BOUNDS FOR RANKIN–SELBERG INTEGRALS AND QUANTUM UNIQUE ERGODICITY FOR POWERFUL LEVELS

PAUL D. NELSON, AMEYA PITALE, AND ABHISHEK SAHA

ABSTRACT. Let f be a classical holomorphic newform of level q and even weight k. We show that the pushforward to the full level modular curve of the mass of f equidistributes as $qk \to \infty$. This generalizes known results in the case that q is squarefree. We obtain a power savings in the rate of equidistribution as q becomes sufficiently "powerful" (far away from being squarefree), and in particular in the "depth aspect" as q traverses the powers of a fixed prime.

We compare the difficulty of such equidistribution problems to that of corresponding subconvexity problems by deriving explicit extensions of Watson's formula to certain triple product integrals involving forms of non-squarefree level. By a theorem of Ichino and a lemma of Michel–Venkatesh, this amounts to a detailed study of Rankin–Selberg integrals $\int |f|^2 E$ attached to newforms f of arbitrary level and Eisenstein series E of full level.

We find that the local factors of such integrals participate in many amusing analogies with global L-functions. For instance, we observe that the mass equidistribution conjecture with a power savings in the depth aspect is equivalent to knowing either a global subconvexity bound or what we call a "local subconvexity bound"; a consequence of our local calculations is what we call a "local Lindelöf hypothesis".

1. Introduction

1.1. **Main result.** Let $f : \mathbb{H} \to \mathbb{C}$ be a classical holomorphic newform of weight $k \in 2\mathbb{N}$ on $\Gamma_0(q)$, $q \in \mathbb{N}$ (see Section 3.1 for definitions). The pushforward to $Y_0(1) = \mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$ of the L^2-mass of f is the finite measure given by

$$\mu_f(\phi) = \int_{\Gamma_0(q)\backslash\mathbb{H}} y^k |f|^2(z)\phi(z)\,\frac{dx\,dy}{y^2}$$

for each bounded measurable function ϕ on $Y_0(1)$. Its value $\mu_f(1)$ at the constant function 1 is (one possible normalization of) the Petersson norm of f. Let $d\mu(z) = y^{-2}dx\,dy$ denote the standard hyperbolic volume measure on $Y_0(1)$, and let

$$D_f(\phi) := \frac{\mu_f(\phi)}{\mu_f(1)} - \frac{\mu(\phi)}{\mu(1)}.$$

The quantity $D_f(\phi)$ compares the probability measures attached to μ_f and μ against a test function ϕ.

The problem of bounding $D_f(\phi)$ for fixed ϕ as the parameters of f vary is a natural analogue of the Rudnick–Sarnak quantum unique ergodicity conjecture [37]. It was raised explicitly in the

---

2010 *Mathematics Subject Classification.* Primary 11F11; Secondary 11F70, 22E50, 58J51.

The first author is supported by NSF grant OISE-1064866 and partially supported by grant SNF-137488. The second author is supported by NSF grant DMS 1100541.

$q = 1$, $k \to \infty$ aspect by Luo–Sarnak [30] and in the $k =$ constant, $q \to \infty$ aspect by Kowalski–Michel–VanderKam [28]; in each case it was conjectured that $D_f(\phi) \to 0$. Such a conjecture is reasonable because a theorem of Watson [46] and subsequent generalizations (see Sections 1.2 and 3.2) have shown that it follows in many cases from the (unproven) Generalized Lindelöf Hypothesis, itself a consequence of the Generalized Riemann Hypothesis.

The first unconditional result for general (non-dihedral) f was obtained by Holowinsky and Soundararajan [18], who showed that $D_f(\phi) \to 0$ for fixed q $(= 1)$ and varying $k \to \infty$; we refer to their paper and [39] for further historical background. The case of varying squarefree levels was addressed in [33], where it was shown that $D_f(\phi) \to 0$ as $qk \to \infty$ provided that q is squarefree.

Our aim in this paper is to address the remaining case in which the varying level q need not be squarefree. We obtain the expected result, thereby settling the remaining cases of the conjecture in [28]:

Theorem 1.1. *Fix a bounded continuous function ϕ on $Y_0(1)$. Let f traverse a sequence of holomorphic newforms of weight k on $\Gamma_0(q)$ with $k \in 2\mathbb{N}$, $q \in \mathbb{N}$. Then $D_f(\phi) \to 0$ whenever $qk \to \infty$.*

Theorem 1.1 is a consequence of the following more precise result and a standard approximation argument (see Section 3.6 and [33, Section 1.6]).

Theorem 1.2. *Fix a Maass eigencuspform or incomplete Eisenstein series ϕ on $Y_0(1)$. Let f traverse a sequence of holomorphic newforms of weight k on $\Gamma_0(q)$ with $k \in 2\mathbb{N}$, $q \in \mathbb{N}$. There exist effective positive constants δ_1, δ_2 so that*[1]

$$D_f(\phi) \ll_\phi (q/q_0)^{-\delta_1} \log(qk)^{-\delta_2}, \tag{1}$$

where q_0 denotes the largest squarefree divisor of q.[2]

A potentially surprising aspect of Theorem 1.2 is the unconditional power savings in the rate of equidistribution when q/q_0 grows faster than a certain fixed power of $\log(q_0 k)$, or in words, when the level is sufficiently *powerful*. A special case that illustrates the new phenomena is the *depth aspect*, in which k is fixed and $q = p^n$ is the power of a fixed prime p with $n \to \infty$.

By contrast, suppose that q is squarefree, so that $q = q_0$. Then the logarithmic rate of decay $D_f(\phi) \ll_\phi \log(qk)^{-\delta_2}$ in Theorem 1.2 is consistent with that obtained in [18, 33], and the problem of improving this logarithmic decay to a power savings $D_f(\phi) \ll_\phi (qk)^{-\delta_3}$ $(\delta_3 > 0)$ is equivalent to the (still open) subconvexity problem for certain fixed GL(1) or GL(2) twists of the adjoint lift of f to GL(3) (see Section 1.2).

Explaining this "surprise" is a major theme of this paper. It amounts to a detailed study of certain Rankin–Selberg zeta integrals $J_f(s)$ arising as proportionality constants in a formula for $D_f(\phi)$ given by Ichino [19], as simplified by a lemma of Michel–Venkatesh [31, Lemma 3.4.2]. In classical terms, $J_f(s)$ is proportional uniformly for $\mathrm{Re}(s) \geq \delta > 0$ to the meromorphic continuation of the ratio

$$\frac{1}{[\Gamma_0(q) : \Gamma_0(1)]} \frac{\int_{\Gamma_0(q)\backslash\mathbb{H}} y^k |f|^2(z) \left(\sum_{\gamma \in \Gamma_\infty \backslash \Gamma_0(1)} (\mathrm{Im}\gamma z)^s\right) \frac{dx\,dy}{y^2}}{\int_{\Gamma_0(q)\backslash\mathbb{H}} y^k |f|^2(z) \left(\sum_{\gamma \in \Gamma_\infty \backslash \Gamma_0(q)} (\mathrm{Im}\gamma z)^s\right) \frac{dx\,dy}{y^2}}, \tag{2}$$

[1] We use the notation $A \ll_{x,y,z} B$ to signify that there exists a positive constant C, depending at most upon x, y, z, so that $|A| \leq C|B|$.

[2] If q has the prime factorization $q = \prod_p p^{a_p}$, then q_0 has the prime factorization $q_0 = \prod_p p^{\min(a_p,1)}$.

defined initially for $\mathrm{Re}(s) > 1$. The quantity $J_f(s)$ factors as a product over the primes dividing the level:

$$J_f(s) = \prod_{p|q} J_p(s),$$

with each $J_p(s)$ a p-adic zeta integral (see (29)) that differs mildly from a polynomial function of $p^{\pm s}$ and satisfies a functional equation under $s \mapsto 1 - s$.

We find the analytic properties of such integrals to be unexpectedly rich and to participate in many amusing analogies. For instance, we show that the problem of obtaining a positive value of δ_1 in Theorem 1.2 is equivalent to knowing *either* a "global" subconvex bound for an L-value *or* what we call a *local subconvex bound* for $J_f(s)$ (see e.g. Observation 1.4). The main technical result of this paper is a proof of (what we call) the *local Lindelöf hypothesis* for $J_f(s)$, which, naturally, saves nearly a factor of $q^{1/4}$ over the *local convexity bound* on the critical line $\mathrm{Re}(s) = 1/2$ (see Section 1.6). We observe numerically that $J_f(s)$ seems to satisfy a *local Riemann hypothesis* (see Section 1.7), the significance of which remains unclear to us.

Remark 1.3. We comment on the nature of the constants δ_1, δ_2 appearing in Theorem 1.2. One may choose δ_2 very explicitly as in [18, 33], while δ_1 depends upon a bound $\theta \in [0, 7/64]$ (see [27]) towards the Ramanujan conjecture for Maass forms on $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$, with any improvement over the trivial bound $\theta \leq 1/2$ sufficing to yield a positive value of δ_1. For example, in the simplest case that $q = p^{2m}$ is an even power of a prime (the "even depth aspect"), our method leads to the bound

$$D_f(\phi) \ll_k m^{O(1)} (p^m)^{-1/2+\theta} \ll_{k,\varepsilon} (p^m)^{-1/2+\theta+\varepsilon}.$$

Our calculations show that the Ramanujan conjecture for Maass forms together with the Lindelöf hypothesis for fixed GL(1) and GL(2) twists of the adjoint lift of f would imply the stronger bound $D_f(\phi) \ll_{\varepsilon,k} (p^m)^{-1+\varepsilon}$, which should be optimal[3] as far as the exponent is concerned.

Our paper is organized as follows. The remainder of Section 1 is an extended introduction that explains the main ideas of our work. In Section 2, we undertake a detailed study of the local Rankin–Selberg integral attached to a spherical Eisenstein series and the L^2-mass of a newform of arbitrary level. Our calculations yield an explicit extension of Watson's formula (see Theorem 3.1) to certain collections of newforms of not necessarily squarefree level. In Section 3, we study the Fourier coefficients of highly ramified newforms at arbitrary cusps of $\Gamma_0(q)$ (see Section 1.9 for an overview) and apply a variant of the Holowinsky–Soundararajan method to deduce Theorem 1.2.

The results of Section 2 suffice on their own to imply Theorem 1.2 when the level q is sufficiently powerful (e.g., if $q = p^n$ with p fixed and $n \to \infty$). At the other extreme, Theorem 1.2 is already known when q is squarefree (see [33]). It is the myriad of intermediate possibilities (e.g., when $q = q_0 p^n$ is the product of a large squarefree integer q_0 and a large prime power p^n) that justifies Section 3.

1.2. **Equidistribution vs. subconvexity.** The motivating *quantum unique ergodicity* (QUE) conjecture, put forth by Rudnick and Sarnak, predicts that the L^2-normalized Laplace eigenfunctions ϕ on a negatively curved compact Riemannian manifold have equidistributed L^2-mass in the large eigenvalue limit. The *arithmetic QUE* conjecture concerns the special case that ϕ traverses a sequence of joint Hecke-Laplace eigenfunctions on an arithmetic manifold. A formula of Watson

[3]That is to say, it should be the optimal bound that holds for *all* f of level p^{2m}. Stronger bounds will hold, for instance, for ramified character twists of forms of lower level.

showed in many cases that the arithmetic QUE conjecture for surfaces, in a sufficiently strong quantitative form, is *equivalent* to a case of the central *subconvexity problem* in the analytic theory of L-functions. A principal motivation for this work was to investigate the extent to which this equivalence survives the passage to variants of arithmetic QUE not covered by Watson's formula.

In the prototypical case that f is a Maass eigencuspform on $Y_0(1)$ with Laplace eigenvalue λ, the definitions of μ_f and D_f given in Section 1.1 still make sense (take $k = 0$), and the equidistribution problem is to improve upon the trivial bound

$$D_f(\phi) \ll_\phi 1 \tag{3}$$

for the *period* $D_f(\phi)$ in the $\lambda \to \infty$ limit. Watson's formula implies that if ϕ is a fixed Maass eigencuspform on $Y_0(1)$, then $D_f(\phi)$ is closely related to a central L-*value*:

$$|D_f(\phi)|^2 = \lambda^{-1+o(1)} L(f \times f \times \phi, 1/2). \tag{4}$$

For quite general (finite parts of) L-functions $L(\pi, s)$, which we always normalize to satisfy a functional equation under $s \mapsto 1 - s$, there is a commonly accepted notion of a trivial bound for the central value $L(\pi, 1/2)$. It is called the *convexity bound*, and takes the form $L(\pi, 1/2) \ll C(\pi)^{1/4+o(1)}$ where $C(\pi) \in \mathbb{R}_{\geq 1}$ is the analytic conductor attached to π by Iwaniec–Sarnak [24]. The *subconvexity problem* is to improve this to $L(\pi, 1/2) \ll C(\pi)^{1/4-\delta}$ for some positive constant δ, while the Grand Lindelöf Hypothesis — itself a consequence of the Grand Riemann Hypothesis — predicts the sharper bound $L(\pi, 1/2) \ll C(\pi)^{o(1)}$. The subconvexity problem remains open in general for the triple product L-functions considered in this paper. We refer to [24, 38, 39] for further background.

For the L-value appearing in (4), the convexity bound reads

$$L(f \times f \times \phi, 1/2) \ll_\phi \lambda^{1+o(1)}. \tag{5}$$

Thus under the correspondence between periods and L-values afforded by Watson's formula (4), the trivial bound (3) for the period essentially[4] coincides with the trivial bound (5) for the L-value; strong bounds for the period imply strong bounds for the L-value, and vice versa.

This matching between trivial bounds for periods and trivial bounds for L-values holds up in the weight and squarefree level aspects: for f a holomorphic newform of weight k and squarefree level q, a generalization[5] of Watson's formula due to Ichino [19] that was pinned down precisely in [33] asserts that for each fixed Maass eigencuspform or unitary Eisenstein series ϕ on $Y_0(1)$, one has

$$|D_f(\phi)|^2 = (qk)^{-1+o(1)} L(f \times f \times \phi, 1/2). \tag{6}$$

Here the convexity bound reads $L(f \times f \times \phi, 1/2) \ll (qk)^{1+o(1)}$. Thus in the eigenvalue, weight, and squarefree level aspects, the trivial bounds for periods and L-values essentially coincide; in other words, the equidistribution and subconvexity problems are essentially equivalent.

We find that this equivalence does *not* survive the passage to non-squarefree levels. A simple yet somewhat artificial way to see this is to consider a sequence of twists $f_p = f_1 \otimes \chi_p$ of a fixed form f_1 of level 1 by quadratic Dirichlet characters χ_p of varying prime conductor p. The form f_p has trivial central character and level p^2. For each ϕ as above, one has

$$L(f_p \times f_p \times \phi, s) = L(f_1 \times f_1 \times \phi, s)$$

[4] That is to say, it coincides up to a bounded multiple of an arbitrarily small power of λ.

[5] Watson's original formula would suffice when $q = 1$.

for all $s \in \mathbb{C}$. Thus it does not even make sense to speak of the "subconvexity problem" corresponding to the equidistribution problem for the measures μ_{f_p}, as only one L-value is involved. The artificial nature of this example suggests that one could conceivably still have such an equivalence by restricting to forms that are *twist-minimal* (have minimal conductor among their GL(1) twists), but this turns out not to be the case; we find that the equidistribution problem is (in general) substantially *easier* than the subconvexity problem (see Section 1.6).

1.3. **Local Rankin–Selberg integrals.** The ideas involved in clarifying the relationship between the equidistribution and subconvexity problems discussed in Section 1.2 are exemplified by the following special case. Let f be a holomorphic newform of fixed weight k and prime power level $q = p^n$, with a fixed prime p and varying exponent $n \to \infty$. Recall the full-level Eisenstein series E_s, defined for $\mathrm{Re}(s) > 1$ by the absolutely and uniformly convergent series

$$E_s(z) = \sum_{\gamma \in \Gamma_\infty \backslash \Gamma_0(1)} (\mathrm{Im}\gamma z)^s, \quad \Gamma_\infty = \{\pm \left[\begin{smallmatrix} 1 & n \\ & 1 \end{smallmatrix}\right] : n \in \mathbb{Z}\}$$

and in general by meromorphic continuation. It is known that $s \mapsto E_s$ has no poles in $\mathrm{Re}(s) \geq 1/2$ except a simple pole at $s = 1$ with constant residue. Those E_s with $\mathrm{Re}(s) = 1/2$ are called unitary Eisenstein series, and furnish the continuous spectrum of $L^2(Y_0(1))$. We fix $t \in \mathbb{R}$ with $t \neq 0$, and take $\phi = E_{1/2+it}$; although ϕ is not bounded, it is a natural function against which to test the measure μ_f.

The period $\mu_f(E_{1/2+it})$ is related to the L-value $L(f \times f, 1/2 + it)$, but not directly. The "usual" integral representation for $L(f \times f, 1/2 + it)$ involves an Eisenstein series for the group $\Gamma_0(q)$, so that the integral cleanly unfolds (initially for $\mathrm{Re}(s) > 1$, in general by analytic continuation):

$$\begin{aligned}\int_{\Gamma_0(q)\backslash\mathbb{H}} y^k |f|^2(z) \left(\sum_{\gamma \in \Gamma_\infty \backslash \Gamma_0(q)} (\mathrm{Im}\gamma z)^s \right) \frac{dx\,dy}{y^2} &= \int_{x=0}^{1} \int_{y=0}^{\infty} y^{k-1+s} |f|^2(z) \frac{dx\,dy}{y} \\ &= \frac{\Gamma(s+k-1)}{(4\pi)^{s+k-1}} \sum_{n \in \mathbb{N}} \frac{\lambda_f(n)^2}{n^s} \\ &\approx \frac{\Gamma(s+k-1)}{(4\pi)^{s+k-1}} \frac{L(f \times f, s)}{\zeta(2s)},\end{aligned}$$

where $f(z) = \sum_{n=0}^{\infty} \lambda_f(n) n^{(k-1)/2} e^{2\pi i n z}$ and $\approx$ denotes equality up to some very simple Euler factors at p that are bounded from above and below by absolute constants when $\mathrm{Re}(s) = 1/2$ (see Section 2.6).

On the other hand, the full-level Eisenstein series E_s is defined relative to $\Gamma_0(1)$. Since f is invariant only under the smaller group $\Gamma_0(q)$, the unfolding for $\mu_f(E_{1/2+it})$ is not so clean; instead of giving a simple multiple of the L-value , it gives its multiple by a more complicated proportionality factor $J_f(s)$ satisfying (2). The square of a precise form of this relation implies (with $\phi = E_s$ and $s = 1/2 + it$)

$$|D_f(\phi)|^2 = q^{o(1)} \, |J_f(s) J_f(1-s)| \, L(f \times f \times \phi, 1/2). \tag{7}$$

Here the implied constant in $o(1)$ is allowed to depend upon the weight k and the fixed form ϕ, and $L(f \times f \times \phi, 1/2) = L(f \times f, 1/2 + it) L(f \times f, 1/2 - it) = |L(f \times f, 1/2 + it)|^2$.

The content of Ichino's formula [19], when combined with a lemma [31, Lemma 3.4.2] of Michel–Venkatesh, is that the relation (7) continues to hold when ϕ is a Maass eigencuspform provided that $s = s_{\phi,p}$ is chosen so the pth Hecke eigenvalue of ϕ is $p^{s-1/2} + p^{1/2-s}$. With this normalization, the Ramanujan conjecture asserts $\mathrm{Re}(s) = 1/2$; it is known unconditionally that $|\mathrm{Re}(s) - 1/2| \leq 7/64 < 1/2$ (see [27]), so in particular $0 < \mathrm{Re}(s) < 1$. Thus in all cases, the relative difficulty of the equidistribution problem for μ_f and the subconvexity problem for twists of $f \times f$ (in the $n \to \infty$ limit) is governed by the analytic behavior of $J_f(s)$ in the strip $\mathrm{Re}(s) \in (1/2 - 7/64, 1/2 + 7/64) \subset (0,1)$.

The quantity $J_f(s)$ is best studied p-adically. Let $W : \mathrm{PGL}_2(\mathbb{Q}_p) \to \mathbb{C}$ be an L^2-normalized Whittaker newform for f at p; in classical terms, this function packages all p-power-indexed Fourier coefficients of f at all cusps of $\Gamma_0(q)$ (see Section 3.4). Then the relation (7) holds with the definition

$$J_f(s) := \int_{k \in \mathrm{GL}_2(\mathbb{Z}_p)} \int_{y \in \mathbb{Q}_p^\times} \left| W\left(\begin{bmatrix} y & \\ & 1 \end{bmatrix} k \right) \right|^2 |y|^s \frac{d^\times y}{|y|} \, dk. \tag{8}$$

We refer to Sections 2.1 and 2.3 for precise definitions and normalizations. When $q = p^1$ is squarefree, there are explicit formulas for W with which one may easily show that

$$J_f(s) = p^{s-1} \frac{\zeta_p(s)\zeta_p(s+1)}{\zeta_p(2s)\zeta_p(1)}, \quad \zeta_p(s) := (1 - p^{-s})^{-1},$$

which is consistent with a special case of the relation (6). When $q = p^n$ with $n \geq 2$, such as is the case when f is supercuspidal at p, the function W is more difficult to describe explicitly, and so it is not immediately clear whether a comparably simple formula exists for $J_f(s)$.

1.4. **Local convexity and subconvexity.** In Section 2.4 we prove what we call a *local convexity bound* for the local integral $J_f(s)$ as given by (8). The terminology is justified by the proof, which we now illustrate. We continue to assume that f is a newform of prime power level $q = p^n$, and let π be the representation of $\mathrm{GL}_2(\mathbb{Q}_p)$ generated by f. The local $\mathrm{GL}(2) \times \mathrm{GL}(2)$ functional equation (see Proposition 2.12, or [25]) asserts that the normalized local Rankin–Selberg integral

$$J_f^*(s) := \frac{\zeta_p(2s)}{L(\pi \times \pi, s)} J_f(s) \tag{9}$$

satisfies

$$J_f^*(s) = C^{s-1/2} J_f^*(1-s), \tag{10}$$

where $C = C(f \times f)$ is the conductor of the Rankin–Selberg self-convolution of f; the latter is a power of p that satisfies $1 \leq C \leq p^{n+1}$ (see Proposition 2.5).

Our assumption that W is L^2-normalized implies the trivial bound $J_f^*(s) \ll 1$ for $\mathrm{Re}(s) = 1$, which we may transfer to the bound $J_f^*(s) \ll C^{-1/2}$ for $\mathrm{Re}(s) = 0$ via the functional equation (10). Interpolating these two bounds by the Phragmen–Lindelöf principle, and using that $J_f^*(s) \asymp J_f(s)$ uniformly for $\mathrm{Re}(s) \geq \delta > 0$, we deduce $J_f(s) \ll C^{-1/2+\mathrm{Re}(s)/2}$ uniformly for $\mathrm{Re}(s)$ in any compact subset of $(0,1)$. If $\mathrm{Re}(s) = 1/2$, which under the Ramanujan conjecture we may always assume to be the case in applications, then the local convexity bound just deduced reads

$$C^{1/2} J_f(s) \ll C^{1/4}. \tag{11}$$

The proof we have just sketched of (11) is analogous to that of the (global) convexity bound for $L(f \times f \times \phi, 1/2)$, which augments a trivial bound in the region of absolute convergence with the functional equation and the Phragmen–Lindelöf principle (see [23, Sec 5.2]). We refer to a bound that improves upon (11) by a positive of power of q as a *local subconvex bound*, and to the problem of producing such a bound as a local subconvexity problem.

