

# Sup-norms of eigenfunctions in the level aspect for compact arithmetic surfaces

Abhishek Saha

ABSTRACT. Let  $D$  be an indefinite quaternion division algebra over  $\mathbb{Q}$ . We approach the problem of bounding the sup-norms of automorphic forms  $\phi$  on  $D^\times(\mathbb{A})$  that belong to irreducible automorphic representations and transform via characters of unit groups of orders of  $D$ . We obtain a non-trivial upper bound for  $\|\phi\|_\infty$  in the level aspect that is valid for *arbitrary* orders. This generalizes and strengthens previously known upper bounds for  $\|\phi\|_\infty$  in the setting of newforms for Eichler orders. In the special case when the index of the order in a maximal order is a *squarefull* integer  $N$ , our result specializes to  $\|\phi\|_\infty \ll_{\pi_\infty, \epsilon} N^{1/3+\epsilon} \|\phi\|_2$ .

A key application of our result is to automorphic forms  $\phi$  which correspond at the ramified primes to either minimal vectors (in the sense of [11]), or  $p$ -adic microlocal lifts (in the sense of [15]). For such forms, our bound specializes to  $\|\phi\|_\infty \ll_\epsilon C^{\frac{1}{6}+\epsilon} \|\phi\|_2$  where  $C$  is the conductor of the representation  $\pi$  generated by  $\phi$ . This improves upon the previously known *local bound*  $\|\phi\|_\infty \ll_{\lambda, \epsilon} C^{\frac{1}{4}+\epsilon} \|\phi\|_2$  in these cases.

## 1. INTRODUCTION

**1.1. Background.** Let  $\phi = \otimes_v \phi_v$  be a cuspidal automorphic form on  $D^\times(\mathbb{A})$  where  $D$  is an indefinite quaternion algebra over  $\mathbb{Q}$ . The sup-norm problem, which has seen a lot of interest recently, is concerned with bounding  $\frac{\|\phi\|_\infty}{\|\phi\|_2}$  in terms of natural parameters of  $\phi$ . When the primary focus is the dependance of the bound on parameters related to the ramified primes or to the underlying level structures associated to  $\phi$  (with the dependance on the archimedean parameters suppressed), this is known as the level-aspect sup-norm problem.

In the split case where  $D^\times = \mathrm{GL}_2$ , this problem has been heavily studied in the special case that  $\phi$  is a global *newform* [5, 21, 7, 8, 22, 18, 17]. In this case,  $\phi$  transforms by a character of the subgroup<sup>1</sup>  $\Gamma_0(N)$  of norm 1 units in the standard Eichler order of level  $N$ , where  $N$  equals the conductor of the representation  $\pi$  generated by  $\phi$ . The best upper bound currently known in the case of newforms on  $\mathrm{GL}_2$  is due to the present author [17] and in the *trivial character* case this bound reads<sup>2</sup>

$$(1) \quad \|\phi\|_\infty \ll_{\pi_\infty, \epsilon} N_0^{1/6+\epsilon} N_1^{1/3+\epsilon} \|\phi\|_2,$$

for any  $\epsilon > 0$ , where we write  $N = N_0 N_1$  with  $N_0$  the largest integer such that  $N_0^2$  divides  $N$ . More recently, the sup-norm problem has also been considered for newforms on  $\mathrm{GL}_2$  over number fields, we refer the reader to [4, 1] for this.

<sup>1</sup>Technically, we need to consider the adelic counterpart of this subgroup.

<sup>2</sup>As usual, 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  $|A| \leq C|B|$ .

In the compact situation where  $D$  is a *division* algebra, less work has been done. As in the split case, the *trivial bound* in the level aspect is

$$(2) \quad \|\phi\|_\infty \ll_{\pi_\infty, \epsilon} N^{1/2+\epsilon} \|\phi\|_2,$$

for any automorphic form  $\phi$  that transforms by a character of the unit group of an order of level  $N$  (see below for a more precise definition of this). The first improvement in this setting was due to Templier [20, 21], who proved that for  $\phi$  a global newform with respect to an Eichler order of level  $N$ , one has the bound

$$(3) \quad \|\phi\|_\infty \ll_{D, \pi_\infty, \epsilon} N^{11/24+\epsilon} \|\phi\|_2.$$

More recently, Marshall [14] proved the bound (again, only in the setting of newforms for Eichler orders of level  $N$ ):

$$(4) \quad \|\phi\|_\infty \ll_{D, \pi_\infty} \left( N_1 \prod_{p|N} (1+p^{-1}) \right)^{1/2} \|\phi\|_2,$$

with  $N_1$  as in (1). (It was noted in [14] that the same bound also holds in the split case of  $D^\times = \mathrm{GL}_2$ , provided one restricts the domain to a fixed compact set.) Note that Marshall's bound is better than Templier's when  $N$  is squarefull, but worse when  $N$  is squarefree. This reflects the fact that Marshall's bound is the local bound, which essentially coincides with the trivial bound for newforms of squarefree level, but is stronger than the trivial bound in general. We discuss this distinction in more detail in Section 1.4; see also Remark 3.2 for a more precise formulation.

**1.2. The main result.** For the rest of this paper, let  $D$  be a fixed indefinite quaternion division algebra over  $\mathbb{Q}$ . We now describe (a simplified version of) the main result of this paper, which deals with the compact case and for the first time, improves upon the trivial bound (2) for completely arbitrary orders. Let  $\mathcal{O}^{\max}$  denote a maximal order of  $D$ . For any order  $\mathcal{O} \subseteq \mathcal{O}^{\max}$  of  $D$  define the level

$$N_{\mathcal{O}} = [\mathcal{O}^{\max} : \mathcal{O}].$$

Given an order  $\mathcal{O} \subseteq \mathcal{O}^{\max}$ , define the (adelic) congruence subgroup  $K_{\mathcal{O}}$  of  $D^\times(\mathbb{A}_{\mathbf{f}})$  (where  $\mathbb{A}_{\mathbf{f}}$  denotes the finite adeles) by

$$K_{\mathcal{O}} = \prod_p \mathcal{O}_p^\times$$

where the product is taken over all primes, and where we denote  $\mathcal{O}_p = \mathcal{O} \otimes_{\mathbb{Z}} \mathbb{Z}_p$ .

Let  $\pi$  be an irreducible automorphic representation of  $D^\times(\mathbb{A})$  with unitary central character. Since we are in the division algebra case,  $\pi$  is unitary and cuspidal. By the multiplicity one theorem, we can uniquely identify  $V_\pi$  with a certain space of automorphic forms on  $D^\times(\mathbb{A})$ . For  $\phi \in V_\pi$ , we define as usual

$$\|\phi\|_2 = \int_{D^\times \backslash D^\times(\mathbb{A})/\mathbb{A}^\times} |\phi(g)|^2 dg.$$

Let  $\mathcal{O} \subseteq \mathcal{O}^{\max}$  be an order and  $\chi$  be a character of  $K_{\mathcal{O}}$ . Note that we may write  $\chi = \prod_p \chi_p$  where  $p$  traverses the primes, and  $\chi_p$  is a character of  $\mathcal{O}_p^\times$  with  $\chi_p$  trivial for almost all  $p$ . The compactness of  $K_{\mathcal{O}}$  and the continuity<sup>3</sup> of  $\chi$  automatically

---

<sup>3</sup>All our characters are assumed to be continuous.

implies that  $\chi$  is of finite order. We define  $V_\pi(K_{\mathcal{O}}, \chi) \subseteq V_\pi$  to be the subspace of functions  $\phi \in V_\pi$  that have the transformation property

$$(5) \quad \phi(gk) = \chi(k)\phi(g) \text{ for all } k \in K_{\mathcal{O}}, g \in D^\times(\mathbb{A}).$$

Given any non-zero  $\phi \in V_\pi(K_{\mathcal{O}}, \chi)$ , we wish to bound the sup-norm  $\frac{\|\phi\|_\infty}{\|\phi\|_2}$  in terms of the level  $N_{\mathcal{O}}$ .

As is clear from the earlier discussion, previous work on this topic has been focussed on the case where  $\phi$  is a newform and  $\mathcal{O}$  is an Eichler order with level  $N_{\mathcal{O}}$  equal to the conductor of  $\pi$ . This restriction to newforms and Eichler orders is quite limiting as it does not capture the behavior of several natural families of automorphic forms. For example, there is an emerging theory of automorphic forms of minimal type [11, 9, 10, 12]; such forms transform naturally with respect to characters of unit groups of certain *non-Eichler* Bass orders. The aim of this paper is to prove for the first time a non-trivial upper bound for the sup-norm in the case of general orders.

**Theorem A.** (see Theorem 1) *Let  $\mathcal{O}$  be an order of  $D$  and denote  $N = [\mathcal{O}^{\max} : \mathcal{O}]$ . Let  $N_1$  be as in (1). Let  $\phi \in V_\pi(K_{\mathcal{O}}, \chi)$  where  $\pi$  is an irreducible automorphic representation of  $D^\times(\mathbb{A})$  with unitary central character, and  $\chi$  is a character of  $K_{\mathcal{O}}$ . Suppose that  $\phi_\infty$  corresponds to the vector of lowest non-negative weight<sup>4</sup> in  $\pi_\infty$ . Then we have*

$$\|\phi\|_\infty \ll_{D, \pi_\infty, \epsilon} N^\epsilon \min(\max(N^{1/3}, N_1^{1/2}), N^{11/24}) \|\phi\|_2.$$

For squarefree  $N$ , our Theorem implies that  $\|\phi\|_\infty \ll_{D, \pi_\infty, \epsilon} N^{11/24+\epsilon} \|\phi\|_2$ . However, when  $N$  is squarefull (i.e., every prime that divides  $N$  does so with exponent at least 2) Theorem A gives a much stronger ‘‘Weyl-type’’ exponent.

**Corollary.** *Let the notations and assumptions be as in Theorem A and assume that  $N$  is squarefull. Then we have*

$$\|\phi\|_\infty \ll_{D, \pi_\infty, \epsilon} N^{\frac{1}{3}+\epsilon} \|\phi\|_2.$$

For an explanation why we get such different exponents for squarefree and squarefull levels, see Section 1.5 of this introduction. An interesting fact is that this squarefree/squarefull dichotomy in the sup-norm bound is also present in the case of newforms on  $\mathrm{GL}_2$  (see Section 1.3 of [17]), but for utterly different reasons!

**Remark 1.1.** *Our main result (Theorem 1) is significantly more general than Theorem A above, because it does not require that the space generated by  $\phi$  under the action of  $K_{\mathcal{O}}$  is one-dimensional.*

**1.3. A classical reformulation.** In this subsection, we reformulate Theorem A in the language of Hecke-Laplace eigenfunctions on the upper-half plane, which may be helpful for those who are more familiar with the classical language.

We let  $d$  denote the reduced discriminant of  $D$  and we fix an isomorphism  $\iota_\infty : D \otimes_{\mathbb{Q}} \mathbb{R} \simeq M_2(\mathbb{R})$ . This leads to an embedding  $D \hookrightarrow M_2(\mathbb{R})$  which we also denote by  $\iota_\infty$ .

---

<sup>4</sup>This assumption on  $\phi_\infty$  is merely for convenience; our result can be stated without this assumption but then the implied constant will depend on  $\phi_\infty$ .

For any order  $\mathcal{O}$  of  $D$ , associate a discrete subgroup  $\Gamma_{\mathcal{O}}$  of  $\mathrm{SL}_2(\mathbb{R})$  as follows:

$$\Gamma_{\mathcal{O}} = \{\gamma \in \iota_{\infty}(\mathcal{O}), \det(\gamma) = 1\}.$$

Note that  $\Gamma_{\mathcal{O}} \backslash \mathbb{H}$  is a compact hyperbolic surface.

Now, let  $\chi = \prod_p \chi_p$  be a unitary character of  $K_{\mathcal{O}}$  as in the previous subsection. We have the identity

$$(6) \quad \iota_{\infty}^{-1}(\Gamma_{\mathcal{O}}) = D^{\times}(\mathbb{Q}) \cap D^{\times}(\mathbb{R})^+ K_{\mathcal{O}},$$

where  $D^{\times}(\mathbb{R})^+$  consists of the elements of  $D^{\times}(\mathbb{R}) \simeq \mathrm{GL}_2(\mathbb{R})$  with positive reduced norm (positive determinant). We extend  $\chi$  to  $D^{\times}(\mathbb{R})^+ K_{\mathcal{O}}$  by making it trivial on  $D^{\times}(\mathbb{R})^+$ . Hence (6) allows us to define a character on  $\Gamma_{\mathcal{O}}$  which (by abusing notation) we also denote by  $\chi$ . Let  $N_{\chi}$  be an integer such that

$$(7) \quad \mathcal{O}^{\max}(N_{\chi}) := \mathbb{Z} + N_{\chi} \mathcal{O}^{\max} \subseteq \mathcal{O}, \quad \chi_p|_{1+N_{\chi}\mathcal{O}_p^{\max}} = 1 \text{ for all primes } p.$$

(The existence of  $N_{\chi}$  is clear.) The character  $\chi : \Gamma_{\mathcal{O}} \rightarrow S^1$  is trivial on the principal congruence subgroup  $\Gamma_{\mathcal{O}^{\max}(N_{\chi})}$ , which is a normal subgroup of  $\Gamma_{\mathcal{O}}$ . In particular,  $\chi$  is a *congruence character*, i.e., it is trivial on a principal congruence subgroup.

We let  $C^{\infty}(\Gamma_{\mathcal{O}} \backslash \mathbb{H}, \chi^{-1})$  denote the space of smooth functions  $f : \mathbb{H} \rightarrow \mathbb{C}$  such that

$$f(\gamma z) = \chi^{-1}(\gamma) f(z)$$

for all  $\gamma \in \Gamma_{\mathcal{O}}$ . For each  $f \in C^{\infty}(\Gamma_{\mathcal{O}} \backslash \mathbb{H}, \chi^{-1})$ , we define

$$\|f\|_2 = \left( \mathrm{vol}(\Gamma_{\mathcal{O}} \backslash \mathbb{H})^{-1} \int_{\Gamma_{\mathcal{O}} \backslash \mathbb{H}} |f(z)|^2 dz \right)^{1/2}$$

where  $dz = c \frac{dx dy}{y^2}$  is any  $\mathrm{SL}_2(\mathbb{R})$ -invariant measure on  $\mathbb{H}$ . On  $C^{\infty}(\Gamma_{\mathcal{O}} \backslash \mathbb{H}, \chi^{-1})$  there exist *Hecke operators*  $T_n$  for each positive integer  $n$  defined as follows:

$$(T_n f)(z) = \sum_{\gamma \in \Gamma_{\mathcal{O}^{\max}(N_{\chi})} \backslash S_{\mathcal{O}^{\max}(N_{\chi})}(n)} f(\gamma z),$$

where  $N_{\chi}$  is chosen as above, and the subset  $S_{\mathcal{O}^{\max}(N_{\chi})}(n)$  of  $\mathrm{GL}_2(\mathbb{Q})$  is defined by

$$S_{\mathcal{O}^{\max}(N_{\chi})}(n) = \{\gamma \in \iota_{\infty}(\mathcal{O}^{\max}(N_{\chi})), \det(\gamma) = n\}.$$

It can be checked that the definition of  $T_n$  given above is *independent* of all choices, including the choice of  $N_{\chi}$ , and is well-defined on the space  $C^{\infty}(\Gamma_{\mathcal{O}} \backslash \mathbb{H}, \chi^{-1})$ . For all  $(n, dN_{\chi}) = 1$ , these operators are normal and commuting.

Now, let  $\phi$  and  $\pi$  be as in Theorem A. Assume that  $\phi$  is right-invariant by  $\iota_{\infty}^{-1}(\mathrm{SO}(2))$ . Define a function  $f_{\phi} : \mathbb{H} \mapsto \mathbb{C}$  via the equation

$$f_{\phi}(z) = \phi(g_{\infty})$$

where  $g_{\infty} \in D^{\times}(\mathbb{R})^+$  is any matrix such that  $\iota_{\infty}(g_{\infty})i = z$ . Then  $f_{\phi}$  has the following properties:

- (1)  $f_{\phi} \in C^{\infty}(\Gamma_{\mathcal{O}} \backslash \mathbb{H}, \chi^{-1})$ .
- (2)  $f_{\phi}$  satisfies  $(\Delta + \lambda)f_{\phi} = 0$  where  $\Delta := y^{-2}(\partial_x^2 + \partial_y^2)$ .
- (3)  $f_{\phi}$  is a simultaneous eigenfunction of the Hecke operators  $T_n$  for all positive integers  $n$  with  $(n, dN_{\chi}) = 1$ .

