implications follows:

(1) $L(s,\pi_v\times\tilde{\pi}_v)$ is holomorphic on $\Re s>2\max_i|\sigma_i|$.

(2) Bounding $\max_i |\sigma_\pi(v,i)|$ is therefore equivalent to establishing a pole-free right half-plane for $L(s,\pi_v \times \tilde{\pi}_v)$.

(3) This last property can be expressed dually, by Mellin inversion. Let $x \mapsto \lambda_{\pi_v \times \tilde{\pi}_v}(x)$ denote the unique U_v-invariant function on $F_v^\times$ whose Mellin transform at an unramified quasi-character $\chi_v = |\cdot|_v^s$, for $\Re s$ sufficiently large, equals $L(s,\pi_v \times \tilde{\pi}_v)$. Here, U_v denotes the maximal compact subgroup of $F_v^\times$. Then (2) is equivalent to determining the asymptotic behavior close to 0 of the *coefficient function* $\lambda_{\pi_v \times \pi_v}$.

At a finite place $v = \mathfrak{p}$, one has $\lambda_{\pi_\mathfrak{p} \times \tilde{\pi}_\mathfrak{p}}(x) = \lambda_{\pi \times \tilde{\pi}}(\mathfrak{p}^r)$ for $|x|_\mathfrak{p} = \mathrm{N}(\mathfrak{p}^r)$, whence the name. We see then that the location of poles of $L(s,\pi_v \times \tilde{\pi}_v)$ is, via Mellin inversion, equivalent to the asymptotic behavior of $\lambda_{\pi_v \times \tilde{\pi}_v}(x)$, as $x \to 0$. When $v = \mathfrak{p}$ this amounts to bounding $\lambda_{\pi \times \tilde{\pi}}(\mathfrak{p}^r)$ for $r \to \infty$. The interest in the above formulation is in its applicability at every place, Archimedean or not.

6.3. **Relation to Whittaker functions.** We have now translated the problem of deducing bounds towards the Ramanujan conjecture to a problem related to certain local L-functions. Since the discussion of temperedness in Section 4.1 was in terms the asymptotic behavior of matrix coefficients (or Whittaker functions), it seems appropriate to understand the link between the two languages.

We begin by reviewing this relation in the simplest case, for the local standard L-function on GL_2 over $F = \mathbb{Q}$. Similarly to the above, denote by $x \mapsto \lambda_{\pi_v}(x)$ the unique U_v-invariant function on $\mathbb{Q}_v^\times$ whose Mellin transform at an unramified quasi-character $|\cdot|_v^s$ is $L(s,\pi_v)$. Let W_π be the (appropriately normalized) global Whittaker function for the cusp form $\pi = \bigotimes_v \pi_v$ on $\mathrm{GL}_2(\mathbb{A}_\mathbb{Q})$. At primes p where π_p is unramified, if $x \in \mathbb{Q}_p^\times$ is such that $|x|_p = p^r$, then

$$\lambda_{\pi_p}(x) = \mathrm{Tr}\left(\mathrm{sym}^r \begin{pmatrix} \alpha_\pi(p,1) & \\ & \alpha_\pi(p,2) \end{pmatrix}\right) = |x|^{-1/2} W_\pi \begin{pmatrix} x & \\ & 1 \end{pmatrix},$$

the last equality resulting from the Casselman–Shalika–Shintani formula [138]. If π_∞ is unramified, then $L(s,\pi_\infty) = \Gamma_\mathbb{R}(s+\nu)\Gamma_\mathbb{R}(s-\nu)$ and one may use a standard integral transformation formula to obtain

$$\lambda_{\pi_\infty}(x) = K_\nu(2\pi|x|) = 2^{-1}|x|^{-1/2} W_{0,\nu}(4\pi|x|) = |x|^{-1/2} W_\pi \begin{pmatrix} x & \\ & 1 \end{pmatrix},$$

where K_ν is the K-Bessel function and $W_{0,\nu}$ is the classical weight 0 Whittaker function. In this classical situation, it is transparent that bounding the L-series coefficient (or, more generally, the v-adic coefficient function) is equivalent to controlling the v-adic asymptotics at 0 of W_π.

The situation is much more complicated when dealing with higher degree L-functions. Fortunately, for the Rankin–Selberg L-function on $\mathrm{GL}_n \times \mathrm{GL}_n$, a very pleasant local integral representation allows us to make a concrete and direct link between the coefficient function and the Whittaker function. To explain, let us assume for simplicity that $F = \mathbb{Q}$, $v = \infty$, and that the cusp form $\pi = \bigotimes_v \pi_v$ is unramified at infinity. Then $\lambda_{\pi_v \times \tilde{\pi}_v}$ is defined as the unique even function on $\mathbb{R}^\times$ such that

$$\lambda_{\pi_v \times \tilde{\pi}_v}(x) = \frac{1}{2\pi i} \int_{(2)} L(s,\pi_v \times \tilde{\pi}_v)|x|^{-s} ds.$$

Let A be the subgroup of $\mathrm{GL}_n(\mathbb{R})$ consisting of those diagonal matrices of the form

$$a = \mathrm{diag}(a_1, a_2, \ldots, a_{n-1}, 1),$$

where $a_j \in \mathbb{R}_+^\times$ for $1 \leqslant j \leqslant n-1$, endowed with the Haar measure

$$d^\times a = \prod_{j=1}^{n-1} \frac{da_j}{a_j}.$$

Let W_{π_v} be the unique up to scaling $\mathrm{O}_n(\mathbb{R})$-invariant Whittaker function associated with π_∞ (with respect to the additive character $\psi(x) = e^{2\pi i x}$). When W_{π_v} is appropriately normalized, a formula of Stade [139] states that

$$L(s, \pi_v \times \tilde{\pi}_v) = \Gamma_{\mathbb{R}}(ns) \int_A |W_{\pi_v}(a)|^2 \det(a)^s \delta(a)^{-1} d^\times a, \tag{14}$$

where

$$\delta(a) = \prod_{j=1}^{n-1} \left(\frac{a_j}{a_{j+1}} \right)^{j(n-j)}.$$

Now let A_1 denote the subgroup of A consisting of matrices having determinant 1. Then an arbitrary element $a \in A$ decomposes as $a = a_x . a_1$ where $a_1 \in A_1$ and $a_x = \mathrm{diag}(x, \dots, x, 1)$ for $x \in \mathbb{R}_+^\times$. Give A_1 the measure μ_1 such that $d^\times a = d\mu_1(a_1) dx/x$. For $x \in \mathbb{R}^\times$ we write $\Psi(x) = \frac{2}{n} \exp(-\pi |x|^{2/n})$ and

$$\mathcal{W}_{\pi_v \times \tilde{\pi}_v}(x) = \int_{A_1} |W_{\pi_v}(a_{|x|} a_1)|^2 d\mu_1(a_1).$$

This is an integration over a level set of the determinant in A. Then (14) expresses the Archimedean L-function $L(s, \pi_v \times \tilde{\pi}_v)$ as the product of the Mellin transforms of Ψ and $\mathcal{W}_{\pi_v \times \tilde{\pi}_v}$. It follows that

$$\lambda_{\pi_v \times \tilde{\pi}_v} = \Psi * \mathcal{W}_{\pi_v \times \tilde{\pi}_v}, \tag{15}$$

the multiplicative convolution of functions on $\mathbb{R}^\times$.

We conclude from the above identities that $\lambda_{\pi_v \times \tilde{\pi}_v}$ is essentially an *average* of $|W_{\pi_v}|^2$ over a level set of the determinant. In particular, the function $\lambda_{\pi_v \times \tilde{\pi}_v}$ for $v = \infty$ is non-negative, which is not trivial to deduce otherwise. For unramified finite places $\mathfrak{p}$, positivity of the coefficients $\lambda_{\pi \times \tilde{\pi}}(\mathfrak{p}^r)$ follows directly from the definition, and for ramified finite places it has been checked in [116].

7. Techniques over $\mathbb{Q}$

In this section, we restrict our attention to the ground field $\mathbb{Q}$ in an effort to describe the basic L-function techniques in their simplest incarnation. There are two approaches, according to the use of the Rankin–Selberg or symmetric square L-functions. While the geometric significance of these techniques remains rather mysterious, Kloosterman sums (and integrals) continue to play a vital role. Some of the material in this section can also be found, for instance, in [4, 100].

7.1. With positivity.

We will show how to improve the Jacquet–Shalika bound (9) to the Luo–Rudnick–Sarnak bounds (11) for cuspidal automorphic representations π on $\mathrm{GL}_n(\mathbb{A}_\mathbb{Q})$. The principal ingredient is the work of Jacquet, Piatetski-Shapiro and Shalika [66] establishing the nice analytic properties of higher degree Rankin–Selberg L-functions, in particular the functional equation and the absolute convergence on $\Re s > 1$.

The original argument which produces the Luo–Rudnick–Sarnak bounds (11) at a rational prime p was actually introduced by Rankin [113] (and somewhat later by

Selberg [127]) in the setting of GL_2, and later generalized by Serre [131] to GL_n. We give our own version of this argument. From the observations of Section 6.2 we see that $\max_i |\sigma_\pi(p,i)| \leqslant \delta$ would follow from $\lambda_{\pi\times\tilde{\pi}}(p^r) \ll (p^r)^{2\delta}$ for sufficiently large $r \in \mathbb{N}$. We fix a parameter $T > 1$ and consider the mean value

$$\mathbb{E}_T(p^r) = \sum_{m\geqslant 1} \lambda_{\pi\times\tilde{\pi}}(m)\Phi_\infty\left(\frac{m}{p^r}\right),$$

where Φ_∞ is a smooth bump function on $\mathbb{R}_+^\times$ supported in the ball of radius $1/T$ about 1. In other words, we restrict the sum to $|m - p^r| \leqslant p^r/T$. By positivity of $\lambda_{\pi\times\tilde{\pi}}(m)$, an upper bound for $\mathbb{E}_T(p^r)$ gives an upper bound for each individual term, in particular the term corresponding to $m = p^r$. This may seem like a wasteful technique, but it gives non-trivial results.

Now the functional equation of $L(s, \pi \times \tilde{\pi})$ yields a kind of Poisson summation formula for the coefficients $\lambda_{\pi\times\tilde{\pi}}(m)$. This allows us to write $\mathbb{E}_T(p^r)$ as a main term, coming from the pole of $L(s, \pi \times \tilde{\pi})$ at $s = 1$, and a "dual sum" similar to $\mathbb{E}_T(p^r)$ with a different Archimedean weight function Φ_∞^* whose definition involves the quotient of Γ-functions

$$\gamma(s, \pi_\infty \times \tilde{\pi}_\infty) = L(s, \pi_\infty \times \tilde{\pi}_\infty)/L(1 - s, \pi_\infty \times \tilde{\pi}_\infty).$$

This integral transform can be estimated using classical analysis. Perhaps the most conceptual approach is through a stationary phase argument, although Serre quotes a fundamental result of Chandrasekharan and Narasimhan [27, Theorem 4.1] that is based on a differencing technique.[16] In any case, it remains to optimize T—the Archimedean length of the sum—in terms of p^r. If T is too large, the main term becomes too large; if T is too small, the dual sum cannot be controlled. The optimal value of T gives (11) at the prime p.

It is perhaps not clear how to modify the above argument to prove the same bounds at an Archimedean place. Indeed, the quest to improve the trivial bound at infinity using Rankin–Selberg theory was the driving impetus for Luo, Rudnick, and Sarnak to find an alternative approach to that sketched above. Their success came swiftly after the discovery of a link between this problem and the non-vanishing of L-functions in a family of character twists. We do not discuss their approach here, which the reader can find clearly explained in the original articles [91, 92] as well as the more recent and slightly more general treatment [10].

Instead, we follow an idea of Iwaniec [62] and replace the short Archimedean interval in $\mathbb{E}_T(p^r)$ by its q-adic analogue, a long arithmetic progression mod q^ℓ. In other words, we impose the congruence condition $m \equiv \pm 1 \pmod{q^\ell}$, the ± being present so as to give a well-defined condition on ideals and not just integers. Here q is a fixed prime which plays the role of the infinite place above (it is where we put a test function), and ℓ gets large just as T did previously. We thus consider

$$\mathbb{E}_{q^\ell}(Y) = \sum_{(m,q)=1} \lambda_{\pi\times\tilde{\pi}}(m)\Phi_S(mY, m),$$

where

$$\Phi_S(x, y) = \lambda_{\pi_\infty\times\tilde{\pi}_\infty}(x)\mathbf{1}_{\pm 1\,(\mathrm{mod}\ q^\ell)}(y).$$

[16] The prototype of this result, but with more restrictive assumptions, was proved by Landau [84] already in 1915.

