THE SUP-NORM PROBLEM FOR AUTOMORPHIC FORMS IN HIGHER RANK

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Radu Toma
aus
Bukarest, Rumänien

Bonn 2024

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät
der Rheinischen Friedrich-Wilhelms-Universität Bonn

Gutachter: Prof. Dr. Valentin Blomer
Gutachterin: Prof. Dr. Catharina Stroppel

Tag der Promotion: 16.09.2024
Erscheinungsjahr: 2024

Abstract

We study the sup-norm of Hecke-Maaß cusp forms on certain locally symmetric spaces of $\mathrm{SL}_n(\mathbb{R})$. The latter correspond either to cocompact lattices defined by orders in division algebras, or to the Hecke congruence subgroups of $\mathrm{SL}_n(\mathbb{Z})$, yielding non-compact spaces. The main results are sub-baseline bounds uniform in the volume of the space and, in the compact case, also in the spectral parameter. These bounds are the first of their kind for $n > 2$. The methods involve a thorough study of level structures in higher rank, including a new reduction theory with level in the non-compact case. This is used to solve the core counting problem by soft, generalisable arguments, based on rigidity principles.

Acknowledgements

I would like to thank my supervisor, Prof. Dr. Valentin Blomer, for guiding virtually all of my mathematical studies, from my first lectures in Göttingen to the completion and defence of the present thesis. He has always suggested truly beautiful projects to me and his feedback on my work has always come with great questions, ideas, and with staggering efficiency. I am also fortunate to have received a myriad of his priceless pieces of advice on mathematical and professional matters. For all this, his constant and kind support during the last eight years, I am deeply grateful.

Another mentor I was lucky to have during my doctoral studies is Dr. Edgar Assing, one of the most knowledgeable people around and one of my best friends. All of the many walks, going for coffee and pastries, talking about mathematics, have played a crucial role in broadening my knowledge and managing my motivation. I thank him and his family, my good friends Federica, Livia, and Marta, for their support and friendship.

I have also benefited immensely from my research stay in Lausanne, visiting Prof. Dr. Philippe Michel. It came at a time of difficulty and low spirits and provided the perfect environment for me to get my motivation and productivity back. I learned a lot and made some of my fondest memories of the last years there. I am very grateful to the whole TAN lab and also to the BIGS program that made this possible.

For conversations about mathematics, research, writing and more during my doctoral studies, I also thank Alex from Bucharest; Viktor, Salvador, Xianyu from Göttingen; all the Homies from Cambridge; the Blomer AG; Sil, Sid, and Jan; Petru and Asbjørn; Professors Brumley, Harcos, and Maga.

Most of all, I am grateful to my parents, Ștefan and Ștefania, and my sister, Maria. They gave me so much invaluable freedom, support, and love. This is dedicated to them.

Contents

1 Introduction **1**
- 1.1 The sup-norm problem . . . 1
- 1.2 Automorphic forms . . . 4
- 1.3 Arithmetic quantum chaos . . . 9
- 1.4 Methods . . . 15
- 1.5 Outline . . . 19

2 The sup-norm of newforms **20**
- 2.1 Introduction . . . 20
- 2.2 Preliminaries on lattices . . . 28
- 2.3 The amplified pretrace formula . . . 34
- 2.4 Higher rank Atkin-Lehner operators . . . 39
- 2.5 Reduction of the domain . . . 48
- 2.6 Counting matrices . . . 53
- 2.7 Final steps . . . 69

3 The cocompact case **71**
- 3.1 Introduction . . . 71
- 3.2 Division algebras and arithmetic subgroups . . . 79
- 3.3 The amplified pretrace formula . . . 84
- 3.4 Counting in the discriminant aspect . . . 89
- 3.5 Counting in the spectral aspect . . . 92
- 3.6 Proof of Theorem 3.1 . . . 94
- 3.7 Quaternion algebras over number fields . . . 96
- 3.8 Remarks on the case of composite degree . . . 101

Bibliography **103**

1. Introduction

The theory of automorphic forms describes the spectrum and cohomology of locally symmetric spaces and lies at the confluence of analysis, representation theory, and geometry. It plays a fundamental role in number theory, applying to many classical questions about quadratic forms, sphere packings, elliptic curves, and more.

In this thesis, we study automorphic forms from an analytic point of view. In broad strokes, our goal is to understand the mass distribution of Laplace eigenfunctions on certain arithmetic spaces. The focus here lies on the sup-norm problem, which asks for strong upper bounds on these eigenfunctions in terms of natural parameters, such as the Laplace eigenvalue or the volume of the space.

There are many works dedicated to the sup-norm problem for locally symmetric spaces attached to the group SL(2). The main contribution of this thesis consists in solving many new instances of this problem for automorphic forms in higher rank, i.e. for $\mathrm{SL}(n)$ with $n > 2$. These results and more are contained in Chapter 2 and Chapter 3, each based on a scientific article.

This introduction should provide a guiding overview, a binding element between both works. Its structure is as follows:

1. Section 1.1 presents the basics of the sup-norm problem and the main results of this thesis.
2. Section 1.2 presents the analytic theory of automorphic forms through examples.
3. Section 1.3 discusses the heuristics and motivation of the sup-norm problem in more detail.
4. Section 1.4 gives an overview of the methods.

1.1 THE SUP-NORM PROBLEM

This thesis is concerned with the following families of locally symmetric spaces with number theoretic significance. First, for $n \geq 2$ and $N \geq 1$, let

$$X_n(N) = \Gamma_0^n(N) \backslash \mathrm{SL}_n(\mathbb{R})/\mathrm{SO}(n),$$

where $\Gamma_0^n(N)$ is the Hecke congruence subgroup of $\mathrm{SL}_n(\mathbb{Z})$ defined by congruence conditions modulo N. Second, let

$$X_{\mathcal{O}} = \mathcal{O}^1 \backslash \mathrm{SL}_n(\mathbb{R})/\mathrm{SO}(n),$$

where $\mathcal{O}$ is an order in a central division algebra A of degree n over $\mathbb{Q}$, split over $\mathbb{R}$, and $\mathcal{O}^1$ denotes its norm 1 units embedded in $\mathrm{SL}_n(\mathbb{R})$. We discuss their interpretation as spaces of lattices in Section 1.2.4.

Note that $X_{\mathcal{O}}$ is compact, while $X_n(N)$ is non-compact. We equip them with measures invariant under the action of $\mathrm{SL}_n(\mathbb{R})$, inherited from a fixed Haar measure. It is a fact that they have finite volume, given in terms of arithmetic data, N and the discriminant $\mathrm{disc}(\mathcal{O})$, respectively.

These spaces also inherit an algebra $\mathcal{D}$ of invariant differential operators. Additionally, they posses a commutative algebra $\mathcal{H}$ of normal operators, called Hecke operators, that commute with elements of $\mathcal{D}$.

Our main objects of study are Hecke-Maaß forms on the locally symmetric space X, either $X_n(N)$ or $X_{\mathcal{O}}$. These are joint eigenfunctions of $\mathcal{D}$ and $\mathcal{H}$. In particular, a Hecke-Maaß form $\phi : X \to \mathbb{C}$ satisfies

$$\Delta\phi = \lambda\phi,$$

where $\lambda \geq 0$ and $\Delta \in \mathcal{D}$ is the positive Laplace operator on X.

For non-compact $X = X_n(N)$, we additionally require moderate growth conditions, and we informally define a Hecke-Maaß *cusp form* by further stipulating rapid decay at infinity. Thus, a cusp form can be shown to be L^2-integrable.

The basic goal of this thesis is to compare the L^∞-norm $\|\phi\|_\infty$ and the L^2-norm $\|\phi\|_2$ of Hecke-Maaß (cusp) forms ϕ, as their parameters vary. This is a natural and well-studied problem in harmonic analysis. It is also motivated by ideas in physics, being a central question in the theory of quantum chaos. Moreover, in our setting, it has applications in number theory, for instance in the theory of L-functions. We review some of these aspects in Section 1.3.

More precisely, if X is one of the locally symmetric spaces of $\mathrm{SL}_n(\mathbb{R})$ above, we aim at improving the so-called *baseline bound*

$$\frac{\|\phi\|_\infty}{\|\phi\|_2} \leq c \cdot \lambda^{n(n-1)/8} \operatorname{vol}(X)^0. \tag{1.1.1}$$

The constant $c > 0$ might depend on n, but not on the eigenvalue λ or the volume of X. This bound is expected to hold for more general compact locally symmetric spaces, but certainly not trivial to prove. The source of this baseline bound and the following refinement is discussed starting with Section 1.3.2.

In the non-compact case $X = X_n(N)$, the statement needs to be refined to

$$\frac{\|\phi\,|_{\Omega_N}\|_\infty}{\|\phi\|_2} \leq c \cdot \lambda^{n(n-1)/8} \operatorname{vol}(X)^0,$$

where we restrict ϕ to a compact subset $\Omega_N \subset X$, intuitively called the bulk of the space. The constant c might depend partially on the choice of Ω_N, but again not on λ or $\mathrm{vol}(X)$.

The general conjecture is that the baseline bound can be improved for Hecke-Maaß forms – this is the *sup-norm problem*. A strong bound would have the shape

$$\frac{\|\phi\|_\infty}{\|\phi\|_2} \leq c \cdot \lambda^{n(n-1)/8-\delta_1} \operatorname{vol}(X)^{-\delta_2}, \tag{1.1.2}$$

for positive δ_1, δ_2. We call this a *sub-baseline bound*, and denote its statement by $H(\delta_1, \delta_2)$. When restricting to a compact set $\Omega_N \subset X$ as in (1.1), we denote the analogous statement by $H_{\Omega_N}(\delta_1, \delta_2)$.

The first to give a solution to the sup-norm problem were Iwaniec and Sarnak in their breakthrough [IS95], still the most important and influential paper in this field. They prove the bound $H(\delta_1, *)$ for $X = X_2(1)$ and $X_{\mathcal{O}}$ for certain orders $\mathcal{O}$, for any $\delta_1 < 1/24$. The volume dependence is not made explicit and we call this a result in the spectral aspect.

The next milestone was solving the sup-norm problem in the volume aspect, also called the *level aspect*. This was done by Blomer and Holowinsky in [BH10], where they proved a global sub-baseline bound $H(\delta_1, \delta_2)$ for Hecke-Maaß cusp forms on $X = X_2(N)$ with N square-free, with any $\delta_1 < 1/4600$ and $\delta_2 < 1/2300$. More on the history of this aspect is provided in the introductions of the papers in the next chapters.

The natural generalisation to higher rank, i.e. to $\mathrm{SL}_n(\mathbb{R})$ with $n > 2$, followed in the work of Blomer and Maga [BM16], and Marshall [Mar14]. Their results are again only in the spectral aspect, of the form $H_\Omega(\delta_1, *)$ with some inexplicit $\delta_1 > 0$ and no uniformity in the volume.

This thesis presents the first level aspect results in higher rank for the spaces $X_n(N)$ and $X_{\mathcal{O}}$. We summarise our main theorems as follows. The first one is discussed in Chapter 2.

Theorem 1.1. *If n and N are prime, then $H_{\Omega_N}(0, \delta_2)$ holds for Hecke-Maaß cusp forms on $X_n(N)$ with any $\delta_2 < (2n^2)^{-1}$.*

The second one is discussed in Chapter 3.

Theorem 1.2. *If n is an odd prime, then $H(\delta_1, \delta_2)$ holds for Hecke-Maaß forms on $X_{\mathcal{O}}$, where $\mathcal{O}$ is any corresponding locally norm-maximal order, with any $\delta_1 < (16n^3)^{-1}$ and $\delta_2 < (8n^4)^{-1}$.*

Although both main theorems hold only for prime degree n as stated, we prove additional results that relax this hypothesis in different ways, assuming other analytic or algebraic conditions. Other generalisations include the extension of Theorem 1.2 to the group PGL(2) over number fields, also new in this level of uniformity.

The methods we use are soft and should lend themselves well to generalisation. We perform a thorough study of level structures in higher rank and we gather a number of results along the way that we consider of independent

interest. These include certain volume computations with discriminants in Chapter 3 and a generalisation of classical reduction theory of lattices in Chapter 2. The latter also involves group theoretic results, highlighted below, which amount to a determination of the symmetries of the spaces $X_n(N)$.

Theorem 1.3. *The normaliser of the Hecke congruence subgroup $\Gamma_0^n(N)$ inside $\mathrm{SL}_n(\mathbb{R})$ is trivial for $n > 2$ and any $N \in \mathbb{Z}_{\geq 1}$.*

1.2 AUTOMORPHIC FORMS

In this section we sketch out the definition of Hecke-Maaß forms and occasionally focus in on features that become important in the sup-norm problem.

1.2.1 Fourier analysis

It is useful to organise the spectral theory of automorphic forms around intuition from classical Fourier analysis.

We consider the real line $\mathbb{R}$ with addition as a Lie group and equip it with the standard Lebesgue measure. The latter is a Haar measure, invariant under the group operations. We have a natural algebra of differential operators, consisting of polynomials in d/dx. The subalgebra of operators invariant under the group operations is generated by $\Delta := d^2/dx^2$, called the Laplace operator.

Inside the Lie group $\mathbb{R}$ we find the discrete subgroup $\mathbb{Z}$. The quotient space $\mathbb{Z}\backslash\mathbb{R}$ is a compact smooth manifold, namely the circle, which inherits the invariant measure and differential operators from $\mathbb{R}$.

The theory of Fourier series is now the study of the Laplacian Δ on the space $L^2(\mathbb{Z}\backslash\mathbb{R})$. The upshot is that we can decompose this space into subspaces spanned by eigenfunctions of Δ, which are the harmonics $x \mapsto \exp(2\pi i n x)$ for $n \in \mathbb{Z}$. Notice that this is a discrete direct decomposition.

On the other hand, the Fourier transform is part of the spectral theory of Δ operating on the space $L^2(\mathbb{R})$. We have the same kind of decomposition, given by Fourier inversion, into subspaces generated by eigenfunctions $x \mapsto \exp(2\pi i \xi x)$ for $\xi \in \mathbb{R}$. However, this is now a continuous direct integral decomposition and these eigenfunctions are no longer L^2-integrable.

1.2.2 Locally symmetric spaces

We now replace $\mathbb{R}$ with a semisimple Lie group G, for example $\mathrm{SL}_2(\mathbb{R})$. It comes with a Haar measure and an algebra of invariant differential operators. For $G = \mathrm{SL}_2(\mathbb{R})$, the latter is again generated by a single operator, called the Casimir element Ω_G.

The geometric spaces we consider in this thesis are called locally symmetric spaces. Choosing a maximal compact subgroup K and a discrete subgroup Γ of G, these are double quotients of the form $\Gamma\backslash G/K$. They form an important and well-studied class of Riemannian manifolds.

For example, if $G = \mathrm{SL}_2(\mathbb{R})$ we can take $K = \mathrm{SO}(2)$ and the spaces we obtain in this way, varying the lattice Γ, are finite volume hyperbolic surfaces. Taking other examples, we obtain Euclidean spaces, the spheres, higher dimensional hyperbolic spaces, and more.

The geometry and spectral theory of locally symmetric spaces is very rich and many questions still remain about their spectrum, for instance Selberg's eigenvalue conjecture. However, we can describe the spectral decomposition at a structural level.

In the simplest case of our example $G = \mathrm{SL}_2(\mathbb{R})$, if $\Gamma\backslash G/K$ is compact, then the spectral theory of Ω_C on $L^2(\Gamma\backslash G/K)$ resembles the case of the circle $\mathbb{Z}\backslash\mathbb{R}$. We again have a discrete direct sum decomposition into spaces generated by eigenfunctions of Ω_C. On the other hand, for $\Gamma = \mathrm{SL}_2(\mathbb{Z})$, the quotient is non-compact and the spectral decomposition is now a combination of a discrete and a continuous part, the latter analogous to the direct integral decomposition of $L^2(\mathbb{R})$. A brief overview of this is given in the next section.

1.2.3 *Classical Maaß forms*

We study the case of $G = \mathrm{SL}_2(\mathbb{R})$, $K = \mathrm{SO}(2)$, and $\Gamma = \mathrm{SL}_2(\mathbb{Z})$ in more detail. Letting G act by Möbius transformations on the imaginary unit i, we obtain a bijection between G/K and the upper half plane

$$\mathbb{H} = \{x + iy \in \mathbb{C} \mid y > 0\}.$$

Transporting the geometry of G, i.e. the invariant Riemannian metric, turns $\mathbb{H}$ into the well-known model of the hyperbolic plane.

The Casimir element Ω_C for $\mathrm{SL}_2(\mathbb{R})$ now descends to $\mathbb{H}$ as the Laplace-Beltrami operator

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

The invariant measure is given, up to scaling, by

$$d\mu = \frac{dx\,dy}{y^2}.$$

The locally symmetric space $X = \Gamma\backslash G/K$ we obtain has finite volume in the inherited measure. It is often called the modular curve, since it is the moduli space of rank 2 lattices up to some notion of isomorphism. We discuss this interpretation more in the next section.

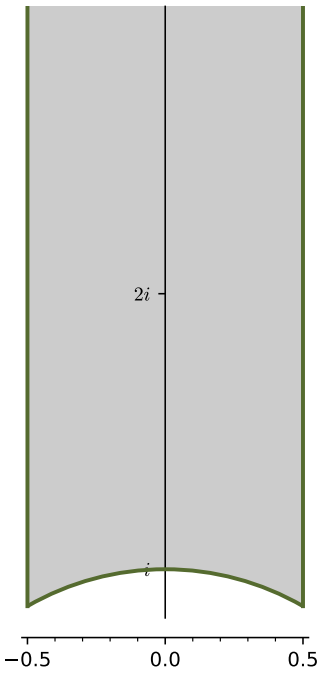


Figure 1.1: A fundamental domain for $\mathrm{SL}_2(\mathbb{Z})$ acting on $\mathbb{H}$

The space X is non-compact, exhibiting a topological end called a *cusp*. Observing Figure 1.1, the strip defines a fundamental domain for X in the upper half plane and we exit to infinity as the imaginary part grows. We note that, in the hyperbolic geometry of $\mathbb{H}$, the strip becomes thinner and thinner, having width proportional to the inverse imaginary part. This explains the term "cusp".

A Maaß form ϕ is now a function $X \to \mathbb{C}$ which is an eigenfunction of Δ, that is,

$$\Delta\phi = \lambda\phi,$$

and satisfies certain moderate growth conditions. By positivity of the Laplace-Beltrami operator, we have $\lambda \geq 0$.

We are particularly interested in cuspidal Maaß forms, or simply *cusp forms*. They are informally defined by additionally asking for vanishing at the cusp, that is, tending to zero as we escape to infinity. As such, one can show that they are L^2-integrable and the space of cusp forms provides almost all of the discrete spectrum of X. The only other component in the discrete decomposition is the space of constant functions, and we write

$$L^2_{\text{disc}} = L^2_{\text{cusp}} \oplus \mathbb{C} \cdot 1.$$

We note that the set of eigenvalues of cusp forms is unbounded, which is a non-trivial fact.

The rest of the spectrum is made up of non-cuspidal Maaß forms called Eisenstein series. These are not L^2-integrable, but they describe the continuous spectrum of X, analogous to the harmonics $\exp(2\pi i\xi x)$ for $L^2(\mathbb{R})$. They do not make a big appearance in this thesis, so we complete their discussion simply by noting the spectral decomposition

$$L^2(X) = L^2_{\text{cusp}} \oplus \mathbb{C} \cdot 1 \oplus L^2_{\text{cont}} = L^2_{\text{cusp}} \oplus L^2_{\text{Eis}}.$$

Here we put the constant function into the so-called Eisenstein spectrum, since it really arises as a residue of the meromorphic family of Eisenstein series.

1.2.4 Spaces of lattices

In this thesis we focus on the families $X_n(N)$ and X_O of locally symmetric spaces, some of the most prominent in the analytic theory of automorphic forms. The motivation to study them is partly given by their interpretation as parametrising spaces for basic geometric objects, namely lattices with various structures.

Recall that a unimodular lattice of rank n is $\mathbb{Z}$-module isomorphic to $\mathbb{Z}^n$ with determinant 1. It can be realised as a set

$$L = \mathbb{Z}^n \cdot g \subset \mathbb{R}^n,$$

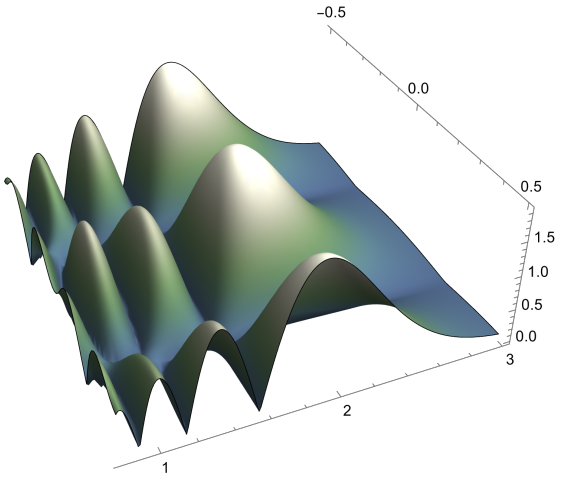


Figure 1.2: The absolute value of a Maaß cusp form on X, eigenvalue $\lambda \approx 1/4 + 13^2$. The fundamental domain in Figure 1.1 is tilted here, showing $\Re z \in [-0.5, 0.5]$ and $\Im z \in [0.8, 3]$. Notice the decay going into the cusp.

where $\mathbb{Z}^n$ is the set of integral row vectors and g is a matrix in $\mathrm{SL}_n(\mathbb{R})$ acting on these vectors.

Generalising the modular curve, the locally symmetric space

$$X_n(1) = \mathrm{SL}_n(\mathbb{Z}) \backslash \mathrm{SL}_n(\mathbb{R}) / \mathrm{SO}(n)$$

parametrises all rank n unimodular lattices, up to isometry. It plays a central role in the theory of automorphic forms.

Often in number theory, problems come with important additional information in the form of a *level structure*. For example, questions about integers might involve some congruence conditions modulo N. More generally, we consider lattices together with a distinguished sublattice, such as $N\mathbb{Z} \subset \mathbb{Z}$, or $N\mathbb{Z} \times \mathbb{Z} \subset \mathbb{Z}^2$. To construct moduli spaces as in the latter example with level structures, we proceed as follows.

For now, let $N = p$ be a prime, and define

$$\Gamma_0(p) := \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in \mathrm{SL}_2(\mathbb{Z}) : p \mid c \right\}.$$

By the theory of elementary divisors, we can check that the space

$$\left\{ (L, L') \mid [L : L'] = p \right\} = \left\{ (\mathbb{Z}^2 g, \mathbb{Z}^2 \operatorname{diag}(p, 1) g) \mid g \in \mathrm{SL}_2(\mathbb{R}) \right\}$$

parametrising pairs of a unimodular lattice and a sublattice of index p can be interpreted as the quotient $\Gamma_0(p) \backslash \mathrm{SL}_2(\mathbb{R})$ and we let

$$X_2(p) = \Gamma_0(p) \backslash \mathrm{SL}_2(\mathbb{R}) / \mathrm{SO}(2).$$

This is a locally symmetric space of finite volume.

The group $\Gamma_0(N)$ can be defined for any $N \in \mathbb{Z}$ analogously and is called the *Hecke congruence subgroup*. Automorphic forms on $X_2(N)$ are said to have level N.

The same philosophy applies for higher rank lattices, where we define

$$\Gamma_0^n(N) = \{\gamma \in \mathrm{SL}_n(\mathbb{Z}) \mid \text{last row of } \gamma \text{ is} \equiv (0, \dots, 0, *) \mod N\}$$

to be the analogue of the Hecke congruence subgroup[1] and the spaces $X_n(N)$ analogously. Automorphic forms on $X_n(N)$ are the main focus of Chapter 2. We remark that the groups $\Gamma_0^n(N)$ are the basis for the theory of newforms, which is useful when discussing automorphic L-functions, generalisations of the Riemann zeta function.

While the spaces $X_n(N)$ are non-compact, we also study compact locally symmetric spaces in this thesis. These can also be interpreted as spaces of lattices with additional algebraic structure, having symmetries with respect to some number field or, more generally, a division algebra. These are the spaces considered in Chapter 3.

Let A be a central division algebra over $\mathbb{Q}$ of degree n and assume that $A \otimes \mathbb{R} = \mathcal{M}_n(\mathbb{R})$. Though these can be constructed concretely, we encourage the reader to think of the matrix algebra $\mathcal{M}_n(\mathbb{Q})$, except that all non-zero elements are invertible.

An order $\mathcal{O} \subset A$ is a $\mathbb{Z}$-module of full rank that is also a ring with unity and it behaves very much like $\mathcal{M}_n(\mathbb{Z})$ inside $\mathcal{M}_n(\mathbb{Q})$. Another illustrative example is that of the ring of integers inside a number field, also a division algebra, though commutative.

These algebras come equipped with a generalisation of the norm and trace. This allows us to define the multiplicative group $\mathcal{O}^1$ of norm 1 units of $\mathcal{O}$. This group can be embedded into $\mathrm{SL}_n(\mathbb{R})$ by the splitting condition on A and it then gives a discrete subgroup such that

$$X_{\mathcal{O}} = \mathcal{O}^1 \backslash \mathrm{SL}_n(\mathbb{R})/\mathrm{SO}(n)$$

is compact – the division algebra condition is crucial here. We can also generalise level structures by taking suborders $\mathcal{O}$ inside some maximal order. In the non-compact case of $X_n(N)$, the maximal order is $\mathcal{M}_n(\mathbb{Z})$ and the suborders are those defined by the congruence condition on the last row.

1.2.5 Hecke operators

The spaces considered are part of the special class of arithmetic locally symmetric spaces. These posses additional symmetries called Hecke operators.

[1] We often drop the superscript n if it is understood from context.

We first give a simple interpretation as averaging operators on the space of lattices. If f is defined on $X_n(1)$, then

$$T_p f(L) = \sum_{[L:L']=p} f(L'),$$

defines a new function on lattices by averaging the value of f over sublattices of index p. This is a special version of a Hecke operator.

From a different, more general perspective, the idea is to average a left Γ-invariant function f over certain shifts αx of a given element $x \in G$. To restore left invariance under Γ, it becomes natural to consider double cosets decomposing into a *finite* union of left cosets, such as

$$\Gamma g \Gamma = \bigcup_i \Gamma \alpha_i.$$

We then have

$$T_g f(x) = \sum_i f(\alpha_i \cdot x),$$

a well-defined averaging operator on the functions on $\Gamma\backslash G/K$, thanks to our finiteness condition.

We obtain Hecke operators T_g for any g in the commensurator of Γ inside G. An essential feature, which for our purposes could be the definition, of an *arithmetic* subgroup Γ is that its commensurator is dense in G. The model for this is $\Gamma = \mathbb{G}(\mathbb{Z})$, $G = \mathbb{G}(\mathbb{R})$, and $\mathrm{Comm}_G(\Gamma) = \mathbb{G}(\mathbb{Q})$, for an algebraic group $\mathbb{G}$ like $\mathrm{PGL}(n)$.

The Hecke operators are normal, commute with the invariant differential operators and the algebra they generate is commutative. Thus, up to some technicalities, we can now define *Hecke-Maaß forms* on an arithmetic locally symmetric space X. These are joint eigenfunctions of the invariant differential operators, with eigenvalues described by a tuple of numbers μ called the spectral parameter, and the Hecke operators, satisfying moderate growth conditions. If X is non-compact, we can again informally ask for vanishing at infinity to define Hecke-Maaß cusp forms.

1.3 ARITHMETIC QUANTUM CHAOS

Not only is understanding the spectrum of a space an essential question in geometry and analysis, but it is also an integral part of physics. Indeed, particles can be described in quantum mechanics through their wave function, which is governed by Schrödinger's equation. In its simplest form, the latter reduces to the Laplace eigenfunction equation.

In the so-called classical limit, one studies high energy states that correspond to eigenfunctions with very large Laplace eigenvalue, expecting them to approximate classical physics. For instance, when the underlying geometry of the space is "chaotic", then one expects high energy waves to also behave chaotically. This is the idea behind Berry's random wave conjecture from the 1970's. The hyperbolic surfaces commonly occurring in number theory, like $X_2(1)$, satisfy such a condition. These considerations then gave rise in the 90's to the theory of *arithmetic quantum chaos*.

Promoted by Peter Sarnak and his collaborators, new conjectures about the asymptotic behaviour of automorphic forms appeared. In essence, we expect Hecke-Maaß forms on groups such as $\mathrm{SL}_2(\mathbb{R})$ to distribute randomly, uniformly on the space, as their eigenvalues grow. For more intuition, we note that an opposite effect would be scarring, where the essential support of an eigenfunction asymptotically describes some proper submanifold, such as a geodesic.

The sup-norm problem is the following "point-wise" approach to quantum chaos.

Problem. If Hecke-Maaß cusp forms become uniformly distributed as their eigenvalues grow, then we expect them not to have any large peaks. Therefore, prove that their sup-norm is small when compared to their L^2-norm.

In the following sections we consider some examples and make the statement more precise, as in Section 1.1.

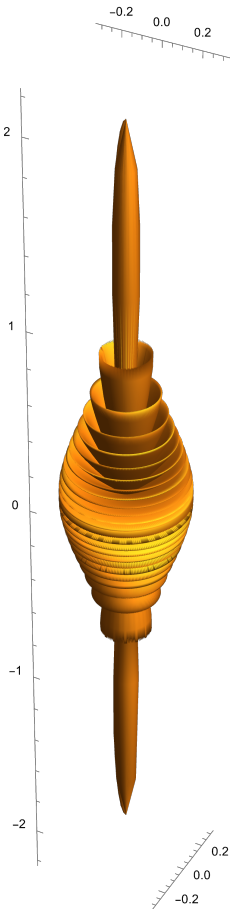


Figure 1.3: Large sup-norms: the spherical harmonic Y_0^{25}

1.3.1 Harmonics of spheres

Take the example of the sphere $S^2 \subset \mathbb{R}^3$, a locally symmetric space stemming from the Lie group SO(3). The eigenvalues of the spherical Laplacian are of the form $l(l+1)$ for $l \in \mathbb{Z}_{\geq 0}$ and, crucially, they have a very high multiplicity. The eigenspace for $l(l+1)$ has dimension $2l+1$ and is spanned by the classical spherical harmonics Y_l^m, in standard notation, of degree l and order m, where $|m| \leq l$.

As we prove below, the sup-norm of an eigenfunction on S^2 with eigenvalue $l(l+1)$ can be as large as $\sqrt{l}$. This is the case for the harmonic Y_0^l depicted in Figure 1.3[2]. This is the most extreme behaviour on two-dimensional spaces.

[2]The plot shows $|Y_0^{25}(x)| \cdot x$ for all $x \in S^2$. The large values occur thus at the north and south pole.

In fact, there can also be strong concentration of these eigenfunctions on the geodesics of the sphere, the great circles. For instance, the harmonic Y_l^l is concentrated around the equator. This is in sharp contrast with the heuristic of quantum chaos, and the main reason is the stability of such geodesics: the geodesic flow is very far from being chaotic on the sphere.