In other words,  $f_\phi$  is a *Maass form* for  $\Gamma_{\mathcal{O}}$  with character  $\chi^{-1}$  and Laplace eigenvalue  $\lambda$  that is a Hecke eigenform at the good primes. *Theorem A can be reformulated as an upper-bound on the sup-norms of such  $f_\phi$ :*

$$\|f_\phi\|_\infty \ll_{D,\lambda,\epsilon} N^\epsilon \min(\max(N^{1/3}, N_1^{1/2}), N^{11/24}) \|f_\phi\|_2.$$

This follows from the fact that  $\sup_{g \in D^\times(\mathbb{A})} |\phi(g)| = \sup_{z \in \mathbb{H}} |f_\phi(z)|$ .

**Remark 1.2.** *In fact, every Maass form  $f$  for  $\Gamma_{\mathcal{O}}$  with character  $\chi^{-1}$  and Laplace eigenvalue  $\lambda$  that is a Hecke eigenform at the good primes, can be obtained in the above way, i.e.,  $f = f_\phi$  for some  $\phi$  as in Theorem A. This can be proved using the technique of adelization. We omit the details of this in the interest of brevity.*

**1.4. The local bound and application to minimal vectors.** *For this subsection, we assume for simplicity that  $\pi$  has trivial central character.* We compare Theorem A with the *local bound* in the level aspect for automorphic forms  $\phi$  inside automorphic representations  $\pi$  of  $D^\times(\mathbb{A})$ . By the local bound for  $\phi$ , we mean the immediate bound emerging from the adelic pre-trace formula where the local test function at each ramified prime is chosen to be the restriction (to a maximal compact subgroup) of the *matrix coefficient* of  $\phi_p$ . In fact, an explicit computation performed in [14, 17] for the case of newforms together with the principle of formal degrees, allows us to write down this bound in terms of the conductor of  $\pi$ .

More precisely, let  $\pi$  be as in Theorem A such that  $\pi$  has trivial central character and  $\pi_p$  is one-dimensional at each prime dividing  $d$ . Then, for all  $\phi \in V_\pi$  that satisfy a mild condition<sup>5</sup>, we have the following bound:

$$(8) \quad \|\phi\|_\infty \ll_\epsilon C_1^{\frac{1}{2}+\epsilon} \|\phi\|_2$$

where  $C_1$  is the smallest integer such that  $\text{cond}(\pi) |C_1^2$ . We call (8) the local bound (in the level aspect). A more refined local bound is given in Remark 3.2 of this paper, under slightly more restrictive assumptions on  $\phi$ .

**Remark 1.3.** *The quantity  $C_1$  is equal to  $\text{cond}(\pi \times \tilde{\pi})^{1/2}$ . One reason that  $C_1 = \text{cond}(\pi \times \tilde{\pi})^{1/2}$  shows up in (8) is that (when  $\pi$  is discrete series)  $C_1$  approximately equals the formal degree of  $\pi$ ; see the calculations in [14, 17] or [10, Appendix A].*

The local bound (8) is essentially due to Marshall [14]. It seems reasonable to call (8) the local bound because (to quote Marshall in [14]) it appears to be “the best bound that may be proved by only considering the behaviour of  $\phi$  in one small open set at a time, without taking the global structure of the space into account”. We note that the bound (8) is also true in the non-compact setting of automorphic forms on  $\text{GL}_2(\mathbb{A})$ , provided one restricts the domain of  $\phi$  to a fixed compact set. It seems worthwhile here to comment on the relationship between the local bound (8) and the trivial bound (2). It can be shown easily that the local bound (8) is always *at least as strong* as the trivial bound (2). However, these two bounds have somewhat different flavours: the trivial bound applies to *all* forms that transform by unitary characters of compact subgroups of a particular volume (and does not depend on the conductors of the associated representations) while the local bound applies to

<sup>5</sup>The condition is that for some  $g \in D^\times(\mathbb{A}_f)$  we have  $\int_{gK_{\mathcal{O}_{\max}g^{-1}}} |\langle \pi(h)\phi, \phi \rangle|^2 dh \gg_\epsilon C_1^{-1-\epsilon} \langle \phi, \phi \rangle^2$ . This is a mild technical condition that is satisfied by several families of automorphic forms, including newforms, automorphic forms corresponding to minimal vectors,  $p$ -adic microlocal lifts, and so on. For a more down-to-earth but slightly stronger condition, see Remark 3.2.

forms whose associated representations have a particular *conductor* (and does not depend on some choice of subgroup that transforms the form by a character).

A central problem in this field (which also generalizes to higher rank automorphic forms) is to improve upon the local bound (8) for natural families of automorphic forms  $\phi$ . An obvious strategy to try to do this would be to first improve upon the trivial bound for some class of subgroups (as we do in Theorem A in wide generality), and then use this result (for some carefully chosen subgroup) to improve upon the local bound. This naive strategy works best when the local component  $\phi_p$  for each ramified prime  $p$  is an eigenvector of a relatively large neighbourhood of the identity. A key class of  $\phi_p$  for which this is true is the family of *minimal vectors*. Minimal vectors may be viewed as  $p$ -adic analogues of holomorphic vectors at infinity and have several remarkable properties, which were proved in our recent work [11] (where the analytic properties of the corresponding global automorphic forms of minimal type were studied for the first time in the setting of  $\mathrm{GL}_2$ ). The main result of [11] proved an optimal sup-norm bound for such forms in the setting of  $\mathrm{GL}_2$ .

However, the techniques used in [11] relied on a very special property of the Whittaker/Fourier expansion of  $\phi$  which only works in the non-compact setting. Therefore, the proof cannot carry over to the compact case, i.e., to our case of indefinite quaternion division algebras  $D$ , as no Whittaker/Fourier expansions exist here. A major impetus behind this paper was to improve upon (8) for automorphic forms of minimal type on *compact arithmetic surfaces*. One consequence of Theorem A is that we can now do this.

**Theorem B.** (see Theorem 2) *Let  $\pi$  be an irreducible, automorphic representation of  $D^\times(\mathbb{A})$  with trivial central character whose local component at each prime dividing  $d$  is one-dimensional, and let  $C$  denote the (arithmetic) conductor of  $\pi$ . Assume that  $C$  is the fourth power of an odd integer and suppose, for each prime  $p$  dividing  $C$ , that  $\pi_p$  is a supercuspidal representation. Let  $\phi$  in the space of  $\pi$  correspond to a minimal vector at each prime dividing  $C$ , a spherical vector at all other primes, and a vector of minimal weight at infinity. Then*

$$(9) \quad \|\phi\|_\infty \ll_{D,\pi,\epsilon} C^{\frac{1}{6}+\epsilon} \|\phi\|_2.$$

We remark that the condition on the conductor being the fourth power of an odd integer is a convenient one that was assumed for the definition of minimal vectors in our work [11]. However, this restriction has been partially removed in more recent work of Hu and Nelson [10] where they define and study properties of minimal vectors for all supercuspidal representations of  $D^\times$  where  $D$  is a (split or division) quaternion algebra over a  $p$ -adic field with  $p \neq 2$ . Using their work, we prove a more general version of Theorem B (Theorem 2) that applies to *any* odd conductor  $C$ .

The quantity  $C^{1/6}$  on the right side of (9) represents one-third of the progress from the local bound  $C^{1/4}$  extracted from the right side of (8) (we observe that in the setting of Theorem B,  $C_1 = C^{1/2}$ ) to the conjectured<sup>6</sup> true bound of  $C^\epsilon$ . Theorem B therefore gives a Weyl-type exponent, which appears to be a natural limit for the problem of improving upon the local bound with current tools, at least in cases where no Fourier expansions are available.

---

<sup>6</sup>To be fair, not a lot of evidence exists for this conjecture beyond the fact that it the best possible bound one can hope to prove, and no theoretical obstructions to achieving it have been found.

Theorem A also leads to a sub-local bound in certain other settings. These other settings include the case of “ $p$ -adic microlocal lifts” in the sense of [15]. The  $p$ -adic microlocal lifts may be naturally viewed as the principal series analogue of minimal vectors. Indeed, for the corresponding global automorphic forms, we are also able to prove a Weyl strength sub-local bound, see (43). We also obtain a bound for *newforms* that generalizes and strengthens previously known results; see Theorem 3.<sup>7</sup>

Finally, we remark that the results of this paper appear to be the first time that the local bound in the conductor aspect has been improved upon for squarefull conductors, for any kind of automorphic form on a compact domain. (In the non-compact case, this had been achieved in our previous paper [11].) It seems also worth mentioning here the very recent work of Hu [9] which generalizes [11] and proves a sub-local bound in the depth aspect for automorphic forms of minimal type on  $\mathrm{GL}_n$  under the assumption that the corresponding local representations have “generic” induction datum.

**1.5. Key ideas.** The heart of this paper is our solution to a counting problem for the number of elements that are “close” to the identity inside a (varying) quaternion order. This counting problem arises naturally in the amplification method for the level-aspect sup-norm problem. Roughly speaking, given an order  $\mathcal{O} \subseteq \mathcal{O}^{\max}$  of  $D$ , we are interested in good upper bounds for the integer

$$N_{\mathcal{O}}(m; z) = |\{\alpha \in S_{\mathcal{O}}(m) : u(z, \alpha z) \ll 1\}|$$

where

$$S_{\mathcal{O}}(m) = \{\gamma \in \iota_{\infty}(\mathcal{O}), \det(\gamma) = m\},$$

and  $u(z_1, z_2)$  denotes the hyperbolic distance.

Above,  $z$  is any point on the upper-half plane  $\mathbb{H}$ . Note, however, that for any  $\gamma \in \Gamma_{\mathcal{O}^{\max}}$ , we have  $N_{\mathcal{O}}(m; z) = N_{\mathcal{O}'}(m; \gamma z)$ , for an order  $\mathcal{O}'$  that is *conjugate* to  $\mathcal{O}$  by an element of  $(\mathcal{O}^{\max})^{\times}$ . This allows us to move  $z$  to a *fixed compact set*  $\mathcal{J}$ , namely  $\mathcal{J}$  equal to some choice of fundamental domain for the action of  $\Gamma_{\mathcal{O}^{\max}}$  on  $\mathbb{H}$ , at the cost of changing the order  $\mathcal{O}$  to a suitable  $(\mathcal{O}^{\max})^{\times}$ -conjugate of it. Now suppose that for each  $m$  and  $z \in \mathcal{J}$ , we are able to prove a bound for  $N_{\mathcal{O}}(m; z)$  that depends only on  $m$ ,  $\mathcal{J}$ , and the  $(\mathcal{O}^{\max})^{\times}$ -conjugacy class of  $\mathcal{O}$ . Then we have actually proved a bound that is valid for all  $z \in \mathbb{H}$ . This reduction is a key piece in our strategy and can be viewed as a workaround for the situation when  $\Gamma_{\mathcal{O}}$  is not a normal subgroup of  $\Gamma_{\mathcal{O}^{\max}}$  (c.f. the comments in Section 1.3 of [21]).

In fact we prove two separate bounds for  $N_{\mathcal{O}}(m; z)$  for  $z \in \mathcal{J}$ . The primary one among them (Proposition 2.8) is valid for general *lattices*  $\mathcal{O}$  (and does not use the multiplicative structure of the order at all!). The analysis behind the proof of this proposition, carried out in Sections 2.1 - 2.3, may be of independent interest. Roughly speaking, we use a workhorse lemma on small linear combinations of integers to reduce the counting problem to several elementary ones involving simple linear congruences. The reader may wonder at this point why we do not instead use standard lattice counting results such as Proposition 2.1 of [2]. The reason is that those counting results are typically in terms of the successive minima of the lattice, which is not a preserved quantity for  $(\mathcal{O}^{\max})^{\times}$ -conjugates of the lattice. In contrast, our method allows us to obtain a strong upper bound (Proposition 2.8) in terms of

---

<sup>7</sup>A much stronger bound in the setting of newforms of trivial character in the *depth aspect* will appear in a sequel to this paper.

the *invariant factors* of the lattice in  $\mathcal{O}^{\max}$  (the invariant factors are the same for all  $(\mathcal{O}^{\max})^\times$ -conjugate lattices).

However, the bound obtained in Proposition 2.8 is sufficient for our purposes only when the lattice is *balanced* in the sense that its largest invariant factor is not very large (relative to the level). This raises the question: how do we deal with unbalanced lattices? For this, we observe another useful fact: the sup-norm of an automorphic form  $\phi$  does not change when the form is replaced by some  $D^\times(\mathbb{A})$ -translate of it. Now, given  $\phi$  as in Theorem A, a  $D^\times(\mathbb{A})$ -translate of  $\phi$  transforms by a character of an order  $\mathcal{O}'$  that is *locally isomorphic* to (in the same genus as) the order  $\mathcal{O}$  that we started off with. This leads us to investigate which orders have the key property of having a locally isomorphic order whose largest invariant factor is not very large. We solve this problem by a careful case-by-case analysis (see Section 3.7) relying on the explicit classification of local *Gorenstein orders* due to Brzezinski. The answer (essentially) is that any order of level  $N$  is locally isomorphic to an order whose largest invariant factor divides  $N_1$ . This result may be of independent interest.

The upshot of all this is that the *only* orders for which our general lattice counting result (Proposition 2.8) does not lead to a non-trivial sup-norm bound are those whose levels are close to squarefree. To deal with this remaining case, we follow Templier’s method [21, 20] and prove a second counting result (Proposition 2.14), which uses the ring structure of the order to prove that elements that are close to the centralizer of some point must lie in a quadratic subfield. The combination of these two counting results lead directly to Theorem A above, and explain the shape the bound therein takes. Indeed, the term  $\max(N^{1/3}, N_1^{1/2})$  in Theorem A comes from our first counting result, while the term  $N^{11/24}$  comes from our second counting result.

Once the counting results are in place, we feed it into the amplification machinery to prove our main Theorems, closely following the adelic language employed in our previous paper [17]. It might be worth mentioning here that we use the slightly improved amplifier introduced by Blomer–Harcos–Milićević in [3] rather than the amplifier used in works like [8, 17], which saves us some technical difficulties.

**1.6. Additional remarks.** For simplicity, we have only proved a level aspect bound in Theorem A. It should be possible to extend the argument to prove a non-trivial hybrid bound, however we do not attempt to do so here. It may also be possible to extend some parts of this paper (with additional work) to the case of number fields, and to certain higher rank groups. This is because our counting argument for general lattices is elementary and highly flexible, and should generalise to anisotropic lattices of higher rank.

We end this introduction with a final remark. As explained in Section 1.4 of this paper, our main result leads to an improvement of the local bound (8) in the level aspect for various families of automorphic forms, particularly the ones of minimal type studied in [11, 10]. This uses crucially the fact that minimal vectors in supercuspidal representations  $\pi_p$  are eigenvectors for the action of a relatively large subgroup (having volume around  $\text{cond}(\pi_p)^{-1/2}$ ). In contrast, newvectors in  $\pi$  behave well only under the action of a much smaller subgroup (having volume around  $\text{cond}(\pi_p)^{-1}$ ). Therefore, the approach of this paper does not immediately lead to an improvement over the local bound for newforms in the depth aspect where the conductor varies over powers of a fixed prime. However, all hope is not lost –



it turns out that one can augment this approach with suitable results on *decay of matrix coefficients*. This is the topic of ongoing work of the author with Yueke Hu, which will be published in a sequel to this paper. Our method there will allow us to beat the local bound for newforms in the depth aspect for the first time.

**Acknowledgements.** I would like to thank Yueke Hu and Paul Nelson for useful discussions, and Raphael Steiner for his generous help with the proof of Lemma 2.1. I would also like to thank the anonymous referee for his suggestion to rephrase the argument of Section 2.3 in terms of matrix manipulations and for other suggestions which have improved the readability of this paper.

## 2. A COUNTING PROBLEM FOR LATTICES

**2.1. A lemma on linear combinations of integers.** The object of this subsection is to prove an elementary but very useful lemma on the existence of “small” linear combinations of integers coprime to another integer. It is possible that some version of this lemma has appeared elsewhere, but we were unable to find a suitable reference. The proof below is essentially due to Raphael Steiner (private correspondence, March 2018) and we are grateful to him for his help and for allowing us to include his proof here.

**Lemma 2.1.** *For any  $\epsilon > 0$ , and any  $c \geq 1$ , there is a positive constant  $C_{c,\epsilon}$  such that for all  $(n+2)$ -tuples of integers  $(a_0, a_1, \dots, a_n, N)$ , with  $N > 0$ ,  $\gcd(a_0, a_1, \dots, a_n, N) = 1$ , there exists at least  $c^n$  distinct  $n$ -tuples of integers  $(p_1, p_2, \dots, p_n)$  with*

- (1)  $0 \leq p_i \leq C_{c,\epsilon} N^\epsilon$  for  $1 \leq i \leq n$ ,
- (2)  $\gcd(a_0 + a_1 p_1 + a_2 p_2 + \dots + a_n p_n, N) = 1$ .

