Some analytic aspects of automorphic forms on GL(2) of minimal type

Yueke Hu, Paul D. Nelson and Abhishek Saha

Abstract. Let π be a cuspidal automorphic representation of $\mathrm{PGL}_2(\mathbb{A}_{\mathbb{Q}})$ of arithmetic conductor C and archimedean parameter T, and let ϕ be an L^2-normalized automorphic form in the space of π. The sup-norm problem asks for bounds on $\|\phi\|_\infty$ in terms of C and T. The quantum unique ergodicity (QUE) problem concerns the limiting behavior of the L^2-mass $|\phi|^2(g)\,dg$ of ϕ. Previous work on these problems in the conductor-aspect has focused on the case that ϕ is a newform.

In this work, we study these problems for a class of automorphic forms that are not newforms. Precisely, we assume that for each prime divisor p of C, the local component π_p is supercuspidal (and satisfies some additional technical hypotheses), and consider automorphic forms ϕ for which the local components $\phi_p \in \pi_p$ are "minimal" vectors. Such vectors may be understood as non-archimedean analogues of lowest weight vectors in holomorphic discrete series representations of $\mathrm{PGL}_2(\mathbb{R})$.

For automorphic forms as above, we prove a sup-norm bound that is sharper than what is known in the newform case. In particular, if π_∞ is a holomorphic discrete series of lowest weight k, we obtain the optimal bound $C^{1/8-\epsilon}k^{1/4-\epsilon} \ll_\epsilon |\phi|_\infty \ll_\epsilon C^{1/8+\epsilon}k^{1/4+\epsilon}$. We prove also that these forms give analytic test vectors for the QUE period, thereby demonstrating the equivalence between the strong QUE and the subconvexity problems for this class of vectors. This finding contrasts the known failure of this equivalence [31] for newforms of powerful level.

1. Introduction

1.1. Overview. Let π be a cuspidal automorphic representation of $\mathrm{GL}_2(\mathbb{A}_{\mathbb{Q}})$. Many problems in the analytic number theory of π depend upon the choice of a specific L^2-normalized automorphic form ϕ in the space of π. For example, the sup norm, L^p-norm and quantum unique ergodicity (QUE) problems have this feature, while the subconvexity problem does not. In such problems, it is customary to work with factorizable vectors $\phi = \otimes \phi_v$ for which

$$\phi_\infty = \text{lowest nonnegative weight vector in } \pi_\infty, \quad \phi_p = \text{newvector in } \pi_p. \quad (1)$$

But other reasonable choices are often possible, useful, and more natural.

A basic illustration of this principle is given by Lindenstrauss's proof of the QUE theorem. One formulation of that theorem is that as π traverses a sequence as above for which π_∞ belongs to the principal series, the L^2-masses of the vectors ϕ given by (1) equidistribute. A key step in the proof is to replace ϕ_∞ by another vector $\widetilde{\phi}_\infty$ (the microlocal lift) whose limit measures acquire additional invariance. Further illustration of this principle is given by period-based approaches to the subconvexity and shifted convolution problems (see e.g. [1, 2, 26]).

This work explores a particular choice for the local components ϕ_p which turn out to have several remarkable properties. Briefly, assuming that π_p is supercuspidal and that its conductor is a fourth power, we consider ϕ_p which are analogues of the lowest weight vectors in holomorphic discrete series representations of $\mathrm{PGL}_2(\mathbb{R})$; see Section 1.4 for a more detailed description of these vectors from this point of view and Definition 2.13 for the formal definition. We aim to demonstrate the strength of our analogy from the analytic perspective by illustrating with two examples: the sup norm problem and the QUE problem.

For lack of better terminology, we refer to these vectors as *minimal vectors* or *vectors of minimal type*. (When π_p belongs to the principal series, analogous vectors were studied in [30].) Minimal vectors are implicit in the type theory approach to the construction of supercuspidal representations, as in the works of Howe [14, 15], Kutzko [24], Moy [27], Bushnell [6], and others. On the other hand, their analytic properties, in the sense of the problems recalled above, do not appear to have been explored. The purpose of the present work is to fill this gap.

Before describing in detail the vectors to be studied, we indicate some of the intended applications.

1.2. The sup norm problem in the level aspect.

Assume that $\phi = \otimes \phi_v$, with ϕ_∞ a vector of lowest non-negative weight and ϕ_p spherical for all primes $p \nmid C$. Then ϕ corresponds to either a Hecke–Maass cusp form f of weight $k \in \{0, 1\}$ and Laplace eigenvalue λ or to a holomorphic Hecke eigencuspform f of weight $k \in \mathbb{Z}_{>0}$ (with respect to some congruence subgroup). The GL(2) sup-norm problem asks for bounds on $\|\phi\|_\infty = \|y^{k/2} f\|_\infty$ in terms of C and k/λ and has been much studied recently. (A variant of this problem asks for bounds on $\|f|_\Omega\|_\infty$, where Ω is a fixed compact set. This formulation avoids the cusps and focusses on behavior at the bulk. We do not discuss this variant in the present paper.)

In the case $C = 1$ and f a Hecke–Maass cusp form of weight 0 for $\mathrm{SL}_2(\mathbb{Z})$, Iwaniec and Sarnak [18] proved the pioneering result

$$\lambda^{1/12-\epsilon} \ll_\epsilon \|f\|_\infty \ll_\epsilon \lambda^{5/24+\epsilon}.$$

Their proof combined the Fourier expansion with a subtle amplification argument. On the other hand, for f a holomorphic cuspidal eigenform of weight k for $\mathrm{SL}_2(\mathbb{Z})$, the Fourier expansion alone turns out to be sufficient to get the optimal exponent in

the weight aspect; this was worked out by Xia [45], who proved

$$k^{1/4-\epsilon} \ll_\epsilon \|y^{k/2} f\|_\infty \ll_\epsilon k^{1/4+\epsilon}. \tag{2}$$

For $C > 1$, one needs to make a choice for ϕ_p at each prime p dividing C. The customary choice has been to take the newvector at each prime. The corresponding forms f are (Hecke–Maass or holomorphic) *newforms* with respect to the group $\Gamma_1(C)$. For such newforms and for *squarefree* C there were several results [5, 10, 11, 39, 41] culminating in the bound $\|\phi\|_\infty \ll_{k/\lambda,\epsilon} C^{1/3+\epsilon}$ due to Harcos and Templier. (Here, for simplicity, we have quoted the bound only in the conductor-aspect, noting that a hybrid result was proved by Templier in [41].) This bound was generalized to the case of powerful (non-squarefree) C by the third author [35]. In the special case of trivial central character, and again focussing only on the conductor aspect, the results of [35] give

$$\|\phi\|_\infty \ll_{k/\lambda,\epsilon} C_0^{1/6+\epsilon} C_1^{1/3+\epsilon}, \tag{3}$$

where we write $C = C_0 C_1$ with C_0 the largest integer such that C_0^2 divides C_1. Note that $C_0^{1/6} C_1^{1/3}$ always lies between $C^{1/4}$ and $C^{1/3}$.

The above bounds have been recently extended to the case of newforms over number fields, initially covering only squarefree conductor [3, 4] and more recently, for all conductors by Edgar Assing (to appear in his forthcoming Bristol thesis). As for lower bounds, one only knows the trivial bound $1 \ll \|\phi\|_\infty$ in general; however in the case when the central character is highly ramified, there exist results giving large lower bounds [34, 40] due to the unusual behavior of local Whittaker newforms (the corresponding best-known upper bounds are also worse in these cases).

Thus, the state-of-the-art for the GL(2) sup-norm problem may seem quite satisfying. Nonetheless there is a key deficiency in all the works so far — they focus exclusively on newforms. The situation for Hecke eigenforms that correspond at the ramified places to interesting local vectors that are not newvectors remains completely unexplored. One aim of this paper is to explore the sup norm problem when ϕ_p is a minimal vector at each prime p dividing C. As indicated above, these local vectors may be viewed as p-adic analogues of holomorphic vectors at infinity. The corresponding global automorphic forms ϕ will be referred to as automorphic forms of minimal type. For such forms, we prove a level aspect sup-norm bound that is strongly analogous to the weight aspect bound (2).

Theorem 1.1 (See Theorem 4.4 for a more general hybrid version). *Let $\pi \simeq \otimes_v \pi_v$ be an irreducible, unitary, cuspidal automorphic representation of $\mathrm{GL}_2(\mathbb{A})$ with trivial central character and conductor C. Assume that $C = N^4$ is the fourth power of an odd integer N and suppose, for each prime p dividing C, that π_p is a supercuspidal representation. Let ϕ be an L^2-normalized automorphic form in the space of π that is of minimal type. Then*

$$C^{\frac{1}{8}-\epsilon} \ll_{k/\lambda,\epsilon} \|\phi\|_\infty \ll_{k/\lambda,\epsilon} C^{\frac{1}{8}+\min(\frac{1}{32},\frac{\delta_\pi}{2})+\epsilon}.$$

Above, δ_π is any exponent towards the Ramanujan conjecture for π; in particular we may take $\delta_\pi = 0$ if π_∞ is holomorphic and $\delta_\pi = 7/64$ otherwise.

The upper-bound in Theorem 1.1 is much stronger than what is known when ϕ is a newform (with the same assumptions on π as above). In the newform case, the best known upper bound [35] is $C^{1/4+\epsilon}$, which is just the "local bound" in the level aspect (both for newforms as well as for the minimal automorphic forms considered here). The bound obtained in this paper gives the first instance of an automorphic form of powerful level for which the local sup-norm bound in the level aspect has been improved upon. Furthermore, our bound is *optimal* in the case when ϕ corresponds to a holomorphic cusp form, and the proof (as we will see) relies only on the Whittaker/Fourier expansion. Thus, it is very close to Xia's result [45] in many respects.

1.3. Period integrals for QUE. Going back to the holomorphic newform case, assume that the local components of ϕ are given by (1), that π has trivial central character, and that π_∞ is a holomorphic discrete series of lowest weight k. Then ϕ corresponds to a holomorphic newform f of weight k with respect to $\Gamma_0(C)$. For each Hecke–Maass cusp form g of weight 0 for $\mathrm{SL}_2(\mathbb{Z})$, define

$$D_f(g) = \frac{\int_{\Gamma_0(C)\backslash\mathbb{H}} y^k |f|^2(z) g(z) \frac{dx\,dy}{y^2}}{\int_{\Gamma_0(C)\backslash\mathbb{H}} y^k |f|^2(z) \frac{dx\,dy}{y^2}}.$$

The problem of proving $D_f(g) \to 0$ for fixed g as the parameters C and k of f grow is a natural analogue of the Rudnick–Sarnak quantum unique ergodicity (QUE) conjecture [33]. It was proved by Holowinsky and Soundararajan [13] that $D_f(g) \to 0$ for fixed C $(= 1)$ and varying $k \to \infty$; we refer to their paper and [37] for further historical background. The case of varying squarefree levels was addressed in [29], where it was shown that $D_f(g) \to 0$ as $Ck \to \infty$ provided that C is squarefree. Finally, it was proved in [31] that $D_f(g) \to 0$ whenever $Ck \to \infty$ (without any restriction on C). In fact, the main result of [31] gave an unconditional power savings bound $D_f(g) \ll_g C_0^{-\delta_1} \log(Ck)^{-\delta_2}$ for some positive constants δ_1, δ_2, where as before, C_0 denotes the largest integer such that $C_0^2|C$. Further extensions of this result to the case when g is not of full level were obtained in [16].

There is a marked difference above between the case when C is squarefree and the case when C is powerful. For C squarefree, a generalization of Watson's formula (see [29] for a precise version) asserts that for each g as above, corresponding to an automorphic representation σ_g, one has

$$|D_f(g)|^2 = (Ck)^{-1+o(1)} L(\pi \times \pi \times \sigma_g, 1/2). \tag{4}$$

Here the convexity bound reads $L(\pi \times \pi \times \sigma_g, 1/2) \ll (Ck)^{1+o(1)}$. Thus, for squarefree levels, the subconvexity and QUE problems are essentially equivalent. A

major point of [31] was that this equivalence is no longer true for powerful levels. For example, in the case when C is a perfect square, the results of [31] imply that $\left|D_f(g)\right|^2 \ll_{g,k} C^{\theta-1}L(\pi \times \pi \times \sigma_g, 1/2)$ where $\theta = 7/64$. The convexity bound in this case gives $L(\pi \times \pi \times \sigma_g, 1/2) \ll_k C^{1/2+o(1)}$. So in this case, the convexity bound alone is enough to imply QUE with power savings in the level aspect! More generally, as shown in [31], the QUE problem is significantly *easier* than the subconvexity problem in the case of newforms of powerful level (in contrast to the squarefree case, where these problems are essentially equivalent).

One may ask whether the equivalence between QUE and subconvexity might be recovered for powerful levels by replacing the newform with a different choice of vector. We show that this is indeed the case for automorphic forms having a local component of minimal type in a supercuspidal representation of fourth power conductor. For a related observation when the local component belongs to a principal series representation, see [30, Rmk. 30].

Let π, $C = N^4$ and ϕ be as in Theorem 1.1. We assume that π_∞ is a holomorphic discrete series of lowest weight k. We can associate to ϕ a holomorphic modular form f defined by $f(z) = j(g_\infty, i)^k \phi(g_\infty)$ where $g_\infty \in \mathrm{SL}_2(\mathbb{R})$ is any matrix such that $g_\infty i = z$. We let Γ denote any congruence subgroup such that $f|_k \gamma = f$ for all $\gamma \in \Gamma$ (we will see later that we may take $\Gamma = \Gamma(N^2)$). We prove the following result.

Theorem 1.2. *Let g be a Hecke–Maass cuspform for $\mathrm{SL}_2(\mathbb{Z})$, and let σ_g be the automorphic representation generated by (the adelization of) g. Then*

$$\frac{\left|\int_{\Gamma\backslash\mathbb{H}} g(z)|f|^2(z)y^k \frac{dx\,dy}{y^2}\right|^2}{\left(\int_{\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}} |g|^2(z)\frac{dx\,dy}{y^2}\right)\left(\int_{\Gamma\backslash\mathbb{H}} |f|^2(z)y^k \frac{dx\,dy}{y^2}\right)^2}$$
$$= \frac{1}{8}\frac{\Lambda(\pi \times \pi \times \sigma_g, 1/2)}{\Lambda(\mathrm{ad}\,\sigma_g, 1)\Lambda(\mathrm{ad}\,\pi, 1)^2}\prod_{p|C} I_p,$$

where each local factor I_p satisfies

$$I_p \asymp \mathrm{Cond}(\pi_\mathfrak{p} \times \pi_p)^{-1/2}.$$

In the above case, the convexity bound reads

$$\Lambda(\pi \times \pi \times \sigma_g, 1/2) \ll_{g,k} C^{1/2+o(1)} = \mathrm{Cond}(\pi \times \pi)^{1/2+o(1)}.$$

So Theorem 1.2 shows that for the family of cusp forms coming from minimal vectors, the QUE and subconvexity problems are essentially equivalent. In fact, our local results imply more general identities in which g is allowed to have some level.