1.5. **QUE versus local and global subconvexity.** The upshot of the above considerations is the following. Preserve the notation and assumptions of Sections 1.3 and 1.4. Assume also, for simplicity, that $\mathrm{Re}(s) = 1/2$. We may rewrite the formula (7) in the suggestive form

$$|D_f(\phi)|^2 = q^{o(1)} \left| \frac{C^{1/2} J_f(s)}{C^{1/4}} \right|^2 \frac{L(f \times f \times \phi, 1/2)}{C^{1/2}}. \tag{12}$$

Here the local and global convexity bounds read

$$\frac{C^{1/2} J_f(s)}{C^{1/4}} \ll 1 \quad \text{resp.} \quad \frac{L(f \times f \times \phi, 1/2)}{C^{1/2}} \ll C^{o(1)}, \tag{13}$$

where the implied constants are allowed to depend upon k, s and ϕ. Now, note that the intersection of the convexity bounds (13) is essentially[6] equivalent, via (12), to the trivial bound $D_f(\phi) \ll 1$ for the QUE problem. For emphasis, we summarize as follows:

Observation 1.4. *Fix a prime p, an even integer k, a complex number s, and either a Maass eigencuspform ϕ with pth normalized Hecke eigenvalue $p^{s-1/2}+p^{1/2-s}$ or a unitary Eisenstein series $\phi = E_s$ on $Y_0(1)$. Suppose, for simplicity, that $\mathrm{Re}(s) = 1/2$. Then the following are equivalent (with all implied constants allowed to depend upon p, k, and ϕ):*

(1) *(Equidistribution in the depth aspect with a power savings) There exists $\delta > 0$ so that $D_f(\phi) \ll q^{-\delta}$ for all holomorphic newforms f of weight k and prime power level $q = p^n$.*
(2) *There exists $\delta > 0$ so that for each holomorphic newform f of weight k and prime power level $q = p^n$, at least one of the following bounds hold:*
   (a) *(Global subconvexity without excessive conductor-dropping)*[7]
$$\frac{L(f \times f \times \phi, 1/2)}{C^{1/2}} \ll q^{-\delta},$$
   (b) *(Local subconvexity)*
$$\frac{C^{1/2} J_f(s)}{C^{1/4}} \ll q^{-\delta}. \tag{14}$$

Remark 1.5. We have stated the above equivalence as an observation (rather than as, say, a theorem) because one of the main results of this paper is that "local subconvexity" holds in a strong form (see Section 1.6).

[6] That is to say, it is equivalent up to $q^{o(1)}$.

[7] This estimate is implied by a global subconvex bound, which saves a small negative power of C rather than of q, together with a condition of the form $\log(C) \geq \alpha \log(q)$ for some fixed $\alpha > 0$. Note, for instance, that $C \geq q$ if f is twist-minimal, in which case we may drop the phrase "without excessive conductor dropping".

1.6. **Local Lindelöf hypothesis.** One might argue that the more interesting objects in the identity (12) are the global period $D_f(\phi)$ and the global L-value $L(f \times f \times \phi, 1/2)$, rather than the local period $J_f(s)$. One would like to compare precisely the difficulty of the QUE problem and the global subconvexity problem. In order to do so via (12), one must understand the true order of magnitude of $J_f(s)$. Suppose once again, for simplicity, that $\mathrm{Re}(s) = 1/2$. A global heuristic[8] suggested that one should have $J_f(s) \approx q^{-1/2+o(1)}$ in a mean-square sense. This expectation would be consistent with the individual bound

$$C^{1/2} J_f(s) \ll (C/q)^{1/2} q^{o(1)}, \tag{15}$$

which we term the *local Lindelöf hypothesis*.

In the special case $q = p^n$ relevant for Observation 1.4, we remark that $C/q \le p$ with equality if and only if n is odd (see Proposition 2.5), so that one should regard the RHS of (15) as being essentially bounded as far as the depth aspect is concerned. We may rewrite the bound (15) in the form $C^{1/2} J_f(s) \ll C^{1/4} (C/q^2)^{1/4} q^{o(1)}$; since $C/q^2 \ll_p q^{-1}$, we see that (15) implies (14) in a strong sense. This makes clear the analogy with the (global) Lindelöf hypothesis, as described in Section 1.2.

One of the main technical results of this paper is a proof of the bound (15) for all newforms on PGL(2). The proof goes by an explicit case-by-case calculation of $J_f(s)$, and yields the more precise bound

$$C^{1/2} J_f(s) \le 10^{3\omega(q)} \tau(q/\sqrt{C}) (C/q)^{1/2}, \tag{16}$$

where $\tau(n)$ (resp. $\omega(n)$) denotes the number of positive divisors (resp. prime divisors) of n. We remark that $q/\sqrt{C}$ is always integral, and equals 1 if and only if q is squarefree. As a byproduct of our explicit calculations, we obtain a precise generalization of Watson's formula to certain triple product integrals involving newforms of non-squarefree level (see Theorem 3.1).

By the discussion of Section 1.4, it follows that the global convexity bound is remarkably *stronger* than the trivial bound for the QUE problem, or in other words, that the subconvexity problem for $L(f \times f \times \phi, 1/2)$ in the depth aspect (f of level p^n, p fixed, $n \to \infty$) is much harder than the corresponding equidistribution problem, in contrast to the essential equivalence of their difficulty in the eigenvalue, weight and squarefree level aspects.

The above situation is somewhat reminiscent of how the problem of establishing the equidistribution of Heegner points of discriminant D on $Y_0(1)$ ($D \to -\infty$) is essentially equivalent to a subconvexity problem when D traverses a sequence of fundamental discriminants (c.f. [7]), but reduces to any nontrivial bound for the pth Hecke eigenvalue of Maass forms on $Y_0(1)$ when $D = D_0 p^{2n}$ for some fixed fundamental discriminant D_0 and some increasing prime power p^n.

Remark 1.6. Let f_1 and f_2 be a pair of L^2-normalized holomorphic newforms, of the same fixed weight, on $\Gamma_0(p^n)$ with $n \ge 2$. One knows that

$$C := C(f_1 \times f_2) \le p^{2n}. \tag{17}$$

There is a sense in which C measures the difference between the representations of $\mathrm{PGL}_2(\mathbb{Q}_p)$ generated by f_1 and f_2, and that for typical f_1 and f_2, the upper bound in (17) is attained. This perspective is consistent with the much stronger bound $C \le p^{n+1}$ that holds on the thin diagonal

[8] The heuristic involved a computation by the first-named author, without appeal to triple product formulas, of an average of $|D_f(\phi)|^2$ over f of level p^{2m} (see [32]).

subset $f_1 = f_2$, and also with the explicit formulas for C given in [4]. We expect that the problems of improving upon the Cauchy–Schwarz bound $\int \overline{f_1} f_2 \phi \ll_\phi 1$ (integral is over $\Gamma_0(q)\backslash\mathbb{H}$ with respect to the hyperbolic probability measure) and the convexity bound $L(f_1 \times f_2 \times \phi, 1/2) \ll C^{1/2}$ should have comparable difficulty if and only if the upper bound in (17) is essentially attained. If reasonable, this expectation suggests a correlation between the smallness of C and the discrepancy of difficulty between the corresponding equidistribution and subconvexity problems.

1.7. **Local Riemann hypothesis.** Maintain the assumption that f is a newform of level p^n that generates a representation π of $\mathrm{PGL}_2(\mathbb{Q}_p)$. Numerical experiments strongly suggest that the normalized local Rankin–Selberg integral $J_f^*(s)$ (see (9)), which is an essentially palindromic polynomial[9] in $p^{\pm s}$, has all its zeros on the line $\mathrm{Re}(s) = 1/2$.[10]

We suspect that this "local Riemann Hypothesis" should follow from known properties of the classical polynomials implicit in our formulas for $J_f^*(s)$ (see Theorem 2.7), but it would be interesting to have a more conceptual explanation, or a proof that does not rely upon our brute-force computations. It seems reasonable to expect that such an alternative explanation would lead to a different proof of the local Lindelöf bound (15).

Example 1.7. Suppose that π has "Type 1" according to the classification recalled in Section 2.2. Let p^{2g} ($g \geq 1$) be the conductor of π. Suppose that p^{2g} is also the conductor of $\pi \times \pi$; equivalently, π is twist-minimal. Then (the calculations leading to) Theorem 2.7 imply that $J_f^*(s)$ differs by a unit in $\mathbb{C}[p^{\pm s}]$ from $F(p^{-s})$, where F is the integral polynomial

$$F(t) = 1 + \sum_{j=1}^{g-1} (p^j - p^{j-1}) t^{2j} + p^g t^{2g} \in \mathbb{Z}[t].$$

Example 1.8. Suppose that π has "Type 2" (see Section 2.2) and conductor p^{2g+1} ($g \geq 1$). Then as above, $J_f^*(s)$ differs by a unit in $\mathbb{C}[p^{\pm s}]$ from $F(p^{-s})$ with

$$F(t) = \sum_{j=0}^{g} p^j t^{2j} - \sum_{j=0}^{g-1} p^j t^{2j+1} \in \mathbb{Z}[t].$$

In either example, F satisfies the formal properties of the L-function of a smooth projective curve of genus g over $\mathbb{F}_p$; for example, the roots of F come in complex conjugate pairs, they have absolute value $p^{-1/2}$, and F satisfies the functional equation $F(1/pt) = p^{-g} t^{-2g} F(t)$. The geometric significance of this, if any, is unclear.

1.8. **A sketch of the proof.** The essential inputs to our method for proving (16) are the local functional equations for $\mathrm{GL}(2)$ and $\mathrm{GL}(2) \times \mathrm{GL}(2)$, and some knowledge of the behavior of representations of $\mathrm{GL}(2)$ under twisting by $\mathrm{GL}(1)$; specifically, for μ on $\mathrm{GL}(1)$ and π on $\mathrm{PGL}(2)$, we use that the formula $C(\pi\mu) = C(\pi)$ holds whenever $C(\mu)^2 < C(\pi)$. Here and below, $C(\cdot)$ is the conductor of a representation.

---

[9] The local functional equation for $\mathrm{GL}(2) \times \mathrm{GL}(2)$ implies that $J_f^*(s) = P_f(p^s)$ for some $P_f(t)$ in $\mathbb{C}[t, 1/t]$ satisfying $P_f(t) = p^{-N/2} t^N P_f(p/t)$ where the integer N is defined by the equation $C(f \times f) = p^N$.

[10] More precisely, it seems that $J_f^*(s)$ has its zeros on $\mathrm{Re}(s) = 1/2$ unless π is a ramified quadratic twist of a highly non-tempered spherical representations, specifically unless $\pi = \beta|.|^{s_0} \boxplus \beta|.|^{-s_0}$ with $s_0 \in \mathbb{R}$, $|s_0| > 1/4 + \delta_p$ (see Section 2.2 for notation); here δ_p is a positive real that satisfies $\delta_p \to 0$ as $p \to \infty$. By the classical bound $|s_0| \leq 1/4$, the latter possibility does not occur.

Write $F = \mathbb{Q}_p$, $|.|$ = the standard p-adic absolute value, $U = \{x \in F : |x| = 1\} = \mathbb{Z}_p^\times$, $G = \mathrm{GL}_2(F)$, $n(x) = \left[\begin{smallmatrix} 1 & x \\ & 1 \end{smallmatrix}\right]$ for $x \in F$, $a(y) = \left[\begin{smallmatrix} y & \\ & 1 \end{smallmatrix}\right]$ for $y \in F^\times$, $N = \{n(x) : x \in F\}$, $K = \mathrm{GL}_2(\mathbb{Z}_p)$, and $Z = \{\left[\begin{smallmatrix} z & \\ & z \end{smallmatrix}\right] : z \in F^\times\}$. We sketch a proof of the bound (16) in the simplest case that f is a newform of prime power level $q = p^n$, and π, the local representation at p attached to f, is a supercuspidal representation of G with trivial central character, realized in its Whittaker model with L^2-normalized newform W. Let $f_3 : ZN\backslash G \to \mathbb{C}$ be the function given by $f_3(n(x)a(y)k) = |y|^s$ in the Iwasawa decomposition. We wish to compute the local integral $J_f(s) = \int_{ZN\backslash G} |W|^2 f_3$. It is convenient to do so in the Bruhat decomposition, where our measures are normalized so that

$$\frac{\zeta_p(1)}{\zeta_p(2)} \int_{ZN\backslash G} |W|^2 f_3 = \int_{x\in F} \max(1, |x|)^{-2s} \int_{y\in F^\times} |W|^2(a(y)wn(x))|y|^{s-1}\, d^\times y\, dx. \tag{18}$$

Because the LHS of (18) satisfies the $\mathrm{GL}(2)\times\mathrm{GL}(2)$ functional equation, it suffices to determine the coefficients of the *positive* powers of p^s occuring on the RHS. The left N-equivariance of W implies that no such positive powers arise from the integral over $|x| \geq C(\pi)^{1/2}$, an implication which in classical terms amounts to the calculation of the widths of the cusps of $\Gamma_0(q)$ (see Section 3.4). In the remaining range $|x| < C(\pi)^{1/2}$, we show that $W(a(y)wn(x))$ is supported on the coset $|y| = C(\pi)$ of the unit group U in $F^\times$. Thus by the invariance of the inner product on π, the integral over $F^\times$ in (18) is simply $C(\pi)^{s-1}$. Integrating over x gives

$$\frac{\zeta_p(1)}{\zeta_p(2)} \int_{ZN\backslash G} |W|^2 f_3 = C(\pi)^{s-1} \left(\int_{\substack{x\in F \\ |x|<C(\pi)^{1/2}}} \max(1, |x|)^{-2s} dx \right) + \sum_{m\in\mathbb{Z}_{\geq 0}} \frac{c_m}{p^{ms}} \tag{19}$$

for some coefficients c_m. After determining c_m via the $\mathrm{GL}(2)\times\mathrm{GL}(2)$ functional equation, we end up with a formula for $\int |W|^2 f_3$ in terms of $C(\pi)$ and $C(\pi \times \pi)$ that shows, by inspection, that $\int |W|^2 f_3$ satisfies the desired bounds.[11]

A key ingredient in the above argument was the support condition on $W(a(y)wn(x))$ for $|x| < C(\pi)^{1/2}$. We derive it via a Fourier decomposition over the the character group of U and invariance properties of W. Indeed, the $\mathrm{GL}(2)$ functional equation implies

$$W(a(y)wn(x)) = \sum_{\substack{\mu\in\hat{U} \\ C(\pi\mu)=|y|}} \mu(y)\varepsilon(\pi\mu)G(x,\mu), \tag{20}$$

where $G(x,\mu) = \int_{y\in F^\times} \psi(xy)\mu(y)W(a(y)) = \int_{y\in U} \psi(xy)\mu(y)$ and $\varepsilon(\pi\mu) = \varepsilon(\pi\mu, 1/2)$ is the local ε-factor (see Section 2.5). The characters μ contributing nontrivially to (20) all satisfy $G(x,\mu) \neq 0$, which implies $C(\mu) \leq x$; in that case our assumption $|x|^2 < C(\pi)$ and our knowledge of the twisting behavior of π implies $C(\pi\mu) = C(\pi)$. It follows that $W(a(y)wn(x)) = 0$ unless $|y| = C(\pi)$.

Remark 1.9. It seems worthwhile to note that one may also compute the RHS of (18) in "bulldozer" fashion, as follows. Suppose for simplicity that π is supercuspidal. We may view the integral over $y \in F^\times$ as the inner product of the functions $W(a(y)wn(x))$ and $W(a(y)wn(x))|y|^s$, whose Mellin transforms are (by definition) local zeta integrals; applying the Plancherel theorem on $F^\times$ and the

[11]It would be possible to establish this by a slightly softer argument, but we believe that having precise formulas is of independent interest.

GL(2) functional equation, we arrive at the formula

$$\frac{\zeta_p(1)}{\zeta_p(2)} \int_{ZN\backslash G} |W|^2 f_3 = \sum_{\mu \in \hat{U}} C(\pi\mu)^{s-1} \int_{x \in F} \frac{|G(x,\mu)|^2}{\max(1,|x|)^{2s}} dx. \tag{21}$$

This also follows from (20) by the Plancherel theorem on U. Substituting into (21) the fact that $C(\pi\mu) \leq \max(C(\pi), C(\mu)^2)$ with equality if $C(\mu)^2 \neq C(\pi)$, evaluating $|G(x,\mu)|$, and summing some geometric series, we find that

$$\frac{\zeta_p(1)}{\zeta_p(2)} \int_{ZN\backslash G} |W|^2 f_3 = p^{n(s-1)} \left\{ 1 + \sum_{1 \leq a < n/2} \frac{\zeta_p(1)^{-1}}{p^{(2s-1)a}} \right\} + p^{-r} + \zeta_p(1) \sum_{C(\mu)^2 = C(\pi)} \frac{C(\pi\mu)^{s-1}}{C(\pi)^s}, \tag{22}$$

where $C(\pi) = p^n$ and $r = \lfloor n/2 \rfloor + 1$. This identity agrees with (19), and shows that the only barrier to obtaining immediately an explicit result is the potentially subtle behavior of the conductors of twists of π by characters of conductor $C(\pi)^{1/2}$ (see also Remark 3.16). It suggests another approach to our local calculations (write $\pi = \pi_0\mu_0$ with π_0 twist-minimal and compute away), but one that would be more difficult to implement when π is a ramified twist of a principal series or Steinberg representation.

The approach sketched in this remark has the virtue of applying to arbitrary vectors $W \in \pi$, leading to formulas generalizing those that we have given in this paper in the special case that W is the newvector.

1.9. **Fourier expansions at arbitrary cusps.** Let f be a newform on $\Gamma_0(q)$, $q \in \mathbb{N}$. In order to apply a variant of the Holowinsky–Soundararajan method in Section 3, we require some knowledge of the sizes of the normalized Fourier coefficients $\lambda(\ell; \mathfrak{a})$ of f at an *arbitrary* cusp $\mathfrak{a}$ of $\Gamma_0(q)$. It is perhaps not widely known that such Fourier coefficients are *not* multiplicative in general; this lack of multiplicativity introduces an additional complication in our arguments. More importantly, we need some knowledge of the sizes of the coefficients $\lambda(\ell; \mathfrak{a})$ when $\ell \mid q^\infty$. For example, the "Hecke bound" $\lambda(\ell; \mathfrak{a}) \ll \ell^{1/2}$ would not suffice for our purposes.

Let $\lambda(\ell) = \lambda(\ell; \infty)$ denote the ℓth normalized Fourier coefficient of f at the cusp ∞. A complete description of the coefficients $\lambda(\ell)$ is given by Atkin and Lehner [1]; for our purposes, it is most significant to note that $\lambda(p^\alpha) = 0$ for each $\alpha \geq 1$ if p is a prime for which $p^2 | q$.

If $\mathfrak{a}$ is the image of ∞ under an Atkin–Lehner operator (an element of the normalizer of $\Gamma_0(q)$ in $\mathrm{PGL}_2^+(\mathbb{Q})$), then the coefficients $\lambda(\ell)$ and $\lambda(\ell; \mathfrak{a})$ are related in a simple way; this is always the case when q is squarefree, in which case the Atkin–Lehner operators act transitively on the set of cusps. Similarly, there is a simple relationship between the Fourier coefficients $\lambda(\ell, \mathfrak{a})$, $\lambda(\ell, \mathfrak{a}')$ of f at each pair of cusps $\mathfrak{a}$, $\mathfrak{a}'$ related by an Atkin–Lehner operator (see [13]). However, such considerations do not suffice to describe $\lambda(\ell; \mathfrak{a})$ explicitly when $\mathfrak{a}$ is not in the Atkin–Lehner orbit of ∞.

Our calculations in Section 2 lead to a precise description of $\lambda(\ell; \mathfrak{a})$ for arbitrary cusps $\mathfrak{a}$, at least in a mildly averaged sense. This may be of independent interest. To give some flavor for the results obtained, suppose that $q = p^n$ with $n \geq 2$. The nature of the coefficients $\lambda(\ell; \mathfrak{a})$ depends heavily upon the *denominator* p^k of the cusp $\mathfrak{a}$, as defined in Section 3.4; briefly, k is the unique integer in $[0, n]$ with the property that $\mathfrak{a}$ is in the $\Gamma_0(p^n)$-orbit of some fraction $a/p^k \in \mathbb{R} \subset \mathbb{P}^1(\mathbb{R})$ with $(a, p) = 1$. The Atkin–Lehner/Fricke involution swaps the cusps of denominator p^k and p^{n-k}.

Say that f is *p-trivial* at a cusp $\mathfrak{a}$ if $\lambda(p^\alpha; \mathfrak{a}) = 0$ for all $\alpha \geq 1$. For example, the result of Atkin–Lehner mentioned above asserts that f is p-trivial at ∞. We observe the "purity" phenomenon:

f is p-trivial at $\mathfrak{a}$ unless n is even and the denominator p^k of $\mathfrak{a}$ satisfies $k = n/2$ (see Proposition 3.12). In the latter case, let us call $\mathfrak{a}$ a *middle cusp*.

In Section 3.4, we compute for each $\alpha \geq 0$ the mean square of $\lambda_f(p^\alpha; \mathfrak{a})$ over all middle cusps $\mathfrak{a}$; an accurate evaluation of this mean square, together with the aforementioned "purity", turns out to be equivalent to our local Lindelöf hypothesis described above (see Remark 3.16). We observe that the "Deligne bound" $|\lambda(\ell; \mathfrak{a})| \leq \tau(\ell)$ can fail in the strong form $\lambda(p^\alpha; \mathfrak{a}) \gg p^{\alpha/4}$ for some $\alpha > 0$ when f is not twist-minimal (see Remark 3.14). In general, $\lambda(\ell; \mathfrak{a})$ may be evaluated exactly in terms of GL(2) Gauss sums (e.g., combine (20) and (49) when π is supercuspidal). We suppress further discussion of this point for sake of brevity.

1.10. **Further remarks.** Our calculations in Section 2, being local, apply in greater generality than we have used them. For example, they imply that the pushforward to $Y_0(1)$ of the L^2-mass of a Hecke-Maass newform on $\Gamma_0(p^n)$ of bounded Laplace eigenvalue equidistributes as $p^n \to \infty$ with $n \geq 2$. They extend also to non-split quaternion algebras, where Ichino's formula applies but the Holowinsky–Soundararajan method does not, due to the absence of Fourier expansions. For example, one could establish that Maass or holomorphic newforms of increasing level on compact arithmetic surfaces satisfy an analogue of Theorem 1.2 provided that their level is sufficiently powerful (c.f. the remarks at the end of Section 1.2); in that context, no unconditional result for forms of increasing squarefree level is known. For automorphic forms of increasing squared-prime level p^2 on definite quaternion algebras, an analogue of Theorem 1.2 had been derived earlier by the first-named author (see [32]) via a different method (i.e., without triple product formulas), but the bounds obtained there are quantitatively weaker than those that would follow from the present work.