1.3.2 Heuristic upper and lower bounds

We now study in a little more detail the case of locally symmetric spaces for compact groups, generalising the sphere. This is meant to give some intuition and informed guesses for the spaces of non-compact type considered in this thesis.

Let H be a compact Lie group and X a locally symmetric space of H. Assume that X has finite volume in the invariant measure.

Let E_λ be the eigenspace for the eigenvalue λ of the Laplacian of X. Thanks to compactness, this space is a finite-dimensional representation of H, where H acts as in the right-regular representation.

Since evaluation at a point $x \in X$ is a continuous linear functional, by the Riesz representation theorem there is $F_x \in E_\lambda$ such that

$$f(x) = \langle f, F_x \rangle,$$

for any $f \in E_\lambda$. The action of H, denoted by ρ, is unitary and we have

$$f(xh) = \langle \rho_h f, F_x \rangle = \langle f, \rho_{h^{-1}} F_x \rangle,$$

so that $F_{xh} = \rho_{h^{-1}} F_x$. Again by unitarity and also by the transitivity of the action of H on X, we find that $\|F_x\|$ is independent of x.

Note now that $F_x(x) = \|F_x\|^2$, by definition. Additionally, Cauchy-Schwarz implies that

$$F_x(y) \leq \|F_x\| \cdot \left\|F_y\right\| = \|F_x\|^2,$$

so that $F_x(x)$ is also the sup-norm of F_x.

Finally, we decompose F_x into an orthonormal basis (f_i), and by definition we have

$$F_x(x) = \sum_i \langle F_x, f_i \rangle f_i(x) = \sum_i |f_i(x)|^2.$$

We can now integrate this identity over X. Since $\left\|f_i\right\| = 1$ and $F_x(x) = \|F_x\|^2$ is independent of x, we obtain

$$\|F_x\|^2 \cdot \mathrm{vol}(X) = \dim E_\lambda.$$

Since $\|F_x\|^2 = \|F_x\|_\infty$, it follows that

$$\frac{\|F_x\|_\infty}{\|F_x\|} = \sqrt{\frac{\dim E_\lambda}{\mathrm{vol}\, X}},$$

which gives a lower bound for the sup-norm problem in this setting. Conversely, if $f \in E_\lambda$, we use Cauchy-Schwarz to find that

$$\frac{f(x)}{\|f\|} = \frac{\langle f, F_x \rangle}{\|f\|} \leq \|F_x\| = \sqrt{\frac{\dim E_\lambda}{\operatorname{vol} X}}$$

for all x. This is thus an upper bound for the sup-norms of all eigenfunctions.

For the sphere, where $H = \mathrm{SO}(3)$, we observed that $\dim E_\lambda$ is roughly $\sqrt{\lambda}$. Thus the sup-norm of eigenfunctions is, up to the volume, at most $\lambda^{1/4}$ and this is also achieved. Unfortunately, these arguments do not directly apply to locally symmetric spaces of $\mathrm{SL}_n(\mathbb{R})$. However, they give some rough indication of what one could expect.

1.3.3 *Baseline bounds for Hecke-Maaß forms*

We now consider non-compact Lie groups. For the following discussion it is useful to introduce some common notation in analytic number theory.

Notation. We write $f(x) \ll g(x)$ when $f(x) \leq c \cdot g(x)$ for some constant $c > 0$, at least for x large. If $f(x) \ll g(x) \ll f(x)$, then we use the notation $f(x) \asymp g(x)$. This implicit constant c can change from one sign $\ll$ to the next, and this is what makes it practical. Moreover, when we wish to point out that c depends on some parameter P, we write $f(x) \ll_P g(x)$.

First of all, the fact that we can find an eigenfunction $f \in E_\lambda$ such that

$$\frac{\|f\|_\infty}{\|f\|} \geq \sqrt{\frac{\dim E_\lambda}{\operatorname{vol} X}}$$

was already pointed out in [Sar04] for any compact Riemmanian manifold. However, we could further restrict to joint eigenspaces of not just the Laplacian, but also of the whole algebra of invariant differential operators and, for arithmetic spaces, the algebra of Hecke operators. In that case, we can often assume that $\dim E_\lambda = 1$.

The upper bound in the previous section then suggests that the most optimistic bound we could hope for and, perhaps, should aim towards is

$$\|f\|_\infty \ll_\varepsilon \lambda^\varepsilon \cdot \operatorname{vol}(X)^{-1/2},$$

for joint eigenfunctions f, for small $\varepsilon > 0$. The factor λ^ε replacing the naïve 1 gives a more realistic guess. While pointing in the right direction, this overly optimistic conjecture needs many refinements (see [Sar04]).

On the other hand, using relatively low-resolution information about the spectrum generally suffices to obtain a weaker bound, which serves as a baseline for our problem. This follows namely from a local Weyl law. The

latter is an analytic way of estimating the dimension of E_λ by averaging over a neighbourhood of λ or, more generally, of the spectral parameter. It has the imprecise shape

$$\sum_{\lambda_i=\lambda+O(1)} \dim E_{\lambda_i} \asymp \lambda^{n(n-1)/4} \operatorname{vol}(X),$$

for locally symmetric spaces of $G = \mathrm{SL}_n(\mathbb{R})$.

Thus, at least heuristically, we get the bound

$$\frac{\|f\|_\infty}{\|f\|} \ll \sqrt{\frac{\dim E_\lambda}{\operatorname{vol} X}} \ll_n \lambda^{n(n-1)/8} \operatorname{vol}(X)^0. \tag{1.3.1}$$

This is called the *baseline*, or convexity, bound throughout the sup-norm problem literature and this thesis. It provides the benchmark for results: any significant improvements of it towards the optimistic bound generally require new insights and have, until now, come as an application of not just analysis, but also number theory. The shape of such an improvement, which we call a *sub-baseline bound*, is

$$\frac{\|f\|_\infty}{\|f\|} \ll_n \operatorname{vol}(X)^{-\delta_1} \cdot \lambda^{n(n-1)/8-\delta_2}, \tag{1.3.2}$$

for positive δ_1 and δ_2. As in Section 1.1, denote the statement of (1.3.2) by $H(\delta_1, \delta_2)$, for short.

1.3.4 Cutting off the cusps

To discuss results in higher rank, we introduce one last technical refinement of the sup-norm problem. This is necessary since, for non-compact spaces, even the baseline bound $H(0, 0)$ does not hold as stated.

The issue is the following. As the argument goes to infinity through some cusp, there is an analytic phenomenon that creates very large bumps before a cusp form finally decays to zero. This is explained in more detail in [BT20]. Intuitively, the cusps are geometrically so small that they create a bottleneck effect. The phenomenon is already noticeable, though weaker, for $n = 2$, as we depict in Figure 1.4.

More precisely, most Hecke-Maaß cusp forms satisfy

$$\frac{\|f\|_\infty}{\|f\|} \gg_{\varepsilon,n} \lambda^{n(n-1)(n-2)/24-\varepsilon}.$$

The exponent here is cubic in n, compared to the quadratic polynomial in the expected bound (1.3.1). A global sub-baseline bound is thus apparently hopeless.

It is however still expected that Maaß forms are small on most of the space. Since the cusp obscures this expectation in the global sup-norm, the

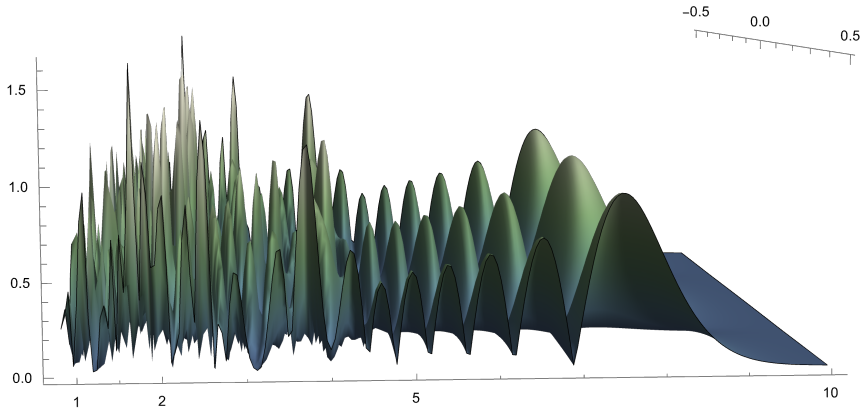


Figure 1.4: A Maaß cusp form on $X_2(1)$, eigenvalue $\lambda \approx 1/4 + 50^2$. The lower axis is the imaginary axis. Notice the bumps on the right-hand side before the decay in the cusp.

conjectures were refined by studying the restriction of forms to compact pieces of the space. For a compact $\Omega \subset X$, the bound

$$\frac{\left\|f\mid_{\Omega}\right\|_{\infty}}{\|f\|} \ll_{\Omega} \mathrm{vol}(X)^{0} \cdot \lambda^{n(n-1)/8}$$

now holds (see [Sar04]), noting that the implied constant depends on Ω.

In this thesis, we think of Ω covering most of the space X. However, for a sup-norm bound to be meaningful in the volume parameter, the implied constant should not depend on $\mathrm{vol}(X)$. Therefore, for the spaces $X_n(N)$ we more carefully define a compact piece, the bulk of the space, as explained in Chapter 2.

In short, consider the projection map $X_n(N) \to X_n(1)$ and let $\Omega \subset X_n(1)$, a fixed subset that does not depend on N. We then define Ω_N to be the preimage of Ω.

A sensible conjecture, uniformly treating both the spectral and the level aspect, is that

$$\frac{\left\|f\mid_{\Omega_N}\right\|_{\infty}}{\|f\|} \ll_{\Omega} \mathrm{vol}(X)^{-\delta_1} \cdot \lambda^{n(n-1)/8-\delta_2}, \tag{1.3.3}$$

which we denote by $H_{\Omega_N}(\delta_1, \delta_2)$, for some $\delta_1, \delta_2 > 0$. This is now the precise formulation of the sup-norm problem considered in this thesis. We remark that, if X is compact, the corresponding conjecture applies to the global sup-norm, as in $H(\delta_1, \delta_2)$.

1.4 METHODS

In the remainder of this introduction, we discuss the general strategy for proving the main results. The following structure of the proof is shared by both papers.

1.4.1 A general approach

One of the most important ways to tackle the sup-norm problem is the application of the amplified pretrace formula. This was first used by Iwaniec and Sarnak [IS95] and it allows us to transform the analytic question of bounding eigenfunctions into a counting problem that we can solve with number theory and geometry. More generally, such pretrace formulae and their refinements (traces, relative traces, amplified traces) are a cornerstone of analytic number theory.

The first idea is to simply write down the spectral decomposition of a well-chosen test function. We can examine this in the toy case of $L^2(\mathbb{Z}\backslash\mathbb{R})$. Take a smooth and compactly supported function f on $\mathbb{R}$. The classical Poisson summation formula now states that

$$\sum_{n\in\mathbb{Z}} f(n) = \sum_{m\in\mathbb{Z}} \hat{f}(m),$$

a beautiful relation between a function and its Fourier transform. As we explain below, the left-hand side contains geometric information, while the right-hand side is of spectral nature.

First, from the aperiod function f we define a new function $K(x, y)$, called an automorphic kernel, by averaging over the lattice $\mathbb{Z}$. More precisely, we write

$$K(x, y) = \sum_{n\in\mathbb{Z}} f(-x + n + y),$$

for $x, y \in \mathbb{R}$. In both these variables, this is a compactly supported, smooth function on $\mathbb{Z}\backslash\mathbb{R}$. As such, we can spectrally decompose it with respect to the Laplace eigenfunctions $e_n(t) = \exp(2\pi i n t)$. Doing so in the variable y, we obtain

$$K(x, y) = \sum_{m\in\mathbb{Z}} \langle K(x, \cdot), e_m\rangle\, e_m(y).$$

It is easy to see that, by a calculation of the shape $\sum_{\mathbb{Z}} \int_{\mathbb{Z}\backslash\mathbb{R}} = \int_{\mathbb{R}}$, we have

$$K(x, y) = \sum_{m\in\mathbb{Z}} \hat{f}(m)\overline{e_m(x)}e_m(y).$$

The Poisson summation formula is obtained by setting $x = y$. Observe that the n-sum is a so-called geometric average over the lattice $\mathbb{Z} \subset \mathbb{R}$, while the m-sum is an average over the spectrum of $\mathbb{Z}\backslash\mathbb{R}$.

The Poisson summation formula has the shape of the *pretrace formula* for more general locally symmetric spaces $X = \Gamma\backslash G/K$. For a test function k on $K\backslash G/K$, we have

$$\sum_{\gamma\in\Gamma} k(x^{-1}\gamma x) = \sum_i \hat{k}(\lambda_i)|\phi_i(x)|^2 + \dots,$$

where $\hat{k}$ is the spherical transform of k. The spectral side is meant formally as the sum over the discrete spectrum plus an integral over the continuous spectrum. The latter is denoted by an ellipsis to avoid technicalities. By non-negativity as described below, we do not need to treat it explicitly.

The point of the formula above is that we get a handle on $|\phi_i(x)|^2$ for any x. Note also that any cuspidal joint eigenfunction ϕ can be embedded in the spectral average as one of the basis elements, say, $\phi = \phi_0$. The choice of test function is then made so that $\hat{k}$ is 1 at the spectral parameter of the eigenfunction ϕ_0 that we are studying. On the other parameters it should be non-negative. This allows us to drop all terms but one and get

$$\sum_{\gamma\in\Gamma} k(x^{-1}\gamma x) \geq |\phi_i(x)|^2.$$

However, it remains to control the function k on the geometric side. As in the Poisson summation formula, this involves a balancing act between k and its spherical transform. Ideally, both should be perfectly localised, and yet the Heisenberg uncertainty principle prevents that.

The necessary analysis was worked out by Blomer and Maga [BM15]. The upshot is that we have a bound for k in terms of the spectral parameter of ϕ_0, which actually gives the baseline bound in the eigenvalue. We can also arrange that k has compact support. Applying the triangle inequality now gives us a counting problem

$$|\phi_i(x)|^2 \leq \|k\|_\infty \cdot |\{\gamma \in \Gamma \mid x^{-1}\gamma x \in \mathrm{supp}(k)\}|.$$

The latter is a finite set because Γ is a discrete subgroup.

Solving this counting problem is at the core of the method. The cardinality in question depends heavily on the point x and, of course, on Γ itself and its covolume.

However, at this stage it turns out that even the best bounds for the counting problem are not sufficient to give a sub-baseline bound. It is here that number theory plays a crucial role by getting us out of this deadlock. Namely, we can use Hecke operators to further amplify the contribution of ϕ_0 in the spectral average. The technique is, appropriately, called *amplification*, and it was invented by Iwaniec.

Being quite involved, we only sketch out the intuition behind amplification. For this, let $g \in G$ be in the commensurator of Γ, so that we can define the

Hecke operator T_g. All the eigenfunctions in the spectral decomposition can be taken to be Hecke eigenfunctions as well, and we write $\lambda_i(g)$ for the eigenvalue of ϕ_i.

These eigenvalues are of utmost importance in number theory and many conjectures have been made about them. For instance, normalising the operator T_g correspondingly, the Generalised Ramanujan Conjecture predicts that $\lambda_i(g) \ll 1$ in the cuspidal spectrum. For most forms ϕ_i, we should also have the lower bound $\lambda_i(g) \gg 1$, at least on average.

Furthermore, for different forms, the Hecke eigenvalues should be uncorrelated. For instance, in the case of $X_n(1)$, we can rewrite T_p as the operator attached to $g = \mathrm{diag}(p, 1, \dots, 1)$, where p is a prime. The corresponding eigenvalues are denoted by $\lambda_i(p)$. We then expect that, for some large parameter L,

$$\sum_{1 \le p \le L} \lambda_i(p)\overline{\lambda_j(p)}$$

is very small in terms of L for $i \neq j$.

The strategy is now roughly the following. We apply the average of Hecke operators

$$\sum_{p \le L} \overline{\lambda_0(p)} \cdot T_p$$

to the pretrace formula. Since all components are joint eigenfunctions, we obtain a spectral average where the contribution of ϕ_0 is now

$$(\sum_{p \le L} |\lambda_0(p)|^2) \cdot \hat{k}(\lambda_0) \cdot |\phi_0(x)|^2 \asymp L \cdot \hat{k}(\lambda_0) \cdot |\phi_0(x)|^2,$$

and that of the other forms much smaller. This is the leveraging mechanism and it is obviously stronger for a larger parameter L.[3] However, it comes at a cost on the geometric side.

We compute by definition that the Hecke operator applied to the variable y gives

$$T_g K(x, y) = \sum_{\alpha \in \Gamma \backslash \Gamma g \Gamma} \sum_{\gamma \in \Gamma} k(x^{-1}\gamma\alpha y) = \sum_{\gamma \in \Gamma g \Gamma} k(x^{-1}\gamma y).$$

This is now of the same shape as $K(x, y)$, yet we have enlarged the set over which we average. The counting problem, as described above, therefore becomes more difficult.

In essence, amplification is a tactical sacrifice. If the number theoretical or geometric techniques are strong enough, it can eventually lead to success.

The naïve technique above is, however, not adequate for many reasons. There are many technicalities involved in constructing a useful amplifier and

[3] To obtain an inequality for $\phi_0(x)$, as we did before from the non-negativity of $\hat{k}$, we must tweak the definition of the amplifier. There are also issues of normalisation. We return to them in the sections on amplification in the next chapters.

they are in essence a combinatorial or p-adic analytic problem. For $\mathrm{SL}_n(\mathbb{R})$, these were also dealt with in [BM15] and are partly explained, as needed, in the corresponding sections of this thesis.

1.4.2 Rigidity principles

The crux of the matter now becomes the counting problem, which one needs to solve uniformly in all parameters. Depending on the situation, we need to count certain integral elements $\gamma \in \mathcal{M}_n(\mathbb{R})$ with determinant m (applying the amplifier means we have to consider $1 \ll m \ll L$, not just $m = 1$), such that

$$x^{-1}\gamma x = O(m^{1/n}).$$

It essentially means that γ lies in some bounded set, a ball skewed by conjugation by x.

The principle behind our approach in this thesis is one of rigidity. Though it does not allow extremely precise counting, this principle is soft enough to be useful in great generality. While it applies in some sense to the proofs of both main theorems, the relevant structures are very different. We thus give here the contours of our strategy and leave the more specific details to the corresponding chapters.

The most basic and surprisingly useful rigidity principle in analytic number theory could be stated as follows: an integer of absolute value strictly less than one must be zero. It is ubiquitous because, in this field, we often have closeness conditions of integral structures and the flexibility of tweaking parameters to have tighter conditions. In the case of the sup-norm problem in the level aspect, these integral structures are the orders $\mathcal{M}_n(\mathbb{Z})$ or $\mathcal{O}$, the closeness condition is given above, and the parameter we can tweak is the length L of the amplifier.

A more sophisticated example of a rigidity principle deals with collinearity. Suppose we have a line of length L in the plane $\mathbb{R}^2$, and three integral points that are ε-close to the line. Suppose further that the area of the triangle defined by these three points is bounded from below by some integer $D \geq 1$ (note that $D = 1$ is always valid). By the closeness condition, this area is also bounded from above roughly by $\varepsilon \cdot L$. Therefore, if εL is much smaller than D, then the points must actually be collinear.

In our work, the plane is replaced by the high-dimensional space of the algebra $\mathcal{M}_n(\mathbb{R})$. The line is replaced by some smaller subvariety, for example a vector space of dimension n. Finally, the bound D is given by the discriminant of the order we are considering, which is directly related to the level or the volume of the corresponding space. The parameters L and ε are then given by the choice of test function and the amplifier, where we have some degrees of freedom.

For instance, we show in Chapter 2, Proposition 2.6.2, that a matrix of $\Gamma_0(N)$-shape satisfying the counting conditions is determined by its last row, at least under certain assumptions. Similarly, in Chapter 3, Lemma 3.4.2, we prove

that the $\mathbb{Q}$-algebra generated by the elements $\gamma \in \mathcal{O}$ satisfying the counting conditions is actually a proper subalgebra of the division algebra A. Assuming the degree of A is prime, this implies that the subalgebra must be a field. This now introduces a great amount of structure that allows for stronger counting techniques.

There is one crucial issue that can introduce difficulties in our counting problem. This is the point x, which particularly in the non-compact case of Chapter 2, has an enormous effect on the success prospects of the strategy outlined above. In the volume aspect, where x varies on larger and larger spaces, this understanding this dependence is essential.

Solving such issues in higher rank is a new feature of our work. Much of the originality of this thesis lies in the ideas required in this problem. They are discussed in detail in the corresponding chapters. These include the aforementioned generalisation of reduction theory and the study of Atkin-Lehner operators in higher rank, which are personal favourites of the author.

1.5 OUTLINE

Chapter 2 deals with the sup-norm problem for $X_n(N)$ and Chapter 3 with the sup-norm problem for $X_{\mathcal{O}}$. These are, respectively, the papers [Tom24] and [Tom23].

The two chapters follow a similar structure of setting up the problem, discussing necessary preliminaries and structures, applying the amplified pretrace formula, and solving the counting problem. The papers are kept in the same form as they were submitted for publication or already published. As such, they can be read independently.

2. The sup-norm of newforms

This chapter reproduces the scientific article [Tom24]:
R. Toma. *The sup-norm problem for newforms of large level on* PGL(n). 2024. arXiv: `2401.02741 [math.NT]`.

Abstract

Let N be a prime and ϕ be a Hecke-Maaß cuspidal newform for the Hecke congruence subgroup $\Gamma_0(N)$ in $\mathrm{SL}_n(\mathbb{R})$. Let Ω be an adelic compactum and let Ω_N be its projection to $\Gamma_0(N)\backslash \mathrm{SL}_n(\mathbb{R})/\mathrm{SO}(n)$. For any prime n, we prove sub-baseline bounds for the sup-norm of ϕ restricted to Ω_N. Conditionally on GRH, we generalise this result to all $n \geq 2$. The methods involve a new reduction theory with level structure, based on generalisations of Atkin-Lehner operators.

2.1 INTRODUCTION

Let $n \geq 2$ be an integer. This article is concerned with bounding the sup-norm of Hecke-Maaß forms on the space

$$X_n(N) = \Gamma_0(N)\backslash \mathrm{SL}_n(\mathbb{R})/\mathrm{SO}(n)$$

in terms of the parameter N, called the *level*. Here, $\Gamma_0(N) \leq \mathrm{SL}_n(\mathbb{Z})$ is the subgroup of integral matrices with last row congruent to $(0, \dots, 0, *)$ modulo N, where $*$ stands for any non-zero residue class.

We normalise the invariant measure on $X_n(N)$ so that it has volume asymptotically equal to $N^{(n-1)+o(1)}$. Now let ϕ be a Hecke-Maaß form on this space, that is, a square-integrable joint eigenfunction of the invariant differential operators and the unramified Hecke algebra. Assuming that $\|\phi\|_2 = 1$, the sup-norm problem asks for non-trivial bounds on $\|\phi\|_\infty$. Several parameters can be considered for this question, the most studied being the spectral parameter and the level.

2.1.1 Some history

This problem has a rich history and the first breakthrough in the eigenvalue aspect for $n = 2$ was achieved by Iwaniec and Sarnak [IS95]. They prove that $\|\phi\|_\infty \ll_{N,\varepsilon} \lambda^{5/24+\varepsilon}$ for any $\varepsilon > 0$. This is an improvement over the so-called local bound $\|\phi\|_\infty \ll_N \lambda^{1/4}$. Their method of using an amplified pretrace

formula remains one of the main tools for obtaining such non-trivial, sub-local bounds.

In the level aspect, the baseline bound expected to hold is $\|\phi\|_\infty \ll_{\lambda,\varepsilon} N^\varepsilon$ for ϕ a newform. The first improvement for $n = 2$ is due to Blomer and Holowinsky [BH10], with important refinements by Harcos and Templier [HT12], [HT13], and the current record bound $\|\phi\|_\infty \ll_{\lambda,\varepsilon} N^{1/4+\varepsilon}$ is due to Khayutin, Nelson and Steiner [KNS22]. These papers deal with the case of square-free level N, and bounds for general N were achieved in [Sah17]. The fact that much of the work on this problem historically focused on square-free levels is in large part a consequence of using Atkin-Lehner operators. This aspect of the problem forms one of the main topics of this paper.

Though many other variations of the problem exist, we consider now its development in higher rank, that is, for $n > 2$. In the spectral aspect we only mention here a selection, namely the work of Blomer and Pohl [BP16] (for Sp_4), Blomer and Maga [BM16] (for SL_n), and Marshall [Mar14] (for more general Lie groups). They achieve power savings over the local bound for any $n \geq 2$, though they only consider the sup-norm of automorphic forms restricted to a fixed compact set. The implicit constants in their bounds thus depend on this set. An investigation of the global sup-norm is the topic of Blomer, Harcos and Maga's paper [BHM20].

The present article deals with the sup-norm problem in higher rank, in the level aspect. Despite the progress described above, there are very few results in this setting. The first result, due to Hu [Hu18], considers the case of prime-power levels $N = p^c$, where c is large, with ϕ corresponding to a so-called minimal vector, thus not applying to newforms. These forms are more suitable for the p-adic methods employed by Hu. Similar to many results in the spectral aspect, the bounds are given for the sup-norm of the restriction to a fixed adelic compact set, which we explain below in a classical language before stating the main theorem in this paper.

The second result [Tom23] is due to the author of this paper and concerns automorphic forms on a different family of locally symmetric spaces $\Gamma\backslash \mathrm{SL}_n(\mathbb{R})/\mathrm{SO}(n)$, where Γ is a subgroup coming from the units of an order in a division algebra of degree n. These spaces are compact and the bounds provided are global and in terms of their volume. The degree n is restricted to prime numbers and results can only be extended partially to odd degrees.

Moreover, the argument is based on the fact that proper subalgebras of division algebras of prime degree are automatically fields, and that zero is the only element of norm zero. The situation is decidedly different for the matrix algebra, whose orders give rise to the groups $\Gamma_0(N)$, and thus the methods of [Tom23] seem to be insufficient in this case.

Not only throughout the history of the sup-norm problem, but also of the subconvexity problem, the level aspect, particularly for prime or square-free levels, is often the last one to be successfully tackled. Given its significance in number theory, this suggests a serious, general difficulty and a need for new

ideas.

2.1.2 Statement of results

In this paper, we consider Hecke-Maaß cuspidal newforms on $X_n(N)$ for $n \geq 2$ and N prime. Let $\Omega \subset \mathrm{SL}_n(\mathbb{R})/\mathrm{SO}(n)$ be a fixed compact set and define

$$\Omega_N \subset X_n(N)$$

as the set of $z \in X_n(N)$ such that there is $\gamma \in \mathrm{SL}_n(\mathbb{Z})$ with $\gamma z \in \Omega$. It is easy to check that $\mathrm{vol}(\Omega_N) \asymp_\Omega \mathrm{vol}(X_n(N))$. We investigate the sup-norm of forms restricted to Ω_N. In adelic language, this corresponds to restricting to a fixed compact domain in $\mathrm{PGL}_n(\mathbb{A}_\mathbb{Q})$, as in [Hu18].

We prove two new results, the first of which applying to all $n \geq 2$ *prime*.

Theorem 2.1. *Let n and N be primes. Let ϕ be a Hecke-Maaß cuspidal newform on $X_n(N)$ with spectral parameter μ and define $\Omega_N \subset X_n(N)$ with respect to a fixed compact set $\Omega \in \mathrm{SL}_n(\mathbb{R})/\mathrm{SO}(n)$. For large N, we have the bound*

$$\left\|\phi|_{\Omega_N}\right\|_\infty \ll_{\Omega,n,\mu,\varepsilon} N^{-\frac{1}{2n^2}+\varepsilon}.$$

The proof involves understanding the geometric structure of the problem as well as handling rather delicate diophantine conditions. It is the latter that are not yet well enough understood in the case where n is not prime. However, the geometric ideas introduced in this paper are valid in full generality and already capture a significant part of the problem. To support this claim, we present below results for all $n \geq 2$, even improving those above numerically, assuming the existence of an efficient amplifier.

For this, let $\lambda(p)$ be the Hecke eigenvalue of ϕ for the Hecke operator T_p, where p is a prime not dividing N, normalised so that $\lambda(p) \ll p^{(n-1)/2}$ under the Ramanujan-Petersson conjecture. See Section 2.3.1 for a precise definition.

Hypothesis. Let $\delta > 0$ be any positive constant and $N \gg_{\delta,\mu} 1$ be large enough. If $L \gg N^\delta$, then

$$\sum_{p \in \mathcal{P}} \frac{|\lambda(p)|}{p^{(n-1)/2}} \gg_\varepsilon L^{3/4-\varepsilon}. \tag{2.1.1}$$

We prove in Lemma 2.3.1 that condition (2.1.1) is true assuming the Grand Riemann Hypothesis. It is similar to condition (1.24) in [IS95], which is checked in [Hua19] for dihedral Maaß forms and in [You18] for Eisenstein series and leads to an improved exponent in the bound of Iwaniec and Sarnak, as explained in [IS95, Remark 1.6].

Theorem 2.2. *Let $n \geq 2$ and N be a prime. Let ϕ be a Hecke-Maaß cuspidal newform on $X_n(N)$ with spectral parameter μ and define $\Omega_N \subset X_n(N)$ with respect to a fixed compact set $\Omega \in \mathrm{SL}_n(\mathbb{R})/\mathrm{SO}(n)$. Assuming hypothesis* (2.1.1)*, we have the bound*

$$\left\|\phi|_{\Omega_N}\right\|_\infty \ll_{\Omega,n,\mu,\varepsilon} N^{-\frac{1}{4n}+\varepsilon}.$$

In particular, the bound holds under the Grand Riemann Hypothesis.

Considering previous work on the sup-norm problem in higher rank, the main contribution of this paper is a new counting argument, based on the reduction of the domain Ω_N using generalised Atkin-Lehner operators, which might be of independent interest. These arguments significantly generalise and give a new perspective on the geometric methods of Harcos and Templier [HT13], which generated many strong results for the sup-norm problem on GL(2) (e.g. [Blo+20], [Sah17], [Ass a]). They also seem to be fundamentally different and provide stronger results than in the spectral aspect in higher rank, where savings are inverse super-exponential in n [Gil20], as opposed to our inverse polynomial savings. In any case, the methods presented here provide the first steps in tackling the level aspect in higher rank and, we believe, a useful framework for proving more general and possibly stronger results in the future.

2.1.3 Methods

For proving both main theorems, we employ an amplified pretrace formula to transform the analytic issue of bounding the sup-norm into a counting problem. This is one of the most common methods of studying the sup-norm of automorphic forms and goes back to the influential paper [IS95].

As in Proposition 2.3.3 below, we reduce the problem of bounding $\phi(z)$ for $z \in \mathrm{SL}_n(\mathbb{R})$ to counting matrices in sets of the form

$$H(z, m, N) := \{\gamma \in \mathcal{M}_n(\mathbb{Z}, N) \mid \det \gamma = m,\ z^{-1}\gamma z = O(m^{1/n})\},$$

where m is running over different, potentially sparse, sets of integers. Here, $\mathcal{M}_n(\mathbb{Z}, N)$ is the set of integral matrices with last row congruent to $(0, \dots, 0, *)$ modulo N. This is an order in the algebra of rational matrices.