It is very likely that, by combining Theorem 1.2 with the arguments of [31, Sec. 3], one could establish the estimate $D_f(g) \ll \log(Ck)^{-\delta}$ for small $\delta > 0$ and fixed g, but we do not pursue this here.

1.4. Automorphic forms of minimal type. In the rest of this introduction, we explain in detail the concept of an automorphic form of minimal type and touch upon some of the key ideas that power our results.

Let $\pi \simeq \otimes_v \pi_v$ be an irreducible, unitary, cuspidal representation of $\mathrm{GL}_2(\mathbb{A}_\mathbb{Q})$ of conductor C. We assume henceforth for simplicity that the central character of π is trivial. An automorphic form $\phi = \otimes_v \phi_v$ in the space of π can be constructed out of any choice of local vectors $\phi_v \in \pi_v$ such that ϕ_p is spherical ($\mathrm{GL}_2(\mathbb{Z}_p)$-fixed) at almost all primes p. It is very natural to choose ϕ_p to be the (unique up to multiples) spherical vector at *all* primes not dividing the conductor C, and we will always do so. At the archimedean place, we will choose ϕ_∞ to be a vector of minimal non-negative weight k, i.e. with the property

$$\pi_\infty \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \phi_\infty = e^{ik\theta}\phi_\infty, \tag{5}$$

where k is the smallest non-negative integer (which in our case must be an even integer as the central character is trivial) for which the above equality holds for some ϕ_∞. Note that $k = 0$ if π_∞ is a principal series representation and $k \geq 2$ if π_∞ is a discrete series representation.

Now, consider the primes p dividing C. What should we take ϕ_p to be? One standard possibility would be to take ϕ_p to be the newvector, i.e. fixed by a congruence subgroup of the form $\begin{bmatrix} 1+p^c\mathbb{Z}_p & \mathbb{Z}_p \\ p^c\mathbb{Z}_p & \mathbb{Z}_p^\times \end{bmatrix}$, where c is taken as small as possible, whence newform theory implies $c = v_p(C)$.

The minimal vectors studied in this paper may be viewed as an alternative to the newvector in many cases. As we now explain, they may be regarded as non-archimedean analogues of the holomorphic (lowest weight) vector at infinity for a discrete series. Let $T_\infty := \mathbb{R}^\times \mathrm{SO}(2)$ be the standard maximal non-split torus inside $\mathrm{GL}_2(\mathbb{R})$; we have the isomorphism $T_\infty \cong \mathbb{C}^\times$ sending $r\begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix}$ to re^{it}. Let θ_{π_∞} be the character on $\mathbb{C}^\times$ given by $\theta_{\pi_\infty}: re^{it} \mapsto e^{ikt}$ which we may view as a character on T_∞. Then the equality (5) may be restated as

$$\pi_\infty(t_\infty)\phi_\infty = \theta_{\pi_\infty}(t_\infty)\phi_\infty, \quad t_\infty \in T_\infty. \tag{6}$$

The character θ_{π_∞} depends only on k and is therefore an invariant attached to π_∞.

Let us further explicate the relation between π_∞ and θ_{π_∞} when π_∞ is a discrete series representation. Let ξ_{π_∞} be the character on $\mathbb{C}^\times$ given by $re^{it} \mapsto e^{i(k-1)t}$. By a special case of the local Langlands correspondence — see [21, (3.4)] and note that $\pi_\infty \simeq D_{k-1}$ in the notation of [21] — the L-parameter of π_∞ under the local Langlands correspondence is the representation $\mathrm{Ind}_{W_\mathbb{C}}^{W_\mathbb{R}}\xi_{\pi_\infty}$ of the real Weil group $W_\mathbb{R}$; equivalently, the representation π_∞ is obtained by *automorphic induction* from the character ξ_{π_∞} of $\mathbb{C}^\times$. Let $\eta_\mathbb{C}$ be the character on $\mathbb{C}^\times$ given by $re^{it} \mapsto e^{it}$ which we may think of as the simplest extension of the sign character on $\mathbb{R}^\times$ to $\mathbb{C}^\times$. Then we have $\theta_{\pi_\infty} = \eta_\mathbb{C}\xi_{\pi_\infty}$.

Next, take p to be a prime dividing C. Then there is a unique unramified quadratic field extension E_p of $\mathbb{Q}_p$ which should replace $\mathbb{C}$ in our analogy. As in the archimedean case, we can specify a maximal non-split torus T_p inside $\mathrm{GL}_2(\mathbb{Z}_p)$ such that $T_p \simeq E_p^\times$; without loss of generality we may assume that T_p is in canonical form (see Definition 2.2). Now, suppose that π_p is a supercuspidal representation of even minimal (exponent of) conductor. Then, similarly to above, π_p is obtained by automorphic induction from some regular character ξ_{π_p} of $E_p^\times$ (see [42, Prop. 3.5]). Let η_{E_p} be the unique unramified extension to $E_p^\times$ of the quadratic character on $\mathbb{Q}_p^\times$ associated to the extension $E_p/\mathbb{Q}_p$ by local class field theory. We view η_{E_p} as the non-archimedean analogue of the character $\eta_{\mathbb{C}}$ defined earlier. Define the character θ_{π_p} on $T_p \simeq E_p^\times$ by $\theta_{\pi_p} = \eta_{E_p}\xi_{\pi_p}$, which is then the analogue of the character θ_{π_∞} on $T_\infty \simeq \mathbb{C}^\times$ defined above. Analogously to (6), we define a *minimal vector* to be any non-zero element ϕ_p in the space of π_p such that

$$\pi_p(t_p)\phi_p = \theta_{\pi_p}(t_p)\phi_p, \quad t_p \in T_p. \tag{7}$$

The comparison of (6) and (7) shows that minimal vectors are the non-archimedean analogues of the lowest weight (holomorphic) vectors in archimedean discrete series representations. The minimal vectors also occur naturally from the point of view of microlocal analysis, in that they are approximate eigenvectors under the action by small elements of the group; they are in this sense analogous also to the p-adic microlocal lifts considered in [30]. We remark here that given a character χ_p of $T_p \simeq E_p^\times$, a T_p-eigenvector with eigencharacter χ_p is a vector ϕ_p that satisfies $\pi_p(t_p)\phi_p = \chi_p(t_p)\phi_p$ for each $t_p \in T_p$. The choice $\chi_p = \theta_{\pi_p}$ corresponds to our case, whereby the vector acquires some remarkable properties.

The Saito–Tunnell theorem [36, 43] implies that a minimal vector, *if it exists*, is unique up to multiples (once the group T_p is fixed); moreover, a minimal vector exists if and only if $\epsilon(1/2, \pi_p \otimes \mathrm{AI}(\theta_{\pi_p}^{-1})) = 1$ (where AI denotes automorphic induction from $E_p^\times$). We verify in Proposition 2.12 below that if p is odd, $v_p(C)$ is a multiple of 4, and π_p is supercuspidal, then a minimal vector (as we have defined it) indeed exists. Precisely, given such a π_p, the character θ_{π_p} of T_p can be extended to a character χ_{π_p} of the compact-mod-centre group $L := T_p(1 + p^n M_2(\mathbb{Z}_p))$ (where $n = \frac{v_p(C)}{4}$) with the property that $\pi_p \simeq c\text{-Ind}_L^G \chi_{\pi_p}$. The restriction of π_p to L contains χ_{π_p}, which gives an immediate proof of existence. Incidentally, the pair (L, χ_{π_p}) is in some sense the smallest possible among all inducing pairs for π_p and constitutes a minimal K-type in the sense of Moy [27]. Therefore, a minimal vector, in our setup, is precisely one that generates the (one-dimensional) minimal K-type associated to π_p. This is one of the reasons for our use of the term "minimal" to describe these vectors.

Returning to the global setup, we suppose that $C = N^4$ is the fourth power of an odd integer, and π_p is supercuspidal at all primes dividing C. Then, by choosing ϕ_p to be a minimal vector at each prime p dividing C, we can construct a

global automorphic form $\phi = \otimes_v \phi_v$ in the space of π; we call this an automorphic form of minimal type. It is precisely for such forms ϕ that our Theorem 1.1 applies.

We end this subsection with a brief discussion of what an automorphic form ϕ of minimal type looks like classically. We can associate to ϕ a function f on $\mathbb{H}$ defined by $f(z) = j(g_\infty, i)^k \phi(g_\infty)$ where $g_\infty \in \mathrm{SL}_2(\mathbb{R})$ is any matrix such that $g_\infty i = z$. Then there exists an integer D and a character χ_π on the "toric" congruence group

$$\Gamma_{T,D}(N) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) : a \equiv d \;(\mathrm{mod}\,N), \; c \equiv -bD \;(\mathrm{mod}\,N) \right\}$$

such that

$$f|_k \gamma = \chi_\pi(\gamma) f, \quad \gamma \in \Gamma_{T,D}(N).$$

The character χ_π turns out to be trivial on the principal congruence subgroup of level N^2 which is contained in $\Gamma_{T,D}(N)$; see Remark 4.2 for more details. Thus, f is a (very special) member of the space of (holomorphic or Maass) Hecke eigencuspforms of weight $k \in 2\mathbb{Z}$ with respect to the principal congruence subgroup of level N^2. Theorem 1.1 gives the optimal sup-norm bound in the conductor aspect (assuming the Ramanujan conjecture) for such forms f, just as (2) gives the optimal sup-norm bound in the weight aspect for holomorphic eigencuspforms. This fits nicely with our analogy between holomorphic vectors at infinity and minimal vectors at a finite prime.

1.5. The Whittaker expansion. The strong bound in Theorem 1.1 is obtained purely from the Whittaker (Fourier) expansion, and depends on an important property of minimal vectors. We now explain this.

As before, let $\pi \simeq \otimes_v \pi_v$ be an irreducible, unitary, cuspidal representation of $\mathrm{GL}_2(\mathbb{A}_\mathbb{Q})$ of conductor $C = N^4 = \prod_p p^{4n_p}$ and of trivial central character. We begin with some general discussion, which applies to any automorphic form ϕ in the space of π. The Whittaker expansion for ϕ, which we want to exploit to bound $|\phi(g)|$, looks as follows,

$$\phi(g) = \sum_{q \in \mathbb{Q}_{\neq 0}} W_\phi\left(\begin{pmatrix} q & \\ & 1 \end{pmatrix} g \right),$$

where W_ϕ is the global Whittaker function attached to ϕ. Let $g = g_\mathbf{f} g_\infty \in G(\mathbb{A})$, where $g_\mathbf{f}$ denotes the finite part of g and g_∞ denotes the infinite component. There is an integer $Q(g_\mathbf{f})$, depending on $g_\mathbf{f}$, such that the Whittaker expansion above is supported only on those q whose denominator divides $Q(g_\mathbf{f})$. Moreover, the sum decays very quickly after a certain point $|q| > T(g_\infty)$ due to the exponential decay of the Bessel function. The upshot is that

$$\phi(g) = \sum_{m \in \mathbb{Z}_{\neq 0}} W_\phi\left(\begin{pmatrix} m/Q(g_\mathbf{f}) & \\ & 1 \end{pmatrix} g \right) \tag{8}$$

with only the terms $|m| < Q(g_\mathbf{f}) T(g_\infty)$ contributing essentially.

Now, suppose that ϕ is an automorphic form of minimal type. We let $g_{\mathbf{f}}$ vary over the set $\prod_{p|C} \mathrm{GL}_2(\mathbb{Z}_p)$ and g_∞ vary over the set $\begin{pmatrix} y & x \\ & 1 \end{pmatrix}$ with $y \geq \sqrt{3}/2$. This gives a generating domain, similar to the one used in [35], and leads to $Q(g_{\mathbf{f}}) = N^2$. Using this alone, a standard argument (see the discussion in Section 1.4 of [35]) gives the bound $|\phi(g)| \ll_{k/\lambda,\epsilon} C^{1/4+\epsilon}$, which is as good as the best known bound in the case of newforms. Incidentally, it turns out that $C^{1/4+\epsilon}$ is the "local bound" in our case just as it is in the case of newforms of conductor C. This follows from Corollary 3.4. Here, we use the term "local bound" in the sense of [25].

Theorem 1.1 of course, goes beyond the local bound, and indeed gives the optimal bound under Ramanujan. What allows us to do this is the following key property of the local Whittaker function W_{ϕ_p} associated to a minimal vector, namely, for each $k \in \mathrm{GL}_2(\mathbb{Z}_p)$ there exists some $\mathbf{a}_k \in \mathbb{Z}_p^\times$ such that $W_{\phi_p}\left(\begin{pmatrix} q & \\ & 1 \end{pmatrix}k\right) \neq 0$ for $q \in \mathbb{Q}_p^\times$ if and only if $p^{2n_p}q$ belongs to $\mathbb{Z}_p^\times$ and satisfies $p^{2n_p}q \equiv \mathbf{a}_k \pmod{p^{n_p}}$. In sharp contrast, the formula for the Whittaker function of a newvector involves a sum of twisted GL_2-epsilon factors [34, Section 2.7], with a likely cancellation that seems difficult to prove.

Using the factorization of global Whittaker functions, it follows that (8) takes the form

$$\phi(g) = \sum_{m \equiv A \bmod N} W_\phi\left(\begin{pmatrix} m/N^2 & \\ & 1 \end{pmatrix} g\right) \tag{9}$$

for some integer A. In other words, the Whittaker function of an automorphic form of minimal type is supported on an *arithmetic progression*.

This last point can also be explained classically. Suppose that π_∞ is a holomorphic discrete series of lowest weight k, in which case ϕ corresponds to a holomorphic modular form f with respect to the group $\Gamma_{T,D}(N)$. Then the above discussion implies that the Fourier expansion of f at any cusp $\alpha = \sigma(\infty)$ takes the form

$$(f|_k\sigma)(z) = \sum_{\substack{n>0 \\ n \equiv b \bmod N}} a_f(n;\alpha)e^{2\pi i n z/N^2}. \tag{10}$$

It is precisely the fact that the Fourier coefficients above are supported on an arithmetic progression that allows us to get the additional savings beyond the local bound.