The subscript S signifies the finite set of places $\{q,\infty\}$, so that Φ_S is an S-adic test function on $\{\pm 1\}\backslash(\mathbb{R}^\times \times \mathbb{Z}_q^\times)$. Note that an upper bound on $\mathbb{E}_{q^\ell}(Y)$ controls the asymptotics at 0 of the Archimedean coefficient function $\lambda_{\pi_\infty\times\tilde{\pi}_\infty}$ by dropping all but the first term and letting $Y\to 0$. With this modified but perfectly analogous set-up, we can now imitate the above argument with the role of the parameter T played by q^ℓ.

7.2. **Without positivity.** If $n\leqslant 4$, one has essentially the same analytic information for symmetric square L-function $L(s,\pi,\mathrm{sym}^2)$ as for $L(s,\pi\times\tilde{\pi})$. Both of these L-functions enjoy the property that the local parameters appear with multiplicity two in the expression for their coefficients, but the former has much smaller degree than the latter. By using the degree $n(n+1)/2$ symmetric square L-function, one could hope to replace the Luo–Rudnick–Sarnak bounds (11) by the stronger

$$\max_i |\sigma_\pi(v,i)| \leqslant \frac{1}{2} - \frac{1}{\frac{1}{2}n(n+1)+1}, \qquad n\leqslant 4. \tag{16}$$

This improvement does not come for free, however, because we lose positivity of coefficients—a crucial input.

An ingenious trick of Duke and Iwaniec [44] resolves this problem. We illustrate their idea in the setting where one wants to bound the local parameters at a fixed rational prime p. We again let q denote another prime and put $S=\{q,\infty\}$. At the real place we fix an even bump function ϕ_∞ supported about ± 1 in $\mathbb{R}^\times$. Then

$$\Phi_S(x,y)=\phi_\infty(x)\mathbf{1}_{\pm 1\ (\mathrm{mod}\ q)}(y)$$

defines an S-adic test function on $\{\pm 1\}\backslash(\mathbb{R}^\times\times\mathbb{Z}_q^\times)$ which we use to define the mean value

$$\mathbb{E}_q(p^\ell)=\sum_{(m,q)=1}\lambda_{\mathrm{sym}^2\pi}(m)\Phi_S\left(m,\frac{m}{p^r}\right).$$

By contrast to the last section, where we considered mean values of Rankin–Selberg coefficients, an upper bound on $\mathbb{E}_q(p^\ell)$ does not imply an upper bound on any given term, since they are no longer necessarily positive. Duke and Iwaniec therefore consider the average

$$\frac{1}{Q}\sum_{Q\leqslant q\leqslant 2Q}\mathbb{E}_q(p^\ell)$$

over all primes q in a dyadic interval. Exchanging the order of summation, this is

$$\sum_{m\geqslant 1}\lambda_{\mathrm{sym}^2\pi}(m)w(m-p^r)\phi_\infty(m), \quad w(a):=\frac{1}{Q}\sum_{\substack{Q\leqslant q\leqslant 2Q\\ q\mid a}}1.$$

The weight function $w(a)$ is essentially a divisor function. The main point here is the trivial observation that 0 is distinguished from all other integers by having exceptionally many divisors. We have $w(0)\asymp 1$, but $w(m-p^r)\ll Q^{\varepsilon-1}$ for all $m\neq p^r$. Hence the preceding display equals

$$w(0)\lambda_{\mathrm{sym}^2\pi}(p^r)+\text{small error},$$

and we can proceed as above, arriving at (16).

7.3. **The role of hyper-Kloosterman sums.** It is instructive to see how the Archimedean and non-Archimedean integral transforms structurally resemble each other. We highlight these similarities now using the real place, the complex place being similar.

Assume for simplicity that π_v, where $v = \mathbb{R}$, is unramified and that $\mu_\pi(v,i) = 0$ for all i. In this case, the Gamma quotient appearing in the integral transform Φ_v^* mentioned in Section 7.1 takes on the particularly simple form

$$\gamma(s, \pi_v \times \tilde{\pi}_v) = \left(\frac{\Gamma_\mathbb{R}(s)}{\Gamma_\mathbb{R}(1-s)} \right)^m . \tag{17}$$

Now it is well known that $\Gamma_\mathbb{R}(s)/\Gamma_\mathbb{R}(1-s)$ is the Mellin transform of the additive character on $\mathbb{R}^\times$. Indeed,

$$\int_{\mathbb{R}^\times} e^{2\pi i x} |x|^s \frac{dx}{|x|} = \frac{\Gamma_\mathbb{R}(s)}{\Gamma_\mathbb{R}(1-s)}, \tag{18}$$

a classical formula whose importance in number theory was revealed in Tate's thesis [142, Chapters 2.4 and 2.5]. Unitary characters $|\cdot|^s = |\cdot|^{\sigma+it}$ correspond to $\sigma = 0$, in which case the right-hand side exists unless $s = 0$. From this formula it follows that the product (17) can be interpreted as the Mellin transform of the m-fold multiplicative convolution of the additive character with itself.

The q-adic analogue of (18) is a Gauß sum,

$$G(\chi) = \frac{1}{q^\ell} \sum_{x\,(q^\ell)} e^{2\pi i x/q^\ell} \chi(x).$$

Indeed, this is just the Mellin transform of the additive character against the Dirichlet character χ mod q^ℓ. The analogue of (17) is then $G(\chi)^m$, and one again may interpret this latter product as the Mellin transform of the m-fold multiplicative convolution of the additive character,

$$\frac{1}{q^{\ell(m-1)}} \sum_{x_1 \cdots x_m = y} e^{2\pi i (x_1 + \cdots + x_m)/q^\ell} .$$

While its Archimedean counterpart has no common appellation, the above sum is called a hyper-Kloosterman sum, since in the special case $m = 2$ it reduces to an ordinary Kloosterman sum. When restricting mean values by the congruence condition $m \equiv \pm 1 \bmod q^\ell$, the hyper-Kloosterman sum appears in the kernel of the resulting transform Φ_S^*.

Such structural analogies between q-adic exponential sums and Archimedean oscillatory integrals are quite familiar in the literature. For example, the fact that Kloosterman sums are the q-adic analogue of the K-Bessel function was observed[17] long ago by Petersson [105, p. 178]. They are helpful in developing intuition, in so far as they motivate an equal treatment of all places. In our setting, we are interested in tight upper bounds on the integral transforms coming from the application of the functional equation of L-functions. At the real place this is done by stationary phase. At the prime q, as long as $\ell \geqslant 2$ one can estimate the hyper-Kloosterman sum mod q^ℓ in an elementary way; see for example [64, Chapter 12].

[17] He writes: *Thus one can say that Kloosterman sums are in the same sense a generalization of the simpler Gauß sums, as the Bessel integral is a generalization of the reciprocal Gamma function.* [Translation by the authors.]

If one needs or prefers to choose $\ell = 1$, so that the modulus is a prime, one has to invoke Deligne's bounds for exponential sums [38, p. 219].

8. Techniques over number fields

Let us now move from $\mathbb{Q}$ to a general algebraic number field F of degree d over $\mathbb{Q}$. The class group of F may now be non-trivial, but since it is finite, as long as we regard the number field as fixed, no new difficulties arise in the methods of the previous section from the lack of unique factorization. The units, on the other hand, may cause substantial problems. Here we describe how the methods over $\mathbb{Q}$ can be generalized to general number fields, despite the complicating presence of an infinite unit group. Our goal here is to view the techniques affording this extension in the light of the analogy with the higher rank situation.

8.1. **Units in number fields.** We begin by describing the naïve number field replacement of the rational congruence condition $n \equiv \pm 1 \bmod q^\ell$ that was used to define the mean value sum $\mathbb{E}_{q^\ell}(Y)$ in Section 7.1. The condition we must invent should be on the set $\mathcal{PI}_F$ of non-zero principal integral ideals of F, which we identify with the $\mathcal{O}_F^\times$-orbits on $\mathcal{O}_F \setminus \{0\}$. We start with an integral ideal $\mathfrak{m}$ and let $\mathcal{O}_F^\times \pmod{\mathfrak{m}}$ denote the image of the unit group $\mathcal{O}_F^\times$ in the invertible residue classes $(\mathcal{O}_F/\mathfrak{m})^\times$. Then the delta function at the identity element in the quotient

$$(\mathcal{O}_F/\mathfrak{m})^\times/\mathcal{O}_F^\times \pmod{\mathfrak{m}}$$

produces a well-defined condition on $\mathcal{PI}_F$: a principal integral ideal is given weight 1 or 0 according to whether or not it contains a generator congruent to 1 mod $\mathfrak{m}$. The argument using positivity sketched in Section 7.1 would go through unchanged provided this latter condition is roughly as strong as the original condition $x \equiv \pm 1 \pmod{m}$ over $\mathbb{Q}$. In other words, we want the above quotient to be big. If, given an $\epsilon > 0$, one has

$$\# \, \mathcal{O}_F^\times \pmod{\mathfrak{m}} \ll_\varepsilon \mathrm{N}(\mathfrak{m})^\varepsilon \tag{19}$$

for an infinite number of $\mathfrak{m}$, then we indeed obtain the Luo–Rudnick–Sarnak bounds (11) for all places v, finite or infinite.

As $\mathfrak{m}$ varies over integral ideals in $\mathcal{O}_F$, how likely is it that $\mathcal{O}_F^\times \pmod{\mathfrak{m}}$ is small? A heuristic can be found in the classical Artin primitive root conjecture which states that given an integer $a > 1$, the multiplicative group generated by a has image mod p equal to all of $(\mathbb{Z}/p\mathbb{Z})^\times$ for a positive proportion of all primes p. Under the assumption of the Generalized Riemann Hypothesis this has been proved by Hooley [59]. There are similar statements over number fields (see, for example, [89]) to the effect that $\mathcal{O}_F^\times \pmod{\mathfrak{p}}$ will often be all, or nearly all, of $(\mathcal{O}_F/\mathfrak{p})^\times$. This is precisely the situation the condition (19) wants to avoid.

Fortunately, it is possible to construct an infinite sequence of square-free ideals $\mathfrak{m}$ that satisfy (19). This was proved by Rohrlich [115] using several non-trivial results as input, not the least of which was a version of the Bombieri–Vinogradov theorem over number fields due to Murty and Murty [101]. Very roughly, one takes $\mathfrak{m}$ to be a product of prime ideals $\mathfrak{p}$ such that $\mathrm{N}(\mathfrak{p}) - 1$ has no large prime factor. Thanks to Rohrlich's theorem, the methods outlined in Section 7.1 can be used to prove the bounds (11). The argument by non-vanishing of L-functions in families of character twists, as originally developed by Luo, Rudnick, and Sarnak, also critically relies on the construction of special moduli due to Rohrlich.

As satisfactory as the above situation is, one drawback remains. The sequence of ideals constructed by Rohrlich is very sparse; there are certainly fewer than $\log X$ such ideals of norm at most X. This renders Rohrlich's construction incompatible with the method of Duke and Iwaniec in Section 7.2 which requires an average over q in a dyadic interval. To run their argument over a number field, the authors in [9] defined a suitable test function *for every ideal* $\mathfrak{m}$, regardless of the size of $\mathcal{O}_F^\times \pmod{\mathfrak{m}}$.