2.1.3.1 LATTICES

To start, we give the sets $H(z, m, N)$ an interpretation in terms of lattices, which motivates the development of new tools introduced below. This is natural, since we recall that the space $X_n(1)$ parametrises shapes of unimodular lattices by associating to $z \in \mathrm{SL}_n(\mathbb{R})$ the lattice

$$L = \mathbb{Z}^n \cdot z \subset \mathbb{R}^n.$$

Here we understand $\mathbb{R}^n$ and $\mathbb{Z}^n$ as sets of row vectors. In this interpretation, the matrix z gives a specific basis for L. If N is prime, the space $X_n(N)$ now parametrises *pairs* (L, L_N) of lattices, up to simultaneous rotation by $\mathrm{SO}(n)$, where

$$L_N = \mathbb{Z}^n \cdot \mathrm{diag}(N, \dots, N, 1)z = (N\mathbb{Z} \times \cdots N\mathbb{Z} \times \mathbb{Z}) \cdot z,$$

is a sublattice of L.

Let $e_1, \dots, e_n$ be the standard basis for $\mathbb{R}^n$. We evaluate the condition

$$z^{-1}\gamma z = O(m^{1/n})$$

at the vectors e_i, after multiplying from the left by z. This amounts to the conditions

$$e_i \cdot \gamma z \in B(m^{1/n} \|e_i \cdot z\|)$$

for each i, where $B(r)$ is a Euclidean ball of radius $O(r)$ around 0. Note that, since γ is an integral matrix, $e_i \cdot \gamma z$ is a lattice point in L determining the i-th row of γ. Moreover, it is important to observe that $e_n \cdot \gamma z$ is additionally a lattice point in the sublattice L_N. On the other side, $e_i \cdot z$ is simply one of the basis vectors in the basis of L determined by z.

To count the number of relevant γ, we can therefore bound the number of possibilities for each of their rows and by the conditions above we reduce to counting lattice points in balls. However, this naïve strategy needs to be refined by an application of the Gram-Schmidt process, which we make precise in Section 2.6.2. By its very nature, this involves the Iwasawa coordinates of z.

In any case, it is apparent that the dependence on z manifests itself in two ways already at this level. Firstly, there might be many lattice points that we count because the basis vectors $e_i \cdot z$ which control the size of the balls are large. Secondly, the lattices L and L_N might be very dense, in the sense that they could have very short vectors relative to their covolume.

Understanding such issues is one of the main goals of reduction theory and the geometry of numbers. However, the level structure needs to be taken into consideration and, indeed, puts serious restrictions on the prospect of success for the amplified pretrace formula strategy. We develop a novel reduction theory with level structure in Section 2.5 and we describe the main ideas below.

2.1.3.2 GENERALISED ATKIN-LEHNER OPERATORS AND REDUCTION

In a nutshell, classical reduction theory provides a way to fit a fundamental domain for $X_n(1)$ inside a Siegel set (for the cusp at infinity). If $z \in \mathrm{SL}_n(\mathbb{R})$ lies in such a fundamental domain, its rows then provide a reduced basis for the lattice L, that is, a basis of vectors that are as short and as orthogonal as possible.

We also obtain in this way an interpretation of the Iwasawa coordinates of z in terms of the successive minima of L. See Section 2.2.3 for more details. This is not only important for implementing the refined counting strategy described above, but also for compensating with other tools when the latter fails.

For instance, solving the matrix counting problem optimally and plugging the result into the amplified pretrace formula *cannot* yield sub-baseline bounds when z is high enough in the cusp. One then compensates by using the Fourier

expansion, which gives strong bounds in terms of the Iwasawa y-coordinates following from the cuspidality of our automorphic form ϕ. This is common for many of the previous works [IS95, Lemma A.1], [HT12, Lemma 5.1], etc.

In the level aspect, already $z = \mathrm{id}_n$ has to be treated using the Fourier bound and notice that this point certainly lies in a standard bulk Ω_N of $X_n(N)$ for Ω a compact neighbourhood of the identity. From one perspective, which we do not explicate here further, this is because of the contribution of Eisenstein series on the spectral side of the pretrace formula. In our framework, the reason is that, even though L is a perfectly balanced lattice and z gives an actual orthogonal basis of shortest vectors, the sublattice L_N is maximally imbalanced.

A desirable reduction theory with level structure might thus fulfil the following. It should provide a basis for the lattice L that, while perhaps not reduced, gives useful information about shortest vectors in the sublattice L_N and about the Iwasawa coordinates, meaning the Gram-Schmidt process for the basis. It should also permit some understanding of the successive minima of both L and L_N. Of course, preserving the level structure means changing bases is only allowed by matrices in $\Gamma_0(N)$. However, there are additional symmetries at our disposal.

It was recognised early on in the treatment of the sup-norm problem in the level aspect that Atkin-Lehner operators would be useful for such reductions. It is classically not hard to see that one can fit the fundamental domain for $X_2(N)$, where N is square-free, quotiented out by the action of these operators in a Siegel set of *finite* volume. This is because the Atkin-Lehner operators for N square-free conjugate all cusps to the cusp at infinity. Unfortunately, for powerful levels there is a deficiency of Atkin-Lehner operators and this forms an important reason why the first and many results on the sup-norm problem are restricted to square-free levels.

The first authors to connect these group theoretic facts to lattices were Harcos and Templier in [HT12, Lemma 2.2]. For example, at the level of lattices, the Fricke involution for prime levels can be understood as switching the lattices in the pair (L, L_N). Together with ideas from reduction theory, this allows us to trade imbalancedness of L or L_N for closeness of z to the cusp (see loc. cit.). Effectively, when the matrix counting results are weak, the Fourier bound gets better.

Generalising the case $n = 2$, we study the symmetries of $X_n(N)$. The point of departure from the classical case is the observation that $\mathrm{PGL}(n)$ for $n > 2$ has an additional outer automorphism, given by $z \mapsto z^{-T}$. This corresponds to taking duals, either at the level of lattices, or at the level of automorphic forms. In this paper, we use this to introduce in Section 2.4 a higher-rank Atkin-Lehner operator corresponding to the *Fricke involution*. It has probably been implicitly present in the theory of newforms, yet an explicit definition seems hard to find in the literature.

Definition. Let

$$A_N = N^{-1/n} \operatorname{diag}(1, \dots, 1, N)$$

and define the *Fricke involution* $W_N : L^2(X_n(N)) \longrightarrow L^2(X_n(N))$ as

$$W_N \phi(z) = \phi(A_N \cdot z^{-T}).$$

We also perform an investigation of other potential generalisations of Atkin-Lehner operators. First, we prove that the normaliser of $\Gamma_0(N)$ inside $\mathrm{PGL}_n(\mathbb{R})$, the source of Atkin-Lehner operators for $n = 2$, is *trivial* for $n > 3$. We refer to Section 2.4.1.

Theorem 2.3. *For $n > 2$, the normaliser of $\Gamma_0(N)$ inside $\mathrm{PGL}_n(\mathbb{R})$ is trivial.*

We then provide a different perspective on the classical Atkin-Lehner operators and show in Proposition 2.4.4 that the only possible generalisation in this interpretation is the Fricke involution. On the one hand, this is in contrast to the case of square-free levels in PGL(2), but it is also a reflection of the remarkable *lack* of such symmetries for powerful levels. Therefore, we first only consider the case of prime level in this paper, similar to the common restrictions in the rank-one case.

The main result of our reduction theory is given in Proposition 2.5.2. It satisfies the intuition from the $n = 2$ case, where the bulk of the reduced fundamental domain is at $\mathfrak{I}(z) \asymp 1/N$. In general, there are the Iwasawa coordinates $y_1, \dots, y_{n-1}$ and the bulk can be found at

$$y_1 \asymp \frac{1}{N},\ y_2 \asymp \dots \asymp y_{n-1} \asymp 1.$$

In this region, we prove that reduced z satisfy that both L and L_N are balanced in Lemma 2.6.1. As noted above, there is also the exceptional region Ω of the bulk, where counting results would be too weak due to imbalancedness of the lattices, but the Fourier bound suffices due to closeness to the cusp.

However, the reduction of the full fundamental domain for $\Gamma_0(N)$ is more complex, as can be seen from the case work in Section 2.5.2. It seems that more refined information can be extracted and doing so would be an important next step in the study of the sup-norm problem in the level aspect.

In higher rank, the reduction process involves the outer automorphism included in the Fricke involution and thus dualising lattices. We are therefore required to develop tools for keeping track of sizes of vectors in the lattices associated to z and its conjugate under the Fricke involution, as well as their duals. This is the content of Section 2.5.1 and Table 2.1. We have found the language of wedge products particularly useful for this because of its flexibility in relating lengths of vectors in lattices and their duals with Iwasawa coordinates.

As a historical interlude, we point out some connections of the above considerations with previous work. The Atkin-Lehner involutions were already

used in the breakthrough [BH10], but balancedness of lattices was interpreted in terms of Diophantine approximation properties of the Iwasawa coordinates, using terminology from the circle method.

The language of lattices was used directly in [HT12], [HT13], and subsequent works, and lead to strong numerical improvements to the bounds. However, the counting problem is interpreted using coordinates not truly inherent to lattices. Many computations in the GL(2) case use, in fact, the "sporadic" symplectic nature of this group. This is not available in higher degree and the direct use of coordinates seems to be very cumbersome.

For the family of groups $\mathrm{PGL}(n)$, some ideas reminiscent of the more general strategy used here can be seen in [BHM20, Sec. 3.2]. We refer also to [Ven06], where certain aspects of the geometry of $X_n(N)$ are studied using lattices as well.

2.1.3.3 DETECTING SPARSE SEQUENCES OF DETERMINANTS

The upshot of the reduction theory with level and the iterative counting strategy is that we get bounds for the set

$$\bigcup_{1\leq m\leq \Lambda} H(z,m,N)$$

for a parameter Λ small enough in terms of N, uniformly in the balanced part of Ω_N. The motto of the counting strategy under these conditions is a rigidity principle: *the last row of $\gamma \in H(z,m,N)$ determines the whole matrix*.

However, the unconditional amplifier of [BM15] gives rise to a counting problem where matrices have perfect power determinants, for instance, n-th powers. Such a sequence of determinants is too sparse and the method above, averaging over all determinants, produces gross over-counting. Similar issues are well-known already in the classical case $n = 2$ (see e.g. the special treatment of square determinants in [HT13]).

The appearance of sparse sequences of determinants on the geometric side is due to the lack of good lower bounds for Hecke eigenvalues. Indeed, such bounds are precisely what Hypothesis (2.1.1) provides. Unconditionally, there is thankfully a substitute obtained from Hecke relations, such as $\lambda(p)^2-\lambda(p^2) = 1$ in suitable normalisation for $n = 2$, from which one derives that at least one of the two eigenvalues is bounded from below. Introducing the Hecke operator T_{p^2} in this way results in sequences of square determinants, and we have similar phenomena in higher degree.

We are able to detect perfect power determinants by using a refinement of the counting strategy above (see Section 2.6.3). The problem reduces to counting solutions to an equation of the shape

$$\chi_\gamma(X) - Y^\nu = 0$$

for $1 \leq \nu \leq n$, where χ_γ is the characteristic polynomial of γ. If this equation is irreducible, then a powerful theorem of Heath-Brown [HB02] provides an adequate non-trivial bound.

To treat the case where the polynomial is reducible, we assume that n is prime to simplify the classification of these degenerate cases. We can thus reduce to counting matrices with $\chi_\gamma(X) = (X - m)^n$. For $n = 2$, this is the special case of parabolic matrices that was also handled in [HT12, Lemma 4.1].

Finally, resolving this problem involves some group theoretic investigations once more. We classify the cusps of $X_n(N)$ as in Lemma 2.6.8, of which there are n many, and observe the action of the Fricke involution on them. The cusp corresponding to the identity element, informally the cusp at infinity, can be dealt with by the counting methods already introduced. The one corresponding to the long Weyl element is conjugated to the identity by the Fricke involution.

Counting at "intermediate" cusps presents new challenges, which might be a consequence of the lack of more symmetries of $X_n(N)$ for $n > 2$. Although much of what is developed in this paper appears to the author to be conceptually necessary and inherent to the problem, this last step is solved by a trick, as one might call it. We use the specific shape of the amplifier of Blomer and Maga. Namely, we take advantage of the fact that, for certain Hecke sets attached to primes p and q, the determinantal divisors are asymmetric in terms of p and q, as in (2.3.1). This eventually collapses an average over two primes to one over a single prime (the case $p = q$), and leads to the required power saving.

Notation

By the Vinogradov notation $f(x) \ll g(x)$ for two functions f, g it is meant that $|f(x)| \leq C \cdot |g(x)|$, at least for large enough x, for some $C > 0$ called the implied constant. Similarly, for a matrix X and a scalar function $f(X)$ we say that $X = O(f(X))$ when $\|X\| \leq C \cdot f(X)$ for some constant $C > 0$ and some choice of matrix norm $\|\cdot\|$.

We use $\ll_P$ to say that the implied constant depends on a parameter P, yet we do not always add the subscript if it is clear from context in order to avoid clutter. For instance, dependency on the compact space $\Omega \subset \mathrm{SL}_n(\mathbb{R})$ includes dependency on n.

2.2 PRELIMINARIES ON LATTICES

Consider the real vector space $V = \mathbb{R}^n$ with standard inner product $\langle v, w \rangle = v \cdot w^T$, where we think of $v, w \in V$ as *row* vectors in the standard basis $e_1, \ldots, e_n$. Let z be a matrix in $\mathrm{GL}_n(\mathbb{R})$ and define L_z to be the lattice $\mathbb{Z}^n \cdot z$ inside V. Note that $e_i \cdot z$ is equal to the i-th row of z. We also define the inner product and

norm

$$\langle v, w\rangle_z = \langle vz, wz\rangle, \quad \|v\|_z = \sqrt{\langle vz, vz\rangle},$$

for $v, w \in V$.

The *dual* lattice L_z^* is defined as the set of vectors w such that $\langle v, w\rangle \in \mathbb{Z}$ for all $v \in L_z$. It is straight-forward to compute that

$$L_z^* = L_{z^{-T}}.$$

We also note that $L_z = L_w$ for any $w \in \mathrm{GL}_n(\mathbb{Z}) \cdot z$.

2.2.1 Exterior powers

If k is a positive integer, the k-th exterior power of L_z is denoted by $\bigwedge^k L_z$ and is defined as the $\mathbb{Z}$-span of the wedge products $v_1 \wedge \cdots \wedge v_k$ for all $v_1, \ldots, v_k \in L_z$. It is a lattice inside $\bigwedge^k V$. The inner product is given by

$$\langle v_1 \wedge \cdots \wedge v_k, w_1 \wedge \cdots \wedge w_k\rangle = \det(\langle v_i, w_j\rangle)_{1\le i,j\le k}$$

and extended linearly.

We have an isomorphism

$$\bigwedge^{n-1} V \cong V,$$

by sending $w \in \bigwedge^{n-1} V$ to $v \in V$ such that, for all $u \in V$,

$$w \wedge u = \langle v, u\rangle.$$

We make implicit use of the fact that $\bigwedge^n \mathbb{R}^n \cong \mathbb{R}$ and of an intermediary isomorphism with the dual space V^*. The isomorphism above is an isometry.

Indeed, we can check that an orthonormal basis is sent to an orthonormal basis. Let $(e_1, \ldots, e_n)$ be the standard orthonormal basis of V. Then

$$(e_1 \wedge \cdots \wedge e_{n-1},\ e_1 \wedge \cdots \wedge e_{n-2} \wedge e_n,\ \ldots,\ e_2 \wedge \cdots \wedge e_n),$$

is an orthonormal basis of $\bigwedge^{n-1} V$, formed by respectively removing each vector e_i from the wedge product $e_1 \wedge \ldots e_n$. It is then easy to check that

$$e_1 \wedge \ldots \wedge e_{n-1} \mapsto e_n,\ e_1 \wedge \ldots \wedge e_{n-2} \wedge e_n \mapsto -e_{n-1},\ \ldots,\ e_2 \wedge \ldots \wedge e_n \mapsto (-1)^{n-1} e_1.$$

Lemma 2.2.1. *The lattice $\bigwedge^{n-1} L_z$ is isometric to the lattice $L_{\det(z)\cdot z^{-T}}$.*

Proof. We use the isomorphism $\bigwedge^{n-1} V \cong V$ described in the paragraphs above. The wedge product has the property that $v_1 z \wedge \ldots \wedge v_n z = \det(z) \cdot v_1 \wedge \ldots \wedge v_n$ for n row vectors (v_i). This allows us to check that, under the given isomorphism,

$$e_1 z \wedge \ldots \wedge e_{n-1} z \mapsto \det(z) \cdot e_n z^{-T},$$

and analogously for the other basis vectors above. □

2.2.2 Successive minima

Throughout this paper, we consider successive minima of lattices L_z with respect to the unit ball $B^1 \subset V$ given by the standard inner product. When considering the exterior products of these lattices, successive minima are defined with respect to the compounds of the unit ball, as in the work of Mahler [Mah55] (refer also to [Eve19], Section 3, for a modern treatment).

More precisely, the *k-th compound* of B^1, denoted here by B^k, is defined as the convex hull of the points $x_1 \wedge \cdots \wedge x_k$, for all $x_1, \ldots, x_k \in B^1$. Mahler notes that B^k is a bounded, convex body in $\bigwedge^k \mathbb{R}^n$, though generally not a sphere (see Section 4 in [Mah55]). Nevertheless, since B^k is bounded and 0 is an inner point of B^k, there are constants $c_{k,n}, C_{k,n} > 0$ such that

$$B(n,k,c_{k,n}) \subset B^k \subset B(n,k,C_{k,n}),$$

where $B(n,k,r)$ is the ball of radius r inside $\bigwedge^k \mathbb{R}^n$. As such, the lenght l of the shortest non-zero vector in $\bigwedge^k L_z$ can be approximated as

$$l \asymp_{n,k} \mu_1,$$

where μ_1 is the first successive minimum of $\bigwedge^k L_z$ with respect to B^k.

A theorem of Mahler (Theorem 3 in [Mah55]; Theorem 3.2 in [Eve19]) relates the successive minima of a lattice to those of its exterior powers. We state here a special case, relevant in this paper.

Lemma 2.2.2. *Let L be a lattice in $\mathbb{R}^n$ and let $\lambda_1, \ldots, \lambda_n$ be its successive minima with respect to the unit ball B^1. Let μ_1 be the first successive minimum of the lattice $\bigwedge^k L$ with respect to B^k. Then*

$$\mu_1 \asymp_{n,k} \lambda_1 \cdots \lambda_k.$$

As explained above, this lemma implies that, if l is the length of the shortest non-zero vector in $\bigwedge^k L$, then

$$l \asymp_{n,k} \lambda_1 \cdots \lambda_k.$$

We use this relation in Section 2.5.2.

We also recall here a classical theorem of Minkowski (see [Cas97, Theorem VIII.1]), stating that

$$d(L) \ll_n \lambda_1 \cdots \lambda_n \ll_n d(L), \tag{2.2.1}$$

where $d(L)$ is the determinant of the lattice, e.g. $d(L_z) = \det(z)$. In particular, for a lattice of determinant 1, called a *unimodular lattice*, we have

$$\lambda_1 \ll_n 1, \tag{2.2.2}$$

using the inequalities $\lambda_1 \leq \lambda_i$, for all i.

The detailed study of successive minima of L_z is crucial in this paper due to the following well-known lemma (see e.g. [BHM16, Lemma 1]), which we apply when counting integral matrices, as explained at the end of Section 2.3.

Lemma 2.2.3. *Let $L \subset \mathbb{R}^n$ be a lattice and let $\lambda_1 \leq \ldots \leq \lambda_n$ be its successive minima with respect to the unit ball. Let $B \subset \mathbb{R}^n$ be a ball of radius R and arbitrary centre. We have the inequality*

$$|L \cap B| \ll_n 1 + \frac{R}{\lambda_1} + \frac{R^2}{\lambda_1\lambda_2} + \cdots + \frac{R^n}{\lambda_1 \cdots \lambda_n}.$$

2.2.3 Iwasawa coordinates and reduction theory

Let $\mathbb{H} = \mathbb{H}_n$ be the generalised upper half plane, that is

$$\mathbb{H} = \mathrm{GL}_n(\mathbb{R})/(\mathrm{O}(n) \cdot \mathbb{R}^\times) \cong \mathrm{SL}_n(\mathbb{R})/\mathrm{SO}(n).$$

In particular, the statement $z \in \mathbb{H}$ is taken to imply $z \in \mathrm{SL}_n(\mathbb{R})$.

By the Iwasawa decomposition (see Section 1.2 in [Gol06]), we can take elements in $\mathbb{H}$ to be of the form $z = n(x) \cdot a(y)$, where $n(x) = (x_{ij})_{1 \leq i,j \leq n} \in \mathrm{SL}_n(\mathbb{R})$ is upper triangular unipotent, meaning that it satisfies

$$x_{ij} = \begin{cases} 0, & j < i; \\ 1, & i = j; \end{cases}$$

and $a(y)$ is diagonal, parametrised as

$$a(y) = \mathrm{diag}(d_1, \ldots, d_n) = \mathrm{diag}(dy_1 \cdots y_{n-1},\ \ldots,\ dy_1y_2,\ dy_1,\ d),$$

where $d, y_1, \ldots, y_{n-1} \in \mathbb{R}_{>0}$ such that

$$\det a(y) = d^n y_1^{n-1} y_2^{n-2} \cdots y_{n-1} = 1.$$

Define the Siegel set $\mathfrak{S}$ to be the set of all $z = n(x)a(y) \in \mathrm{SL}_n(\mathbb{R})$ such that

$$|x_{ij}| \leq \frac{1}{2}$$

for all $i < j$ and

$$y_i \geq \frac{\sqrt{3}}{2},$$

for all i, using the Iwasawa coordinates defined above. Reduction theory (see [Bor19, Theorem I.1.4] or [Gol06, Proposition 1.3.2]) shows that

$$\mathrm{SL}_n(\mathbb{R}) = \mathrm{SL}_n(\mathbb{Z}) \cdot \mathfrak{S}.$$

If $z \in \mathfrak{S}$, we say that $(e_1 z, \ldots, e_n z)$ is a *reduced basis* for L_z. We also remark that reduction theory allows us to pick $e_n z$ to be any vector of shortest length in L_z (this is, indeed, part of the reduction algorithm).

Remark 2.2.4. It is useful in later sections to note an embedding of $\mathrm{SL}_{n-1}(\mathbb{R})$ into $\mathrm{SL}_n(\mathbb{R})$ and the connection between the two systems of Iwasawa coordinates. More precisely, we can write $z = n(x)a(y) \in \mathbb{H}$ as

$$z = \begin{pmatrix} dy_1 \cdot w & * \\ 0 & d \end{pmatrix},$$

where $w \in \mathbb{H}_{n-1}$ is a matrix in $\mathrm{GL}_{n-1}(\mathbb{R})$. Though not normalised, we can use a variant of the Iwasawa coordinates (it is the one used in Definition 1.2.3 in [Gol06]) to write $w = n(x') \cdot a(y')$, where

$$a(y') = \mathrm{diag}(y_2 \cdots y_{n-1}, \ldots, y_2, 1).$$

Multiplication of z by parabolic matrices

$$g = \begin{pmatrix} h & 0 \\ 0 & 1 \end{pmatrix} \in \mathrm{SL}_n(\mathbb{Z})$$

with $h \in \mathrm{SL}_{n-1}(\mathbb{Z})$, acts on w by sending it to $h \cdot w$ and otherwise leaves the last row of z invariant. Reduction theory in degree $n-1$ now implies that there is a parabolic block matrix $g \in \mathrm{SL}_n(\mathbb{Z})$ as above so that $g \cdot z = n(x) \cdot a(y)$ with $y_i \geq \sqrt{3}/2$ for $i = 2, \ldots, n-1$.

More generally one could define a Siegel set $\mathfrak{S}_\eta$ for any $\eta > 0$ as the set of all $z = n(x)a(y) \in \mathrm{SL}_n(\mathbb{R})$ such that $|x_{ij}| \leq 1/2$ and $y_i \geq \eta$. The following is a well-known fact in reduction theory, which we state and prove in the version needed in this paper.

Lemma 2.2.5. *If $z = n(x)a(y) \in \mathfrak{S}_\eta$ and $\lambda_1 \leq \ldots \leq \lambda_n$ are the successive minima of L_z, then*

$$\lambda_i \asymp_{n,\eta} \|e_{n+1-i}\|_z \asymp_{n,\eta} d_{n+1-i}. \tag{2.2.3}$$

Proof. Notice that we can find $n(x') \in \mathrm{SL}_n(\mathbb{R})$ upper triangular unipotent such that

$$z = n(x)a(y) = a(y)n(x').$$

One can easily check that

$$x'_{ij} = x_{ij} \cdot d_j/d_i = x_{ij} \cdot (y_{n-i} \cdots y_{n-j+1})^{-1} \ll_\eta 1$$

for $i < j$ when $z \in \mathfrak{S}_\eta$. Thus, every entry of $n(x')$ is bounded uniformly in terms of η and so the operator norm of $n(x')$ with respect to the Euclidean norm is bounded in terms of η and n. Since the entries of the inverse $n(x')^{-1}$ are polynomials in the x'_{ij}, we see analogously that its operator norm is also bounded and we can deduce that

$$\|v\|_{n(x')} \asymp_{n,\eta} \|v\|$$

for all vectors $v \in \mathbb{R}^n$. Using coordinates with respect to the standard basis $e_1, \dots, e_n$, we have

$$\|(c_1, \dots, c_n)\|_z = \|(c_1, \dots, c_n)\|_{a(y)n(x')} = \\ \|(d_1 c_1, \dots, d_n c_n)\|_{n(x')} \asymp \|(d_1 c_1, \dots, d_n c_n)\| .$$

Now $e_n \cdot z, \dots, e_1 \cdot z$ are linearly independent vectors in L_z, which implies that $\lambda_i \leq \|e_{n+1-i}\|_z$. Conversely, suppose that $v_1, \dots, v_k \in L_z$ are linearly independent vectors with $\max \|v_i\| = \lambda_k$. In particular, for any i we have $v_i = (c_{i1}, \dots, c_{in}) \cdot z$ with $c_{ij} \in \mathbb{Z}$ and there is at least one $i \in \{1, \dots, k\}$ such that $c_{ij} \neq 0$ for some $j \leq n+1-k$ (we are just expressing the fact that $v_1, \dots, v_k$ cannot be contained in the linear span of the $k-1$ vectors $e_{n+2-k}z, \dots, e_n z$). As such, we have

$$\begin{aligned} \lambda_k \geq \|v_i\| &= \|(c_{i1}, \dots, c_{in})\|_z \asymp_{n,\eta} \|(d_1 c_{i1}, \dots, d_n c_{in})\| \\ &\geq d_j = \frac{d_j}{d_{n+1-k}} \cdot d_{n+1-k} \geq \eta^{n+1-k-j} d_{n+1-k}. \end{aligned}$$

□

We recall also another standard lemma, which informally says that a reduced basis behaves similarly to an orthogonal basis.

Lemma 2.2.6. *Let $(v_1, \dots, v_n)$ be a reduced basis of a lattice L. Let $v \in L$ and write $v = \sum_{i=1}^n a_i v_i$ with $a_i \in \mathbb{Z}$. Then $a_i \ll_n \|v\| / \|v_i\|$.*

Proof. See Lemma 1 in [Ven06]. □

Finally, if $\Omega \subset \mathbb{H}$ is a compact set (in particular, it projects to a compact set in the space of lattices $\mathrm{SL}_n(\mathbb{Z})\backslash\mathbb{H}$) and $z \in \Omega$, then $\lambda_1 \gg_\Omega 1$ by Mahler's criterion [Bor19, Corollary I.1.9]. The other successive minima must then also be bounded from below, so $\lambda_i \gg 1$. By (2.2.1), we have that

$$1 \ll \lambda_2^{n-1} \leq \lambda_2 \cdots \lambda_n \ll 1/\lambda_1 \ll 1$$

since z has determinant 1. Thus $\lambda_2 \asymp 1$ and inductively we find $\lambda_i \asymp_\Omega 1$ for all i. We may say L_z is an Ω-balanced lattice.

For any $z \in \mathbb{H}$ we say that z *reduces* to Ω if there is $w \in \Omega$ such that $L_z = L_w$, in other words if there is $\gamma \in \mathrm{SL}_n(\mathbb{Z})$ such that $z = \gamma w$. The discussion in the paragraph above proves the following lemma.

Lemma 2.2.7. *Suppose that $z \in \mathbb{H}$ reduces to a compact set Ω and let $\lambda_1, \dots, \lambda_n$ be the successive minima of L_z. Then $\lambda_i \asymp_\Omega 1$ for all $i \in \{1, \dots, n\}$, where the implicit constant depends only on Ω.*

2.3 THE AMPLIFIED PRETRACE FORMULA

We follow the amplification scheme of Blomer and Maga [BM15], using their archimedean test function but giving also a version that simplifies the sum over Hecke eigenvalues by assuming a conjecture about their sizes.

Let $G = \mathrm{SL}_n(\mathbb{R})$, $K = \mathrm{SO}(n)$, $\Gamma = \Gamma_0(N)$, and let ϕ be the cuspidal Hecke-Maaß form of level N that we wish to bound. Let $\mu = (\mu_1, \dots, \mu_n)$ be the spectral parameters of ϕ. We may embed ϕ into a basis of the space of Hecke-Maaß cusp forms for $\Gamma_0(N)$. More precisely, we have a spectral decomposition

$$L^2(\Gamma_0(N)\backslash\mathbb{H}) = \int V_\omega d\omega = L^2_{\mathrm{cusp}} \oplus L^2_{\mathrm{Eis}},$$

where every V_ω is a one-dimensional space generated by an eigenform ϕ_ω of the algebra of invariant differential operators and the Hecke algebra. Let μ_ω be the spectral parameter of ϕ_ω and assume that $\phi = \phi_{\omega_0}$. Note moreover that L^2_{cusp} has a discrete decomposition.

Recall the Cartan decomposition $G = KAK$, where A is the subgroup of diagonal matrices. The latter has a Lie algebra $\mathfrak{a}$, on which the Weyl group W of G acts. We define the Cartan projection $C(g) \in \mathfrak{a}/W$ of an element $g \in G$ via the Cartan decomposition $g = k_1 \exp(C(g))k_2$, where $k_1, k_2 \in K$. Now pick a W-invariant norm $\|\cdot\|$ on $\mathfrak{a}$. We note that, if $\|C(g)\| \ll 1$, then by exponentiating we have

$$g = k + O(1),$$

where $k \in K$ and $O(1)$ stands for a matrix whose norm (by equivalence, any norm) is $O(1)$.