1.6. Further remarks. The minimal vectors have many other important properties that we have not discussed above. Perhaps their most striking feature is that the matrix coefficient associated to an L^2-normalized minimal vector is a *character* of the supporting subgroup (see Proposition 3.2). This matrix coefficient formula can be easily used to calculate the local integrals of Gan–Gross–Prasad type involving a minimal vector (as in the proof of Theorem 1.2). More generally, one might hope to use such vectors in classical period formulas (e.g. in Waldspurger's formula or

the triple product formula) with a view towards applications to subconvexity, mass equidistribution, L^p-norms, arithmetic of special L-values, and so on; Theorem 1.2 may be understood as a first step in that direction.

The fact that the matrix coefficient of a minimal vector turns out to be a character also has another very interesting interpretation, which further justifies our use of the word "minimal." By formal degree considerations, the integral of the square of the matrix coefficient associated to an L^2-normalized vector in a square-integrable local representation π_p of conductor p^{c_p} is *independent* of the choice of vector, and equals roughly $p^{-c_p/2}$. The matrix coefficient of an L^2-normalized minimal vector is a character and so has *maximum* possible absolute value on the support (since the absolute value of a matrix coefficient of an L^2-normalized vector can never exceed 1, by the triangle inequality). Therefore the minimal vectors have the property that their matrix coefficients have as small support as possible!

Incidentally, this last fact makes such a matrix coefficient a great choice as a test function in the pre-trace formula for amplification purposes, since small support translates to more congruences for counting purposes. Indeed, while Theorem 1.1 does not rely on any sort of amplification, one could consider the analogous sup-norm problem for automorphic forms of minimal type on a compact quotient of the upper half-plane. In this case, while there is no Whittaker expansion, an amplification argument should allow one to achieve an upper bound for the sup-norm in the conductor aspect that improves upon the local bound. One could also consider analogous problems for quaternion algebras ramified at infinity, where similarly strong bounds may be expected from amplification. We suppress further discussion of this topic in the interest of brevity.

Next, we say a few words about the restriction to C being a fourth power of an odd integer, and π_p being supercuspidal at all primes dividing p. These conditions can in fact be removed when p is not equal to 2, provided one is happy to slightly relax the definition of minimal vector. To give an example, consider the case of an odd prime p such that π_p is supercuspidal of even minimal (exponent of) conductor but $v_p(C) \equiv 2 \pmod 4$. In this case, no vector satisfying (7) exists (the Saito–Tunnell criterion is not satisfied). However, if one were to slightly perturb (7) by multiplying θ_{π_p} by any character of conductor p, then vectors satisfying the resulting equality indeed exist. Similar discussion (roughly in the spirit of [30]) applies to principal series representations (one needs to take $E_p = \mathbb{Q}_p \times \mathbb{Q}_p$ in this case) as well as dihedral supercuspidals with odd minimal (exponent of) conductor, for which we should take E_p to be a ramified quadratic extension of $\mathbb{Q}_p$. Indeed, if $p \neq 2$, every case can be covered, leading to a comprehensive theory of such "almost-minimal" vectors that takes care of every type of representation. The sup-norms of the resulting automorphic forms of almost-minimal type can be studied similarly, though the bounds will be sometimes slightly worse than what we get.

The case of $p = 2$ is much more subtle due to the presence of non-dihedral supercuspidals, and currently it is not clear to us how to define minimal vectors in

that case. One general possibility in every case might be to consider a vector inside a minimal K-type. The details of this theory over GL(2) can be found in [24]. Such a definition should in fact work not just for GL(2) but for all reductive groups, using a well known theorem of Moy–Prasad [28] on the existence of unrefined minimal K-types for irreducible, admissible representations of p-adic reductive groups. It would be very interesting to see if these ideas can be used to study the sup-norm problem in the level aspect for higher rank groups.

Notations. We collect here some general notations that will be used throughout this paper. Additional notations will be defined where they first appear in the paper.

Let $\mathbb{H}$ denote the upper half plane and $\mathrm{GL}_2(\mathbb{R})^+$ the group of real two-by-two matrices with positive determinant. For $z \in \mathbb{H}$, $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{GL}_2(\mathbb{R})^+$, we let $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)z = \frac{az+b}{cz+d} \in \mathbb{H}$ be the point obtained by Möbius transformation. Given a function f on $\mathbb{H}$, an integer k, and some $\gamma = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{GL}_2(\mathbb{R})^+$, we define a function $f|_k\gamma$ on $\mathbb{H}$ via $(f|_k\gamma)(z) = \det(\gamma)^{k/2}(cz + d)^{-k} f(\gamma z)$.

For any two complex numbers α, z, we let $K_\alpha(z)$ denote the modified Bessel function of the second kind. The symbol φ denotes the Euler totient function.

For elements x, y, t in some ring R, we define the following two-by-two matrices over R:

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

We use the notation $A \ll_{x,y,\dots} B$ to signify that there exists a positive constant C, depending at most upon $x, y, \dots$ so that $|A| \le C|B|$. The absence of the subscripts $x, y, \dots$ will mean that the constant C is universal. We will use $A \asymp B$ to mean that $B \ll A \ll B$. The symbol ϵ will denote a small positive quantity. The values of ϵ and that of the constant implicit in $\ll_\epsilon$ may change from line to line.

We shall always assume every character is continuous (but not necessarily unitary). The convention used for our Hermitian inner products is that they are linear in the first variable.

2. Minimal vectors and their Whittaker functions

This section will be purely local.

2.1. Preliminaries on fields, characters and representations.

Let F denote a non-archimedean local field of characteristic zero. We assume throughout that F has *odd* residue cardinality q. Let $\mathfrak{o}$ be its ring of integers, and $\mathfrak{p}$ its maximal ideal. Fix a uniformizer ϖ of $\mathfrak{o}$ (a choice of generator of $\mathfrak{p}$). Let $|\cdot|$ denote the absolute value on F normalized so that $|\varpi| = q^{-1}$. For each $x \in F^{\times}$, let $v(x)$ denote the integer such that $|x| = q^{-v(x)}$. For a non-negative integer m, we define the subgroup U_m of $\mathfrak{o}^{\times}$ to be the set of elements $x \in \mathfrak{o}^{\times}$ such that $v(x - 1) \geq m$.

We denote the unique unramified quadratic field extension of F by E. Since q is odd, we note that $E = F(\sqrt{\delta})$ for any element $\delta \in \mathfrak{o}^{\times} \setminus (\mathfrak{o}^{\times})^2$. We denote the ring of integers of E by $\mathfrak{o}_E$. The valuation v and the absolute value $|\ |$ naturally extend to the field E. Note that ϖ is a uniformizer of $\mathfrak{o}_E$. We let $x \mapsto \bar{x}$ denote the unique non-trivial automorphism of E/F.

We let η denote the unique unramified quadratic character of $F^{\times}$; equivalently, η is the character associated to the extension E/F via local class field theory. For each character χ of $F^{\times}$, we let $a(\chi)$ denote the smallest integer such that χ is trivial on the subgroup $U_{a(\sigma)}$. Similarly, for a character χ of $E^{\times}$, we let $a(\chi)$ denote the smallest integer such that χ is trivial on the subgroup $\{x \in \mathfrak{o}_E^{\times} : v(x - 1) \geq a(\chi)\}$.

We fix once and for all an additive character ψ of F such that ψ is trivial on $\mathfrak{o}$ but not on $\varpi^{-1}\mathfrak{o}$. We let $\psi_E := \psi \circ \mathrm{tr}_{E/F}$ be the corresponding additive character on E.

Throughout this section, we denote $G = \mathrm{GL}_2(F)$ and $K = \mathrm{GL}_2(\mathfrak{o})$. Define subgroups $N = \{n(x) : x \in F\}$, $A = \{a(y) : y \in F^{\times}\}$, $Z = \{z(t) : t \in F^{\times}\}$, $B_1 = NA$, and $B = ZNA = G \cap \left[\begin{smallmatrix} * & * \\ & * \end{smallmatrix}\right]$ of G. For each integer r, denote

$$K_1(r) = K \cap \begin{bmatrix} 1 + \mathfrak{p}^r & \mathfrak{o} \\ \mathfrak{p}^r & \mathfrak{o} \end{bmatrix}, \quad K(r) = K \cap \begin{bmatrix} 1 + \mathfrak{p}^r & \mathfrak{p}^r \\ \mathfrak{p}^r & 1 + \mathfrak{p}^r \end{bmatrix},$$

$$B_1(r) = K \cap \begin{bmatrix} 1 + \mathfrak{p}^r & \mathfrak{p}^r \\ 0 & 1 \end{bmatrix}.$$

We note our normalization of Haar measures. 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.

For π an irreducible admissible generic representation of G, we let ω_{π} denote the central character of π. We define $a(\pi)$ to be the smallest non-negative integer such that π has a $K_1(\mathfrak{p}^{a(\pi)})$-fixed vector. It is known that π can be realized as a

unique subrepresentation of the space of functions W on G satisfying $W(n(x)g) = \psi(x)W(g)$. This is the Whittaker model of π and will be denoted $\mathcal{W}(\pi, \psi)$.

If π is unitary, there is a unique (up to multiples) G-invariant inner product $\langle \, , \, \rangle$ on it. In this case, for a vector $v_0 \in \pi$, we define the matrix coefficient Φ_{v_0} on G as follows:

$$\Phi_{v_0}(g) = \frac{\langle \pi(g)v_0, v_0 \rangle}{\langle v_0, v_0 \rangle},$$

which is clearly unchanged if v_0 is multiplied by a constant and is also independent of the normalization of inner product. We will normalize the inner product in the model $\mathcal{W}(\pi, \psi)$ as follows:

$$\langle W_1, W_2 \rangle = \int_{F^\times} W_1(a(t))\overline{W_2(a(t))}d^\times t. \tag{11}$$

The following lemma will be useful for us.

Lemma 2.1. *Let π be an irreducible admissible supercuspidal representation of G such that $a(\omega_\pi) < a(\pi)/2$. Then π is twist-minimal, i.e. $a(\pi \otimes \chi) \geq a(\pi)$ for each character χ of $F^\times$.*

Proof. Suppose, on the contrary, that $\pi \simeq \sigma \otimes \chi^{-1}$ with σ minimal, and $a(\sigma) < a(\pi)$. As σ and π are supercuspidal, we have $2 \leq a(\sigma) < a(\pi)$. By a result of Tunnell [42, Prop. 3.4], we have $a(\pi) = a(\sigma \otimes \chi^{-1}) = 2a(\chi)$; so $a(\chi) > 1$. Since q is odd, we have that $a(\chi^2) = a(\chi) = a(\pi)/2$. As $a(\omega_\pi) < a(\pi)/2$, it follows that $a(\omega_\pi \chi^2) = a(\chi^2)$. On the other hand, we have $\omega_\pi = \omega_\sigma \chi^{-2}$, i.e. $\omega_\sigma = \omega_\pi \chi^2$. Therefore, $a(\omega_\sigma) = a(\chi^2) = a(\chi) = a(\pi)/2 > a(\sigma)/2$, which contradicts Proposition 3.4 of [42]. $\square$

2.2. Inert tori and their eigenvectors. For α, β, γ elements of F, denote

$$S = S_{\alpha, \beta, \gamma} = \begin{pmatrix} \alpha & \beta/2 \\ \beta/2 & \gamma \end{pmatrix}$$

and define

$$T_{\alpha, \beta, \gamma} := \{g \in \mathrm{GL}_2(F) : {}^t g S g = \det(g)S\}.$$

Definition 2.2. A subgroup T of G is called an *inert torus* if $T = T_{\alpha, \beta, \gamma}$ such that $\delta := \beta^2 - 4\alpha\gamma$ satisfies[1] $E = F(\sqrt{\delta})$. An inert torus T is said to be in *canonical form* if $T = T_{\alpha, 0, 1}$ for some $\alpha \in \mathfrak{o}^\times$, $-\alpha \notin (\mathfrak{o}^\times)^2$.

If $T = T_{\alpha, \beta, \gamma}$ is an inert torus, then the map

$$x + y\sqrt{\delta}/2 \mapsto \left(\begin{pmatrix} x + y\beta/2 & y\gamma \\ -y\alpha & x - y\beta/2 \end{pmatrix} \right) \tag{12}$$

---

[1]Equivalently, δ is not a square in F and $v(\delta)$ is even.

gives an explicit isomorphism from $E^\times$ to T. If T is an inert torus in canonical form, that $\delta = -4\alpha$ and (12) takes $\mathfrak{o}_E^\times$ isomorphically onto $T(\mathfrak{o}) = T \cap K$. It follows immediately that for an inert torus T in canonical form we have

$$T = ZT(\mathfrak{o}) = \bigsqcup_{n \in \mathbb{Z}} \varpi^n T(\mathfrak{o}).$$

We note down several additional useful properties about inert tori.

Proposition 2.3. (1) *All inert tori in G are conjugate.*

(2) *Let T be an inert torus. Then there exists $g \in G$ such that gTg^{-1} is in canonical form.*

(3) *If T_1, T_2 are two inert tori in canonical form, then there exists $y \in \mathfrak{o}^\times$ such that $T_1 = a(y)T_2 a(y)^{-1}$.*

(4) *Let T be an inert torus in canonical form. Then $G = B_1 T = TB_1$ and $K = B_1(\mathfrak{o})T(\mathfrak{o}) = T(\mathfrak{o})B_1(\mathfrak{o})$.*

Proof. All the parts of the above Proposition follow from elementary computations involving 2 by 2 matrices. Let us start with part (3). If $T_1 = T_{\alpha_1,0,1}$ and $T_2 = T_{\alpha_2,0,1}$, then there exist $m \in \mathfrak{o}^\times$ such that $\alpha_2 = m\alpha_1^2$. So $S_{\alpha_2,0,1} = a(m)S_{\alpha_1,0,1}a(m)$ and therefore $T_2 = a(m^{-1})T_1 a(m)$.

Next we prove part (2). Suppose that T is associated to a matrix S. There exists $h \in \mathrm{GL}_2(F)$ such that ${}^t hSh = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$ for some $\lambda_i \in F^\times$. Write $\lambda_1/\lambda_2 = mn^2$ with $m \in \mathfrak{o}^\times$. Then

$$\lambda_2^{-1} \begin{pmatrix} n^{-1} & \\ & 1 \end{pmatrix} {}^t hSh \begin{pmatrix} n^{-1} & \\ & 1 \end{pmatrix} = \begin{pmatrix} m & 0 \\ 0 & 1 \end{pmatrix}.$$

Consequently, we have $(h(a(n^{-1}))^{-1}T(h(a(n^{-1})) = T_{m,0,1}$. Part (1) follows from Parts (2) and (3).