*More precisely, for any non-empty subset  $I$  of  $\{1, \dots, n\}$ , let  $P_I$  denote the natural projection map from  $\mathbb{Z}^n$  to  $\mathbb{Z}^I$  taking any  $n$ -tuple to its associated sub-tuple corresponding to the indices in  $I$ . Then, if  $S \subset \mathbb{Z}^n$  consists of all the  $n$ -tuples  $(p_1, p_2, \dots, p_n)$  satisfying the conditions (1), (2) above, then  $|P_I(S)| \geq c^{|I|}$  for each subset  $I$  of  $\{1, \dots, n\}$ .*

**Remark 2.2.** *With more sophisticated sieving methods a la [13], one can replace  $N^\epsilon$  by  $\log(N)^2$  in the lemma above.*

**Remark 2.3.** *We encourage the reader to initially focus on the case  $n = 1$  of the lemma above to get a feel for the statement. In this paper, we will need the lemma only in the case  $c = 2$ ,  $n = 2$ .*

*Proof.* We may assume without loss of generality that  $\gcd(a_0, a_1, \dots, a_n) = 1$  and that  $N$  is squarefree. Indeed, if these conditions are not met, we can replace each  $a_i$  by  $a_i/d$  where  $d = \gcd(a_0, a_1, \dots, a_n)$  and we can replace  $N$  by its largest squarefree divisor, so that this modified setup does satisfy the conditions. Any constant  $C_{c,\epsilon}$  that works for this modified setup will also work for the original setup.

We will prove the lemma by induction on  $n$ . Let us prove the base case  $n = 1$ . The starting point for this case is the following fact: *For all  $\epsilon > 0$ , there is a constant  $D_\epsilon$  such that for all positive integers  $a_0, a_1, Q$  with  $\gcd(a_1, Q) = 1$  and  $X > 0$  we have*

$$(10) \quad \sum_{\substack{1 \leq m \leq X \\ \gcd(a_0 + ma_1, Q) = 1}} 1 \geq \frac{\phi(Q)}{Q} X - D_\epsilon Q^{\epsilon/2}.$$

The proof of (10) follows from the following calculation:

$$\begin{aligned} \sum_{\substack{1 \leq m \leq X \\ \gcd(a_0 + ma_1, Q) = 1}} 1 &= \sum_{1 \leq m \leq X} \sum_{\substack{d|Q \\ d|a_0 + ma_1}} \mu(d) = \sum_{d|Q} \mu(d) \sum_{\substack{1 \leq m \leq X \\ m \equiv -\frac{a_0}{a_1} \pmod{d}}} 1 \\ &= \sum_{d|Q} \mu(d) \left( \frac{X}{d} + O(1) \right) = X \sum_{d|Q} \frac{\mu(d)}{d} + \sum_{d|Q} O(1) = \frac{\phi(Q)}{Q} X + O_\epsilon(Q^\epsilon). \end{aligned}$$

Let us now explain how the case  $n = 1$  of the lemma follows from (10). We may assume that  $D_\epsilon > 1$ . We can find a constant  $E_\epsilon$  such that

$$\frac{Q}{\phi(Q)} \leq E_\epsilon Q^{\epsilon/2}$$

for all  $Q$  and all  $\epsilon > 0$ . Now put  $Q = N / \gcd(a_1, N)$  and choose  $C_{c,\epsilon} > (c+1)D_\epsilon E_\epsilon$ . Then picking  $X = C_{c,\epsilon} Q^\epsilon$  in (10) we have that

$$(11) \quad \sum_{\substack{1 \leq m \leq C_{c,\epsilon} Q^\epsilon \\ \gcd(a_0 + ma_1, Q) = 1}} 1 \geq \frac{\phi(Q)}{Q} C_{c,\epsilon} Q^\epsilon - D_\epsilon Q^{\epsilon/2} \geq c D_\epsilon Q^{\epsilon/2} \geq c.$$

So there exist at least  $c$  distinct integers  $m_1, m_2, \dots, m_c$ , such that for  $p_1 \in \{m_1, m_2, \dots, m_c\}$ , we have

$$p_1 \leq C_{c,\epsilon} N^\epsilon, \quad \gcd(a_0 + p_1 a_1, N / \gcd(a_1, N)) = 1.$$

However, since  $\gcd(a_0, a_1) = 1$ , we have that

$$\gcd(a_0 + p_1 a_1, N / \gcd(a_1, N)) = 1 \Rightarrow \gcd(a_0 + p_1 a_1, N) = 1.$$

The proof of the case  $n = 1$  of the Lemma is complete.

We now prove the induction step. Assume that the lemma is proved for  $n = k$ . We will prove the lemma for  $n = k+1$ . Suppose we have integers  $a_0, a_1, \dots, a_{k+1}, N$ , with  $\gcd(a_0, a_1, \dots, a_{k+1}) = 1$ . By replacing  $a_{k+1}$  by its residue modulo  $N$  if necessary, we may assume that  $0 \leq a_{k+1} \leq N$ .

Since  $a_{k+1} \leq N$ , by the case  $n = k$  of the Lemma, we can find a set  $S_k \subset \mathbb{Z}^k$  with the following properties:

- For each  $(p_1, \dots, p_k) \in S_k$ , we have  $0 \leq p_i \leq C_{c,\epsilon} N^\epsilon$  and  $\gcd(a_0 + a_1 p_1 + \dots + a_k p_k, a_{k+1}) = 1$ .
- For each non-empty subset  $I$  of  $\{1, \dots, k\}$ ,  $|P_I(S)| \geq c^{|I|}$ .

We now construct a set  $S_{k+1} \subset \mathbb{Z}^{k+1}$ . Namely, given any  $\tilde{s} = (p_1, \dots, p_k) \in S_k$ , we use the case  $n = 1$  of the lemma to find  $c$  distinct integers  $p_{\tilde{s},i}$  for  $1 \leq i \leq c$ , such that  $0 \leq p_{\tilde{s},i} \leq C_{c,\epsilon} N^\epsilon$  and such that

$$\gcd((a_0 + a_1 p_1 + \dots + a_k p_k) + a_{k+1} p_{\tilde{s},i}, N) = 1.$$

Define

$$S_{k+1} = \{(\tilde{s}, p_{\tilde{s},i}) : \tilde{s} \in S_k, 1 \leq i \leq c\}.$$

It is clear that  $S_{k+1}$  satisfies the required conditions and thus the induction step is complete.  $\square$

**2.2. Lattices in quaternion orders.** Let  $D$  be an indefinite quaternion division algebra over  $\mathbb{Q}$ . We let  $d$  denote the reduced discriminant of  $D$ , i.e., the product of all primes such that  $D_p$  is a division algebra. Fix once and for all a maximal order  $\mathcal{O}^{\max}$  of  $D$ , and an isomorphism<sup>8</sup>  $\iota_\infty : D_\infty \rightarrow M(2, \mathbb{R})$ .

For  $\alpha \in D$ , let  $\alpha \mapsto \bar{\alpha}$  be the standard involution of  $D$  and let the reduced norm  $\text{nr}$  and trace  $\text{Tr}$  be given by

$$\text{nr}(\alpha) = \alpha\bar{\alpha}, \quad \text{Tr}(\alpha) = \alpha + \bar{\alpha}.$$

Given a subset  $\mathcal{L}$  of  $\mathcal{O}^{\max}$ , and an integer  $m$ , we define

$$\begin{aligned} \mathcal{L}(m) &= \{\alpha \in \mathcal{L} : \text{nr}(\alpha) = m\}, \\ \mathcal{L}_m &= \{\alpha \in \mathcal{L} : \text{Tr}(\alpha) = m\}. \end{aligned}$$

Thus,  $\mathcal{O}_0^{\max}$  denotes the trace 0 elements of  $\mathcal{O}^{\max}$ , and  $\mathcal{O}^{\max}(1)$  is the subgroup of  $(\mathcal{O}^{\max})^\times$  with reduced norm 1. We fix, once and for all, three elements  $i_1, i_2, i_3$  in  $\mathcal{O}_0^{\max}$  such that

$$\mathcal{O}_0^{\max} = \mathbb{Z}i_1 \oplus \mathbb{Z}i_2 \oplus \mathbb{Z}i_3.$$

So we have

$$D = \mathbb{Q} + \mathbb{Q}i_1 + \mathbb{Q}i_2 + \mathbb{Q}i_3.$$

Fix a compact subset  $\mathcal{J}$  of  $\mathbb{H}$ .<sup>9</sup> Given a subset  $\mathcal{L}$  of  $D$ , and an element  $z \in \mathbb{H}$ ,  $\delta > 0$ , define for each positive integer  $m$  the set

$$\mathcal{L}(m; z, \delta) = \{\alpha \in \mathcal{L}(m) : u(z, \iota_\infty(\alpha)z) \leq \delta\}.$$

The reader should think of  $\delta \asymp 1$  as fixed, since all constants will be allowed to depend on  $\delta$  (in fact, for our eventual applications, we will actually fix  $\delta = 1$ , however it will be useful to keep this variable  $\delta$  around for now). Our goal is to bound the cardinality of  $\mathcal{L}(m; z, \delta)$  (in terms of some basic invariants of  $\mathcal{L}$ ) whenever  $\mathcal{L} \subseteq \mathcal{O}^{\max}$  is a *lattice* containing 1.

Let  $\mathcal{L} \subseteq \mathcal{O}^{\max}$  be a lattice containing 1. We denote

$$N = [\mathcal{O}^{\max} : \mathcal{L}]$$

and call  $N$  the level of  $\mathcal{L}$ . By the structure theorem for finitely generated abelian groups, the finite group  $\mathcal{O}_0^{\max}/L_0$  is isomorphic to  $(\mathbb{Z}/M_1\mathbb{Z}) \times (\mathbb{Z}/M_2\mathbb{Z}) \times (\mathbb{Z}/M_3\mathbb{Z})$  for some uniquely defined positive integers  $M_1|M_2|M_3$ , which are sometimes known as invariant factors. We will refer to these integers as the shape of  $\mathcal{L}$ .

**Definition 2.4.** *Given positive integers  $M_1, M_2, M_3$  such that  $M_1|M_2|M_3$ , a lattice  $\mathcal{L}$  of  $D$  is said to have shape  $(M_1, M_2, M_3)$  if  $1 \in \mathcal{L}$ ,  $\mathcal{L} \subseteq \mathcal{O}^{\max}$  and there exist elements  $\Delta_1, \Delta_2, \Delta_3 \in \mathcal{O}_0^{\max}$  such that:*

- (1)  $\mathcal{O}_0^{\max} = \mathbb{Z}\Delta_1 \oplus \mathbb{Z}\Delta_2 \oplus \mathbb{Z}\Delta_3$ ,
- (2)  $\mathcal{L}_0 = M_1\mathbb{Z}\Delta_1 \oplus M_2\mathbb{Z}\Delta_2 \oplus M_3\mathbb{Z}\Delta_3$ .

Furthermore, we denote  $M = M_1M_2M_3$  and call it the level of  $\mathcal{L}_0$ .

**Remark 2.5.** *Let  $\mathcal{L} \subseteq \mathcal{O}^{\max}$  be a lattice of shape  $(M_1, M_2, M_3)$  and level  $N$ . If  $x \in D$  satisfies  $x\mathcal{L}x^{-1} \subseteq \mathcal{O}^{\max}$ , then one may ask about the shape and level of  $\mathcal{L}' := x\mathcal{L}x^{-1}$ .*

*It is easy to see that  $\mathcal{L}'$  always has level  $N$  but its shape might be different in general. However, if  $x \in (\mathcal{O}^{\max})^\times$ , then  $\mathcal{L}'$  also has shape  $(M_1, M_2, M_3)$ .*

<sup>8</sup>Such an isomorphism  $\iota_\infty$  is unique up to conjugation by  $\text{GL}_2(\mathbb{R})$ .

<sup>9</sup>Later in this paper, we will take  $\mathcal{J}$  to be a fundamental domain for the action of  $\iota_\infty(\mathcal{O}^{\max}(1))$  on  $\mathbb{H}$ .

It turns out to be more convenient for us to descend to the sublattice  $\mathbb{Z} \oplus \mathcal{L}_0$ , for which the next lemma is essential.

**Lemma 2.6.** *Let  $\mathcal{L}$  be a lattice in  $D$  such that  $\mathcal{L} \subseteq \mathcal{O}^{\max}$  and  $1 \in \mathcal{L}$ . Then*

$$[\mathcal{L} : \mathbb{Z} \oplus \mathcal{L}_0] \in \{1, 2\}.$$

*Proof.* Given an element  $\ell \in \mathcal{L}$ , we have  $\text{Tr}(\ell) \in \mathbb{Z}$  and furthermore,  $\ell$  belongs to  $\mathbb{Z} \oplus \mathcal{L}_0$  if and only if  $\text{Tr}(\ell) \in 2\mathbb{Z}$ . So if  $\ell_1$  and  $\ell_2$  are two elements in  $\mathcal{L}$ , neither of which belong to  $\mathbb{Z} \oplus \mathcal{L}_0$ , then  $\ell_1 + \ell_2 \in \mathbb{Z} \oplus \mathcal{L}_0$ . The statement follows.  $\square$

Let  $\mathcal{L}$  be a lattice as in Lemma 2.6, and let  $N$  be the level of  $\mathcal{L}$ , and  $M$  the level of  $\mathcal{L}_0$ . Using Lemma 2.6 and the fact that  $\mathcal{O}_0^{\max} + \mathbb{Z}$  has index two in  $\mathcal{O}^{\max}$ , we observe that

$$(12) \quad M = [\mathcal{O}_0^{\max} : \mathcal{L}_0] = \frac{[\mathcal{O}^{\max} : \mathbb{Z} \oplus \mathcal{L}_0]}{2} = \frac{N}{e}, \quad e \in \{1, 2\}$$

where  $e$  equals 2 if and only if  $\mathcal{L} = \mathbb{Z} \oplus \mathcal{L}_0$ . So  $N \asymp M = M_1 M_2 M_3$ .

**Remark 2.7.** *Consider the lattice  $\mathcal{L}$  in  $\mathcal{O}^{\max}$ . The invariant factors of  $\mathcal{L}$  with respect to  $\mathcal{O}^{\max}$  are  $(1, M'_1, M'_2, M'_3)$ , for some integers  $M'_1 | M'_2 | M'_3$ . Now, using Lemma 2.6, we obtain for  $i = 1, 2, 3$ ,*

$$(13) \quad M'_i = e_i M_i, \quad e_i \in \{1, 2\}, \quad e_1 e_2 e_3 = e \in \{1, 2\}.$$

We now state our main counting result.