2.3.1 *The Hecke algebra and Hecke eigenvalues*

We now briefly review some aspects of the structure of the unramified Hecke algebra. Let p be a prime not dividing N and $\mathbf{a} = (a_1, \dots, a_n) \in \mathbb{Z}^n$. The double coset

$$\Gamma \,\mathrm{diag}(p^{a_1}, \dots, p^{a_n})\Gamma = \bigcup_j \Gamma\alpha_j$$

defines a Hecke operator

$$T_{\mathbf{a}}(p)(\psi)(z) = \sum_j \psi(\alpha_j \cdot z),$$

where ψ is any function on $\Gamma\backslash\mathbb{H}$. We define the standard Hecke operator as

$$T(p) = T_{(1,0,\dots,0)}(p).$$

One computes that the adjoint of $T(p)$ is the operator $T'(p) = T_{(1,\dots,1,0)}(p)$. Let $\lambda(p, \phi_\omega)$ be the eigenvalue of ϕ_ω under $T(p)$, so that $\overline{\lambda(p, \phi_\omega)}$ is its eigenvalue

under $T'(p)$. By [BM15, Lemma 4.4] we have

$$T(p) \cdot T'(p) = a \cdot T_{(2,1,\dots,1,0)}(p) + b \cdot p^{n-1} \,\mathrm{id},$$

where $a, b \ll 1$. Furthermore, if p and q are distinct primes not dividing N, then we have the multiplication rule on double cosets

$$\Gamma \,\mathrm{diag}(p, 1, \dots, 1)\Gamma \cdot \Gamma \,\mathrm{diag}(q, \dots, q, 1)\Gamma = \Gamma \,\mathrm{diag}(pq, q, \dots, q, 1)\Gamma \qquad (2.3.1)$$

in the Hecke algebra, corresponding to the composition $T(p) \cdot T'(q)$ (see [BM15, Section 6]).

Let now $L > 0$ be a parameter and $\mathcal{P}$ be the set of primes contained in $[L, 2L]$, not dividing N. Define

$$A_\omega = \left| \sum_{p \in \mathcal{P}} \frac{\lambda(p, \omega)}{p^{(n-1)/2}} \cdot x_p \right|^2,$$

where $x_p = |\lambda(p, \omega_0)|/\lambda(p, \omega_0)$.

We use here the normalised eigenvalues $\lambda(p, \omega)/p^{(n-1)/2}$ as defined in [Gol06, (9.3.5)]. Note that

$$A_{\omega_0} = \left| \sum_{p} |\lambda(p, \omega_0)/p^{(n-1)/2}| \right|^2.$$

A lower bound for this quantity is given in Hypothesis (2.1.1). We now prove it follows from GRH.

Lemma 2.3.1. *Let $\delta > 0$ be any positive constant and $N \gg_\delta 1$ be large enough. Assuming the Grand Riemann Hypothesis, if $L > N^\delta$, then*

$$\sum_{p \in \mathcal{P}} \frac{|\lambda(p, \omega_0)|}{p^{(n-1)/2}} \gg_\varepsilon L^{3/4-\varepsilon}. \qquad (2.3.2)$$

Proof. The following are standard computations and we refer to Sections 5.1, 5.3, 5.6, 5.7 in [IK04] for more details. Let $\lambda(p) = \lambda(p, \omega_0)/p^{(n-1)/2}$ and note that these give the coefficients of the L-function attached to ϕ or, equivalently, to the automorphic representation π generated by ϕ. Let $L_{\mathrm{RS}}(s) = L(s, \pi \times \tilde{\pi})$ be the Rankin-Selberg L-function and define $\Lambda_{\mathrm{RS}}(n)$ to be its coefficients, so that

$$\frac{L'_{\mathrm{RS}}}{L_{\mathrm{RS}}}(s) = \sum_{n=1}^{\infty} \frac{\Lambda_{\mathrm{RS}}(n)}{n^s}.$$

Then we have $\Lambda_{\mathrm{RS}}(p) = |\lambda(p)|^2 \log p$.

The prime number theorem under GRH states that

$$\sum_{n\leq x}\Lambda_{\mathrm{RS}}(n) = x + O_{\varepsilon,\mu}(x^{1/2+\varepsilon}\cdot N^{\varepsilon}). \tag{2.3.3}$$

For $y \leq \sqrt{x}$, we obtain that

$$\sum_{x\leq n\leq x+y}\Lambda_{\mathrm{RS}}(n) \ll_{\varepsilon,\mu} x^{1/2+\varepsilon}N^{\varepsilon}.$$

Now we note that $\Lambda_{\mathrm{RS}}(n) \geq 0$ for all n by the definition of the Rankin-Selberg convolution. It follows from the prime number theorem above by dropping all but one term that

$$\lambda(p)^2 \ll \Lambda_{\mathrm{RS}}(p) \ll x^{1/2+\varepsilon}N^{\varepsilon}$$

for $p \asymp x$.

Let $x \gg N^{\delta}$ for some $\delta > 0$. The bound above and (2.3.3) imply that

$$x^{1-\varepsilon} \ll \sum_{p\asymp x}|\lambda(p)|^2 \ll x^{1/4+\varepsilon}\sum_{p\asymp x}|\lambda(p)|.$$

This proves the claim. □

Remark 2.3.2. It is expected that a stronger version of (2.1.1) holds, that is, with exponent 1 instead of 3/4. To prove this we would require the Ramanujan-Petersson conjecture. This would improve the saving in Theorem 2.2 by doubling the exponent.

2.3.2 Amplifiers

Let $\mathcal{M}_n(\mathbb{Z}, N)$ be the set of integral matrices with last row congruent to $(0,\dots,0,*)$ modulo N. For $(m,N)=1$ let

$$H(m,N) := \{\gamma \in \mathcal{M}_n(\mathbb{Z},N) \mid \det\gamma = m\}$$

and

$$H(z,m,N) := \{\gamma \in \mathcal{M}_n(\mathbb{Z},N) \mid \det\gamma = m,\ z^{-1}\gamma z = O(m^{1/n})\},$$

where the implicit constant depends on n, dependence which we suppress throughout the arguments.

Proposition 2.3.3. *Let ϕ be a Hecke-Maaß form for $\Gamma_0(N) \leq \mathrm{SL}_n(\mathbb{R})$ with spectral parameter μ, let $L \gg N^{\delta}$ for some $\delta > 0$ be a parameter and let $\mathcal{P}$ be the set of primes in $[L, 2L]$, not dividing N. Then, assuming Hypothesis* (2.1.1)*, we have the bound*

$$L^{3/2-\varepsilon}|\phi(z)|^2 \ll_{\mu,\varepsilon} |\mathcal{P}|\cdot|H(z,1,N)| + \frac{1}{L^{n-1}}\sum_{p,q\in\mathcal{P}}|H(z,p\cdot q^{n-1},N)|.$$

Proof. We choose the archimedean test function $f_\mu : C_c^\infty(K\backslash G/K) \longrightarrow \mathbb{C}$ defined in [BM15, Section 3]. It has compact support and is bounded $f_\mu \ll_{\mu,n} 1$ in terms of μ, where the dependence on μ is continuous.[1] Its spherical transform $\tilde{f}_\mu$ satisfies

$$\tilde{f}_\mu(\mu) \geq 1$$

and is non-negative on all possible spectral parameters occurring in the decomposition of $L^2(\Gamma_0(N)\backslash\mathbb{H})$. Finally, when writing $f_\mu(g)$ for $g \in \mathrm{GL}_n(\mathbb{R})$, where $\det(g) > 0$, we mean $f_\mu(g/\det(g)^{1/n})$, thus extending the domain of f_μ by postulating its invariance under scalars.

Now consider

$$\int A_\omega \cdot \tilde{f}_\mu(\mu_\omega)\phi_\omega(z)\overline{\phi_\omega(w)}d\omega,$$

expand every A_ω and group terms into expressions of the form

$$\frac{1}{(pq)^{(n-1)/2}} \int \lambda(p,\omega)x_p \cdot \overline{\lambda(q,\omega)x_q} \cdot \tilde{f}_\mu(\mu_\omega)\phi_\omega(z)\overline{\phi_\omega(w)}d\omega,$$

which is equal to

$$S_{p,q} = \frac{x_p\overline{x_q}}{(pq)^{(n-1)/2}} \cdot T(p)T'(q) \cdot \int \tilde{f}_\mu(\mu_\omega)\phi_\omega(z)\overline{\phi_\omega(w)}d\omega,$$

where the Hecke operators act in the variable z. We apply the pretrace formula to obtain the geometric side

$$S_{p,q} = \frac{x_p\overline{x_q}}{(pq)^{(n-1)/2}} \cdot T(p)T'(q) \sum_{\gamma\in\Gamma} f_\mu(z^{-1}\gamma w),$$

where again we write $\Gamma_0(N) = \Gamma$ for brevity. Note that for any double coset $\Gamma g\Gamma$, the corresponding Hecke operator T_g acts on the variable z by

$$T_g \sum_{\gamma\in\Gamma} f_\mu(z^{-1}\gamma w) = \sum_{\gamma\in\Gamma g\Gamma} f_\mu(z^{-1}\gamma w),$$

by definition and sum unfolding. Moreover, using the compact support of f_μ, we can bound the right-hand side by

$$\sum_{\gamma\in\Gamma g\Gamma} f_\mu(z^{-1}\gamma w) \ll_\mu |\{\gamma \in \Gamma g\Gamma \mid z^{-1}\gamma w = \det(\gamma)^{1/n}(k + O(1)),\ k \in K\}|$$

using the triangle inequality. Since K is compact, we can simplify $k + O(1)$ to $O(1)$, where the implicit constant depends on n.

[1] In fact, there is an explicit bound for the function f_μ. However, it is only useful in the spectral aspect. For our purposes, we may simply bound f_μ by a constant depending on μ, but independent of the level.

We now write the compositions $T(p)\cdot T'(q)$ as linear combinations of Hecke operators T_g. Let $z = w$ and assume that $p \neq q$. Recalling that $T(p)\cdot T'(q)$ is the Hecke operator corresponding to

$$\Gamma \operatorname{diag}(pq, q, \dots, q, 1)\Gamma,$$

and that $x_p \ll 1$ for all $p \in \mathcal{P}$, we bound

$$S_{p,q} \ll_\mu \frac{1}{L^{n-1}} \cdot |H(z, pq^{n-1}, N)|.$$

Note that we have made this upper bound larger by forgetting the structure of the double coset and simply retaining the information about the determinant, which is an invariant of the double coset. Analogously we obtain

$$S_{p,p} \ll_\mu \frac{1}{L^{n-1}} \cdot |H(z, p^n, N)| + |H(z, 1, N)|.$$

We now put together the bounds above and observe that non-negativity of $\tilde{f}_\mu$ and of A_ω gives

$$A_{\omega_0}|\phi(z)|^2 \leq \int A_\omega \cdot \tilde{f}_\mu(\mu_\omega)|\phi_\omega(z)|^2 d\omega.$$

Finally, we get a lower bound on A_{ω_0} by Hypothesis (2.1.1). □

For unconditional bounds, one may work with the amplifier given in [BM15, (6.2)]. It uses Hecke operators attached to higher powers of primes for providing an alternative to Hypothesis (2.1.1). In fact, we give the slightly more precise version of this amplifier by including information on the determinantal divisors. Recall that the *j-th determinantal divisor* $\Delta_j(\gamma)$ of an integral matrix γ is equal to the greatest common divisor of all $j \times j$ minors.

Proposition 2.3.4. *With the same notation as in Proposition 2.3.3, we have the unconditional bound*

$$L^{2-\varepsilon}|\phi(z)|^2 \ll_{\mu,\varepsilon} |\mathcal{P}| \cdot |H(z, 1, N)| + \sum_{\nu=1}^{n} \frac{1}{L^{(n-1)\nu}} \sum_{p,q\in\mathcal{P}} |\overline{H}(z, p^\nu, q^{(n-1)\nu}, N)|,$$

where $\overline{H}(z, p^\nu, q^{(n-1)\nu}, N)$ *consists of matrices* $\gamma \in H(z, p^\nu q^{(n-1)\nu}, N)$ *satisfying the additional conditions*

$$\Delta_j(\gamma) = (q^{n-1})^{j-1},$$

for all $1 \leq j \leq n-1$.

Remark 2.3.5. Blomer and Maga only preserve the condition on Δ_1 and Δ_2 (see their definition of $S(m, l)$). These and the additional ones in the proposition above follow directly using the crucial property of the determinantal divisors, namely their invariance under right or left multiplication by elements of $\mathrm{SL}_n(\mathbb{Z})$ (see e.g. [New72, Thm. II.8]). Except for the proof of Proposition 2.6.10, these conditions are not used and we mostly consider the larger set $H(z, m, N)$ for simplicity of notation.

2.4 HIGHER RANK ATKIN-LEHNER OPERATORS

In this section we consider possible generalisations of Atkin-Lehner operators to the spaces $X_n(N)$ for $n > 2$. We consider this to be of independent interest and therefore do a thorough investigation of all cases, regardless of the restrictions imposed in the rest of this paper. In fact, the results in this section motivate these restrictions, as one of the main conclusions is the uniqueness of the generalised Fricke involution among the potential symmetries of $X_n(N)$ considered here for $n > 2$.

2.4.1 The normaliser of the Hecke congruence subgroup

In the theory of automorphic forms on $\mathrm{SL}_2(\mathbb{R})$, an Atkin-Lehner operator S is an involution on space of left-$\Gamma_0(N)$ invariant functions. It is obtained by setting $Sf(z) = f(gz)$ for all $z \in \mathbb{H}$, where g lies in the normaliser of $\Gamma_0(N)$ inside $\mathrm{SL}_2(\mathbb{R})$. This is a natural method of producing automorphisms, since the invariance of $f(z)$ under a group Γ is equivalent to the invariance of $f(gz)$ under $g^{-1}\Gamma g$. The normaliser has been computed by Atkin and Lehner in [AL70] and an example of a non-trivial normalising element is

$$g = \begin{pmatrix} & -1 \\ N & \end{pmatrix},$$

which induces the so-called *Fricke involution*. In fact, the normaliser gives all automorphism of the modular curve $X_2(N)$, in more standard notation $X_0(N)$, for all N up to finitely many exceptions (see [KM88]).

Thus, searching for symmetries of automorphic forms in higher rank should involve computing the normalisers of $\Gamma_0(N) \leq \mathrm{SL}_n(\mathbb{R})$ for $n > 2$. Unfortunately, this method can only produce the identity operator, since we prove below that these normalisers, in contrast to the case $n = 2$, are trivial. In the following we denote by $\mathrm{GL}_n^+(\mathbb{Q})$ the subgroup of invertible matrices with positive determinant.

Theorem 2.4. *For $n > 2$, the normaliser of $\Gamma_0(N)$ inside $\mathrm{GL}_n^+(\mathbb{Q})$ is trivial, that is, equal to $\mathbb{Q}_{>0} \cdot \Gamma_0(N)$.*

For simplicity and clarity of the argument, since we work with some explicit coordinates, we prove the theorem in the case of $n = 3$. The way to generalise the proof should be apparent to the reader.

Consider the *left* action of $G := \mathrm{GL}_3^+(\mathbb{Q})$ on full $\mathbb{Z}$-lattices in $\mathbb{R}^3$ (using column vectors).[2] Let $L_1 = \langle e_1, e_2, e_3 \rangle$ be the standard lattice for a basis (e_1, e_2, e_3) of $\mathbb{R}^3$

[2] As opposed to the rest of the present paper, in this independent section we let G act from the left on vectors. This allows for some simplifications of the arguments. In fact, from the point of view of lattices, this is the more natural setting for $\Gamma_0(N)$. For instance, when N is prime, it is easier to see that $\mathrm{SL}_n(\mathbb{R})/\Gamma_0(N)$ parametrises pairs of unimodular lattices together with a sublattice of index N. On the other hand, in the theory of automorphic forms, the dual picture is more standard.

and consider $\mathcal{L} = G \cdot L_1$, the orbit of L_1 under the action of G.

Note that the stabiliser of L_1 under this action is the group $\mathrm{SL}_3(\mathbb{Z})$. More generally, for $M \in \mathbb{N}$, let $L_M = \langle e_1, e_2, Me_3 \rangle$, or in other words,

$$L_M = \begin{pmatrix} 1 & & \\ & 1 & \\ & & M \end{pmatrix} \cdot L_1.$$

If we let $A_M = \mathrm{diag}(1, 1, M)$, then the stabiliser of L_M is

$$\mathrm{Stab}(L_M) = A_M \,\mathrm{Stab}(L_1) A_M^{-1} = \left\{ \begin{pmatrix} a_{11} & a_{12} & \frac{a_{13}}{M} \\ a_{21} & a_{22} & \frac{a_{23}}{M} \\ Ma_{31} & Ma_{32} & a_{33} \end{pmatrix} : (a_{ij}) \in \mathrm{SL}_3(\mathbb{Z}) \right\}.$$

It follows that $\mathrm{Stab}(L_1) \cap \mathrm{Stab}(L_M) = \Gamma_0(M)$. Since $\Gamma_0(N) \subset \Gamma_0(M)$ for all $M \mid N$, we also have that

$$\bigcap_{M|N} \mathrm{Stab}(L_M) = \Gamma_0(N).$$

The following lemma provides a converse for this observation.

Lemma 2.4.1. *The set of lattices fixed by $\Gamma_0(N)$ is*

$$\bigcup_{M|N} \{qL_M : q \in \mathbb{Q}_{>0}\}.$$

Proof. Let $L = g \cdot L_1 \in \mathcal{L}$, where $g \in \mathrm{GL}_3^+(\mathbb{Q})$, and assume that $\Gamma_0(N)$ fixes L. Then $g^{-1}\Gamma_0(N)g$ fixes L_1, so we must have $g^{-1}\Gamma_0(N)g \subset \mathrm{SL}_3(\mathbb{Z})$.

Scaling g by a positive rational number, we may assume that $g \in \mathcal{M}_{3\times 3}(\mathbb{Z})$. Let then H be the Hermite normal form of g, so that

$$H = gU,$$

with $U \in \mathrm{SL}_3(\mathbb{Z})$ and H lower triangular. We have $HL_1 = gUL_1 = gL_1 = L$. So we may further assume that $g = H$ and is thus lower triangular. More explicitly, write

$$H = \begin{pmatrix} \alpha_1 & 0 & 0 \\ \beta_1 & \beta_2 & 0 \\ \gamma_1 & \gamma_2 & \gamma_3 \end{pmatrix} \in \mathcal{M}_{3\times 3}(\mathbb{Z}).$$

We test the inclusion $H^{-1}\xi H \in \mathrm{SL}_3(\mathbb{Z})$ with various matrices $\xi \in \Gamma_0(N)$. Observe that

$$H^{-1}\begin{pmatrix}1 & 1 & \\ & 1 & \\ & & 1\end{pmatrix}H \in \mathrm{SL}_3(\mathbb{Z}) \quad \text{implies that} \quad \frac{\beta_1}{\alpha_1}, \frac{\beta_2}{\alpha_1}, \frac{\beta_1\gamma_2 - \gamma_1\beta_2}{\alpha_1\gamma_3} \in \mathbb{Z};$$

$$H^{-1}\begin{pmatrix}1 & & 1 \\ & 1 & \\ & & 1\end{pmatrix}H \in \mathrm{SL}_3(\mathbb{Z}) \quad \text{implies that} \quad \frac{\gamma_1}{\alpha_1}, \frac{\gamma_2}{\alpha_1}, \frac{\gamma_3}{\alpha_1} \in \mathbb{Z};$$

$$H^{-1}\begin{pmatrix}1 & & \\ 1 & 1 & \\ & & 1\end{pmatrix}H \in \mathrm{SL}_3(\mathbb{Z}) \quad \text{implies that} \quad \frac{\alpha_1}{\beta_2}, \frac{\alpha_1}{\beta_2}\cdot\frac{\gamma_2}{\gamma_3} \in \mathbb{Z};$$

$$H^{-1}\begin{pmatrix}1 & & \\ & 1 & \\ N & & 1\end{pmatrix}H \in \mathrm{SL}_3(\mathbb{Z}) \quad \text{implies that} \quad N\frac{\alpha_1}{\gamma_3} \in \mathbb{Z}.$$

Since $\frac{\beta_2}{\alpha_1}, \frac{\alpha_1}{\beta_2} \in \mathbb{Z}$, we must have $\frac{\beta_2}{\alpha_1} = \pm 1$. Since $\frac{\gamma_3}{\alpha_1}, N\frac{\alpha_1}{\gamma_3} \in \mathbb{Z}$, we must have $\frac{\gamma_3}{\alpha_1} = \pm M$, where $M \mid N$. Using the rest of the findings above, we may do column manipulations and obtain

$$H = \alpha_1\begin{pmatrix}1 & 0 & 0 \\ \frac{\beta_1}{\alpha_1} & \frac{\beta_2}{\alpha_1} & 0 \\ \frac{\gamma_1}{\alpha_1} & \frac{\gamma_2}{\alpha_1} & \frac{\gamma_3}{\alpha_1}\end{pmatrix} = \alpha_1\begin{pmatrix}1 & & \\ & 1 & \\ & & M\end{pmatrix}U',$$

with $U' \in \mathrm{SL}_3(\mathbb{Z})$. Thus $L = HL_1 = L_M$ up to $\mathbb{Q}_{>0}$ scalars. □

Proof of Theorem 2.4. Let $g \in \mathrm{GL}_3^+(\mathbb{Q})$ such that $g^{-1}\Gamma_0(N)g = \Gamma_0(N)$. Since $\Gamma_0(N)$ fixes the lattices L_M for all divisors M of N, we find that $\Gamma_0(N)$ must also fix the lattices gL_M for $M \mid N$. By the previous lemma, for each divisor M of N there is a rational number q_M and a divisor $f(M) \mid N$ such that

$$gL_M = q_M L_{f(M)}$$

for all $M \mid N$.

By the definition of L_M and using the fact that $\mathrm{Stab}(L_1) = \mathrm{SL}_3(\mathbb{Z})$, we can deduce that

$$q_M^{-1}\begin{pmatrix}1 & & \\ & 1 & \\ & & f(M)^{-1}\end{pmatrix}\cdot g \cdot \begin{pmatrix}1 & & \\ & 1 & \\ & & M\end{pmatrix} \in \mathrm{SL}_3(\mathbb{Z}), \tag{2.4.1}$$

for all $M|N$.

Rescaling g by $q_1 \in \mathbb{Q}$ we may assume that $q_1 = 1$. Taking $M = 1$ in (2.4.1) and applying determinants, we deduce that $\det(g) = f(1)$. Applying determinants to all other equations, we find that

$$q_M^3 = \frac{f(1)M}{f(M)}.$$

In particular, for $M = N$, we have $q_N^3 f(N) = Nf(1)$. Since $f(N) \mid N$, we must have $q_N \in \mathbb{Z}$.

Let us make (2.4.1) more explicit. Taking $M = 1$, we have

$$g = \begin{pmatrix} * & * & * \\ * & * & * \\ f(1)* & f(1)* & f(1)* \end{pmatrix},$$

where $*$ denotes unknown *integers*. In particular, the last column of g is integral. If we now take $M = N$, we have

$$g = \begin{pmatrix} q_N* & q_N* & * \\ q_N* & q_N* & * \\ q_N f(N)* & q_N f(N)* & * \end{pmatrix}.$$

Using the properties of the determinant and that $*$ denotes integers, we deduce that $q_N^2 | \det(g) = f(1)$.

Let $f(1) = q_N^2 k$ for some $k \in \mathbb{Z}$. Now the last row of g is divisible by $q_N^2 k$ and the first two columns are divisible by q_N. By the same method we infer that $q_N k \cdot q_N \cdot q_N = q_N^3 k$ divides $\det(g) = f(1) = q_N^2 k$. Therefore $q_N = 1$, which implies that $f(N) = Nf(1)$. Since $f(N) \mid N$, it follows that $f(1) = 1$ and $f(N) = N$. Putting everything together, it follows that $g \in \Gamma_0(N)$. □

Remark 2.4.2. The case $n > 3$ can be done similarly. In essence, what makes the case $n > 2$ differ from $n = 2$ is the imbalance between the number of columns with divisibility conditions and the number of rows with such conditions. This leads to the different exponents of q_N in the proof and ultimately to the triviality of the solutions to our equations.

Theorem 2.3 on the normaliser of $\Gamma_0(N)$ in the real group $\mathrm{PGL}_n(\mathbb{R})$ now follows as a corollary to Theorem 2.4.

Proof of Theorem 2.3. We use the results of [Bor66], which imply that the normaliser of $\Gamma_0(N)$, being commensurable with the arithmetic group $\mathrm{PGL}_n(\mathbb{Z})$, lies in $\mathrm{PGL}_n(\mathbb{Q})$. □

2.4.2 *A different perspective*

We have seen in the last section that $n = 2$ is singular in the sequence of families $\Gamma_0(N) \leq \mathrm{SL}_n(\mathbb{Z})$ of congruence subgroups. To arrive at a general definition of Atkin-Lehner operators, it is useful to note another way in which the group $\mathrm{PGL}_2(\mathbb{R})$ is distinguished, as described below.

An important automorphism of matrices in $\mathrm{SL}_n(\mathbb{R})$ is the map $g \mapsto g^{-T}$, sending a matrix to its inverse transpose. As already noted in the present paper, this map sends a lattice L_g to its dual, but is also used to define the dual form of an automorphic form for $\mathrm{SL}_n(\mathbb{Z})$ (see section 9.2 in [Gol06]) or also the contragredient representation of a $\mathrm{GL}(n)$ automorphic representation.

For PGL(2), dual forms are not commonly mentioned because this automorphism is, in fact, inner in this case. Indeed, if we take

$$w = \begin{pmatrix} & -1 \\ 1 & \end{pmatrix}$$

to be the non-trivial Weyl element, then we easily compute that

$$wg^{-T}w^{-1} = -\frac{1}{\det(g)}g. \tag{2.4.2}$$

Thus, the map $z \mapsto z^{-T}$ induces the identity on $\mathrm{PGL}_2(\mathbb{Z})\backslash \mathrm{PGL}_2(\mathbb{R})/\mathrm{PO}(2)$.

We can artificially introduce the dual map into the theory of Atkin-Lehner operators. For instance, one could write the Fricke involution W_N as

$$W_N f(z) = f\left(\begin{pmatrix} N & -1 \\ & \end{pmatrix} z\right) = f\left(\begin{pmatrix} N & -1 \\ & \end{pmatrix} wz^{-T}w\right) = f\left(\begin{pmatrix} 1 & \\ & N \end{pmatrix} z^{-T}\right).$$

Though slightly cumbersome in rank 1, this approach leads to the right definition of Fricke involutions for $n > 2$.

Let $g \in \mathrm{GL}_n(\mathbb{R})$ such that

$$g^{-1}\Gamma_0(N)g = \Gamma_0(N)^T. \tag{2.4.3}$$

Then the map $f(z) \mapsto f(gz^{-T})$ is an operator on the space of automorphic forms for $\Gamma_0(N)$, which we call by definition an *Atkin-Lehner operator*. As in the previous example, all Atkin-Lehner operators for $n = 2$ can be interpreted as above. More precisely, taking a matrix in the normaliser of $\Gamma_0(N) \leq \mathrm{SL}_2(\mathbb{R})$ and multiplying from the right by the non-trivial Weyl element gives a matrix g satisfying (2.4.3).

We now provide an example of Atkin-Lehner operators for all n. The author was informed that Gergely Harcos has also, independently, found an example in the case $n = 3$.

Definition 1. Let

$$A_N = N^{-1/n}\,\mathrm{diag}(1, \dots, 1, N)$$

and define the *Fricke involution* $W_N : L^2(X_n(N)) \longrightarrow L^2(X_n(N))$ as

$$W_N\phi(z) = \phi(A_N \cdot z^{-T}).$$

We often also refer to the Fricke involution at the group level and denote

$$z' := A_N \cdot z^{-T}.$$

It is easy to check that A_N satisfies (2.4.3). The operator defined above is obviously an involution and the expected properties hold.

Lemma 2.4.3. *The Fricke involution W_N preserves the space of cuspidal newforms and is self-adjoint. If T_g is the Hecke operator associated to the coset $\Gamma_0(N)g\Gamma_0(N)$, where $(\det(g), N) = 1$, then*

$$T_g W_N = W_N T_g^*.$$

If an automorphic form ϕ has spectral parameters $(\mu_1, \dots, \mu_n)$, then $W_N\phi$ has parameters $(-\mu_n, \dots, -\mu_1)$.

Proof. We first prove that $T_g W_N = W_N T_g^*$. By a variant of the Smith normal form, we may assume that g is diagonal and by a variant of the transposition anti-automorphism for $\Gamma_0(N)$ (generalising Lemma 4.5.2 and Theorem 4.5.3 in [Miy89], we may assume that there are matrices α_i, $i = 1, \dots, k$, for some k, such that

$$\Gamma_0(N)g\Gamma_0(N) = \bigcup_i \Gamma_0(N)\alpha_i = \bigcup_i \alpha_i\Gamma_0(N).$$

Then by definition we have

$$T_g W_N f(z) = \sum_i W_N f(\alpha_i z) = \sum_i f(A_N \cdot \alpha_i^{-T} z^{-T}) = \sum_i f(\beta_i \cdot A_N \cdot z^{-T}) = W_N \sum_i f(\beta_i z), \quad (2.4.4)$$

where $\beta_i = A_N \alpha_i^{-T} A_N^{-1}$. The proof is finished by showing that $\bigcup_i \Gamma_0(N)\beta_i = \Gamma_0(N)g^{-1}\Gamma_0(N)$, since this double coset corresponds to T_g^* (s. [Gol06, Thm. 9.6.3]. Indeed,

$$\begin{aligned}
\bigcup_i \Gamma_0(N)\beta_i &= \bigcup_i \Gamma_0(N) W_N \alpha_i^{-T} A_N^{-1} \\
&= \bigcup_i A_N \Gamma_0(N)^T A_N^{-1} A_N \alpha_i^{-T} A_N^{-1} \\
&= A_N \left[\bigcup_i \Gamma_0(N)\alpha_i \right]^{-T} A_N^{-1} \\
&= A_N \Gamma_0(N)^T g^{-1} \Gamma_0(N)^T A_N^{-1} \\
&= \Gamma_0(N) g^{-1} \Gamma_0(N).
\end{aligned}$$

Here we made use of fundamental property (2.4.3) of A_N and of the fact that g is diagonal, thus commuting with A_N.

Next, we prove that W_N is self-adjoint. This can easily be seen by using a known fact about the dual forms for $\mathrm{SL}_n(\mathbb{Z})$. Namely, the map $f(z) \mapsto f(wz^{-T}w^{-1})$, where w is the long Weyl element, is self-adjoint (one can compute directly in explicit coordinates given in [Gol06], Proposition 9.2.1 or Proposition 6.3.1). We can interpret the Fricke involution as

$$W_N f(z) = f(mwz^{-T}w^{-1}),$$

where $m = A_N w^{-1}$, that is, as the composition of the dualising map above with the left-action of m. Since the measure on $\mathbb{H}^n$ is $\mathrm{GL}_n(\mathbb{R})$-invariant, we can make the same explicit computations and change of coordinates as for the dualising map. Since A_N is diagonal, we easily deduce the conclusion $W_N^* = W_N$. Moreover, this interpretation of the Fricke involution and [BHM20, (45)] also prove the statement about the spectral parameters of ϕ.