Finally, let us prove part (4). For $g = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$, put

$$u_1 = \frac{\alpha(ad - bc)}{\alpha a^2 + c^2}, \quad m_1 = -\frac{ab\alpha + cd}{\alpha a^2 + c^2},$$

$$u_2 = \frac{c^2 + d^2\alpha}{\alpha(ad - bc)}, \quad m_2 = -\frac{ac + \alpha bd}{\alpha(ad - bc)}.$$

Then an easy calculation shows that $g \begin{pmatrix} u_1 & m_1 \\ & 1 \end{pmatrix} \in T$ and $\begin{pmatrix} u_2 & m_2 \\ & 1 \end{pmatrix} g \in T$. Furthermore, if $g \in K$ then it is immediate that $u_1, u_2 \in \mathfrak{o}^\times$, $m_1, m_2 \in \mathfrak{o}$. $\square$

Now let $T \subset G$ be an inert torus and let $\theta \colon E^\times \to \mathbb{C}^\times$ be a character such that $\theta|_{F^\times} = 1$. Using the isomorphism (12), we view θ as a character of T (note that this entails *fixing* a choice of square root $\sqrt{\delta}$ in $E^\times$ which we henceforth do without

comment). Let π be an irreducible admissible generic representation of G with *trivial central character*. A non-zero vector $v \in \pi$ is said to be a (T, θ)-eigenvector if

$$\pi(t)v = \theta(t)v, \quad \text{for all } t \in T.$$

It is known that the space of (T, θ)-eigenvectors in π has dimension less than or equal to 1, and it has dimension 1 if and only if the epsilon factor $\epsilon(1/2, \pi \otimes \mathrm{AI}(\theta^{-1}))$ (which is equal to ± 1) equals 1, where $\mathrm{AI}(\theta^{-1})$ is the representation of G obtained from θ^{-1} by automorphic induction; see [19, 36, 38].

The precise choice of T is unimportant, because any two inert tori are conjugate in G. If T_1, T_2 are two inert tori with $T_2 = gT_1g^{-1}$, and v_1 is a (T_1, θ)-eigenvector, then $\pi(g)v_1$ is a (T_2, θ)-eigenvector. In particular, we may assume, by taking a suitable conjugate of T, that our inert torus T is in canonical form $T = T_{\alpha,0,1}$ (see part (2) of Proposition 2.3). In this case, we have $\sqrt{\delta} = 2\sqrt{-\alpha}$. We define the shorthand notation

$$w_\alpha := \begin{pmatrix} 0 & 1 \\ -\alpha & 0 \end{pmatrix}.$$

The isomorphism (12) now reads

$$x + y\sqrt{-\alpha} \mapsto x + yw_\alpha = \begin{pmatrix} x & y \\ -\alpha y & x \end{pmatrix}. \tag{13}$$

The goal of the rest of Section 2 is to delve into a particularly important case in which (T, θ)-eigenvectors exist and explicate some remarkable properties in that case.

2.3. Compact induction and minimal vectors.

Definition 2.4. Given an inert torus $T = T_{\alpha,0,1}$ in canonical form, we define for each non-negative integer r, the congruence subgroup $K_T(r)$ of K as follows:

$$K_T(r) = \left\{ g = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in K : a - d \in \mathfrak{p}^r, \ c + b\alpha \in \mathfrak{p}^r \right\} = T(\mathfrak{o})K(r).$$

Using the calculations in the proof of Proposition 2.3, part (4) it can be seen that

$$K_T(r) = T(\mathfrak{o})B_1(r) = B_1(r)T(\mathfrak{o}). \tag{14}$$

Since $B_1(r)$ intersects T trivially, it follows that the index of $K_T(r)$ in K is $\asymp q^{2r}$.

Lemma 2.5. *Let $T = T_{\alpha,0,1}$ be an inert torus in canonical form. Let θ be a character of $E^\times$ such that $a(\theta) = 2n$ and $\theta|_{F^\times} = 1$. Then there exists $a_{\theta,T} \in \mathfrak{o}^\times$ such that $\psi_E(\varpi^{-n}a_{\theta,T}\sqrt{-\alpha}u) = \theta(1 + \varpi^n u)$ for all $u \in \mathfrak{o}_E$.*

Proof. Note that $\psi'(x) := \theta(1 + \varpi^n\sqrt{-\alpha}x)$ is an additive character on $\mathfrak{o}$. So, there must exist $y \in F$ such that $\psi'(x) = \psi(xy)$ for all $x \in \mathfrak{o}$. Comparing conductors,

we see that $v(y) = -n$. So we may put $y = -2a_{\theta,T}\alpha\varpi^{-n}$ for some $a_{\theta,T} \in \mathfrak{o}^\times$. We claim that this $a_{\theta,T}$ works. Indeed, let $u = a + b\sqrt{-\alpha} \in \mathfrak{o}_E$, with $a, b \in \mathfrak{o}$. Then $\psi_E(\varpi^{-n}a_{\theta,T}\sqrt{-\alpha}u) = \psi(-2\varpi^{-n}a_{\theta,T}\alpha b) = \psi'(b) = \theta(1 + \varpi^n\sqrt{-\alpha}b) = \theta(1 + \varpi^n u)$ where in the last step we have used that $\theta|_{F^\times} = 1$ and $a(\theta) = 2n$. $\square$

This enables the following definition.

Definition 2.6. Given an inert torus $T = T_{\alpha,0,1}$ in canonical form and a character θ on $E^\times$ (which we view as a character of T) with $a(\theta) = 2n$ and $\theta|_{F^\times} = 1$, we extend the character θ to a function $\chi_{\theta,T}$ on the group $ZK_T(n) = TK(n)$ as follows:

$$\chi_{\theta,T}(t(1 + \varpi^n g)) = \theta(t)\psi(\varpi^{-n}a_{\theta,T}\operatorname{Tr}(w_\alpha g)),$$

or equivalently

$$\chi_{\theta,T}\left(t\begin{pmatrix} 1 + \varpi^n y_1 & \varpi^n x_1 \\ \varpi^n x_2 & 1 + \varpi^n y_2 \end{pmatrix}\right) = \theta(t)\psi(\varpi^{-n}a_{\theta,T}(x_2 - \alpha x_1)).$$

To see that the above formulae are well-defined, we note that $T \cap K(n) = 1 + \varpi^n\mathfrak{o}_E$ under the identification given by (13).

Lemma 2.7. *The function $\chi_{\theta,T}$ is a multiplicative character of $ZK_T(n)$.*

Proof. First, we claim that for all $k \in K(n), t \in T$,

$$\chi_{\theta,T}(kt) = \chi_{\theta,T}(t)\chi_{\theta,T}(k) = \chi_{\theta,T}(tk). \tag{15}$$

To prove (15), write $k = 1 + \varpi^n g$ and let $g' = t^{-1}gt$ so that $kt = t(1 + \varpi^n g')$. Then

$$\begin{aligned}
\chi_{\theta,T}(kt) &= \chi_{\theta,T}(t)\chi_{\theta,T}(1 + \varpi^n g') \\
&= \chi_{\theta,T}(t)\psi(\varpi^{-n}a_{\theta,T}\operatorname{Tr}(w_\alpha t^{-1}gt)) \\
&= \chi_{\theta,T}(t)\psi(\varpi^{-n}a_{\theta,T}\operatorname{Tr}(t^{-1}w_\alpha gt)) \\
&= \chi_{\theta,T}(t)\psi(\varpi^{-n}a_{\theta,T}\operatorname{Tr}(w_\alpha g)) \\
&= \chi_{\theta,T}(t)\chi_{\theta,T}(k).
\end{aligned}$$

Next note that $\chi_{\theta,T}$ is multiplicative on the group $K(n)$ by using that $a(\theta) = 2n$. The multiplicativity of $\chi_{\theta,T}$ follows immediately by combining this fact with (15). $\square$

Remark 2.8. Using $ZK_T(n) = TB_1(n)$, we can also write $\chi_{\theta,T}$ as

$$\chi_{\theta,T}\left(t\begin{pmatrix} y & \varpi^n x \\ 0 & 1 \end{pmatrix}\right) = \theta\big(t\big(1 + \sqrt{-\alpha}\varpi^n x/2\big)\big), \quad x \in \mathfrak{o}, y \in U_n. \tag{16}$$

Further one can define the character $\chi_{\theta,T}$ on $ZK_T(n)$ directly in terms of the entries of the matrix as follows:

$$\chi_{\theta,T}\left(\begin{pmatrix} a & b \\ c & d \end{pmatrix}\right) = \theta\big((a\sqrt{-\alpha} + c)(ab\alpha + cd - 2\sqrt{-\alpha}(ad - bc))\big).$$

By [23] (see [7] for a recent treatment), the supercuspidal representations of G are obtained by compact induction from subgroups that are compact modulo Z. Precisely, let π be an irreducible supercuspidal representation of G. Then there exists a maximal compact-mod-center subgroup ZK of G, and an irreducible complex representation ξ of ZK, such that $\pi \simeq c\text{-Ind}_{ZK}^{G}\xi$ where c-Ind denotes compact induction [7, 15.5, 15.8]. As shown in [24], the representation ξ is itself induced from a smaller representation which is often one-dimensional. In the special case we are interested in, one can make all this very simple and explicit.

Proposition 2.9. *Let π be a supercuspidal representation of G with trivial central character satisfying $a(\pi) = 4n$ for some positive integer n. There exists a character θ_π of $E^\times$ with $a(\theta_\pi) = 2n$ and $\theta_\pi|_{F^\times} = 1$ such that for any inert torus T in canonical form, we have*

$$\pi \simeq c\text{-Ind}_{ZK_T(n)}^{G}\chi_{\theta_\pi,T}.$$

Proof. This follows from the results of Kutzko [23,24] but for our purposes it is more convenient to appeal to the treatment in [7, Chapter 5]. As q is odd, π is associated to a pair $(E/F, \chi)$ as in [7, 20.2] and the assumption $a(\pi) = 4n$ means that the quadratic extension E/F is unramified (and hence coincides with our setup) and furthermore that the integer $l(\chi)$ (in the terminology of [7]) equals $2n - 1$. Hence, defining $\theta_\pi = \chi$, the result follows from [7, 19.3] (note that our character $\chi_{\theta_\pi,T}$ is denoted Λ there). $\square$

Definition 2.10. Given a supercuspidal representation π of G with trivial central character satisfying $a(\pi) = 4n$ for some positive integer n, and an inert torus T in canonical form, we let χ_π denote the character $\chi_{\theta_\pi,T}$ of $ZK_T(n)$. Thus,

$$\pi \simeq c\text{-Ind}_{ZK_T(n)}^{G}\chi_\pi.$$

Remark 2.11. As q is odd, and $a(\pi)$ is even, the representation π is a dihedral supercuspidal representation associated to a character ξ_π of $E^\times$ (see, e.g. [38, Sec. 1.2]). Therefore, it is natural to ask for the relation between the characters θ_π and ξ_π. This is given by $\theta_\pi = \xi_\pi \eta_E$ where η_E is the unique unramified quadratic character on $E^\times$; see [7, 34.4].

Proposition 2.12. *Let π, θ_π be as in Proposition 2.9 and let T, χ_π be as in Definition 2.10. Then there exists a unique up to multiples element $v \in \pi$ such that*

$$\pi(k)v = \chi_\pi(k)v, \quad \text{for } k \in ZK_T(n). \tag{17}$$

In particular v is a (T, θ_π)-eigenvector.

Proof. Recall that any element in $\pi = c\text{-Ind}_{ZK_T(n)}^{G}\chi_\pi$ is a function ϕ on G such that $\phi(kg) = \chi_\pi(k)\phi(g)$ for $k \in ZK_T(n)$, with the group G acting by right translation.

In particular we can take ϕ to be the function

$$\phi(g) = \begin{cases} \chi_\pi(g), & \text{if } g \in ZK_T(n); \\ 0, & \text{otherwise.} \end{cases} \tag{18}$$

Then it's clear that

$$\pi(k)\phi = \chi_\pi(k)\phi, \quad \text{for } k \in ZK_T(n). \tag{19}$$

The uniqueness assertion follows from the general fact that the space of (T, θ)-eigen-vectors has dimension at most 1. $\qquad\square$

The above proposition allows us to make the following definition.

Definition 2.13. Let π be a supercuspidal representation of trivial central character satisfying $a(\pi) = 4n$ for some positive integer n. By a *minimal vector* in π, we will mean a nonzero vector satisfying (17) for some inert torus T in canonical form.

As we have seen, minimal vectors exist. In fact, whenever we fix an inert torus T in canonical form, there is a unique up to multiples (T, θ_π)-eigenvector; we will call such a vector a *minimal vector for T*. By part (2) of Proposition 2.3, it follows that the set of all minimal vectors (without fixing T) lie in a single $A(\mathfrak{o}^\times)$-orbit.

As it turns out, minimal vectors have remarkable properties which make them extremely special. Indeed (as pointed out in the introduction) a minimal vector may be viewed as the true non-archimedean analogue of the lowest weight vectors in (archimedean) holomorphic discrete series representations. As shown in Section 3.2, the matrix coefficient associated to such a vector has the amazing property of being a *character* on the support. This implies that the minimal vectors are those for which the associated matrix coefficient function has smallest possible support. Another important reason for singling out the (T, θ)-eigenvectors associated to the character $\theta = \theta_\pi$ is that the conductor of the degree 4 L-function $L(s, \pi \otimes \text{AI}(\theta^{-1}))$ is smallest when $\theta = \theta_\pi$.

2.4. Main results. For the rest of Section 2, we let π be a supercuspidal represent-ation of trivial central character satisfying $a(\pi) = 4n$ for some positive integer n. Note that such a representation is automatically unitary. Our first result shows that minimal vectors have a remarkably simple description in the Whittaker model.