After completing an earlier draft of this paper, we learned of some interesting parallels in the literature of some of the analogies presented hitherto. Lemma 2.1 of Soundararajan and Young [45] gives something resembling a "local Riemann hypothesis" for a certain Dirichlet series, studied earlier by Bykovskii and Zagier, attached to (not necessarily fundamental) quadratic discriminants.[12] Section 9 of a paper of Einsiedler, Lindenstrauss, Michel and Venkatesh [8] establishes what they refer to as "local subconvexity" for certain local toric periods, the proof of one aspect of which resembles that of what we describe here as "local convexity". It would be interesting to understand whether our work can be understood together with these parallels in a unified manner.

1.11. **Acknowledgements.** We thank Ralf Schmidt for helping us with the representation theory of $\mathrm{PGL}_2(\mathbb{Q}_p)$, and Philippe Michel for pointing us to the crucial lemma [31, Lemma 3.4.2]. We thank Nahid Walji and Matthew Young for helpful comments on an earlier draft of this paper. Finally, we would like to thank the referee for many helpful comments which have improved the correctness, clarity, and exposition of this paper.

2. Local calculations

2.1. Notation and preliminaries.

[12] We thank M. Young for bringing this similarity to our attention.

2.1.1. *Groups, measures.* Let p be a prime number, and $F = \mathbb{Q}_p$.[13] Let $\mathfrak{o}$ be its ring of integers, and $\mathfrak{p}$ its maximal ideal. Fix a generator ϖ of $\mathfrak{p}$. Let $|.|$ or $|.|_p$ denote the absolute value on F normalized so that $|\varpi| = p^{-1}$.

Let $G = \mathrm{GL}_2(F)$ and $K = \mathrm{GL}_2(\mathfrak{o})$. For each integral ideal $\mathfrak{a}$ of $\mathfrak{o}$, let $K_0(\mathfrak{a})$ and $K_1(\mathfrak{a})$ denote the usual congruence subgroups of K:

$$K_0(\mathfrak{a}) = K \cap \begin{bmatrix} \mathfrak{o} & \mathfrak{o} \\ \mathfrak{a} & \mathfrak{o} \end{bmatrix}, \quad K_1(\mathfrak{a}) = K \cap \begin{bmatrix} 1+\mathfrak{a} & \mathfrak{o} \\ \mathfrak{a} & \mathfrak{o} \end{bmatrix}.$$

In particular, $K_0(\mathfrak{o}) = K_1(\mathfrak{o}) = K$. Write

$$w = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad a(y) = \begin{bmatrix} y & \\ & 1 \end{bmatrix}, \quad n(x) = \begin{bmatrix} 1 & x \\ & 1 \end{bmatrix}, \quad z(t) = \begin{bmatrix} t & \\ & t \end{bmatrix}$$

for $x \in F, y \in F^\times, t \in F^\times$. Define subgroups $N = \{n(x) : x \in F\}$, $A = \{a(y) : y \in F^\times\}$, $Z = \{z(t) : t \in F^\times\}$, and $B = ZNA = G \cap \left[\begin{smallmatrix} * & * \\ & * \end{smallmatrix}\right]$ of G.

We normalize Haar measures as in [31, Section 3.1]: The measure dx on the additive group F assigns volume 1 to $\mathfrak{o}$, and transports to a measure on N. The measure $d^\times y$ on the multiplicative group $F^\times$ assigns volume 1 to $\mathfrak{o}^\times$, and transports to measures on A and Z. We obtain a left Haar measure $d_L b$ on B via

$$d_L(z(u)n(x)a(y)) = |y|^{-1}\, d^\times u\, dx\, d^\times y.$$

Let dk be the probability Haar measure on K. The Iwasawa decomposition $G = BK$ gives a left Haar measure $dg = d_L b\, dk$ on G; with respect to the Bruhat decomposition $G = B \sqcup BwN$, this measure takes the form

$$dg = \frac{\zeta_p(2)}{\zeta_p(1)} |y|^{-1}\, d^\times u\, d^\times y\, dx'\, dx \qquad \text{for } g = n(x')a(y)z(u)wn(x), \tag{23}$$

where $\zeta_p(s) = (1-p^{-s})^{-1}$ (see [31, (3.1.6)]).

2.1.2. *Representations, models.* Fix an additive character $\psi : F \to \mathbb{C}^1$ with conductor $\mathfrak{o}$. For each generic representation σ of G, let $\mathcal{W}(\sigma, \psi)$ denote the Whittaker model of σ with respect to ψ (see [26]). For two characters χ_1, χ_2 on $F^\times$, let $\chi_1 \boxplus \chi_2$ denote the principal series representation on G that is unitarily induced from the corresponding representation of B; this consists of smooth functions f on G satisfying

$$f\left(\begin{bmatrix} a & b \\ 0 & d \end{bmatrix} g\right) = |a/d|^{\frac{1}{2}} \chi_1(a)\chi_2(d) f(g).$$

2.1.3. *Conductors, L-functions, ε-factors.* For each character[14] σ of $F^\times$, there exists a minimal integer $a(\sigma)$ such that $\sigma(1+t) = 1$ for all $t \in \mathfrak{p}^{a(\sigma)}$. For each irreducible admissible representation

[13] Most of this section reads correctly in the more general case that p is an arbitrary prime power and F is a non-archimedean local field of characteristic zero whose residue field has cardinality p. We work in the restricted generality that we need for our global applications only because we have not checked that the calculations in the Type 3 case of the proof of Theorem 2.7 carry through in this more general context when $p = 2$.

[14] We adopt the convention that a *character* of a topological group is a continuous (but not necessarily unitary) homomorphism into $\mathbb{C}^\times$.

σ of G, there exists a minimal integer $a(\sigma)$ such that σ has a $K_1(\mathfrak{p}^{a(\sigma)})$-fixed vector. In either case, the integer $p^{a(\sigma)}$ is called the local analytic conductor[15] of σ; we denote it by $C(\sigma)$.

For a representation σ of G and a character χ of $F^\times$, write $\sigma\chi$ for the representation $\sigma \otimes (\chi \circ \det)$ of G.

Let $L(\sigma, s)$ (resp. $\varepsilon(\sigma, \psi, s)$) denote the L-function (resp. ε-factor) of an irreducible admissible representation σ of G or a character σ of $F^\times$. These local factors are defined in [26]. For σ an irreducible admissible representation of G, let $L(\mathrm{ad}\sigma, s)$ denote the adjoint L-function of σ, or equivalently, the standard L-function of the adjoint lift of σ to an admissible representation of $\mathrm{PGL}_3(F)$.

If σ_1, σ_2 are two irreducible admissible representations of G, the local Rankin–Selberg factors $L(\sigma_1 \times \sigma_2, s)$ and $\varepsilon(\sigma_1 \times \sigma_2, \psi, s)$ are defined in [25]. The local analytic conductor $C(\sigma_1 \times \sigma_2)$ is a nonnegative integral power of p, and can be defined by the formula $\varepsilon(\sigma_1 \times \sigma_2, \psi, s) = C(\sigma_1 \times \sigma_2)^{1/2-s} \varepsilon(\sigma_1 \times \sigma_2, \psi, 1/2)$; we also let $a(\sigma_1 \times \sigma_2)$ denote the nonnegative integer for which $C(\sigma_1 \times \sigma_2) = p^{a(\sigma_1 \times \sigma_2)}$.

2.1.4. *Temperedness.* Let π be a generic irreducible admissible unitarizable representation of G with trivial central character. The quantity

$$\lambda(\pi) = \begin{cases} 0 & \text{if } \pi \text{ is tempered,} \\ |s_0| & \text{if } \pi \cong \beta\,|\cdot|^{s_0} \boxplus \beta^{-1}\,|\cdot|^{-s_0}, \quad s_0 \in \mathbb{R}, \quad \beta \text{ unitary,} \end{cases} \tag{24}$$

measures the temperedness of π. When π arises as the local factor of a cuspidal automorphic representation of $\mathrm{GL}_2(\mathbb{A})$, it is known that $\lambda(\pi) \le 7/64$ (see [27]). For our purposes, it suffices to assume that $\lambda(\pi) < 1/4$. We record this assumption as follows:

Condition 2.1. π is a generic irreducible admissible unitarizable representation of G with trivial central character and $\lambda(\pi) < 1/4$.

2.1.5. *Classification of representations.* Let π satisfy Condition 2.1. Write $n = a(\pi)$, and suppose that $n \ge 2$. We recall a certain classification of such π. The classification is standard, although our labeling is not (and we are not aware of a standard labeling).

- **Type 1.** These are the supercuspidal representations satisfying $\pi \cong \pi\eta$, where η is the unique non-trivial unramified quadratic character of $F^\times$. Equivalently, π is the dihedral supercuspidal representation $\rho(E/F, \xi)$ associated to the unramified quadratic extension E of F and a character ξ of $E^\times$ that is not $\mathrm{Gal}(E/F)$-invariant.
- **Type 2.** These are the supercuspidal representations satisfying $\pi \not\cong \pi\eta$, with η as above.
- **Type 3.** In this case π is a ramified quadratic twist of a spherical representation:
$$\pi \cong \beta\,|\cdot|^{s_0} \boxplus \beta\,|\cdot|^{-s_0}, \quad s_0 \in i\mathbb{R} \cup (-1/4, 1/4), \quad \beta \text{ ramified}, \ \ \beta^2 = 1.$$
We denote $\beta_{s_0} = (p^{s_0} + p^{-s_0})^2$.
- **Type 4.** In this case π is a ramified principal series that is not of Type 3:
$$\pi \cong \beta \boxplus \beta^{-1}, \qquad \beta \text{ ramified, unitary character of } F^\times, \qquad \beta^2 \text{ ramified.}$$

[15]In the rest of this paper, we will often drop the words "local analytic" for brevity and call this simply the "conductor".

- **Type 5.** In this case π is a ramified quadratic twist of the Steinberg representation:
$$\pi \cong \beta \mathrm{St}_{\mathrm{GL}(2)}, \qquad \beta \text{ ramified, } \beta^2 = 1.$$

Remark 2.2. If p is odd, then each supercuspidal representation is dihedral, i.e., constructed via the Weil representation from a quadratic extension E of F and a non-$\mathrm{Gal}(E/F)$-invariant character ξ of $E^\times$. Such representations are of Type 1 if E/F is unramified and of Type 2 if E/F is ramified. If p is even, there exist non-dihedral supercuspidals; these are also of Type 2.

Remark 2.3. For representations of Type 3 or 5, the ramified quadratic character β satisfies $a(\beta) = 1$ if p is odd and $a(\beta) \in \{2, 3\}$ if p is even.

Remark 2.4. If π is of Type 1, 3, 4, or 5, then n is even. If π is of Type 2, then n can be either odd or even.

2.1.6. *Properties of the adjoint conductor.* Let π be a generic irreducible admissible unitarizable representation of G with trivial central character. Write $n = a(\pi)$ and $N = a(\pi \times \pi)$. In Section 2.6, we will establish the following result concerning the integer N and its relation to n. We state it here because it will be useful in interpreting the results to follow.

Proposition 2.5. *The integer N is even and satisfies $N \leq n + 1$. Furthermore, the following conditions on π are equivalent:*

1. *$N = n + 1$.*
2. *n is odd.*
3. *Either*
   - (a) *π is the Steinberg representation or an unramified quadratic twist thereof (in which case $n = 1$), or*
   - (b) *π is a representation of Type 2 for which n is odd.*

2.1.7. *Definition of Ichino integral.* Let s be a complex parameter, and $\pi_3 = |.|^{s-1/2} \boxplus |.|^{1/2-s}$ the corresponding principal series representation of G. It is well-known that π_3 is irreducible and unitarizable if and only if $\mathrm{Re}(s) = 1/2$ or $s \in (0, 1)$; suppose that this is the case. Fix a non-zero K-invariant vector $x_3 \in \pi_3$, which is then unique up to a scalar. We recall, for later use, the following formula for the normalized Hecke eigenvalues of x_3:

$$\lambda_{s,m} = \sum_{\substack{i,j \in \mathbb{Z}_{\geq 0} \\ i+j=m}} \alpha^i \beta^j = \begin{cases} \frac{\alpha^{m+1} - \beta^{m+1}}{\alpha - \beta} & m \geq 0 \\ 0 & m < 0 \end{cases} \quad \text{with } \alpha = p^{s-1/2}, \beta = p^{1/2-s}. \tag{25}$$

Let π be a representation of G satisfying Condition 2.1. Let $x \in \pi$ be a newvector, i.e., a nonzero vector on the unique line of $K_0(\mathfrak{p}^{a(\pi)})$-invariant vectors in π. Fix arbitrary G-invariant inner products $\langle , \rangle$ on π and π_3. It follows from [19, Lemma 2.1] that what we will call the *local Ichino integral*

$$I(s) = I(s; \pi) = \int_{Z \backslash G} \left(\frac{\langle gx, x \rangle}{\langle x, x \rangle} \right)^2 \frac{\langle gx_3, x_3 \rangle}{\langle x_3, x_3 \rangle} \, dg \tag{26}$$

converges absolutely provided that either $\mathrm{Re}(s) = 1/2$ or $s \in (2\lambda(\pi), 1 - 2\lambda(\pi))$; we will see later that it extends to a meromorphic function of $s \in \mathbb{C}$. Note that $I(s)$ depends only upon π, s, and our normalization of measures, and not upon the precise choice of x, x_3, or the inner products $\langle , \rangle$

on π, π_3. While not immediately obvious, it can be shown (using (31) below, for instance) that the right hand side of (26) is nonnegative.

It will be convenient to work with the normalized quantity

$$I^*(s) = I^*(s;\pi) = \left(\frac{L(\pi \times \pi \times \pi_3, 1/2)\zeta_p(2)^2}{L(\mathrm{ad}\pi_3, 1)L(\mathrm{ad}\pi, 1)^2}\right)^{-1} I(s).$$

We note that $L(\pi \times \pi \times \pi_3, 1/2) = L(\pi \times \pi, s)L(\pi \times \pi, 1-s)$ and $L(\mathrm{ad}\pi_3, 1)^{-1} = (1-p^{2s-2})(1-p^{-1})(1-p^{-2s})$.

2.2. **Statement of results.** Let π be a representation of G satisfying Condition 2.1, and let $s \in \mathbb{C}$. Our main local result is an explicit formula for the normalized local Ichino integral $I^* = I^*(s) = I^*(s;\pi)$; as a consequence, we deduce optimal bounds for the latter. The proofs will occupy the remaining subsections of Section 2. We will use the notation

$$n = a(\pi), \quad N = a(\pi \times \pi), \quad n' = n - \frac{N}{2}.$$

Proposition 2.5 implies that n' is an integer satisfying $\frac{N}{2} - 1 = n' = \frac{n-1}{2}$ if n is odd and $\frac{N}{2} \le n' \le n$ if n is even.

When $n \in \{0, 1\}$, the value of I^* is already known (see [20, Theorem 1.2] and [33, Lemma 4.2]):

Theorem 2.6. *Suppose that $n = 0$ or $n = 1$. Then $I^* = p^{-n}$.*

We turn to the case $n \ge 2$. Our formulas will depend upon the classification of π recalled in Section 2.1.5 and the notation $\lambda_{s,m}$ introduced in (25).

Theorem 2.7. *Suppose that $n \ge 2$. Then $I^* = p^{-n} \cdot L(\mathrm{ad}\pi, 1)^2 \cdot Q_{\pi,p}(s)^2$ with*

$$Q_{\pi,p}(s) = \begin{cases} \lambda_{s,n'} - p^{-1}\lambda_{s,n'-2} & \text{for Type 1,} \\ \lambda_{s,n'} - p^{-1/2}\lambda_{s,n'-1} & \text{for Type 2,} \\ \lambda_{s,n'} - 2p^{-1/2}\lambda_{s,n'-1} + p^{-1}\lambda_{s,n'-2} & \text{for Type 4,} \\ \lambda_{s,n'} - p^{-1/2}(1+p^{-1})\lambda_{s,n'-1} + p^{-2}\lambda_{s,n'-2} & \text{for Type 5.} \end{cases}$$

In the remaining case that π is of Type 3, we have $N = 0$, $n = n' = 2a(\beta) \in \{2, 4, 6\}$ and

$$Q_{\pi,p}(s) = \begin{cases} \lambda_{s,2} - p^{-1/2}\beta_{s_0}\lambda_{s,1} + p^{-1}(2\beta_{s_0} - 2 - p^{-1}), & p \text{ odd,} \\ \lambda_{s,n} - p^{-1/2}\beta_{s_0}\lambda_{s,n-1} + 2p^{-1}(\beta_{s_0} - 1)\lambda_{s,n-2} - p^{-3/2}\beta_{s_0}\lambda_{s,n-3} + p^{-2}\lambda_{s,n-4}, & p \text{ even.} \end{cases}$$

Corollary 2.8 (Local Lindelöf hypothesis)**.** *Let $\theta = |\mathrm{Re}(s - 1/2)|$. Then $I^* < 10^5 p^{-n}\tau(p^{n'})^2 p^{2\theta n'}$.*

Proof. The case $n \in \{0, 1\}$ follows easily from Theorem 2.6, so suppose that $n \ge 2$. The value of $L(\mathrm{ad}\pi, 1)$ can be read off from the formula for $L(\pi \times \pi, s)^{-1}\zeta_p(2s) = (1 + p^{-s})^{-1}L(\mathrm{ad}\pi, s)^{-1}$ given in Table 1 below. We see that $L(\mathrm{ad}\pi, 1) \le 30 < 10^{\frac{3}{2}}$ in every case. The formulas for $Q_{\pi,p}(s)$ provided above imply the bound $|Q_{\pi,p}(s)| \le 10\tau(p^{n'})p^{\theta n'}$. The result now follows from Theorem 2.7, noting that $n \ge 2$ implies $n' \ge 1$. ∎

2.3. **An identity of local integrals.** In this section we apply a lemma of Michel–Venkatesh to establish the meromorphic continuation of the local Ichino integral $I(s) = I(s;\pi)$ defined in Section 2.1.7 and to reduce its study to that of a Rankin–Selberg integral involving the Whittaker newform of π.

Let π be a generic irreducible admissible unitarizable representation of G with trivial central character, realized in its ψ-Whittaker model: $\pi = \mathcal{W}(\pi, \psi)$. By [26, Lemma 2.19.1], the formula

$$\langle W_1, W_2 \rangle = \int_{F^\times} W_1(a(y))\overline{W_2(a(y))}\, d^\times y \quad (W_1, W_2 \in \pi) \tag{27}$$

defines a G-invariant hermitian pairing on π.

Definition 2.9. The *normalized Whittaker newform* $W \in \pi$ is the unique vector invariant under $K_0(\mathfrak{p}^{a(\pi)})$ that satisfies $\langle W, W\rangle = 1$ and $W(1) > 0$.

Remark 2.10. One can check that $W(1) = 1$ whenever $a(\pi) \geq 2$.

Let $s \in \mathbb{C}$ be a complex parameter. We realize $\pi_3 = |.|^{s-1/2} \boxplus |.|^{1/2-s}$ in its induced model, and let $f_s \in \pi_3$ denote the unique K-invariant vector that satisfies $f_s(1) = 1$. Define the *local Rankin–Selberg integral*

$$J(s) = \int_{NZ\backslash G} W(g)W(a(-1)g)f_s(g)\, dg, \tag{28}$$

where $W \in \pi$ is the normalized Whittaker newform. It is well-known that the RHS converges absolutely in some nonempty vertical strip and extends to a meromorphic function of s on the complex plane (see [25]). Using the identity $W(a(-1)g) = \overline{W(g)}$ and the Iwasawa decomposition, we can rewrite this definition as

$$J(s) = \int_{k\in K}\int_{y\in F^\times} |W|^2(a(y)k)|y|^{s-1}\, d^\times y\, dk. \tag{29}$$

or alternatively, using the Bruhat decomposition (see (23)), as

$$J(s) = \frac{\zeta_p(2)}{\zeta_p(1)} \int_{x\in F} \max(1,|x|)^{-2s} \int_{y\in F^\times} |W|^2(a(y)wn(x))|y|^{s-1}\, d^\times y\, dx. \tag{30}$$

The following important result is a consequence of Lemma 3.4.2 in [31].

Proposition 2.11. *Suppose that π satisfies Condition 2.1. The integral $I(s)$, defined initially for $\mathrm{Re}(s) = 1/2$ or $s \in (2\lambda(\pi), 1-2\lambda(\pi))$, extends to a meromorphic function of s on the entire complex plane. We have an identity of meromorphic functions*

$$I(s) = (1-p^{-1})^{-1}J(s)J(1-s). \tag{31}$$

Proof. Denote by $\mathcal{D} = \{s \in \mathbb{C} : \mathrm{Re}(s) = 1/2\} \cup (2\lambda(\pi), 1-2\lambda(\pi))$ the cross on which $I(s)$ was defined, and let $s \in \mathcal{D}$. Then π_3 is irreducible and unitarizable. We normalize the (unique up to scaling) G-invariant hermitian pairing $\langle,\rangle$ on π_3 so that $\langle f_s, f_s\rangle = 1$.[16] With this normalization,

[16]In the tempered case $\mathrm{Re}(s) = 1/2$, we have explicitly

$$\langle f, f'\rangle = \int_{k\in K} f(k)\overline{f'(k)}\, dk \quad (f, f' \in \pi_3).$$

When $s \in (0,1)$, the formula for the pairing is slightly more complicated.

the definition (26) reads

$$I(s) = \int_{Z\backslash G} \langle gW, W\rangle^2 \langle gf_s, f_s\rangle \, dg. \tag{32}$$

This integral converges absolutely and locally uniformly on $\mathcal{D}$.

We observe that $\langle gf_s, f_s\rangle$ extends to an entire function of s, and in fact a polynomial function of $p^{\pm s}$; explicitly,

$$\langle k_1 a(\varpi^m) k_2 f_s, f_s\rangle = p^{-m/2}(1+p^{-1})^{-1}(\lambda_{s,m} - p^{-1}\lambda_{s,m-2})$$

with $\lambda_{s,m}$ as in (25) for all $k_1, k_2 \in K$, $m \geq 1$. Moreover, we have the majorization $|\langle gf_s, f_s\rangle| \leq \langle gf_\sigma, f_\sigma\rangle \in \mathbb{R}_{\geq 0}$ with $\sigma = \mathrm{Re}(s)$. Consequently, the integral (32) converges normally and defines a holomorphic function on the strip $\mathcal{D}' = \{s \in \mathbb{C} : \mathrm{Re}(s) \in (2\lambda(\pi), 1 - 2\lambda(\pi))\}$.