**Proposition 2.8.** *Let  $\mathcal{L} \subseteq \mathcal{O}^{\max}$  be a lattice containing 1. Suppose that  $\mathcal{L}$  has shape  $(M_1, M_2, M_3)$  and level  $N$ . Let  $z \in \mathcal{J}$  and  $1 \leq L \leq N^{O(1)}$ . Then the following hold.*

$$(14) \quad \sum_{1 \leq m \leq L} |\mathcal{L}(m; z, \delta)| \ll_{\epsilon, \delta} N^\epsilon \left( L^{1/2} + \frac{L}{M_1} + \frac{L^{3/2}}{M_1 M_2} + \frac{L^2}{N} \right).$$

$$(15) \quad \sum_{1 \leq m \leq L} |\mathcal{L}(m^2; z, \delta)| \ll_{\epsilon, \delta} N^\epsilon \left( L + \frac{L^2}{M_1 M_2} + \frac{L^3}{N} \right).$$

**Remark 2.9.** *The constants implicit in the bounds above depend only on  $\epsilon, \delta$  and the fixed objects  $D, \mathcal{O}^{\max}, \iota_\infty, \mathcal{J}$ .*

**Remark 2.10.** *Note that the bounds do not depend on the elements  $\Delta_1, \Delta_2, \Delta_3$ . Hence the bounds obtained are uniform over all  $(\mathcal{O}^{\max})^\times$ -conjugates of  $\mathcal{L}$ . This will be key for us later on.*

**Remark 2.11.** *Because of (12) one can replace  $N$  by  $M$  in the theorem above, if one wishes. Furthermore, because of (13), one can replace  $M_1, M_2$  in the theorem above by  $M'_1, M'_2$  respectively, if one wishes to.*

**Remark 2.12.** *The bound obtained in Proposition 2.8 is not sufficient for our purposes when  $M_1 M_2$  is small in relation to  $N$ . So in Section 2.4, we will prove another counting result under the additional assumption that  $\mathcal{L}$  is an order.*

### 2.3. Proof of Proposition 2.8.

**Lemma 2.13.** *For any  $\delta > 0$ , there exists a constant  $T$  (depending on  $\delta$ ,  $\mathcal{J}$  and  $\iota_\infty$ ) with the following property: For  $m > 0$ ,  $z \in \mathcal{J}$ , and  $\alpha = a_0 + a_1 i_1 + a_2 i_2 + a_3 i_3 \in D(m)$  satisfying  $u(z, \iota_\infty(\alpha)z) \leq \delta$ , we have*

$$|a_i| \leq Tm^{1/2} \text{ for } i = 0, 1, 2, 3.$$

*Proof.* It suffices to show that the set of all  $\frac{\alpha}{\sqrt{m}}$  as above lies in a compact set depending only on  $\delta$ ,  $\mathcal{J}$  and  $\iota_\infty$ . The subset  $\Omega_{\mathcal{J}, \delta}$  of  $\mathrm{SL}_2(\mathbb{R})$  given by

$$\Omega_{\mathcal{J}, \delta} = \{\gamma \in \mathrm{SL}_2(\mathbb{R}) : u(z, \gamma z) \leq \delta \text{ for all } z \in \mathcal{J}\}$$

is compact since the stabilizer of each point is compact, and  $\mathcal{J}$  is compact. Therefore the subset  $\iota_\infty^{-1}(\Omega_{\mathcal{J}, \delta})$  is a compact subset of  $D(1)$  that contains all the elements  $\frac{\alpha}{\sqrt{m}} = \frac{a_0 + a_1 i_1 + a_2 i_2 + a_3 i_3}{\sqrt{m}}$  as in the Lemma. The result follows.  $\square$

We now prove Proposition 2.8. Using Lemma 2.6, we may assume that  $\mathcal{L} = \mathbb{Z} \oplus \mathcal{L}_0$ . Indeed, putting  $\mathcal{L}' = \mathbb{Z} \oplus \mathcal{L}_0$ , we see that

$$|\mathcal{L}'(m; z, \delta)| \leq |\mathcal{L}(m; z, \delta)| \leq |\mathcal{L}'(4m; z, \delta)|.$$

So by shrinking  $\mathcal{L}$  if necessary, we will assume throughout the rest of this subsection that

$$(16) \quad \mathcal{L} = \mathbb{Z} \oplus \mathcal{L}_0.$$

Now, using Lemma 2.13, we see that Proposition 2.8 would follow from the following statement:

*Let  $\mathcal{L} = \mathbb{Z} \oplus M_1 \mathbb{Z} \Delta_1 \oplus M_2 \mathbb{Z} \Delta_2 \oplus M_3 \mathbb{Z} \Delta_3$  be a lattice of shape  $(M_1, M_2, M_3)$  and level  $N$ , where  $\Delta_1, \Delta_2, \Delta_3 \in \mathcal{O}_0^{\max}$  and  $M_1 | M_2 | M_3$ . For  $T > 0$ , define*

$$\mathcal{L}(m, T) = \{a_0 + a_1 i_1 + a_2 i_2 + a_3 i_3 \in \mathcal{L}(m) : a_i \in \mathbb{Z}, |a_i| \leq Tm^{1/2} \text{ for } 0 \leq i \leq 3\}.$$

*Then for  $1 \leq L \leq N^{O(1)}$  we have:*

$$(17) \quad \#\left(\bigcup_{1 \leq m \leq L} \mathcal{L}(m, T)\right) \ll_{\epsilon, T} N^\epsilon \left(L^{1/2} + \frac{L}{M_1} + \frac{L^{3/2}}{M_1 M_2} + \frac{L^2}{N}\right).$$

$$(18) \quad \#\left(\bigcup_{\substack{1 \leq m \leq L \\ \sqrt{m} \in \mathbb{Z}}} \mathcal{L}(m, T)\right) \ll_{\epsilon, T} N^\epsilon \left(L^{1/2} + \frac{L}{M_1 M_2} + \frac{L^{3/2}}{N}\right).$$

We now begin the proof of the bounds (17), (18). This will complete the proof of Proposition 2.8. For brevity, we drop  $T$  from the  $\ll$  symbol in the rest of this subsection (so all constants henceforth are allowed to depend on  $T$ ).

Since  $(\Delta_1, \Delta_2, \Delta_3)$  and  $(i_1, i_2, i_3)$  are integral bases for  $\mathcal{O}_0$ , it follows that there exists a matrix  $\delta \in \mathrm{GL}_3(\mathbb{Z})$  such that

$$(19) \quad [\Delta_1, \Delta_2, \Delta_3]^t = \delta [i_1, i_2, i_3]^t.$$

We note that

$$(20) \quad \det(\delta) = \pm 1.$$

In particular,  $\mathrm{gcd}(\delta_{1,1}, \delta_{1,2}, \delta_{1,3}) = 1$ . Therefore, using Lemma 2.1, we fix integers  $r_2, r_3$ , both  $\ll_\epsilon N^\epsilon$ , such that

$$(21) \quad R = \delta_{1,1} + r_2 \delta_{1,2} + r_3 \delta_{1,3} \text{ satisfies } \mathrm{gcd}(R, N) = 1.$$

Define  $R_2, R_3$  via the equation

$$(22) \quad \delta [1, r_2, r_3]^t = [R, R_2, R_3]^t.$$

Fix an integer  $\bar{R}$  such that  $R\bar{R} \equiv 1 \pmod{N}$ . Now define

$$h = \begin{bmatrix} 1 & 0 & 0 \\ -R_2\bar{R} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Let the integers  $S_1, S_2, S_3$  be defined via the equation

$$[0, 1, 0]h\delta = [S_1, S_2, S_3].$$

We claim that  $\gcd(S_1, S_2, S_3, N) = 1$ . Suppose this were not true. Then we would be able to find a prime  $p|N$  such that

$$[0, 1, 0]h\delta \equiv [0, 0, 0] \pmod{p}.$$

Since  $\delta \in \mathrm{GL}_3(\mathbb{Z})$ , this implies that

$$(23) \quad [0, 1, 0]h \equiv [0, 0, 0] \pmod{p}.$$

However, the left-hand-side of (23) is a vector of the form  $[*, 1, 0]$ , leading to a contradiction. This contradiction proves that  $\gcd(S_1, S_2, S_3, N) = 1$ .

Using Lemma 2.1, we now fix integers  $s_2, s_3$ , both  $\ll_\epsilon N^\epsilon$  and such that

$$s_2 \neq r_2, \quad S = S_1 + s_2S_2 + s_3S_3 \text{ satisfies } \gcd(S, N) = 1.$$

Define the element  $S'$  via

$$S' = [1, 0, 0]\delta[1, s_2, s_3]^t.$$

Put

$$g = \begin{bmatrix} 1 & 1 & 0 \\ r_2 & s_2 & 0 \\ r_3 & s_3 & 1 \end{bmatrix}.$$

By our assumption that  $s_2 \neq r_2$ , we have  $\det(g) \neq 0$ . Moreover, from (22) and the definitions of  $h, S, S'$ , we see that

$$(24) \quad h\delta g \equiv \begin{bmatrix} R & S' & \delta_{1,3} \\ 0 & S & -R_2\bar{R}\delta_{1,3} + \delta_{2,3} \\ * & * & * \end{bmatrix} \pmod{N}.$$

We now begin the proof of (17). Let

$$\alpha \in \bigcup_{1 \leq m \leq L} \mathcal{L}(m, T).$$

Our strategy will be to associate to each such  $\alpha$  a quadruple  $(a_0, A_\alpha, B_\alpha, a_3)$  such that the function  $\alpha \mapsto (a_0, A_\alpha, B_\alpha, a_3)$  is injective. A bound for the cardinality of  $\bigcup_{1 \leq m \leq L} \mathcal{L}(m, T)$  will then follow by bounding the number of distinct tuples  $(a_0, A_\alpha, B_\alpha, a_3)$  that are possible.

Write

$$\alpha = a_0 + a_1i_1 + a_2i_2 + a_3i_3.$$

We have that  $|a_i| \ll L^{1/2}$ . Furthermore, using the definition of  $\mathcal{L}$  together with (19), it follows that there exist integers  $u_1, u_2, u_3$  such that

$$(25) \quad [a_1, a_2, a_3] = [u_1M_1, u_2M_2, u_3M_3]\delta.$$

Define  $A_\alpha, B_\alpha$  via the equation

$$(26) \quad [A_\alpha, B_\alpha, a_3] = [a_1, a_2, a_3]g = [u_1M_1, u_2M_2, u_3M_3]\delta g.$$

We saw already that  $g$  is an invertible matrix and therefore the mapping  $\alpha \mapsto (a_0, A_\alpha, B_\alpha, a_3)$  is injective, as required.

It remains to bound the number of distinct tuples of integers  $(a_0, A_\alpha, B_\alpha, a_3)$  that are possible. To achieve that, we will prove certain bounds and congruences satisfied by such tuples. First of all, using (26) and the fact that  $r_2, r_3$  are both  $\ll_\epsilon N^\epsilon$ , it follows that  $|A_\alpha| \ll_\epsilon L^{1/2}N^\epsilon$  and  $A_\alpha \equiv 0 \pmod{M_1}$ . *So there are  $\ll_\epsilon N^\epsilon \left(1 + \frac{L^{1/2}}{M_1}\right)$  choices for  $A_\alpha$ . Henceforth, assume such a choice has been made.*

Next, using (24) and (26) we see that

$$(27) \quad [A_\alpha, B_\alpha, a_3] = [u_1M_1, u_2M_2, u_3M_3]h^{-1}h\delta g \\ \equiv [u_1M_1 + R_2\bar{R}u_2M_2, u_2M_2, 0] \begin{bmatrix} R & S' & \delta_{1,3} \\ 0 & S & -R_2\bar{R}\delta_{1,3} + \delta_{2,3} \\ * & * & * \end{bmatrix} \pmod{M_3}$$

$$(28) \quad \equiv [u_1M_1, 0, 0] \begin{bmatrix} R & S' & \delta_{1,3} \\ * & * & * \\ * & * & * \end{bmatrix} \pmod{M_2}.$$

The last congruence (28) above gives us  $B_\alpha \equiv A_\alpha\bar{R}S' \pmod{M_2}$ . *Since  $|B_\alpha| \ll_\epsilon L^{1/2}N^\epsilon$ , there are  $\ll_\epsilon N^\epsilon \left(1 + \frac{L^{1/2}}{M_2}\right)$  choices for  $B_\alpha$ . Henceforth, assume such a choice has been made.*

Furthermore, (27) directly leads to an expression of the form

$$[u_1M_1 + R_2\bar{R}u_2M_2, u_2M_2] \equiv [A_\alpha\bar{R}, \bar{S}(B_\alpha - A_\alpha\bar{R}S')] \pmod{M_3}.$$

Therefore, another application of (27) tells us that  $a_3$  is known modulo  $M_3$  in terms of choices that have already been made. Indeed, if one were to actually perform the above steps, one would arrive at the explicit expression

$$(29) \quad a_3 \equiv A_\alpha\bar{R}\delta_{1,3} + \bar{S}(B_\alpha - A_\alpha\bar{R}S')(\delta_{2,3} - R_2\bar{R}\delta_{1,3}) \pmod{M_3}.$$

*Since  $|a_3| \ll_\epsilon L^{1/2}$  there are  $\ll_\epsilon N^\epsilon \left(1 + \frac{L^{1/2}}{M_3}\right)$  choices for  $a_3$ .*

*Finally, as  $|a_0| \ll L^{1/2}$ , there are  $\ll L^{1/2}$  possible choices for  $a_0$ .*

Combining all the above italicized statements, we see that there are

$$\ll_\epsilon N^\epsilon L^{1/2} \left(1 + \frac{L^{1/2}}{M_1}\right) \left(1 + \frac{L^{1/2}}{M_2}\right) \left(1 + \frac{L^{1/2}}{M_3}\right) \\ \ll_\epsilon N^\epsilon \left(L^{1/2} + \frac{L}{M_1} + \frac{L^{3/2}}{M_1M_2} + \frac{L^2}{N}\right).$$

possible different choices for triples  $(a_0, A_\alpha, B_\alpha, a_3)$  as  $\alpha$  varies in  $\bigcup_{1 \leq m \leq L} \mathcal{L}(m, T)$ . By the injectivity noted earlier, this completes the proof of (17).

To prove (18), define  $\alpha_0 := (\alpha - \bar{\alpha})/2 = a_1i_1 + a_2i_2 + a_3i_3$ . Now, only consider those  $\alpha$  such that  $m = \text{nr}(\alpha) = \ell^2$  for some  $\ell \in \mathbb{Z}$ ,  $0 \leq \ell \leq L^{1/2}$ . Then we get

$$(30) \quad (a_0 - \ell)(a_0 + \ell) = \alpha_0^2 = -\text{nr}(\alpha_0).$$

Now,  $\text{nr}(\alpha_0)$  is an integer and  $|\text{nr}(\alpha_0)| \ll N^{O(1)}$ . So, if  $\text{nr}(\alpha_0) \neq 0$  then (30) tells us that there are  $\ll_\epsilon N^\epsilon$  possibilities for  $a_0$ . If  $\text{nr}(\alpha_0) = 0$  then we must have  $\alpha_0 = 0$  (since  $D$  is a *division algebra*) and so there are  $\ll L^{1/2}$  possibilities for  $a_0 = \alpha$ .

Putting it together, we see that the number of elements in  $\bigcup_{\substack{1 \leq m \leq L \\ \sqrt{m} \in \mathbb{Z}}} \mathcal{L}(m, T)$  is

$$\ll_{\epsilon} L^{1/2} + N^{\epsilon} \left( 1 + \frac{L^{1/2}}{M_1} + \frac{L}{M_1 M_2} + \frac{L^{3/2}}{N} \right).$$

This completes the proof of (18).

**2.4. A supplementary counting result for orders.** In this subsection, we give another counting result to supplement Proposition 2.8, but one that is applicable only if  $\mathcal{L} = \mathcal{O}$  is an *order*.

**Proposition 2.14.** *Let  $\mathcal{O} \subseteq \mathcal{O}^{\max}$  be an order of level  $N$ . There is a constant  $C$  (depending on  $\delta$ ,  $\mathcal{J}$  and  $\iota_{\infty}$ ) such that for  $z \in \mathcal{J}$  and  $1 \leq m < CN^{\frac{1}{3}}$ , we have*

$$(31) \quad |\mathcal{O}(m; z, \delta)| \ll_{\epsilon, \delta} m^{\epsilon}.$$

Our proof of Proposition 2.14 is broadly similar to that of Proposition 6.5 of [21] (see also [20]). The proof will follow from the following sequence of lemmas. Throughout the proof, we will use the notations introduced in Section 2.2 and we will assume (without loss of generality) that (16) holds.