Proposition 2.14. *Let $W_0 \in \mathcal{W}(\pi, \psi)$ be a minimal vector (with respect to some inert torus T in canonical form) in the Whittaker model of π. Then the restriction of W_0 to A is, for $\mathbf{a} = -a_{\theta,T}\alpha$ and some normalizing constant $c \in \mathbb{C}$, given by*

$$W_0(a(y)) = \begin{cases} c, & \text{if } y \in \varpi^{-2n}\mathbf{a}U_n; \\ 0, & \text{otherwise.} \end{cases}$$

Proof. We define an intertwining operator from $c\text{-Ind}_{ZK_T(n)}^G \chi_\pi$ to $\mathcal{W}(\pi, \psi)$ via

$$\phi \mapsto W_\phi(g) = \int_F \phi\left(\begin{pmatrix} -\frac{\varpi^{2n}}{a_{\theta,T}\alpha} & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} g\right)\psi(-x)dx. \qquad (20)$$

To see that the operator above is non-trivial, we compute directly the special values of the Whittaker function for the minimal vector, as defined in (18). In particular

$$W_0(a(y)) = \int_F \phi\left(\begin{pmatrix} -\frac{\varpi^{2n}}{a_{\theta,T}\alpha} & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}\begin{pmatrix} y & 0 \\ 0 & 1 \end{pmatrix}\right)\psi(-x)dx. \qquad (21)$$

Recall that $\phi(g) = 0$ unless $g \in ZK_T(n)$. So to ensure that

$$\begin{pmatrix} -\frac{\varpi^{2n}}{a_{\theta,T}\alpha} & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix}\begin{pmatrix} y & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -y\frac{\varpi^{2n}}{a_{\theta,T}\alpha} & -x\frac{\varpi^{2n}}{a_{\theta,T}\alpha} \\ 0 & 1 \end{pmatrix} \in ZK_T(n),$$

we need $v(y) = -2n$, $-y\frac{\varpi^{2n}}{a_{\theta,T}\alpha} - 1 \in \mathfrak{p}^n$ and $-x\frac{\varpi^{2n}}{a_{\theta,T}\alpha} \in \mathfrak{p}^n$. The conditions on y give $y \in -\varpi^{-2n}a_{\theta,T}\alpha U_n$. Thus $W_0(a(y)) = 0$ if $y \notin -\varpi^{-2n}a_{\theta,T}\alpha U_n$. On the other hand if $y \in -\varpi^{-2n}a_{\theta,T}\alpha U_n$,

$$\begin{pmatrix} -y\frac{\varpi^{2n}}{a_{\theta,T}\alpha} & -x\frac{\varpi^{2n}}{a_{\theta,T}\alpha} \\ 0 & 1 \end{pmatrix} = I_2 - \varpi^n\begin{pmatrix} \varpi^{-n} + y\frac{\varpi^n}{a_{\theta,T}\alpha} & x\frac{\varpi^n}{a_{\theta,T}\alpha} \\ 0 & 0 \end{pmatrix}.$$

By definition of ϕ in (18) and Definition 2.6,

$$W_0(a(y)) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (22)$$

$$= \int_{v(x)\geq -n} \psi \circ \text{tr}\left(-\varpi^{-n}\begin{pmatrix} 0 & a_{\theta,T} \\ -a_{\theta,T}\alpha & 0 \end{pmatrix}\begin{pmatrix} y\frac{\varpi^n}{a_{\theta,T}\alpha} + \varpi^{-n} & x\frac{\varpi^n}{a_{\theta,T}\alpha} \\ 0 & 0 \end{pmatrix}\right)\psi(-x)\,dx$$

$$= \int_{v(x)\geq -n} \psi\left(\text{tr}\begin{pmatrix} 0 & 0 \\ y + \varpi^{-2n}a_{\theta,T}\alpha & x \end{pmatrix}\right)\psi(-x)\,dx$$

$$= \int_{v(x)\geq -n} \psi(x)\psi(-x)\,dx \qquad\qquad\qquad\qquad\qquad\qquad\qquad (23)$$

is a non-zero constant independent of y in the support. $\qquad\square$

Remark 2.15. Recall that different inert tori in canonical form are $A(\mathfrak{o}^\times)$ conjugate. Moreover, it is well known that a vector in $\mathcal{W}(\pi, \psi)$ is uniquely specified by its restriction to A (the so-called Kirillov model). Therefore, Proposition 2.14 gives us an alternative way to characterize minimal vectors: *these are precisely those vectors which in the Kirillov model are equal to the characteristic function of $\varpi^{-2n}aU_n$ for some $a \in \mathfrak{o}^\times$.*

Remark 2.16. Using (11), it is clear that one can pick $c = (|\mathfrak{o}^\times / U_n|)^{1/2}$ in Proposition 2.14 for W_0 to be L^2−normalized.

Proposition 2.14 has some key consequences which will be crucial for our global results.

Corollary 2.17. *Let T be an inert torus in canonical form and let $W_0 \in \mathcal{W}(\pi, \psi)$ be a minimal vector for T in the Whittaker model. Let $\mathbf{a} \in \mathfrak{o}^\times$ be as in Proposition 2.14. Let $g \in G$ and (using Proposition 2.3) write $g = \left(\begin{smallmatrix} y & m \\ & 1 \end{smallmatrix} \right) t$ for $t \in T$, $y \in F^\times$, $m \in F$. Then we have*

$$\frac{W_0(g)}{\langle W_0, W_0 \rangle^{1/2}} = |\mathfrak{o}^\times / U_n|^{1/2} \times \begin{cases} \theta_\pi(t) \psi(m), & \text{if } y \in \varpi^{-2n} \mathbf{a} U_n, \\ 0, & \text{otherwise.} \end{cases} \qquad (24)$$

Proof. This is immediate as W_0 is a (T, θ_π)-eigenvector. □

Corollary 2.18. *Let $W_0 \in \mathcal{W}(\pi, \psi)$ be a minimal vector in the Whittaker model of π. Then*

$$\frac{\sup_{g \in G} |W_0(g)|}{\langle W_0, W_0 \rangle^{1/2}} \asymp q^{n/2}.$$

Proof. This is immediate from the previous corollary. □

Corollary 2.19. *Let $W_0 \in \mathcal{W}(\pi, \psi)$ be a minimal vector in the Whittaker model of π and let $k \in K$. Then there exists some $b \in \mathfrak{o}^\times / U_n$ such that*

$$\frac{|W_0(a(y)k)|^2}{|\mathfrak{o}^\times / U_n|} = \begin{cases} \langle W_0, W_0 \rangle, & \text{if } y \in \varpi^{-2n}(b + \mathfrak{p}^n), \\ 0, & \text{otherwise.} \end{cases}$$

Proof. By assumption, W_0 is a (T, θ_π)-eigenvector for some inert torus T in canonical form. Using the last part of Proposition 2.3, we can write $k = \left(\begin{smallmatrix} z & m \\ & 1 \end{smallmatrix} \right) t$ for $t \in T$, $z \in \mathfrak{o}^\times$, $m \in \mathfrak{o}$. So using Corollary 2.17 we see that $\frac{|W_0(a(y)k)|^2}{|\mathfrak{o}^\times / U_n|}$ equals $\langle W_0, W_0 \rangle$ if $y \in \varpi^{-2n} z^{-1} \mathbf{a} U_n$ and equals 0 otherwise. □

3. The QUE test vector property

Here we revisit the discussion of Section 1.3 in a local context, and establish the local results underlying the proof of Theorem 1.2.

3.1. Generalities. We continue to use the notations of the previous section. In particular, the base field F has odd residue characteristic (indeed, some of the results we will state below fail in the stated forms for even residual characteristic). Let π_1, π_2, π_3 be generic irreducible unitary representations of G with $\prod_{i=1}^3 \omega_{\pi_i} = 1$. We assume that they arise as local components of cuspidal automorphic representations; this

implies sufficient bounds towards temperedness to give the absolute convergence of the matrix coefficient integrals

$$\mathcal{H} \colon \pi_1 \otimes \pi_2 \otimes \pi_3 \to \mathbb{C}$$

$$\mathcal{H}(v_1, v_2, v_3) := \int_{g \in Z \backslash G} \langle g v_1, v_1 \rangle \langle g v_2, v_2 \rangle \langle g v_3, v_3 \rangle$$

for smooth vectors $v_i \in \pi_i$. One calls $\pi_1 \otimes \pi_2 \otimes \pi_3$ *distinguished* if $\mathcal{H}$ is not identically zero. By a result of Prasad [32],

$$\pi_1 \otimes \pi_2 \otimes \pi_3 \text{ is distinguished} \iff \epsilon(\pi_1 \otimes \pi_2 \otimes \pi_3, 1/2) = 1. \qquad (25)$$

We focus here on the case in which

$$\pi_1 = \overline{\pi_2} =: \pi \qquad (26)$$

and in which the conductor of π is large compared to that of π_3. This case is the relevant one when considering the quantum unique ergodicity (QUE) problem for global automorphic forms having $v \in \pi$ as a local component. One then encounters, after an application of Ichino's formula, the local integrals

$$\mathcal{H}(v, \overline{v}, u), \qquad (27)$$

where u is an "essentially fixed" unit vector, while either the conductor of π or the residue field cardinality of F tends off to ∞. As explained at length in [31], the size of

$$C(\pi \otimes \overline{\pi})^{1/2} \mathcal{H}(v, \overline{v}, u) \qquad (28)$$

quantifies the relative difficulty of the QUE and subconvexity problems.

When $a(\pi) = 1$ and $a(\pi_3) = 0$ and v is a newvector, it was shown in [29] that the quantity (28) has size $\asymp 1$. This corresponds globally to the QUE and subconvexity problems for a sequence of squarefree level newforms having approximately equivalent difficulty.

It was observed in [31] that if $a(\pi) \geq 2, a(\pi_3) = 0$ and $v \in \pi$ is an L^2-normalized newvector, then (28) is rather small; globally, this says that the QUE problem for newforms of powerful level is substantially *easier* than the corresponding subconvexity problem. Related results were obtained in [16] when $a(\pi_3) > 0$.

It is natural to ask whether the equivalence of difficulty in the squarefree level case may be restored in the case of powerful levels by choosing the test vector more carefully. This was shown in [30, Rmk. 30, Rmk. 50] when π belongs to the principal series by taking for v a "p-adic microlocal lift." Below we address the case in which π is supercuspidal, assuming that its conductor satisfies the congruence condition from Section 2. It turns out that a minimal vector works for this case.

3.2. Matrix coefficients of minimal vectors. In this subsection, we assume that π is a supercuspidal representation of G with trivial central character and conductor of the form $a(\pi) = 4n$ for some positive integer n. We look at the matrix coefficient associated to a minimal vector for π.

The matrix coefficients for representations before and after compact induction can be directly related; see, for example, [22]. We briefly recall this relation. Let $H \subset G$ be an open and closed subgroup containing Z with H/Z compact. Let ρ be an irreducible smooth representation of H with unitary central character and $\pi = c\text{-Ind}_H^G(\rho)$. By the assumption on H/Z, ρ is automatically unitarizable, and we shall denote the unitary pairing on ρ by $\langle \cdot, \cdot \rangle_\rho$. Then one can define the unitary pairing on π by

$$\langle \phi, \psi \rangle = \sum_{x \in H \backslash G} \langle \phi(x), \psi(x) \rangle_\rho. \tag{29}$$

If we let $y \in H \backslash G$ and $\{v_i\}$ be a basis for ρ, the elements

$$f_{y,v_i}(g) = \begin{cases} \rho(h)v_i, & \text{if } g = hy \in Hy; \\ 0, & \text{otherwise,} \end{cases}$$

form a basis for π.

Lemma 3.1. *For* $y, z \in H \backslash G$,

$$\langle \pi(g)f_{y,v_i}, f_{z,v_j} \rangle = \begin{cases} \langle \rho(h)v_i, v_j \rangle_\rho, & \text{if } g = z^{-1}hy \in z^{-1}Hy; \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

Proof. This is a direct consequence of (29) and the definition of our basis elements. $\quad\square$

Proposition 3.2. *Let* v_0 *be a minimal vector in* π *and let* $\Phi_0(g) = \frac{\langle \pi(g)v_0, v_0 \rangle}{\langle v_0, v_0 \rangle}$. *Then,*

$$\Phi_0(g) = \begin{cases} \chi_\pi(g), & \text{if } g \in ZK_T(n), \\ 0, & \text{otherwise.} \end{cases} \tag{31}$$

Proof. This follows from putting $H = ZK_T(n)$, $\rho = \chi_\pi$, and $y = z = 1$ in Lemma 3.1 and using (18). $\quad\square$

Remark 3.3. Thus, we see that the matrix coefficient of a minimal vector has the remarkable property that it is a character of its supporting group.

Corollary 3.4. *Let* v_0 *be a minimal vector and let* $\Phi_0(g) = \frac{\langle \pi(g)v_0, v_0 \rangle}{\langle v_0, v_0 \rangle}$. *Let* $\delta \asymp q^{-2n}$ *be the volume of* $K_T(n)$. *Then* $\int_{Z \backslash G} |\Phi_0(g)|^2 \, dg = \delta$. *Moreover,* $R(\overline{\Phi_0})v_0 = \delta v_0$ *and* $\Phi_0 * \Phi_0 = \delta \Phi_0$ *where we denote as usual*

$$R(\overline{\Phi_0})v := \int_{Z \backslash G} \overline{\Phi_0(g)}(\pi(g)v) \, dg, \quad (\Phi_0 * \Phi_0)(h) := \int_{Z \backslash G} \Phi_0(g^{-1})\Phi_0(gh) \, dg.$$

Proof. This follows immediately from Proposition 2.12 and Proposition 3.2. $\quad\square$

3.3. The main result.

Theorem 3.5. *Assume that π is an irreducible, admissible supercuspidal represen-tation of G with trivial central character and with conductor of the form $a(\pi) = 4n$ for some positive integer n. Let $v \in \pi$ be an L^2-normalized minimal vector. Let π_3 be an irreducible, admissible, unitary representation of $\mathrm{GL}_2(F)$ with trivial central character.*

(1) *We have $C(\pi \otimes \bar{\pi}) = q^{4n}$.*

(2) *Suppose that $u \in \pi_3$ is $K(n)$-fixed. Then*

$$\mathcal{H}(v, \bar{v}, u) = \mathrm{vol}(K_T(n)) \int_{h \in T/Z} \langle hu, u \rangle = \mathrm{vol}(K_T(n)) \int_{h \in T(\mathfrak{o})} \langle hu, u \rangle$$

where the h-integral is taken with respect to the probability Haar measure. In particular, if u is also $T(\mathfrak{o})$-fixed, then

$$C(\pi \otimes \bar{\pi})^{1/2} \mathcal{H}(v, \bar{v}, u) \asymp 1,$$

with absolute implied constants.

(3) *Assume that*

$$a(\pi) \geq 2a(\pi_3). \tag{32}$$

Then $\pi \otimes \bar{\pi} \otimes \pi_3$ is distinguished if and only if $a(\pi_3)$ is even. Furthermore, whenever $a(\pi_3)$ is even, there exists a unit vector $u \in \pi_3$ which is fixed by $K_T(n) = T(\mathfrak{o})K(n)$, and hence (by the previous part) we have

$$C(\pi \otimes \bar{\pi})^{1/2} \mathcal{H}(v, \bar{v}, u) \asymp 1.$$

Proof. In our case, as π has trivial central character, we have $\bar{\pi} \simeq \pi$. Therefore in the proof, we will replace $\bar{\pi}$ by π whenever appropriate.

First of all, π is twist-minimal by Lemma 2.1. The computations in [31, Sec. 2.6] now imply that $C(\pi \otimes \bar{\pi}) = q^{4n}$. This proves part (1). Next, using Proposition 3.2, we see that

$$\mathcal{H}(v, \bar{v}, u) = \int_{K_T(n)} \langle hu, u \rangle \, dh.$$

Note that $K_T(n) = T(\mathfrak{o})K(n)$ and by our normalization $T(\mathfrak{o})$ has volume 1. So, if u is $K(n)$-fixed, we obtain

$$\mathcal{H}(v, \bar{v}, u) = \mathrm{vol}(K_T(n)) \int_{h \in T(\mathfrak{o})} \langle hu, u \rangle$$

as required. This proves part (2) of the theorem.