The relation (31) on the line $\mathrm{Re}(s) = 1/2$ follows from Lemma 3.4.2 in [31] upon noting that $J(s) = \overline{J(1-s)}$ whenever $\mathrm{Re}(s) = 1/2$. Since both sides of (31) vary analytically with s on the strip $\mathcal{D}'$, we obtain at once the meromorphic continuation of $J(s)$ to the complex plane and the general case of the identity (32). $\square$

Proposition 2.11 is significant for our purposes because it reduces the evaluation of the integral $I(s)$, which appears in Ichino's formula, to that of the simpler integral $J(s)$.

2.4. **The local functional equation.** Let $\pi = \mathcal{W}(\pi, \psi)$ be a generic irreducible admissible unitarizable representation of G with trivial central character, let $W \in \pi$ be the normalized Whittaker newform, and let $J(s)$ be the local Rankin–Selberg integral. The main difficulty in computing $J(s)$, and hence $I(s)$, is that $W(g)$ has no simple formula when $a(\pi) \geq 2$. In Section 2.6, we will split the integral (30) defining $J(s)$ into several pieces. Initially, we will be able to evaluate at least half of these pieces. The key tool that will enable us to compute the remaining pieces is the local functional equation for $\mathrm{GL}(2) \times \mathrm{GL}(2)$, which we now recall in a specialized form. It is convenient to define the *normalized local Rankin–Selberg integral*

$$J^*(s) = \frac{J(s)\zeta_p(2s)}{L(\pi \times \pi, s)},$$

and to introduce the shorthand $C = C(\pi \times \pi)$.

Proposition 2.12 (Local functional equation for $\mathrm{GL}(2) \times \mathrm{GL}(2)$)**.** *$J^*(s)$ extends to a polynomial function of $p^{\pm s}$ that satisfies the functional equation $J^*(s) = C^{s-1/2} J^*(1-s)$.*

Proof. This follows from (1.1.5) of [11] by taking the Schwartz function Φ to be the characteristic function of $\mathfrak{o} \times \mathfrak{o}$. We have used here that the epsilon factor $\varepsilon(s, \pi \times \pi, \psi)$ equals $C^{1/2-s}$. This follows from the fact that the local root number of $\pi \times \pi$ is equal to $+1$; see the proof of Prop. 2.1 of [36]. $\square$

Suppose that π satisfies Condition 2.1. From Proposition 2.11 and the definitions of $I^*(s)$ and $J^*(s)$, we readily derive the formula

$$I^*(s) = (1+p^{-1})^2 L(\mathrm{ad}\pi, 1)^2 J^*(s) J^*(1-s). \tag{33}$$

By Proposition 2.12, $J^*(s)$ is an entire function of s. It follows that $I^*(s)$ is also entire as a function of s. By contrast, $I(s)$ may have poles. Using soft analytic techniques, we deduce from Proposition 2.12 the *local convexity bound* described in the introduction.

Corollary 2.13 (Local convexity bound). *For $0 \leq \mathrm{Re}(s) \leq 1$, we have $J^*(s) \ll C^{-1/2+\mathrm{Re}(s)/2}$ and $I^*(s) \ll C^{-1/2}$ with absolute implied constants.*

Proof. By (33), it suffices to prove the first part of the statement. Using (29) and the fact that $W(g)$ is L^2-normalized, we get the trivial bound $J^*(s) \ll 1$ for $\mathrm{Re}(s) = 1$. We transfer this to the bound $J^*(s) \ll C^{-1/2}$ for $\mathrm{Re}(s) = 0$ via Proposition 2.12. We interpolate these two bounds by the Phragmen–Lindelöf theorem, which in this context is nothing more than the maximum modulus principle, to deduce that $J^*(s) \ll C^{-1/2+\mathrm{Re}(s)/2}$ for all s with $0 \leq \mathrm{Re}(s) \leq 1$. □

2.5. **Properties of Whittaker functions.** Let $\pi = \mathcal{W}(\pi, \psi)$ be a generic irreducible admissible unitarizable representation of G with trivial central character, and $W \in \pi$ its normalized Whittaker newform. The purpose of this section is to establish the key properties of W that will be used in our proof of Theorem 2.7.

Lemma 2.14 (Invariance of inner product on Whittaker model). *For each $g_1, g_2 \in G$, one has $\int_{y \in F^\times} |W|^2(a(y)g_1)\, d^\times y = \int_{y \in F^\times} |W|^2(a(y)g_2)\, d^\times y = 1$.*

Proof. Since G acts on $\mathcal{W}(\pi, \psi)$ by right translation, the first identity amounts to the fact that integration along A defines a G-invariant hermitian pairing on $\mathcal{W}(\pi, \psi)$ (see the beginning of Section 2.3). The second identity follows from our assumption that W is L^2-normalized. □

Lemma 2.15 (Support condition[17]). *Write $n = a(\pi)$, and suppose that $n \geq 2$. If $|x|^2 < \max(p^n, |y|)$ and $W(a(y)wn(x)) \neq 0$, then $|y| = p^n$.*

Before embarking on the proof of this lemma, we must introduce some notation and recall the local $\mathrm{GL}(2)$ functional equation. Let μ be a character of the unit group $\mathfrak{o}^\times$. We extend μ to a (unitary) character of $F^\times$ (non-canonically) by setting $\mu(\varpi) = 1$, and henceforth denote this extension also by μ. We may write the standard ε-factor for $\pi\mu$ in the form $\varepsilon(\pi\mu, s, \psi) = \varepsilon(\pi\mu)C(\pi\mu)^{1/2-s}$ for some $\varepsilon(\pi\mu) = \varepsilon(\pi\mu, \psi, 1/2) \in \mathbb{C}^1$, where $C(\pi\mu) = p^{a(\pi\mu)}$ is as in Section 2.1.3; for notational simplicity, we suppress the dependence of $\varepsilon(\pi\mu)$ on our fixed choice of uniformizer ϖ and unramified additive character ψ.

With this notation, the local $\mathrm{GL}(2)$ functional equation (see [26]) asserts that for each vector $W' \in \mathcal{W}(\pi, \psi)$, each character μ of $\mathfrak{o}^\times$, and each complex number $s \in \mathbb{C}$, the local zeta integral

$$Z(W', \mu, s) = \int_{F^\times} W'(a(y))\mu(y)|y|^s\, d^\times y$$

satisfies

$$\frac{Z(W', \mu^{-1}, s)}{L(\pi\mu^{-1}, 1/2 + s)} = \varepsilon(\pi\mu)C(\pi\mu)^s \frac{Z(wW', \mu, -s)}{L(\pi\mu, 1/2 - s)}. \tag{34}$$

Proof of Lemma 2.15. Suppose that $|x|^2 < \max(p^n, |y|)$. If $|x|^2 \geq p^n$, then $|y| > |x|^2 \geq p^n$, hence

$$\max\left(\left|\frac{x}{y}\right|, \left|\frac{x^2}{y}\right|, p^n \left|\frac{1}{y}\right|\right) < 1.$$

[17] The reader looking to understand how our arguments would apply to slightly more general vectors $W' \in \pi$ might complain that this condition is very particular to the newform. We refer to Remark 1.9 for a sketch of an alternative, more robust argument that does not make use of this condition.

It follows that for each unit $u \in \mathfrak{o}^\times$, the matrix

$$(a(y)wn(x))^{-1}n(u\varpi^{-1})(a(y)wn(x)) = \begin{bmatrix} 1+\frac{x}{y}u\varpi^{-1} & \frac{x^2}{y}u\varpi^{-1} \\ -\frac{1}{y}u\varpi^{-1} & 1-\frac{x}{y}u\varpi^{-1} \end{bmatrix}$$

belongs to $K_0(\mathfrak{p}^n)$. Therefore $W(a(y)wn(x)) = \psi(u\varpi^{-1})W(a(y)wn(x))$ for all $u \in \mathfrak{o}^\times$. Since ψ has conductor $\mathfrak{o}$, we see that $W(a(y)wn(x)) = 0$.

It remains to consider the case that $|x|^2 < p^n$. Let $W' = wn(x)W$. We wish to show that $W'(a(y)) = 0$ unless $|y| = p^n$. By Fourier inversion on the unit group $\mathfrak{o}^\times$, it is equivalent to show that for each character μ of $\mathfrak{o}^\times$, the zeta integral $Z(W', \mu^{-1}, s)$ is a constant multiple of p^{ns}, where the constant is allowed to depend upon μ but not upon s.

It is a standard fact (see [42, 40]) that the map $F^\times \ni y \mapsto W(a(y))$, and hence also the map

$$F^\times \ni y \mapsto (wW')(a(y)) = (n(x)W)(a(y)) = W(a(y)n(x)) = \psi(xy)W(a(y)), \tag{35}$$

is supported on $\mathfrak{o}^\times$, so that $c_0(\mu) := Z(wW', \mu, -s)$ is independent of s; it is here that we have used the assumption $n \geq 2$. Therefore the functional equation (34) reads

$$Z(W', \mu^{-1}, s) = c_0(\mu)\varepsilon(\pi\mu)C(\pi\mu)^s \frac{L(\pi\mu^{-1}, 1/2+s)}{L(\pi\mu, 1/2-s)},$$

and we reduce to showing that $c_0(\mu) \neq 0$ implies that $C(\pi\mu) = p^n$ and $L(\pi\mu, s) = L(\pi\mu^{-1}, s) = 1$.

The right-$a(\mathfrak{o}^\times)$-invariance of W implies that (35) is invariant under $\mathfrak{o}^\times \cap (1 + x^{-1}\mathfrak{o})$, hence $c_0(\mu) = 0$ unless $C(\mu) \leq |x|$, in which case $C(\mu)^2 \leq |x|^2 < p^n$ and $C(\pi\mu) = p^n$. If π is of Type 1 or Type 2, we deduce immediately that $L(\pi\mu, s) = L(\pi\mu^{-1}, s) = 1$; in the other cases this holds by inspection. $\square$

Remark 2.16. A slight modification of the above argument implies that under the hypotheses of Lemma 2.15, we have

$$W(a(y)wn(x)) = \sum_{\substack{\mu \in \widehat{\mathfrak{o}^\times} \\ C(\pi\mu) = |y|}} \mu(y)\varepsilon(\pi\mu)G(x, \mu), \tag{36}$$

where $G(x, \mu) = \int_{u \in U} \psi(xu)\mu(u)$ is a Gauss-Ramanujan sum. Note that the characters μ contributing nontrivially to (36) are those for which $G(x, \mu) \neq 0$, which implies that $C(\mu) \leq |x|$.

2.6. **The proofs.** Our aim in this section is to prove Theorem 2.7; along the way, we will also establish Proposition 2.5. Let π satisfy Condition 2.1. Recall the notation $n = a(\pi)$ and $N = a(\pi \times \pi)$. Suppose that $n \geq 2$. By (31), the calculation of $I^*(s)$ reduces to that of $J^*(s)$. Let T_m be the coefficient of p^{ms} therein:

$$J^*(s) = \sum_{m \in \mathbb{Z}} T_m p^{ms}. \tag{37}$$

Recalling that $J^*(s)$ is a polynomial in $p^{\pm s}$ and applying its functional equation $J^*(s) = (p^N)^{s-1/2}J^*(1-s)$ (Proposition 2.12), we see that $T_m = 0$ for almost all m and

$$T_{-m+N} = p^{m-\frac{N}{2}}T_m. \tag{38}$$

Table 1. Relation between T_m and R_m

Representation	$L(\pi \times \pi, s)^{-1}\zeta_p(2s)$	T_m in terms of R_m
Type 1	1	R_m
Type 2	$\frac{1}{1+p^{-s}}$	$\sum_{r=0}^{\infty}(-1)^r R_{m+r}$
Type 3	$\frac{(1-p^{-s})(1-p^{2s_0-s})(1-p^{-2s_0-s})}{1+p^{-s}}$	$R_m - \beta_{s_0}R_{m+1} - R_{m+2} + 2\beta_{s_0}\sum_{r=2}^{\infty}(-1)^r R_{m+r}$
Type 4	$\frac{1-p^{-s}}{1+p^{-s}}$	$R_m + 2\sum_{r=1}^{\infty}(-1)^r R_{m+r}$
Type 5	$\frac{1-p^{-s-1}}{1+p^{-s}}$	$R_m + (1+p^{-1})\sum_{r=1}^{\infty}(-1)^r R_{m+r}$

Setting $s = 1$ in (37) and using the identity $J(1) = 1$, we obtain

$$\sum_m T_m p^m = \frac{\zeta_p(2)}{L(\pi \times \pi, 1)}. \tag{39}$$

Closely related to T_m are the quantities R_m defined by

$$J(s) = \sum_{m \in \mathbb{Z}} R_m p^{ms}. \tag{40}$$

A linear relation between the sequences T_m and R_m follows immediately from the definition

$$J^*(s) = \frac{J(s)\zeta_p(2s)}{L(\pi \times \pi, s)}.$$

For convenience, we explicate this relation case-by-case in Table 1.

Let us now explain our strategy for computing $J^*(s)$. In view of (37), (38) and (39), it suffices to compute T_m for positive m. Using Table 1, we reduce further to computing R_m for positive m.

The definition (30) of $J(s)$ implies that

$$R_m = \frac{\zeta_p(2)}{\zeta_p(1)} \int_{\substack{x \in F \\ |y|/\max(1,|x|)^2 = p^m}} \int_{y \in F^\times} |W|^2(a(y)wn(x))|y|^{-1}\, d^\times y\, dx.$$

Let x, y be as in the integrand above, and suppose that $W(a(y)wn(x)) \neq 0$. Since $m > 0$, we have $|x|^2 < |y|$. By the support condition on W (Lemma 2.15), we deduce that $|y| = p^n$. Therefore

$$R_m = \frac{\zeta_p(2)}{\zeta_p(1)} p^{-n} \int_{\substack{x \in F \\ \max(1,|x|)^2 = p^{n-m}}} \int_{y \in F^\times} |W|^2(a(y)wn(x))\, d^\times y\, dx.$$

By the invariance of the inner product on $\mathcal{W}(\pi, \psi)$ and our assumption that W is L^2-normalized (Lemma 2.14), we deduce that the integral over y is identically 1, and in particular, independent of x. We summarize thusly:

Proposition 2.17. *Let m be a positive integer. Then*

$$R_m = \frac{\zeta_p(2)}{\zeta_p(1)} p^{-n} \operatorname{vol}(\{x \in F : \max(1, |x|)^2 = p^{n-m}\}, dx).$$

Explicitly,

(1) $R_m = 0$ *if either*
- $m > n$, *or*
- $1 \le m < n$ *and* $m - n$ *is odd.*

(2) $R_n = \frac{p^{-n}}{1+p^{-1}}$

(3) $R_m = p^{\frac{-n-m}{2}} \frac{1-p^{-1}}{1+p^{-1}}$ *if* $1 \le m < n$ *and* $m - n$ *is even.*

In order to prove Theorem 2.7, it suffices to evaluate T_m for each m and then use (33) and (37). From (38) and Proposition 2.17 we know that $T_m = 0$ if $m > n$ or $m < N - n$. For the remaining values of m, the evaluation of T_m follows by collecting together (38), (39) Proposition 2.17, and the relations in Table 1. We record the results.

Type 1. In this case π is a dihedral supercuspidal representation $\rho(E/F, \xi)$, associated to the unramified quadratic extension E of F and to a non-Galois-invariant character ξ of $E^\times$. A standard computation [40] shows that $n = 2a(\xi)$ and $N = 2a(\xi^2)$. This shows that n and N are even and $N \le n$. As for T_m, we have $T_m = 0$ if $m > n$ or $m < N - n$ or $N - n \le m \le n$ and $m - n$ is odd; $T_m = \frac{p^{-n}}{1+p^{-1}}$ if $m = n$; $T_m = \frac{p^{-\frac{N}{2}}}{1+p^{-1}}$ if $m = N - n$ and $T_m = p^{\frac{-n-m}{2}} \frac{1-p^{-1}}{1+p^{-1}}$ in the remaining cases.

Type 2. In this case we will prove that

$$T_m = (-1)^{m+n} \, \frac{p^{\lfloor \frac{-m-n}{2} \rfloor}}{1 + p^{-1}} \tag{41}$$

unless we have $m > n$ or $m < N - n$, in which case T_m equals 0. Indeed, from Proposition 2.17 and Table 1, we see that (41) holds for m positive. Now, if N were odd, then we would be able to use (38) to find T_m for all m; however, the resulting formula would contradict (39). We conclude that N is even. Now using (38) and (39) we see that T_m is given by (41) for all m in the range $N - n \le m \le n$ and is 0 otherwise.

Next we show that $N = n + 1$ whenever n is odd. Indeed, if not, then we must have either $N \ge n + 2$ or $N \le n$. In the first case, (41) implies that $T_1 = T_0 = 0$, and hence (by the relation $R_0 = T_1 + T_0$) that $R_0 = 0$. This is a contradiction since (29) shows immediately that $R_0 \ge \int_K |W(k)|^2 dk > 0$ since $W(1) > 0$. In the second case, (41) implies that $T_0 = (-1)^n \, \frac{p^{\lfloor -\frac{n}{2} \rfloor}}{1+p^{-1}}$ and $T_1 = (-1)^{n+1} \, \frac{p^{\lfloor \frac{-n-1}{2} \rfloor}}{1+p^{-1}}$. As n is odd, we have $T_0 = -T_1$ and because $T_0 = R_0 - T_1$, this implies that $R_0 = 0$, once again leading to the same contradiction.

Type 3. In this case, we must have $n = 2a(\beta)$, $N = 0$. We have $T_m = 0$ if $m > n$ or $m < N - n = -n$. First assume that p is odd; so $a(\beta) = 1$. We have $T_2 = \frac{p^{-2}}{1+p^{-1}}$ and $T_1 = -\beta_{s_0} \frac{p^{-2}}{1+p^{-1}}$. From (38), it follows that $T_{-1} = -\beta_{s_0} \frac{p^{-1}}{1+p^{-1}}$ and $T_{-2} = \frac{1}{1+p^{-1}}$. It is left to calculate T_0. For that we use the fact that $\sum T_m p^m = \frac{\zeta_p(2)}{L(\pi \times \pi, 1)}$. This gives us $T_0 = p^{-1} \frac{1-2p^{-1}-p^{-2}+2\beta_{s_0} p^{-1}}{1+p^{-1}}$. The case $p = 2$ is similar, except that now $a(\beta) \in \{2, 3\}$. We compute T_{n-2}, T_{n-3} from Table 1 (since $n \ge 4$). We omit the details.

Type 4. In this case $n = 2a(\beta)$ and $N = 2a(\beta^2)$. So n and N are even and $N \le n$. As always, we have $T_m = 0$ if $m > n$ or $m < N - n$. For the remaining cases, we compute $T_m = \frac{p^{-n}}{1+p^{-1}}$ if

$m = n$; $T_m = \frac{p^{-\frac{N}{2}}}{1+p^{-1}}$ if $m = N - n$; $T_m = p^{\frac{-n-m}{2}}$ if $0 < n - m < 2n - N$ and $m - n$ is even; $T_m = -\frac{2p^{\frac{-n-m-1}{2}}}{1+p^{-1}}$ if $0 < n - m < 2n - N$ and $m - n$ is odd.

Type 5. In this case $N = 2$ and $n = 2a(\beta)$. As always, we have $T_m = 0$ if $m > n$ or $m < N - n$. Moreover $T_m = \frac{p^{-n}}{1+p^{-1}}$ if $m = n$; $T_m = \frac{p^{-\frac{N}{2}}}{1+p^{-1}}$ if $m = N-n$; $T_m = p^{\frac{-n-m}{2}} \frac{1+p^{-2}}{1+p^{-1}}$ if $0 < n-m < 2n-N$ and $m - n$ is even; $T_m = -p^{\frac{-n-m-1}{2}}$ if $0 < n - m < 2n - N$ and $m - n$ is odd .

By substituting the above formulas into (37), we get an explicit formula for $J^*(s)$. This immediately proves Theorem 2.7 using the relation (33). We note here the precise relation between $Q_{\pi,p}(s)$ and $J^*(s)$,

$$p^{\frac{Ns}{2}} Q_{\pi,p}(s) = (1 + p^{-1}) p^{\frac{N}{4}+\frac{n}{2}} J^*(s).$$

Note that along the way we have also proved Proposition 2.5.

Finally, one can easily derive explicit formulas for R_m for all m from those for T_m calculated above and the relations written down in Table 1. For example, for Type 1 representations, we have

(1) $R_m = 0$ if either
   - $m > n$, or $m < N - n$, or
   - $N - n \le m < n$ and $m - n$ is odd.

(2) $R_n = \frac{p^{-n}}{1+p^{-1}}$.

(3) $R_{N-n} = \frac{p^{-N/2}}{1+p^{-1}}$.

(4) $R_m = p^{\frac{-n-m}{2}} \frac{1-p^{-1}}{1+p^{-1}}$ if $N - n \le m < n$ and $m - n$ is even.

The values of R_m for $m \le 0$ are related to the Fourier coefficients at various cusps of a newform corresponding to π (see Section 3.4).

3. Proof of Theorem 1.2

3.1. **Background and notations.** In this subsection we collect some notation that will be used frequently in this section. For complete definitions and proofs, we refer the reader to Serre [41], Shimura [43], Iwaniec [21, 22] and Atkin–Lehner [1]. We note that some of this (boilerplate) subsection is borrowed from [33].

General notations. For an integer n and a prime p, we let n_p denote the largest divisor of n that is a power of p, and let $n_\diamond$ denote the largest integer such that $n_\diamond^2$ divides n. In words, n_p is the "p-part" of n (the maximal p-power divisor), while $n_\diamond^2$ is the "square part" of n (the maximal square divisor). Note that $n_p = |n|_p^{-1}$ where $|n|_p$ denotes the p-adic absolute value. We let n_0 denote the largest squarefree divisor of n. One could also write $n_p = (n, p^\infty)$ and $n_0 = (n, \prod_p p)$. We have $n_\diamond = 1$ if and only if $n_0 = n$ if and only if n is squarefree, but there is in general no simple relation between $n_\diamond$ and n_0.

Given a finite collection of rational numbers $\{\dots, a_i, \dots\}$, the greatest common divisor $(\dots, a_i, \dots)$ (resp. least common multiple $[\dots, a_i, \dots]$) is the unique nonnegative generator of the (principal) $\mathbb{Z}$-submodules $\sum \mathbb{Z}a_i$ (resp. $\cap \mathbb{Z}a_i$) of $\mathbb{Q}$. In particular, if a and b are two positive rational numbers with prime factorizations $a = \prod p^{a_p}$, $b = \prod p^{b_p}$, then we have $(a, b) = \prod p^{\min(a_p, b_p)}$ and $[a, b] = \prod p^{\max(a_p, b_p)}$. We write $a|b$ to denote that the ratio b/a is an integer.