**Lemma 2.15.** *Let  $\mathcal{L}$  be a subset of  $D$  that is closed under multiplication and contains 1. Let  $z \in \mathbb{H}$ ,  $L$  a positive integer, and  $\delta > 0$ . Then the  $\mathbb{Q}$ -algebra generated by all elements in  $\bigcup_{1 \leq m \leq L} \mathcal{L}(m; z, \delta)$  is contained in the  $\mathbb{Q}$ -vector-space spanned by  $\bigcup_{1 \leq m \leq L^2} \mathcal{L}(m; z, 2\delta)$ .*

*Proof.* By basic properties of a quaternion algebra, any element of the  $\mathbb{Q}$ -algebra generated by  $\bigcup_{1 \leq m \leq L} \mathcal{L}(m; z, \delta)$  is a  $\mathbb{Q}$ -linear combination of elements of the form  $\beta = \beta_1 \beta_2$  with  $\beta_1, \beta_2 \in \bigcup_{1 \leq m \leq L} \mathcal{L}(m; z, \delta)$ . So it suffices to show that any such  $\beta$  belongs to  $\bigcup_{1 \leq m \leq L^2} \mathcal{L}(m; z, 2\delta)$ . This is clear as  $\text{nr}(\beta) = \text{nr}(\beta_1) \text{nr}(\beta_2) \leq L^2$  and

$$u(z, \iota_{\infty}(\beta)z) = u(\iota_{\infty}(\beta_1^{-1})z, \iota_{\infty}(\beta_2)z) \leq u(z, \iota_{\infty}(\beta_1)z) + u(z, \iota_{\infty}(\beta_2)z) \leq 2\delta.$$

□

**Lemma 2.16.** *Let  $\mathcal{L}$  be a lattice in  $D$  of level  $N$ ,  $z \in \mathcal{J}$ ,  $L$  a positive integer, and  $\delta > 0$ . Then there is a constant  $C$  (depending on  $\delta$ ,  $\mathcal{J}$  and  $\iota_{\infty}$ ) such that the  $\mathbb{Q}$ -vector-space spanned by  $\bigcup_{1 \leq m \leq L^2} \mathcal{L}(m; z, 2\delta)$  is proper whenever  $L < CN^{1/3}$ .*

*Proof.* Let  $\alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}$  be three arbitrary elements of  $\bigcup_{1 \leq m \leq L^2} \mathcal{L}(m; z, 2\delta)$ . It suffices to show that  $1, \alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}$  are linearly dependant. For  $i = 1, 2, 3$ , write

$$\alpha^{(i)} = a_0^{(i)} + a_1^{(i)} i_1 + a_2^{(i)} i_2 + a_3^{(i)} i_3.$$

Let  $A$  be the 3 by 3 matrix whose  $(i, j)$ th entry is  $a_j^{(i)}$  for  $1 \leq i, j \leq 3$ . We need to show that  $\det(A) = 0$ . Let  $\mathcal{L}$  have shape  $(M_1, M_2, M_3)$  with  $M_1 M_2 M_3 \asymp N$ , and let the integers  $\delta_{i,j}$  be as in (19). Therefore we have integers  $u_j^{(i)}$  such that for  $1 \leq i, j \leq 3$ , we have

$$(32) \quad a_j^{(i)} = u_1^{(i)} M_1 \delta_{1,j} + u_2^{(i)} M_2 \delta_{2,j} + u_3^{(i)} M_3 \delta_{3,j}.$$

Writing the above system of equations in matrix form, we see immediately that  $M_1 M_2 M_3$  divides  $\det(A)$ . On the other hand, by Lemma 2.13, we have that  $\det(A) \ll L^3$  where the implied constant depends on  $\delta$ ,  $\mathcal{J}$  and  $\iota_{\infty}$ . So if  $L^3 \ll N$ , we must have  $\det(A) = 0$ , as desired. □



Now let  $\mathcal{O} \subseteq \mathcal{O}^{\max}$  be an order of level  $N$ . The above two lemmas imply that if  $1 \leq m < CN^{1/3}$ , then all elements of  $\mathcal{O}(m; z, \delta)$  lie in a *quadratic field* (since the only proper  $Q$ -algebras in a given quaternion algebra are  $\mathbb{Q}$  and various embedded quadratic fields). Now the proof of Proposition 2.14 follows from the following lemma and the fact that  $d(m) \ll_{\epsilon} m^{\epsilon}$ .

**Lemma 2.17.** *Let  $F \subset D$  be a quadratic field. Then for any  $z \in \mathbb{H}$ , and any positive integer  $m$ , we have*

$$(33) \quad |F(m; z, \delta)| \ll_{\delta} d(m),$$

where  $d(m)$  denotes the divisor function and the implied constant is independent of  $F$ .

*Proof.* Any element of  $F(m; z, \delta)$  is a product of a unit in  $\mathcal{O}_F^{\times}$  of norm 1, and an element of  $\mathcal{O}_F$  of norm  $m$ , with the latter taken from a fixed set of cardinality  $\ll d(m)$ . So we only need to consider the action of units. By the proof of Lemma 6.4 of [21], the number of norm 1 units  $\kappa \in \mathcal{O}_F^{\times}$  satisfying  $u(z, \iota_{\infty}(\kappa)z) \leq \delta$  is  $\ll_{\delta} 1$ . This completes the proof.  $\square$

### 3. THE MAIN RESULT: STATEMENT AND KEY APPLICATIONS

**3.1. Basic notations.** We continue to use the notations established in Section 2.2, and introduce some new ones below. Let  $\mathbf{f}$  denote the finite places of  $\mathbb{Q}$  (which we identify with the set of primes) and  $\infty$  the archimedean place. We let  $\mathbb{A}$  denote the ring of adèles over  $\mathbb{Q}$ , and  $\mathbb{A}_{\mathbf{f}}$  the ring of finite adèles. Given an algebraic group  $H$  defined over  $\mathbb{Q}$ , a place  $v$  of  $\mathbb{Q}$ , a subset of places  $U$  of  $\mathbb{Q}$ , and a positive integer  $M$ , we denote  $H_v := H(\mathbb{Q}_v)$ ,  $H_U := \prod_{v \in U} H_v$ ,  $H_M := \prod_{p|M} H_p$ . Given an element  $g$  in  $H(\mathbb{Q})$  (resp., in  $H(\mathbb{A})$ ), we will use  $g_p$  to denote the image of  $g$  in  $H_p$  (resp., the  $p$ -component of  $g$ ); more generally for any set of places  $U$ , we let  $g_U$  the image of  $g$  in  $H_U$ .

Recall that  $D$  is an indefinite quaternion division algebra over  $\mathbb{Q}$  with reduced discriminant  $d$  and that we have fixed a maximal order  $\mathcal{O}^{\max}$ . We denote  $G = D^{\times}$  and  $G' = PD^{\times} = D^{\times}/Z$  where  $Z$  denotes the center of  $D^{\times}$ . For each prime  $p$ , let  $K_p = (\mathcal{O}_p^{\max})^{\times}$  and let  $K'_p$  denote the image of  $K_p$  in  $G'_p$ . Thus, for  $p|d$ ,  $K'_p$  has index 2 in the compact group  $G'_p$ .

For each place  $v$  that is not among the primes dividing  $d$ , fix an isomorphism  $\iota_v : D_v \xrightarrow{\cong} M(2, \mathbb{Q}_v)$ . We assume that these isomorphisms are chosen such that for each finite prime  $p \nmid d$ , we have  $\iota_p(\mathcal{O}_p) = M(2, \mathbb{Z}_p)$ . It is well known that all such choices are conjugate to each other by some matrix in  $\mathrm{GL}_2(\mathbb{Z}_p)$ . By abuse of notation, we also use  $\iota_v$  to denote the composition map  $D(\mathbb{Q}) \rightarrow D_v \rightarrow M(2, \mathbb{Q}_v)$ .

We fix the Haar measure on each group  $G_p$  such that  $\mathrm{vol}(K_p) = 1$ . We fix Haar measures on  $\mathbb{Q}_p^{\times}$  such that  $\mathrm{vol}(\mathbb{Z}_p^{\times}) = 1$ . This gives us resulting Haar measures on each group  $G'_p$  such that  $\mathrm{vol}(K'_p) = 1$ . Fix any Haar measure on  $G_{\infty}$ , and take the Haar measure on  $\mathbb{R}^{\times}$  to be equal to  $\frac{dx}{|x|}$  where  $dx$  is the Lebesgue measure. This gives us a Haar measure on  $G'_{\infty}$ . Take the measures on  $G(\mathbb{A})$  and  $G'(\mathbb{A})$  to be given by the product measure.

For each continuous function  $\phi$  on the space  $G(\mathbb{A})$ , we let  $R(g)$  denote the right regular action, given by  $(R(g)\phi)(h) = \phi(hg)$ . If a continuous function  $\phi$  on  $G(\mathbb{A})$

satisfies that  $|\phi|$  is left  $Z(\mathbb{A})G(\mathbb{Q})$  invariant, define

$$\|\phi\|_2 = \int_{G'(\mathbb{Q}) \backslash G'(\mathbb{A})} |\phi(g)|^2 dg.$$

Note above that  $G'(\mathbb{Q}) \backslash G'(\mathbb{A})$  is compact, so convergence of the integral is not an issue.

**3.2. Some facts on orders and their localizations.** We recall some facts we will need. Proofs of these standard facts can be found, e.g., in [23].

For any lattice  $\mathcal{L} \subseteq \mathcal{O}^{\max}$  of  $D$ , we get a local lattice  $\mathcal{L}_p$  of  $D_p$  by localizing at each prime  $p$ . These collection of lattices satisfy

$$(34) \quad \mathcal{L} = \{g \in D : g_p \in \mathcal{L}_p \text{ for all primes } p\}.$$

Conversely, if we are given a collection of local lattices  $\{\mathcal{L}_p\}_{p \in \mathfrak{f}}$ , such that  $\mathcal{L}_p \subseteq \mathcal{O}_p^{\max}$  for all  $p$  and  $\mathcal{L}_p = \mathcal{O}_p^{\max}$  for all but finitely many  $p$ , then there exists a unique lattice  $\mathcal{L} \subseteq \mathcal{O}^{\max}$  of  $D$  defined via (34) and whose localizations at primes  $p$  are precisely the  $\mathcal{L}_p$ . We will refer to  $\mathcal{L}$  as the global lattice corresponding to the collection of local lattices  $\{\mathcal{L}_p\}_{p \in \mathfrak{f}}$ .

Two orders of  $D$  (or of  $D_p$ ) are isomorphic as algebras if and only if they are conjugate by an element of  $D^\times$  (respectively, by an element of  $D_p^\times$ ). Given an order  $\mathcal{O}$  of  $D$ , we define a compact subgroup of  $G(\mathbb{A}_{\mathfrak{f}})$  by

$$K_{\mathcal{O}} = \prod_{p \in \mathfrak{f}} \mathcal{O}_p^\times.$$

We define the shape and level of an order as in Section 2.2. These quantities have obvious local analogues, and so for each order  $\mathcal{O}_p \subseteq \mathcal{O}_p^{\max}$  of  $D_p$ , we can define its shape  $(p^{m_{1,p}}, p^{m_{2,p}}, p^{m_{3,p}})$  and level  $p^{n_p}$ . It is easy to see that  $\text{vol}(\mathcal{O}_p^\times) \asymp p^{-n_p}$ . If  $\mathcal{O}$  is the global order of shape  $(M_1, M_2, M_3)$  and level  $N$  corresponding to the collection of local orders  $\{\mathcal{O}_p\}_{p \in \mathfrak{f}}$  with shape and level as above, then for  $i = 1, 2, 3$  we have:  $M_i = \prod_{p \in \mathfrak{f}} p^{m_{i,p}}$ ,  $N = \prod_{p \in \mathfrak{f}} p^{n_p}$ . From this it follows that

$$(35) \quad N^{-1-\epsilon} \ll_{\epsilon} \text{vol}(K_{\mathcal{O}}) \ll_{\epsilon} N^{-1+\epsilon}.$$

For each  $g \in G(\mathbb{A}_{\mathfrak{f}})$ , and an order  $\mathcal{O}$  of  $D$ , we let  $g \cdot \mathcal{O}$  denote the order whose localization at each prime  $p$  equals  $g_p \mathcal{O}_p g_p^{-1}$ . An order is said to be locally isomorphic to (in the same genus as)  $\mathcal{O}$  if and only if it is equal to  $g \cdot \mathcal{O}$  for some  $g \in G(\mathbb{A}_{\mathfrak{f}})$ . Note that

$$(36) \quad g \cdot \mathcal{O} \subseteq \mathcal{O}^{\max} \iff g_p \mathcal{O}_p g_p^{-1} \in \mathcal{O}_p^{\max} \forall p.$$

Note also that

$$(37) \quad g K_{\mathcal{O}} g^{-1} = K_{g \cdot \mathcal{O}}.$$

Given  $g$  satisfying (36), the orders  $\mathcal{O}$  and  $g \cdot \mathcal{O}$  need not be isomorphic or have the same shape, however they always have the same level. However, note that if  $k \in K_{\mathcal{O}^{\max}}$ , then  $k \cdot \mathcal{O}$  has exactly the same shape as  $\mathcal{O}$ .

**3.3. Statement of main result.** Let  $\pi = \otimes_v \pi_v$  be an irreducible, unitary, cuspidal automorphic representation of  $G(\mathbb{A})$  where we identify  $V_\pi$  with a (unique) subspace of functions on  $G(\mathbb{A})$ , so that  $\pi(g)$  coincides with  $R(g)$  on that subspace. Given a compact open subgroup  $K' = \prod_{p \in \mathfrak{f}} K'_p$  of  $K_{\mathcal{O}^{\max}}$  (where each  $K'_p$  is a subgroup of  $K_p$ , with  $K'_p = K_p$  for almost all  $p$ ) and a finite dimensional representation  $\rho$  of  $K'$ , we say that an automorphic form  $\phi \in V_\pi$  is of  $(K', \rho)$ -type if the right-regular action of  $K'$  on  $\phi$  generates a representation isomorphic to  $\rho$ . Observe that the existence of a form of  $(K', \rho)$ -type implies that the restrictions of  $\rho$  and  $\omega_\pi$  to the centre of  $K'$  must coincide.

We can now state our main theorem.

**Theorem 1.** *Let  $\mathcal{O} \subseteq \mathcal{O}^{\max}$  be an order of  $D$  of level  $N$  and let  $\rho$  be a finite dimensional representation of  $K_{\mathcal{O}}$ . Let  $N_1$  be the smallest positive integer such that  $N$  divides  $N_1^2$ . Let  $\pi = \otimes_v \pi_v$  be an irreducible, unitary, cuspidal automorphic representation of  $G(\mathbb{A})$ . Let  $\phi \in V_\pi$  be of  $(K_{\mathcal{O}}, \rho)$ -type and assume that  $\phi$  is of minimal weight at the archimedean place, i.e.,*

$$(38) \quad \phi \left( g \left( \iota_\infty^{-1} \begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix} \right) \right) = e^{ik\theta} \phi(g)$$

where  $k = 0$  if  $\pi_\infty$  is principal series, and  $k$  is the lowest weight if  $\pi_\infty$  is discrete series. Then

$$(39) \quad \sup_{g \in G(\mathbb{A})} |\phi(g)| \ll_{D, \pi_\infty, \epsilon} N^\epsilon \min(\max(N^{1/3}, N_1^{1/2}), N^{11/24}) \dim(\rho)^{1/2} \|\phi\|_2.$$

Theorem A of the introduction is a special case of Theorem 1, where we take  $\rho$  to be a character. A key flexibility of Theorem 1 comes from the fact that given  $\phi$ , one can optimise which order  $\mathcal{O}$  to use depending on how much information one has about the dimensions of the representations  $\rho$  generated under the action of various  $K_{\mathcal{O}}$ .

In certain cases, however, one may only know the dimension under the action of some  $K'$  that is *not* of the form  $K_{\mathcal{O}}$ . In such cases one can still get a bound by working with any order  $\mathcal{O}$  containing  $K'$ . The following corollary makes this precise.

**Corollary 3.1.** *Let  $\phi$  be an automorphic form in the space of  $\pi$  such that  $\|\phi\|_2 = 1$  and (38) holds. Suppose that  $\phi$  is of  $(K', \rho)$ -type for some compact open subgroup  $K'$  of  $K_{\mathcal{O}^{\max}}$  and some finite dimensional representation  $\rho$  of  $K'$ . Let  $\mathcal{O} \subseteq \mathcal{O}^{\max}$  be any order of  $D$  of level  $N$  such that  $K' \subseteq K_{\mathcal{O}}$ . Let  $N_1$  be the smallest positive integer such that  $N$  divides  $N_1^2$ . Then*

$$(40) \quad \sup_{g \in G(\mathbb{A})} |\phi(g)| \ll_{D, \pi_\infty, \epsilon} N^\epsilon \min(\max(N^{1/3}, N_1^{1/2}), N^{11/24}) \dim(\rho)^{1/2} [K_{\mathcal{O}} : K']^{1/2}.$$

*Proof.* Consider the representation  $\rho'$  generated by  $\phi$  under the action of  $K_{\mathcal{O}}$ . Then from elementary considerations,

$$\dim(\rho') \leq \dim(\rho) [K_{\mathcal{O}} : K'].$$

Now apply Theorem 1 using the fact that  $\phi$  is of  $(K_{\mathcal{O}}, \rho')$ -type.  $\square$

**Remark 3.2.** *Suppose that  $\phi \in V_\pi$  is an automorphic form satisfying (38) and suppose that  $K'$  is a compact open subgroup that acts on  $\phi$  by a character. Taking  $\mathcal{O} = \mathcal{O}^{\max}$  in Corollary 3.1 then gives us the trivial bound:*

$$\|\phi\|_\infty \ll_{D, \pi_\infty, \epsilon} [K_{\mathcal{O}^{\max}} : K']^{1/2} \|\phi\|_2,$$

which is a mild extension of (2).