We now prove part (3). First of all, we verify that

$$\pi \otimes \pi \otimes \pi_3 \text{ is distinguished} \iff a(\pi_3) \text{ is even.} \tag{33}$$

For this, we recall the three possibilities for π_3.

(i) π_3 is a principal series representation with trivial central character, hence induced by a pair $\{\chi, \chi^{-1}\}$ of characters of $F^\times$.

(ii) π_3 is a twist of the Steinberg representation by a character χ of $F^\times$ satisfying $\chi^2 = 1$.

(iii) π_3 is supercuspidal.

In case (i), the conductor exponent $a(\pi_3) = a(\chi) + a(\chi^{-1}) = 2a(\chi)$ is even. On the other hand, the self-duality of π implies that

$$\epsilon(\pi \otimes \pi \otimes \pi_3, 1/2) = \epsilon(\pi \otimes \pi \otimes \chi, 1/2)\epsilon(\pi \otimes \pi \otimes \chi^{-1}, 1/2) = 1,$$

and therefore, using the criterion (25), we see that $\pi \otimes \pi \otimes \pi_3$ is distinguished.

It remains to consider cases (ii) and (iii). We treat both cases simultaneously. Recall that the local Langlands correspondence associates to π a Weil–Deligne representation of the form $\sigma_\xi := \text{Ind}_E^F(\xi)$ for the unramified quadratic extension E/F and character ξ of $E^\times$ (cf. Remark 2.11). The fact that π has trivial central character implies that the restriction of ξ to $F^\times$ equals the unramified quadratic character on $F^\times$ (see, e.g., page 7 of [38]) and therefore

$$\xi^2(y) = \xi(x\bar{x}) = 1, \tag{34}$$

for all $x \in E^\times$, $y \in F^\times$. Furthermore, $a(\pi) = 2a(\xi)$ which leads to $a(\xi) = 2n$. We denote also by σ_3 the Weil–Deligne representation associated to π_3. By rewriting Prasad's criterion (25) in terms of Weil–Deligne representations, our task reduces to showing that

$$\epsilon(\text{Ind}_E^F(\xi) \otimes \text{Ind}_E^F(\xi) \otimes \sigma_3, 1/2) = 1 \iff a(\pi_3) \text{ is even.} \tag{35}$$

To compute these ϵ-factors, we recall (see [32, 8.1.4]) that for any even dimensional Weil–Deligne representation σ, one has

$$\epsilon(\text{Ind}_E^F(\xi) \otimes \sigma, 1/2) = \epsilon(\sigma|_E \otimes \xi, 1/2) \cdot \omega_{E/F}^{(\dim \sigma)/2}(-1). \tag{36}$$

Moreover, denoting by ξ^- the composition of ξ with the nontrivial automorphism $x \mapsto \bar{x}$ of E/F, we have

$$\text{Ind}_E^F(\xi)|_E = \xi \oplus \xi^-. \tag{37}$$

On the other hand, (34) implies that $\xi^- = \xi^{-1} = \bar{\xi}$. Thus

$$\epsilon(\text{Ind}_E^F(\xi) \otimes \text{Ind}_E^F(\xi) \otimes \sigma_3, 1/2) = \epsilon(\xi \otimes (\text{Ind}_E^F(\xi) \otimes \sigma_3)|_E, 1/2) \tag{38}$$

$$= \epsilon(\xi \otimes (\xi \oplus \bar{\xi}) \otimes \sigma_3|_E, 1/2)$$

$$= \epsilon(\xi^2 \otimes \sigma_3|_E, 1/2)\epsilon(\sigma_3|_E, 1/2). \tag{39}$$

(The first equality follows from (36) applied to the four-dimensional Weil–Deligne representation $\text{Ind}_E^F(\xi) \otimes \sigma_3$, the second from (37).)

By (32), $a(\xi) = 2n \geq a(\pi_3)$. On the other hand, as $n \geq 1$ and the residue characteristic of F is odd, we have that $a(\xi) = a(\xi^2)$. So

$$a(\xi^2) \geq a(\pi_3) > a(\pi_3)/2 + 1$$

and hence by [43, Prop. 1.7 and Lemma 3.1], the character ξ^2 appears in $\pi_3|_{E^\times}$ (where we think of $E^\times$ as a subgroup of G). So, by the main theorem of [43], we have $\epsilon(\xi^2 \otimes \sigma_3|_E, 1/2) = 1$. (Observe here that $\sigma_3|_E$ corresponds, under local Langlands, to the base change of π_3 to $\text{GL}_2(E)$). So, to finish the proof of (33), we need to show that the quantity $\epsilon(\sigma_3|_E, 1/2)$ equals 1 if and only if $a(\pi_3)$ is even. For this, first observe that $\epsilon(\sigma_3|_E, 1/2) = \epsilon(\pi_3, 1/2)\epsilon(\pi_3 \otimes \eta, 1/2)$, where η is the unique non-trivial unramified quadratic character. Now, by [38, (11)], we have $\epsilon(\pi_3 \otimes \eta, 1/2) = (-1)^{a(\pi_3)}\epsilon(\pi_3, 1/2)$ and hence $\epsilon(\sigma_3|_E, 1/2) = (-1)^{a(\pi_3)}$, as desired.

Finally, let $a(\pi_3) = 2m$ for some nonnegative integer $m \leq n$. We now take for u the Gross–Prasad test vector in [9, Prop 2.6] defined relative to the torus T. Among other properties, this vector u is invariant by $ZK_T(m)$, hence in particular by $ZK_T(n)$, as required. $\square$

4. Global cusp forms of minimal type

From now on, we move to a global setup. Throughout this section, the letter G will stand for the algebraic group GL_2. We will usually denote a non-archimedean place v by p where p is a rational prime. The set of all non-archimedean places (primes) will be denoted by $\mathbf{f}$. The archimedean place will be denoted by $v = \infty$. Let $K_\infty = \text{SO}_2(\mathbb{R})$ be the standard maximal connected compact subgroup of $G(\mathbb{R})$. We let ψ denote the unique non-trivial additive character on $\mathbb{A}$ that is unramified at all finite places and equals $e^{2\pi i x}$ at $\mathbb{R}$. We normalize the Haar measure on $\mathbb{R}$ to be the Lebesgue measure. We fix measures on all our adelic groups by taking the product of the local measures. We give all discrete groups the counting measure and thus obtain a measure on the appropriate quotient groups.

4.1. Setup and statement of sup-norm result. Let $\pi = \otimes_v \pi_v$ be an irreducible, unitary, cuspidal automorphic representation of $G(\mathbb{A})$ with trivial central character and the following additional property:

- If π_p is ramified then p is odd and π_p is a supercuspidal representation satisfying $a(\pi_p) = 4n_p$ for some positive integer n_p.

We let $\mathbf{c} \subset \mathbf{f}$ denote the set of primes where π_p is ramified. Let $N = \prod_{p \in \mathbf{c}} p^{n_p}$ and $C = N^4 = \prod_{p \in \mathbf{c}} p^{4n_p}$. Thus C is the conductor of the representation π.

Since π has trivial central character, there are two possibilities for π_∞.

Case 1. Principal series representations. In this case, $\pi_\infty \simeq \chi_1 \boxplus \chi_2$, where $\chi_1(y) = |y|^{it} \mathrm{sgn}(y)^m$, $\chi_2(y) = |y|^{-it} \mathrm{sgn}(y)^m$, with $m \in \{0, 1\}$, $t \in \mathbb{R} \cup (-\frac{i}{2}, \frac{i}{2})$. In this case, put

$$k = 0, \quad T = 1 + |t|.$$

Case 2. Holomorphic discrete series representations. In this case π_∞ is the unique irreducible subrepresentation of $\chi_1 \boxplus \chi_2$, where $\chi_1(y) = |y|^{(k-1)/2}$, $\chi_2(y) = |y|^{-(k-1)/2}$ for some positive even integer k. In this case we put

$$T = k.$$

In either case, we will call k the lowest weight. Note that $k = 0$ in Case 1. We say that a vector ϕ_∞ in π_∞ is a lowest weight vector if

$$\pi_\infty \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \phi_\infty = e^{ik\theta} \phi_\infty. \tag{40}$$

Definition 4.1. We say in what follows that a non-zero automorphic form $\phi \in V_\pi$ is of "minimal type" if ϕ is a factorizable vector $\phi = \otimes_v \phi_v$ with $\phi_v \in V_{\pi_v}$ that is lowest weight at the archimedean place and minimal at the finite places. Precisely:

(1) For all $p \in \mathbf{c}$, ϕ_p is a minimal vector in the sense of Definition 2.13.

(2) For all $p \in \mathbf{f}$, $p \notin \mathbf{c}$, ϕ_p is $G(\mathbb{Z}_p)$-invariant.

(3) ϕ_∞ is a lowest weight vector.

We define $\|\phi\|_2 = \int_{Z(\mathbb{A})G(F)\backslash G(\mathbb{A})} |\phi(g)|^2 \, dg$.

Remark 4.2. It is interesting to translate things to a classical setup. Suppose that ϕ is an automorphic form of minimal type. By definition, for each $p \in \mathbf{c}$, ϕ_p is an minimal vector with respect to some inert torus in canonical form $T_p = T_{\alpha_p, 0, 1}$ (as in Definition 2.2) where $\alpha_p \in \mathbb{Z}_p^\times$; let χ_{π_p} be the character on $Z_p K_{T_p}(n_p)$ as defined in Definition 2.6. Let D be an integer such that $D \equiv \alpha_p \pmod{p^{n_p}}$ for all $p \in \mathbf{c}$ and define the congruence subgroup $\Gamma_{T,D}(N)$ of $\mathrm{SL}_2(\mathbb{Z})$ as follows:

$$\Gamma_{T,D}(N) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) : a \equiv d \pmod{N}, \ c \equiv -bD \pmod{N} \right\}.$$

Clearly, the group $\Gamma_{T,D}(N)$ contains the principal congruence subgroup $\Gamma(N)$. Define a character χ on $\Gamma_{T,D}(N)$ by $\chi(\gamma) = \prod_{p|N} \chi_{\pi_p}^{-1}(\gamma)$. Note that χ is trivial on the principal congruence subgroup $\Gamma(N^2)$ but non-trivial on $\Gamma(Nm)$ for any $1 \leq m < N, m|N$.

Then, the function f on $\mathbb{H}$ defined by

$$f(x + iy) = y^{-k/2} \phi\left(\begin{pmatrix} y^{1/2} & xy^{-1/2} \\ & y^{-1/2} \end{pmatrix} \right)$$

has the following properties:

- If we are in Case 1, then f is a real analytic function satisfying $\Delta f = -\lambda f$ and if we are in Case 2 then f is a holomorphic function.

- For all $\gamma \in \Gamma_{T,D}(N)$, $z \in \mathbb{H}$,

$$f|_k \gamma = \chi(\gamma) f. \tag{41}$$

- f decays rapidly at the cusps.

- f is an eigenfunction of all the Hecke operators T_n for $(n, N) = 1$.

It is also clear that $\sup_{g \in G(\mathbb{A})} |\phi(g)| = \sup_{z \in \mathbb{H}} |y^{k/2} f(z)|$.

Let the real numbers $\lambda_\pi(n)$ be the coefficients of the (finite part of the) L-function attached to π, i.e.

$$L_{\mathbf{f}}(s, \pi) = \sum_{n=1}^{\infty} \frac{\lambda_\pi(n)}{n^s}. \tag{42}$$

Note that all our L-functions are normalized so that the functional equation takes $s \to 1 - s$.

Definition 4.3. We fix δ_π to be any real number such that $\lambda_\pi(n) \le d(n) n^{\delta_\pi}$ for all positive integers n where $d(n)$ is the divisor function. In particular, we may uniformly take $\delta_\pi = \frac{7}{64}$ in Case 1, and $\delta_\pi = 0$ in Case 2.

Our main result is as follows.

Theorem 4.4. *Let* $\phi \in V_\pi$ *be of minimal type and satisfy* $\|\phi\|_2 = 1$.

(1) *If we are in Case 1 then*

$$C^{\frac{1}{8}-\epsilon} T^{\frac{1}{6}-\epsilon} \ll_\epsilon \sup_{g \in G(\mathbb{A})} |\phi(g)| \ll_\epsilon C^{\frac{1}{8}+\epsilon} T^{\frac{1}{2}+\epsilon} \min\left(C^{\frac{\delta_\pi}{2}} T^{\delta_\pi}, C^{\frac{1}{32}}\right). \tag{43}$$

(2) *If we are in Case 2, then*

$$C^{\frac{1}{8}-\epsilon} k^{\frac{1}{4}-\epsilon} \ll_\epsilon \sup_{g \in G(\mathbb{A})} |\phi(g)| \ll_\epsilon C^{\frac{1}{8}+\epsilon} k^{\frac{1}{4}+\epsilon}. \tag{44}$$

We will prove this theorem by carefully looking at the Whittaker expansion. Before getting into the details of the proof, let us make a simple but key reduction. Let $\mathcal{F}$ be the subset of $B_1(\mathbb{R})^+$ defined by $\mathcal{F} := \{n(x)a(y) : x \in \mathbb{R}, \ y \ge \sqrt{3}/2\}$. Let

$$\mathcal{J}_N = \prod_{p \mid N} G(\mathbb{Z}_p).$$

Then, using strong approximation, it follows that for any $g \in G(\mathbb{A})$, the double coset $G(\mathbb{Q}) g \prod_{p \nmid N} G(\mathbb{Z}_p)$ has a representative in $\mathcal{J}_N \times \mathcal{F}$. Since ϕ is left $G(\mathbb{Q})$-invariant and right $\prod_{p \nmid N} G(\mathbb{Z}_p)$-invariant, it suffices in Theorem 4.4 to only consider the supremum for g lying in $\mathcal{J}_N \times \mathcal{F}$, i.e. $g = g_{\mathbf{f}} n(x) a(y)$ with $g_{\mathbf{f}} \in \mathcal{J}_N, n(x) a(y) \in \mathcal{F}$.