For each complex number z, we write $e(z) := e^{2\pi i z}$. For each positive integer n, we let $\varphi(n)$ denote the Euler phi function $\varphi(n) = \#(\mathbb{Z}/n)^\times = \#\{a \in \mathbb{Z} : 1 \leq a \leq n, (a, n) = 1\}$. We let $\tau(n)$ denote the number of positive divisors of n and $\omega(n)$ the number of prime divisors of n.

The upper-half plane. We shall make use of notation for the upper half-plane $\mathbb{H} = \{z \in \mathbb{C} : \mathrm{Im}(z) > 0\}$, the modular group $\Gamma = \mathrm{SL}(2, \mathbb{Z}) \circlearrowright \mathbb{H}$ acting by fractional linear transformations, its congruence subgroup $\Gamma_0(q)$ consisting of those elements with lower-left entry divisible by q, the modular curve $Y_0(q) = \Gamma_0(q)\backslash\mathbb{H}$, the Poincaré measure $d\mu = y^{-2}\, dx\, dy$, and the stabilizer $\Gamma_\infty = \{\pm \left(\begin{smallmatrix} 1 & n \\ & 1 \end{smallmatrix}\right) : n \in \mathbb{Z}\}$ in Γ of $\infty \in \mathbb{P}^1(\mathbb{R})$. We denote a typical element of $\mathbb{H}$ as $z = x + iy$ with $x, y \in \mathbb{R}$.

Holomorphic newforms. Let k be a positive even integer, and let α be an element of $\mathrm{GL}(2, \mathbb{R})$ with positive determinant; the element α acts on $\mathbb{H}$ by fractional linear transformations in the usual way. Given a function $f : \mathbb{H} \to \mathbb{C}$, we denote by $f|_k\alpha$ the function $z \mapsto \det(\alpha)^{k/2} j(\alpha, z)^{-k} f(\alpha z)$, where $j\left(\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right], z\right) = cz + d$.

A *holomorphic cusp form* on $\Gamma_0(q)$ of weight k is a holomorphic function $f : \mathbb{H} \to \mathbb{C}$ that satisfies $f|_k\gamma = f$ for all $\gamma \in \Gamma_0(q)$ and vanishes at the cusps of $\Gamma_0(q)$. A *holomorphic newform* is a cusp form that is an eigenform of the algebra of Hecke operators and orthogonal with respect to the Petersson inner product to the oldforms (see [1]). We say that a holomorphic newform f is a *normalized holomorphic newform* if moreover $\lambda_f(1) = 1$ in the Fourier expansion

$$y^{k/2} f(z) = \sum_{n \in \mathbb{N}} \frac{\lambda_f(n)}{\sqrt{n}} \kappa_f(ny) e(nx), \tag{42}$$

where $\kappa_f(y) = y^{k/2} e^{-2\pi y}$; in that case the Fourier coefficients $\lambda_f(n)$ are real, multiplicative, and satisfy [5, 6] the Deligne bound $|\lambda_f(n)| \leq \tau(n)$.

Recall, from Section 1.1, the definitions of the measures μ and μ_f on $Y_0(1)$, given by

$$\mu(\phi) = \int_{\Gamma\backslash\mathbb{H}} \phi(z) \,\frac{dx\, dy}{y^2}, \quad \mu_f(\phi) = \int_{\Gamma_0(q)\backslash\mathbb{H}} \phi(z)|f|^2(z) y^k \,\frac{dx\, dy}{y^2}$$

for all bounded measurable functions ϕ on $Y_0(1)$.

Maass forms. A *Maass cusp form* (of level 1, on $\Gamma_0(1)$, on $Y_0(1)$, ...) is a Γ-invariant eigenfunction of the hyperbolic Laplacian $\Delta := y^{-2}(\partial_x^2 + \partial_y^2)$ on $\mathbb{H}$ that decays rapidly at the cusp of Γ. By the "$\lambda_1 \geq 1/4$" theorem (see [21, Corollary 11.5]) there exists a real number $r \in \mathbb{R}$ such that $(\Delta + 1/4 + r^2)\phi = 0$; our arguments use only that $r \in \mathbb{R} \cup i(-1/2, 1/2)$, which follows from the nonnegativity of Δ.

A *Maass eigencuspform* is a Maass cusp form that is an eigenfunction of the Hecke operators at all finite places and of the involution $T_{-1} : \phi \mapsto [z \mapsto \phi(-\bar{z})]$; these operators commute with one another as well as with Δ. A Maass eigencuspform ϕ has a Fourier expansion

$$\phi(z) = \sum_{n \in \mathbb{Z}_{\neq 0}} \frac{\lambda_\phi(n)}{\sqrt{|n|}} \kappa_\phi(ny) e(nx) \tag{43}$$

where $\kappa_\phi(y) = 2|y|^{1/2} K_{ir}(2\pi|y|) \mathrm{sgn}(y)^{\frac{1-\delta}{2}}$ with K_{ir} the standard K-Bessel function, $\mathrm{sgn}(y) = 1$ or -1 according as y is positive or negative, and $\delta \in \{\pm 1\}$ the T_{-1}-eigenvalue of ϕ; note that the argument ny of $\kappa_\phi(ny)$ in (43) may be negative even if y is positive. A *normalized Maass eigencuspform* further satisfies $\lambda_\phi(1) = 1$; in that case the coefficients $\lambda_\phi(n)$ are real and multiplicative.

Because $f(-\bar{z}) = \overline{f(z)}$ for each normalized holomorphic newform f, we have $\mu_f(\phi) = 0$ whenever $T_{-1}\phi = \delta\phi$ with $\delta = -1$. Thus we shall assume throughout the rest of this paper that $\delta = 1$, i.e., that ϕ is an *even* Maass form.

Eisenstein series. Let $s \in \mathbb{C}$, $z \in \mathbb{H}$. The *real-analytic Eisenstein series* $E(s,z) = \sum_{\Gamma_\infty \backslash \Gamma} \operatorname{Im}(\gamma z)^s$ converges normally for $\operatorname{Re}(s) > 1$ and continues meromorphically to the half-plane $\operatorname{Re}(s) \geq 1/2$ where the map $s \mapsto E(s,z)$ is holomorphic with the exception of a unique simple pole at $s = 1$ of constant residue $\operatorname{res}_{s=1} E(s,z) = \mu(1)^{-1}$. The Eisenstein series satisfies the invariance $E(s,\gamma z) = E(s,z)$ for all $\gamma \in \Gamma$. When $\operatorname{Re}(s) = 1/2$ we call $E(s,z)$ a *unitary Eisenstein series.* We write E_s for the function $E_s(z) = E(s,z)$.

To each $\Psi \in C_c^\infty(\mathbb{R}_+^*)$, we attach the *incomplete Eisenstein series* $E(\Psi,z) = \sum_{\gamma \in \Gamma_\infty \backslash \Gamma} \Psi(\operatorname{Im}(\gamma z))$, which descends to a compactly supported function on $Y_0(1)$. One can express $E(\Psi,z)$ as a weighted contour integral of $E(s,z)$ via Mellin inversion.

3.2. **An extension of Watson's formula.** The general analytic properties of triple product L-functions on GL(2) follow from an integral representation introduced by Garrett [9] and further developed by Piatetski-Shapiro–Rallis [35].

Harris–Kudla [16] established a general "triple product formula" relating the (magnitude squared of the) integral of the product of three automorphic forms (on quaternion algebras) to the central value of their triple product L-function, with proportionality constants given by somewhat complicated local zeta integrals. Gross and Kudla [15] and Watson [46] evaluated sufficiently many of the Harris–Kudla zeta integrals to obtain a completely explicit triple product formula for each triple of newforms having the *same squarefree level.*

Ichino [19] obtained a more general triple product formula of the type considered by Harris–Kudla, but in which the proportionality constants are given by simpler integrals over the group $\mathrm{PGL}_2(\mathbb{Q}_p)$. Sufficiently many of those simpler integrals were computed in [20, Theorem 1.2] and [33, Lemma 4.2] to derive an explicit triple product formula for each triple of newforms of (not necessarily the same) squarefree level (see [33, Remark 4.2]).

Our local calculations in Section 2 give an explicit triple product formula for certain triples of newforms of not necessarily squarefree level. We state only the identity that we shall need.

Conventions regarding L-functions. Let $\pi = \otimes \pi_v$ be one of the symbols ϕ, f, $\operatorname{ad}\phi$, $\operatorname{ad}f$, or $f \times f \times \phi$; here v traverses the set of places of $\mathbb{Q}$. One can attach a local factor $L_v(\pi,s) = L(\pi_v,s)$ for each v. We write $L(\pi,s) = \prod_p L_p(\pi,s)$ for the finite part of the corresponding global L-function and $\Lambda(\pi,s) = L_\infty(\pi,s)L(\pi,s) = \prod_v L_v(\pi,s)$ for its completion. The functional equation relates $L(\pi,s)$ and $L(\pi,1-s)$.

For the convenience of the reader, we collect here some references for the definitions of $L(\pi,s)$ with π as above. Watson [46, Section 3.1] is a good reference for squarefree levels. In general, the standard L-functions attached to $\pi = f$ and $\pi = \phi$ may be found in a number of sources (see for instance [10, 26, 3]). Since ϕ has trivial central character and is everywhere unramified, we may write $L_v(\phi,s) = \zeta_v(s+s_0)\zeta_v(s-s_0)$ for some $s_0 \in \mathbb{C}$, where $\zeta_\infty(s) = \pi^{-s/2}\Gamma(s/2)$ and $\zeta_p(s) = (1-p^{-s})^{-1}$. Then $L_v(f \times f \times \phi, s) = L_v(f \times f, s+s_0)L_v(f \times f, s-s_0)$. It is known that $L_v(f \times f, s)$ factors as $L_v(\operatorname{ad}f, s)\zeta_v(s)$. Finally, the local factors $L_v(\operatorname{ad}f, s)$ may be found in [12].

Theorem 3.1. *Let ϕ be a Maass eigencuspform of level 1. Let f be a holomorphic newform on $\Gamma_0(q)$, $q \in \mathbb{N}$. Then*

$$\frac{\left|\int_{\Gamma_0(q)\backslash\mathbb{H}} \phi(z)|f|^2(z)y^k \, \frac{dx\,dy}{y^2}\right|^2}{\left(\int_{\Gamma\backslash\mathbb{H}} |\phi|^2(z)\, \frac{dx\,dy}{y^2}\right)\left(\int_{\Gamma_0(q)\backslash\mathbb{H}} |f|^2(z)y^k \, \frac{dx\,dy}{y^2}\right)^2} = \frac{1}{8q}\frac{\Lambda(\phi\times f\times f, \frac{1}{2})}{\Lambda(\mathrm{ad}\phi,1)\Lambda(\mathrm{ad}f,1)^2}\prod_{p|q_\diamond}\left(L_p(\mathrm{ad}f,1)\cdot Q_{f,p}(s_{\phi,p})\right)^2,$$

with $s = s_{\phi,p} \in \mathbb{C}$ chosen so that the pth normalized Hecke eigenvalue of ϕ is $p^{s-1/2}+p^{1/2-s}$ and the local factors $Q_{f,p}(s_{\phi,p})$ as in Theorem 2.7.

Proof. Ichino's generalization of Watson's formula [19] reads

$$\frac{\left|\int_{\Gamma_0(q)\backslash\mathbb{H}} \phi(z)|f|^2(z)y^k \, \frac{dx\,dy}{y^2}\right|^2}{\left(\int_{\Gamma\backslash\mathbb{H}} |\phi|^2(z)\, \frac{dx\,dy}{y^2}\right)\left(\int_{\Gamma_0(q)\backslash\mathbb{H}} |f|^2(z)y^k \, \frac{dx\,dy}{y^2}\right)^2} = \frac{1}{8}\frac{\Lambda(f\times f\times\phi,1/2)}{\Lambda(\mathrm{ad}\phi,1)\Lambda(\mathrm{ad}f,1)^2}\prod I_v^*, \tag{44}$$

where I_p^* was defined and explicitly calculated in Section 2 and $I_\infty^* \in \{0,1,2\}$ (see [46]). In our case, $I_\infty^* = 1$. The result now follows from Theorems 2.6 and 2.7. □

Remark 3.2. A conclusion analogous to that of Theorem 3.1 holds also when $\phi = E_s$ is an Eisenstein series, in which case the computation follows more directly from the Rankin–Selberg method and the calculations of Section 2. See also [31, Section 4.4].

3.3. **Bound for $D_f(\phi)$ in terms of L-functions.** We briefly recall the setup for Theorem 1.2. Let f be a holomorphic newform of weight $k \in 2\mathbb{N}$ on $\Gamma_0(q)$. We assume without loss of generality that f is a normalized newform. Fix a Maass eigencuspform or incomplete Eisenstein series ϕ on $Y_0(1) = \Gamma_0(1)\backslash\mathbb{H}$. We wish to prove the bound asserted by Theorem 1.2, i.e., that

$$D_f(\phi) := \frac{\mu_f(\phi)}{\mu_f(1)} - \frac{\mu(\phi)}{\mu(1)} \ll_\phi (q/q_0)^{-\delta_1}\log(qk)^{-\delta_2}$$

for some $\delta_1, \delta_2 > 0$, with q_0 the largest squarefree divisor of q. For simplicity, we treat in detail only the case that ϕ is a Maass eigencuspform, since the changes required to treat incomplete Eisenstein series are exactly as in [33].[18]

We collect first an upper bound for $D_f(\phi)$ obtained by combining the extension of Watson's formula (Theorem 3.1) with Soundararajan's weak subconvex bounds [44].

Proposition 3.3. *For each holomorphic newform f on $\Gamma_0(q)$ and each Maass eigencuspform ϕ (of level 1), we have*

$$|D_f(\phi)|^2 \ll_\phi \frac{1}{q}\frac{\Lambda(f\times f\times\phi,1/2)}{\Lambda(\mathrm{ad}f,1)^2}\ 10^{5\omega(q/\sqrt{C})}\tau(q/\sqrt{C})^2(q/\sqrt{C})^{2\theta}.$$

where $\theta \in [0, 7/64]$ (see [27]) is a bound towards the Ramanujan conjecture for Maass forms on $\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}$.

[18] However, one obtains different numerical values for δ_1, δ_2 when ϕ is an incomplete Eisenstein series; see the statement of Theorem 3.19.

Proof. Let C be the (finite) conductor of $f \times f$. Then C is a perfect square, and $\sqrt{C}$ divides q. The result now follows from Theorem 3.1 and the bounds of Corollary 2.8. $\square$

The analytic conductor of $f \times f \times \phi$ is $\asymp C^2 k^4$, so the arguments of Soundararajan [44] imply that

$$L(f \times f \times \phi, 1/2) \ll \frac{\sqrt{C}k}{\log(Ck)^{1-\varepsilon}}.$$

By Stirling's formula as in [44, Proof of Cor 1], we deduce:

Proposition 3.4.

$$|D_f(\phi)|^2 \ll_\phi \frac{1}{L(\mathrm{ad} f, 1)^2} \frac{10^{5\omega(q/\sqrt{C})}}{\log(Ck)^{1-\varepsilon}} \frac{\tau(q/\sqrt{C})^2}{(q/\sqrt{C})^{1-2\theta}}. \tag{45}$$

Note that $q/\sqrt{C} \in \mathbb{N}$ (cf. Prop 2.5). Furthermore, when q is squarefree, we have $C = q^2$, so that the third factor on the RHS of (45) is absent.

Remark 3.5. The same bound holds when ϕ is a unitary Eisenstein series, and with uniform implied constants. By Mellin inversion, the bound holds also when ϕ is an incomplete Eisenstein series (c.f. [44, Proof of Cor 1] or [33, Proof of Prop 5.3]).

3.4. **Cusps of $\Gamma_0(q)$ and Fourier expansions.** We collect some (to the best of our knowledge, non-standard) information concerning the Fourier expansions of newforms at arbitrary cusps of $\Gamma_0(q)$ (§3.4.2). To illuminate that discussion, we take some time to recall in detail certain comparatively standard facts concerning the cusps of $\Gamma_0(q)$ themselves (§3.4.1).

3.4.1. *Background on cusps.* The group $G := \mathrm{PGL}_2^+(\mathbb{R})$ acts on the upper half-plane $\mathbb{H}$ and its boundary $\mathbb{P}^1(\mathbb{R})$ by fractional linear transformations. For each lattice (i.e., discrete subgroup of finite covolume) $\Delta < G := \mathrm{PGL}_2^+(\mathbb{R})$, let $\mathcal{P}(\Delta)$ denote the set of boundary points $\mathfrak{a} \in \mathbb{P}^1(\mathbb{R})$ stabilized by a nonscalar element of Δ; one might call $\mathcal{P}(\Delta)$ the set of *parabolic vertices* of Δ. Equivalently, for each $\mathfrak{a} \in \mathbb{P}^1(\mathbb{R})$, let $U_\mathfrak{a}$ denote the unipotent radical of the parabolic subgroup $P_\mathfrak{a} = \mathrm{Stab}_G(\mathfrak{a})$. Then $\mathcal{P}(\Delta) = \{\mathfrak{a} \in \mathbb{P}^1(\mathbb{R}) : \mathrm{vol}(U_\mathfrak{a}/U_\mathfrak{a} \cap \Delta) < \infty\}$.

The group Δ acts on $\mathcal{P}(\Delta)$, and the orbit space $\mathcal{C}(\Delta) := \Delta\backslash\mathcal{P}(\Delta)$ is called the set of *cusps* of Δ. One may take as representatives for $\mathcal{C}(\Delta)$ the set of parabolic vertices of a given fundamental polygon for $\Delta\backslash\mathbb{H}$. Intrinsically, $\mathcal{C}(\Delta)$ is in bijection with the set of Δ-conjugacy classes of parabolic subgroups $P < G$ whose unipotent radical U satisfies $\mathrm{vol}(U/U \cap \Delta) < \infty$.

Recall that $\Gamma = \Gamma_0(1) = \mathrm{SL}_2(\mathbb{Z})$, and set henceforth $\Gamma' = \Gamma_0(q)$. Then $\mathcal{P}(\Gamma) = \mathcal{P}(\Gamma') = \mathbb{P}^1(\mathbb{Q})$. The action of Γ on $\mathcal{P}(\Gamma)$ is transitive, and the stabilizer in Γ (as well as in Γ') of $\infty \in \mathcal{P}(\Gamma)$ is $\Gamma_\infty = \{\pm \left(\begin{smallmatrix} 1 & n \\ & 1 \end{smallmatrix}\right) : n \in \mathbb{Z}\}$. Thus we have the left Γ-set $\mathcal{P}(\Gamma) = \Gamma/\Gamma_\infty$, the left Γ'-set $\mathcal{P}(\Gamma') = \Gamma/\Gamma_\infty$ and their orbit spaces $\mathcal{C}(\Gamma) = \Gamma\backslash\Gamma/\Gamma_\infty = \{1\}$, $\mathcal{C}(\Gamma') = \Gamma'\backslash\Gamma/\Gamma_\infty$.

For an arbitrary ring R, the group Γ has a natural right action on the set $\mathbb{P}^1(R)$, realized as row vectors: $[x : y] \cdot \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = [ax + cy : bx + dy]$. The congruence subgroup $\Gamma_0(q)$ is then the stabilizer in Γ of $[0 : 1] \in \mathbb{P}^1(\mathbb{Z}/q)$. The group Γ acts transitively on $\mathbb{P}^1(\mathbb{Z}) = \mathbb{P}^1(\mathbb{Q})$, hence on $\mathbb{P}^1(\mathbb{Z}/q)$, and so we may identify $\Gamma'\backslash\Gamma = \mathbb{P}^1(\mathbb{Z}/q)$ as right Γ-sets. Under this identification, $\alpha = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \Gamma$ corresponds to $[c : d] \in \mathbb{P}^1(\mathbb{Z}/q)$. Two row vectors $[c : d]$ and $[c' : d']$ with $(c, d) = (c', d') = 1$ represent the same element of $\mathbb{P}^1(\mathbb{Z}/q)$ if and only if there exists $\lambda \in (\mathbb{Z}/q)^\times$ for which $c' = \lambda c$ and $d' = \lambda d$. Thus $\mathbb{P}^1(\mathbb{Z}/q)$ may be identified with the set of diagonal $(\mathbb{Z}/q)^\times$-orbits on the set of ordered pairs $[c : d]$ of relatively prime residue classes $c, d \in \mathbb{Z}/q$. In each such orbit there is a pair

$[c:d]$ for which c divides q;[19] if $[c,d]$ is one such pair, then all such pairs arise as $[c:\lambda d]$ for some $\lambda \in (\mathbb{Z}/q)^\times$ that satisfies $\lambda c \equiv c \pmod{q}$, or equivalently $\lambda \equiv 1 \pmod{q/c}$. Thus as c traverses the set of positive divisors of q and d traverses $\{d \in \mathbb{Z}/(q/c) : (d,c,q/c) = 1\}$, the vector $[c:d]$ traverses $\mathbb{P}^1(\mathbb{Z}/q)$.[20]

The element $\left(\begin{smallmatrix} 1 & n \\ & 1 \end{smallmatrix}\right)$ of Γ_∞ sends $[c:d] \in \mathbb{P}^1(\mathbb{Z}/q)$ to $[c:d+nc]$. The orbit of $[c:d]$ in $\mathbb{P}^1(\mathbb{Z}/q)$ may then be identified with the set of all $[c:d']$ where $d' \in \mathbb{Z}/(q/c)$ and $d' \equiv d \pmod{c}$. In summary, each section of the map $\Gamma \ni \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \mapsto [c:d] \in \mathbb{P}^1(\mathbb{Z}/q)$ gives rise to a commutative diagram

$$\begin{array}{ccc}
\Gamma'\backslash\Gamma & \longrightarrow & \Gamma'\backslash\Gamma/\Gamma_\infty = \mathcal{C}(\Gamma') \\
\| & & \| \\
\mathbb{P}^1(\mathbb{Z}/q) & \longrightarrow & \mathbb{P}^1(\mathbb{Z}/q)/\Gamma_\infty \\
\| & & \| \\
\{[c:d] : c|q, d \in \mathbb{Z}/(q/c), (d,c,q/c) = 1\} & \longrightarrow & \{[c:d] : c|q, d \in (\mathbb{Z}/(c,q/c))^\times\}.
\end{array}$$

When $c|q$ and $d \in (\mathbb{Z}/(c,q/c))^\times$, we henceforth write $\mathfrak{a}_{d/c} \in \mathcal{C}(\Gamma')$ for the corresponding cusp. It corresponds to $a/c \in \mathbb{P}^1(\mathbb{Q})$ where a is an integer with $(a,c) = 1$ and $ad \equiv 1 \pmod{(c,q/c)}$.

Thinking of $[c:d]$ as the "fraction" d/c, we define the *denominator* of the cusp $\mathfrak{a}_{d/c}$ to be c, which is by assumption a positive divisor of q.