To describe the construction we take $\mathfrak{m}$ to be a prime ideal $\mathfrak{q}$ for simplicity. Let $\phi_\mathfrak{q}$ be the characteristic function of $1+\mathfrak{q}$, the degree one neighborhood of 1 in $F_\mathfrak{q}^\times$. Similarly we let ϕ_∞ be a smooth positive function supported in a small ball about 1 in $F_\infty^\times = \prod_{v|\infty} F_v^\times$. Put $S=\{\infty,\mathfrak{q}\}$ and write $\phi_S = \phi_\mathfrak{q}\phi_\infty$. By setting

$$\Phi_S(x) = \sum_{u\in\mathcal{O}_S^\times} \phi_S(ux),$$

we obtain a function on $F_S^\times = F_\mathfrak{q}^\times \times F_\infty^\times$ invariant under $\mathcal{O}_S^\times$, the S-units of F. While there is nothing uncommon about the above construction, which takes a factorizable S-adic neighborhood of 1 and averages it, the interplay between the place q and the Archimedean places ∞ that result from it seems quite miraculous. If $1\in F_\mathfrak{q}^\times$ has large image under $\mathcal{O}_S^\times$—this is another way of saying that the units cover many invertible residue classes mod $\mathfrak{q}$—then $\mathrm{supp}(\phi_\infty)\subset F_\infty^\times$ has small image under $\mathcal{O}_F^\times$ mod $\mathfrak{q}$. Indeed the volume of the support of Φ_S is equal to that of ϕ_S, which is about $\mathrm{N}(\mathfrak{q})$ regardless of the arithmetic nature of the chosen ideal $\mathfrak{q}$. The Archimedean and non-Archimedean places compensate each other!

8.2. **Exploring all directions.** We continue to let F denote a fixed number field. Let d be its degree. We set the module $\mathfrak{m}=1$ so that, unlike the previous section, there is no congruence condition, and now look at the interplay of the Archimedean places *amongst each other.*

We begin by examining the most immediate generalization of the mean value $\mathbb{E}_T(p^r)$, whose test function was taken to be purely Archimedean. Over a number field, we would define a mean value of coefficients $\lambda_{\pi\times\tilde\pi}(\mathfrak{a})$ over ideals $\mathfrak{a}$ subject to

$$|\mathrm{N}(\mathfrak{a}) - \mathrm{N}(\mathfrak{p}^r)| \leqslant \mathrm{N}(\mathfrak{p}^r)/T. \tag{20}$$

The positivity argument of Section (7.1) then produces

$$\max_i |\sigma_\pi(v,i)| \leqslant \frac{1}{2} - \frac{1}{dn^2+1}, \tag{21}$$

a bound that is implicit in Serre's letter [131] to Deshouillers. While better than the local bound (9) of Jacquet and Shalika, the quality of (21) worsens with d. The reason for this degradation is that the factor

$$\gamma(s,\pi_\infty\times\tilde\pi_\infty) = \prod_{v|\infty}\gamma(s,\pi_v\times\tilde\pi_v)$$

defining the Archimedean integral transform oscillates more rapidly since it consists now of dn^2 Gamma quotients.

To better understand the dependence on the degree d, we observe that d can be otherwise viewed as the number of dimensions of the ambient space on which one defines an appropriate test function. Indeed, one can embed the principal fractional ideals $\mathcal{P}_F$ of F as a discrete subset of the d-dimensional space $X_F = \mathcal{O}_F^\times\backslash F_\infty^\times$ by identifying $\mathcal{P}_F$ with the quotient $\mathcal{O}_F^\times\backslash F^\times$. To understand the structure of X_F, let

$$\begin{aligned} \mathrm{N}_\infty : \quad F_\infty^\times &\longrightarrow \mathbb{R}_+^\times \\ (x_v)_{v|\infty} &\longmapsto \prod_{v|\infty} |x_v|_v. \end{aligned}$$

This is a surjective homomorphism whose kernel we denote by F_∞^1. Then N_∞ factors through X_F, and when applied to a principal fractional ideal $\mathfrak{a} \in \mathcal{P}_F$, it evaluates to the norm $\mathrm{N}(\mathfrak{a})$. We can identify the fibers of N_∞ with $X^1 = \mathcal{O}_F^\times\backslash F_\infty^1$, and the connected components of this latter space are $d-1$ dimensional tori. The method leading to (21) uses test functions on X_F which are constant on the fibers of N_∞. By contrast, in [9] and [10] we use test functions which can concentrate in all possible directions of X_F. As in the previous section, one simply averages a factorizable test function on $F_\infty^\times$ over the units $\mathcal{O}_F^\times$. This greater flexibility allows one to eliminate the dependence of the bound (21) on d.

Even when F has only *one* Archimedean place (but F is not $\mathbb{Q}$), there is still something interesting to say. This is the case of an imaginary quadratic field. The two dimensions in $F_\infty^\times = \mathbb{C}^\times$ are given by the modulus and argument of a non-zero complex number; the condition (20) only makes use of the former. We can understand this alternatively by means of the automorphic dual of $\mathrm{GL}_1(F_\infty) = F_\infty^\times$, the characters of X_F. For simplicity we take $F = \mathbb{Q}(i)$. The mod 1 unitary Hecke characters of $\mathbb{Q}(i)$ are of the form

$$\chi_{m,t}(z) = (z/|z|)^{4m}|z|^{it},$$

where $m \in \mathbb{Z}$ and $t \in \mathbb{R}$. The analytic conductor of $\chi_{m,t}$ is defined to be $1+m^2+t^2$; it is the measure of complexity of $\chi_{m,t}$. To see how the bound (21) arises in this situation, observe that the condition (20) (or a smoothed version thereof) defines an annulus in $\mathbb{C}^\times$. Since this condition is rotation-invariant, only characters $\chi_{0,t}$ appear in the Fourier expansion, and by a suitable smoothing we can ensure that $t \in [-T, T]$. We therefore have a test function with support of size $1/T$ whose spectrum is composed of characters of analytic conductor in $[1, T^2]$. This is not optimal, for we can shrink the support without increasing conductors by additionally restricting $\arg(z)$ to $[-1/T, 1/T]$. The support now has size about $1/T^2$, while the characters $\chi_{m,t}$ appearing in the Fourier expansion have parameters t, m in $[-T, T]$, hence conductor again in $[1, T^2]$.

8.3. **Analogy with higher rank.** We mentioned in Section 5.2 that the number field situation can be viewed as a kind of dress rehearsal for the case of higher rank groups such as GL_n. There are real differences, though. For instance, the dependence in the degree of the number field in (21) is relatively easy to eliminate, as we saw in the previous paragraph. It is a different matter entirely to mitigate (not to mention eliminate) the dependence on n in the Luo–Rudnick–Sarnak bounds (11). Here we attempt to convey the nature of this difficulty and to speculate on its source.

The method described in Section 7.1 is based on the determination of the asymptotic behavior close to 0 of the local coefficient function $\lambda_{\pi_v\times\tilde{\pi}_v}$. When v is Archimedean and π_v is spherical, the formula (15) shows that $\lambda_{\pi_v\times\tilde{\pi}_v}$ is the average of the (square of the) Whittaker function over the fibers of the determinant map on A. As this integration takes place over all but one of the possible directions in A, a considerable loss of information is incurred. As in the previous section, it is the failure to exploit all directions in the large space A which is (at least partly) responsible for the pejoration in n of the existing bounds towards the Ramanujan conjecture.

It would be interesting to formalize this, establishing, in the spirit of Section 4.1, an equivalence between the decay rate and the L^p properties of Whittaker functions when they are averaged over level sets of the determinant.

9. Perspectives

9.1. **Kloosterman sums.**

Returning to the discussion in Section 3.4, we saw that in the classical case $n = 2$ an approach through trace formulas combined with best possible bounds for certain exponential sums (the Riemann hypothesis over finite fields) provided rather strong bounds towards the Ramanujan conjecture. It is natural to see what insight may be gained by developing this trace formula approach for higher rank.

Looking at the simplest case, one could try to bound the Satake parameters $\alpha_\pi(p, i)$ of a cusp form π on $\mathrm{GL}_3/\mathbb{Q}$ at an unramified prime p using the Kuznetsov trace formula on GL_3. From the Bruhat decomposition, this formula is expressed in terms of certain exponential sums, also called Kloosterman sums, associated with elements in the Weyl group for GL_3. Under certain relative primality conditions, the Kloosterman sum S_{w_ℓ} associated with the long Weyl group element w_ℓ factorizes as a product of two standard GL_2 Kloosterman sums (see [20] for details). It is this sum S_{w_ℓ}, rather than the hyper-Kloosterman sum from Section 7.3, which appears in the GL_3 Kuznetsov trace formula. An application of Weil's bound to each GL_2 Kloosterman sum in S_{w_ℓ} yields the bound $\alpha_\pi(p, i) = O(p^{1/2})$, which is, up to the implied constant, simply the Jacquet–Shalika bound. For $n \geqslant 4$, this naïve trace formula approach becomes even worse. One would like to build a more geometric intuition to the relation between the higher rank Kloosterman sums, reduction theory, and effective equidistribution of periodic unipotent orbits, as in Sections 3.3 and 3.4.

The above argument can be modified for the Archimedean place to show $|\Re\mu_\pi(\infty, i)| \leqslant 1/2$ for π on $\mathrm{GL}_3/\mathbb{Q}$ unramified at infinity. While this does not improve the Jacquet–Shalika bounds, the argument can be used to show the following density result. For simplicity we restrict our attention to everywhere unramified cusp forms $\pi = \bigotimes_v \pi_v$ of $\mathrm{GL}_3(\mathbb{A}_\mathbb{Q})$. These correspond bijectively to cuspidal smooth functions ϕ on the locally symmetric space $\mathrm{SL}_3(\mathbb{Z})\backslash\mathrm{SL}_3(\mathbb{R})/\mathrm{SO}(3)$ which are Hecke eigenfunctions and eigenfunctions of the rank 2 polynomial algebra of invariant differential operators. As an extension of the classical terminology for $\mathrm{GL}_2/\mathbb{Q}$, these functions ϕ are also referred to as Maaß forms. Furthermore, we call ϕ *exceptional* if it violates the Ramanujan conjecture at infinity. The local parameters at infinity of an exceptional Maaß form are given by $(-2i\gamma, \rho + i\gamma, -\rho + i\gamma)$ where $\rho \in (0, 1/2)$ and $\gamma \in \mathbb{R}$. The smaller ρ the closer π_∞ is to being tempered. One can show [8, Theorem 2]

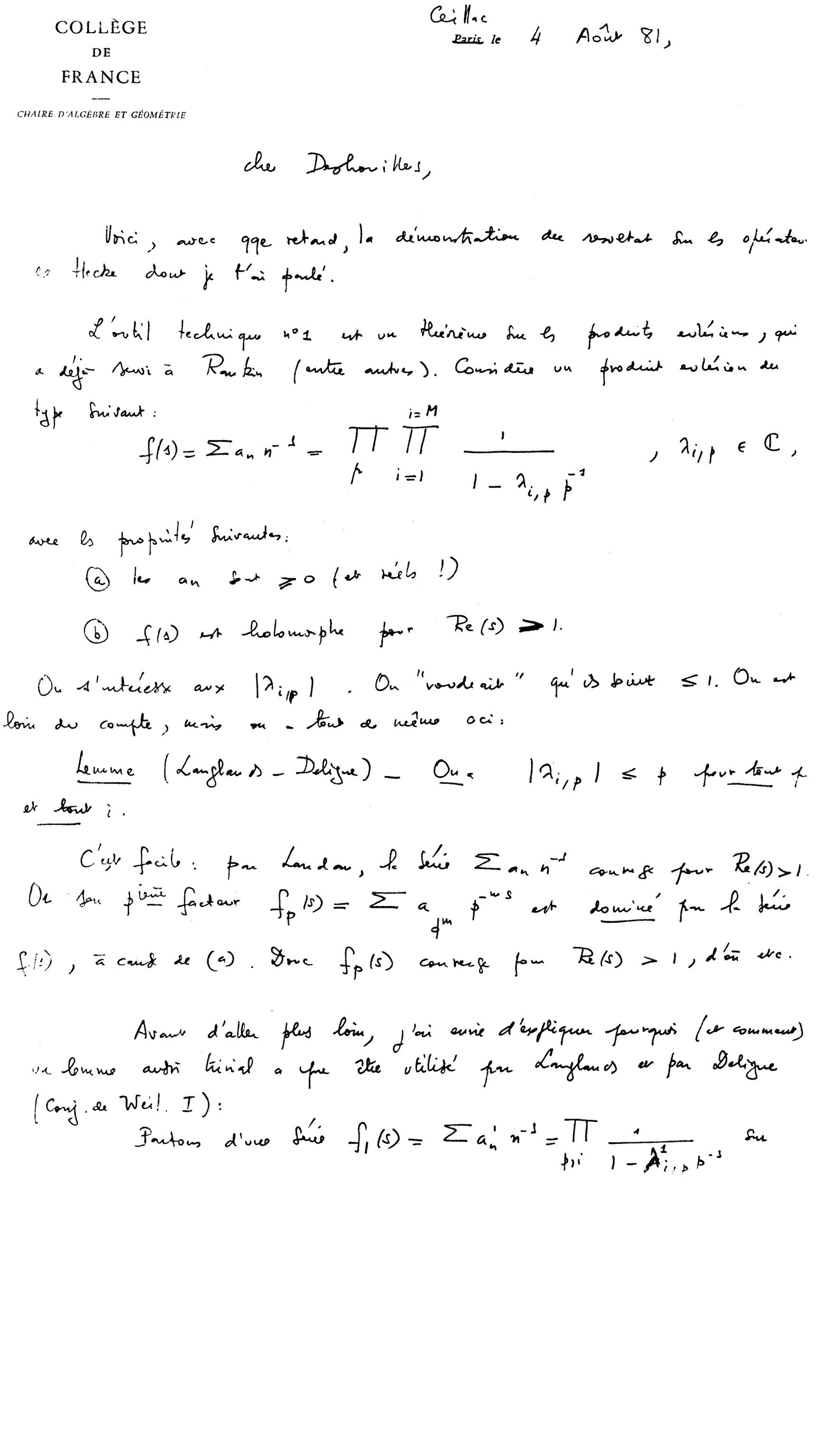
$$\sum_{X \leqslant \gamma_j \leqslant X+1} X^{4\rho_j} \ll_\varepsilon X^{2+\varepsilon}$$

for any parameter $X > 1$, where the sum is over a basis of exceptional Maaß forms ϕ_j for $\mathrm{SL}_3(\mathbb{Z})$. Of course one expects the set which the above sum is over to be empty, but the best available estimate on the size of the summing set is

$O(X^2)$, a result due to Lapid and Müller [88]. The content of the above estimate is then, informally speaking, that exceptional Maaß forms become rarer the more exceptional they are.