4.2. Generalities on the Whittaker expansion and proof of the lower bounds.

Let π, ϕ be as in the statement of Theorem 4.4. Let $g_{\mathbf{f}} = \prod_{v \in \mathbf{f}} g_v \in G(\mathbb{A}_{\mathbf{f}})$, $x \in \mathbb{R}$, $y \in \mathbb{R}^+$. Then the Whittaker expansion for ϕ says that

$$\phi(g_{\mathbf{f}}n(x)a(y)) = \sum_{q \in \mathbb{Q}_{\neq 0}} W_\phi(a(q)g_{\mathbf{f}}n(x)a(y)), \tag{45}$$

where W_ϕ is the global Whittaker newform corresponding to ϕ given explicitly by

$$W_\phi(g) = \int_{x \in \mathbb{A}/\mathbb{Q}} \phi(n(x)g)\psi(-x)\,dx. \tag{46}$$

For each unramified prime p, i.e. for $p \in \mathbf{f} - \mathbf{c}$, let the function $W_p(g)$ on $G(\mathbb{Q}_p)$ be equal to the unique right $G(\mathbb{Z}_p)$-invariant function in the Whittaker model of π_p normalized so that $W_{\pi_p}(1) = 1$. It is well known that for $(m, N) = 1$ we have

$$m^{1/2} \prod_{p \in \mathbf{f} - \mathbf{c}} W_p(a(m)) = \lambda_\pi(m),$$

where $\lambda_\pi(m)$ is defined by (42). For each ramified prime p, i.e. for $p \in \mathbf{c}$, let the function $W_p(g)$ on $G(\mathbb{Q}_p)$ be equal to $\frac{W_{0,p}(g)}{\langle W_{0,p}, W_{0,p} \rangle^{1/2}}$ where $W_{0,p}$ is an element corresponding to ϕ_p in the Whittaker model for π_p. The function $W_p(g)$ in this case is given explicitly by the right hand side of (24). Finally for $v = \infty$, let the function $W_\infty(g)$ on $G(\mathbb{R})$ be the element of the Whittaker model of π_∞ corresponding to ϕ_∞, normalized so that $W_\infty(a(y)) = \kappa(y)$ for all $y \in \mathbb{R}$ where

$$\kappa(y) := \begin{cases} |y|^{1/2} K_{it}(2\pi|y|)\operatorname{sgn}(y)^m & \text{in Case 1,} \\ y^{k/2}e^{-2\pi y}((1 + \operatorname{sgn}(y))/2) & \text{in Case 2.} \end{cases} \tag{47}$$

Put

$$c_\infty = \langle W_\infty, W_\infty \rangle^{1/2} = \left(\int_{\mathbb{R}^\times} |\kappa(y)|^2 \frac{dy}{|y|} \right)^{1/2}.$$

It is a well known fact (see, e.g. [40, Lemma 5.3] or [34, (27)]) that

$$\frac{\sup_{g \in G(\mathbb{R})} |W_\infty(g)|}{c_\infty} = \frac{\sup_{y>0} \kappa(y)}{c_\infty} \asymp \begin{cases} T^{1/6} & \text{in Case 1,} \\ k^{1/4} & \text{in Case 2.} \end{cases} \tag{48}$$

By Lemma 2.2.3 of [26], the function W_ϕ factors as follows. For $g_{\mathbf{f}} = \prod_{v \in \mathbf{f}} g_v \in G(\mathbb{A}_{\mathbf{f}})$, $x \in \mathbb{R}$, $y \in \mathbb{R}^+$, we have

$$W_\phi(g_{\mathbf{f}}n(x)a(y)) = \sqrt{\frac{2\zeta(2)}{L_{\mathbf{f}}(1, \pi, \operatorname{Ad})}} \times \frac{e^{2\pi ix}\kappa(y)}{c_\infty} \times \prod_{p \in \mathbf{f}} W_p(g_p), \tag{49}$$

where $L_{\mathbf{f}}(1, \pi, \operatorname{Ad}) = \prod_{p < \infty} L(1, \pi_p, \operatorname{Ad})$ denotes the finite part of the global adjoint L-function for π. By a result of Hoffstein–Lockhart [12], we have

$$(CT)^{-\epsilon} \ll_\epsilon L_{\mathbf{f}}(1, \pi, \operatorname{Ad}) \ll_\epsilon (CT)^\epsilon. \tag{50}$$

Remark 4.5. To deduce (49) from Lemma 2.2.3 of [26], note from Table 1 of [31] that $\frac{L(1,\pi_p,\mathrm{Ad})\zeta_p(1)}{\zeta_p(2)} = 1$ for all $p \in \mathbf{c}$.

Using (46), (48), (49), and (50), we conclude that

$$\sup_{g \in G(\mathbb{A})} |\phi(g)| \gg \sup_{g \in G(\mathbb{A})} |W_\phi(g)|$$

$$\gg_\epsilon (CT)^{-\epsilon} h(\pi_\infty) \prod_{p \in \mathbf{c}} \sup_{g \in G(\mathbb{Q}_p)} |W_p(g)|,$$

where $h(\pi_\infty) = T^{1/6}$ in Case 1 and $h(\pi_\infty) = k^{1/4}$ in Case 2. By Corollary 2.17, we have

$$\prod_{p \in \mathbf{c}} \sup_{g \in G(\mathbb{Q}_p)} |W_p(g)| \gg_\epsilon C^{1/8-\epsilon}.$$

This completes the proof of the lower bounds in Theorem 4.4!

Next, recall that for $(m, N) = 1$, we have $\lambda_\pi(m) = m^{1/2} \prod_{p \in \mathbf{f}-\mathbf{c}} W_p(a(m))$. From Definition 4.3, we have

$$\lambda_\pi(m) \ll_\epsilon m^{\delta_\pi + \epsilon}. \tag{51}$$

We will need the following property of the coefficients $\lambda_\pi(n)$ to get an improved bound in Case 1.

Proposition 4.6. *Let* $1 \le r \le 4$ *be an integer. Then*

$$\sum_{1 \le |n| \le X} |\lambda_\pi(n)|^{2r} \ll_\epsilon X(NTX)^\epsilon.$$

Proof. This follows by first taking the sym^r-lift of π to GL_{r+1} which is known to exist [8, 20] and then using the analytic properties of $L(s, \mathrm{sym}^r \pi \otimes \mathrm{sym}^r \bar{\pi})$. For a detailed proof in the case $r = 2$, we refer the reader to [12, Lemma 2.1]. The proofs in the other cases are essentially identical. $\qquad\square$

Let $g_\mathbf{f} \in \mathcal{J}_N$. For each $m \in \mathbb{Z}$, we define

$$\lambda'(m; g_\mathbf{f}) = \prod_{p \in \mathbf{c}} W_p\big(a(m/N^2)g_p\big).$$

By Corollary 2.19, there exists some integer $b = b(g_\mathbf{f})$ coprime to N, such that

$$|\lambda'(m; g_\mathbf{f})| = \begin{cases} \sqrt{\varphi(N)}, & \text{if } m \equiv b \pmod{N}; \\ 0, & \text{otherwise.} \end{cases} \tag{52}$$

Therefore, for any $g_{\mathbf{f}} \in \mathcal{J}_N$ and $x \in \mathbb{R}$, $y \in \mathbb{R}^+$, the expansion (45) together with the above discussion gives us:

$$
\phi(g_{\mathbf{f}}n(x)a(y)) = \sqrt{\frac{2\zeta(2)}{L_{\mathbf{f}}(1, \pi, \mathrm{Ad})}}
$$

$$
\times \frac{1}{c_\infty} \sum_{\substack{m \in \mathbb{Z} \\ m \equiv b \bmod N}} m^{-1/2} e^{2\pi imx/N^2} \kappa(my/N^2) \lambda_\pi(m) \lambda'(m; g_{\mathbf{f}}). \quad (53)
$$

In particular, the Whittaker expansion of ϕ is supported on an arithmetic progression! It is this remarkable feature that will allow us to prove a strong upper bound. As a key first step, using (50) and the triangle inequality, we note the bound

$$
|\phi(g_{\mathbf{f}}n(x)a(y))| \ll_\epsilon (CT)^\epsilon \frac{N^{1/2}}{c_\infty} \sum_{\substack{m \in \mathbb{Z} \\ m \equiv b \bmod N}} m^{-1/2} |\kappa(my/N^2)| |\lambda_\pi(m)|. \quad (54)
$$

4.3. Proof of the upper bounds.

We can now prove the upper bounds in Theorem 4.4. Throughout this subsection, let $g_{\mathbf{f}} \in \mathcal{J}_N$ and $x \in \mathbb{R}$, $y \in \mathbb{R}^+$, with $y \geq \frac{\sqrt{3}}{2}$. As noted at the end of Section 4.1, it is sufficient to restrict to $g = g_{\mathbf{f}}n(x)a(y)$ with $g_{\mathbf{f}}$, x, y as above.

Case 1. In this case we have

$$
|\kappa(my/N^2)| = N^{-1}(my)^{1/2} |K_{it}(2\pi|my|/N^2)|.
$$

By [40, Lemma 5.3], $c_\infty \gg e^{-\pi t/2}$. So (54) gives

$$
|\phi(g_{\mathbf{f}}n(x)a(y))| \ll_\epsilon (CT)^\epsilon e^{\pi t/2} \left(\frac{y}{N}\right)^{1/2} \sum_{\substack{m \in \mathbb{Z} \\ m \equiv b \bmod N}} |\lambda_\pi(m)| |K_{it}(2\pi|my|/N^2)|.
$$

$$
(55)
$$

We need to prove the following two bounds:

$$
|\phi(g_{\mathbf{f}}n(x)a(y))| \ll_\epsilon C^{\frac{1}{8}+\frac{\delta_\pi}{2}+\epsilon} T^{\frac{1}{2}+\delta_\pi+\epsilon}, \quad (56)
$$

$$
|\phi(g_{\mathbf{f}}n(x)a(y))| \ll_\epsilon C^{\frac{1}{8}+\frac{1}{32}+\epsilon} T^{\frac{1}{2}+\epsilon}. \quad (57)
$$

Let $f(y) = \min(T^{1/6}, |\frac{y}{T} - 1|^{-1/4})$. Then it is known that

$$
e^{\pi t/2} |K_{it}(y)| \ll T^{-1/2} f(y);
$$

see, e.g. [41, (3.1)]. Furthermore, the quantity $\lambda_\pi(m)|K_{it}(2\pi|my|/N^2)|$ decays exponentially for $m \gg R$ where $R = N^{2+\epsilon}(T + T^{1/3+\epsilon})/2\pi y$. Therefore, if $R \ll 1$,

the right side of (55) is negligible. So we henceforth assume that $R \gg 1$, i.e. $y \ll N^{2+\epsilon} T^{1+\epsilon}$. Furthermore, for the same reason, we can restrict the sum in (55) to $|m| < R$.

Let us now prove (56). We obtain from (51) and (55)

$$
|\phi(g_{\mathfrak{f}} n(x) a(y))|
$$

$$
\ll_\epsilon (CT)^\epsilon T^{-1/2} \Big(\frac{y}{N}\Big)^{1/2} \sum_{\substack{1 \leq |m| \leq R \\ m \equiv b \bmod N}} m^{\delta_\pi + \epsilon} f(2\pi|my|/N^2)
$$

$$
\ll_\epsilon (CT)^\epsilon T^{-1/2} \Big(\frac{y}{N}\Big)^{1/2} N^{\delta_\pi} \sum_{\substack{0 < |m| \leq R/N \\ m \in \frac{b}{N} + \mathbb{Z}}} m^{\delta_\pi + \epsilon} f(2\pi|my|/N)
$$

$$
\ll_\epsilon (CT)^\epsilon T^{-\frac{1}{2}} y^{\frac{1}{2}} N^{\delta_\pi - \frac{1}{2}} \max\left(1, (R/N)^{\delta_\pi}\right) \left(T^{\frac{1}{6}} + \int_0^{\frac{R}{N}} \left|\frac{2\pi xy}{NT} - 1\right|^{-\frac{1}{4}} dx\right)
$$

$$
\ll_\epsilon (CT)^\epsilon T^{-\frac{1}{2}} y^{\frac{1}{2}} N^{\delta_\pi - \frac{1}{2}} \max\left(1, (R/N)^{\delta_\pi}\right) \left(T^{\frac{1}{6}} + \frac{NT}{y}\right)
$$

$$
\ll_\epsilon (CT)^\epsilon N^{\frac{1}{2} + 2\delta_\pi} T^{\frac{1}{2} + \delta_\pi},
$$

where in the last step we have used $1 \ll y \ll N^2 T^{1+\epsilon}$. This completes the proof of (56) since $N^{\frac{1}{2} + 2\delta_\pi} = C^{\frac{1}{8} + \frac{\delta_\pi}{2}}$.

Let us now prove (57). We obtain from Proposition 4.6 and (55), together with Holder's inequality:

$$
|\phi(g_{\mathfrak{f}} n(x) a(y))| \ll_\epsilon (CT)^\epsilon T^{-1/2} \Big(\frac{y}{N}\Big)^{1/2}
$$

$$
\times \left(\sum_{1 \leq |m| \leq R} |\lambda_\pi(m)|^8\right)^{1/8} \times \left(\sum_{\substack{0 < |m| \leq R \\ m \equiv b \bmod N}} f(2\pi|my|/N^2)^{8/7}\right)^{7/8}
$$

$$
\ll_\epsilon (CT)^\epsilon T^{-1/2} \Big(\frac{y}{N}\Big)^{1/2} R^{1/8} \left(\sum_{\substack{1 \leq |m| \leq R/N \\ m \in \frac{b}{N} + \mathbb{Z}}} f(2\pi|my|/N)^{8/7}\right)^{7/8}
$$

$$
\ll_\epsilon (CT)^\epsilon T^{-\frac{1}{2}} y^{\frac{1}{2}} N^{-\frac{1}{2}} R^{1/8} \left(T^{\frac{4}{21}} + \int_0^{\frac{R}{N}} \left|\frac{2\pi xy}{NT} - 1\right|^{-\frac{2}{7}} dx\right)^{7/8}
$$

$$
\ll_\epsilon (CT)^\epsilon T^{-\frac{3}{8}} y^{\frac{3}{8}} N^{-\frac{1}{4}} \left(T^{\frac{1}{6}} + \Big(\frac{NT}{y}\Big)^{7/8}\right)
$$

$$
\ll_\epsilon (CT)^\epsilon N^{\frac{5}{8}} T^{\frac{1}{2}},
$$

which is equivalent to (57).