On the other hand, suppose that  $\pi$  has trivial central character and is spherical at all  $p|d$ . Denote the conductor of  $\pi$  by  $C$  and let  $C_1$  be the smallest integer such that  $C|C_1^2$ . Suppose that  $\phi \in V_\pi$  has the property that some  $G(\mathbb{A}_f)$  translate of it is fixed by the “principal congruence subgroup”  $K_{\mathcal{O}^{\max}(C_1)}$  (see (7) for the definition). Then by the results of [19, p. 96-97], the action of  $K_{\mathcal{O}}$  on  $\phi$  generates a representation of dimension  $\leq C_1 \prod_{p|C} (1 + p^{-1})$ , and so by Corollary 3.1:

$$(41) \quad \|\phi\|_\infty \ll_{D, \pi_\infty} C_1^{1/2} \left( \prod_{p|C} (1 + p^{-1}) \right)^{1/2}.$$

This is a refinement of the local bound (8) for such  $\phi$ . We note here that the class of  $\phi$  having the property mentioned above is quite broad and includes the usual newforms, the automorphic forms of minimal type, and the  $p$ -adic microlocal lifts, as well as  $G(\mathbb{A}_f)$ -translates of all these.

**3.4. The case of automorphic forms of minimal type.** We now explain how Theorem 1 implies Theorem B. In fact, we will provide a more general version of Theorem B in Theorem 2 below. Before stating the theorem, let us briefly recall the concept of a minimal vector. Let  $p$  be an odd prime and  $\pi_p$  be a twist-minimal supercuspidal representation of  $\mathrm{GL}_2(\mathbb{Q}_p)$  of conductor  $p^{c_p}$ . (The twist-minimal condition is automatic whenever  $\pi_p$  has trivial central character, or more generally whenever  $c_p \neq 2m_p$  where  $p^{m_p}$  is the conductor of the central character of  $\pi_p$ .) Note that as  $\pi_p$  is supercuspidal, we must have  $c_p \geq 2$ . We define integers  $n_p, d_p$  as follows depending on the congruence class of  $c_p \pmod 4$ :

- (1) If  $c_p \equiv 0 \pmod 4$ , then  $n_p = \frac{c_p}{2}$ ,  $d_p = 0$ .
- (2) If  $c_p \equiv 2 \pmod 4$ , then  $n_p = \frac{c_p}{2} - 1$ , and  $d_p = 1$ .
- (3) If  $c_p \equiv 1 \pmod 2$ , then  $n_p = \frac{c_p}{2} + \frac{1}{2}$ ,  $d_p = 0$ .

The concept of a minimal vector was first introduced in [11] where we focussed only on the first case above, i.e., the case  $c_p \equiv 0 \pmod 4$ . In this case, the minimal vector is a unique (up to multiples) vector  $\phi_p$  in the space of  $\pi_p$  that can be described as follows: Let  $\alpha_p \in \mathbb{Z}_p^\times$  be a non-square. Define the order  $\tilde{\mathcal{O}}_p$  of  $\mathrm{GL}_2(\mathbb{Z}_p)$  via

$$\tilde{\mathcal{O}}_p = \mathbb{Z}_p + \mathbb{Z}_p \begin{pmatrix} 0 & 1 \\ \alpha_p & 0 \end{pmatrix} + p^{c_p/4} M_2(\mathbb{Z}_p)$$

and let  $\mathcal{O}_p = \iota_p^{-1}(\tilde{\mathcal{O}}_p)$  be the corresponding order of  $D_p$ . Then there exists a character  $\chi_{\pi_p}$  of  $\mathcal{O}_p^\times$  (defined in Definition 2.10 of [11]) such that  $\pi_p(k_p)\phi_p = \chi_{\pi_p}(k_p)\phi_p$ , for all  $k_p \in \mathcal{O}_p^\times$ . This property defines the minimal vector uniquely up to multiples (the definition depends on the isomorphism  $\iota_p$  and the element  $\alpha_p$  but a different choice of these simply corresponds to a  $\mathcal{O}_p^\times$  translate of  $\phi_p$ ).

In a recent work [10], Hu and Nelson extended the concept of a minimal vector to cases (2) and (3) above, so that now there is a well-defined notion of a minimal vector for all twist-minimal supercuspidal representations of  $\mathrm{GL}_2(\mathbb{Q}_p)$  for  $p$  odd. We remark here that the twist-minimality is merely for convenience since the minimal vector in the general case is defined in terms of the twist-minimal case. In principle, the case of  $p = 2$  can also be dealt with similarly but in this case the computations get more technical and these have not been performed so far. The analogous vectors for the case of principal series representations has also been dealt with in separate work of Nelson [15]; in this case the relevant vectors are known as  $p$ -adic microlocal lifts.

Going back to the case of a twist-minimal supercuspidal representation  $\pi_p$  of  $\mathrm{GL}_2(\mathbb{Q}_p)$  for  $p$  odd, we define a ‘‘Type 2 minimal vector’’ as in [10]. If  $c_p \not\equiv 2 \pmod{4}$ , then the relevant space is one dimensional and so any minimal vector is automatically of Type 2. In the case  $c_p \equiv 2 \pmod{4}$ , the space of minimal vectors is  $p$ -dimensional (except for the case  $c_p = 2$ , when the space is  $p - 1$  dimensional) and has a basis consisting of Type 2 minimal vectors.

A Type 2 minimal vector  $\phi_p$  in the space of  $\pi_p$  has the property that there exists an order  $\mathcal{O}_p \in \mathcal{O}_p^{\max}$  of level  $p^{n_p}$  such that the action of  $\mathcal{O}_p^\times$  on  $\phi_p$  generates an irreducible representation  $\rho_p$  of  $\mathcal{O}_p^\times$  with dimension  $p^{d_p}$ , except in the special case  $c_p = 2$ , in which the representation has dimension  $p - 1$ . Now Theorem 1 leads to the following theorem.

**Theorem 2.** *Let  $\pi = \otimes_v \pi_v$  be an irreducible, unitary, cuspidal automorphic representation of  $G(\mathbb{A})$ . Assume that*

- *If  $p|d$ , then  $\pi_p$  has a (non-zero) vector fixed by  $K_p$ . (This implies that  $\pi_p$  is one-dimensional for each  $p|d$ .)*
- *If  $p \nmid d$ , and the representation  $\pi_p$  of  $G_p \cong \mathrm{GL}_2(\mathbb{Q}_p)$  is ramified, then  $p$  is odd and  $\pi_p$  is a twist-minimal supercuspidal representation with conductor  $p^{c_p}$ .*

Define  $C = \prod_p p^{c_p}$  with the product taken over the ramified primes, so that  $C$  is the (‘‘away from  $d$ ’’ part of the) conductor of  $\pi$  and equal to an odd squarefull integer. Let  $\phi = \otimes_v \phi_v$  be a non-zero automorphic form in the space of  $\pi$  such that  $\phi_p$  is  $K_p$  fixed for all  $p \nmid C$ ,  $\phi_\infty$  is a vector of smallest non-negative weight, and  $\phi_p$  is a Type 2 minimal vector for each  $p|C$ . Then we have

$$(42) \quad \sup_{g \in G(\mathbb{A})} |\phi(g)| \ll_{D, \pi_\infty, \epsilon} C_1^{1/3+\epsilon} \|\phi\|_2 \prod_{\substack{p|C \\ c_p \equiv 2 \pmod{4}}} p^{1/6}.$$

*Proof.* For each  $p|C$ , we have a local order  $\mathcal{O}_p$  of level  $p^{n_p}$  such that the action of  $\mathcal{O}_p^\times$  on  $\phi_p$  generates a representation  $\rho_p$  of dimension  $p^{d_p}$ , except if  $c_p = 2$ , when the dimension is  $p^{d_p} \left(1 - \frac{1}{p}\right)$ . Let  $\mathcal{O}$  be the corresponding global order (put  $\mathcal{O}_p = \mathcal{O}_p^{\max}$  if  $p \nmid C$ ) and  $\rho = \otimes_p \rho_p$  the corresponding representation of  $K_\mathcal{O}$ . Then the dimension of  $\rho$  is

$$r := \prod_{\substack{p|C \\ c_p \equiv 2 \pmod{4}}} p \prod_{c_p=2} \left(1 - \frac{1}{p}\right)$$

and the level of  $\mathcal{O}$  is  $\prod_{p|C} p^{n_p}$ . Now the result follows immediately from Theorem 1.  $\square$

Theorem 2 improves upon the local bound (8) except when  $\sqrt{C}$  is a squarefree integer (in which case we recover the local bound).

**3.5. Bounds for  $p$ -adic microlocal lifts and for newforms.** In fact, Theorem 1 implies sub-local bounds in the level aspect for certain families of automorphic forms in addition to the ones of minimal type described above. For example, consider the case where  $\pi$  has trivial central character and whose (away-from- $d$ ) conductor  $C$  equals  $N^4$  where  $N = \prod_p p^{n_p}$  is an odd integer. For two characters  $\chi_1, \chi_2$  on  $\mathbb{Q}_p^\times$ , let  $\chi_1 \boxplus \chi_2$  denote the principal series representation on  $\mathrm{GL}_2(\mathbb{Q}_p)$  that is unitarily induced from the corresponding representation of its Borel subgroup. Now assume

that for each  $p|N$ ,  $\pi_p$  is of the form  $\chi_p \boxplus \chi_p^{-1}$  with  $a(\chi_p) = 2n_p$ . Let  $\phi_p \in V_{\pi_p}$  at these primes  $p$  correspond to  $p$ -adic microlocal lifts in the sense of [15].

Consider the group

$$K_p^*(n_p) = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{GL}_2(\mathbb{Z}_p), b \equiv c \equiv 0 \pmod{p^{n_p}} \right\}.$$

Then, by [15, Lemma 22], we see that a  $p$ -adic microlocal lift  $\phi_p \in V_{\pi_p}$  is characterized by the property that  $\pi_p(\iota_p^{-1}(k))\phi_p = \chi_p^{\pm 1}(ab^{-1})\phi_p$  for all  $k = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in K_p^*(n_p)$ .

On the other hand, it is easy to check that  $\iota_p^{-1}(K_p^*(n_p)) = \mathcal{O}_p^\times$  where  $\mathcal{O}_p$  is an order of level  $p^{2n_p}$ . Therefore (by identical arguments as in the proof of Theorem 2), we obtain

$$(43) \quad \|\phi\|_\infty \ll_{D, \pi_\infty, \epsilon} C^{\frac{1}{6} + \epsilon},$$

which is a sub-local bound for sup-norms of “automorphic forms of  $p$ -adic microlocal type”.

**Remark 3.3.** *Minimal vectors and  $p$ -adic microlocal lifts are analogues of each other, the only difference being that the former live in supercuspidal representations and the latter live in principal series representations. Both these classes of vectors may be viewed as special cases (in the  $p$ -adic setting) of the more general class of “localized” vectors. See also [16] for a discussion of localized vectors in the archimedean setting, where they are termed “microlocalized” vectors.*

Finally, we discuss what sort of bound Theorem 1 gives us for *newforms*. We obtain the following general result:

**Theorem 3.** *Let  $\pi = \otimes_v \pi_v$  be an irreducible, unitary, cuspidal automorphic representation of  $G(\mathbb{A})$  with conductor  $C$ . Let  $M$  be the conductor of the central character of  $\pi$ . Let  $\phi$  in the space of  $\pi$  be a global newform, i.e.,  $\phi = \otimes_v \phi_v$  with  $\phi_p$  spherical if  $p \nmid C$ ,  $\phi_p$  equal to the local newvector for  $p|C$ , and  $\phi_\infty$  a vector of smallest non-negative weight. Then we have*

$$(44) \quad \sup_{g \in G(\mathbb{A})} |\phi(g)| \ll_{D, \pi_\infty, \epsilon} C^\epsilon \min(\max(C^{\frac{1}{3}}, C_1^{\frac{1}{2}}), (C_1^2/C)^{-\frac{1}{24}} \mathrm{lcm}(M, C_1)^{\frac{1}{2}}) \|\phi\|_2.$$

*Proof.* For any integer  $N$ , let  $\mathcal{O}_0(N)$  denote the Eichler order of level  $N$ . We first apply Theorem 1 with  $\mathcal{O} = \mathcal{O}_0(C)$ . Since  $\phi$  transforms by character under the action of  $K_{\mathcal{O}_0(C)}$ , we obtain that

$$\sup_{g \in G(\mathbb{A})} |\phi(g)| \ll_{D, \pi_\infty, \epsilon} C^\epsilon \max(C^{1/3}, C_1^{1/2}).$$

Next, we apply the theorem with  $\mathcal{O} = \mathcal{O}_0(C')$  where  $C' = C_1^2/C$  is the squarefree integer obtained by taking the product of all primes which divide  $C$  to an odd power. Then, it was shown in [17, Sec. 2.7] that the action of  $K_{\mathcal{O}}$  on (a suitable right-translate of)  $\phi$  generates a representation of dimension  $\ll \frac{\mathrm{lcm}(M, C_1)}{C'}$ . Now applying Theorem 1 (and using the fact that right-translating does not change the sup-norm), we get that

$$\sup_{g \in G(\mathbb{A})} |\phi(g)| \ll_{D, \pi_\infty, \epsilon} C^\epsilon (C')^{-1/24} \mathrm{lcm}(M, C_1)^{1/2}.$$

This completes the proof. □

Theorem 3 generalizes several previously known bounds for the supnorms of newforms on  $G(\mathbb{A})$ , and its proof clearly demonstrates the flexibility of Theorem 1. Note however, that when the central character of  $\pi$  is trivial, then Theorem 3 reduces to

$$(45) \quad \|\phi\|_\infty \ll_{D, \pi_\infty, \epsilon} C_1^{\frac{1}{2} + \epsilon} (C')^{-1/24}$$

where  $C' = C_1^2/C$  is the squarefree integer obtained by taking the product of all primes which divide  $C$  to an odd power. This bound (45) *fails* to improve upon the local bound (8) for any family of newforms  $\phi \in V_\pi$  for which  $\frac{C'}{C} \rightarrow 0$ .

In particular, a key outstanding case concerns the problem of beating the local bound for newforms of trivial central character in the depth aspect  $C = p^n$ ,  $n \rightarrow \infty$  where  $p$  is a fixed prime. This case will be treated in a sequel to this paper written with Y. Hu, where we will introduce a new technique in the setting described above (under an additional assumption that  $p$  is odd) which will enable us to replace the exponent  $1/2$  in (8) by the exponent  $5/24$  in this particular aspect.

**3.6. Preparations for the proof.** We now begin the proof of Theorem 1. The main part of the proof will be completed in Section 4 (which will crucially rely on the counting results from Section 2). In this subsection, we make a few simple but key observations, which will allow us to impose additional hypotheses without any loss of generality.

First of all, the property of being of minimal weight at the archimedean place, strictly speaking, depends on the local isomorphism  $\iota_\infty$  which has been fixed by us. However, a different choice of  $\iota_\infty$  simply corresponds to replacing  $\phi$  by a  $G_\infty$ -translate of it, and (by definition) the sup-norm of this translated form coincides with the sup-norm of  $\phi$ . Therefore, fixing  $\iota_\infty$  does not change the sup-norm. Now we fix, once and for all, a compact fundamental domain  $\mathcal{J}$  for the action of

$$\Gamma_{\mathcal{O}^{\max}} = \{\gamma \in \iota_\infty(\mathcal{O}^{\max}), \det(\gamma) = 1\}$$

on  $\mathbb{H}$ . Any element of  $G(\mathbb{A})$  can be left-multiplied by a suitable element of  $Z(\mathbb{A})G(\mathbb{Q})$  and right-multiplied by a suitable element of  $K_{\mathcal{O}^{\max}}$  to get an element  $g_\infty \in G_\infty$  such that  $\det(\iota_\infty(g_\infty)) > 0$  and  $\iota_\infty(g_\infty)(i)$  lies in  $\mathcal{J}$ . Since  $|\phi(g)|$  is  $Z(\mathbb{A})G(\mathbb{Q})$  invariant, we may assume, for the purposes of proving (39), that  $g = \prod_v g_v$  satisfies

$$(46) \quad g_p \in K_p \text{ for all } p \in \mathfrak{f}, \quad \det(\iota_\infty(g_\infty)) > 0, \quad \text{and } \iota_\infty(g_\infty)(i) \in \mathcal{J},$$

where  $\mathcal{J}$  is our fixed compact set.