9.2. **Final thoughts.** The generalized Ramanujan conjecture is one of the great open conjectures in mathematics. It has served as a guide post to the work of multiple generations of mathematicians working in the fields of analytic number theory, automorphic forms and arithmetic geometry. Work on it has provided a means of communication between disparate branches—and styles—of mathematics. One need only think of Deligne's proof of the classical Ramanujan conjecture, where he was inspired [37, §3] by ideas of Rankin and Langlands. It is not unreasonable to hope that algebro-geometric methods, such as those developed by Taylor [30, 145] and his school, can be a "catalyst"—to use Deligne's word—towards an understanding, and someday a proof of, the Ramanujan-Petersson conjecture for the mysterious Maaß forms we described in Sections 2 and 3. In the meantime, it is not enough to wait for a proof of the functoriality conjectures of Langlands on the symmetric power lifts from GL_2. While it is true that the current world records on bounds towards the Ramanujan conjecture have come as corollaries of newly proven cases of functoriality, the methods proving these low rank cases seem to have been completely exhausted, and it is not clear what will take their place. The time may have come where the search for a better understanding of the Ramanujan conjecture (and of periods of automorphic forms, more generally) might *itself* lead to a greater understanding, and perhaps a breakthrough, in functoriality. The two conjectures are that intimately linked.

J.-P. Serre's 1981 letter to J.-M. Deshouillers



COLLÈGE
DE
FRANCE

CHAIRE D'ALGÈBRE ET GÉOMÉTRIE

Ceillac
Paris le 4 Août 81,

Cher Deshouillers,

Voici, avec qqe retard, la démonstration du résultat sur les opérateurs de Hecke dont je t'ai parlé.

L'outil technique n°1 est un théorème sur les produits eulériens, qui a déjà servi à Rankin (entre autres). Considère un produit eulérien du type suivant :

$$f(s) = \sum a_n n^{-s} = \prod_{p} \prod_{i=1}^{i=M} \frac{1}{1-\lambda_{i,p}\,p^{-s}}, \quad \lambda_{i,p} \in \mathbb{C},$$

avec les propriétés suivantes :

ⓐ les a_n sont $\geqslant 0$ (et réels !)

ⓑ $f(s)$ est holomorphe pour $\mathrm{Re}(s) > 1$.

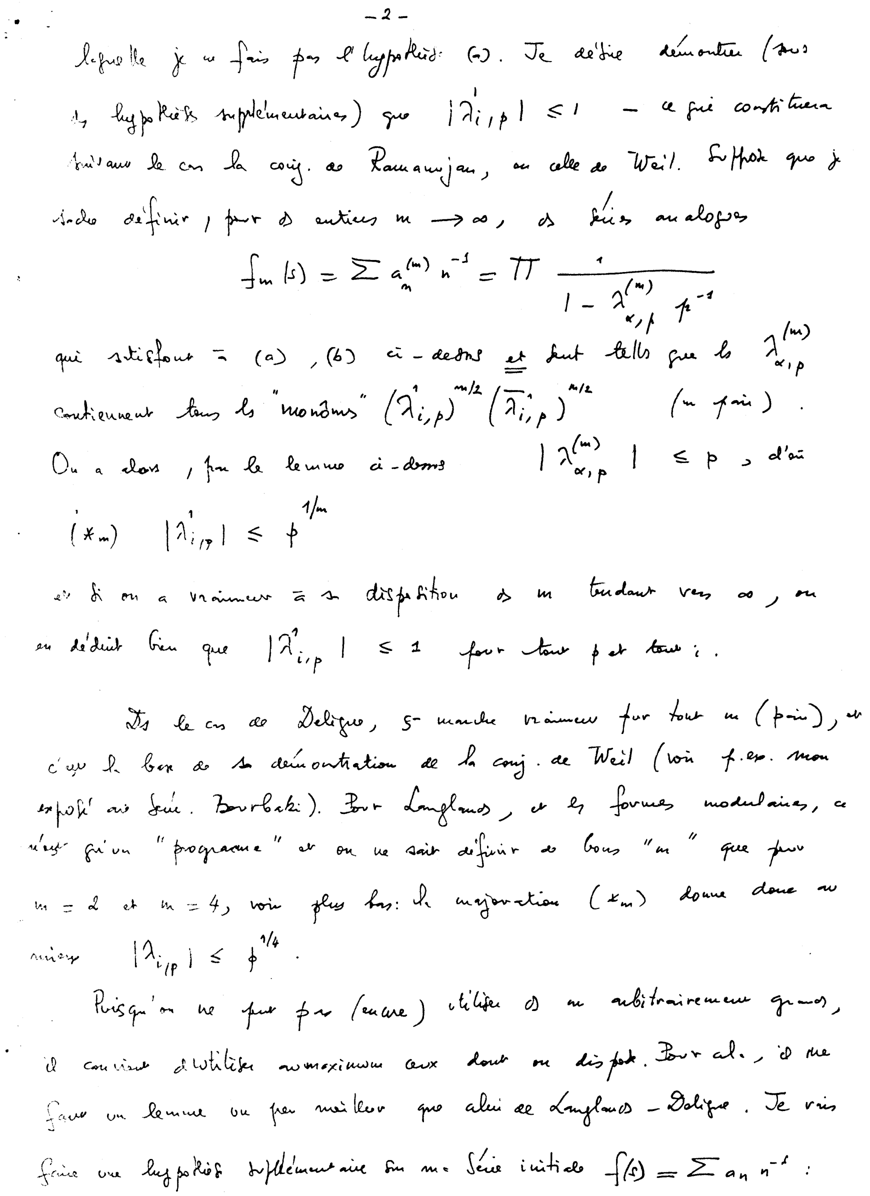
On s'intéresse aux $|\lambda_{i,p}|$. On "voudrait" qu'ils soient ≤ 1. On est loin du compte, mais on a tout de même ceci :

Lemme (Langlands – Deligne) — *On a* $|\lambda_{i,p}| \leq p$ pour *tout* p et *tout* i.

C'est facile : par Landau, la série $\sum a_n n^{-s}$ converge pour $\mathrm{Re}(s) > 1$. Or son $p^{\text{ième}}$ facteur $f_p(s) = \sum a_{p^m} p^{-ms}$ est *dominé* par la série $f(s)$, à cause de (a). Donc $f_p(s)$ converge pour $\mathrm{Re}(s) > 1$, d'où etc.

Avant d'aller plus loin, j'ai envie d'expliquer pourquoi (et comment) ce lemme aussi trivial a pu être utilisé par Langlands et par Deligne (Conj. de Weil. I) :

Partons d'une série $f_1(s) = \sum a^1_n n^{-s} = \prod_{p,i} \frac{1}{1-\lambda^1_{i,p}\,p^{-s}}$ sur



– 2 –

laquelle je ne fais pas l'hypothèse (a). Je désire démontrer (sous des hypothèses supplémentaires) que $|\lambda'_{i,p}| \leq 1$ – ce qui constituera suivant le cas la conj. de Ramanujan, ou celle de Weil. Supposé que je sache définir, pour des entiers $m \to \infty$, des séries analogues

$$f_m(s) = \sum a_n^{(m)} n^{-s} = \prod \frac{1}{1 - \lambda^{(m)}_{\alpha,p}\, p^{-s}}$$

qui satisfont à (a), (b) ci-dessus <u>et</u> sont tels que les $\lambda^{(m)}_{\alpha,p}$ contiennent tous les "monômes" $(\lambda'_{i,p})^{m/2} (\bar{\lambda}'_{i,p})^{m/2}$ (m pair). On a alors, par le lemme ci-dessus $|\lambda^{(m)}_{\alpha,p}| \leq p$, d'où

$(*_m)$ $|\lambda'_{i,p}| \leq p^{1/m}$

et si on a vraiment à sa disposition des m tendant vers ∞, on en déduit bien que $|\lambda'_{i,p}| \leq 1$ pour tout p et tout i.

Dans le cas de Deligne, ça marche vraiment pour tout m (pair), et c'est la base de sa démonstration de la conj. de Weil (voir p.ex. mon exposé au Sém. Bourbaki). Pour Langlands, et les formes modulaires, ce n'est qu'un "programme" et on ne sait définir de bons "m" que pour $m = 2$ et $m = 4$, voir plus bas: la majoration $(*_m)$ donne donc au mieux $|\lambda_{i,p}| \leq p^{1/4}$.

Puisqu'on ne peut pas (encore) utiliser des m arbitrairement grands, il convient d'utiliser au maximum ceux dont on dispose. Pour cela, il me faut un lemme un peu meilleur que celui de Langlands – Deligne. Je vais faire une hypothèse supplémentaire sur ma série initiale $f(s) = \sum a_n n^{-s}$:

— 3 —

(c) Il existe un produit ~~extérieur~~ de facteurs gamma et d'exponentielles

$$\Delta(s) = A^s \prod_\nu \Gamma(\alpha_\nu s + \beta_\nu) \qquad \left|\begin{array}{l} \alpha_\nu \text{ réel } > 0 \\ \beta_\nu \in \mathbb{C} \end{array}\right.$$

tel que $\Delta(s) f(s)$ soit méromorphe ds tout $\mathbb{C}$, avec nbre fini de pôles, genre fini, et satisfasse à une éq. fonct. :

$$\Delta(s) f(s) = c . \Delta(1-s) g(1-s), \qquad c \in \mathbb{C}^*,$$

où g est une fonction de même type que f. [en fait, $g = f$ dans toutes les applications].

Je dirai plus brièvement " f a une éq. fonctionnelle ". Je poserai

$$N = 2 \sum_\nu \alpha_\nu$$

et je dirai que N est le "rang" de f. En pratique, $\alpha_\nu = \frac{1}{2}$ pour tout ν, de sorte que N est un entier $\geqslant 1$ et c'est aussi l'entier M du début (nombre de $\lambda_{i,p}$ pour p fixé).