The *width* of a cusp $\mathfrak{a} \in \mathcal{C}(\Gamma')$ is the index $w_\mathfrak{a} = [\mathrm{Stab}_\Gamma(\mathfrak{a}) : \mathrm{Stab}_{\Gamma'}(\mathfrak{a})]$ of its Γ'-stabilizer in its Γ-stabilizer.[21] Equivalently, if we take as a fundamental domain for $\Gamma'\backslash\mathbb{H}$ a union of translates of fundamental domains for Γ, then the width of $\mathfrak{a}$ is the number of such translates that touch $\mathfrak{a}$ (regarded as a Γ'-orbit of parabolic vertices); in other words, it is the cardinality of the fiber above $\mathfrak{a}$ under the projection $\Gamma'\backslash\Gamma \to \mathcal{C}(\Gamma')$. Let us write π for the bottom horizontal arrow in the above diagram. Then the width of $\mathfrak{a}_{d/c}$ is

$$\#\pi^{-1}(\mathfrak{a}_{d/c}) = \frac{(q/c)(c,q/c)^{-1}\varphi((c,q/c))}{\varphi((c,q/c))} = \frac{q/c}{(c,q/c)} = \frac{q}{(c^2,q)} = [q/c^2, 1].$$

We now write simply $\mathcal{C} = \mathcal{C}(\Gamma')$ for the set of cusps of Γ', which we enumerate as $\mathcal{C} = \{\mathfrak{a}_j\}_j$. Write c_j for the denominator of $\mathfrak{a}_j$, and $w_j = [q/c_j^2, 1]$ for its width. For each positive divisor c of q, let

$$\mathcal{C}[c] := \{\mathfrak{a}_j \in \mathcal{C} : c_j = c\}$$

denote the set of cusps of denominator c. It follows from the above diagram that $\#\mathcal{C}[c] = \varphi((c,q/c))$.

Choose an element $\tau_j \in \Gamma$ representing the double coset $\mathfrak{a}_j \in \Gamma'\backslash\Gamma/\Gamma_\infty$. If $\mathfrak{a}_j = \mathfrak{a}_{d/c}$, then we may take $\tau_j = \left(\begin{smallmatrix} * & * \\ c & d' \end{smallmatrix}\right)$ for any integer d' for which $(d',c) = 1$ and $d' \equiv d \pmod{(c,q/c)}$. The τ_j so-obtained form a set of representatives for $\Gamma'\backslash\Gamma/\Gamma_\infty$. Intrinsically, the width of $\mathfrak{a}_j$ is given by $w_j = [\Gamma_\infty : \Gamma_\infty \cap \tau_j^{-1}\Gamma'\tau_j]$. The *scaling matrix* of $\mathfrak{a}_j$ is

$$\sigma_j = \tau_j \begin{bmatrix} w_j & \\ & 1 \end{bmatrix} \tag{46}$$

[19] We say that the residue class $c \in (\mathbb{Z}/q)$ divides q if its unique representative $c' \in [1,q]$ divides q.

[20] Note that $d \mapsto (d,c,q/c)$ is a well-defined function on $\mathbb{Z}/(q/c)$.

[21] For more general subgroups than $\Gamma_0(q)$, one should replace "Γ'-stabilizer" with "$\Gamma' \cdot \{\pm 1\}$-stabilizer".

which has the property $B \cap \sigma_j^{-1}\Gamma'\sigma_j = \Gamma_\infty$ with $B = \{(\begin{smallmatrix} * & * \\ & * \end{smallmatrix})\} < G$. To put it another way, for each $z \in \mathbb{H}$, let us write $z_j = x_j + iy_j$ for the change of variable $z_j := \sigma_j^{-1}z$ and $\Gamma'_j = \mathrm{Stab}_{\Gamma'}(\mathfrak{a}_j)$. Then each element $\gamma \in \Gamma'$ satisfying $(\gamma z)_j = z_j + 1$ generates Γ'_j. In other words, $z \mapsto z_j$ is a proper isometry of $\mathbb{H}$ under which $z_j \mapsto z_j + 1$ corresponds to the action on z by a generator for Γ'_j.

3.4.2. *Fourier expansions.* We now turn to explicating the Fourier expansion of $|f|^2$ at the cusp $\mathfrak{a}_j$, or equivalently that of $|f|^2(z)$ regarded as a function of the variable z_j. Recall the weight k slash operation: for $\alpha \in \mathrm{GL}_2^+(\mathbb{R})$, set $f|_k\alpha(z) = \det(\alpha)^{k/2}j(\alpha,z)^{-k}f(\alpha z)$, where $j\left(\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right], z\right) = cz + d$. We then have $|f|^2(z)y^k = |f|^2(\sigma_j z_j)\mathrm{Im}(\sigma_j z_j)^k = |f|_k\sigma_j|^2\,(z_j)y_j^k$, and may write

$$f|_k\sigma_j(z_j) = y_j^{-k/2}\sum_{n\in\mathbb{N}}\frac{\lambda_j(n)}{\sqrt{n}}\kappa(ny_j)e(nx_j) \tag{47}$$

for $\kappa(y) = y^{k/2}e^{-2\pi y}$ $(y \in \mathbb{R}_+^\times)$ and some coefficients $\lambda_j(n) \in \mathbb{C}$. In the special case, $\mathfrak{a}_j = \infty$, we note that $\lambda_j(n) = \lambda(n)$. In general, the notation $\lambda_j(n)$ is slightly misleading because $\lambda_j(n)$ depends not only on the cusp $\mathfrak{a}_j$, but also on the choice of scaling matrix τ_j. However, if $\lambda'_j(n)$ denotes the coefficient obtained by a different choice τ'_j, then one has $\lambda'_j(n) = e(bn/w_j)\lambda_j(n)$ for some integer b.

The coefficients $\lambda_j(n)$ seem easiest to describe by working adelically. For background on adeles and adelization of automorphic forms, we refer the reader to [10]. We recall the following notation from Section 2:

$$w = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad a(y) = \begin{bmatrix} y & \\ & 1 \end{bmatrix}, \quad n(x) = \begin{bmatrix} 1 & x \\ & 1 \end{bmatrix}, \quad \text{and } z(y) = \begin{bmatrix} y & \\ & y \end{bmatrix}.$$

Let $\hat{\mathbb{Z}} = \varprojlim \mathbb{Z}/n = \prod \mathbb{Z}_p$, $\hat{\mathbb{Q}} = \hat{\mathbb{Z}} \otimes_{\mathbb{Z}} \mathbb{Q} = \prod' \mathbb{Q}_p$ and $\mathbb{A} = \mathbb{R} \times \hat{\mathbb{Q}}$. To f one attaches a function $F : \mathrm{GL}_2(\mathbb{A}) \to \mathbb{C}$ in the following standard way. By strong approximation, every element of $\mathrm{GL}_2(\mathbb{A})$ may be expressed in the form $\gamma g_\infty \kappa_0$ for some $\gamma \in \mathrm{GL}_2(\mathbb{Q})$, $g_\infty \in \mathrm{GL}_2(\mathbb{R})^+$ and $\kappa_0 \in K_0(q) = \{\left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right] \in \mathrm{GL}_2(\hat{\mathbb{Z}}) : q \mid c\}$. Then $F(\gamma g_\infty \kappa_0) = f|_k g_\infty(i)$. Recall that $\sigma_j \in \mathrm{GL}_2(\mathbb{Q})^+$. Let $i_\infty : \mathrm{GL}_2(\mathbb{Q}) \hookrightarrow \mathrm{GL}_2(\mathbb{R}) \hookrightarrow \mathrm{GL}_2(\mathbb{A})$ and $i_{\mathrm{fin}} : \mathrm{GL}_2(\mathbb{Q}) \hookrightarrow \mathrm{GL}_2(\hat{\mathbb{Q}}) \hookrightarrow \mathrm{GL}_2(\mathbb{A})$ be the natural inclusions. If $g_z \in \mathrm{GL}_2(\mathbb{R})^+$ is chosen so that $g_z i = z$, then $f|_k\sigma_j(z) = f|_k\sigma_j g_z(i) = F(\iota_\infty(\sigma_j)g_z) = F(g_z\iota_{\mathrm{fin}}(\sigma_j^{-1}))$ by the left-$G(\mathbb{Q})$-invariance of F. For $g \in \mathrm{GL}_2(\mathbb{A})$, one has a Fourier expansion

$$F(g) = \sum_{n\in\mathbb{Q}_{\neq 0}} W(a(n)g),$$

where W is a global Whittaker newform corresponding to f; it is given explicitly by $W(g) = \int_{x\in\mathbb{A}/\mathbb{Q}} F(n(x)g)\psi(-x)\,dx$ where the integral is taken with respect to an invariant probability measure. It satisfies $W(n(x)g) = \psi(x)W(g)$ for all $x \in \mathbb{A}$, where $0 \neq \psi = \prod \psi_v \in \mathrm{Hom}(\mathbb{A}/\mathbb{Q}, \mathbb{C}^1)$ is the additive character for which $\psi_\infty(x) = e^{2\pi ix}$. The function W factors as $\prod W_v$ over the places of $\mathbb{Q}$. We may pin down this factorization uniquely by requiring that $W_\infty(a(y)) = \kappa(y)$ and $W_p(1) = 1$ for all primes p. Writing $z = x + iy$, we may and shall assume that $g_z = n(x)a(y)$. Then

$$y^{k/2}f|_k\sigma_j(z) = F(g_z\iota_{\mathrm{fin}}(\sigma_j^{-1})) = \sum_{n\in\mathbb{Q}_{\neq 0}}\kappa(ny)e(nx)\prod_p W_p(a(n)\sigma_j^{-1}).$$

Here we identify σ_j with its image under the natural inclusion $G(\mathbb{Q}) \hookrightarrow G(\mathbb{Q}_p)$. If $p \nmid q$, then W_p is unramified at p and $\sigma_j \in \mathrm{GL}_2(\mathbb{Z}_p)$, since σ_j differs from $\tau_j \in \mathrm{SL}_2(\mathbb{Z})$ by a diagonal matrix with integral entries dividing q (and hence with determinant coprime to p); thus $W_p(a(n)\sigma_j^{-1}) = W_p(a(n))$. If we also have $p \nmid n$, then $a(n) \in \mathrm{GL}_2(\mathbb{Z}_p)$, and so $W_p(a(n)) = 1$. Therefore the expansion (47) holds with

$$\lambda_j(n) = \sqrt{n} \prod_{p \mid [n,q]} W_p(a(n)\sigma_j^{-1}) = \sqrt{n} \prod_{p \mid q} W_p(a(n)\sigma_j^{-1}) \prod_{p \mid \frac{n}{(n,q^\infty)}} W_p(a(n)). \tag{48}$$

Let us spell out (48) a bit more precisely. Write $\tau_j = \left[\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right]$, so that $\mathfrak{a} = \mathfrak{a}_{d/c}$ in the notation introduced above. The Bruhat decomposition of τ_j^{-1} reads

$$\tau_j^{-1} = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} -c & \\ & -c \end{bmatrix} n(-d/c)a(1/c^2)wn(-a/c),$$

so that for $y \in \mathbb{Q}_p^\times$, we have

$$\begin{aligned} W_p(a(y)\sigma_j^{-1}) &= W_p(a(y/[q/c^2,1])n(-d/c)a(1/c^2)wn(-a/c)) \\ &= W_p(n(-yd/[q/c,c])a(y/[q,c^2])wn(-a/c)) \\ &= \psi_p\left(\frac{-dy}{[q/c,c]}\right) W_p(a(y/[q,c^2])wn(-a/c)). \end{aligned}$$

Note also that

$$\prod_{p \mid \frac{n}{(n,q^\infty)}} W_p(a(n)) = \prod_{p \mid \frac{n}{(n,q^\infty)}} W_p\left(a\left(\frac{n}{(n,q^\infty)}\right)\right) = \lambda\left(\frac{n}{(n,q^\infty)}\right).$$

Recall here that $\lambda(m)$ is our notation for the coefficient $\lambda_j(m)$ at the distinguished cusp $\mathfrak{a}_j = \infty$. From the above calculations, we deduce that

$$\lambda_j(n) = \sqrt{n} \cdot e\left(\frac{dn}{[q/c,c]}\right) \lambda\left(\frac{n}{(n,q^\infty)}\right) \prod_{p \mid q} W_p(a(n/[q,c^2])wn(-a/c)). \tag{49}$$

One can check that λ_j is not multiplicative in general; for example, it can happen that $\lambda_j(1) \neq 1$, or even that $\lambda_j(mn)\lambda_j(1) \neq \lambda_j(m)\lambda_j(n)$ for pairs of coprime integers m, n. To circumvent this lack of multiplicativity, we work with the root-mean-square of λ_j taken over all cusps of a given denominator. For each positive divisor c of q, define

$$\lambda_{[c]}(n) = \left(\frac{1}{\#\mathcal{C}[c]} \sum_{\mathfrak{a}_j \in \mathcal{C}[c]} |\lambda_j(n)|^2\right)^{1/2}. \tag{50}$$

An explicit formula in terms of GL(2) Gauss sums for the RHS of (49), and hence for $\lambda_j(n)$, may be derived following the method of Section 2.5. For our purposes, it suffices (by Cauchy–Schwarz; see Section 3.5) to evaluate the simpler averages $\lambda_{[c]}(n)$. It turns out that these averages are multiplicative in a certain non-conventional sense:

Definition 3.6. Let us call an arithmetic function $f : \mathbb{N} \to \mathbb{C}$ *factorizable* if it can be written as a product $f = \prod_p f_p$ over the primes, where the $f_p : \mathbb{N} \to \mathbb{C}$ satisfy

(1) $f_p(n) = f_p(n_p)$ for all $n \in \mathbb{N}$ and all p,[22] and

[22] Recall that n_p denote the largest divisor of n that is a power of p.

(2) $f_p(1) = 1$ for all but finitely many p.

Remark 3.7. Every multiplicative[23] function is factorizable (take $f_p(n) = f(n_p)$), and every factorizable function f satisfies

$$f(mn)f(1) = f(m)f(n) \text{ whenever } (m,n) = 1, \tag{51}$$

but neither of these implications is reversible. A factorizable function f is multiplicative if and only if $f(1) = 1$. Many non-factorizable functions f satisfy (51), but if $f(1) \neq 0$, then f is factorizable if and only if it satisfies (51), in which case $n \mapsto f(n)/f(1)$ is multiplicative.

Lemma 3.8. *Let c be a positive divisor of q. The function $n \mapsto \lambda_{[c]}(n)$ is factorizable:*

$$\lambda_{[c]}(n) = \prod_p \lambda_{[c],p}(n)$$

for each $n \in \mathbb{N}$, where $\lambda_{[c],p} : \mathbb{N} \to \mathbb{R}_{\geq 0}$ is defined by

$$\lambda_{[c],p}(n) = \begin{cases} |\lambda(n_p)| = |n|_p^{-1/2} |W_p|(a(n)) & p \nmid q, \\ |n|_p^{-1/2} \left(\int_{u \in \mathbb{Z}_p^\times} |W_p|^2 \left(a\left(\frac{un}{[q,c^2]}\right) wn(1/c) \right) d^\times u \right)^{1/2} & p \mid q. \end{cases} \tag{52}$$

Proof. For each $\mathfrak{a}_j \in \mathcal{C}[c]$, let us write $\tau_j = \left(\begin{smallmatrix} a_j & * \\ c & * \end{smallmatrix}\right)$. Recall that as $\mathfrak{a}_j$ traverses $\mathcal{C}[c]$, the lower-right entry of τ_j traverses $(\mathbb{Z}/(c,q/c))^\times$, hence so does the upper-left entry a_j. The formula (49) and the definition (50) imply that

$$\lambda_{[c]}(n) = n^{1/2} \left| \lambda\left(\frac{n}{(n,q^\infty)}\right) \right| \left(\frac{1}{\#\mathcal{C}[c]} \sum_{\mathfrak{a}_j \in \mathcal{C}[c]} \prod_{p \mid q} |W_p|^2 (a(n/[q,c^2]) wn(-a_j/c)) \right)^{1/2}. \tag{53}$$

We treat the three factors on the RHS successively; in doing so, we shall repeatedly invoke the right-$a(\mathbb{Z}_p^\times)$-invariance of W_p for each prime p. The first factor may be written $n^{1/2} = \prod_p |n|_p^{-1/2}$. The second is $\left|\lambda\left(\frac{n}{(n,q^\infty)}\right)\right| = \prod_{p \nmid q} |W_p|(a(n))$. For the third, note that the average over $\mathcal{C}[c]$, or equivalently, over $a_j \in (\mathbb{Z}/(c,q/c))^\times$, lifts to an Eulerian integral over $\prod_{p \mid q} \mathbb{Z}_p^\times$:

$$\frac{1}{\#\mathcal{C}[c]} \sum_{\mathfrak{a}_j \in \mathcal{C}[c]} \prod_{p \mid q} |W_p|^2 (a(n/[q,c^2]) wn(-a_j/c)) = \prod_{p \mid q} \int_{u \in \mathbb{Z}_p^\times} |W_p|^2 (a(n/[q,c^2]) wn(-u/c))\, d^\times u.$$

The identity $wn(-u/c) \equiv a(-1/u) wn(1/c) \pmod{Z(\mathbb{Q}_p)}$ and substitution $u \mapsto -1/u$ allows us to rewrite the above as

$$\prod_{p \mid q} \int_{u \in \mathbb{Z}_p^\times} |W_p|^2 (a\left(\frac{un}{[q,c^2]}\right) wn(1/c))\, d^\times u.$$

Collecting the identities obtained for each of the three factors in (53), we deduce

$$\lambda_{[c]}(n) = \prod_p \left(|n|_p^{-1/2} \times \begin{cases} |W_p|(a(n)) & p \nmid q \\ \left(\int_{u \in \mathbb{Z}_p^\times} |W_p|^2 (a\left(\frac{un}{[q,c^2]}\right) wn(1/c))\, d^\times u \right)^{1/2} & p \mid q \end{cases} \right).$$

This establishes the claimed formula $\lambda_{[c]}(n) = \prod_p \lambda_{[c],p}(n)$. It is clear from the definition that $\lambda_{[c],p}(n) = \lambda_{[c],p}(n_p)$ for all p and that $\lambda_{[c],p}(1) = 1$ for all p not dividing q. ∎

[23]Recall that an arithmetic function f is multiplicative if $f(mn) = f(m)f(n)$ whenever $(m,n) = 1$.

Remark 3.9. It follows from the right-$a(\mathbb{Z}_p^\times)$-invariance of W that $\lambda_{[c],p}(n) = \lambda_{[c_p],p}(n_p)$.

Lemma 3.10. *For each prime p, each $c|q$ and each $n \in \mathbb{N}$, we have $\lambda_{[c],p}(n) = \lambda_{[q/c],p}(n)$.*

Proof. Let $w_q = \left[\begin{smallmatrix} 0 & 1 \\ -q & 0 \end{smallmatrix}\right]$. Then w_q acts as the Atkin–Lehner operator on the newvector W_p, and so $W_p(gw_q) = \pm W_p(g)$ for all $g \in \mathrm{GL}_2(\mathbb{Q}_p)$. Since

$$a\left(\frac{y}{[q, c^2]}\right) wn(1/c)w_q = z\left(\frac{q}{c}\right) n(-u)a\left(\frac{-y}{[q,(q/c)^2]}\right) wn(1/(q/c))a(-1)$$

for each $y \in \mathbb{Q}_p^\times$, the lemma follows from the left-$Z(\mathbb{Q}_p)N(\mathbb{Q}_p)$-equivariance and right-$A(\mathbb{Z}_p)$-invariance of W_p. □

Remark 3.11. When q is a prime power, the classical content of the proof of the above lemma is that for $ad \equiv 1 \ (q)$, the operator $z \mapsto -1/(qz)$ takes a/c to $-a^{-1}/(qc^{-1}) \equiv -d/(qc^{-1}) \pmod{\mathbb{Z}}$.

We are now in a position to compute $\lambda_{[c],p}$ exactly. We do this in Proposition 3.12. The quantities $R_{m,p}$ that appear in the statement below are the coefficients "R_m" that were defined in (40) and later computed exactly[24] for all representations of $\mathrm{PGL}_2(\mathbb{Q}_p)$ with conductor at least p^2.

Proposition 3.12. *Let c be a positive divisor of q, p a prime divisor of q, and n a natural number. Write $n = up^k$ with $(u, p) = 1$ and $k \geq 0$.*

1. $\lambda_{[c],p}(n) = \lambda_{[c],p}(p^k)$.
2. *If p^2 does not divide q, then $\lambda_{[c],p}(p^k) = p^{-k/2}$.*
3. *If p^2 divides q and $c_p^2 \neq q_p$, then $\lambda_{[c],p}(p^k) = 1$ if $k = 0$ and vanishes otherwise.*
4. *If p^2 divides q and $c_p^2 = q_p$, then*

$$\lambda_{[c],p}(p^k)^2 = \begin{cases} \left(\frac{1+p^{-1}}{1-p^{-1}}\right) q_p^{\frac{1}{2}} R_{-k,p} & \text{if } k > 0 \\ \left(\frac{1+p^{-1}}{1-p^{-1}}\right)\left(q_p^{\frac{1}{2}} R_{0,p} - \frac{1}{p+1}\right) & \text{if } k = 0. \end{cases}$$

By the formulas for $R_{-k,p}$ from Section 2, we deduce immediately:

Corollary 3.13. *For each prime p for which p^2 divides q, each positive divisor c of q, and each nonnegative integer k, we have*

$$\lambda_{[c],p}(p^k) \ll p^{k/4}$$

with an absolute implied constant.

Remark 3.14. In general, one cannot hope to improve upon the above inequality in the range $0 \leq k \leq n - N$ where the integer N is such that $p^N = C_p$, the p-part of the conductor of $f \times f$. This is clear from the formulas for R_m from Section 2. In particular, the "Deligne bound" $|\lambda_j(p^k)| \leq \tau(p^k)$ does not hold in general.

Proof of Proposition 3.12. Part (1) follows immediately from the definition of $\lambda_{[c],p}$. Part (2) follows from standard formulas for the local Whittaker function attached to a Steinberg representation (see [15, Lemma 2.1]).

We now turn to (3) and (4). *To simplify notation, we restrict henceforth to the case that q and c are powers of p*; the general case then follows by the observation of Remark 3.9. The proofs of (3)

[24]We wrote down exact formulas only for T_m but similar ones for R_m can be easily worked out using Table 1.

and (4) will each make use of the following consequence of the support condition on W_p established in Lemma 2.15 and the $\mathrm{GL}_2(\mathbb{Q}_p)$-invariance of the Whittaker inner product: for each $v \in \mathbb{Z}$ and $x \in \mathbb{Q}_p$ with $|x|^2 < q$, we have

$$\int_{u \in \mathbb{Z}_p^\times} |W_p|^2 \left(a(up^v/q)wn(x)\right) \, d^\times u = \delta_v := \begin{cases} 1 & v = 0, \\ 0 & v \neq 0. \end{cases} \tag{54}$$

For part (3), suppose that $p^2 | q$ and $c^2 \neq q$. By the "functional equation" $\lambda_{[c],q}(p^k) = \lambda_{[q/c],p}(p^k)$ of Lemma 3.10, we may assume without loss of generality that c^2 (properly) divides q. Then $[q, c^2] = q$, so

$$\frac{\lambda_{[c],p}(p^k)^2}{p^k} = \int_{u \in \mathbb{Z}_p^\times} |W_p|^2 \left(a(up^k/q)wn(1/c)\right) d^\times u.$$

Since $|(1/c)|_p^2 = c^2 < q$, the identity (54) implies $\lambda_{[c],p}(p^k)^2 = p^k \delta_k = \delta_k$, as desired.