To prove cuspidality it is best to work adelically, though this can be reduced again to noting the relation between W_N and the dualising map. Namely, the form $W_N\phi$ generates the contragredient of the representation generated by ϕ, which is known to be cuspidal (see e.g. [Bum97, Prop. 3.3.4]). From this perspective, it is also easy to see that $W_N\phi$ is a newform. In the interest of brevity, we leave out the details of adelisation. □

In this interpretation of Atkin-Lehner operators, the group structure coming from the normaliser is not obvious any more. Indeed, using (2.4.3), we cannot even recover the identity for $n > 2$. Finding an even more general definition proves difficult, since the available types of automorphisms on invertible matrices are scarce.

As explained in [McD78], all automorphisms in the case $n > 2$ are constructed out of inner automorphisms, radial automorphisms, and the inverse-transpose automorphism. Inner automorphisms cannot contribute, since we have proved that the normaliser of $\Gamma_0(N)$ is trivial; radial automorphisms are trivial in our context, since we consider only automorphic forms that are invariant under the centre of $\mathrm{GL}_n(\mathbb{R})$; and the inverse-transpose automorphism is precisely the basis for the definition given in this note.

2.4.3 Uniqueness of the Fricke involution

The theory of Atkin-Lehner operators for $\Gamma_0(N)$ shows some weaknesses already in the well-understood case $n = 2$. Indeed, one can only define Atkin-Lehner operators for divisors M of the level N, such that M and N/M are coprime. More precisely, there are no operators induced by matrices with determinant equal $M \mid N$, such that $(M, N/M) \neq 1$ (see [AL70, p. 138]).

This phenomenon creates difficulties in applications when considering powerful levels, as already noted in the historical context of the sup-norm problem. In the present section, we see that these difficulties only get more problematic in higher rank (see Remark 2.4.5). In fact, the only Atkin-Lehner operator for $n > 2$, according to our definition, is the Fricke involution.

Proposition 2.4.4. *Let $g \in \mathrm{GL}_n^+(\mathbb{Q})$ satisfy $g^{-1}\Gamma_0(N)g = \Gamma_0(N)^T$. Then, after scaling by a suitable rational number, g is integral, the last row and the last column of g are divisible by N, and $\det(g) = N$. Equivalently,*

$$g \in \mathbb{Q}_{>0} \cdot \Gamma_0(N)\,\mathrm{diag}(1, \dots, 1, N).$$

Proof. We apply the same ideas as in the proof of Theorem 2.4. Again the proof is done for $n = 3$, merely for simplicity and clarity. One can check that $\Gamma_0(N)^T$ stabilises the lattices

$$L_{M^{-1}} = \langle e_1, e_2, M^{-1}e_3 \rangle = \mathrm{diag}(1, 1, M^{-1})L_1$$

for all divisors $M \mid N$. It follows that $\Gamma_0(N)$ must stabilise (up to scalars) the lattices $gL_{M^{-1}}$.

By Lemma 2.4.1 determining the fixed points of $\Gamma_0(N)$, we have

$$gL_{M^{-1}} = q_M L_{f(M)},$$

with $f(M) \mid N$. We normalise g by a rational number so that $q_1 = 1$. The equations above imply that

$$g \in q_M \,\mathrm{diag}(1, 1, f(M))\,\mathrm{SL}_3(\mathbb{Z})\,\mathrm{diag}(1, 1, M), \tag{2.4.5}$$

using that the stabiliser of L_1 is $\mathrm{SL}_3(\mathbb{Z})$. Let us take determinants and deduce that

$$\det g = q_M^3 \cdot f(M) \cdot M. \tag{2.4.6}$$

By our assumption, $\det g = f(1)$.

Take $M = N$ in (2.4.6) and note that

$$q_N^{-3} = \frac{f(N)N}{f(1)}.$$

Since $f(1) \mid N$, we deduce that $q_N^{-3} \in \mathbb{Z}$, so $d := q_N^{-1} \in \mathbb{Z}$. Using this notation we have $d^3 f(1) = f(N)N$.

Now we use the matrix equation for $M = 1$ and $M = N$ to find that

$$g = \begin{pmatrix} & & \\ & & \\ f(1)* & f(1)* & f(1)* \end{pmatrix} \quad \text{and} \quad g = \begin{pmatrix} & & \frac{N}{d}* \\ & & \frac{N}{d}* \\ \frac{f(N)}{d}* & \frac{f(N)}{d}* & \frac{f(N)N}{d}*, \end{pmatrix} \tag{2.4.7}$$

where $*$ stands for unknown integers and the rest of the matrices are also filled by integers.

We claim that

$$d \mid N.$$

Indeed, say there is a prime p such that $p^k \mid d$, but $p^k \nmid N$. Then $p^k \nmid f(N)$ since $f(N) \mid N$, and thus $p^{2k} \nmid Nf(N)$. But we know that $d^3 f(1) = Nf(N)$, so we must have $p^{3k} \mid Nf(N)$, which is a contradiction unless $k = 0$.

Now suppose p is a prime dividing d such that $p^k \parallel d$ is the maximal power of p dividing d, with $k \geqslant 1$. As in the last paragraph, it would follow that $p^{3k} \mid f(N)N$ and $p^k \mid N$. Since $f(N) \mid N$, we deduce that p divides N/d.

We now use the divisibility conditions from the right of (2.4.7) for the last column of g and the divisibility conditions from the left of (2.4.7) for the first two entries of the last row of g. Putting everything together we obtain

$$g = \begin{pmatrix} & & p* \\ & & p* \\ f(1)* & f(1)* & f(1)p^2* \end{pmatrix}.$$

It would follow that $f(1) \cdot p \mid \det(g) = f(1)$, but this is a contradiction. Therefore $d = 1$.

We infer that $f(1) = Nf(N)$, so considering divisibility we must have $f(1) = N$ and $f(N) = 1$. This implies that $\det g = N$ and that the last row and column of g are divisible by N.

Thus g is of the form

$$g = \begin{pmatrix} \alpha_1 & \alpha_2 & N\alpha_3 \\ \beta_1 & \beta_2 & N\beta_3 \\ N\gamma_1 & N\gamma_2 & N\gamma_3 \end{pmatrix}$$

with $\alpha_i, \beta_i, \gamma_i \in \mathbb{Z}$. Since $\det(g) = N$, it must be that γ_3 is coprime to N and that $(\alpha_3, \beta_3, \gamma_3) = 1$. In fact, put these together to have $(N\alpha_3, N\beta_3, \gamma_3) = 1$.

Now take $x, y, z \in \mathbb{Z}$ such that

$$xN\alpha_3 + yN\beta_3 + z\gamma_3 = 1.$$

Then $(xN, yN, z) = 1$, so we can find a matrix $u \in \Gamma_0(N)$ with last row equal to (xN, yN, z). It follows from the above that the entry in the lower right corner of $u \cdot g$ is equal to N. By doing row manipulations we can find $u' \in \Gamma_0(N)$ such that

$$u'g = \begin{pmatrix} * & * & 0 \\ * & * & 0 \\ N* & N* & N \end{pmatrix}.$$

In this form, it is obvious that we can find another $u'' \in \Gamma_0(N)$ so that $u''g = \mathrm{diag}(1, 1, N)$. □

Remark 2.4.5. Let us note what changes in the proof in the case $n = 2$ and how this leads to the lack of Atkin-Lehner operators for powerful levels. In the notation above, we would have the equation $d^2 f(1) = f(N)N$, where the exponent of d is equal to n in general. We can still prove that $d \mid N$, yet the next paragraph in the proof differs slightly.

We suppose p is a prime dividing d such that $p^k \parallel d$ is the maximal power of p dividing d, with $k \geq 1$. As in the proof above, we deduce that $p^{2k} \mid f(N)N$ and $p^k \mid N$. If we were to continue the proof as above and deduce that $d = 1$, we would need the step showing that p divides N/d. This is not true in this case any more. For example, if N is square free, then $k \leq 1$ and the claim in

the step may not hold for certain choices of $f(N)$. In fact, solving the matrix equations eventually leads to the matrices found by Atkin and Lehner (after suitably multiplying by the long Weyl element).

If N is powerful, then we could have that a higher power of p divides N. For certain choices of d, we can indeed deduce that $p \mid N/d$ and produce a contradiction. These choices of d correspond to divisors M of N, such that $(M, N/M) \neq 1$. Indeed, suppose that $\det(g) = f(1) =: M$, $p \mid M$ and $p \mid N/M$. Then p divides $d = f(N)N/M$. If $p^k \parallel d$, then applying the p-adic valuation to $d^2M = f(N)N$ and recalling that $f(N) \mid N$ shows that $p \mid N/d$. We proceed as in the proof above and derive a contradiction. This shows that there are no Atkin-Lehner operators for such divisors M as above.

2.5 REDUCTION OF THE DOMAIN

After studying generalised Atkin-Lehner operators, we showcase their main application in this section. More precisely, we study fundamental domains for the action of these operators on $X_n(N)$. Though very natural at a geometric level, we first note how this is relevant to the sup-norm problem.

The value of $\phi(z)$ is independent of which element in the orbit $\Gamma_0(N) \cdot z$ we choose instead of z. Similarly, the number and shape modulo N of the matrices we are considering in the amplified pretrace formula in Proposition 2.3.3 is invariant under shifting by elements of $\Gamma_0(N)$, which would merely amount to conjugating $H(z, m, N)$.

Consider now the action of the Fricke involution $W_N(\phi)(z) = \phi(z')$. If $Y \subset \Gamma_0(N)$ is a subset, we denote by Y' the image of Y under the map $z \mapsto z'$. It is clear that we obtain a bound for a Hecke-Maaß form ϕ on $Y \cup Y'$ if we have a bound for both ϕ and $W_N(\phi)$ on the subset Y.

Recall now that $W_N(\phi)$ has essentially the same properties as ϕ by Lemma 2.4.3. Since the amplifiers, Proposition 2.3.3 and Proposition 2.3.4, and the Fourier bound, Proposition 2.7.1, apply similarly to both forms, we are free to choose any representative in

$$\Gamma_0(N)z \cup \Gamma_0(N)z'$$

when attacking the counting problem.[3]

In this section we propose a system for making this selection of representative. In other words, we construct an approximate fundamental domain for the action of $\Gamma_0(N)$ and the Fricke involution, at least in the bulk. It can be seen as a reduction theory with level structure, for which we often use the shorter term Fricke reduction.

[3]Indeed, the implied constant depending on μ in the amplifier is also of the same size, as the computation of spectral parameters in Lemma 2.4.3 shows.

2.5.1 Two lattices

Throughout the following sections we assume that N is a prime.

Recall that for $z \in \mathrm{SL}_n(\mathbb{R})$ we write

$$z' := A_N z^{-T} = N^{-1/n} \operatorname{diag}(1, \dots, 1, N) z^{-T}.$$

We consider the lattices L_z and $L_{z'}$ in the notation and terminology established in Section 2.2. Note that both lattices have determinant 1. We define the sets

$$A(z) = \{\|e_n\|_{\gamma z} \mid \gamma \in \Gamma_0(N)\},$$
$$B(z) = \{\|e_2 \wedge \cdots \wedge e_n\|_{\gamma z} \mid \gamma \in \Gamma_0(N)\}.$$

In the following paragaphs we show how the union of $A(z)$, $B(z)$, $A(z')$, $B(z')$ provides the lengths of all primitive vectors in L_z, $L_{z'}$, and their duals.

First, we claim that the union of lengths

$$\{\|e_n\|_{\gamma z} \mid \gamma \in \Gamma_0(N)\} \cup \{\|e_2 \wedge \cdots \wedge e_n\|_{\gamma z'} \mid \gamma \in \Gamma_0(N)\}$$

exhausts the lengths of all primitive vectors in L_z. For this we use the fact that any primitive vector in $\mathbb{Z}^n$ is the last row (in fact, any row or any column) of some matrix in $\mathrm{SL}_n(\mathbb{Z})$. Consequently, the vectors $e_n \gamma$ give all primitive vectors in $N\mathbb{Z} \times \cdots \times N\mathbb{Z} \times \mathbb{Z}$ in the lattice L_z.

For the second set, note using Lemma 2.2.1 that

$$\|e_2 \wedge \cdots \wedge e_n\|_{\gamma z'} = \|e_1\|_{\gamma^{-T} A_N^{-1} z} = N^{1/n} \|(a_1, \dots, a_n)\|_z, \tag{2.5.1}$$

where $(a_1, \dots, a_{n-1}, N a_n)$ is the top row of γ^{-T}. We prove in Lemma 2.5.1 below that we obtain this way all primitive vectors $(a_1, \dots, a_n)$ in L_z, for which

$$\gcd(\gcd(a_1, \dots, a_{n-1}), N) = 1.$$

Since N is prime, the greatest common divisor of $\gcd(a_1, \dots, a_{n-1})$ and N can only be 1 or N, and thus, considering the paragraph above we have exhausted all primitive vectors in L_z.

Lemma 2.5.1. *For N prime, if $v = (a_1, \dots, a_{n-1}, N a_n) \in \mathbb{Z}^n$ is a primitive vector, then there is $\gamma \in \Gamma_0(N)$ such that v is the first row of γ^T.*

Proof. Let $g \in \mathrm{SL}_n(\mathbb{Z})$ be any matrix with first row v. Multiplying g from the left by block matrices of the form

$$\begin{pmatrix} 1 & \\ & h \end{pmatrix},$$

where $h \in \mathrm{SL}_{n-1}(\mathbb{Z})$, leaves the first row invariant. We shall inductively apply such row operations on g to make its last column be of the form $(c_1, \dots, c_n)$, where $c_1 = N a_n$ and N divides $c_1, \dots, c_{n-1}$.

Indeed, if $N \mid c_i$ for any $i \in \{2, \dots, n\}$, then we can permute rows to assume that $N \mid c_2$. Otherwise we can assume that $\gcd(N, c_3) = 1$. Let $\bar{c}_3$ be any representative of the inverse of c_3 modulo N. Bézout's lemma provides a matrix $h' \in \mathrm{SL}_2(\mathbb{Z})$ with top row $(N, \bar{c}_3)$. Using h of the form

$$h = \begin{pmatrix} h' & \\ & 1_{n-3} \end{pmatrix},$$

as above, we may now assume that $c_2 \equiv 1$ modulo N. Another transformation of the same type, where h' now has top row $(1, -\bar{c}_3)$, allows us to assume that $N \mid c_2$. We conclude by induction. □

Next, the union of lengths

$$\{\|e_2 \wedge \cdots \wedge e_n\|_{\gamma z} \mid \gamma \in \Gamma_0(N)\} \cup \{\|e_n\|_{\gamma z'} \mid \gamma \in \Gamma_0(N)\}$$

exhausts the lengths of all primitive vectors in $L_z^* = L_{z^{-T}}$. Indeed, Lemma 2.2.1 gives that

$$\|e_2 \wedge \cdots \wedge e_n\|_{\gamma z} = \|e_1\|_{\gamma^{-T} z^{-T}} = \|(a_1, \dots a_{n-1}, N a_n)\|_{z^{-T}}, \tag{2.5.2}$$

where $(a_1, \dots, N a_n)$ is the first row of γ^{-T}. As above, we obtain this way all primitive vectors in $\mathbb{Z}^{n-1} \times N\mathbb{Z}$ in the lattice $L_{z^{-T}}$. Furthermore,

$$\|e_n\|_{\gamma A_N z^{-T}} = N^{1-1/n} \|(a_1, \dots, a_n)\|_{z^{-T}}, \tag{2.5.3}$$

for $(a_1, \dots, a_n)$ primitive with $\gcd(a_n, N) = 1$. Since N is prime, this shows the claim.

The above considerations are collected for an overview in Table 2.1. Each

L_z	$A(z)$	$N^{-1/n} \cdot B(z')$
L_z^*	$N^{-1+1/n} \cdot A(z')$	$B(z)$
$L_{z'}$	$A(z')$	$N^{-1/n} \cdot B(z)$
$L_{z'}^*$	$N^{-1+1/n} \cdot A(z)$	$B(z')$

Table 2.1: Lattices and sets of lengths of primitive vectors.

row corresponds to a lattice and the union of the two sets in that row is the set of the lengths of all primitive vectors in the corresponding lattice. By multiplication of a set by a scalar we mean multiplication of each element in the set by the given scalar. We use here that $z \mapsto z'$ is an involution on unimodular lattices.

2.5.2 *Fricke reduction*

Let us consider minima of the lattices in the previous section. Write

$$\alpha(z) = \min A(z), \quad \beta(z) = \min B(z).$$

As in Table 2.1, the minimal non-zero length in the lattice L_z is found either in $A(z)$, equal in this case to $\alpha(z)$, or in $B(z')$, equal to $N^{-1/n}\beta(z')$.

More generally, let x be any of the letters α or β. Let L be any of the lattices L_z, $L_{z'}$, L_z^*, $L_{z'}^*$. Then the minimal length in L is an x-expression if it is of the form $N^\eta x(w)$, where η is a non-positive number and w is either z or z'. From Table 2.1 and the discussion of that section, we see that there are only two possibilities for each lattice, namely a unique α-expression or a unique β-expression.

Definition 2. Let X and Y denote the *Greek* letters A or B, and analogously for their lowercase variants. We say that $z \in \mathcal{L}(X, Y)$ if the smallest length in L_z is the unique x-expression and the smallest length in L_z^* is the unique y-expression. Similarly, $z \in \mathcal{L}'(X, Y)$ if the smallest lengths in $L_{z'}$ and $L_{z'}^*$ are the x-expression and the y-expression, respectively.

Example 1. If $z \in \mathcal{L}(B, A)$, then the smallest length in L_z is given by $N^{-1/n}\beta(z')$ and the smallest length in L_z^* is given by $N^{-1+1/n}\alpha(z')$.

Let $z \in \mathbb{H}$. For the study of the sup-norm and our counting problem, we are allowed to choose any conjugate of z in the orbit $\Gamma_0(N) \cdot z$ and also switch between z and z', as explained at the beginning of Section 2.5. Now it is clear by construction that every z is contained in some $\mathcal{L}(X, Y)$. We then make the choice of conjugate to obtain a well-positioned z, where we have control over its successive minima and Iwasawa coordinates, based on which set $\mathcal{L}(X, Y)$ contains z.

2.5.2.1 CASE I

Let

$$z \in \bigcup_{X \in \{A,B\}} \mathcal{L}(A, X) \cup \mathcal{L}'(A, X).$$

By switching between z and z' if needed, we can assume that $z \in \mathcal{L}(A, X)$, for some $X \in \{A, B\}$. In this case, the minimal length in L_z is $\alpha(z)$. Shifting z by $\gamma \in \Gamma_0(N)$ if needed, we assume that $\alpha(z) = \|e_n\|_z$. In Iwasawa coordinates $z = n(x) \cdot a(y)$ as in Section 2.2.3, we have $\alpha(z) = d$.

Let γ be of the form

$$\gamma = \begin{pmatrix} h & \\ & 1 \end{pmatrix} \in \Gamma_0(N),$$

where $h \in \mathrm{SL}_{n-1}(\mathbb{Z})$. Note that $e_n \cdot \gamma = e_n$, so we can make the same assumptions about γz as about z above. As in Remark 2.2.4, shifting by γ as above if needed, we may now additionally assume that $z = n(x)a(y)$ satisfies $y_i \geq \sqrt{3}/2$ for $i = 2, \ldots, n-1$.

By Lemma 2.2.2, if λ_1 and λ_2 are the first two successive minima of L_z, then the shortest length l in $\bigwedge^2 L_z$ satisfies

$$l \asymp_n \lambda_1 \cdot \lambda_2.$$

In particular, $l \gg \lambda_1^2$. This implies that

$$d^2 y_1 = \|e_{n-1} \wedge e_n\|_z \gg \alpha(z)^2 = d^2.$$

We deduce that $y_1 \gg_n 1$.[4]

2.5.2.2 CASE II

Let

$$z \in [\mathcal{L}(B,B) \cap \mathcal{L}'(B,A)] \cup [\mathcal{L}(B,A) \cap \mathcal{L}'(B,B)] \cup [\mathcal{L}(B,B) \cap \mathcal{L}'(B,B)].$$

Applying the Fricke involution if needed, we can assume that z lies in $\mathcal{L}'(B,B)$ and in $\mathcal{L}(B,*)$. Then the minimal length in $L_{z'}^*$, is given by $b(z')$ and the minimal length in L_z is $N^{-1/n} b(z')$.

By Minkowski's theorem, more precisely equation (2.2.2), applied to $L_{z'}^*$, we find that $b(z') \ll_n 1$. This now implies that the minimal length in L_z is $N^{-1/n} b(z') \ll N^{-1/n}$.

2.5.2.3 CASE III

Let

$$z \in \mathcal{L}(B,A) \cap \mathcal{L}'(B,A).$$

Applying the Fricke involution if required, we may assume that $\alpha(z') \leq \alpha(z)$. Shifting $z = n(x)a(y)$ by a suitable $\gamma \in \Gamma_0(N)$ as in Case I, i.e. Section 2.5.2.1, we also assume that $\alpha(z) = \|e_n\|_z = d$ and that $y_i \gg 1$ for $i = 2, \ldots, n-1$.

Note now that the minimal lenght in $L_{z'}^*$ is $N^{-1+1/n}\alpha(z)$. Note also that $L_{z'}^*$ is the lattice corresponding to

$$z'^{-T} = A_N^{-1} z.$$

We now compute that

$$\|e_{n-1} \wedge e_n\|_{z'^{-T}} = \|e_{n-1} \wedge e_n\|_{A_N^{-1} z} = N^{-1+2/n} d^2 y_1.$$

Using Lemma 2.2.2, we deduce that the minimal lenght l in $\bigwedge^2 L_{z'}^*$ satisfies $l \gg \mu_1^2$, where μ_1 is the first successive minimum of $L_{z'}^*$. Putting everything together we arrive at

$$N^{-1+2/n} d^2 y_1 \gg N^{-2+2/n} d^2,$$

which implies that $y_1 \gg_n N^{-1}$.

[4] This can be viewed as a soft version of Hermite reduction, that is, reduction to a Siegel set. Indeed, here we also take the last row to be the shortest vector and then use induction, as in the classical proof of reduction.

2.5.2.4 FRICKE REDUCTION OF POINTS THAT REDUCE TO A COMPACTUM

We summarise the cases described above in the context of points z that reduce to a fixed compact set $\Omega \subset \mathbb{H}$.

Proposition 2.5.2. *Let $z \in \mathbb{H}$ and let $\Omega \subset \mathbb{H}$ be a compact set. For $N \gg_\Omega 1$ prime, large enough, there is*

$$w \in \{\gamma z \mid \gamma \in \Gamma_0(N)\} \cup \{\gamma z' \mid \gamma \in \Gamma_0(N)\},$$

*where $z' = A_N z^{-T}$, with Iwasawa coordinates $w = n(x)a(y)$ which, if z reduces to Ω, satisfy **either***

$$y_i \asymp_\Omega 1$$

for all $i = 1, \dots, n-1$, in which case

$$w \in \bigcup_{X \in \{A,B\}} \mathcal{L}(A, X) \cup \mathcal{L}'(A, X)$$

or

$$y_1 \asymp_\Omega \frac{1}{N} \quad \textit{and} \quad y_i \asymp_\Omega 1$$

for $i = 2, \dots, n-1$, in which case

$$w \in \mathcal{L}(B, A) \cap \mathcal{L}'(B, A).$$

Proof. By Lemma 2.2.7, we eliminate Case II, since there the minimal length in L_z is $\ll N^{-1/n}$.

In Case I we find w as in the statement such that $y_i \gg_n 1$ for all $i = 1, \dots, n-1$. Thus w lies in a Siegel set and Lemma 2.2.5 together with Lemma 2.2.7 implies that $y_i \asymp_{n,\Omega} 1$ for all i.

In Case III we find w such that $\alpha(w') \leq \alpha(w)$, $y_1 \gg_n N^{-1}$, and $y_i \gg 1$ for $i = 2, \dots, n-1$. Since $w \in \mathcal{L}(B, A)$, the minimal length in L_w^* is $N^{-1+1/n}\alpha(w')$ and by Lemma 2.2.7 we deduce that $\alpha(w') \gg_\Omega N^{1-1/n}$. Since $\alpha(w') \leq \alpha(w)$, we also have that $\alpha(w) \gg N^{1-1/n}$.

Now $\alpha(w) = \|e_n\|_z = d$. Writing out the definition of d, we see that

$$d^{-n} = y_1^{n-1} \cdots y_{n-1} \ll N^{-(n-1)}.$$

Combining this with the bounds above for the y-coordinates, we deduce that $y_1 \asymp N^{-1}$ and $y_i \asymp 1$ for $i = 2, \dots, n-1$, where the implicit constants depend on Ω. □

2.6 COUNTING MATRICES

2.6.1 An overview

When applying the amplified pretrace formula, e.g. Proposition 2.3.4, we arrive at the problem of counting matrices in $H(z, m, N)$. We give a brief overview of

the counting strategy in the simplest case of $n = 2$. The perspective taken in this paper is new even in this case. We recall some ideas already introduced in Section 2.1.3.1.

Let $z \in \mathrm{SL}_2(\mathbb{R})$, for which we assume the Iwasawa form

$$z = \begin{pmatrix} \sqrt{y} & x/\sqrt{y} \\ 0 & 1/\sqrt{y} \end{pmatrix}$$

and let $\gamma \in H(z, m, N)$. The bound

$$z^{-1}\gamma z = O(m^{1/2}) \tag{2.6.1}$$

implies the conditions

$$e_i \cdot \gamma z \in B(m^{1/2} \|e_i \cdot z\|)$$

for $i = 1, 2$, where $B(r)$ is a Euclidean ball of radius $O(r)$ around 0.

We assume now that z lies in what we call the balanced bulk, as in the second alternative in Proposition 2.5.2, meaning that z reduces to some compact Ω and $y \asymp 1/N$. Let

$$z_N = \mathrm{diag}(N, 1) \cdot z,$$

which defines a sublattice of index N of L_z. We prove in Lemma 2.6.1 that the lattices defined by z, z', and z_N are balanced. That is to say that their respective successive minima and covolume satisfy

$$\lambda_1 \asymp \lambda_2 \asymp \sqrt{\mathrm{vol}}.$$

Notice also that

$$\|e_2\|_z = \left\|(0, 1/\sqrt{y})\right\| \asymp \sqrt{N} \asymp \sqrt{\mathrm{vol}(L_{z_N})},$$

making $e_2 \cdot z$ one of the shortest vectors in L_{z_N}.

This is helpful since we now count the possibilities for $e_2 \cdot \gamma$, a vector in the sublattice $N\mathbb{Z} \times \mathbb{Z}$. We do this by applying Lemma 2.2.3, which counts lattice points in balls. Since z_N is balanced, the bound we obtain is roughly the volume of the ball $B(m^{1/2} \|e_2\|_z)$ divided by the covolume of the lattice z_N. This gives $\ll m$ possibilities.

For $e_1 \cdot \gamma$, we notice that $\|e_1\|_z$ is equal to $y + x/y \asymp 1/N + Nx$. Unfortunately, if z is a balanced lattice, one can compute that we must have a bound $x \gg 1/\sqrt{N}$. Thus the norm above can be rather large. Even though L_z is balanced, the size of the ball would give a hopelessly large bound.

Fortunately, we notice that

$$e_1 \cdot z - x e_2 \cdot z = (\sqrt{y}, x/\sqrt{y}) - x(0, 1/\sqrt{y}) = (\sqrt{y}, 0),$$

by the Iwasawa decomposition or the Gram-Schmidt process. The conditions above can be combined to show that

$$e_1 \cdot \gamma z - x e_2 \cdot \gamma z \in B(m^{1/2} \left\|(\sqrt{y}, 0)\right\|).$$

Since $y \asymp 1/N$, we see that if $m \ll N^{1-\varepsilon}$, the ball we obtain has a small radius of size $o(1)$. Since L_z is a balanced lattice, we can only have at most one lattice point in such a small ball, regardless of its centre. For every vector $e_2 \cdot \gamma$ fixed as above, this leaves at most one possibility for $e_1 \cdot \gamma$. Therefore, the second row $e_2 \cdot \gamma$ already fixes the whole matrix γ.

This strategy gives a bound

$$\#\bigcup_{l=1}^{m} H(z, l, N) \ll m$$

if m is small enough in terms of N. A glance at Proposition 2.3.4 shows that this bound is insufficient to obtain a saving when averaging over square determinants $l = p^2q^2$ and thus $m = L^4$, in the notation of the proposition.

To refine the process above, we only partially fix the second row of γ. This seems difficult to do in standard coordinates, that is, working with the exact entries of γ. Instead, we choose a reduced basis, v_1 and v_2, for the balanced lattice L_{z_N}. An upshot of Fricke reduction is that we can choose $v_2 = e_2 \cdot z$ (we already noticed above that $e_2 \cdot z$ is a shortest vector in L_{z_N}).

We now write $e_2 \cdot \gamma z \in L_{z_N}$ in coordinates using v_1 and v_2. By our conditions and the balancedness of the lattice, the coefficients for both basis vectors are bounded by $\sqrt{m}$. In a first step, we only choose the coefficient of v_1, giving us $\sqrt{m}$ possibilities.

We now ask how many matrices γ have such a coefficient. For two such matrices γ_1, γ_2, the difference $\gamma_1 - \gamma_2$ would have last row equal to $c \cdot e_2$ with $c \ll \sqrt{m}$. It would also satisfy (2.6.1). These two observations imply that the strategy above applies to this difference. The principle that the last row fixes the matrix now gives that $\gamma_1 - \gamma_2 = c \cdot \mathrm{id}_2$.

Applying the determinant to $\gamma_1 = \gamma_2 + c \cdot \mathrm{id}_2$ and assuming that γ_1 has a square determinant imply that $-c$ gives a solution to

$$\chi_{\gamma_2}(X) = Y^2.$$

We employ a theorem of Heath-Brown to count solutions to such equations and obtain adequate bounds for the amplified pretrace formula in the non-degenerate case.

The degenerate case is precisely when the characteristic polynomial of γ_2 is a square. This means that γ_2 is a parabolic matrix and therefore fixes a cusp. For $\Gamma_0(N)$ with N prime, there are two such cusps and these are conjugated by the Fricke involution. This allows us to assume that γ_2 fixes the cusp at infinity and is therefore an upper triangular matrix, up to conjugation. The

strategy above can be adapted slightly for us to apply, again, the principle that the last row determines the matrix. In this case, the last row is the same as that of a multiple of the identity matrix and we are done.

2.6.2 The iterative strategy

In this section we generalise the process described above for $n = 2$.

Let $z = n(x)a(y) \in \mathrm{SL}_n(\mathbb{R})$ be a matrix in Iwasawa form. Let $\gamma \in \mathcal{M}_n(\mathbb{Z}, N)$ with $\det \gamma = m$ and

$$z^{-1}\gamma z = O(m^{1/n}).$$

We can now multiply the previous equation with its transpose and obtain

$$z^{-1} \cdot \gamma \cdot z \cdot z^T \cdot \gamma^T \cdot z^{-T} = O(m^{2/n}). \tag{2.6.2}$$

Notice now that $\gamma \cdot z \cdot z^T \cdot \gamma^T$ is the Gram matrix of the rows of γ with respect to the scalar product defined by z.