Lemme (à la Rankin) — Sous les hypothèses (a), (b), (c), on a

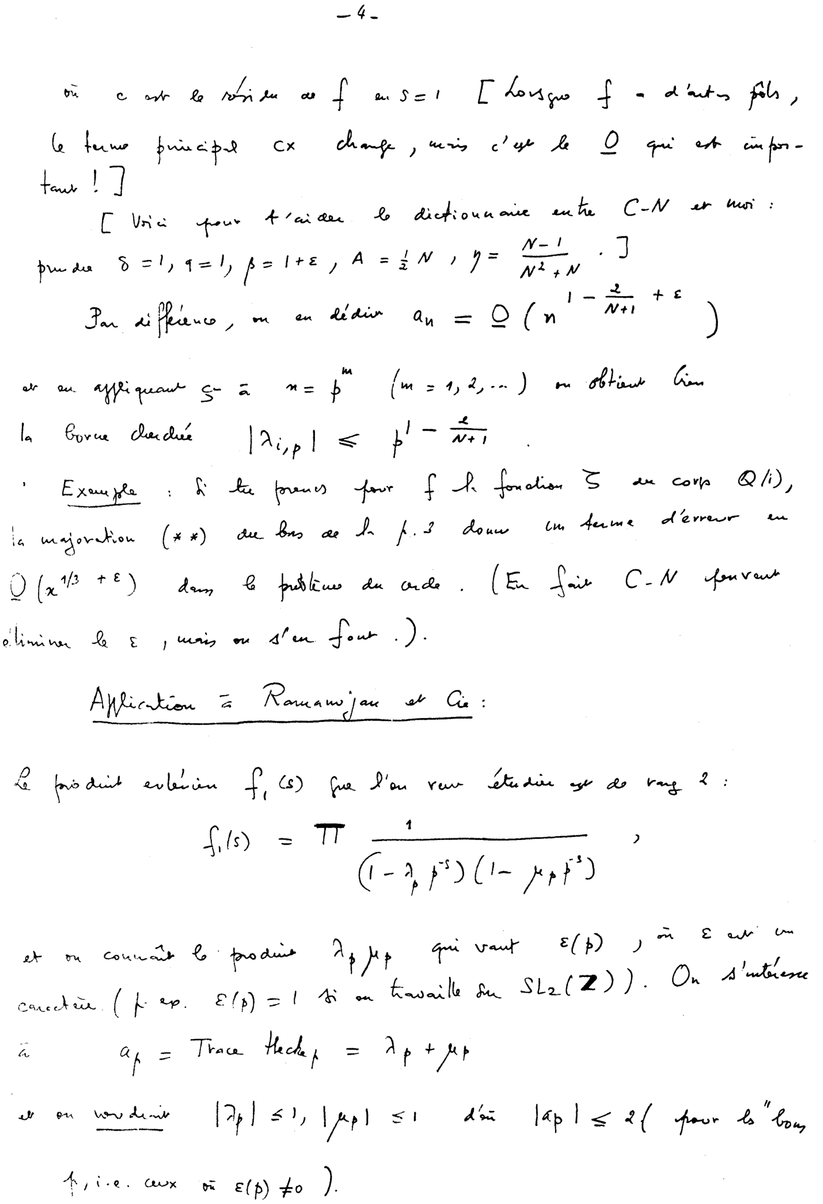
$$|\lambda_{i,p}| \leqslant p^{1 - \frac{2}{N+1}} .$$

Dém. Posons

$$A(x) = \sum_{n \leq x} a_n .$$

Supposons pour simplifier que $\Delta(s) f(s)$ n'ait qu'un pôle simple en $s = 1$ (et $s = 0$). (Le cas général est analogue.) D'après un th. de Landau (qui malheureusement suppose les β_ν réels), généralisé par Chandrasekharan – Narasimhan (Ann. Math. 76, 1962, p. 106, th. 4.1), on a

$$(**) \qquad A(x) = c . x + \underline{O}\left(x^{1 - \frac{2}{N+1} + \varepsilon}\right) \quad \text{pour tout } \varepsilon > 0$$



-4-

où c est le résidu de f en $s=1$ [Lorsque f a d'autres pôles, le terme principal cx change, mais c'est le O qui est important !]

[Voici pour t'aider le dictionnaire entre C-N et moi : prendre $\delta=1$, $q=1$, $\beta=1+\varepsilon$, $A=\frac{1}{2}N$, $\eta=\frac{N-1}{N^2+N}$.]

Par différence, on en déduit

$$a_n = O\left(n^{1-\frac{2}{N+1}+\varepsilon}\right)$$

et en appliquant ça à $n=p^m$ $(m=1,2,\dots)$ on obtient bien la borne cherchée $|\lambda_{i,p}| \leqslant p^{1-\frac{2}{N+1}}$.

Exemple : si tu prends pour f la fonction ζ du corps $Q(i)$, la majoration $(**)$ du bas de la p. 3 donne un terme d'erreur en $O(x^{1/3+\varepsilon})$ dans le problème du cercle. (En fait C-N peuvent éliminer le ε, mais on s'en fout.).

Application à Ramanujan et Cie :

Le produit eulérien $f_1(s)$ que l'on veut étudier est de rang 2 :

$$f_1(s) = \prod \frac{1}{(1-\lambda_p p^{-s})(1-\mu_p p^{-s})} ,$$

et on connaît le produit $\lambda_p \mu_p$ qui vaut $\varepsilon(p)$, où ε est un caractère (p. ex. $\varepsilon(p)=1$ si on travaille sur $SL_2(\mathbf{Z})$). On s'intéresse à

$$a_p = \text{Trace Hecke}_p = \lambda_p + \mu_p$$

et on *voudrait* $|\lambda_p| \leq 1$, $|\mu_p| \leq 1$ d'où $|a_p| \leq 2$ (pour les "bons" p, i.e. ceux où $\varepsilon(p) \neq 0$).

Bien sûr, la condition de positivité n'est pas satisfaite. Mais on peut passer à ce qu'on aurait envie d'appeler $f_1 \otimes \bar{f}_1$, le produit ensoriel de rang 4 dont les "λ_p" sont $\lambda_p \cdot \bar{\lambda}_p$, $\lambda_p \cdot \bar{\mu}_p$, $\mu_p \cdot \bar{\lambda}_p$, $\mu_p \cdot \bar{\mu}_p$. Rankin a démontré que ce produit a l'éq. fonct. (et le prolgt analytique) désirés. Le lemme ci-dessus s'applique avec $N = 4$ et donne donc

$$|\lambda_p \bar{\lambda}_p| \leq p^{1-\frac{2}{5}} = p^{\frac{3}{5}}$$

d'où

$$|\lambda_p| \leq p^{\frac{3}{10}} \quad , \quad |a_p| \leq p^{3/10} + p^{-3/10} \leq 2p^{3/10} .$$

Appliqué à $a_p = \tau(p)/p^{11/2}$, cela donne la majoration de Rankin, à savoir

$$\tau(p) \leq 2p^{6-\frac{1}{5}} -$$

Heureusement, on peut faire mieux. Considère d'abord la série $f_2 =$ "$\mathrm{Sym}^2 f_1$", de rang 3, donnée par

$$f_2 = \prod_p \frac{1}{(1-\lambda_p^2 p^{-s})(1-\lambda_p \mu_p p^{-s})(1-\mu_p^2 p^{-s})} .$$

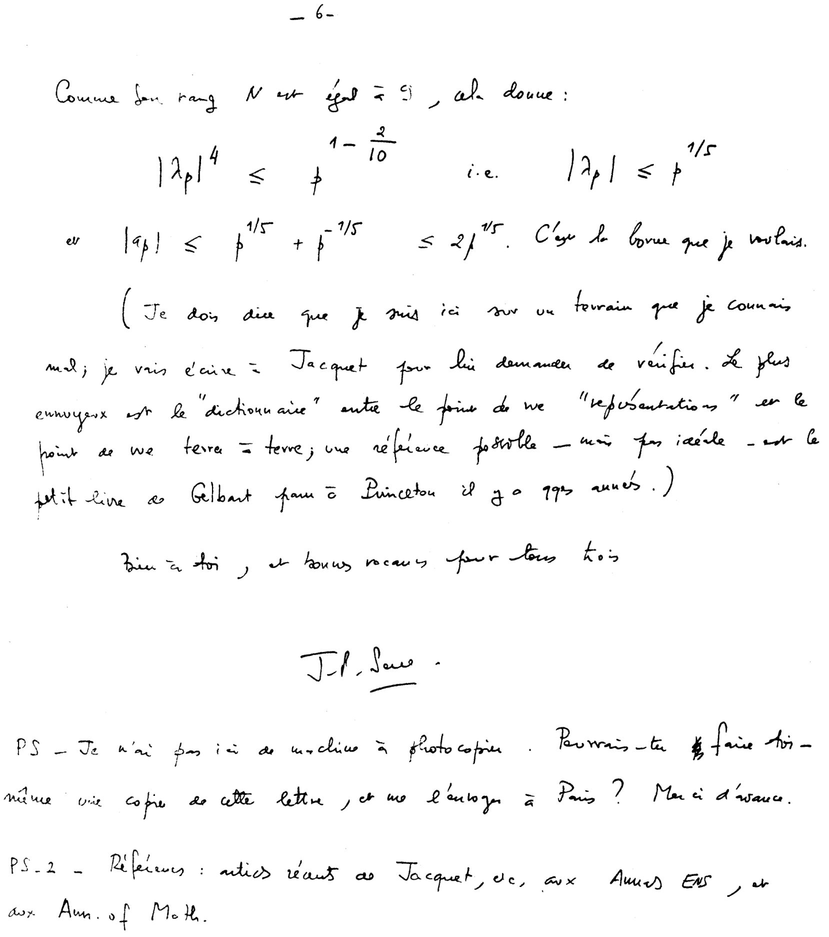
Jacquet, Shalika, etc ont montré que f_2 correspond à une forme parabolique sur $\underline{SL_3}$ (du moins si f_1 n'est pas "de type CM", cas facile à traiter directement). Il en résulte que la série

par d'autres ths. de Jacquet, etc

[illegible]

$$f_4 = "f_2 \otimes \bar{f}_2"$$

de rang 9, dont les "λ_p" sont les 9 produits de $(\lambda_p^2, \lambda_p \mu_p, \mu_p^2)$ avec $(\bar{\lambda}_p^2, \bar{\lambda}_p \bar{\mu}_p, \bar{\mu}_p^2)$, est une série qui satisfait à nos conditions (a), (b), (c).



– 6 –

Comme son rang N est égal à 9, cela donne :

$$|\lambda_p|^4 \leq p^{1-\frac{2}{10}} \qquad \text{i.e.} \qquad |\lambda_p| \leq p^{1/5}$$

et $|a_p| \leq p^{1/5} + p^{-1/5} \leq 2p^{1/5}$. C'est la borne que je voulais.

(Je dois dire que je suis ici sur un terrain que je connais mal; je vais écrire à Jacquet pour lui demander de vérifier. Le plus ennuyeux est le "dictionnaire" entre le point de vue "représentations" et le point de vue terre à terre; une référence possible – mais pas idéale – est le petit livre de Gelbart paru à Princeton il y a qqs années.)

Bien à toi, et bonnes vacances pour tous trois

J-P. Serre.

PS – Je n'ai pas ici de machine à photocopier. Pourrais-tu faire toi-même une copie de cette lettre, et me l'envoyer à Paris? Merci d'avance.

PS-2 – Références : articles récents de Jacquet, etc, aux Annales ENS, et aux Ann. of Math.

Acknowledgments

The authors would like to thank Peter Sarnak and Laurent Clozel for very detailed comments on a preliminary version of this article, and Andrew Granville for constructive advice on the presentation of the material.

Thanks are also due to Samuel Patterson for pointing us to the reference [106], Stefan Baur for drawing Figures 1 and 2, and Christopher Ambrose, Alex Dahl and Marc Palm for spotting various errors and making useful suggestions.

We would also like to thank Jean-Marc Deshouillers for providing us with a copy of Serre's letter [131], which played an important role in the development of techniques yielding bounds towards the Ramanujan conjecture on GL_n. We are most grateful to Jean-Pierre Serre for allowing us to include his letter as an appendix to the present article, and we hope that it is both of mathematical and historical interest for the mathematical community.

About the authors

Valentin Blomer and Farrell Brumley work in analytic number theory and automorphic forms. Blomer is professor of mathematics at the University of Göttingen. Brumley is currently Maître de Conférences at Université Paris 13. Their collaborative work began in Nancy, France, and was fostered by long stays in Princeton and Lausanne.

References

[1] Dan Abramovich, *A linear lower bound on the gonality of modular curves*, Internat. Math. Res. Notices **20** (1996), 1005–1011, DOI 10.1155/S1073792896000621. MR1422373 (98b:11063)

[2] James Arthur, *A note on the automorphic Langlands group*, Canad. Math. Bull. **45** (2002), no. 4, 466–482, DOI 10.4153/CMB-2002-049-1. Dedicated to Robert V. Moody. MR1941222 (2004a:11120)

[3] J. Arthur, *The Endoscopic Classification of Representations: Orthogonal and Symplectic Groups*, Colloquium Publication Series, AMS (to appear).