Case 2. In this case, we have $\delta_\pi = 0$ and

$$|\kappa(my/N^2)| = N^{-k}(my)^{k/2}e^{-2\pi my/N^2}.$$

By [40, Lemma 5.3], c_∞ equals $(4\pi)^{-k/2}\Gamma(k)^{1/2}$. So (54) gives

$$|\phi(g_\mathfrak{f}n(x)a(y))|$$

$$\ll_\epsilon (CT)^\epsilon \frac{(4\pi y)^{k/2}N^{-k+1/2}}{\Gamma(k)^{1/2}} \sum_{\substack{m\in\mathbb{Z} \\ m\equiv b \bmod N}} e^{-2\pi my/N^2}m^{(k-1)/2+\epsilon}$$

$$\ll_\epsilon (CT)^\epsilon \left(\frac{y}{N}\right)^{1/2}\frac{2^{k/2}}{\Gamma(k)^{1/2}} \sum_{n\in\frac{b}{N}+\mathbb{N}} e^{-2\pi ny/N}(2\pi ny/N)^{(k-1)/2+\epsilon}.$$

To estimate the above sum we proceed similarly to [45]. Indeed, if we take the relevant series in [45] and replace $y \mapsto y/N$, $k \mapsto k/2$, and take the summation over $b/N + \mathbb{Z}_{\geq 0}$ instead of $\mathbb{Z}_{>0}$, we get our series above. Observe also that the function $\xi \mapsto e^{-2\pi\xi y/N}(2\pi\xi y/N)^{(k-1)/2+\epsilon}$ obtains its maximum at

$$\xi = \frac{k/2 - 1/2 + \epsilon}{2\pi(y/N)}.$$

So the general term in the series above is decreasing if $\xi < b/N$, i.e. if $\frac{y}{N} \gg \frac{Nk}{b}$. Now the argument of [45], *mutatis mutandis*, leads to

$$|\phi(g_\mathfrak{f}n(x)a(y))| \ll_\epsilon \begin{cases} (CT)^\epsilon\left(\frac{k^{1/4+\epsilon}}{(y/N)^{1/2}} + \frac{k^\epsilon(y/N)^{1/2}}{k^{1/4}}\right) & \text{if } \frac{y}{N} \ll \frac{Nk}{b}, \\ (CT)^\epsilon\left(\frac{k^{1/4+\epsilon}}{(y/N)^{1/2}} + \frac{k^{1/4+\epsilon}N^{1/2}}{b^{1/2}}\right) & \text{if } \frac{y}{N} \gg \frac{Nk}{b}. \end{cases} \tag{58}$$

As $y \gg 1$, in either case we have

$$|\phi(g_\mathfrak{f}n(x)a(y))| \ll_\epsilon C^{1/8+\epsilon}k^{1/4+\epsilon},$$

completing the proof in Case 2.

4.4. The proof of Theorem 1.2. We explain in this final subsection the proof of Theorem 1.2. Let the notations be as in the statement of Theorem 1.2 and let $\sigma_g = \otimes_p \sigma_p$. Ichino's generalization of Watson's formula [17] reads

$$\frac{\left|\int_{\Gamma\backslash\mathbb{H}} g(z)|f|^2(z)y^k\frac{dx\,dy}{y^2}\right|^2}{\left(\int_{\mathrm{SL}_2(\mathbb{Z})\backslash\mathbb{H}}|g|^2(z)\frac{dx\,dy}{y^2}\right)\left(\int_{\Gamma\backslash\mathbb{H}}|f|^2(z)y^k\frac{dx\,dy}{y^2}\right)^2}$$

$$= \frac{1}{8}\frac{\Lambda(\pi\times\pi\times\sigma_g, 1/2)}{\Lambda(\mathrm{ad}\,\sigma_g, 1)\Lambda(\mathrm{ad}\,\pi, 1)^2}I_\infty\prod_{p|C}I_p.$$

By an explicit calculation, the archimedean quantity I_∞ is equal to 1 in our case; see [44]. The local quantities I_p are defined for each prime $p|C$ as follows:

$$I_p = \left(\frac{L(\pi_p \times \pi_p \times \sigma_p, 1/2)\zeta_p(2)^2}{L(\operatorname{ad}\sigma_p, 1)L(\operatorname{ad}\pi_p, 1)^2} \right)^{-1} \mathcal{H}_p(v_p, \overline{v}_p, u_p),$$

where we are using the notation of Section 3, and v_p denotes the minimal vector in π_p, and u_p denotes the unramified vector in σ_p. In particular, u_p satisfies the condition in part (2) of Theorem 3.5 and therefore we have

$$\mathcal{H}_p(v_p, \overline{v}_p, u_p) \operatorname{Cond}(\pi_\mathfrak{p} \times \pi_p)^{1/2} \asymp 1.$$

On the other hand, it follows from well-known bounds on the Satake parameters that

$$\frac{L(\pi_p \times \pi_p \times \sigma_p, 1/2)\zeta_p(2)^2}{L(\operatorname{ad}\sigma_p, 1)L(\operatorname{ad}\pi_p, 1)^2} \asymp 1.$$

Therefore $I_p\operatorname{Cond}(\pi_\mathfrak{p} \times \pi_p)^{1/2} \asymp 1$ as required.

References

[1] J. Bernstein and A. Reznikov, Subconvexity bounds for triple L-functions and representation theory, *Ann. of Math. (2)*, **172** (2010), no. 3, 1679–1718. Zbl 1225.11068 MR 2726097

[2] V. Blomer and G. Harcos, The spectral decomposition of shifted convolution sums, *Duke Math. J.*, **144** (2008), no. 2, 321–339. Zbl 1246.11108 MR 2437682

[3] V. Blomer, G. Harcos, and D. Milićević, Bounds for eigenforms on arithmetic hyperbolic 3-manifolds, *Duke Math. J.*, **165** (2016), no. 4, 625–659. Zbl 1339.11062 MR 3474814

[4] V. Blomer, G. Harcos, P. Maga, and D. Milićević, The sup-norm problem for GL(2) over number fields, *J. Eur. Math. Soc. (JEMS)*, 30 August, 2019. DOI: 10.4171/JEMS/916

[5] V. Blomer and R. Holowinsky. Bounding sup-norms of cusp forms of large level, *Invent. Math.*, **179** (2010), no. 3, 645–681. Zbl 1243.11059 MR 2587342

[6] C. J. Bushnell, Hereditary orders, Gauss sums and supercuspidal representations of GL_N, *J. Reine Angew. Math.*, **375/376** (1987), 184–210. Zbl 0601.12025 MR 882297

[7] C. J. Bushnell and G. Henniart, *The Local Langlands Conjecture for* GL(2), Grundlehren der Mathematischen Wissenschaften, 335, Springer-Verlag, Berlin, 2006. Zbl 1100.11041 MR 2234120

[8] S. Gelbart and H. Jacquet, A relation between automorphic representations of GL(2) and GL(3), *Ann. Sci. École Norm. Sup. (4)*, **11** (1978), no. 4, 471–542. Zbl 0406.10022 MR 533066

[9] B. H. Gross and D. Prasad, Test vectors for linear forms, *Math. Ann.*, **291** (1991), no. 2, 343–355. Zbl 0768.22004 MR 1129372

[10] G. Harcos and N. Templier, On the sup-norm of Maass cusp forms of large level. II, *Int. Math. Res. Not. IMRN*, (2012), no. 20, 4764–4774. Zbl 1332.11048 MR 2989618

[11] G. Harcos and N. Templier, On the sup-norm of Maass cusp forms of large level. III, *Math. Ann.*, **356** (2013), no. 1, 209–216. Zbl 1332.11049 MR 3038127

[12] J. Hoffstein and P. Lockhart, Coefficients of Maass forms and the Siegel zero. With an appendix by Dorian Goldfeld, Hoffstein, and Daniel Lieman, *Ann. of Math. (2)*, **140** (1994), no. 1, 161–181. Zbl 0814.11032 MR 1289494

[13] R. Holowinsky and K. Soundararajan, Mass equidistribution for Hecke eigenforms, *Ann. of Math. (2)*, **172** (2010), no. 2, 1517–1528. Zbl 1211.11050 MR 2680499

[14] R. E. Howe, Some qualitative results on the representation theory of Gl_n over a p-adic field, *Pacific J. Math.*, **73** (1977), no. 2, 479–538. Zbl 0385.22009 MR 492088

[15] R. E. Howe, Tamely ramified supercuspidal representations of Gl_n, *Pacific J. Math.*, **73** (1977), no. 2, 437–460. Zbl 0404.22019 MR 492087

[16] Y. Hu, Triple product formula and mass equidistribution on modular curves of level N, *Int. Math. Res. Not. IMRN*, (2018), no. 9, 2899–2943. Zbl 07013470 MR 3801500

[17] A. Ichino, Trilinear forms and the central values of triple product L-functions, *Duke Math. J.*, **145** (2008), no. 2, 281–307. Zbl 1222.11065 MR 2449948

[18] H. Iwaniec and P. Sarnak, L^∞ norms of eigenfunctions of arithmetic surfaces, *Ann. of Math. (2)*, **141** (1995), no. 2, 301–320. Zbl 0838.11058 MR 1324136

[19] H. Jacquet and R. P. Langlands, *Automorphic forms on* GL(2), Lecture Notes in Mathematics, 114, Springer-Verlag, Berlin, 1970. Zbl 0236.12010 MR 401654

[20] H. H. Kim, Functoriality for the exterior square of GL_4 and the symmetric fourth of GL_2. With Appendix 1 by Dinakar Ramakrishnan and Appendix 2 by Kim and Peter Sarnak, *J. Amer. Math. Soc.*, **16** (2003), no. 1, 139–183 (electronic). Zbl 1018.11024 MR 1937203

[21] A. W. Knapp, Local Langlands correspondence: the Archimedean case, in *Motives (Seattle, WA, 1991)*, 393–410, Proc. Sympos. Pure Math., 55, Amer. Math. Soc., Providence, RI, 1994. Zbl 0811.11071 MR 1265560

[22] A. Knightly and C. Ragsdale, Matrix coefficients of depth-zero supercuspidal representations of GL(2), *Involve. A Journal of Mathematics*, **7** (2014), no. 5, 669–690. Zbl 1297.22019 MR 3245843

[23] P. C. Kutzko, On the supercuspidal representations of Gl_2, *Amer. J. Math.*, **100** (1978), no. 1, 43–60. Zbl 0417.22012 MR 507253

[24] P. C. Kutzko, On the supercuspidal representations of Gl_2. II, *Amer. J. Math.*, **100** (1978), no. 4, 705–716. Zbl 0421.22012 MR 507254

[25] S. Marshall, Local bounds for L^p norms of Maass forms in the level aspect, *Algebra Number Theory*, **10** (2016), no. 4, 803–812. Zbl 1382.35175 MR 3519096

[26] P. Michel and A. Venkatesh, The subconvexity problem for GL_2, *Publ. Math. Inst. Hautes Études Sci.*, (2010), no. 111, 171–271. Zbl 1376.11040 MR 2653249

[27] A. Moy, A conjecture on minimal K-types for GL_n over a p-adic field, in *Representation theory and number theory in connection with the local Langlands conjecture (Augsburg, 1985)*, 249–254, Contemp. Math., 86, Amer. Math. Soc., Providence, RI, 1989. Zbl 0685.22009 MR 987030

[28] A. Moy and G. Prasad, Unrefined minimal K-types for p-adic groups, *Invent. Math.*, **116** (1994), no. 1-3, 393–408. Zbl 0804.22008 MR 1253198

[29] P. D. Nelson, Equidistribution of cusp forms in the level aspect, *Duke Math. J.*, **160** (2011), no. 3, 467–501. Zbl 1273.11069 MR 2852367

[30] P. D. Nelson, Microlocal lifts and quantum unique ergodicity on $\mathrm{GL}_2(\mathbb{Q}_p)$, *Algebra Number Theory*, **12** (2018), no. 9, 2033?2064. Zbl 1406.58019 MR 3894428

[31] P. D. Nelson, A. Pitale, and A. Saha, Bounds for Rankin-Selberg integrals and quantum unique ergodicity for powerful levels, *J. Amer. Math. Soc.*, **27** (2014), no. 1, 147–191. Zbl 1322.11051 MR 3110797

[32] D. Prasad, Trilinear forms for representations of GL(2) and local ϵ-factors, *Compositio Math.*, **75** (1990), no. 1, 1–46. Zbl 0731.22013 MR 1059954

[33] Z. Rudnick and P. Sarnak, The behaviour of eigenstates of arithmetic hyperbolic manifolds, *Comm. Math. Phys.*, **161** (1994), no. 1, 195–213. Zbl 0836.58043 MR 1266075

[34] A. Saha, Large values of newforms on GL(2) with highly ramified central character, *Int. Math. Res. Not. IMRN*, (2016), no. 13, 4103–4131. Zbl 1404.11055 MR 3544630

[35] A. Saha, Hybrid sup-norm bounds for Maass newforms of powerful level, *Algebra and Number Theory*, **11** (2017), no. 5, 1009–1045. Zbl 06748165 MR 3671430

[36] H. Saito, On Tunnell's formula for characters of GL(2), *Compositio Math.*, **85** (1993), no. 1, 99–108. Zbl 0795.22009 MR 1199206

[37] P. Sarnak, Recent Progress on QUE, 2009. Available at: http://www.math.princeton.edu/sarnak/SarnakQUE.pdf

[38] R. Schmidt, Some remarks on local newforms for GL(2), *J. Ramanujan Math. Soc.*, **17** (2002), no. 2, 115–147. Zbl 0997.11040 MR 1913897

[39] N. Templier, On the sup-norm of Maass cusp forms of large level, *Selecta Math. (N.S.)*, **16** (2010), no. 3, 501–531. Zbl 1260.11032 MR 2734340

[40] N. Templier, Large values of modular forms, *Camb. J. Math.*, **2** (2014), no. 1, 91–116. Zbl 1307.11062 MR 3272013

[41] N. Templier, Hybrid sup-norm bounds for Hecke–Maass cusp forms, *J. Eur. Math. Soc. (JEMS)*, **17** (2015), no. 8, 2069–2082. Zbl 1376.11030 MR 3372076

[42] J. B. Tunnell, On the local Langlands conjecture for $GL(2)$, *Invent. Math.*, **46** (1978), no. 2, 179–200. Zbl 0385.12006 MR 476703

[43] J. B. Tunnell, Local ϵ-factors and characters of GL(2), *Amer. J. Math.*, **105** (1983), no. 6, 1277–1307. Zbl 0532.12015 MR 721997

[44] T. C. Watson, Rankin triple products and quantum chaos, 2008. arXiv:0810.0425

[45] H. Xia, On L^∞ norms of holomorphic cusp forms, *J. Number Theory*, **124** (2007), no. 2, 325–327. Zbl 1121.11038 MR 2321365

Y. Hu, Department of Mathematics, ETH Zürich,
Raemistrasse 101, 8092 Zürich, Switzerland

E-mail: huyueke2012@gmail.com

P. D. Nelson, Department of Mathematics, ETH Zürich,
Raemistrasse 101, 8092 Zürich, Switzerland

E-mail: paul.nelson@math.ethz.ch

A. Saha, School of Mathematical Sciences, Queen Mary University of London,
London E1 4NS, UK

E-mail: abhishek.saha@qmul.ac.uk