Secondly, suppose that  $\mathcal{O}' \subseteq \mathcal{O}^{\max}$  is any order in the same genus as  $\mathcal{O}$ . So there exists  $h \in G(\mathbb{A}_\mathfrak{f})$  such that  $h \cdot \mathcal{O} = \mathcal{O}'$  (recall the notations from Section 3.2). Clearly  $\mathcal{O}'$  has the same level as  $\mathcal{O}$ . Let the finite dimensional representation  $\rho'$  of  $K_{\mathcal{O}'}$  be defined via  $\rho'(k) = \rho(h^{-1}kh)$  (So  $\rho$  and  $\rho'$  are isomorphic). Now define the automorphic form  $\phi' = \pi(h)\phi$ . Then  $\phi'$  is of  $(K_{\mathcal{O}'}, \rho')$  type and of minimal weight at the archimedean place. Further, it has the same sup-norm as  $\phi$ , being a translate. So it suffices to prove the Theorem for  $\phi'$  (which allows us to change the order from  $\mathcal{O}$  to  $\mathcal{O}'$ ). However, a very useful algebraic fact, that we will prove below in Section 3.7, is that each genus of orders contains an order with shape  $(M_1, M_2, M_3)$  and level  $N$  such that  $M_3|N_1$ . So, for the purpose of proving Theorem 1, we can and will assume the following:

$$(47) \quad \text{If } (M_1, M_2, M_3) \text{ is the shape of } \mathcal{O}, \text{ then } M_3 \text{ divides } N_1.$$

Thirdly, we may assume, for the purpose of proving Theorem 1, that

$$(48) \quad \rho \text{ is irreducible.}$$

Indeed, suppose we have proved the Theorem under (48). Now for the general case, we simply write  $\phi$  as an orthogonal sum of automorphic forms  $\phi_i$ , each of which generates an irreducible representation  $\rho_i$  under the action of  $K_{\mathcal{O}}$ . Now apply the already proved result to each  $\phi_i$ , follow it by the triangle inequality and then Cauchy Schwartz, to obtain the desired result for  $\phi$  (using the facts that  $\dim(\rho) = \sum_i \dim(\rho_i)$  and  $\|\phi\|_2^2 = \sum_i \|\phi_i\|_2^2$ ).

### 3.7. A result on balanced representatives for orders.

**Definition 3.4.** *Given a pair of lattices  $\mathcal{L}_1, \mathcal{L}_2$  in  $D$  such that  $\mathcal{L}_1 \subseteq \mathcal{L}_2$ ,*

- (1) *The invariant factors of  $\mathcal{L}_1$  in  $\mathcal{L}_2$  are the unique quadruple of positive integers  $(a_1, a_2, a_3, a_4)$  such that  $a_1|a_2|a_3|a_4$  and*

$$\mathcal{L}_2/\mathcal{L}_1 \simeq (\mathbb{Z}/a_1\mathbb{Z}) \times (\mathbb{Z}/a_2\mathbb{Z}) \times (\mathbb{Z}/a_3\mathbb{Z}) \times (\mathbb{Z}/a_4\mathbb{Z}).$$

- (2)  *$\mathcal{L}_1$  is balanced in  $\mathcal{L}_2$  if the invariant factors  $(a_1, a_2, a_3, a_4)$  have the following property: If  $t_1$  denotes the smallest integer such that  $a_1 a_2 a_3 a_4$  divides  $t_1^2$ , then  $a_4$  divides  $t_1$ .*

Note that if  $\mathcal{L}_1$  and  $\mathcal{L}_2$  are orders, then the smallest invariant factor  $a_1$  equals 1.

**Remark 3.5.** *Let  $\mathcal{O} \subseteq \mathcal{O}^{\max}$  be an order with shape  $(M_1, M_2, M_3)$  and level  $N$ , and let  $N_1$  be the smallest integer such that  $N|N_1^2$ . Now, suppose that  $\mathcal{O}$  is balanced in  $\mathcal{O}^{\max}$ . Then by Remark 2.7, we see that  $M_3|N_1$ . In particular assumption (47) holds.*

The object of this subsection is to prove the following result, which was used in the previous subsection to show that we can always assume (47) without any loss of generality.

**Proposition 3.6.** *Let  $\mathcal{O}$  be an order in  $D$ . Then there exists  $g \in G(\mathbb{A}_{\mathbf{f}})$  such that  $g \cdot \mathcal{O}$  is balanced in  $\mathcal{O}^{\max}$ .*

To prove the above Proposition, we first of all recall (see, e.g., [23, Chapter 24]) that the order  $\mathcal{O}$  can be written as  $\mathcal{O} = \mathbb{Z} + f\mathcal{O}^{\text{gor}}$  where  $f \in \mathbb{Z}$  and  $\mathcal{O}^{\text{gor}}$  is a Gorenstein order. If for some  $g \in G(\mathbb{A}_{\mathbf{f}})$  we know that  $g \cdot \mathcal{O}^{\text{gor}} \subseteq \mathcal{O}^{\max}$  and that the invariant factors of  $g \cdot \mathcal{O}^{\text{gor}}$  in  $\mathcal{O}^{\max}$  are  $(1, a_2, a_3, a_4)$  then it easily follows that the invariant factors of  $g \cdot \mathcal{O}$  in  $\mathcal{O}^{\max}$  are  $(1, fa_2, fa_3, fa_4)$ . In particular,  $g \cdot \mathcal{O}$  is balanced in  $\mathcal{O}^{\max}$  whenever  $g \cdot \mathcal{O}^{\text{gor}}$  is. So it suffices to prove Proposition 3.6 for Gorenstein orders.

Being Gorenstein is a local property. Now from the local-global principle for orders (see Section 3.2), Proposition 3.6 follows from the next statement.

**Proposition 3.7.** *Let  $\mathcal{O}_p \subseteq \mathcal{O}_p^{\max}$  be a Gorenstein order of  $D_p$ . Then there exists an order  $\mathcal{O}'_p$  of  $D_p$  with the following properties:*

- (1)  $\mathcal{O}'_p \simeq \mathcal{O}_p$ ,  
(2)  $\mathcal{O}'_p \subseteq \mathcal{O}_p^{\max}$ ,  
(3) *If  $(m_1, m_2, m_3)$  are the unique triple of non-negative integers such that  $m_1 \leq m_2 \leq m_3$  and there is an isomorphism as  $\mathbb{Z}_p$ -modules*

$$\mathcal{O}_p^{\max}/\mathcal{O}'_p \simeq (\mathbb{Z}_p/p^{m_1}\mathbb{Z}_p) \times (\mathbb{Z}_p/p^{m_2}\mathbb{Z}_p) \times (\mathbb{Z}_p/p^{m_3}\mathbb{Z}_p), \text{ then}$$

$$(49) \quad m_3 \leq \left\lceil \frac{m_1 + m_2 + m_3}{2} \right\rceil \text{ holds.}$$



We now prove Proposition 3.7. We rely heavily on the work of Brzezinski [6] who gives a complete list of Gorenstein orders (up to isomorphism) and their resolutions in terms of explicit linear combinations of generators of  $\mathcal{O}^{\max}$ . It is therefore easy (albeit tedious) to compute the triple  $(m_1, m_2, m_3)$  for each order in his list (by bringing the corresponding matrices to Smith normal form). We do this and observe that most orders in his list already satisfy (49); the ones that aren't can be conjugated by a simple element and made to satisfy it. We give the key details below, omitting some of the routine calculations.

First consider the case when  $D_p \simeq M_2(\mathbb{Q}_p)$ . Put  $x_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ ,  $x_2 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ ,  $x_3 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$ . Note that  $\mathcal{O}_p^{\max} = \langle 1, x_1, x_2, x_3 \rangle$ . According to Prop. 5.4 of [6],  $\mathcal{O}_p$  is isomorphic to one of the cases (a) – (d'\_3) described there. We denote  $r_n = \begin{pmatrix} p^{\lfloor \frac{n}{2} \rfloor} & 0 \\ 0 & 1 \end{pmatrix}$ . We write down the required order  $\mathcal{O}'_p$  in each case, using the notation from Proposition 5.4 of [6].

Case (a). In this case we take  $\mathcal{O}'_p = r_n E_n^{(1)} r_n^{-1}$ .

Case (b). In this case we take  $\mathcal{O}'_p = E_n^{(-1)}$ .

Case (c). In this case we take  $\mathcal{O}'_p = E_n^{(0)}$ .

Case (d\_1). In this case we take  $\mathcal{O}'_p = r_n E_{n,s}^{(1)} r_n^{-1}$ .

Case (d\_2). In this case we take  $\mathcal{O}'_p = E_{n,s}^{(-1)}$ .

Case (d\_3). In this case we take  $\mathcal{O}'_p = E_{n,s}^{(0)}$ .

Case (d'\_3). In this case we take  $\mathcal{O}'_p = E_{2,s^+}^{(0)}$ .

Next, we consider the case when  $p \nmid d$ , i.e.  $D_p$  is a division algebra. Let  $x_1, x_2, x_3$  be as in [6, (5.5)]. Then  $\mathcal{O}_p^{\max} = \langle 1, x_1, x_2, x_3 \rangle$ . According to Prop. 5.6 of [6],  $\mathcal{O}_p$  is isomorphic to one of the cases (a) – (c\_2) described there.

Case (a). In this case we take  $\mathcal{O}'_p = \Gamma_n^{(-1)}$ .

Case (b). In this case we take  $\mathcal{O}'_p = \Gamma_n^{(0)}$ .

Case (c\_1). In this case we take  $\mathcal{O}'_p = \Gamma_{n,s}^{(-1)}$ .

Case (c\_2). In this case we take  $\mathcal{O}'_p = \Gamma_{n,s}^{(0)}$ .

In all cases above, the description of  $\mathcal{O}'_p$  given in [6] provides an explicit  $\mathbb{Z}_p$ -basis for  $\mathcal{O}'_p$  in terms of a  $\mathbb{Z}_p$ -basis for  $\mathcal{O}_p^{\max}$ . We reduce the resulting matrix into Smith normal form via elementary operations, and observe that the invariant factors  $(1, p^{m_1}, p^{m_2}, p^{m_3})$  of the resulting matrix always satisfies (49). (We remark here that in the case (d\_2) for  $p \nmid d$ , and the cases (b), (c\_1) for  $p \mid d$ , the reference [6] gives five generators for  $\mathcal{O}'_p$ . However, once the generator matrix is brought into Smith normal form, we get exactly four non-zero rows; these correspond to a  $\mathbb{Z}_p$ -basis of the form required.)

This completes the proof of Proposition 3.7, and therefore of Proposition 3.6.

#### 4. AMPLIFICATION

In this Section, we complete the proof of Theorem 1. Throughout this section, we assume the setup of Section 3.3. We fix an order  $\mathcal{O} \subseteq \mathcal{O}^{\max}$  of level  $N$ , an automorphic form  $\phi$  in  $V_\pi$  with

$$\|\phi\|_2 = 1,$$

and a finite dimensional representation  $\rho$  of  $K_{\mathcal{O}}$  such that the conditions of Theorem 1 are satisfied. Given the above data, and some  $g \in G(\mathbb{A})$  satisfying (46), our goal in this section is to prove

$$(50) \quad |\phi(g)| \ll_{D, \pi_{\infty}, \epsilon} N^{\epsilon} \min(\max(N^{1/3}, N_1^{1/2}), N^{11/24}) \dim(\rho)^{1/2},$$

which will complete the proof of Theorem 1. As explained previously, we can and will assume (without loss of generality) that (47) and (48) hold.

**4.1. Test functions.** Our main tool is the amplification method. From the adelic point of view, amplification corresponds to an appropriate choice of test function  $\kappa$  on  $G(\mathbb{A})$  which increases the contribution of the particular automorphic form  $\phi$  in the resulting pre-trace formula. In this subsection, we describe this test function  $\kappa$  (which will depend on  $\phi$  and  $\mathcal{O}$ ) and note its key properties.

Recall that  $\mathbf{f}$  denotes the set of finite primes, which we identify with the non-archimedean places of  $\mathbb{Q}$ . The representation  $\rho$  is isomorphic to  $\otimes_{p \in \mathbf{f}} \rho_p$  where  $\rho_p$  is a representation of  $\mathcal{O}_p^{\times}$  with  $\rho_p$  trivial for almost all  $p$ . We choose a finite subset  $S \subset \mathbf{f}$  with the following properties:

- (1)  $S$  contains all primes dividing  $dN$ ,
- (2) If  $p \notin S$ , then  $\rho_p$  is trivial.

For convenience, we denote  $G_S = \prod_{p \in S} G_p$ ,  $\mathbb{Q}_S^{\times} = \prod_{p \in S} \mathbb{Q}_p^{\times}$ , and  $\mathcal{O}_S^{\times} = \prod_{p \in S} \mathcal{O}_p^{\times}$ . Put  $\rho_S = \otimes_{p \in S} \rho_p$ . So  $\rho_S$  is an irreducible representation of  $\mathcal{O}_S^{\times}$  with  $\dim(\rho_S) = \dim(\rho)$  and  $V_{\rho_S}$  is the subspace of  $V_{\pi}$  generated by the action of  $\mathcal{O}_S^{\times}$  on  $\phi$ . It follows that  $\rho_S$  is unitary with respect to the Petersson inner product. Let  $\mathbf{ur} = \mathbf{f} \setminus S$  be the set of primes not in  $S$ . We will choose  $\kappa$  of the form  $\kappa = \kappa_S \kappa_{\mathbf{ur}} \kappa_{\infty}$ .

We define the function  $\kappa_S$  on  $G_S$  as follows:

$$\kappa_S(g_S) = \begin{cases} 0 & \text{if } g_S \notin \mathbb{Q}_S^{\times} \mathcal{O}_S^{\times}, \\ \omega_{\pi}^{-1}(z) \langle \phi, \pi(k) \phi \rangle & \text{if } g_S = zk, \quad z \in \mathbb{Q}_S^{\times}, k \in \mathcal{O}_S^{\times}. \end{cases}$$

Our assumptions and basic properties of finite dimensional irreducible representations of compact groups imply that

$$(51) \quad \begin{aligned} R(\kappa_S) \phi &:= \int_{\mathbb{Q}_S^{\times} \backslash G_S} \kappa_S(g) (R(g) \phi) dg \\ &= \int_{\mathcal{O}_S^{\times}} \langle \phi, \rho_S(k) \phi \rangle (\rho_S(k) \phi) dk \\ &= \frac{\text{vol}(K_{\mathcal{O}})}{\dim(\rho)} \phi, \end{aligned}$$

where we have used the fact that  $\text{vol}(\mathcal{O}_S^{\times}) = \text{vol}(K_{\mathcal{O}})$  (since  $\mathcal{O}_p^{\times} = (\mathcal{O}_p^{\max})^{\times}$  for primes  $p \notin S$ ). Observe also that the formula in the first line above gives us a self-adjoint, non-negative operator  $R(\kappa_S)$  on the space of all automorphic forms on  $D^{\times}(\mathbb{A})$  which have central character  $\omega_{\pi}$ .

Next, we consider the primes  $p \in \mathbf{ur}$ . Note that  $\pi_p$  is unramified for each such prime (indeed for such  $p$ ,  $\rho_p$  is trivial and hence  $\phi$  is  $K_p$ -fixed). Let  $\mathcal{H}_{\mathbf{ur}}$  be the set of all compactly supported functions on  $\prod_{p \in \mathbf{ur}} \text{GL}_2(\mathbb{Q}_p)$  that are bi- $\text{GL}_2(\mathbb{Z}_p)$  invariant for each  $p \in \mathbf{ur}$  and transform under the action of the centre by  $\omega_{\pi}^{-1}$ . For each positive integer  $\ell$  satisfying  $(\ell, S) = 1$ , define the functions  $\kappa_{\ell}$  in  $\mathcal{H}_{\mathbf{ur}}$  as in Section 3.5 of [17]; these correspond to the usual Hecke operators  $T_{\ell}$ .

Recall that for each  $p \in \mathbf{ur}$ , we have fixed an isomorphism  $\iota_p : D_p \xrightarrow{\cong} M(2, \mathbb{Q}_p)$  such that  $\iota_p(\mathcal{O}_p) = M(2, \mathbb{Z}_p)$ . Put  $G_{\mathbf{ur}} := \prod_{p \in \mathbf{ur}} G_p$ ,  $K_{\mathbf{ur}} = \prod_{p \in \mathbf{ur}} \mathcal{O}_p^\times$ . Using the local isomorphisms  $\iota_p$ , we identify  $\mathcal{H}_{\mathbf{ur}}$  with the set of compactly supported functions on  $G_{\mathbf{ur}}$  that are bi- $K_{\mathbf{ur}}$  invariant and transform under the action of the centre by  $\omega_\pi^{-1}$ . (This identification does not depend on the choice of the local isomorphisms.) In particular, we can now identify the functions  $\kappa_\ell$  with functions on  $G_{\mathbf{ur}}$ . For each  $\kappa \in \mathcal{H}_{\mathbf{ur}}$ , we obtain in the usual manner an operator  $R(\kappa)$  on the space of all automorphic forms on  $D^\times(\mathbb{A})$  which have central character  $\omega_\pi$ . We have the standard involution  $\kappa \mapsto \kappa^*$  on  $\mathcal{H}_{\mathbf{ur}}$  given by  $\kappa^*(g) = \overline{\kappa(g^{-1})}$  which makes  $R(\kappa^*)$  the adjoint of  $R(\kappa)$ . Given elements  $\kappa_1$  and  $\kappa_2$  in  $\mathcal{H}_{\mathbf{ur}}$ , we define their convolution  $\kappa_1 * \kappa_2 \in \mathcal{H}_{\mathbf{ur}}$  to be the function defined as follows:

$$(52) \quad (\kappa_1 * \kappa_2)(h) = \int_{Z \backslash G} \kappa_1(g^{-1}) \kappa_2(gh) dg.$$

Note that  $R(\kappa_1 * \kappa_2) = R(\kappa_1)R(\kappa_2)$ .