[4] Nicolas Bergeron, *Le spectre des surfaces hyperboliques*, Savoirs Actuels (Les Ulis). [Current Scholarship (Les Ulis)], EDP Sciences, Les Ulis, 2011 (French). MR2857626

[5] Nicolas Bergeron and Laurent Clozel, *Spectre automorphe des variétés hyperboliques et applications topologiques*, Astérisque **303** (2005), xx+218 (French, with English and French summaries). MR2245761 (2007j:22031)

[6] N. Bergeron and L. Clozel, *Quelques conséquences des travaux d'Arthur pour le spectre et la topologie des variétés hyperboliques*, Invent. Math., to appear.

[7] Valentin Blomer, *Ternary quadratic forms, and sums of three squares with restricted variables*, Anatomy of integers, CRM Proc. Lecture Notes, vol. 46, Amer. Math. Soc., Providence, RI, 2008, pp. 1–17. MR2437962 (2010i:11046)

[8] V. Blomer, *Applications of the Kuznetsov formula on* GL(3), preprint `arXiv:1205.1781`.

[9] Valentin Blomer and Farrell Brumley, *On the Ramanujan conjecture over number fields*, Ann. of Math. (2) **174** (2011), no. 1, 581–605, DOI 10.4007/annals.2011.174.1.18. MR2811610

[10] V. Blomer and F. Brumley, *Non-vanishing of L-functions, the Ramanujan conjecture, and families of Hecke characters*, Canadian J. Math, to appear.

[11] Valentin Blomer and Gergely Harcos, *Hybrid bounds for twisted L-functions*, J. Reine Angew. Math. **621** (2008), 53–79, DOI 10.1515/CRELLE.2008.058. MR2431250 (2009e:11094)

[12] Valentin Blomer and Gergely Harcos, *Twisted L-functions over number fields and Hilbert's eleventh problem*, Geom. Funct. Anal. **20** (2010), no. 1, 1–52, DOI 10.1007/s00039-010-0063-x. MR2647133 (2011g:11090)

[13] Andrew R. Booker and Andreas Strömbergsson, *Numerical computations with the trace formula and the Selberg eigenvalue conjecture*, J. Reine Angew. Math. **607** (2007), 113–161, DOI 10.1515/CRELLE.2007.047. MR2338122 (2008g:11089)

[14] Andrew R. Booker, Andreas Strömbergsson, and Akshay Venkatesh, *Effective computation of Maass cusp forms*, Int. Math. Res. Not., posted on 2006, Art. ID 71281, 34, DOI 10.1155/IMRN/2006/71281. MR2249995 (2007i:11073)

[15] Robert Brooks, *Some relations between spectral geometry and number theory*, Topology '90 (Columbus, OH, 1990), Ohio State Univ. Math. Res. Inst. Publ., vol. 1, de Gruyter, Berlin, 1992, pp. 61–75. MR1184403 (93j:58134)

[16] R. W. Bruggeman, *Fourier coefficients of cusp forms*, Invent. Math. **45** (1978), no. 1, 1–18. MR0472701 (57 #12394)

[17] R. W. Bruggeman, *Automorphic forms*, Elementary and analytic theory of numbers (Warsaw, 1982), Banach Center Publ., vol. 17, PWN, Warsaw, 1985, pp. 31–74. MR840472 (87h:11037)

[18] F. Brumley and A. Venkatesh, *Effective equidistribution on homogeneous spaces and applications*, in preparation.

[19] Daniel Bump, *Automorphic forms and representations*, Cambridge Studies in Advanced Mathematics, vol. 55, Cambridge University Press, Cambridge, 1997. MR1431508 (97k:11080)

[20] Daniel Bump, Solomon Friedberg, and Dorian Goldfeld, *Poincaré series and Kloosterman sums for* $\mathrm{SL}(3,\mathbf{Z})$, Acta Arith. **50** (1988), no. 1, 31–89. MR945275 (89j:11047)

[21] Daniel Bump and David Ginzburg, *Symmetric square L-functions on* $\mathrm{GL}(r)$, Ann. of Math. (2) **136** (1992), no. 1, 137–205, DOI 10.2307/2946548. MR1173928 (93i:11058)

[22] M. Burger, J.-S. Li, and P. Sarnak, *Ramanujan duals and automorphic spectrum*, Bull. Amer. Math. Soc. (N.S.) **26** (1992), no. 2, 253–257, DOI 10.1090/S0273-0979-1992-00267-7. MR1118700 (92h:22023)

[23] D. A. Burgess, *On character sums and L-series. II*, Proc. London Math. Soc. (3) **13** (1963), 524–536. MR0148626 (26 #6133)

[24] Kevin Buzzard, Mark Dickinson, Nick Shepherd-Barron, and Richard Taylor, *On icosahedral Artin representations*, Duke Math. J. **109** (2001), no. 2, 283–318, DOI 10.1215/S0012-7094-01-10922-8. MR1845181 (2002k:11078)

[25] A. Caraiani, *Local-global compatibility and the action of monodromy on nearby cycles,* Duke Math. J. 161 (2012), 2311-2413.

[26] Henri Carayol, *Limites dégénérées de séries discrètes, formes automorphes et variétés de Griffiths-Schmid: le cas du groupe* $\mathrm{U}(2,1)$, Compositio Math. **111** (1998), no. 1, 51–88, DOI 10.1023/A:1000282229017 (French, with English summary). MR1611063 (99k:22027)

[27] K. Chandrasekharan and Raghavan Narasimhan, *Functional equations with multiple gamma factors and the average order of arithmetical functions*, Ann. of Math. (2) **76** (1962), 93–136. MR0140491 (25 #3911)

[28] L. Clozel, *Spectral theory of automorphic forms,* IAS/Park City Lecture Notes, Park City, Utah, 2002, 41-94.

[29] Laurent Clozel, *Motifs et formes automorphes: applications du principe de fonctorialité*, Automorphic forms, Shimura varieties, and L-functions, Vol. I (Ann Arbor, MI, 1988), Perspect. Math., vol. 10, Academic Press, Boston, MA, 1990, pp. 77–159 (French). MR1044819 (91k:11042)

[30] Laurent Clozel, Michael Harris, and Richard Taylor, *Automorphy for some l-adic lifts of automorphic mod l Galois representations*, Publ. Math. Inst. Hautes Études Sci. **108** (2008), 1–181, DOI 10.1007/s10240-008-0016-1. With Appendix A, summarizing unpublished work of Russ Mann, and Appendix B by Marie-France Vignéras. MR2470687 (2010j:11082)

[31] Laurent Clozel, Hee Oh, and Emmanuel Ullmo, *Hecke operators and equidistribution of Hecke points*, Invent. Math. **144** (2001), no. 2, 327–351, DOI 10.1007/s002220100126. MR1827734 (2002m:11044)

[32] Laurent Clozel and Emmanuel Ullmo, *Équidistribution des points de Hecke*, Contributions to automorphic forms, geometry, and number theory, Johns Hopkins Univ. Press, Baltimore, MD, 2004, pp. 193–254 (French). MR2058609 (2005f:11090)

[33] J. W. Cogdell, *Langlands conjectures for* GL_n, An introduction to the Langlands program (Jerusalem, 2001), Birkhäuser Boston, Boston, MA, 2003, pp. 229–249. MR1990381

[34] James W. Cogdell, *On sums of three squares*, J. Théor. Nombres Bordeaux **15** (2003), no. 1, 33–44 (English, with English and French summaries). Les XXIIèmes Journées Arithmetiques (Lille, 2001). MR2018999 (2005d:11072)

[35] J. W. Cogdell, H. H. Kim, I. I. Piatetski-Shapiro, and F. Shahidi, *Functoriality for the classical groups*, Publ. Math. Inst. Hautes Études Sci. **99** (2004), 163–233, DOI 10.1007/s10240-004-0020-z. MR2075885 (2006a:22010)

[36] M. Cowling, U. Haagerup, and R. Howe, *Almost L^2 matrix coefficients*, J. Reine Angew. Math. **387** (1988), 97–110. MR946351 (89i:22008)

[37] Pierre Deligne, *La conjecture de Weil. I*, Inst. Hautes Études Sci. Publ. Math. **43** (1974), 273–307 (French). MR0340258 (49 #5013)

[38] P. Deligne, *Cohomologie étale*, Lecture Notes in Mathematics, Vol. 569, Springer-Verlag, Berlin, 1977. Séminaire de Géométrie Algébrique du Bois-Marie SGA $4\frac{1}{2}$; Avec la collaboration de J. F. Boutot, A. Grothendieck, L. Illusie et J. L. Verdier. MR0463174 (57 #3132)

[39] P. Deligne, *Formes modulaires et représentations l-adiques*, Séminaire Bourbaki vol. 1968/69 Exposés 347-363, Lecture Notes in Mathematics 179, Springer-Verlag.

[40] Pierre Deligne and Jean-Pierre Serre, *Formes modulaires de poids* 1, Ann. Sci. École Norm. Sup. (4) **7** (1974), 507–530 (1975) (French). MR0379379 (52 #284)

[41] Patrick Delorme, *Théorème de Paley-Wiener pour les fonctions de Whittaker sur un groupe réductif p-adique*, J. Inst. Math. Jussieu **11** (2012), no. 3, 501–568, DOI 10.1017/S1474748011000193 (French, with English and French summaries). MR2931317

[42] Fred Diamond and Jerry Shurman, *A first course in modular forms*, Graduate Texts in Mathematics, vol. 228, Springer-Verlag, New York, 2005. MR2112196 (2006f:11045)

[43] W. Duke, J. B. Friedlander, and H. Iwaniec, *The subconvexity problem for Artin L-functions*, Invent. Math. **149** (2002), no. 3, 489–577, DOI 10.1007/s002220200223. MR1923476 (2004e:11046)

[44] W. Duke and H. Iwaniec, *Estimates for coefficients of L-functions. I*, Automorphic forms and analytic number theory (Montreal, PQ, 1989), Univ. Montréal, Montreal, QC, 1990, pp. 43–47. MR1111010 (92f:11068)

[45] William Duke and Rainer Schulze-Pillot, *Representation of integers by positive ternary quadratic forms and equidistribution of lattice points on ellipsoids*, Invent. Math. **99** (1990), no. 1, 49–57, DOI 10.1007/BF01234411. MR1029390 (90m:11051)

[46] Martin Eichler, *Quaternäre quadratische Formen und die Riemannsche Vermutung für die Kongruenzzetafunktion*, Arch. Math. **5** (1954), 355–366 (German). MR0063406 (16,116d)

[47] T. Estermann, *Vereinfachter Beweis eines Satzes von Kloosterman*, Abh. Math. Sem. Univ. Hamburg 7 (1929) 82-98.

[48] D. Flath, *Decomposition of representations into tensor products*, Automorphic forms, representations and L-functions (Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977), Part 1, Proc. Sympos. Pure Math., XXXIII, Amer. Math. Soc., Providence, R.I., 1979, pp. 179–183. MR546596 (81f:22028)

[49] Stephen S. Gelbart, *Automorphic forms on adèle groups*, Princeton University Press, Princeton, N.J., 1975. Annals of Mathematics Studies, No. 83. MR0379375 (52 #280)

[50] Stephen Gelbart and Hervé Jacquet, *A relation between automorphic representations of* GL(2) *and* GL(3), Ann. Sci. École Norm. Sup. (4) **11** (1978), no. 4, 471–542. MR533066 (81e:10025)

[51] A. Ghosh, A. Gorodnik, and A. Nevo, *Diophantine approximation and automorphic spectrum*, preprint.