Denote the rows of γ by $\gamma_1, \dots, \gamma_n$, and denote the rows of $n(x)^{-1}\gamma$ by $v_1, \dots, v_n$. We compute that

$$z^{-1}\gamma z z^T \gamma^T z^{-T} = \begin{pmatrix} \|v_1\|_z^2 \cdot d_1^{-2} & \langle v_1, v_2\rangle_z \cdot (d_1 d_2)^{-1} & \dots & \langle v_1, v_n\rangle_z \cdot (d_1 d_n)^{-1} \\ * & \|v_2\|_z^2 \cdot d_2^{-2} & \dots & \langle v_2, v_n\rangle_z \cdot (d_2 d_n)^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \dots & \|v_n\|_z^2 \cdot d_n^{-2} \end{pmatrix},$$

where the matrix should be completed by noting that it is symmetric. Observe now that the condition (2.6.2) reduces to

$$\|v_i\|_z \ll m^{1/n} \cdot d_i, \tag{2.6.3}$$

for all $i = 1, \dots, n$, since the off-diagonal conditions simply follow by the Cauchy-Schwarz inequality.

The strategy for counting the number of matrices γ is to iteratively count the number of possibilities for its rows. More precisely, we first count the number of possible $\gamma_n = v_n$ by a lattice point counting argument, that is Lemma 2.2.3, since $\gamma_n \in \mathbb{Z}^n$. For each such fixed possibility, we then count the number of possible γ_{n-1} by using the condition on v_{n-1} in (2.6.3). For this observe that

$$v_{n-1} = \gamma_{n-1} - \xi \cdot \gamma_n,$$

where $\xi \in \mathbb{R}$ can be computed from the x-coordinates of z (in fact, $\xi = x_{n-1,n}$). Thus, having fixed γ_n, the condition can be interpreted as saying that γ_{n-1} is a lattice point inside a ball with shifted centre. We can use that the bounds in 2.2.3 are independent of the centre of the ball. In the results below, we

ultimately choose m small enough so that the ball can only contain one lattice point.

We continue this process iteratively, using that $n(x)^{-1}$ is upper triangular unipotent. We bound the number of γ by multiplying together the number of possibilities for each row. As before, we only used the inequality $\det(\gamma) \leq m$ and therefore we cannot detect, at this point, the sparseness of the sequence of determinants. This latter issue only shows up when using the unconditional amplifier and is dealt with in the next section.

To get the point z into a good position for applying the strategy above, we make the reduction given by Proposition 2.5.2 and assume the second alternative in the statement. In this case, we study the properties of all lattices derived from z relevant for this and the next sections. For the other alternative we use a bound derived from the Whittaker expansion, for which we refer to Section 2.7.1.

Definition 3. For any $z \in \mathrm{SL}_n(\mathbb{R})$ define

$$z_N = \mathrm{diag}(N, \dots, N, 1) \cdot z.$$

Lemma 2.6.1. *Let N be a prime and let $z \in \mathbb{H}$ reduce to a compactum Ω. Assume that z has Iwasawa coordinates*

$$y_1 \asymp_\Omega \frac{1}{N} \quad \textit{and} \quad y_i \asymp_\Omega 1$$

for $i = 2, \dots, n-1$, with $d = \alpha(z) = \|e_n\|_z$ and satisfies

$$z \in \mathcal{L}(B, A) \cap \mathcal{L}'(B, A).$$

Then the successive minima of z and z' are all $\asymp_\Omega 1$ and the successive minima of z_N and $(z')_N$ are all $\asymp_\Omega N^{(n-1)/n}$.

Proof. Throughout this proof all implied constants are allowed to depend on Ω and, implicitly, n. We call a lattice L *balanced* if $\lambda_1 \asymp d(L)^{1/n}$, where $\lambda_1 \leq \dots \leq \lambda_n$ are the successive minima of L. By Minkowski's theorem (2.2.1), generalising Lemma 2.2.7, this is equivalent to $\lambda_i \asymp d(L)^{1/n}$ for all $i = 1, \dots, n$. This, together with Lemma 2.2.1 on the dual lattice and Lemma 2.2.2 on the successive minima of exterior products, implies that L is balanced if and only if the dual L^* is balanced. Note also that the property of being balanced is invariant under scaling.

Computing the determinants, we thus aim to prove that $L_z, L_{z'}, L_{z_N}, L_{(z')_N}$ are balanced lattices. That L_z is balanced is part of the assumption (see again Lemma 2.2.7). We also compute from the Iwasawa coordinates and the fact that $\det(z) = 1$ that

$$d^n \asymp N^{n-1} = \det(z_N).$$

Now since $z \in \mathcal{L}'(B, A)$, Table 2.1 shows that the first successive minimum of $L_{z'}^*$ is equal to $N^{-1+1/n}\alpha(z) \asymp 1$. Therefore $L_{z'}^*$ is balanced and so is $L_{z'}$.

Next, compute explicitly that

$$(z_N)^{-T} = N^{-1}\,\mathrm{diag}(1,\dots,1,N)z^{-T} = N^{-1+1/n}z'.$$

By the above, it follows that $L^*_{z_N}$ is balanced and so is L_{z_N}.

We finally note that

$$(z')_N = N^{1-1/n}z^{-T}$$

so the same reasoning implies that $L_{(z')_N}$ is balanced. □

The following is the main and simplest counting result of this paper and implements the strategy discussed above.

Proposition 2.6.2. *Let N be a prime and let $z \in \mathbb{H}$ reduce to a compactum Ω. Assume that z has Iwasawa coordinates*

$$y_1 \asymp_\Omega \frac{1}{N} \quad \textit{and} \quad y_i \asymp_\Omega 1$$

for $i = 2,\dots,n-1$, and $d = \alpha(z) = \|e_n\|_z$, and satisfies

$$z \in \mathcal{L}(B,A) \cap \mathcal{L}'(B,A).$$

Then

$$|\{\gamma \in \mathcal{M}_n(\mathbb{Z},N) \mid \det(\gamma) \ll \Lambda^n,\ z^{-1}\gamma z = O(\Lambda)\}| \ll_{n,\Omega} \Lambda^n(1+\Lambda^n/N)^{n-1}.$$

Proof. The bottom row $e_n \cdot z$ has congruence conditions and thus lies in the lattice corresponding to z_N. By Lemma 2.6.1, this is a balanced lattice, meaning that we can approximate all successive minima of L_{z_N} by $\det(z_N)^{1/n} = N^{(n-1)/n} \asymp d$. In fact, the proof of Lemma 2.6.1 shows that the minimum of z_N is equal to the minimum of $N^{1-1/n}(z')^{-T}$, which is $\alpha(z) = d$. Thus $e_n z_N$ is a vector of shortest length in L_{z_N}.

Recall now the condition

$$\|\gamma_n\|_z \ll \Lambda d_n = \Lambda d$$

from (2.6.3). By Lemma 2.2.3, there are at most

$$1 + \frac{\Lambda d}{d} + \frac{(\Lambda d)^2}{d^2} + \cdots + \frac{(\Lambda d)^n}{d^n} \ll_n \Lambda^n$$

possibilities for the row $\gamma_n = e_n \cdot \gamma$.

We continue bounding the number of possibilities for γ_i inductively, $i < n$. More precisely, we suppose that γ_j with $i < j \le n$ are fixed. Then, by using the fact that $n(x)^{-1}$ is unipotent upper triangular in condition (2.6.3), the number of possibilities left for γ_i is bounded by the number of lattice points in L_z in a ball of radius $L \cdot d_i$ with fixed centre determined by the γ_j, $i < j$, and $n(x)$.

Next, note that the successive minima of L_z are all $\asymp_{n,\Omega} 1$, by Lemma 2.2.7. Furthermore,

$$d_i^n = (dy_1 \cdots y_{n-i})^n \asymp_{n,\Omega} 1/N.$$

By Lemma 2.2.3, there are at most

$$\ll_{n,\Omega} 1 + \Lambda d_i + \cdots + (\Lambda d_i)^n \ll_{n,\Omega} 1 + \frac{\Lambda^n}{N}$$

possibilities for γ_i.

Putting all bounds together, we bound the number of matrices γ by

$$\ll_{n,\Omega} \Lambda^n (1 + \Lambda^n/N)^{n-1}.$$

□

Remark 2.6.3. The last part of the proof above shows that, as long as Λ is small enough in terms of N, the choice of last row of γ already determines the whole matrix.[5]

2.6.3 Detecting determinants that are higher powers

The bound supplied by Proposition 2.6.2 is too weak to suffice in the unconditional amplifier, Proposition 2.3.4, where powers $\nu > 1$ show up and introduce sparseness into the average. Taking Remark 2.6.3 into consideration, we see that the approach in the previous section is over-counting the possibilities for the last row of γ. Motivated by this observation, we refine the argument by counting the lattice points γ_n only up to the contribution of the vector e_n. This latter contribution and the shape of the determinant (being a ν-th power) give rise to a diophantine equation that has the right amount of solutions in the generic case. We then consider the degenerate case separately. To simplify the latter, we eventually make the assumption that the degree n is prime.

For talking about the non-degenerate case, denote by $\chi_\gamma(X) = \det(X \cdot \mathrm{id}_n - \gamma)$ the characteristic polynomial of a matrix γ. We call $\gamma \in M_n(\mathbb{Q})$ *non-degenerate* if the polynomials

$$(-1)^n \chi_\gamma(X) - Y^\nu \in \mathbb{Q}[X, Y]$$

are irreducible over $\mathbb{Q}$ for all $1 \leq \nu \leq n$. Define

$$H_*(z, m, N) = \{\gamma \in H(z, m, N) \mid \gamma \text{ non-degenerate}\}.$$

Proposition 2.6.4. *Assume the same conditions as in Proposition 2.6.2. Additionally, let $L \ll N^{1/n^2-\varepsilon}$ and $N \gg_\Omega 1$ be large enough. Then*

$$\sum_{m \asymp L^n} |H_*(z, m^\nu, N)| \ll L^{(n-1)\nu} \cdot L^{1+\varepsilon}$$

for any $1 \leq \nu \leq n$.

[5] We also remark that numerical experiments in dimension $n = 2$ seem to indicate that the bound we obtain for the possibilities for the last row might be sharp.

Proof. Let $\gamma \in H_*(z, m^n, N)$, $m \asymp L^n$, and consider again the number of possibilities for the last row γ_n. For this, let $b_1, \dots, b_{n-1}, e_n z$ be a reduced basis for L_{z_N} (see Section 2.2.3 for the definition, which we apply to $N^{-(n-1)/n} z_N \in \mathrm{SL}_n(\mathbb{R})$). From the proof of Proposition 2.6.2, we note again that e_n is a vector of shortest length in L_{z_N}, where

$$\|e_n\|_z = \|e_n\|_{z_N} = d \asymp_{n,\Omega} N^{(n-1)/n}.$$

By Minkowski's theorem, we also have $\|b_i\| \asymp_{n,\Omega} N^{(n-1)/n}$.

Now $\gamma_n \in L_{z_N}$, so it can be written as

$$\gamma_n = \sum_{i=1}^{n-1} a_i b_i + a_n e_n z$$

with $a_i \in \mathbb{Z}$. By Lemma 2.2.6 and recalling the condition $\|\gamma_n\|_z \ll L^v d$ from (2.6.3), we deduce that $a_i \ll L^v$, for all $1 \le i \le n$.

There are $L^{(n-1)v}$ possibilities for $a_1, \dots, a_{n-1}$. Choose any such combination of coefficients and assume there exist $\gamma \in H_*(z, m^v, N)$ and $\gamma' \in H_*(z, l^v, N)$, for $m, l \asymp L^n$, such that $\gamma_n = \sum_{i=1}^{n-1} a_i b_i + a_n e_n z$ and

$$\gamma'_n - \gamma_n = \lambda e_n.$$

Then $\lambda \in \mathbb{Z}$ and $\lambda \ll L^v$. Observe also that the matrix $\gamma - \gamma'$ satisfies the same geometric conditions (2.6.3) as γ and γ', simply by the triangle inequality (with a doubled implied constant, of course).

We now apply the same iterative process as in the proof of Proposition 2.6.2. We note however that, under the present conditions, each step yields at most one possibility. Indeed, fix the last row of $\gamma - \gamma'$, having the form λe_n, by fixing $\lambda \ll L^v$. Next, the number of possibilities for the row $(\gamma - \gamma')_{n-1}$ is bounded by the number of L_z-lattice points in a ball of radius $L^v \cdot N^{-1/n}$ centred at $x_{n-1,n} \cdot \lambda e_n z$, where $x_{n-1,n}$ is one of the x-coordinates of z. By assumption, the radius is bounded by $N^{-\varepsilon}$. However, if N is large enough, this is greater than the first successive minimum of z, which is $\asymp_\Omega 1$. There is thus only one possible lattice point.

On the other hand, it is clear that the multiple $\lambda \cdot \mathrm{id}_n$ of the identity matrix lies in the set $H(z, \lambda, N)$. Since $\lambda \ll L^v$, we see that $\lambda e_{n-1} \cdot z$ satisfies the condition of the lattice point above (again, condition (2.6.3)). Consequently, it follows that

$$(\gamma - \gamma')_{n-1} = \lambda \cdot e_{n-1}.$$

Iterating this argument and keeping in mind the computations in the proof of Proposition 2.6.2, we deduce that

$$\gamma - \gamma' = \lambda \cdot \mathrm{id}_n .$$

It remains to count the possibilities for λ. Considering the determinant of γ', we have

$$l^\nu = \det(\gamma') = \det(\gamma - \lambda \cdot \mathrm{id}_n) = (-1)^n \chi_\gamma(\lambda).$$

Therefore, $(\lambda, l) \in \mathbb{Z}$ are a solution to the equation

$$(-1)^n \chi_\gamma(X) - Y^\nu = 0.$$

Since this polynomial is defined over $\mathbb{Z}$ and irreducible over $\mathbb{Q}$ by assumption, we count the number of such solutions using Heath-Brown's Theorem 3 in [HB02]. In the notation there, after homogenising the polynomial, we set $B_1 = L^\nu$ for the bound on λ, then $B_2 = L^n$ for the bound on l, and finally $B_3 = 1$ for the bound on the additional variable. Then we compute $T = L^{n\nu}$ and $V = L^{\nu+n}$. Heath-Brown's result then gives the bound

$$\frac{V^{1/n+\varepsilon}}{T^{1/n^2}} = L^{1+\varepsilon}$$

on the number of solutions we are considering. This bounds in particular the number of possibilities for λ over all relevant determinants and so finishes the proof. □

We are now left with counting degenerate matrices. This is reminiscent of treating the special case of parabolic matrices in [HT12, Lemma 2]. For this we restrict to prime degrees, allowing for a clean classification of the degenerate case.

Let $n \geq 2$ be prime. Since χ_γ is a polynomial of degree n over $\mathbb{Q}$, a result of Schinzel [Sch67] shows that

$$(-1)^n \chi_\gamma(X) - Y^\nu$$

is irreducible, unless $\nu = n$ and

$$\chi_\gamma(X) = \alpha(X - \beta)^n$$

for $\alpha, \beta \in \mathbb{Q}$. In the first case, it is irreducible over $\mathbb{C}$ if and only if it is irreducible over $\mathbb{Q}$. In the latter case, we have $\alpha = 1$ by normalisation and $\beta^n = \det(\gamma)$.

The irreducibility criterion above and Proposition 2.6.4, by following its proof again verbatim, imply the following bounds.

Corollary 2.6.5. *Assume the same conditions as in Proposition 2.6.4 and, additionally, let n be prime. Then*

$$\sum_{m \asymp L^n} |H(z, m^\nu, N)| \ll L^{(n-1)\nu} \cdot L^{1+\varepsilon}$$

for any $1 \leq \nu \leq n-1$.

We have thus reduced the problem to counting matrices $\gamma \in H(z, m^n, N)$ for some $m \asymp L^n$, such that

$$\chi_\gamma(X) = (X - \beta)^n.$$

Since $\beta \in \mathbb{Q}$, it follows that $\beta = \pm m \in \mathbb{Z}$ (there is no sign for odd n). Denote the subset of such matrices by $H_{\mathrm{par}}(z, m^n, N)$.

The method of proof in Proposition 2.6.4 provides even more. We recall at this point that the determinants m^ν appearing in the counting problem have a particular shape, namely $m = p \cdot q^{n-1}$, where p and q are primes of size L (see the amplifier in Proposition 2.3.4). We are thus averaging over a set of size L^2. However, we can consider the special case $p = q$ to reduce this size.

Corollary 2.6.6. *Assume the same conditions as in Corollary 2.6.5. Then*

$$\sum_{p \asymp L} |H_{par}(z, p^{n^2}, N)| \ll L^{(n-1)n} \cdot L.$$

Proof. We follow the proof of Proposition 2.6.4, but first we fix the determinant p^{n^2}, where $p \asymp L$. There are, of course, at most L such determinants. Now the number of choices for a potential last row of $\gamma \in H_{\mathrm{par}}(z, p^{n^2}, N)$ up to the contribution of e_n, i.e. up to the last component, is bounded by $L^{(n-1)n}$. Choose γ and γ' two matrices in $H_{\mathrm{par}}(z, p^{n^2}, N)$ with the same last row up to the last component.

As in the proof of Proposition 2.6.4, we find that

$$\gamma - \gamma' = \lambda \cdot \mathrm{id}_n .$$

We apply again the determinant to this equation and obtain that

$$(\lambda - p^n)^n = p^{n^2}.$$

It follows that there are only two possibilities for λ and this proves the statement. □

We observe that the actual average of size L^2 would have given a bound of the form $L^{n(n-1)} \cdot L^2$, which is on the edge of what is needed for a saving. The next section significantly refines the argument to treat this issue.

2.6.4 Counting at different cusps

Corollary 2.6.6 allows us now to reduce the problem further. We are now counting matrices γ in the set

$$\bigcup_{\substack{p,q \asymp L \\ p \neq q}} H_{\mathrm{par}}(z, (pq^{n-1})^n, N).$$

By Theorem III.12 in [New72], there is $h \in \mathrm{SL}_n(\mathbb{Z})$ such that

$$h\gamma h^{-1} = \begin{pmatrix} m & * & * \\ & \ddots & * \\ & & m \end{pmatrix} \tag{2.6.4}$$

is upper triangular with m on the diagonal. Indeed, χ_γ splits into linear factors and thus the blocks in [New72, Thm. III.12] are one dimensional.

In the simplest case, we could assume that $h \in \Gamma_0(N)$. The next lemma shows that this is almost the same as assuming that $h = 1$ and that γ has the same last row as the identity matrix, in which case we apply the philosophy from Remark 2.6.3, namely that the last row determines the matrix. However, we remark here already that there are other possibilities for h that correspond to different *cusps*, as in Lemma 2.6.8 below, for which counting becomes more difficult.

Lemma 2.6.7. *Assume the same conditions as in Proposition 2.6.4 and let $\gamma \in H(z, m^n, N)$ or $\gamma \in H(z', m^n, N)$ for $m \asymp L^n$. If there exists $h \in \Gamma_0(N)$ such $h\gamma h^{-1}$ has last row equal to $m \cdot e_n = (0, \dots, 0, m)$, then $\gamma = m \cdot \mathrm{id}_n$.*

Proof. Assume that $\gamma \in H(z, m^n, N)$. Since $h \in \Gamma_0(N)$, it is easy to see from the definition that $\gamma \in H(z, m^n, N)$ implies $\eta := h\gamma h^{-1} \in H(hz, m^n, N)$. Consider the Iwasawa coordinates of $hz = n(x)a(y)$. Multiplying h from the left by a matrix of the form

$$\begin{pmatrix} \xi & \\ & 1 \end{pmatrix} \in \Gamma_0(N)$$

with $\xi \in \mathrm{SL}_n(\mathbb{Z})$, we may assume that $y_i \gg 1$ for $i = 2, \dots, n-1$ (see Remark 2.2.4). Under such a modification, we may also still assume that the last row $e_n\eta$ has the form $(0, \dots, 0, m) = m \cdot e_n$.

To obtain from this bounds on the entries of $a(y)$ we note that, since $\det(hz) = 1$,

$$\|e_2 \wedge \cdots \wedge e_n\|_{hz} = (dy_1 \cdots y_{n-1})^{-1} \geq \beta(z),$$

recalling the definition of $\beta(z)$ in Section 2.5.2 and that $h \in \Gamma_0(N)$. By assumption, $z \in \mathcal{L}'(B, A)$, which by Table 2.1 implies that $\beta(z) = N^{1/n}\lambda_1$ for λ_1 the first successive minimum of $L_{z'}$. Lemma 2.6.1 shows now that $\beta(z) \asymp N^{1/n}$. As such, we have

$$dy_1 \ll dy_1y_2 \ll \ldots \ll dy_1 \cdots y_{n-1} \ll N^{-1/n}.$$

This is now a similar situation in the proofs of the counting results Proposition 2.6.2 and Proposition 2.6.4, except that d might be large. However, the last row of η is already fixed to be $m \cdot e_n$. As in Proposition 2.6.4, the assumption $L \ll N^{1/n^2-\varepsilon}$ and the bound above on the entries of $a(y)$ imply that the last row of η determines the whole matrix. Therefore, $\eta = m \cdot \mathrm{id}_n$ and so, undoing conjugation, $\gamma = m \cdot \mathrm{id}_n$.

The case $\gamma \in H(z', m^n, N)$ follows analogously. What changes is, for instance, that $\beta(z') = N^{1/n}\lambda_1$ for λ_1 the minimum of L_z. We then continue by using Lemma 2.6.1 again. □

We investigate now the cusps of $\Gamma_0(N)$ with respect to the minimal parabolic. Define therefore $U_n(\mathbb{Z})$ to be the subgroup of $\mathrm{SL}_n(\mathbb{Z})$ of unipotent upper triangular matrices, that is, with ones on the diagonal.

Let also $W_n \leq \mathrm{SL}_n(\mathbb{Z})$ denote the subgroup of permutation matrices. We call two such matrices equivalent if they have the same last row and denote by $\overline{W}_n$ the set of equivalence classes. By considering $\mathrm{SL}_{n-1}(\mathbb{Z})$ embedded inside $\Gamma_0(N)$, it is easy to see that

$$\overline{W}_n \cong \Gamma_0(N) \cap W_n \backslash W_n.$$

and note also that $|\overline{W}_n| = n$.

Lemma 2.6.8. *Let N be prime. Then any system of representatives for $\overline{W}_n$ is a system of representatives for the double quotient*

$$\Gamma_0(N) \backslash \mathrm{SL}_n(\mathbb{Z}) / U_n(\mathbb{Z}).$$

Proof. Let $\xi \in \mathrm{SL}_n(\mathbb{Z})$ and let $(a_1, \ldots, a_n)$ be the first column of ξ, a primitive vector in $\mathbb{Z}^n$. First, we reduce a_n to either 0 or 1 by acting from the left by $\Gamma_0(N)$.

Indeed, assume that $\gcd(a_n, N) = 1$. Then the vector $(Na_1, \ldots, Na_{n-1}, a_n)$ is also primitive. Therefore there is a primitive $(b_1, \ldots, b_n) \in \mathbb{Z}^n$ such that

$$Na_1b_1 + \ldots + Na_{n-1}b_{n-1} + a_nb_n = 1.$$

From this it is clear that $\gcd(N, b_n) = 1$ so that $(Nb_1, \ldots, Nb_{n-1}, b_n)$ is primitive. Let $\gamma \in \mathrm{SL}_n(\mathbb{Z})$ be a matrix with the latter as its last row. Then $\gamma \in \Gamma_0(N)$ and $\gamma\xi$ has last row of the form $(1, *, \ldots, *)$.

Since N is prime, negating the assumption above means that $N \mid a_n$. Now let $d = \gcd(a_1, \ldots, a_{n-1})$. Then $\gcd(a_n, d) = 1$ and there exists a primitive vector $(b_1, \ldots, b_{n-1})$ such that

$$b_1a_1 + \ldots b_{n-1}a_{n-1} = d.$$

Therefore

$$\sum_{i=1}^{n-1} (a_nb_i) \cdot a_i + (-d)a_n = 0.$$

The vector $(a_nb_1, \ldots, a_nb_{n-1}, -d)$ is primitive by the observations above, so there is $\gamma \in \mathrm{SL}_n(\mathbb{Z})$ with this vector as its last row. Again, $\gamma \in \Gamma_0(N)$ since $N \mid a_n$ and the last row of $\gamma\xi$ has the form $(0, *, \ldots, *)$.

Assume now that ξ has last row of the form $(1, *, \ldots, *)$. It is clear that we can multiply ξ from the right by a matrix in $U_n(\mathbb{Z})$ such that the resulting last

row is simply $(1, 0, \dots, 0)$. Call this new matrix ξ again and take $w \in W_n$ a permutation matrix with the same last row (for instance the so-called long Weyl element). In other words, $e_n\xi = e_n w$, where e_n is the n-th standard basis vector $(0, \dots, 0, 1)$. The matrix $w\xi^{-1}$ preserves e_n so it must have e_n as its last row. In particular, $w\xi^{-1} \in \Gamma_0(N)$ and we are done in this case.

On the other hand, let ξ have last row of the form $(0, *, \dots, *)$. Using the embedding of $\mathrm{SL}_{n-1}(\mathbb{Z})$ in the upper left corner of $\Gamma_0(N)$, we may modify ξ so that its first column is of the form $(1, 0, \dots, 0)$, by similar arguments. This now allows an inductive procedure, considering the lower right $n-1 \times n-1$ block of ξ. We see that one can always reduce the last row of ξ to be a standard basis vector and the paragraph above shows how to obtain a permutation matrix from ξ.

To check that no two such representatives in $\overline{W}_n$ produce the same double coset is easy. For $w_1, w_2 \in W_n$, if $w_1 = \gamma w_2 u$ with $\gamma \in \Gamma_0(N)$ and $u \in U_n(\mathbb{Z})$, then $\gamma = w_1 u^{-1} w_2$. One now computes the shape of $U_n(\mathbb{Z})$ transformed by permutation of rows and of columns. We leave out the details of this argument. □

Remark 2.6.9. We make the following simple observation that becomes very useful in the arguments below. Let $w_k \in \overline{W}_n$ be a representative with last row equal to e_k. We can take $w_n = \mathrm{id}_n$. We can also take w_1 to be the long Weyl element

$$w_1 = \begin{pmatrix} & & & & 1 \\ & & & 1 & \\ & & \iddots & & \\ & 1 & & & \\ 1 & & & & \end{pmatrix}$$

with ones on the anti-diagonal. Finally, for any $k \neq 1$, we can choose the representative w_k to have first row (and thus also first column) equal to e_1.

We finally state the main result for degenerate matrices below and recall the additional condition on the determinantal divisors appearing in the amplifier, Proposition 2.3.4.

Proposition 2.6.10. *Assume the same conditions as in Corollary 2.6.5. For N large enough, the set of matrices γ possibly occurring in $H_{par}(z, (pq^{n-1})^n, N)$ for some primes $p, q \asymp L$, $p \neq q$, satisfying additionally that*

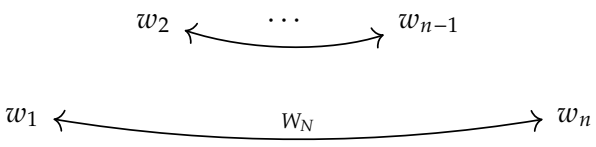
$$\Delta_{n-1}(\gamma) = q^{(n-1)(n-2)}$$

is empty.

It is perhaps useful at this point to give a brief overview of the proof. We make a case distinction, based on the cusp classification above. If h in (2.6.4) corresponds to the identity w_n, then we are done by Lemma 2.6.7. If

h corresponds to the long Weyl element w_1, we apply the Fricke involution, which effectively switches the cases w_n and w_1, and so the same lemma, available for both z and z', finishes this case as well.

In fact, the Fricke involution generally exchanges the cases w_k and w_{n+1-k}.



However, the usual counting argument, choosing vectors step-by-step from the bottom of the matrix going upwards, seems difficult to implement in the intermediate cases $1 < k < n$. It is here that the assumption $p \neq q$, together with the seemingly harmless choice of representatives w_k in Remark 2.6.9, comes in. Indeed, the choice of representatives is akin to a very weak balancedness assumption on the new, unknown basis for the lattice that appears in the counting problem. This assumption implies that at least one element of the superdiagonal of the upper triangular matrix in (2.6.4) is zero. Computing Δ_{n-1}, this is enough to derive a contradiction to $p \neq q$.

Proof. Let $m = pq^{n-1}$. As in (2.6.4), there is $h \in \mathrm{SL}_n(\mathbb{Z})$ such that $h\gamma h^{-1}$ is upper-triangular with diagonal $(m, \dots, m)$. By Lemma 2.6.8 we can write $h^{-1} = \sigma^{-1} w u^{-1}$ with $\sigma \in \Gamma_0(N)$, $u \in U_n(\mathbb{Z})$, and $w \in \overline{W}_n$.

Next, conjugating by u, we easily see that

$$w^T \sigma\gamma\sigma^{-1} w = \begin{pmatrix} m & * & * \\ & \ddots & * \\ & & m \end{pmatrix} =: \eta \tag{2.6.5}$$

is also of the same form. Now if $w = w_n = \mathrm{id}_n$, meaning that the last row of w is e_n as in Remark 2.6.9, we are done by Lemma 2.6.7. The latter implies that $\gamma = pq^{n-1} \cdot \mathrm{id}_n$, which does not have the required determinantal divisors and leads to a contradiction.

If $w = w_1$ is the long Weyl element, we apply the Fricke involution. By transposing the condition

$$z^{-1}\gamma z = O(m).$$

we see that

$$A_N(\sigma\gamma\sigma^{-1})^T A_N^{-1}$$

lies in $H(\tilde{\sigma} z', m^n, N)$ with some $\tilde{\sigma} \in \Gamma_0(N)$.

Next, observe that

$$(\sigma\gamma\sigma^{-1})^T = w\eta^T w^T$$

is again upper triangular. By Lemma 2.6.7, we deduce that

$$A_N(\sigma\gamma\sigma^{-1})^T A_N^{-1} = m \cdot \mathrm{id}_n$$

and thus $\gamma = m \cdot \mathrm{id}_n$, which is a contradiction again.

Finally, let $w = w_k$ with $1 < k < n$. Notice first that (2.6.5) and the congruences modulo N satisfied by γ and σ imply that the k-th row of η also satisfies congruences. Indeed, $w_k \eta w_k^T$ is a matrix of the $\Gamma_0(N)$ shape. More precisely, $N \mid \eta_{kj}$ for $j > k$. Since $k < n$, we have in particular $N \mid \eta_{k,k+1}$.