For each positive integer  $m$  such that  $(m, S) = 1$ , we let  $\lambda_\pi(m)$  be the coefficient of  $m^{-s}$  in the Dirichlet series corresponding to  $L(s, \pi)$ , where we normalize the  $L$ -function to have functional equation  $s \mapsto 1 - s$ .

Let  $\Lambda \geq 1$  be a real number. We define

$$\mathcal{P} = \{\ell : \ell \text{ prime}, \ell \in \mathbf{ur}, \Lambda \leq \ell \leq 2\Lambda\}.$$

Define for each integer  $r$  satisfying  $(r, S) = 1$ ,  $c_r = \frac{|\lambda_\pi(r)|}{\lambda_\pi(r)}$ . We put  $\delta_{\mathbf{ur}} = \sum_{r \in \mathcal{P}} c_r \kappa_r$ , and  $\gamma_{\mathbf{ur}} = \sum_{r \in \mathcal{P}} c_{r^2} \kappa_{r^2}$ . Finally, put

$$\kappa_{\mathbf{ur}} = \delta_{\mathbf{ur}} * \delta_{\mathbf{ur}}^* + \gamma_{\mathbf{ur}} * \gamma_{\mathbf{ur}}^*.$$

It is clear that  $R(\kappa_{\mathbf{ur}})$  is a normal, non-negative operator. Moreover, by a standard argument (see (5.6-5.8) of [3] and Section 3.7 of [17]) we get that

$$(53) \quad R(\kappa_{\mathbf{ur}})\phi = \lambda_{\mathbf{ur}}\phi, \quad \lambda_{\mathbf{ur}} \gg_\epsilon \Lambda^{2-\epsilon}.$$

Furthermore, the well-known relation

$$\kappa_m * \kappa_n^* = \sum_{t | \gcd(m, n)} \left( \prod_{p|t} \omega_{\pi_p}(t) \right) \left( \prod_{p|n} \omega_{\pi_p}^{-1}(n) \right) \kappa_{mn/t^2},$$

gives us that

$$(54) \quad \kappa_{\mathbf{ur}} = \sum_{1 \leq l \leq 16\Lambda^4} y_l \kappa_l$$

where the complex numbers  $y_l$  satisfy:

$$(55) \quad |y_l| \ll \begin{cases} \Lambda, & l = 1, \\ 1, & l = \ell_1 \ell_2 \text{ or } l = \ell_1^2 \ell_2^2 \text{ with } \ell_1, \ell_2 \in \mathcal{P}, \\ 0, & \text{otherwise.} \end{cases}$$

Finally, we consider the infinite place. As we are not looking for a bound in the archimedean aspect, the choice of  $\kappa_\infty$  is unimportant. However for definiteness, let us fix the function  $\kappa_\infty$  on  $G_\infty$  as in [17] (see end of Section 3.5). In particular this choice has the property that

$$\kappa_\infty(g) \neq 0 \Rightarrow \det(\iota_\infty(g_\infty)) > 0, \quad u(\iota_\infty(g_\infty)) \leq 1$$

where for  $g \in \mathrm{GL}_2(\mathbb{R})^+$ , we let  $u(g) = \frac{|g(i)-i|^2}{4\mathrm{Im}(g(i))}$  denote the hyperbolic distance from  $g(i)$  to  $i$ . Furthermore, the operator  $R(\kappa_\infty)$  is self-adjoint, non-negative and satisfies

$$(56) \quad R(\kappa_\infty)\phi_\infty = \lambda_\infty\phi_\infty, \quad \lambda_\infty \gg_{\pi_\infty} 1.$$

**4.2. The automorphic kernel and spectral expansion.** Note that the function  $\kappa$  on  $G(\mathbb{A})$  defined in the previous subsection transforms by  $\omega_\pi^{-1}$  under the action of the centre  $\mathbb{A}^\times$ . In particular,  $\kappa$  is  $\mathbb{Q}^\times$ -invariant. Define the automorphic kernel  $K_\kappa(g_1, g_2)$  for  $g_1, g_2 \in G(\mathbb{A})$  via

$$K_\kappa(g_1, g_2) = \sum_{\gamma \in G'(\mathbb{Q})} \kappa(g_1^{-1}\gamma g_2).$$

Using (35), (51), (53), and (56), we see that

$$R(\kappa)\phi = \lambda\phi, \quad \text{where } \lambda \gg_{\pi_\infty, \epsilon} \frac{N^{-1-\epsilon}\Lambda^{2-\epsilon}}{\dim(\rho)}.$$

Now, spectrally expanding  $K_\kappa(g, g)$  and using the non-negativity of the operator  $R(\kappa)$ , we obtain,

$$(57) \quad \dim(\rho)^{-1}N^{-1-\epsilon}\Lambda^{2-\epsilon}|\phi(g)|^2 \ll_{\pi_\infty, \epsilon} K_\kappa(g, g).$$

Write  $K_S = \prod_{p \in S} K_p$ . Now, let  $g \in G(\mathbb{A})$  satisfy (46); so  $g_S \in K_S$ . Put  $z = g_\infty(i) \in \mathcal{J}$ . Using (54), we see that

$$(58) \quad K_\kappa(g, g) = \sum_{1 \leq \ell \leq 16\Lambda^4} y_\ell \sum_{\gamma \in G'(\mathbb{Q})} \kappa_S(g_S^{-1}\gamma_S g_S)\kappa_\ell(\gamma)\kappa_\infty(g_\infty^{-1}\gamma_\infty g_\infty).$$

**Definition 4.1.** Given a lattice  $\mathcal{L}$ ,  $z \in \mathbb{H}$ , and a positive integer  $\ell$ , let  $S(\ell, \mathcal{L}; z)$  be the set of  $\gamma \in G'(\mathbb{Q})$  satisfying the following properties:

- (1)  $\gamma_p \in \mathbb{Q}_p^\times \mathcal{L}_p(\ell)$  for all  $p \in \mathfrak{f}$ , where  $\mathcal{L}_p(\ell) = \{\alpha \in \mathcal{L}_p : \mathrm{nr}(\alpha) \in \ell\mathbb{Z}_p^\times\}$ .
- (2)  $\det(\iota_\infty(\gamma_\infty)) > 0$ ,  $u(z, \iota_\infty(\gamma_\infty)z) \leq 1$ .

Our definition of  $\kappa$  implies that if  $\kappa_S(g_S^{-1}\gamma_S g_S)\kappa_\ell(\gamma)\kappa_\infty(g_\infty^{-1}\gamma_\infty g_\infty) \neq 0$  then we must have  $\gamma \in S(\ell, g \cdot \mathcal{O}; z)$ . Since  $\kappa_\ell(\gamma) \leq \ell^{-1/2}$ , and  $|\kappa_S| \leq 1$ ,  $|\kappa_\infty| \leq 1$ , the triangle inequality on (58), together with (55) and (57), now gives us

$$(59) \quad \begin{aligned} |\phi(g)|^2 &\ll_{\pi_\infty, \epsilon} \dim(\rho)N^{1+\epsilon}\Lambda^{-2+\epsilon} \sum_{1 \leq \ell \leq 16\Lambda^4} \frac{|y_\ell|}{\sqrt{\ell}} |S(\ell, g \cdot \mathcal{O}; z)| \\ &\ll \dim(\rho)N^{1+\epsilon}\Lambda^\epsilon \left( \Lambda^{-1} |S(1, g \cdot \mathcal{O}; z)| + \Lambda^{-3} \sum_{\ell_1, \ell_2 \in \mathcal{P}} |S(\ell_1 \ell_2, g \cdot \mathcal{O}; z)| \right. \\ &\quad \left. + \Lambda^{-4} \sum_{\ell_1, \ell_2 \in \mathcal{P}} |S(\ell_1^2 \ell_2^2, g \cdot \mathcal{O}; z)| \right). \end{aligned}$$

**4.3. The endgame.** We can now wrap up the proof, beginning with a simple proposition that links it all back to Section 2.

**Proposition 4.2.** For any lattice  $\mathcal{L}$  and non-zero integer  $\ell$ , the natural map  $G(\mathbb{Q}) \rightarrow G'(\mathbb{Q})$  induces a bijection of finite sets  $\pm 1 \setminus \mathcal{L}(\ell; z, 1) \cong S(\ell, \mathcal{L}; z)$ . In particular

$$|S(\ell, \mathcal{L}; z)| = \frac{1}{2} |\mathcal{L}(\ell; z, 1)|.$$

*Proof.* It is clear that any element of  $\mathcal{L}(\ell; z, 1)$  satisfies the two conditions defining  $S(\ell, \mathcal{L}; z)$ . Furthermore, if two elements  $\gamma_1, \gamma_2$  in  $\mathcal{L}(\ell; z, 1)$  represent the same class in  $G'(\mathbb{Q})$ , then putting  $t\gamma_1 = \gamma_2$ , we obtain (taking norms) that  $t^2 = 1$  which means that  $\gamma_1 = \pm\gamma_2$ . Therefore we get an injective map  $\pm 1 \backslash \mathcal{L}(\ell; z, 1) \rightarrow S(\ell, \mathcal{L}; z)$ . To complete the proof, we need to show that this map is surjective. Let  $\gamma \in G'(\mathbb{Q})$  be an element whose image in  $G'(\mathbb{Q})$  lies in  $S(\ell, \mathcal{L}; z)$ . We need to prove that there exists  $t_0 \in \mathbb{Q}^\times$  such that  $t_0\gamma \in \mathcal{L}(\ell; z, 1)$ . By Definition 4.1, we can find for each prime  $p$ , an element  $t_p \in \mathbb{Q}_p^\times$  such that  $\gamma_p \in t_p\mathcal{L}_p(\ell)$  for all primes  $p$ . By strong approximation for  $\mathbb{Q}$ , we can choose  $t_0 \in \mathbb{Q}$  such that  $t_0t_p \in \mathbb{Z}_p^\times$  for all primes  $p$ . Now consider the element  $t_0\gamma$ . For each prime  $p$ , we have that the  $p$ -component of  $t_0\gamma$  lies in  $\mathcal{L}_p$ . So by (34), we have  $t_0\gamma \in \mathcal{L}$ . Furthermore, we have that the  $p$ -component of  $\text{nr}(t_0\gamma)$  lies in  $\ell\mathbb{Z}_p^\times$  for all primes  $p$ , and hence  $\text{nr}(t_0\gamma) = \pm\ell$ . But by assumption  $\text{nr}(\gamma) > 0$ . Hence  $\text{nr}(t_0\gamma) = \ell$ . It follows that  $t_0\gamma \in \mathcal{L}(\ell)$ . Since  $u(z, \iota_\infty(\gamma)z) \leq 1$ , it is now immediate that  $t_0\gamma \in \mathcal{L}(\ell; z, 1)$ .  $\square$

Now let us go back to (59). We will prove two bounds. For the first, we choose  $\Lambda = \frac{1}{2}C^{1/4}N^{1/12}$  and apply Proposition 2.14 to (59). (Note here that  $g \cdot \mathcal{O}$  has the same level  $N$  as  $\mathcal{O}$ ). This gives us

$$(60) \quad |\phi(g)| \ll_{\pi_\infty, \epsilon} \dim(\rho)^{1/2} N^{\frac{11}{24} + \epsilon}.$$

For the second bound, we apply Proposition 2.8 to (59). By the assumption (46), the orders  $\mathcal{O}$  and  $g \cdot \mathcal{O}$  have the same shape, which we denote by  $(M_1, M_2, M_3)$ . Furthermore, by (47),  $M_1M_2 \gg N/N_1$ . Now, applying Proposition 2.8 to (59), we get

$$|\phi(g)|^2 \ll_{\pi_\infty, \epsilon} \dim(\rho) N^{1+\epsilon} \left( \Lambda^{-1} + \frac{\Lambda^2}{N} + \frac{N_1}{N} \right).$$

Choosing  $\Lambda = N^{1/3}$  above gives us

$$(61) \quad |\phi(g)| \ll_{\pi_\infty, \epsilon} \dim(\rho)^{1/2} N^\epsilon \max(N^{1/3}, N_1^{1/2}).$$

Combining (60) and (61), we obtain (50).

## REFERENCES

- [1] Edgar Assing. On sup-norm bounds part I: ramified Maaß newforms over number fields. *Preprint*, 2017.
- [2] U. Betke, M. Henk, and J. M. Wills. Successive-minima-type inequalities. *Discrete Comput. Geom.*, 9(2):165–175, 1993.
- [3] Valentin Blomer, Gergely Harcos, and Djordje Milićević. Bounds for eigenforms on arithmetic hyperbolic 3-manifolds. *Duke Math. J.*, 165(4):625–659, 2016.
- [4] Valentin Blomer, Gergely Harcos, Djordje Milićević, and Peter Maga. The sup-norm problem for  $\text{GL}(2)$  over number fields. *J. Eur. Math. Soc. (JEMS)*. To appear.
- [5] Valentin Blomer and Roman Holowinsky. Bounding sup-norms of cusp forms of large level. *Invent. Math.*, 179(3):645–681, 2010.
- [6] J. Brzeziński. On orders in quaternion algebras. *Comm. Algebra*, 11(5):501–522, 1983.
- [7] Gergely Harcos and Nicolas Templier. On the sup-norm of Maass cusp forms of large level: II. *Int. Math. Res. Not. IMRN*, 2012(20):4764–4774, 2012.
- [8] Gergely Harcos and Nicolas Templier. On the sup-norm of Maass cusp forms of large level. III. *Math. Ann.*, 356(1):209–216, 2013.
- [9] Yueke Hu. Sup norm on  $\text{PGL}_n$  in depth aspect. *ArXiv e-prints*, 2018. arXiv:1809.00617.
- [10] Yueke Hu and Paul D. Nelson. New test vector for Waldspurger’s period integral. *ArXiv e-prints*, 2018. arXiv:1810.11564v1.
- [11] Yueke Hu, Paul D. Nelson, and Abhishek Saha. Some analytic aspects of automorphic forms on  $\text{GL}(2)$  of minimal type. *Comm. Math. Helv.*, To appear, 2017.

- [12] Yueke Hu, Jie Shu, and Hongbo Yin. An explicit Gross-Zagier formula related to the Sylvester Conjecture. *ArXiv e-prints*, 2018.
- [13] Henryk Iwaniec. On the problem of Jacobsthal. *Demonstratio Math.*, 11(1):225–231, 1978.
- [14] Simon Marshall. Local bounds for  $L^p$  norms of Maass forms in the level aspect. *Algebra & Number Theory*, 10(4):803–812, 2016.
- [15] Paul D. Nelson. Microlocal lifts and quantum unique ergodicity on  $GL_2(\mathbb{Q}_p)$ . Preprint, 2016.
- [16] Paul D. Nelson and Akshay Venkatesh. The orbit method and analysis of automorphic forms. *ArXiv e-prints*, 2018.
- [17] Abhishek Saha. Hybrid sup-norm bounds for Maass newforms of powerful level. *Algebra & Number Theory*, 11(5):1009–1045, 2017.
- [18] Abhishek Saha. On sup-norms of cusp forms of powerful level. *J. Eur. Math. Soc. (JEMS)*, 19(11):3549–3573, 2017.
- [19] Allan J. Silberger. *PGL<sub>2</sub> over the p-adics: its representations, spherical functions, and Fourier analysis*. Lecture Notes in Mathematics, Vol. 166. Springer-Verlag, Berlin-New York, 1970.
- [20] Nicolas Templier. <http://pi.math.cornell.edu/~templier/papers/errselecta.pdf>.
- [21] Nicolas Templier. On the sup-norm of Maass cusp forms of large level. *Selecta Math. (N.S.)*, 16(3):501–531, 2010.
- [22] Nicolas Templier. Hybrid sup-norm bounds for Hecke–Maass cusp forms. *J. Eur. Math. Soc. (JEMS)*, 17(8):2069–2082, 2015.
- [23] John Voight. *Quaternion algebras*. <https://math.dartmouth.edu/~jvoight/quat-book.pdf>.

SCHOOL OF MATHEMATICAL SCIENCES, QUEEN MARY UNIVERSITY OF LONDON, LONDON E1 4NS, UK