[52] David Ginzburg, Stephen Rallis, and David Soudry, *Generic automorphic forms on* $\mathrm{SO}(2n+1)$*: functorial lift to* $\mathrm{GL}(2n)$*, endoscopy, and base change*, Internat. Math. Res. Notices **14** (2001), 729–764, DOI 10.1155/S1073792801000381. MR1846354 (2002g:11065)

[53] Dorian Goldfeld, *Automorphic forms and L-functions for the group* $\mathrm{GL}(n,\mathbf{R})$, Cambridge Studies in Advanced Mathematics, vol. 99, Cambridge University Press, Cambridge, 2006. With an appendix by Kevin A. Broughan. MR2254662 (2008d:11046)

[54] Dorian Goldfeld and Joseph Hundley, *Automorphic representations and L-functions for the general linear group. Volume I*, Cambridge Studies in Advanced Mathematics, vol. 129, Cambridge University Press, Cambridge, 2011. With exercises and a preface by Xander Faber. MR2807433 (2012i:11054)

[55] Alexander Gorodnik and Amos Nevo, *The ergodic theory of lattice subgroups*, Annals of Mathematics Studies, vol. 172, Princeton University Press, Princeton, NJ, 2010. MR2573139 (2011c:22006)

[56] G. Harder and D. A. Kazhdan, *Automorphic forms on* GL_2 *over function fields (after V. G. Drinfel′d)*, Automorphic forms, representations and L-functions (Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977), Part 2, Proc. Sympos. Pure Math., XXXIII, Amer. Math. Soc., Providence, R.I., 1979, pp. 357–379. MR546624 (83e:12008)

[57] Michael Harris and Richard Taylor, *The geometry and cohomology of some simple Shimura varieties*, Annals of Mathematics Studies, vol. 151, Princeton University Press, Princeton, NJ, 2001. With an appendix by Vladimir G. Berkovich. MR1876802 (2002m:11050)

[58] E. Hecke, *Zur Theorie der elliptischen Modulfunktionen*, Math. Ann. **97** (1927), no. 1, 210–242, DOI 10.1007/BF01447866 (German). MR1512360

[59] Christopher Hooley, *On Artin's conjecture*, J. Reine Angew. Math. **225** (1967), 209–220. MR0207630 (34 #7445)

[60] R. Howe and I. I. Piatetski-Shapiro, *A counterexample to the "generalized Ramanujan conjecture" for (quasi-) split groups*, Automorphic forms, representations and L-functions (Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977), Part 1, Proc. Sympos. Pure Math., XXXIII, Amer. Math. Soc., Providence, R.I., 1979, pp. 315–322. MR546605 (81f:22036)

[61] M. N. Huxley, *Introduction to Kloostermania*, Elementary and analytic theory of numbers (Warsaw, 1982), Banach Center Publ., vol. 17, PWN, Warsaw, 1985, pp. 217–306. MR840479 (87j:11046)

[62] Henryk Iwaniec, *The lowest eigenvalue for congruence groups*, Topics in geometry, Progr. Nonlinear Differential Equations Appl., vol. 20, Birkhäuser Boston, Boston, MA, 1996, pp. 203–212. MR1390315 (97e:11058)

[63] Henryk Iwaniec, *Topics in classical automorphic forms*, Graduate Studies in Mathematics, vol. 17, American Mathematical Society, Providence, RI, 1997. MR1474964 (98e:11051)

[64] Henryk Iwaniec and Emmanuel Kowalski, *Analytic number theory*, American Mathematical Society Colloquium Publications, vol. 53, American Mathematical Society, Providence, RI, 2004. MR2061214 (2005h:11005)

[65] H. Jacquet and R. P. Langlands, *Automorphic forms on* $\mathrm{GL}(2)$, Lecture Notes in Mathematics, Vol. 114, Springer-Verlag, Berlin, 1970. MR0401654 (53 #5481)

[66] H. Jacquet, I. I. Piatetskii-Shapiro, and J. A. Shalika, *Rankin-Selberg convolutions*, Amer. J. Math. **105** (1983), no. 2, 367–464, DOI 10.2307/2374264. MR701565 (85g:11044)

[67] H. Jacquet and J. A. Shalika, *On Euler products and the classification of automorphic representations. I*, Amer. J. Math. **103** (1981), no. 3, 499–558, DOI 10.2307/2374103. MR618323 (82m:10050a)

[68] Chandrashekhar Khare, *Remarks on mod p forms of weight one*, Internat. Math. Res. Notices **3** (1997), 127–133, DOI 10.1155/S1073792897000093. MR1434905 (97m:11070)

[69] Chandrashekhar Khare and Jean-Pierre Wintenberger, *On Serre's conjecture for* 2*-dimensional* mod p *representations of* $\mathrm{Gal}(\overline{\mathbb{Q}}/\mathbb{Q})$, Ann. of Math. (2) **169** (2009), no. 1, 229–253, DOI 10.4007/annals.2009.169.229. MR2480604 (2009m:11077)

[70] Chandrashekhar Khare and Jean-Pierre Wintenberger, *Serre's modularity conjecture. I*, Invent. Math. **178** (2009), no. 3, 485–504, DOI 10.1007/s00222-009-0205-7. MR2551763 (2010k:11087)

[71] Chandrashekhar Khare and Jean-Pierre Wintenberger, *Serre's modularity conjecture. II*, Invent. Math. **178** (2009), no. 3, 505–586, DOI 10.1007/s00222-009-0206-6. MR2551764 (2010k:11088)

[72] Henry H. Kim, *Langlands-Shahidi method and poles of automorphic L-functions: application to exterior square L-functions*, Canad. J. Math. **51** (1999), no. 4, 835–849, DOI 10.4153/CJM-1999-036-0. MR1701344 (2000f:11058)

[73] Henry H. Kim, *Functoriality for the exterior square of* GL_4 *and the symmetric fourth of* GL_2, J. Amer. Math. Soc. **16** (2003), no. 1, 139–183 (electronic), DOI 10.1090/S0894-0347-02-00410-1. With appendix 1 by Dinakar Ramakrishnan and appendix 2 by Kim and Peter Sarnak. MR1937203 (2003k:11083)

[74] Henry H. Kim, *On local L-functions and normalized intertwining operators*, Canad. J. Math. **57** (2005), no. 3, 535–597, DOI 10.4153/CJM-2005-023-x. MR2134402 (2006a:11063)

[75] Henry H. Kim and Freydoon Shahidi, *Cuspidality of symmetric powers with applications*, Duke Math. J. **112** (2002), no. 1, 177–197, DOI 10.1215/S0012-9074-02-11215-0. MR1890650 (2003a:11057)

[76] Henry H. Kim and Freydoon Shahidi, *Functorial products for* $\mathrm{GL}_2 \times \mathrm{GL}_3$ *and the symmetric cube for* GL_2, Ann. of Math. (2) **155** (2002), no. 3, 837–893, DOI 10.2307/3062134. With an appendix by Colin J. Bushnell and Guy Henniart. MR1923967 (2003m:11075)

[77] Felix Klein, *Vorlesungen über das Ikosaeder und die Auflösung der Gleichungen vom fünften Grade*, Birkhäuser Verlag, Basel, 1993 (German, with German summary). Reprint of the 1884 original; Edited, with an introduction and commentary by Peter Slodowy. MR1315530 (96g:01046)

[78] H. Kloosterman, *On the representation of numbers in the form* $ax^2 + by^2 + cz^2 + dt^2$. Acta. Math. 49 (1926), 407-464.

[79] H. Kloosterman, *Asymptotische Formeln für die Fourierkoeffizienten ganzer Modulformen*, Abh. Math. Sem. Univ Hamburg 5 (1927), 337-352.

[80] Neal Koblitz, *Introduction to elliptic curves and modular forms*, 2nd ed., Graduate Texts in Mathematics, vol. 97, Springer-Verlag, New York, 1993. MR1216136 (94a:11078)

[81] Nobushige Kurokawa, *Examples of eigenvalues of Hecke operators on Siegel cusp forms of degree two*, Invent. Math. **49** (1978), no. 2, 149–165, DOI 10.1007/BF01403084. MR511188 (80b:10040)

[82] N. V. Kuznecov, *The Petersson conjecture for cusp forms of weight zero and the Linnik conjecture. Sums of Kloosterman sums*, Mat. Sb. (N.S.) **111(153)** (1980), no. 3, 334–383, 479 (Russian). MR568983 (81m:10053)

[83] J.-P. Labesse and R. P. Langlands, *L-indistinguishability for* SL(2), Canad. J. Math. **31** (1979), no. 4, 726–785, DOI 10.4153/CJM-1979-070-3. MR540902 (81b:22017)

[84] E. Landau, *Über die Anzahl der Gitterpunkte in gewissen Bereichen II*, Nachr. v. d. Gesellschaft d. Wiss. zu Göttingen, Math.-Phys. Klasse 1915, 209-243.

[85] R. P. Langlands, *Problems in the theory of automorphic forms*, Lectures in Modern Analysis and Applications, III, Springer, Berlin, 1970, pp. 18–61. Lecture Notes in Math., Vol. 170. MR0302614 (46 #1758)

[86] Robert P. Langlands, *Base change for* GL(2), Annals of Mathematics Studies, vol. 96, Princeton University Press, Princeton, N.J., 1980. MR574808 (82a:10032)

[87] Erez Lapid and Zhengyu Mao, *On the asymptotics of Whittaker functions*, Represent. Theory **13** (2009), 63–81, DOI 10.1090/S1088-4165-09-00343-4. MR2495561 (2010b:22024)

[88] Erez Lapid and Werner Müller, *Spectral asymptotics for arithmetic quotients of* $\mathrm{SL}(n, \mathbb{R})/\mathrm{SO}(n)$, Duke Math. J. **149** (2009), no. 1, 117–155, DOI 10.1215/00127094-2009-037. MR2541128 (2010h:11083)

[89] H. W. Lenstra Jr., *On Artin's conjecture and Euclid's algorithm in global fields*, Invent. Math. **42** (1977), 201–224. MR0480413 (58 #576)

[90] Elon Lindenstrauss and Akshay Venkatesh, *Existence and Weyl's law for spherical cusp forms*, Geom. Funct. Anal. **17** (2007), no. 1, 220–251, DOI 10.1007/s00039-006-0589-0. MR2306657 (2008c:22016)

[91] W. Luo, Z. Rudnick, and P. Sarnak, *On Selberg's eigenvalue conjecture*, Geom. Funct. Anal. **5** (1995), no. 2, 387–401, DOI 10.1007/BF01895672. MR1334872 (96h:11045)

[92] Wenzhi Luo, Zeév Rudnick, and Peter Sarnak, *On the generalized Ramanujan conjecture for* $\mathrm{GL}(n)$, Automorphic forms, automorphic representations, and arithmetic (Fort Worth, TX, 1996), Proc. Sympos. Pure Math., vol. 66, Amer. Math. Soc., Providence, RI, 1999, pp. 301–310. MR1703764 (2000e:11072)

[93] Hans Maass, *Über eine neue Art von nichtanalytischen automorphen Funktionen und die Bestimmung Dirichletscher Reihen durch Funktionalgleichungen*, Math. Ann. **121** (1949), 141–183 (German). MR0031519 (11,163c)

[94] Philippe Michel, *Analytic number theory and families of automorphic L-functions*, Automorphic forms and applications, IAS/Park City Math. Ser., vol. 12, Amer. Math. Soc., Providence, RI, 2007, pp. 181–295. MR2331346 (2008m:11104)

[95] S. Minakshisundaram and Å. Pleijel, *Some properties of the eigenfunctions of the Laplace-operator on Riemannian manifolds*, Canadian J. Math. **1** (1949), 242–256. MR0031145 (11,108b)

[96] C-P. Mok, *Endoscopic classification of representations of quasi-split unitary groups I,* arXiv:1206.0882v1.

[97] L. Mordell, *On Mr. Ramanujan's empirical expansions of modular functions*, Cambr. Phil. Soc. Proc. 19 (1917), 117-124.