Let us now assume that the superdiagonal of η only contains non-zero elements. That is, $\eta_{j,j+1} \neq 0$ for all $1 \geq j < n$. Recall the condition

$$z^{-1}\gamma z = (w^T \sigma z)^{-1} \eta (w^T \sigma z) = O(m).$$

We can rewrite $w^T \sigma z = n \cdot a$ in Iwasawa coordinates (indeed, conjugating by an orthogonal matrix leaves $O(m)$ invariant), denoting the y-coordinates as usual. It is now a common and important observation that the superdiagonal of upper triangular matrices enjoys a certain additive abelian-like property with respect to matrix multiplication. This observation or direct computation should convince the reader that

$$(na)^{-1}\eta na = \begin{pmatrix} m & y_{n-1}^{-1}\eta_{1,2} & * & \cdots \\ & m & y_{n-2}^{-1}\eta_{2,3} & \cdots \\ & & \vdots & \vdots \\ & & m & y_1^{-1}\eta_{n-1,n} \\ & & & m \end{pmatrix}.$$

Since this is $O(m)$, the assumption that $|\eta_{j,j+1}| \geq 1$ now implies that $y_j \gg 1/m$. Even more and crucially, recall that $N \mid \eta_{k,k+1}$, so that $y_{n-k} \gg N/m$. Putting these together, we obtain the bound

$$y_1 \cdots y_{n-1} \gg \frac{N}{m^{n-1}} \gg N^{1/n+\varepsilon}$$

using the assumption $L \ll N^{1/n^2-\varepsilon}$ and that $m \asymp L^n$.

We return now to a technique used in the proof of Lemma 2.6.7. We observe again that

$$\|e_2 \wedge \cdots \wedge e_n\|_{w^T \sigma z} = (d y_1 \cdots y_{n-1})^{-1}.$$

On the other hand, our choice of representative $w = w_k$ in Remark 2.6.9 implies that the first row of w^T is equal to e_1 and the other rows are permuted between them in some way. This means that

$$\|e_2 \wedge \cdots \wedge e_n\|_{w^T \sigma z} = \|e_2 \wedge \cdots \wedge e_n\|_{\sigma z} \geq \beta(z) \asymp N^{1/n}.$$

Therefore, as in the proof of the aforementioned lemma, we obtain that

$$d y_1 \cdots y_{n-1} \ll N^{1/n}.$$

Recall also that $d = \|e_n\|_{w^T \sigma z}$, and since z defines a balanced lattice, $d \gg 1$. Therefore

$$y_1 \cdots y_{n-1} \ll N^{1/n},$$

which constitutes a contradiction to the previous paragraph for large enough N.

We deduce that the superdiagonal of η must contain some zero. It is now straight-forward to prove that m divides $\Delta_{n-1}(\eta)$. Indeed, the only $(n-1)\times(n-1)$ minor that is not obviously divisible by m is the upper right minor, formed by removing the first column and the last row of η. Proving the claim here is an easy exercise in Laplace, or cofactor, expansion.

Observe now that the invariance properties of determinantal divisors (see [New72, Thm. II.8]) imply that

$$\Delta_{n-1}(\eta) = \Delta_{n-1}(\gamma),$$

since $w, \sigma \in \mathrm{SL}_n(\mathbb{Z})$. Since $p \mid m$, it follows from the paragraph above and our assumption on the determinantal divisors that

$$p \mid q^{(n-1)(n-2)}.$$

If $n > 2$, this implies that $p = q$, which is a contradiction to the assumption. □

Remark 2.6.11. Notice that the case $n = 2$ does not involve any intermediate Weyl elements. Indeed, there are only two cusps and both reduce as above to counting upper-triangular matrices directly. A more general result (for square-free levels) is contained in a slightly different language in [HT12, Lemma 4.1].

The counting results of this section taken together produce the following corollary. It gives a solution to the counting problem for prime n that can be successfully applied to the sup-norm problem through the amplifier in Proposition 2.3.4.

Corollary 2.6.12. *Let n and N be a prime, and let $z \in \mathbb{H}$ reduce to a compactum Ω. Assume that z has Iwasawa coordinates*

$$y_1 \asymp_\Omega \frac{1}{N} \quad \text{and} \quad y_i \asymp_\Omega 1$$

for $i = 2, \dots, n-1$, and $d = \alpha(z) = \|e_n\|_z$, and satisfies

$$z \in \mathcal{L}(B, A) \cap \mathcal{L}'(B, A).$$

Let $L \ll N^{1/n^2-\varepsilon}$ and assume that $N \gg_{\Omega,\varepsilon} 1$ is large enough. Then

$$\sum_{p,q \asymp L} |H(z, p^\nu, q^{(n-1)\nu}, N)| \ll L^{(n-1)\nu} \cdot L^{1+\varepsilon}$$

for any $1 \le \nu \le n$.

2.7 FINAL STEPS

2.7.1 The Fourier bound

To prove a bound in the first domain given by the reduction in Proposition 2.5.2, we use the Whittaker expansion and bounds for the first Fourier coefficient of newforms of level N. Here we state a version of the bound that is unconditional, yet sufficient for our purposes.

Proposition 2.7.1. *Let ϕ be an L^2-normalised Hecke-Maaß newform of prime level N and spectral parameter μ, and let $z \in \Omega$ for some compactum $\Omega \subset \mathbb{H}$. For $\varepsilon > 0$ we have*

$$\phi(z) \ll_{\Omega,\mu\varepsilon} N^{-1/4+1/4n+\varepsilon}.$$

Proof. We use the bound given in Theorem 3 of [BHM20], making the necessary adjustments from the level 1 results to level N. The proof is very similar, so we refer to [BHM20] for more details and mostly remark on what changes need to be made.

Note first that the method of proof involves the Whittaker expansion [BHM20, (46)]. An automorphic form for the group $\Gamma_0(N)$ enjoys the same type of Whittaker expansion, since $\mathrm{SL}_{n-1}(\mathbb{Z})$ embeds in the upper left $(n-1)\times(n-1)$ block of $\Gamma_0(N)$, so that one can follow the same arguments given in, for instance, [Gol06, Theorem 5.3.2] in level 1. To follow the arguments in [BHM20] further, we normalise ϕ arithmetically, so that the first coefficient in the expansion is 1.

Next, the bound [BHM20, (49)] for $L(1+\varepsilon, \pi \times \tilde{\pi})$ holds similarly, with an additional N^{ε} on the right-hand side. Here, we let π be the automorphic representation generated by ϕ. Finally, to account for the factor between arithmetically normalised forms and L^2-normalised forms, we note the display before [BHM20, (66)]. More precisely, if we assume ϕ to be arithmetically normalised, as in [BHM20, (46)], then standard Rankin-Selberg theory shows that

$$\left\|\phi\right\|^2 \asymp_{\mu} \mathrm{vol}(\Gamma_0(N)\backslash\mathbb{H}) \cdot \mathrm{res}_{s=1} L(s, \pi \times \tilde{\pi}).$$

By [Bru06, Theorem 3], as in the two displays after (43) in [Lap13, Appendix], we can use the lower bound

$$\mathrm{res}_{s=1} L(s, \pi \times \tilde{\pi}) \gg C(\pi \times \tilde{\pi})^{-1/2+1/2n-\varepsilon},$$

where $C(\pi \times \tilde{\pi}) = C(\pi \times \tilde{\pi}, 0)$ is the analytic conductor of $L(s, \pi \times \tilde{\pi})$. We have $C(\pi) \asymp_{\mu} N$ and by [BH97, (2)] the bound

$$C(\pi \times \tilde{\pi}) \ll N^{n\cdot 1+n\cdot 1-1}$$

holds.

It is easy to compute that $\mathrm{vol}(\Gamma_0(N)\backslash\mathbb{H}) \asymp N^{n-1}$. Therefore,

$$\left\|\phi\right\|^2 \gg_{\mu} N^{n-1} \cdot N^{(2n-1)(-1/2+1/2n)-\varepsilon} = N^{1/2-1/2n-\varepsilon}.$$

Going back to ϕ being L^2-normalised by putting together the bound above and [BHM20, (49)] with the indicated adjustments, we deduce the claim. □

Remark 2.7.2. Working more precisely, one could prove that $C(\pi \times \tilde{\pi}) \asymp N^{2n-2}$ and improve the exponent in the bound above. This is not necessary for this paper and we do not prove the claim. In fact, one expects that $\mathrm{res}_{s=1} L(s, \pi \times \tilde{\pi}) \gg N^{\varepsilon}$. This is proven by Hoffstein-Lockhart in the case $n = 2$ and for this reason we have

$$\phi(z) \ll_{\Omega,\mu,\varepsilon} N^{-1/2},$$

for $z \in \Omega$ as in [HT12, Lemma 4], for example.

2.7.2 *Finishing the proof*

First assume Hypothesis (2.1.1). Proposition 2.3.3 and Proposition 2.6.2 together with the prime number theorem imply that

$$\phi(z)^2 \ll_{\mu,\Omega,\varepsilon} L^{-1/2+\varepsilon} + L^{-1/2-n+\varepsilon} \cdot L^n (1 + L^n/N)^{n-1},$$

under the assumptions on z specified in Proposition 2.6.2. Optimising the size of L, we choose $L = N^{1/n}$. In this case, we have

$$\phi(z) \ll L^{-1/4+\varepsilon} \ll N^{-1/4n+\varepsilon}.$$

The Fourier bound, Proposition 2.7.1, certainly implies the same bound

$$\phi(z) \ll N^{-1/4n+\varepsilon}$$

for $n \geq 2$.

These bounds are valid on the subsets of $\mathbb{H}$ given in Proposition 2.5.2. As remarked at the beginning of Section 2.5, these now extend to the whole domain Ω_N, and the proof is finished.

Without assuming Hypothesis (2.1.1), we let n be prime and we apply Proposition 2.3.4 using the counting result Corollary 2.6.12. Similarly to the computation above, we have

$$\phi(z)^2 \ll_{\mu,\Omega,\varepsilon} L^{-1+\varepsilon} + \sum_{\nu=1}^{n} \frac{1}{L^{(n-1)\nu}} \cdot L^{(n-1)\nu} L^{1+\varepsilon}$$

for $L \ll N^{1/n^2-\varepsilon}$. Maximising L, we get

$$\phi(z) \ll L^{-1/2+\varepsilon} \ll N^{-1/2n^2+\varepsilon}.$$

3. The cocompact case

This chapter reproduces the scientific article [Tom23]:
R. Toma. 'Hybrid bounds for the sup-norm of automorphic forms in higher rank'. *Trans. Amer. Math. Soc.* 376.8 (2023), pp. 5573–5600.

Abstract

Let A be a central division algebra of prime degree p over $\mathbb{Q}$. We obtain subconvex hybrid bounds, uniform in both the eigenvalue and the discriminant, for the sup-norm of Hecke-Maaß forms on the compact quotients of $\mathrm{SL}_p(\mathbb{R})/\mathrm{SO}(p)$ by unit groups of orders in A. The exponents in the bounds are explicit and polynomial in p. We also prove subconvex hybrid bounds in the case of certain Eichler-type orders in division algebras of arbitrary odd degree.

3.1 INTRODUCTION

3.1.1 Motivation and historical context

The sup-norm problem arises as a natural question in analysis and quantum physics and has received considerable attention in the number theory community. It is the problem of bounding the L^∞-norm of eigenfunctions on Riemannian manifolds in terms of their L^2-norm. To make this a reasonable endeavour, one chooses some parameters for the eigenfunctions and estimates the quotient of the two norms while these parameters vary.

For example, let X be a compact Riemannian manifold of dimension n and let ϕ be an L^2-normalised eigenfunction of the Laplacian Δ_X with eigenvalue $\lambda > 0$. Motivated by semi-classical analysis, one would like to bound $\|\phi\|_\infty$ in terms of λ when $\lambda \to \infty$. In general, local analysis gives the sharp baseline bound

$$\|\phi\|_\infty \ll \lambda^{(n-1)/4+\varepsilon},$$

for large enough λ. The bound is attained on the round n-spheres.

If ϕ is assumed to be an eigenfunction for a larger algebra of operators, then we can expect better bounds. Indeed, if $X = \Gamma\backslash S$ is a compact locally symmetric space of rank r and $G(S)$ is the groups of isometries of the symmetric space S, then one can consider the algebra of $G(S)$-invariant differential operators. This algebra is generated by r operators, including the Laplacian. If ϕ is an

L^2-normalised joint eigenfunction of these operators, then[1]

$$\|\phi\|_\infty \ll \lambda^{(n-r)/4+\varepsilon}. \tag{3.1.1}$$

For more details on the above paragraphs, see [Sar04].

In special cases when X has arithmetic structure, we expect to obtain better, so-called *subconvex* bounds, i.e. an exponent $(n-r)/4-\delta$ in (3.1.1) with $\delta > 0$. In these cases, there is an additional algebra of commutative normal Hecke operators, which commute with the differential operators above. In this arithmetic setting, the sup-norm problem is to find a subconvex bound for $\|\phi\|_\infty$, where ϕ is a joint eigenfunction of the invariant differential operators and the Hecke algebra. The prototype of such a result is due to Iwaniec and Sarnak [IS95] in the case of $X = \Gamma\backslash\mathfrak{h}^2$, where $\mathfrak{h}^2$ is the hyperbolic plane and $\Gamma \leq \mathrm{SL}_2(\mathbb{R})$ is a cocompact arithmetic subgroup. They show that $\|\phi\|_\infty \ll \lambda^{1/4-1/24+\varepsilon}$ for an L^2-normalised Hecke-Maaß form ϕ.

Another parameter that one could choose is the *volume* of X. This is particularly interesting in the arithmetic case and is reminiscent of the level aspect in the subconvexity problem of L-functions (indeed, the two problems are very much related in methodology and numerology). One thoroughly studied example is the family of non-compact spaces $X_0(N) := \Gamma_0(N)\backslash\mathfrak{h}^2$ of volume $N^{1+o(1)}$, where $\Gamma_0(N)$ is the Hecke congruence subgroup of level N. For an example of an application, the corresponding 'convexity' bound on average (for squarefree N) was used in [AU95] to compare the Arakelov and the Poincaré metrics on $X_0(N)$. A great amount of work was dedicated to achieving sub-baseline bounds in more and more general settings, for example in [BH10], [Tem10], [HT13], [Sah17], [Ass a]. See the introduction of [HS20] for a more complete set of references with the corresponding bounds.

In general, the level aspect seems to factorise into the case of squarefree level and the case of powerful level, in particular prime powers. The latter is called the *depth* aspect and is amenable to techniques from p-adic analysis that are not available in the squarefree case. To describe an example, let N_1 denote the smallest positive integer such that $N \mid N_1^2$. Note that $N_1 = N$ if N is squarefree, yet $N_1 \asymp \sqrt{N}$ if N is a high prime power. As an example of a subconvex bound in the case of $X_0(N)$ and the interplay between squarefree and powerful levels, it was shown in [Sah17] that

$$\|\phi\|_\infty \ll N^{-2/6+\varepsilon} N_1^{1/6} \lambda^{5/24+\varepsilon},$$

for ϕ a newform. This is also an example of a *hybrid* bound. Such a bound is uniform in *both* the spectral and the volume aspect.

Closer to the topic of these notes is the case of cocompact surfaces, where the arithmetic subgroup is given by the norm 1 units $\mathcal{O}^1$ of an order $\mathcal{O}$ in an indefinite division quaternion algebra A over $\mathbb{Q}$. The volume of $\mathcal{O}^1\backslash\mathfrak{h}^2$ is

[1] Here and in the rest of this article, the implied constants are allowed to depend on ε.

approximately equal to the squareroot of the discriminant of $\mathcal{O}$ (see Section 3.2.1). This discriminant is made up of the discriminant of the algebra and the so-called level of $\mathcal{O}$. More precisely, we choose a fixed maximal order $\mathcal{O}_m$ containing $\mathcal{O}$ and define $N = [\mathcal{O}_m : \mathcal{O}]$, which we call the level, and note that $\mathrm{disc}(\mathcal{O}) = N^2 \cdot \mathrm{disc}(\mathcal{O}_m) = N^2 \cdot \mathrm{disc}(A)$. Most results in the literature only give bounds in the parameter N, but the discriminant of A, and therefore the covolume of $\mathcal{O}$, can get arbitrarily large independently of the level. A hybrid bound that is uniform in the full volume aspect was proved by Blomer and Michel [BM13] for the related case of totally definite quaternion algebras.

For Eichler orders of indefinite quaternion algebras, the (p-adic) local bound $\|\phi\|_\infty \ll N^{-1/2+\varepsilon} N_1^{1/2}$ was shown by Marshall [Mar16] for ϕ a newform. For squarefree N this corresponds to the bound N^ε, which was first improved by Templier [Tem10], who obtained the bound $N^{-1/24+\varepsilon}$ for general N. Saha [Sah20, Theorem 3] combines these two bounds to $N^{-11/24+\varepsilon} N_1^{5/12}$ for newforms. Saha remarks in Section 1.6 that the argument should also provide a non-trivial hybrid bound but the details do not seem to be in print.

The depth aspect was recently improved in [HS20] for newforms. If the level is a prime power p^n, then Hu and Saha obtain the bound $p^{n(5/24+\varepsilon)}$.

It should be noted at this point that the bounds above that include the number N_1 are proven using newform theory and, in particular, assuming that the order is Eichler, as stated. Saha and Templier also prove results for general orders (e.g. [Sah20, Theorem 1]), thus not assuming that the forms are newforms, but only that they are Hecke eigenfunctions. Templier [Tem10, Theorem 2] explicitly assumes ϕ is a newform, but the argument does not use newform theory; it does, however, use the structure of an Eichler order. As is apparent in the main results of this paper, described in the next section, our focus is rather on proving bounds in the case of general orders and we therefore never use newforms or the number N_1.

The sup-norm problem has also been pushed in the last decade to the case of higher-rank groups, such as $\mathrm{GL}(n)$ for $n > 2$. For example, Blomer and Maga [BM16] prove that if ϕ is a Hecke-Maaß cusp form for the Hecke congruence group $\Gamma_0(N) \leq \mathrm{SL}_n(\mathbb{Z})$, then

$$\|\phi|_\Omega\|_\infty \ll_\Omega N^\varepsilon \lambda^{\frac{n(n-1)}{8}(1-\delta_n)},$$

for some fixed compact set $\Omega \subset \mathfrak{h}^n = \mathrm{SL}_n(\mathbb{R})/\mathrm{SO}(n)$ and effectively computable $\delta_n > 0$. The local bound (3.1.1) in this case is $n(n-1)/8$.

In higher-rank, the only other bound related to the volume aspect (that is available to the author) is a bound in the depth aspect given by Hu [Hu18]. The result is stated only for automorphic forms corresponding to minimal vectors, which seem to be more suitable for the p-adic analysis of the depth aspect.

3.1.2 The main results and methods

The purpose of this article is to establish the first non-trivial hybrid sup-norm bounds for Hecke-Maaß forms in *unbounded* rank. Note that the uniformity in the volume (not just the level, but also the discriminant of the algebra) is a new feature even in the degree 2 case.

Before stating the main theorem, recall that a *locally norm-maximal* order is an order $\mathcal{O}$ such that $\mathrm{nr}(\mathcal{O}_p^\times) = \mathbb{Z}_p^\times$ for all primes p, where $\mathcal{O}_p$ is the p-adic completion of $\mathcal{O}$. This a technical assumption in the theorems below and we refer to Remark 3.1.1 for a way to remove it.

Theorem 3.1. *Let $p \geq 3$ be a prime and A a central division algebra of degree p over $\mathbb{Q}$ that is split over $\mathbb{R}$. Let $\mathcal{O} \subset A$ be a locally norm-maximal order of covolume $V := \mathrm{vol}\,\mathcal{O}^1\backslash\mathfrak{h}^p$. If ϕ is an L^2-normalised Hecke-Maaß form on $\mathcal{O}^1\backslash\mathfrak{h}^p$ with large eigenvalue λ, then*

$$\|\phi\|_\infty \ll \lambda^{\frac{p(p-1)}{8}-\delta_1+\varepsilon} V^{-\delta_2+\varepsilon}, \tag{3.1.2}$$

where the savings can be taken to be $\delta_1 = (16p^3)^{-1}$ and $\delta_2 = (8p^3(p-1))^{-1}$, and the implied constant depends on p and ε.

The so-called convexity bound is given by setting $\delta_1 = \delta_2 = 0$ in the exponents of (3.1.2), so that we obtain subconvex bounds in both aspects simultaneously. The savings in the exponents also have the advantage of being explicit and polynomial in the degree. In fact, the proof shows a marginally stronger bound in the spectral aspect, but we have simplified the exponent for aesthetic reasons.

There are certain assumptions that can be removed in the statement of Theorem 3.1, as well as in the theorems below. These include allowing automorphic forms that *transform under characters* and orders that are *not locally norm-maximal*, as noted in Remark 3.1.1 after discussing the methods of proof.

The reason we only work over $\mathbb{Q}$ in Theorem 3.1 is explained in Remark 3.2.1. It is no loss of generality, but in fact the only field relevant to our problem for the space $\mathfrak{h}^p$ with $p > 2$. For the case $p = 2$, we prove a bound over totally real number fields.

Theorem 3.2. *Let A be an indefinite division quaternion algebra over a totally real number field F. Let $\mathcal{O} \subset A$ be a locally norm-maximal order of covolume $V := \mathrm{vol}\,\mathcal{O}^1\backslash\mathfrak{h}^2$. If ϕ is an L^2-normalised Hecke-Maaß form on $\mathcal{O}^\times\backslash\mathfrak{h}^2$ with large eigenvalue λ, then*

$$\|\phi\|_\infty \ll \lambda^{\frac{1}{4}-\delta_1+\varepsilon} \cdot V^{-\delta_2+\varepsilon}, \tag{3.1.3}$$

where the savings can be taken to be $\delta_1 = 1/120$ and $\delta_2 = 1/30$, and the implied constant depends on F and ε.

Considering the literature on the indefinite quaternion algebra case, the novelty of the bound is the uniformity in the discriminant of the algebra, and

thus in the full volume of the quotient, while previous bounds only considered the level of the order. It should also be noted that this is the first hybrid bound appearing in print (accessible to the author). However, as a compromise for being hybrid, we should remark that our bounds in Theorem 3.2 are weaker than those of [Tem10] (i.e. $N^{-1/24}$, where we recall that $V \asymp_A N^{1+o(1)}$) or [IS95] (i.e. $\lambda^{1/4-1/24}$) when isolating the relevant aspect.

In the case of quaternion algebras over the rational numbers, a sup-norm bound uniform in the full volume of the hyperbolic surface, in particular uniform in the discriminant, was also recently and independently achieved by Khayutin, Nelson, and Steiner [KNS22]. Among many impressive results, their theorem in the corresponding setting of Theorem 3.2 is that $\|\phi\|_\infty \ll_{\lambda,\varepsilon} V^{-1/4+\varepsilon}$. This is a major improvement over other bounds in the literature, however only for Eichler orders and newforms, and without uniformity in the spectral aspect. Their proof uses different novel methods which, though very powerful in degree 2, do not seem to generalise easily to the higher rank situation.

To give the proof of our main theorems some historical context, we note that it shares some strategies with Section 6 of [Tem10] and Section 2.4 of [Sah20], the latter being inspired by the former. In turn, the proof in [Tem10] was inspired by the work of Silberman and Venkatesh [SV19] on quantum unique ergodicity, which has many similarities to the sup-norm problem. As [SV19] treats more generally division algebras of prime degree, it seems only natural that the previous sup-norm problem arguments should extend to this setting, and this is achieved in these notes. The main new idea for obtaining hybrid bounds is a systematic use of conjugation invariant functions, such as the characteristic polynomial, as we explain below.

We observe that going beyond prime degrees seems to require a new general strategy, as was also noted by Silberman and Venkatesh. The present paper provides a first step in this direction and the methods suffice for treating certain orders of Eichler type (see the discussion in Section 3.8), in arbitrary odd degree. We work with orders $\mathcal{O}$ of type $\mathcal{O}_0(N)$, by which we mean that, at unramified primes p, the completion $\mathcal{O}_p$ is of the form

$$\mathcal{O}_0(N)_p = \left\{\gamma \in M_n(\mathbb{Z}_p) \mid \text{last row of } \gamma \equiv (0,\dots,0,*) \bmod N\mathbb{Z}_p\right\},$$

up to conjugation.

Theorem 3.3. *Let $n \geq 3$ be an odd integer and A a central division algebra of degree n over $\mathbb{Q}$ that is split over $\mathbb{R}$. Let $\mathcal{O} \subset A$ be an order of type $\mathcal{O}_0(N)$ and let $V := \operatorname{vol} \mathcal{O}^1\backslash\mathfrak{h}^n$ be its covolume. If ϕ is an L^2-normalised Hecke-Maaß form on $\mathcal{O}^1\backslash\mathfrak{h}^n$ with large eigenvalue λ, then*

$$\|\phi\|_\infty \ll_A \lambda^{\frac{n(n-1)}{8}-\delta_1+\varepsilon} V^{-\delta_2+\varepsilon}, \tag{3.1.4}$$

where the savings can be taken to be $\delta_1 = (8n^3)^{-1}$ and $\delta_2 = (4n^3(n-1))^{-1}$, and the implied constant depends on n, ε, and the discriminant of A.

Comparing the bound above with the one in Theorem 3.1, the reader may remark that the savings are stronger by a factor of two. This is because, for the special orders to which it applies, the argument can handle counting in the spectral and level aspect at the same time, whilst for Theorem 3.1 we interpolate two different bounds, whence the halving of the savings.

Generalising Theorem 3.3 to handle any type of orders only requires improving the counting argument in the volume aspect. The uniform counting argument in the spectral aspect, given in Section 3.5, is valid for any orders, at least in odd degree.

The following discussion of the methods applies generally to all three theorems stated above. As usual in the treatment of the sup-norm problem, the argument starts with an amplified pretrace formula. We embed the form ϕ into a basis of Hecke-Maaß forms (ϕ_j) for $L^2(\mathcal{O}^1\backslash\mathfrak{h}^p)$ and we spectrally expand an automorphic kernel with respect to this basis. This leads to an equality between a weighted sum (the spectral side) of the form

$$\sum_j A_j|\phi_j(z)|^2$$

and a sum over elements in sets constructed from the group $\mathcal{O}^1$ (the geometric side).

To choose an amplifier essentially means to find suitable non-negative weights A_j so that the contribution of ϕ is large and that of the other forms is little, hopefully negligible. These are constructed using Hecke eigenvalues. This is a problem in analysis and combinatorics (or real and p-adic analysis), and was solved for example in [BM15] for the groups $\mathrm{PGL}_n(\mathbb{R})$. Restricting to unramified places, we are also able to use the amplifier of Blomer and Maga.

After choosing an amplifier, we can then drop all but one terms and obtain a bound for $|\phi_j(z)|^2$ in terms of a sum over certain elements in $\mathcal{O}$, determined by the amplifier, which turns into a counting problem. We count elements $\gamma \in \mathcal{O}$ of norms up to some parameter L, such that the distance between z and γz is small. This is usually done quite explicitly in the non-compact case of congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$. In our case, we rely on the very rigid structure of division algebras of prime degree.

More precisely, we may assume that the elements we are counting lie in a *proper* subalgebra of A at the cost of upper bounds for the parameter L. Here we make crucial use of the degree being prime (and nowhere else in an essential way). In this case, the only proper subalgebras of A are fields, where we have better techniques available. In particular, it suffices to count ideals and units with certain conditions in the ring of integers, and the resulting number of elements is essentially best-possible. Thus, the bound in the theorem is dictated by how large we can take L to be. We note that the scarceness of subalgebras in the prime degree case is also the reason why Silberman and Venkatesh prove their results in this setting (see Section 1.3 in [SV19]).

There are several difficulties to be overcome in the more general setting of this paper. One of them is that the counting problem we obtain when bounding $|\phi_j(z)|^2$ depends on the point z. This point only needs to vary along a compact region, but to make the bounds uniform in the volume, we need to have a good grasp on how the counting depends on z. In this paper we use conjugation invariant functions to completely remove z from the counting problem.

Another difficulty is finding a way to incorporate the discriminant of the algebra, which is not visible in previous approaches. The crucial step is at Lemma 3.4.2, where we deduce that a certain subspace spanned by elements in our order is proper inside the whole algebra. This is done using the determinant method, that is, by proving lower and upper bounds for a well-chosen determinant which contradict each other if the determinant is non-zero and certain parameters are small enough. The new input is the use of a special type of matrix whose determinant is divisible by the discriminant. This is a particularly useful choice of matrix, since it's entries are given by certain traces, which are conjugation invariant. This is a first example of the strategy mentioned in the previous paragraph.

The other conjugation invariant function we extensively make use of is the norm. It is essential for the uniform counting in the spectral aspect and for obtaining bounds in composite degrees. Our arguments not only obtain the properness of the algebra generated by the elements we are counting, but we prove its commutativity directly in certain cases. This is based on two simple but powerful observations: that commutators $k_1k_2 - k_2k_1$ of special orthogonal matrices in arbitrary odd degree are singular; and that commutators of elements in orders of $\mathcal{O}_0(N)$-type (e.g. matrices in $\Gamma_0(N)$) have determinant divisible by N. These observations, together with the fact that the only element in a division algebra with norm 0 is 0, solve the counting problem in the spectral and level aspect simultaneously for $\mathcal{O}_0(N)$-type orders and in the spectral aspect for any order. We refer to Section 3.5 and Section 3.8 for more details.

Another problem worth mentioning concerns the counting argument for units in a field (as explained above, we must also count ideals, but this is easily done by using unique factorisation). In this article, we handle this by bounding the possibilities for the characteristic polynomial of these elements, which again is conjugation invariant. Since we are counting in a (commutative) field, this automatically bounds the number of units. This argument is valid in any degree and can be employed whenever, by other arguments, one can restrict counting to a commutative algebra.

Finally, we remark that the counting argument makes use of the discriminant $\mathrm{disc}(\mathcal{O})$ in lieu of the volume of $\mathcal{O}^1\backslash\mathfrak{h}^p$. It is well-known that these two parameters are essentially equal, as already mentioned in this introduction. This is indeed readily available for quaternion algebras, for which there is extensive literature available. Nevertheless, the author was not able to find a

direct reference for the required formulae, at least for the more general setting in this article. We generalise the quaternion algebra argument to show that $\mathrm{vol}(\mathcal{O}^1\backslash\mathfrak{h}^p) = \mathrm{disc}(\mathcal{O})^{1/2+o(1)}$ in section 3.2.1. The calculations may implicitly be present in other works on zeta functions of division algebras, but since this fact might be useful in particular in the theory of automorphic forms, we provide the details here to serve as reference.

Remark 3.1.1. The statements of the theorems presented in this introduction can be generalised by working adelically. Although this paper considers automorphic forms as classical objects, as in the theory of quantum chaos, both approaches are common in the literature (e.g. [BM15] is written classically and [Sah20] adelically).

For example, taking ϕ in the theorems stated above to be an adelic form, one could drop the assumption that $\mathcal{O}$ is locally norm maximal (see also Remark 3.3.2) and one could allow ϕ to generate a finite dimensional representation under the action of the corresponding adelisation of the unit group $\mathcal{O}^1$. Indeed, this is the approach taken in [Sah20], Theorem 1. Following the argument there, one would perform amplification using only unramified primes and the counting problem would be the same as in the classical case. Some details regarding the classical and adelic formulations of the theorems, in particular for automorphic forms transforming under a finite character of $\mathcal{O}^1$, are given in Section 1.3 of [Sah20].