It remains to consider part (4), in which $c^2 = q$. By definition (see (40)), $R_{-k,p}$ is the coefficient of p^{-ks} in $J_p(s)$ (see (30)). By writing the p-adic integral in (30) as a sum and invoking the right invariance of W_p, we obtain

$$\begin{aligned} \frac{\zeta_p(1)}{\zeta_p(2)} R_{-k,p} &= \int_{\substack{x \in \mathbb{Q}_p \\ |y|/\max(1,|x|)^2 = p^{-k}}} \int_{y \in \mathbb{Q}_p^\times} |y|^{-1} |W_p|^2 \left(a(y)wn(x)\right) \, dx \, d^\times y \\ &= \sum_{t=0}^{\infty} v_t p^{k-2t} \int_{u \in \mathbb{Z}_p^\times} |W_p|^2 \left(a(p^{k-2t})wn(up^{-t})\right) d^\times u, \end{aligned}$$

where $v_0 = \mathrm{vol}(\mathbb{Z}_p, dx) = 1$ and $v_t = \mathrm{vol}(p^{-t}\mathbb{Z}_p^\times, dx) = p^t(1 - p^{-1})$ for $t \geq 1$; the measures here are normalized as in Section 2.1. Set $q = p^n$. By the right-$a(\mathbb{Z}_p)^\times$-invariance of W_p, the inner integral may be written as

$$\int_{u \in \mathbb{Z}_p^\times} |W_p|^2 \left(a(up^{k-2t+n}/q)wn(p^{-t})\right) d^\times u. \tag{55}$$

We consider separately several cases:

- If $t < n/2$, then $|p^{-t}|_p^2 < q$ and $k - 2t + n > 0$, so (54) implies that (55) vanishes.
- If $t > n$, then the identity $wn(x) \equiv n(-x^{-1})a(x^{-2})n_-(x^{-1}) \pmod{Z(\mathbb{Q}_p)}$, where $n_-(x^{-1}) = \left(\begin{smallmatrix} 1 & \\ x^{-1} & 1 \end{smallmatrix}\right)$, shows that
$$\begin{aligned} W_p(a(up^{k-2t+n}/q)wn(p^{-t})) &= W_p(n(-x^{-1}up^{k+n}/q)a(up^{k+n}/q)n_-(p^t)) \\ &= W_p(a(up^{k+n}/q)) = \delta_k. \end{aligned}$$
Thus the integral (55) is δ_k.
- For $n/2 \leq t \leq n$, the definition (52) specializes to
$$\lambda_{[p^t],p}(p^k)^2 = p^k \int_{u \in \mathbb{Z}_p^\times} |W_p|^2 \left(a(up^{k-2t})wn(p^{-t})\right) d^\times u.$$
This shows that (55) equals $p^{-k}\lambda_{[p^t],p}(p^k)^2$. If the lower inequality is strict, i.e., if $t > n/2$, then the proof given above of part (3) of the present proposition shows moreover that $\lambda_{[p^t],p}(p^k)^2 = \delta_k$.

- Combining the previous two cases, we see for $t > n/2$ that (55) equals δ_k.

Collecting together the above calculations, we deduce that

$$\frac{\zeta_p(1)}{\zeta_p(2)} R_{-k,p} = v_{n/2} p^{-n} \lambda_{[p^{n/2}],p}(p^k)^2 + \delta_k \sum_{t > n/2} v_t p^{-2t}.$$

Rearranging, recalling that that $v_t = \frac{p^t}{\zeta_p(1)}$ (for $t \geq 1$), and summing some geometric series, we arrive at

$$\begin{aligned} \lambda_{[p^{n/2}],p}(p^k)^2 &= \frac{p^n}{v_{n/2}} \frac{\zeta_p(1)}{\zeta_p(2)} R_{-k,p} - \frac{p^n}{v_{n/2}} \delta_k \sum_{t > n/2} v_t p^{-2t} \\ &= p^{n/2} \frac{\zeta_p(1)^2}{\zeta_p(2)} R_{-k,p} - \delta_k p^{n/2} \sum_{t > n/2} p^{-t} \\ &= p^{n/2} \frac{1 + p^{-1}}{1 - p^{-1}} R_{-k,p} - \delta_k \frac{p^{-1}}{1 - p^{-1}}, \end{aligned}$$

which is equivalent to the claimed formula. □

Remark 3.15. It is instructive to apply Proposition 3.12 when f is associated to an elliptic curve $E_{/\mathbb{Q}}$ of conductor q. In that case, we have $k = 2$ and $\lambda(n)\sqrt{n} \in \mathbb{Z}$. Since $\mathrm{Aut}(\mathbb{C})$ acts transitively on the set of cusps of given denominator, Proposition 3.12 provides a characterization of the cusps at which the differential form $f(z)dz$ vanishes, complementing some recent work of Brunault [2]. With further work, one may derive from Proposition 3.12 an exact formula for the ramification index at a given cusp of the modular parametrization $X_0(q) \to E$. The resulting formula turns out to depend only on the reduction modulo certain powers of 2 and 3 of the coefficients of the minimal Weierstrass equation for E.

Remark 3.16. One may extend $\lambda_{[c],p}$ to a function on $\mathbb{Q}_p^\times$ via the formula in its original definition (50), and then $\lambda_{[c]} : \mathbb{N} \to \mathbb{R}_{\geq 0}$ to a function $\lambda_{[c]} : \hat{\mathbb{Q}}^\times \to \mathbb{R}_{\geq 0}$ via $(y_p)_p \mapsto \prod \lambda_{[c],p}(y_p)$. Then by directly evaluating $J_p(s)$ in the Iwasawa decomposition, one obtains

$$J_f(s) = \prod_{p|q} J_p(s) = \frac{1}{[\Gamma : \Gamma']} \int_{y \in \prod_{p|q} \mathbb{Q}_p^\times} |y|_{\mathbb{A}}^s \sum_{c|q} [q/c^2, 1]^s \varphi((q/c, c)) \lambda_{[c]}(y)^2 \, d^\times y.$$

Suppose now that $q = p^{2m}$ is a prime power with even exponent. Then the support condition (by Proposition 3.12) that $\lambda_{[c],p}(p^k) = 0$ unless $k = 0$ or $c = p^m$ implies

$$J_f(s) = \frac{p^{2m(s-1)}}{1 + 1/p} \sum_{0 \leq t \leq m-1} \frac{\varphi(p^t)}{p^{2ts}} + \frac{p^{-m-1}}{1 + 1/p} + p^{-m} \frac{1 - 1/p}{1 + 1/p} \sum_{k \geq 0} \frac{\lambda_{[p^m],p}(p^k)^2}{p^{ks}}.$$

Thus the "local Lindelöf bound" in the form $J_f(s) \ll mp^{-m}$ $(\mathrm{Re}(s) = 1/2)$ is "equivalent" to the estimate $\sum_{k \geq 0} \lambda_{[p^m],p}(p^k)^2 / p^{k/2} \ll m$ for the sum of the mean squares of the Fourier coefficients of f at the cusps of $\Gamma_0(p^{2m})$ with denominator p^m. When the representation π of $\mathrm{PGL}_2(\mathbb{Q}_p)$ generated by f is supercuspidal, we note that the identity (22) implies the cute formula

$$\frac{\sum_{C(\mu)^2=C(\pi)} (C(\pi\mu)/C(\pi))^{s-1}}{\sum_{C(\mu)^2=C(\pi)} 1} = \sum_{k\geq 0} \frac{\lambda_{[p^m],p}(p^k)^2}{p^{ks}}$$

for the "moments" of $\{\mu : C(\mu)^2 = C(\pi)\} \ni \mu \mapsto C(\pi\mu)$ (see Section 1.8 for notation).

3.5. **Proof of Theorem 1.2, modulo technicalities.** In this section we follow Holowinsky [17] in bounding $D_f(\phi)$ in terms of shifted convolution sums, to which we apply an extension (Proposition 3.17) of a refinement [33, Thm 3.10] of his bounds for such sums [17, Thm 2]. By combining with the bounds obtained in Section 3.3 and Section 3.4, we deduce Theorem 1.2.

Let $Y \geq 1$ be a parameter (to be chosen later), and let $h \in C_c^\infty(\mathbb{R}_+^\times)$ be an everywhere nonnegative test function with Mellin transform $h^\wedge(s) = \int_0^\infty h(y)y^{-s-1}dy$ such that $h^\wedge(1) = \mu(1)$. The proof of [33, Lem 3.4] shows without modification that

$$Y\mu_f(\phi) = \sum_{\mathfrak{a}_j\in\mathcal{C}} \int_{y_j=0}^\infty h(Yw_jy_j) \int_{x_j=0}^1 \phi(w_jz_j)|f|^2(z)y^k \,\frac{dx_j\,dy_j}{y_j^2} + O_\phi(Y^{1/2}\mu_f(1)). \tag{56}$$

Let

$$I_\phi(l,n,x) = (mn)^{-1/2}\int_{y=0}^\infty h(xy)\kappa_\phi(ly)\kappa_f(my)\kappa_f(ny)\frac{dy}{y^2}, \quad m := n+l,$$

where κ_ϕ and κ_f are as in Section 3.1. Write $w_c := [q/c^2, 1]$ for all $c|q$. By inserting Fourier expansions and applying some trivial bounds as in [33, Lem 3.8], we obtain

(57)

$$\begin{aligned} D_f(\phi) &= \frac{1}{Y\mu_f(1)} \sum_{\substack{l\in\mathbb{Z}_{\neq 0}\\ |l|<Y^{1+\varepsilon}}} \frac{\lambda_\phi(l)}{\sqrt{|l|}} \sum_j \left(\sum_{\substack{n\in\mathbb{N}\\ m:=n+w_jl\in\mathbb{N}}} \lambda_j(m)\lambda_j(n) I_\phi(w_jl,n,Yw_j)\right) + O_{\phi,\varepsilon}(Y^{-1/2}), \\ &= \frac{1}{Y\mu_f(1)} \sum_{\substack{l\in\mathbb{Z}_{\neq 0}\\ |l|<Y^{1+\varepsilon}}} \frac{\lambda_\phi(l)}{\sqrt{|l|}} \sum_{c|q} I_\phi(w_cl,n,Yw_c) \sum_{\substack{n\in\mathbb{N}\\ m:=n+w_cl\in\mathbb{N}}} \left(\sum_{\mathfrak{a}_j\in\mathcal{C}[c]} \lambda_j(m)\lambda_j(n)\right) + O_{\phi,\varepsilon}(Y^{-1/2}). \end{aligned}$$

By Cauchy–Schwarz, we deduce that

$$\begin{aligned} |D_f(\phi)| \leq \frac{1}{Y\mu_f(1)} &\sum_{\substack{l\in\mathbb{Z}_{\neq 0}\\ |l|<Y^{1+\varepsilon}}} \frac{|\lambda_\phi(l)|}{\sqrt{|l|}} \sum_{c|q} \#\mathcal{C}[c]\,|I_\phi(w_cl,n,Yw_c)| \sum_{\substack{n\in\mathbb{N}\\ m:=n+w_cl\in\mathbb{N}}} \lambda_{[c]}(m)\lambda_{[c]}(n) \\ &+ O_{\phi,\varepsilon}(Y^{-1/2}). \end{aligned} \tag{58}$$

The weight $I_\phi(w_cl,n,Yw_c)$ essentially restricts the sum to $\max(m,n) \ll Ykw_c$: indeed, [33, Lemma 3.12] asserts (in slightly different notation) that

$$I_\phi(l,n,x) \ll_A \frac{\Gamma(k-1)}{(4\pi)^{k-1}} \cdot \max\left(1, \frac{\max(m,n)}{xk}\right)^{-A}$$

for every $A > 0$.

In Section 3.7, we prove the following:

Proposition 3.17. *For $l \in \mathbb{Z}_{\neq 0}$, $x \in \mathbb{R}_{\geq 1}$, $\varepsilon \in (0,1)$ and each positive divisor c of q, we have*

$$\sum_{\substack{n \in \mathbb{N} \\ m := n + l \in \mathbb{N} \\ \max(m,n) \leq x}} |\lambda_{[c]}(m) \lambda_{[c]}(n)| \ll_\varepsilon q_\diamond^\varepsilon \log\log(e^e q)^{O(1)} \frac{x \prod_{p \leq x}(1 + 2|\lambda_f(p)|/p)}{\log(ex)^{2-\varepsilon}}.$$

Inserting this bound into (58), summing dyadically (or by parts) as in [33, Proof of Cor 3.14], applying the Rankin–Selberg bound for $\lambda_\phi(l)$ as in [33, Lem 3.17], invoking the Rankin–Selberg formula

$$\mu_f(1) \asymp qk \frac{\Gamma(k-1)}{(4\pi)^{k-1}} L(\mathrm{ad} f, 1)$$

for $\mu_f(1)$, and pulling it all together as in [33, Section 3.3], we obtain

$$D_f(\phi) \ll_{\phi,\varepsilon} Y^{-1/2} + \frac{Y^{1/2+\varepsilon} \log(qk)^\varepsilon q_\diamond^\varepsilon}{q} \sum_{c | q} \frac{[q/c^2, 1]\ \varphi((c, q/c))}{\log([q/c^2, 1] k Y)^{2-\varepsilon}} \prod_{p \leq [q/c^2, 1] k Y} \left(1 + \frac{2\,|\lambda_f(p)|}{p}\right). \tag{59}$$

To control the sum over c in (59), we apply the following lemma, whose (technical) proof we defer to Section 3.7:

Lemma 3.18. *Let $x \geq 2$, $\varepsilon \in (0,1)$, and $q \in \mathbb{N}$. Then*

$$\sum_{c | q} \frac{[q/c^2, 1]\ \varphi((c, q/c))}{\log([q/c^2, 1] x)^{2-\varepsilon}} \ll \frac{q \log\log(e^e q)^{O(1)}}{\log(qx)^{2-\varepsilon}}.$$

with absolute implied constants.

Applying this lemma to (59) gives

$$D_f(\phi) \ll_{\phi,\varepsilon} Y^{-1/2} + \frac{Y^{1/2+\varepsilon} q_\diamond^\varepsilon}{\log(qk)^{2-\varepsilon}} \prod_{p \leq qkY} \left(1 + \frac{2\,|\lambda_f(p)|}{p}\right). \tag{60}$$

The partial product over $qk < p \leq qkY$ contributes negligibly, so choosing Y suitably as in [17] yields

$$D_f(\phi) \ll_{\phi,\varepsilon} \log(qk)^\varepsilon q_\diamond^\varepsilon M_f(qk)^{1/2}, \tag{61}$$

where

$$M_f(x) = \frac{\prod_{p \leq x}(1 + 2|\lambda_f(p)|/p)}{\log(ex)^2 L(\mathrm{ad} f, 1)}.$$

Feeding (45) and (61) into the recipe of [33, Section 5] gives the following result.

Theorem 3.19. *Fix a Maass cusp form or incomplete Eisenstein series ϕ on $Y_0(1)$. Then for a holomorphic newform f of weight $k \in 2\mathbb{N}$ on $\Gamma_0(q)$, $q \in \mathbb{N}$, we have*

$$D_f(\phi) \ll_{\phi,\varepsilon} \log(qk)^\varepsilon \min\left\{ \frac{(q/\sqrt{C})^{-1+2\theta+\varepsilon}}{\log(kC)^\delta L(\mathrm{ad} f, 1)}, q_\diamond^\varepsilon \log(qk)^{1/12} L(\mathrm{ad} f, 1)^{1/4} \right\}.$$

Here $\varepsilon > 0$ is arbitrary, $\mathrm{ad} f$ is the adjoint lift of f, C is the (finite) conductor of $\mathrm{ad} f$, $\theta \in [0, 7/64]$ is a bound towards the Ramanujan conjecture for ϕ at primes dividing q (take $\theta = 0$ if ϕ is incomplete Eisenstein), and $\delta = 1/2$ or 1 according as ϕ is cuspidal or incomplete Eisenstein.

When q is squarefree, one has $q/\sqrt{C} = 1$, and Theorem 3.19 recovers a statement appearing on the final page of [33] from which the main result of that paper, the squarefree case of Theorem 1.1, is deduced in a straightforward manner. In general, Proposition 2.5 implies that C is a square integer satisfying $C \leq q q_0$, where q_0 is the largest squarefree divisor of q. From this one deduces Theorem 1.2 by considering separately the cases that $L(\mathrm{ad} f, 1)$ is large and small, as in [18, Section 3].

3.6. **Proof that Theorem 1.2 implies Theorem 1.1.** We explain briefly how Theorem 1.1 follows from Theorem 1.2. It's known that the class $C_c(Y_0(1))$ of *compactly supported* continuous functions on $Y_0(1)$ is contained in the uniform span of the Maass eigencuspforms and incomplete Eisenstein series (see [22]). Fix a bounded continuous function ϕ on $Y_0(1)$. Let $\varepsilon > 0$ be arbitrary. Choose $T = T(\varepsilon)$ large enough that the ball $B_T := \{z \in \mathbb{H} : -1/2 \leq \mathrm{Re}(z) \leq 1/2,\ \mathrm{Im}(z) > T\}$ has normalized volume $\mu(B_T)/\mu(1) < \varepsilon$. Write $\phi = \phi_1 + \phi_2$, where $\phi_1 \in C_c(Y_0(1))$ and ϕ_2 is supported on B_T. Because ϕ_1 can be uniformly approximated by Maass eigencuspforms and incomplete Eisenstein series, and because the the collection of maps $D_f(\cdot)$ is equicontinuous for the uniform topology, Theorem 1.2 implies that $|D_f(\phi_1)| < \varepsilon$ eventually.[25] Choose a smooth $[0,1]$-valued function h supported on the complement of B_T in $Y_0(1)$ that satisfies $\mu(h)/\mu(1) > 1 - 2\varepsilon$. Theorem 1.2 implies that the positive real number $\mu_f(h)/\mu_f(1)$ eventually exceeds $1 - 3\varepsilon$. By the nonnegativity of μ_f, we deduce that $\mu_f(B_T)/\mu_f(1) < 3\varepsilon$ eventually. Let R be the supremum of $|\phi|$. Then $|\mu_f(\phi_2)/\mu_f(1)| \leq R\mu_f(B_T)/\mu_f(1) \leq 3R\varepsilon$ eventually and $|\mu(\phi_2)/\mu(1)| \leq R\varepsilon$, so that $|D_f(\phi_2)| \leq 4R\varepsilon$ eventually. Thus $|D_f(\phi)| < (1 + 4R)\varepsilon$ eventually. Letting $\varepsilon \to 0$, we obtain Theorem 1.1.

3.7. **Technical arguments.**

Proof of Proposition 3.17. The proof extends that of [33, Theorem 3.10], which in turn refines [17, Theorem 2].

We may assume $1 \leq l \leq x$. Fix $\alpha \in (0, 1/2)$ and set $y = x^\alpha$, $s = \alpha \log\log(x)$, $z = x^{1/s}$. If $x \gg_\alpha 1$ then $10 \leq z \leq y \leq x$, as we henceforth assume. Define finite sets of primes

$$\mathcal{P} = \{p \leq z, p \nmid q\}, \quad \mathcal{P}' = \{p \leq z\} \cup \{p \mid q\}.$$

For each set S of primes, define the *S-part* of a positive integer n, denoted n_S, to be its greatest positive divisor composed entirely of primes in S. We henceforth use the symbol m to denote $n+l$. By the Cauchy–Schwarz inequality, we may bound the contribution to the main sum coming from those terms for which the $\mathcal{P}'$-part of m or of n is $> y$ by

$$\sum_{\substack{\max(m,n) \leq x \\ \max(m_{\mathcal{P}'}, n_{\mathcal{P}'}) > y}} \lambda_{[c]}(m)\lambda_{[c]}(n) \leq 2x \left(\sum_{m \leq x} \frac{\left|\lambda_{[c]}(m)\right|^2}{m} \right)^{1/2} \left(\sum_{\substack{n \leq x \\ n_{\mathcal{P}'} > y}} \frac{\left|\lambda_{[c]}(n)\right|^2}{n} \right)^{1/2}.$$

By Proposition 3.12 and Corollary 3.13, we have

$$\left(\sum_{m \leq x} \frac{\left|\lambda_{[c]}(m)\right|^2}{m} \right) \leq \left(\prod_{p|q} \sum_{k=0}^{\infty} \frac{\lambda_{[c],p}(p^k)^2}{p^k} \right) \sum_{m \leq x} \frac{|\lambda(m)|^2}{m} \ll q_\diamond^\varepsilon \log(x)^3,$$

[25] Here and in what follows, "eventually" means "provided that qk large enough".

and

$$\left(\sum_{\substack{n\leq x\\ n_{\mathcal{P}'}>y}} \frac{\left|\lambda_{[c]}(n)\right|^2}{n}\right) \leq \left(\prod_{p|q}\sum_{k=0}^{\infty}\frac{\lambda_{[c],p}(p^k)^2}{p^k}\right)\sup_{d|q^\infty}\sum_{\substack{n\leq x/d\\ n_{\mathcal{P}}>y/d}}\frac{\left|\lambda(n)\right|^2}{n} \ll q_\diamond^\varepsilon \sup_{d|q^\infty}\sum_{\substack{n\leq x/d\\ n_{\mathcal{P}}>y/d}}\frac{\left|\lambda(n)\right|^2}{n}. \tag{62}$$

To bound the RHS of (62), we consider separately the ranges $d > y^{1/2}$ and $d \leq y^{1/2}$. If $d > y^{1/2}$, then $\sum_{\substack{n\leq x/d\\ n_{\mathcal{P}}>y/d}} \frac{|\lambda(n)|^2}{n} \ll x^{1-\alpha/4}$ thanks to, say, the Deligne bound $|\lambda(n)| \leq \tau(n)$. If $d \leq y^{1/2}$, we apply Cauchy–Schwarz, the Deligne bound, and the estimate $\sum_{\substack{n\leq x\\ n_{\mathcal{P}}>y^{1/2}}} 1 \ll_{A,\alpha} \frac{x}{\log(x)^A}$ for every $A > 0$ which follows from a theorem of Krause [29] (see the discussion in [34, Proof of Lem 6.3]) to deduce that $\sum_{\substack{n\leq x/d\\ n_{\mathcal{P}}>y/d}} \frac{|\lambda(n)|^2}{n} \ll \frac{x}{\log(x)^A}$ (for a different value of A). Combining these estimates, we obtain

$$\sum_{\max(m,n)\leq x} \lambda_{[c]}(m)\lambda_{[c]}(n) \leq \sum_{\substack{\max(m,n)\leq x\\ \max(m_{\mathcal{P}'},n_{\mathcal{P}'})\leq y}} \lambda_{[c]}(m)\lambda_{[c]}(n) + O\left(\frac{q_\diamond^\varepsilon x}{\log(x)^A}\right).$$