[98] W. Müller and B. Speh, *Absolute convergence of the spectral side of the Arthur trace formula for* GL_n, Geom. Funct. Anal. **14** (2004), no. 1, 58–93, DOI 10.1007/s00039-004-0452-0. With an appendix by E. M. Lapid. MR2053600 (2005m:22021)

[99] M. Ram Murty, *On the estimation of eigenvalues of Hecke operators*, Rocky Mountain J. Math. **15** (1985), no. 2, 521–533, DOI 10.1216/RMJ-1985-15-2-521. Number theory (Winnipeg, Man., 1983). MR823263 (87j:11037)

[100] M. Ram Murty, *Applications of symmetric power L-functions*, Lectures on Automorphic L-functions, Fields Inst. Monogr., vol. 20, Amer. Math. Soc., Providence, RI, 2004, pp. 203–283. MR2071508

[101] M. Ram Murty and V. Kumar Murty, *A variant of the Bombieri-Vinogradov theorem*, Number Theory (Montreal, Que., 1985), CMS Conf. Proc., vol. 7, Amer. Math. Soc., Providence, RI, 1987, pp. 243–272. MR894326 (88h:11087)

[102] Maki Nakasuji, *Generalized Ramanujan conjecture over general imaginary quadratic fields*, Forum Math. **24** (2012), no. 1, 85–98, DOI 10.1515/form.2011.050. MR2879972

[103] Bao Châu Ngô, *Le lemme fondamental pour les algèbres de Lie*, Publ. Math. Inst. Hautes Études Sci. **111** (2010), 1–169, DOI 10.1007/s10240-010-0026-7 (French). MR2653248 (2011h:22011)

[104] Hee Oh, *Uniform pointwise bounds for matrix coefficients of unitary representations and applications to Kazhdan constants*, Duke Math. J. **113** (2002), no. 1, 133–192, DOI 10.1215/S0012-7094-02-11314-3. MR1905394 (2003d:22015)

[105] Hans Petersson, *Über die Entwicklungskoeffizienten der automorphen Formen*, Acta Math. **58** (1932), no. 1, 169–215, DOI 10.1007/BF02547776 (German). MR1555346

[106] Hans Petersson, *Konstruktion der sämtlichen Lösungen einer Riemannschen Funktionalgleichung durch Dirichlet-Reihen mit Eulerscher Produktentwicklung. II*, Math. Ann. **117** (1939), 39–64 (German). MR0001768 (1,294c)

[107] Hans Petersson, *Über eine Metrisierung der automorphen Formen und die Theorie der Poincaréschen Reihen*, Math. Ann. **117** (1940), 453–537 (German). MR0002624 (2,87e)

[108] I. I. Pjateckij-Šapiro, *Euler subgroups*, Lie Groups and Their Representations (Proc. Summer School, Bolyai János Math. Soc., Budapest, 1971), Halsted, New York, 1975, pp. 597–620. MR0406935 (53 #10720)

[109] Dinakar Ramakrishnan, *Modularity of the Rankin-Selberg L-series, and multiplicity one for* SL(2), Ann. of Math. (2) **152** (2000), no. 1, 45–111, DOI 10.2307/2661379. MR1792292 (2001g:11077)

[110] D. Ramakrishnan, *A mild Chebotarev theorem for* $\mathrm{GL}(n)$, preprint `arXiv:1003.4498`.

[111] S. Ramanujan, *On certain arithmetical functions*, Trans. Cambridge Phil. Soc. 22 (1916), 159-184.

[112] Burton Randol, *Small eigenvalues of the Laplace operator on compact Riemann surfaces*, Bull. Amer. Math. Soc. **80** (1974), 996–1000. MR0400316 (53 #4151)

[113] R. A. Rankin, *Contributions to the theory of Ramanujan's function* $\tau(n)$ *and similar arithmetical functions. III. A note on the sum function of the Fourier coefficients of integral modular forms*, Proc. Cambridge Philos. Soc. **36** (1940), 150–151. MR0001249 (1,203d)

[114] Walter Roelcke, *Über die Wellengleichung bei Grenzkreisgruppen erster Art*, S.-B. Heidelberger Akad. Wiss. Math.-Nat. Kl. **1953/1955** (1953/1955), 159–267 (1956) (German). MR0081967 (18,476d)

[115] David E. Rohrlich, *Nonvanishing of L-functions for* GL(2), Invent. Math. **97** (1989), no. 2, 381–403, DOI 10.1007/BF01389047. MR1001846 (90g:11062)

[116] Zeév Rudnick and Peter Sarnak, *Zeros of principal L-functions and random matrix theory*, Duke Math. J. **81** (1996), no. 2, 269–322, DOI 10.1215/S0012-7094-96-08115-6. A celebration of John F. Nash, Jr. MR1395406 (97f:11074)

[117] Y. Sakellaridis, A. Venkatesh, *Periods and harmonic analysis on spherical varieties*, preprint `arXiv:1203.0039`.

[118] Hans Salié, *Zur Abschätzung der Fourierkoeffizienten ganzer Modulformen*, Math. Z. **36** (1933), no. 1, 263–278, DOI 10.1007/BF01188622 (German). MR1545344

[119] Peter Sarnak, *Asymptotic behavior of periodic orbits of the horocycle flow and Eisenstein series*, Comm. Pure Appl. Math. **34** (1981), no. 6, 719–739, DOI 10.1002/cpa.3160340602. MR634284 (83m:58060)

[120] Peter C. Sarnak, *Diophantine problems and linear groups*, II (Kyoto, 1990), Math. Soc. Japan, Tokyo, 1991, pp. 459–471. MR1159234 (93g:11054)

[121] Peter Sarnak, *Maass cusp forms with integer coefficients*, A Panorama of Number Theory or the View from Baker's Garden (Zürich, 1999), Cambridge Univ. Press, Cambridge, 2002, pp. 121–127, DOI 10.1017/CBO9780511542961.009. MR1975448 (2004c:11053)

[122] Peter Sarnak, *Spectra of hyperbolic surfaces*, Bull. Amer. Math. Soc. (N.S.) **40** (2003), no. 4, 441–478 (electronic), DOI 10.1090/S0273-0979-03-00991-1. MR1997348 (2004f:11107)

[123] Peter Sarnak, *Notes on the generalized Ramanujan conjectures*, Harmonic Analysis, the Trace Formula, and Shimura Varieties, Clay Math. Proc., vol. 4, Amer. Math. Soc., Providence, RI, 2005, pp. 659–685. MR2192019 (2007a:11067)

[124] Peter Sarnak, *What is... an expander?*, Notices Amer. Math. Soc. **51** (2004), no. 7, 762–763. MR2072849

[125] Peter Sarnak and Jacob Tsimerman, *On Linnik and Selberg's conjecture about sums of Kloosterman sums*, Algebra, Arithmetic, and Geometry: in honor of Yu. I. Manin. Vol. II, Progr. Math., vol. 270, Birkhäuser Boston Inc., Boston, MA, 2009, pp. 619–635, DOI 10.1007/978-0-8176-4747-6_20. MR2641204 (2011g:11152)

[126] Peter Sarnak and Xiao Xi Xue, *Bounds for multiplicities of automorphic representations*, Duke Math. J. **64** (1991), no. 1, 207–227, DOI 10.1215/S0012-7094-91-06410-0. MR1131400 (92h:22026)

[127] Atle Selberg, *Bemerkungen über eine Dirichletsche Reihe, die mit der Theorie der Modulformen nahe verbunden ist*, Arch. Math. Naturvid. **43** (1940), 47–50 (German). MR0002626 (2,88a)

[128] A. 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. MR0088511 (19,531g)

[129] Atle Selberg, *On the estimation of Fourier coefficients of modular forms*, Proc. Sympos. Pure Math., Vol. VIII, Amer. Math. Soc., Providence, R.I., 1965, pp. 1–15. MR0182610 (32 #93)

[130] J.-P. Serre, *A course in arithmetic*, Springer-Verlag, New York, 1973. Translated from the French; Graduate Texts in Mathematics, No. 7. MR0344216 (49 #8956)

[131] J.-P. Serre, letter to Deshouillers (1981).

[132] Freydoon Shahidi, *On the Ramanujan conjecture and finiteness of poles for certain L-functions*, Ann. of Math. (2) **127** (1988), no. 3, 547–584, DOI 10.2307/2007005. MR942520 (89h:11021)

[133] Freydoon Shahidi, *Langlands functoriality conjecture and number theory*, Representation Theory and Automorphic Forms, Progr. Math., vol. 255, Birkhäuser Boston, Boston, MA, 2008, pp. 151–173, DOI 10.1007/978-0-8176-4646-2_5. MR2369498 (2009c:11073)

[134] Freydoon Shahidi, *Arthur packets and the Ramanujan conjecture*, Kyoto J. Math. **51** (2011), no. 1, 1–23, DOI 10.1215/0023608X-2010-018. MR2784745

[135] J. A. Shalika, *The multiplicity one theorem for* GL_n, Ann. of Math. (2) **100** (1974), 171–193. MR0348047 (50 #545)

[136] Goro Shimura, *On the holomorphy of certain Dirichlet series*, Proc. London Math. Soc. (3) **31** (1975), no. 1, 79–98. MR0382176 (52 #3064)

[137] Sug Woo Shin, *Galois representations arising from some compact Shimura varieties*, Ann. of Math. (2) **173** (2011), no. 3, 1645–1741, DOI 10.4007/annals.2011.173.3.9. MR2800722

[138] Takuro Shintani, *On an explicit formula for class-1 "Whittaker functions" on GL_n over P-adic fields*, Proc. Japan Acad. **52** (1976), no. 4, 180–182. MR0407208 (53 #10991)

[139] Eric Stade, *Archimedean L-factors on* $\mathrm{GL}(n) \times \mathrm{GL}(n)$ *and generalized Barnes integrals*, Israel J. Math. **127** (2002), 201–219, DOI 10.1007/BF02784531. MR1900699 (2003f:11071)

[140] Andreas Strömbergsson, *On the uniform equidistribution of long closed horocycles*, Duke Math. J. **123** (2004), no. 3, 507–547, DOI 10.1215/S0012-7094-04-12334-6. MR2068968 (2005f:11105)

[141] S. Takeda, *The twisted symmetric square L-function of* $\mathrm{GL}(r)$, preprint, `arXiv:1005.1979v7`.

[142] J. T. Tate, *Fourier analysis in number fields, and Hecke's zeta-functions*, Algebraic Number Theory (Proc. Instructional Conf., Brighton, 1965), Thompson, Washington, D.C., 1967, pp. 305–347. MR0217026 (36 #121)

[143] J. Tate, *Number theoretic background*, Automorphic Forms, Representations and L-functions (Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977), Part 2, Proc. Sympos. Pure Math., XXXIII, Amer. Math. Soc., Providence, R.I., 1979, pp. 3–26. MR546607 (80m:12009)

[144] Richard Taylor, *Galois representations*, Ann. Fac. Sci. Toulouse Math. (6) **13** (2004), no. 1, 73–119 (English, with English and French summaries). MR2060030 (2005a:11071)

[145] Richard Taylor, *Automorphy for some l-adic lifts of automorphic mod l Galois representations. II*, Publ. Math. Inst. Hautes Études Sci. **108** (2008), 183–239, DOI 10.1007/s10240-008-0015-2. MR2470688 (2010j:11085)

[146] Richard Taylor, *On icosahedral Artin representations. II*, Amer. J. Math. **125** (2003), no. 3, 549–566. MR1981033 (2004e:11057)

[147] Jerrold Tunnell, *Artin's conjecture for representations of octahedral type*, Bull. Amer. Math. Soc. (N.S.) **5** (1981), no. 2, 173–175, DOI 10.1090/S0273-0979-1981-14936-3. MR621884 (82j:12015)

[148] Nolan R. Wallach, *Real reductive groups. II*, Pure and Applied Mathematics, vol. 132, Academic Press Inc., Boston, MA, 1992. MR1170566 (93m:22018)

[149] André Weil, *On some exponential sums*, Proc. Nat. Acad. Sci. U. S. A. **34** (1948), 204–207. MR0027006 (10,234e)

[150] André Weil, *Über die Bestimmung Dirichletscher Reihen durch Funktionalgleichungen*, Math. Ann. **168** (1967), 149–156 (German). MR0207658 (34 #7473)

[151] P.-J., White, *Tempered automorphic representations of the unitary group,* preprint `arXiv:1106.1127`.

[152] E. T. Whittaker and G. N. Watson, *A course of modern analysis*, Cambridge Mathematical Library, Cambridge University Press, Cambridge, 1996. An Introduction to the General Theory of Infinite Processes and of Analytic Functions; with an account of the principal transcendental functions; Reprint of the fourth (1927) edition. MR1424469 (97k:01072)

[153] P. G. Zograf, *Small eigenvalues of automorphic Laplacians in spaces of cusp forms*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI) **134** (1984), 157–168 (Russian, with English summary). Automorphic functions and number theory, II. MR741858 (86a:58116)

Mathematisches Institut, Bunsenstr. 3-5, 37073 Göttingen
E-mail address: `blomer@uni-math.gwdg.de`

Institut Galilée, Université Paris 13, 99 avenue J.-B. Clément, 93430 Villetaneuse, France
E-mail address: `brumley@math.univ-paris13.fr`