To treat the remaining sum, we follow [17] in partitioning it according to the values $m_{\mathcal{P}'}$ and $n_{\mathcal{P}'}$. Specifically, for $a, b, d \in \mathbb{N}$ with $(a, b) = 1$ and $d|l$, let $\mathbb{N}_{abd}$ denote the set of all $n \in \mathbb{N}$ for which $ad = m_{\mathcal{P}'}$ and $bd = n_{\mathcal{P}'}$. Then $\mathbb{N} = \bigsqcup \mathbb{N}_{abd}$. For $n \in \mathbb{N}_{abd}$, we have $\lambda_{[c]}(m)\lambda_{[c]}(n) = \left(\prod_{p\in\mathcal{P}'} \lambda_{[c],p}(ad)\right)\left(\prod_{p\in\mathcal{P}'} \lambda_{[c],p}(bd)\right) \lambda(m/ad)\lambda(n/bd)$ because each prime divisor of q is contained in $\mathcal{P}'$. Recall the notation $\Omega(n) = \sum_{p^\alpha||n} \alpha$ for the number of prime factors of n counted with multiplicity. Since $|\lambda(n)| \leq \tau(n)$ for all $n \in \mathbb{N}$,

$$\tau\left(\frac{m}{ad}\right) = \prod_{p^\alpha||\frac{m}{ad}} (\alpha + 1) \leq 2^{\Omega(m/ad)}, \quad \text{ and } \Omega(m/ad) \leq \frac{\log\left(\frac{m}{ad}\right)}{\log(z)} \leq s,$$

we have $|\lambda(m/ad)| \leq 2^s$. Similarly, $|\lambda(n/bd)| \leq 2^s$. Thus

$$\sum_{\substack{\max(m,n)\leq x\\ \max(m_{\mathcal{P}'},n_{\mathcal{P}'})\leq y}} |\lambda_{[c]}(m)\lambda_{[c]}(n)| \leq 4^s \sum_{d|\ell} \sum_{\substack{a\in\mathbb{N}\\ (a,b)=1\\ \max(ad,bd)\leq y\\ p|abd \implies p\in\mathcal{P}'}}\sum_{b\in\mathbb{N}} \prod_{p\in\mathcal{P}'} \lambda_{[c],p}(ad) \prod_{p\in\mathcal{P}'} \lambda_{[c],p}(bd) \cdot \#(\mathbb{N}_{abd} \cap \mathcal{R}) \tag{63}$$

with $\mathcal{R} := [1, x] \cap [1, x - \ell]$. The factor 4^s is negligible if α is chosen sufficiently small, precisely $4^s \ll_\varepsilon \log(x)^\varepsilon$ for $\alpha \ll_\varepsilon 1$. Set $r = abd^{-1}l$. As in [33] and [17], the large sieve implies

$$\#(\mathbb{N}_{abd} \cap \mathcal{R}) \ll \frac{x/abd + z^2}{\sum_{t\leq z} h(t)},$$

where $h(t)$ is supported on squarefree integers t, multiplicative, and given by

$$h(p) = \begin{cases} 1 & p \mid r \\ 2 & \text{otherwise} \end{cases}$$

on the primes. Note that for all $p \le z$, we have

$$h(p) = \begin{cases} 1 & p \mid r \text{ and } p \le z \\ 2 & \text{otherwise} \end{cases} = \begin{cases} 1 & p \mid r_{\mathcal{P}} \\ 2 & \text{otherwise} \end{cases}.$$

It is standard [14, pp55-59] that

$$\sum_{t \le z} h(t) \gg \frac{\varphi(r_{\mathcal{P}})}{r_{\mathcal{P}}} \log(z)^2.$$

Since $x + abdz^2 \ll x$, $\log(z) \gg \log(x)/\log\log(x) \gg \log(x)^{1-\varepsilon}$ and

$$\frac{\varphi(r_{\mathcal{P}})}{r_{\mathcal{P}}} \gg \log\log(x)^{-1} \log\log(e^e q)^{-1},$$

we obtain

$$\#(\mathbb{N}_{abd} \cap \mathcal{R}) \ll \log\log(e^e q) \frac{1}{abd} \frac{x}{\log(x)^{2-\varepsilon}}.$$

To complete the proof of the proposition, it now suffices to show that

$$\sum_{d \mid \ell} \sum_{\substack{a \in \mathbb{N} \\ (a,b)=1 \\ \max(ad,bd) \le y \\ p \mid abd \implies p \in \mathcal{P}'}} \sum_{b \in \mathbb{N}} \frac{\prod_{p \in \mathcal{P}'} \lambda_{[c],p}(ad) \prod_{p \in \mathcal{P}'} \lambda_{[c],p}(bd)}{abd} \ll q_\diamond^\varepsilon \log(x)^\varepsilon \prod_{p \le x} \left(1 + \frac{2|\lambda_f(p)|}{p}\right). \tag{64}$$

Note first that

$$\sum_{\substack{a \in \mathbb{N} \\ (a,b)=1 \\ \max(ad,bd) \le y \\ p \mid abd \implies p \in \mathcal{P}'}} \sum_{b \in \mathbb{N}} \frac{\prod_{p \in \mathcal{P}'} \lambda_{[c],p}(ad) \prod_{p \in \mathcal{P}'} \lambda_{[c],p}(bd)}{ab} \le \left(\prod_{\substack{p \le z \\ p \nmid q}} \sum_{k \ge 0} \frac{\lambda(p^{k+v_p(d)})}{p^k} \right)^2 \left(\prod_{p \mid q} \sum_{k \ge 0} \frac{\lambda_{[c],p}(p^{k+v_p(d)})}{p^k} \right)^2$$

If $p \nmid q$, then the arguments of [33, Proof of Thm. 3.10] show that $\sum_{k \ge 0} \frac{\lambda(p^{k+v})}{p^k} \le 3v + 3$ if $v \ge 1$ and $\sum_{k \ge 0} \frac{\lambda(p^k)}{p^k} \le \left(1 + \frac{\lambda(p)}{p}\right)\left(1 + \frac{20}{p^2}\right)$. If $p|q$ but $c_p^2 \ne q_p$, we have uniformly $\sum_{k \ge 0} \frac{\lambda_{[c],p}(p^{k+v})}{p^k} \le \frac{1}{1 - p^{-3/2}}$. Finally, if $p^2|q$ and $c_p^2 = q_p$, then Corollary 3.13 shows that $\sum_{k \ge 0} \frac{\lambda_{[c],p}(p^{k+v})}{p^k} \ll p^{\frac{v}{4}}$ where the implied constant is absolute. Putting all this together, and arguing exactly as in [33, Proof of Thm. 3.10], we see that the LHS of (64) is bounded by an absolute constant multiple of

$$\log(x)^\varepsilon \prod_{p \le x} \left(1 + \frac{2|\lambda_f(p)|}{p}\right) \prod_{p \mid q_\diamond} O(1)$$

Since $\prod_{p \mid q_\diamond} O(1) \ll q_\diamond^\varepsilon$ this completes the proof.

□

Proof of Lemma 3.18. This lemma generalizes the bound

$$\sum_{d|q} \frac{d}{\log(dx)^{2-\varepsilon}} \ll \frac{q \log\log(e^e q)}{\log(qx)^{2-\varepsilon}} \tag{65}$$

proved in [33, Lem 3.5], which holds for all squarefree q and all $x \geq 2$, $\varepsilon \in (0,1)$, with an absolute implied constant. The proof of (65) applies a convexity argument to reduce to the case that q is the product of the first r primes, partitions the sum according to the number of divisors of d, and then invokes a weak form of the prime number theorem. Our strategy here is to reduce the general case to that in which q is squarefree, and then apply the known bound (65).

First, note that

$$\sum_{c|q} \frac{[q/c^2,1]\ \varphi((c,q/c))}{\log([q/c^2,1]x)^{2-\varepsilon}} = \sum_{d|q} \varphi((d,q/d)) \frac{[d^2/q,1]}{\log([d^2/q,1]k)^{2-\varepsilon}}.$$

Since

$$\varphi((d,q/d))[d^2/q,1] \leq (d,q/d)[d^2/q,1] = d,$$

we see that it suffices to show

$$\sum_{d|q} \frac{d}{\log([d^2/q,1]k)^{2-\varepsilon}} \ll \frac{q\log\log(e^e q)^{O(1)}}{\log(qk)^{2-\varepsilon}}.$$

From here on, the argument is unfortunately a bit technical. Let $q_1 < \cdots < q_r$ be the distinct prime factors of q. Define maps $B_i : \{d \in \mathbb{N} : d \mid q\} \to \{0,1\}$ by

$$B_i(d) = \begin{cases} 0 & (d, q_i^\infty) \mid q_\diamond \\ 1 & \text{otherwise.} \end{cases}$$

Thus $B_i(d) = 1$ or 0 according as the valuation of d at q_i does or does not exceed half that of q. Let $B = \prod B_i : \{d \in \mathbb{N} : d \mid q\} \to \{0,1\}^r$ be the product map that sends d to the r-tuple $(B_1(d), \dots, B_r(d))$. For each positive divisor $d = \prod q_i^{\alpha_i}$ of q and each $\eta = (\eta_1, \dots, \eta_r) \in \{0,1\}^r$, write $d_\eta = \prod q_i^{\eta_i \alpha_i}$. Our reason for introducing this notation is that for all $d \in B^{-1}(\eta)$, we have $[d^2/q,1] = (d^2/q)_\eta$ and may write $d = q_\diamond (q/q_\diamond)_\eta \prod q_i^{-\delta_i}$ where $\delta_i \geq 0$ for all i. Thus

$$\frac{d}{\log(k[d^2/q,1])^{2-\varepsilon}} = q_\diamond \frac{(q/q_\diamond)_\eta}{\log(k(q/q_\diamond)_\eta)^{2-\varepsilon}} \frac{\log(k(q/q_\diamond)_\eta)^{2-\varepsilon}}{\log(k(d^2/q)_\eta)^{2-\varepsilon}} \prod q_i^{-\delta_i}. \tag{66}$$

Let us now write $q = \prod q_i^{\beta_i}$ and $q_\diamond = \prod q_i^{\gamma_i}$; the definition of $q_\diamond$ implies $\gamma_i = \lfloor \beta_i/2 \rfloor$. Then

$$\frac{\log(k(q/q_\diamond)_\eta)}{\log(k(d^2/q)_\eta)} = \frac{\log(k) + \sum \eta_i(\beta_i - \gamma_i)\log(q_i)}{\log(k) + \sum \eta_i(\beta_i - 2\delta_i)\log(q_i)} \leq \max_{i:\eta_i=1} \frac{\beta_i}{\beta_i - 2\delta_i} \leq \prod_{i:\eta_i=1} \frac{1}{1-2\delta_i/\beta_i}.$$

(In the above, define an empty maximum or an empty product to be 1.) By comparing the sum to an integral, one shows easily that

$$\sum_{0 \leq \delta_i < \beta_i/2} \frac{q_i^{-\delta_i}}{(1-2\delta_i/\beta_i)^{2-\varepsilon}} \leq 1 + \frac{9 + O(1/\log q_i)}{q_i} \leq 1 + O(1/q_i) \leq (1+1/q_i)^{O(1)}.$$

with absolute implied constants. Since $\prod_i (1+1/q_i) \ll \log\log(e^e q)$, we deduce from (66) that

$$\sum_{d \in B^{-1}(\eta)} \frac{d}{\log([d^2/q,1])^{2-\varepsilon}} \ll q_\diamond \frac{(q/q_\diamond)_\eta}{\log(k(q/q_\diamond)_\eta)^{2-\varepsilon}} \log\log(e^e q)^{O(1)}.$$

To complete the proof of the lemma, it suffices now to establish that

$$\sum_{\eta \in \{0,1\}^r} \frac{(q/q_\diamond)_\eta}{\log(k(q/q_\diamond)_\eta)^{2-\varepsilon}} \ll \frac{q/q_\diamond}{\log(k(q/q_\diamond))^{2-\varepsilon}} \log\log(e^e q). \tag{67}$$

As in [33, Proof of Lem 3.5], define $\beta(x) = x/\log(e^e x k)^{2-\varepsilon}$. Then $\beta(x) \asymp x/\log(xk)^{2-\varepsilon}$ for all $x \in \mathbb{R}_{\geq 1}$, so the desired bound (67) is equivalent to

$$\sum_{\eta \in \{0,1\}^r} \frac{\beta((q/q_\diamond)_\eta)}{\beta(q/q_\diamond)} \ll \log\log(e^e q).$$

Since β is increasing on $\mathbb{R}_{\geq 1}$ and the map $\mathbb{R}_{\geq 0} \ni x \mapsto \log\beta(e^x)$ is convex, we have (compare with [33, Proof of Lem 3.5])

$$\begin{aligned}
\frac{\beta((q/q_\diamond)_\eta)}{\beta(q/q_\diamond)} &= \frac{\beta(q_1^{\eta_1(\beta_1-\gamma_1)} \cdots q_r^{\eta_r(\beta_r-\gamma_r)})}{\beta(q_1^{\beta_1-\gamma_1} \cdots q_r^{\beta_r-\gamma_r})} \leq \frac{\beta(q_1^{\eta_1} q_2^{\eta_2(\beta_2-\gamma_2)} \cdots q_r^{\eta_r(\beta_r-\gamma_r)})}{\beta(q_1 q_2^{\beta_2-\gamma_2} \cdots q_r^{\beta_r-\gamma_r})} \\
&\leq \frac{\beta(q_1^{\eta_1} q_2^{\eta_2} q_3^{\eta_3(\beta_3-\gamma_3)} \cdots q_r^{\eta_r(\beta_r-\gamma_r)})}{\beta(q_1 q_2 q_3^{\beta_3-\gamma_3} \cdots q_r^{\beta_r-\gamma_r})} \leq \cdots \\
&\leq \frac{\beta(q_1^{\eta_1} \cdots q_r^{\eta_r})}{\beta(q_1 \cdots q_r)} = \frac{\beta(\prod q_i^{\eta_i})}{\beta(\prod q_i)}.
\end{aligned}$$

But $\prod q_i^{\eta_i}$ is squarefree, so (65) implies

$$\sum_{\eta \in \{0,1\}^r} \frac{\beta(\prod q_i^{\eta_i})}{\beta(\prod q_i)} = \sum_{d \mid \prod q_i} \frac{\beta(d)}{\beta(\prod q_i)} \ll \log\log(e^e \prod q_i) \ll \log\log(e^e q),$$

as desired. ☐

References

[1] A. O. L. Atkin and J. Lehner. Hecke operators on $\Gamma_0(m)$. *Math. Ann.*, 185:134–160, 1970.

[2] François Brunault. On the ramification of modular parametrizations at the cusps. *Preprint.*

[3] Daniel Bump. *Automorphic Forms and Representations*, volume 55 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 1997.

[4] Colin J. Bushnell, Guy M. Henniart, and Philip C. Kutzko. Local Rankin-Selberg convolutions for GL_n: explicit conductor formula. *J. Amer. Math. Soc.*, 11(3):703–730, 1998.

[5] Pierre Deligne. Formes modulaires et reprsentations ℓ-adiques. *Séminaire Bourbaki, Vol. 1968/1969, Exp. 347-363*, 179:139–172, 1971.

[6] Pierre Deligne. La conjecture de Weil. I. *Inst. Hautes Études Sci. Publ. Math.*, 43:273–307, 1974.

[7] W. Duke. Hyperbolic distribution problems and half-integral weight Maass forms. *Invent. Math.*, 92(1):73–90, 1988.

[8] Manfred Einsiedler, Elon Lindenstrauss, Philippe Michel, and Akshay Venkatesh. Distribution of periodic torus orbits and Duke's theorem for cubic fields. *Ann. of Math. (2)*, 173(2):815–885, 2011.

[9] Paul B. Garrett. Decomposition of Eisenstein series: Rankin triple products. *Ann. of Math. (2)*, 125(2):209–235, 1987.

[10] Stephen Gelbart. *Automorphic Forms on Adèle Groups*. Princeton University Press, Princeton, N.J., 1975. Annals of Mathematics Studies, No. 83.

[11] Stephen Gelbart and Hervé Jacquet. A relation between automorphic representations of GL(2) and GL(3). *Ann. Sci. École Norm. Sup. (4)*, 11(4):471–542, 1978.

[12] Stephen Gelbart and Hervé Jacquet. Forms of GL(2) from the analytic point of view. In *Automorphic Forms, Representations and L-functions (Proc. Sympos. Pure Math., Oregon State Univ., Corvallis, Ore., 1977), Part 1*, Proc. Sympos. Pure Math., XXXIII, pages 213–251. Amer. Math. Soc., Providence, R.I., 1979.

[13] Dorian Goldfeld, Joseph Hundley, and Min Lee. Fourier expansions of GL(2) newforms at various cusps. *arXiv e-prints*, 2010. http://arxiv.org/abs/1009.0028.

[14] George Greaves. *Sieves in number theory*, volume 43 of *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]*. Springer-Verlag, Berlin, 2001.

[15] Benedict H. Gross and Stephen S. Kudla. Heights and the central critical values of triple product L-functions. *Compositio Math.*, 81(2):143–209, 1992.

[16] Michael Harris and Stephen S. Kudla. The central critical value of a triple product L-function. *Ann. of Math. (2)*, 133(3):605–672, 1991.

[17] Roman Holowinsky. Sieving for mass equidistribution. *Ann. of Math. (2)*, 172(2):1499–1516, 2010.

[18] Roman Holowinsky and Kannan Soundararajan. Mass equidistribution for Hecke eigenforms. *Ann. of Math. (2)*, 172(2):1517–1528, 2010.

[19] Atsushi Ichino. Trilinear forms and the central values of triple product L-functions. *Duke Math. J.*, 145(2):281–307, 2008.

[20] Atsushi Ichino and Tamutsu Ikeda. On the periods of automorphic forms on special orthogonal groups and the Gross-Prasad conjecture. *Geom. Funct. Anal.*, 19(5):1378–1425, 2010.

[21] Henryk Iwaniec. *Topics in classical automorphic forms*, volume 17 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 1997.

[22] Henryk Iwaniec. *Spectral methods of automorphic forms*, volume 53 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2002.

[23] Henryk Iwaniec and Emmanuel Kowalski. *Analytic number theory*, volume 53 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2004.

[24] Henryk Iwaniec and Peter Sarnak. Perspectives on the analytic theory of L-functions. *Geom. Funct. Anal.*, (Special Volume, Part II):705–741, 2000. GAFA 2000 (Tel Aviv, 1999).

[25] Hervé Jacquet. *Automorphic forms on* GL(2)*. Part II.* Lecture Notes in Mathematics, Vol. 278. Springer-Verlag, Berlin, 1972.

[26] Hervé Jacquet and R. P. Langlands. *Automorphic forms on* GL(2). Lecture Notes in Mathematics, Vol. 114. Springer-Verlag, Berlin, 1970.

[27] Henry H. Kim. Functoriality for the exterior square of GL_4 and the symmetric fourth of GL_2. *J. Amer. Math. Soc.*, 16(1):139–183 (electronic), 2003. With appendix 1 by Dinakar Ramakrishnan and appendix 2 by Kim and Peter Sarnak.

[28] E. Kowalski, P. Michel, and J. VanderKam. Rankin-Selberg L-functions in the level aspect. *Duke Math. J.*, 114(1):123–191, 2002.

[29] Uwe Krause. Abschätzungen für die Funktion $\Psi_K(x, y)$ in algebraischen Zahlkörpern. *Manuscripta Math.*, 69(3):319–331, 1990.

[30] Wenzhi Luo and Peter Sarnak. Mass equidistribution for Hecke eigenforms. *Comm. Pure Appl. Math.*, 56(7):874–891, 2003. Dedicated to the memory of Jürgen K. Moser.

[31] Philippe Michel and Akshay Venkatesh. The subconvexity problem for GL_2. *Publ. Math. Inst. Hautes Études Sci.*, (111):171–271, 2010.

[32] Paul Nelson. Mass distribution of automorphic forms on quaternion algebras. *In preparation*.

[33] Paul Nelson. Equidistribution of cusp forms in the level aspect. *Duke Math. J.*, 160(3):467–501, 2011.

[34] Paul Nelson. Mass equidistribution of Hilbert modular eigenforms. *The Ramanujan Journal*, 27:235–284, 2012.

[35] Ilya Piatetski-Shapiro and Stephen Rallis. Rankin triple L functions. *Compositio Math.*, 64(1):31–115, 1987.

[36] Dipendra Prasad and Dinakar Ramakrishnan. On the global root numbers of $\mathrm{GL}(n) \times \mathrm{GL}(m)$. In *Automorphic forms, automorphic representations, and arithmetic (Fort Worth, TX, 1996)*, volume 66 of *Proc. Sympos. Pure Math.*, pages 311–330. Amer. Math. Soc., Providence, RI, 1999.

[37] Zeév Rudnick and Peter Sarnak. The behaviour of eigenstates of arithmetic hyperbolic manifolds. *Comm. Math. Phys.*, 161(1):195–213, 1994.
[38] Peter Sarnak. Arithmetic quantum chaos. In *The Schur lectures (1992) (Tel Aviv)*, volume 8 of *Israel Math. Conf. Proc.*, pages 183–236. Bar-Ilan Univ., Ramat Gan, 1995.
[39] Peter Sarnak. Recent Progress on QUE. http://www.math.princeton.edu/sarnak/SarnakQUE.pdf, 2009.
[40] Ralf Schmidt. Some remarks on local newforms for GL(2). *J. Ramanujan Math. Soc.*, 17(2):115–147, 2002.
[41] Jean-Pierre Serre. *A course in arithmetic.* Springer-Verlag, New York, 1973. Translated from the French, Graduate Texts in Mathematics, No. 7.
[42] Hideo Shimizu. Some examples of new forms. *J. Fac. Sci. Univ. Tokyo Sect. IA Math.*, 24(1):97–113, 1977.
[43] Goro Shimura. *Introduction to the arithmetic theory of automorphic functions*. Publications of the Mathematical Society of Japan, No. 11. Iwanami Shoten, Publishers, Tokyo, 1971. Kanô Memorial Lectures, No. 1.
[44] Kannan Soundararajan. Weak subconvexity for central values of L-functions. *Ann. of Math. (2)*, 172(2):1469–1498, 2010.
[45] Kannan Soundararajan and Matthew P. Young. The prime geodesic theorem. *arXiv e-prints*, 2010. http://arXiv.org/abs/1011.5486.
[46] Thomas C. Watson. Rankin triple products and quantum chaos. *arXiv e-prints*, 2008. http://arXiv.org/abs/0810.0425.

EPFL, Station 8, CH-1015 Lausanne, Switzerland
E-mail address: paul.nelson@epfl.ch

Department of Mathematics, University of Oklahoma, Norman, OK 73019, USA,
E-mail address: apitale@math.ou.edu

Department of Mathematics, University of Bristol, Bristol BS81TW, UK
E-mail address: abhishek.saha@bris.ac.uk