The counting argument being the main novelty in this article, the classical formulation was chosen here to reduce technicalities and size. Some details on adelisation are given in Section 3.3 to aid with understanding the classical Hecke algebra.

Notation

We recall the Vinogradov notation $f(x) \ll g(x)$ for two functions f, g, meaning that $f(x) \leq C \cdot g(x)$, at least for large enough x, for some $C > 0$ called the implied or implicit constant. We shall often use the notation $g = h + O(\delta)$ for matrices g, h, meaning that each coefficient of $g - h$ is $O(\delta)$. An expression of the form $f(x) = x^{o(1)}$ is to be intepreted as $f(x) \ll_\varepsilon x^\varepsilon$ and $x^{-\varepsilon} \ll_\varepsilon f(x)$ for any $\varepsilon > 0$, where the implied constant in both bounds can depend on ε. Also, we sometimes work with more general degrees $n \in \mathbb{N}$ for the division algebra A, and restrict where necessary to prime degrees p.

Acknowledgements

I would like to thank Valentin Blomer for introducing me to this topic and for his constant support. For many enlightening discussions on these topics, I thank my colleagues Edgar Assing and Bart Michels. For their kind help with understanding their work, I would also like to thank Abhishek Saha and John Voight. Finally, the readability and accuracy of the paper were greatly

improved by the anonymous referee's comments and suggestions, for which I am grateful.

3.2 DIVISION ALGEBRAS AND ARITHMETIC SUBGROUPS

Let A be a central division algebra of degree n over $\mathbb{Q}$. Let $\mathcal{O} \subset A$ be an order, i.e. a subring with 1 that is a full $\mathbb{Z}$-lattice. Suppose that A splits over $\mathbb{R}$, meaning that there is an embedding $A \hookrightarrow M_n(\mathbb{R})$. For an element $x \in A$, the reduced norm $\mathrm{nr}(x)$ and the reduced trace $\mathrm{tr}(x)$ are given by the determinant and the trace, respectively, of its image under this embedding. The group $\mathcal{O}^1 = \{\gamma \in \mathcal{O} : \mathrm{nr}(\gamma) = 1\}$ now embeds into $\mathrm{SL}_n(\mathbb{R})$ as a cocompact arithmetic lattice (see [Mor15], Proposition 6.8.9). In particular, denoting the symmetric space of $\mathrm{SL}_n(\mathbb{R})$ by

$$\mathfrak{h}^n = \mathrm{SL}_n(\mathbb{R})/\mathrm{SO}(n),$$

then the quotient $\mathcal{O}^1\backslash\mathfrak{h}^n$ is compact.

Note that for n odd, A splits automatically over $\mathbb{R}$. Indeed, by the Albert-Brauer-Hasse-Noether theorem, all central simple algebras of finite degree over a number field are cyclic, meaning that they contain a strictly maximal subfield that is a Galois extension of $\mathbb{Q}$ of degree n (for background on these statements and the following, see [Pie82], Theorem 18.6 and sections 13.1 through 13.3). The strictly maximal subfield of A splits A and is Galois of odd degree over $\mathbb{Q}$, so it must be totally real, in particular contained in $\mathbb{R}$.

In the special case $n = 2$, A is called a quaternion algebra and we may replace the ground field $\mathbb{Q}$ by any totally real number field F. Let $[F : \mathbb{Q}] = n$ and denote by $\mathfrak{o}_F$ the ring of integers of F. For A to be split over $\mathbb{R}$, we assume that there is an embedding $\sigma_0 \in \hom(F, \mathbb{R})$ such that $A \otimes_{\sigma_0} \mathbb{R} \cong M_2(\mathbb{R})$. For all other embeddings $\sigma_0 \neq \sigma \in \hom(F, \mathbb{R})$ assume that $A \otimes_\sigma \mathbb{R} \cong \mathcal{H}(\mathbb{R})$, where $\mathcal{H}(\mathbb{R})$ is the Hamilton quaternion algebra. We may view A as embedded (diagonally) into $A_\infty \cong M_2(\mathbb{R}) \times \mathcal{H}(\mathbb{R})^{n-1}$, and similarly for the norm 1 elements,

$$A^1 \hookrightarrow \mathrm{SL}_2(\mathbb{R}) \times \mathrm{SO}(3)^{n-1}.$$

We use ϕ_0 to denote the projection onto the first component $M_2(\mathbb{R})$ and ϕ_i, $i = 1, \ldots, n-1$, to denote the projections onto the Hamiltonian components.

Generalising our setting, let $\mathcal{O}$ be an $\mathfrak{o}_F$-order. By restriction of scalars, the projection $\phi_0(\mathcal{O}^1) \subset \mathrm{SL}_2(\mathbb{R})$ of the group of units of reduced norm 1 onto the split component gives a cocompact arithmetic lattice.

Remark 3.2.1. For $\phi_0(\mathcal{O}^1)$ to be a cocompact arithmetic subgroup in the split component $\mathrm{SL}_2(\mathbb{R})$, it is important that the other components, in this case all isomorphic to SO(3), are compact (see the definition of an arithmetic group in [Mor15], Definition 5.1.19). If A is a central division algebra over a number field $F \neq \mathbb{Q}$ and $\deg(A) = n > 2$, this is not possible any more.

Indeed, the process of restriction of scalars requires us to embed A^1 into the product of its completions at all infinite places. Now a central simple

algebra over $\mathbb{R}$ is isomorphic to a matrix algebra over a division algebra by Wedderburn's theorem. By a theorem of Frobenius (see [Pie82], Corollary 13.1 c) these are either matrix algebras over $\mathbb{R}$ or over the Hamiltonians $\mathcal{H}$. Since $n > 2$, the group of norm 1 units in these algebras cannot be compact any more. Thus, the number field case in higher degree gives rise to non-compact lattices, which are outside the scope of this article.

It will be useful later to note that the tower rule holds for division algebras (also called skew fields). More precisely, the notion of vector space over a division algebra and its dimension is the same as for commutative fields. If $A' \subset A$ is a subalgebra, then A may be viewed as a vector space over A', where A' acts by multiplication from the left (or from the right, according to taste). We denote the dimension of the vector space by $\dim_{A'} A$ as usual.

Let now $A_1 \subset A_2 \subset A_3$ be division algebras. Then

$$\dim_{A_1} A_3 = \dim_{A_2} A_3 \cdot \dim_{A_1} A_2 \tag{3.2.1}$$

holds and is proven as in the commutative case (see [Coh08, Proposition 3.1.1]). Thus, if A is a finite dimensional division algebra over $\mathbb{Q}$, then the dimension over $\mathbb{Q}$ of any subalgebra of A must divide $\dim_{\mathbb{Q}} A$. Moreover, if A is central, then any subfield of A must have dimension over $\mathbb{Q}$ dividing the degree of A (see [Pie82], Corollary 13.1 a).

3.2.1 The volume approximation

For simplicity, we first assume that the ground field is $\mathbb{Q}$ and quote the relevant results for quaternion algebras over number fields at the end of the section.

Let $\mathcal{O}_m$ be a maximal order in A containing $\mathcal{O}$. Because of their lattice structure, it is useful to work with the index $[\mathcal{O}_m : \mathcal{O}]$, which we call the *level* of $\mathcal{O}$ in $\mathcal{O}_m$. Yet the volume of $\mathcal{O}^1\backslash\mathfrak{h}^n$, the relevant parameter in our sup-norm problem, is given by the volume of $\mathcal{O}_m^1\backslash\mathfrak{h}^n$ and the multiplicative index $[\mathcal{O}_m^1 : \mathcal{O}^1]$. Fortunately the two indices are related in an explicit way. For our purposes (and because the exact formulae would involve too many cases in general), it suffices to prove that they are approximately equal. The proper equalities obtained in the proof can be used together with the machinery of zeta functions and Tamagawa numbers to produce a formula for the volume of $\mathcal{O}^1\backslash\mathfrak{h}^n$ (as in [Voi21], 39.2.8).

Lemma 3.2.2. *Let A be a central simple algebra over $\mathbb{Q}$, but not a definite quaternion algebra. Let $\mathcal{O} \subset \mathcal{O}_m$ be two orders in A. If $[\mathcal{O}_m : \mathcal{O}] = N$, then $[\mathcal{O}_m^1 : \mathcal{O}^1] = N^{1+o(1)}$.*

This lemma is a generalisation of the well-know fact that the index of the Hecke congruence group $\Gamma_0(N)$ in $\mathrm{SL}_2(\mathbb{Z})$ is

$$N \prod_{p|N} (1 + 1/p).$$

In this case, A is the matrix algebra $M_2(\mathbb{Q})$, the maximal order is $M_2(\mathbb{Z})$ and $\mathcal{O}$ is the suborder of level N of integral matrices with lower left entry divisible by N.

Before proving Lemma 3.2.2, we use it together with a formula for the covolume of a maximal order to approximate the covolume of $\mathcal{O}$ by the discriminant. Recall that the *discriminant* of an arbitrary $\mathbb{Z}$-order $\mathcal{O}$ is defined as the ideal $\mathrm{disc}(\mathcal{O}) \subset \mathbb{Z}$ generated by the set

$$\{\det(\mathrm{tr}(x_i \cdot x_j))_{1\leq i,j\leq n} \mid x_i \in \mathcal{O}\}.$$

By abuse of notation, we also denote a positive generator of this ideal by $\mathrm{disc}(\mathcal{O})$. For more details on discriminants, see Section 10 in [Rei75].

Proposition 3.2.3. *Let A be a central simple algebra over $\mathbb{Q}$, but not a definite quaternion algebra, and let $\mathcal{O}$ be an order in A. Then* $\mathrm{vol}(\mathcal{O}^1\backslash\mathfrak{h}^n) = \mathrm{disc}(\mathcal{O})^{1/2+o(1)}$.

Proof. Let $\mathcal{O}_m$ be a maximal order containing $\mathcal{O}$. By Theorem 3.7 in [Kle00], we can approximate

$$\mathrm{vol}(\mathcal{O}_m^1\backslash\mathfrak{h}^n) = \mathrm{disc}(\mathcal{O}_m)^{\frac{1}{2}+o(1)}.$$

By Lemma 15.2.15 in [Voi21], we also have

$$\mathrm{disc}(\mathcal{O}) = [\mathcal{O}_m : \mathcal{O}]^2 \cdot \mathrm{disc}(\mathcal{O}_m).$$

Together with Lemma 3.2.2, we now obtain the claimed approximation since $\mathrm{vol}(\mathcal{O}^1\backslash\mathfrak{h}^n) = \mathrm{vol}(\mathcal{O}_m^1\backslash\mathfrak{h}^n) \cdot [\mathcal{O}_m^1 : \mathcal{O}^1]$. □

The proof of Lemma 3.2.2 generalises the argument for quaternion algebras in [Voi21], Lemma 26.6.7, which in turn follows an argument of Körner. We provide here full details for the sake of completeness.

The first ingredient is the strong approximation theorem (see Kneser's article in [BM66]), which allows us to reduce the statement to a local one. We denote by $A_p = A \otimes \mathbb{Q}_p$ and $\mathcal{O}_p = \mathcal{O} \otimes \mathbb{Z}_p$ the completions at a prime p. For all but finitely many primes p, the completion A_p is split, i.e. $A_p \cong \mathcal{M}_n(\mathbb{Q}_p)$ (see Proposition 18.5 coupled with Corollary 17.10.a in [Pie82]). Additionally, for all but finitely many primes p, the completion $\mathcal{O}_p$ is a maximal order of A_p (see Lemma 10.4.4 in [Voi21]). In particular, at these primes we have $\mathcal{O}_{m,p} = \mathcal{O}_p$. The primes where equality does not hold will be referred to as *ramified*.

We embed $\mathcal{O}$ diagonally into $\hat{\mathcal{O}} = \prod_p \mathcal{O}_p$ and, similarly, $\mathcal{O}^1$ into $\hat{\mathcal{O}}^1 = \prod_p \mathcal{O}_p^1$, where p runs over all prime numbers. Then strong approximation implies that $\mathcal{O}^1$ is dense in $\hat{\mathcal{O}}^1$ (see [Voi21], Corollary 18.5.14, and more generally [Kle00], Theorem 4.4). Explicitly, if S a set of finite places, $a_p \in \mathcal{O}_p^1$ and t_p for each $p \in S$, then we can find $x \in \mathcal{O}^1$ such that

$$x \equiv a_v \pmod{p^{t_v}\mathcal{O}_p} \quad (p \in S).$$

Lemma 3.2.4. *For two orders $\mathcal{O} \subset \mathcal{O}_m$ as above, the level and the index of the unit groups can be computed locally, that is,*

$$[\mathcal{O}_m : \mathcal{O}] = \prod_p [\mathcal{O}_{m,p} : \mathcal{O}_p] \quad \text{and} \quad [\mathcal{O}_m^1 : \mathcal{O}^1] = \prod_p [\mathcal{O}_{m,p}^1 : \mathcal{O}_p^1].$$

Proof. Note first that the products contain only finitely many factors not equal to 1, as in the remarks above. Next, we start the proof for the unit groups. The claim follows by showing that the map

$$\mathcal{O}_m^1/\mathcal{O}^1 = \prod_p \mathcal{O}_{m,p}^1/\mathcal{O}_p^1$$

is bijective.

Injectivity follows by noting that $\bigcap_p \mathcal{O}_p = \mathcal{O}$. Surjectivity follows by strong approximation. Indeed, let $(a_p) \in \prod_p \mathcal{O}_{m,p}^1$. Choose an integer N such that $N\mathcal{O}_{m,p} \subset p\mathcal{O}_p$ for all ramified primes p. Strong approximation supplies us with an element $b \in \mathcal{O}_m^1$ such that $b \in a_p + N\mathcal{O}_{m,p}$. Thus $b = a_p \cdot u_p$, where $u_p \in 1 + N\mathcal{O}_{m,p}$, so that $u_p \in \mathcal{O}_p^1$.

The proof for the factorisation of the level is similar, where the corresponding strong approximation theorem is the Chinese Remainder Theorem. □

In the following we work with the localised orders at a prime p, which we suppress in notation for simplicity. We now remove the condition on the norm to work with the full group of units. We have the short exact sequence

$$0 \to \mathcal{O}^1 \to \mathcal{O}^\times \to \mathrm{nr}(\mathcal{O}^\times) \to 0,$$

and similarly for $\mathcal{O}_m$. By defining a non-canonical bijection[2]

$$\mathcal{O}_m^1/\mathcal{O}^1 \times \mathrm{nr}(\mathcal{O}_m^\times)/\mathrm{nr}(\mathcal{O}^\times) \to \mathcal{O}_m^\times/\mathcal{O}^\times,$$

we obtain that

$$|\mathcal{O}_m^1/\mathcal{O}^1| \cdot |\mathrm{nr}(\mathcal{O}_m^\times)/\mathrm{nr}(\mathcal{O}^\times)| = |\mathcal{O}_m^\times/\mathcal{O}^\times|.$$

Lemma 3.2.5. *For $\mathbb{Z}_p$-orders $\mathcal{O} \subset \mathcal{O}_m$, we have*

$$[\mathcal{O}_m^\times : \mathcal{O}^\times] = [\mathcal{O}_m : \mathcal{O}] \cdot p^{o(1)}.$$

Proof. The proof starts as in Lemma 26.6.7 in [Voi21]. Let n be such that $p^n\mathcal{O}_m \subset p\mathcal{O}$. Note that $1 + p\mathcal{O} \subset \mathcal{O}^\times$ by the convergence of the geometric series. We now have

$$[\mathcal{O}_m^\times : \mathcal{O}^\times] = \frac{[\mathcal{O}_m^\times : 1 + p\mathcal{O}_m] \cdot [1 + p\mathcal{O}_m : 1 + p^n\mathcal{O}_m]}{[\mathcal{O}^\times : 1 + p\mathcal{O}] \cdot [1 + p\mathcal{O} : 1 + p^n\mathcal{O}_m]}.$$

[2]Note that the groups in question are not abelian and not necessarily normal, so that we cannot apply the snake lemma directly.

For $\alpha, \beta \in 1 + p\mathcal{O}$, we have $\alpha\beta^{-1} \in 1 + p^n\mathcal{O}_m$ if and only if $\alpha - \beta \in p^n\mathcal{O}_m$, so that

$$[1 + p\mathcal{O} : 1 + p^n\mathcal{O}_m] = [p\mathcal{O} : p^n\mathcal{O}_m] = [\mathcal{O} : p^{n-1}\mathcal{O}_m],$$

and similarly for $\mathcal{O}_m$. If follows that

$$\frac{[1 + p\mathcal{O}_m : 1 + p^n\mathcal{O}_m]}{[1 + p\mathcal{O} : 1 + p^n\mathcal{O}_m]} = \frac{[\mathcal{O}_m : p^{n-1}\mathcal{O}_m]}{[\mathcal{O} : p^{n-1}\mathcal{O}_m]} = [\mathcal{O}_m : \mathcal{O}].$$

To further compute the factors $[\mathcal{O}^\times : 1 + p\mathcal{O}]$, we employ the strategy in [Voi21], Lemma 24.3.12, of introducing the Jacobson radical $\operatorname{rad}\mathcal{O} =: J$. We have $p\mathcal{O} \subset J$ and there is an integer r such that $J^r \subset p\mathcal{O}$ (see [Rei75], Theorem 6.13), which we assume to be minimal. Thus $1 + J \subset \mathcal{O}^\times$ and we obtain a filtration

$$\mathcal{O}^\times \supset 1 + J \supset 1 + J^2 \supset \ldots \supset 1 + p\mathcal{O} \supset 1 + J^r.$$

Being kernels, all subgroups are normal inside their parent groups. It follows that

$$[\mathcal{O}^\times : 1 + p\mathcal{O}] = |\mathcal{O}^\times/1 + J| \cdot |1 + J/1 + J^2| \cdots |1 + J^{r-1}/1 + p\mathcal{O}|.$$

On the additive side, we also have a filtration $\mathcal{O} \supset J \supset \ldots \supset p\mathcal{O}$ and the quotients $\mathcal{O}/J, J/J^2, \ldots, J^{r-1}/p\mathcal{O}$ are $\mathbb{F}_p$-algebras. If R is the rank of $\mathcal{O}$, then

$$p^R = [\mathcal{O} : p\mathcal{O}] = [\mathcal{O} : J][J : J^2] \cdots [J^{r-1} : p\mathcal{O}].$$

We now reduce the multiplicative indices to the additive ones. Indeed $1+J/1+J^2 \cong J/J^2$, and similarly for all powers of J, and $1+J^{r-1}/1+p\mathcal{O} \cong J^{r-1}/p\mathcal{O}$, since $J^{2(r-1)} \subset p\mathcal{O}$ (at least for $r > 1$; the case $r = 1$ is simpler and can be done directly). Therefore,

$$|1 + J/1 + J^2| = |J/J^2|, \ldots, |1 + J^{r-1}/1 + p\mathcal{O}| = |J^{r-1}/p\mathcal{O}|.$$

Thus,

$$[\mathcal{O}^\times : 1 + p\mathcal{O}] = p^R \frac{[\mathcal{O}^\times/1 + J]}{[\mathcal{O} : J]}.$$

Now one can easily see that $\mathcal{O}^\times/1 + J \cong (\mathcal{O}/J)^\times$. The reason for working with the Jacobson radical is that $\mathcal{O}/J$ is a semisimple $\mathbb{F}_p$-algebra, meaning that

$$\mathcal{O}/J \cong M_{d_1}(A_1) \times \cdots \times M_{d_l}(A_l),$$

for some finite division algebras A_i over $\mathbb{F}_p$. Since finite division algebras are fields by Wedderburn's theorem, one can check by counting that $|GL_{d_i}(A_i)| = |M_{d_i}(A_i)|^{1-o(1)}$ (this can be made precise, but the approximation is sufficient for our purposes).

Since $\mathcal{O}$ and $\mathcal{O}_m$ have the same rank, it follows that

$$\frac{[\mathcal{O}_m^\times : 1 + p\mathcal{O}_m]}{[\mathcal{O}^\times : 1 + p\mathcal{O}]} = p^{o(1)}.$$

This finishes the proof. □

For the groups of norms, we note that $\mathbb{Z}_p^\times \subset \mathcal{O}^\times$, and so $(\mathbb{Z}_p^\times)^n \leq \mathrm{nr}(\mathcal{O}^\times) \leq \mathbb{Z}_p^\times$. Now $[\mathbb{Z}_p^\times : \mathbb{Z}_p^\times)^n] \ll n$ by Korollar 5.8 in [Neu92]. This contributes to the global index by $n^{\omega(N)} \ll_n d(N) \ll N^\varepsilon$, where $\omega(N)$ is the number of different primes dividing the level N, $d(N)$ is the number of divisors of N, and ε is any positive real number. This completes the proof of Lemma 3.2.2.

A similar argument, taking into account all of the different embeddings, would also apply for division algebras over number fields. Since we are only interested in quaternion algebras for totally real number fields, we simply note an approximate version of Main Theorem 39.1.8 in [Voi21] (it is given for locally norm-maximal orders; see Remark 39.1.12 for the general case, which is the same observation we make in the previous paragraph). Let F be a totally real number field with ring of integers $\mathfrak{o}_F$. In this case, we recall that the discriminant $\mathrm{disc}(\mathcal{O})$ is an $\mathfrak{o}_F$-ideal and we denote by $N_F(\mathrm{disc}(\mathcal{O}))$ its norm.

Proposition 3.2.6. *Let A be an indefinite quaternion algebra over F and let $\mathcal{O}$ be an order in A. Then* $\mathrm{vol}(\mathcal{O}^1\backslash\mathfrak{h}^2) = N_F(\mathrm{disc}(\mathcal{O}))^{1/2+o(1)}$*, where the implicit constant depends on F.*

3.3 THE AMPLIFIED PRETRACE FORMULA

In this section we again assume that the ground field is $\mathbb{Q}$. The additional technicalities involved in the case of quaternion algebras over number fields are described in Section 3.7.

The space of automorphic forms $L^2(\mathcal{O}^1\backslash\mathfrak{h}^n)$ has a discrete decomposition, admitting a basis of Hecke-Maaß forms $(\phi_j)_{j\in\mathbb{N}}$, that is, eigenfunctions of the algebra of invariant differential operators and of the Hecke algebra (described below). Denote the spectral parameters of each form ϕ_j by $\mu_j \in \mathfrak{a}_{\mathbb{C}}^*$, where $\mathfrak{a}_{\mathbb{C}}^*$ is the complexified dual of the Lie algebra of the diagonal torus A in $\mathrm{SL}_n(\mathbb{R})$. Recalling the main goal of this paper we note that for bounding an individual automorphic form ϕ with spectral parameter μ, we may assume that ϕ is part of the basis (ϕ_j).

For setting up the pretrace formula, we follow the notation in [BM16]. In particular, to any $\lambda \in \mathfrak{a}_{\mathbb{C}}^*$ we attach the quantity $D(\lambda)$ defined in [BM16, (1.2)]. To ease notational clutter due to the fact that the discriminant of $\mathcal{O}$ is denoted by D, we put $S(\lambda) := D(\lambda)$ in this paper. As in Section 2, ibid., we denote $\mu^* := \Re\mu \in \mathfrak{a}^*$ and assume that $\left\|\mu^*\right\|$ is sufficiently large. If λ_ϕ is the Laplace eigenvalue of ϕ, then

$$S(\mu^*) \ll 1 + \lambda_\phi^{n(n-1)/4}, \tag{3.3.1}$$

as in [BM16, (2.3)]. The bounds we obtain in the remainder of the paper use $S(\mu^*)$ instead of the eigenvalue since they are slightly improved this way (we only stated the main theorems using the eigenvalue for simplicity), but also to simplify some exponents.

As usual we denote $K = \mathrm{SO}(n)$. For a function $f \in C_c^\infty(K\backslash G/K)$, the pretrace formula states that

$$\sum_{j\in\mathbb{N}} \tilde{f}(\mu_j)\phi_j(z)\overline{\phi_j(z')} = \sum_{\gamma\in O^1} f(z^{-1}\gamma z'),$$

for $z, z' \in G$, where $\tilde{f}$ is the spherical transform of f.

To calibrate the pretrace formula for our distinguished function ϕ with spectral parameter μ, Blomer and Maga (e.g. see [BM16], Section 3) show that we can find $f_\mu \in C_c^\infty(K\backslash G/K)$ such that the spherical transform $\tilde{f}$ satisfies $\tilde{f}_\mu(\lambda) \geq 0$ for all possible spectral parameters λ and $\tilde{f}_\mu(\mu) \geq 1$.

Remark 3.3.1. We may use the same test function f_μ as Blomer and Maga, since the relevant spectral parameters of automorphic forms for the division algebra A depend only on the archimedean completion, more precisely on the Lie group $\mathrm{SL}_n(\mathbb{R})$. These are included in the set defined in (2.2) of [BM16] by the general theory of joint eigenfunctions and spherical functions on symmetric spaces (see [Hel84], in particular Chapter IV, Section 1, 2, and 8). Generally, the choice of test function f_μ is a local problem at the archimedean place, while applying Hecke operators as we do below is a choice of test function at the finite places (the reader comfortable with the adelic theory of automorphic forms could see Section 4.1 of [Sah20] for the adelic treatment of amplification).

In fact, using bounds of Blomer and Pohl [BP16, Sect. 6], we can also assume certain decay properties of f_μ. More precisely, we can assume the diameter of the support of f_μ to be bounded by any positive constant depending on the (fixed) degree n as we prefer. Moreover, if d denotes the invariant distance function on $\mathfrak{h}^n$, then we have the bound

$$f_\mu(g) \ll S(\mu^*)(1 + S(\mu^*)^{\frac{2}{n(n-1)}} \cdot d(g, 1))^{-1/2}, \tag{3.3.2}$$

which is easily implied by [BM16, (2.4)].

To further amplify the contribution of ϕ in the pretrace formula, Blomer and Maga also construct a general amplifier using Hecke operators (see [BM15], Section 6) for $\mathrm{SL}_n(\mathbb{Z})$. This amplifier applies in our situation as well, as long as we only use unramified places. To explain this statement, we sketch below some facts about the Hecke algebra, for which we note our assumption that the ground field is $\mathbb{Q}$.

First, we define the group U_O as

$$U_O = \mathrm{GL}_n^+(\mathbb{R}) \times \prod_p O_p^\times$$

Note that

$$O^1 = U_O \cap A^\times.$$

Next, as can be seen from Lemma 3.2.4 for instance, the primes p that divide $D := \mathrm{disc}(\mathcal{O})$ are exactly the primes at which A ramifies or $\mathcal{O}_p$ is not maximal, and we call these primes *ramified*.

From now on we assume that $\mathcal{O}$ is *locally norm-maximal*, meaning that $\mathrm{nr}(\mathcal{O}_p^\times) = \mathbb{Z}_p^\times$ for all primes p. This assumption, explained in Remark 3.3.2, is satisfied, for example, by any maximal orders or intersection of two maximal orders, i.e. Eichler orders. We define the semigroup $S_\mathcal{O}$ inside the adelisation $A_\mathbb{A}^\times$ by

$$S_\mathcal{O} = \left(\mathrm{GL}_n^+(\mathbb{R}) \times \prod_p S_p\right) \cap A_\mathbb{A}^\times,$$

where $S_p = \{\alpha \in \mathcal{O}_p : \mathrm{nr}(\alpha) \neq 0\}$ for $p \nmid D$ and $S_p = \mathcal{O}_p^\times$ for $p \mid D$. This distinction means that we only consider the *unramified* Hecke algebra. Finally, let

$$\Delta_\mathcal{O} = S_\mathcal{O} \cap \mathcal{O}.$$

We can now define the *classical* Hecke algebra $R(\mathcal{O}^1, \Delta_\mathcal{O})$, which is generated by double cosets of the form $\mathcal{O}^1 \xi \mathcal{O}^1$, where $\xi \in \Delta_\mathcal{O}$, and similarly the *adelic* Hecke algebra $R(U_\mathcal{O}, S_\mathcal{O})$. For more details, see [Miy89], Sections 2.7 and 5.3.

The adelic point of view is advantageous since we automatically obtain a factorisation of $R(U_\mathcal{O}, S_\mathcal{O})$ as the tensor product $\bigotimes_p R(\mathcal{O}_p^\times, S_p)$ of the local Hecke algebras. Fortunately in our case, there is essentially nothing lost in translation between the classical and the adelic Hecke algebra (both unramified). Indeed, they are isomorphic under the simple correspondence $\mathcal{O}^1 \xi \mathcal{O}^1 \mapsto U_\mathcal{O} \xi U_\mathcal{O}$. This can be seen by carefully applying the argument in the proof of Theorem 5.3.5 in [Miy89]. The proof makes crucial use of approximation theorems.

Remark 3.3.2. The property of being locally norm-maximal implies that the idelic quotient defined by $\mathcal{O}$ has only one connected component. In particular, the dictionary between classical automorphic forms and adelic forms is simpler. For many works in the literature (see for instance the use of Eichler orders in [Tem10], and [SV19], Remark 6.3.1), this is a common "cosmetic" assumption on orders.

Indeed, our counting arguments in prime degree, the key new ideas in this paper, simply make no use of this assumption. Removing it is merely a matter of working directly with adelic automorphic forms and the full unramified Hecke algebra. This is the approach in [Sah20], where the generalisation of Templier's argument for Eichler orders [Tem10] to arbitrary orders is transparent.

In fact, even working classically, the pretrace inequality (3.3.3) below and Remark 3.3.4 are still true for any order $\mathcal{O}$, at least conditionally on a suitable Riemann hypothesis. We sketch this technicality in this paragraph, which should be read preferably after going through the proof of Theorem 3.1. More precisely, to adjust the definition of the Hecke algebra, one may need to impose additionally that $S_p = \mathcal{O}_p^\times$ for all primes p that are not n-th powers modulo D (supposing for simplicity that $(D, n) = 1$). This is the "worst-case" scenario,

since we have the inclusions $(\mathbb{Z}_p^\times)^n \subset \mathrm{nr}(\mathcal{O}_p^\times) \subset \mathbb{Z}_p^\times$. Under this assumption, the proof of Theorem 5.3.5 in [Miy89] goes through. We are now left with a potentially smaller set of primes $\mathcal{P}$, defined just before (3.3.3), where we also assume that these primes are n-th powers modulo D. An application of the generalised Riemann hypothesis and the Chebotarev density theorem would then imply that $|\mathcal{P}| \gg_n L^{1-\varepsilon}D^{-\varepsilon}$, even for L a small power of D, as taken in the proof of the main theorem.

Now at unramified primes p, the local Hecke algebra $R(\mathcal{O}_p^\times, S_p)$ is isomorphic to the Hecke algebra of $\mathrm{GL}_n(\mathbb{Z}_p)$. Therefore, we can use the same Hecke operators as Blomer and Maga.

For $m \in \mathbb{Z}$, let

$$\mathcal{O}(m) := \{\gamma \in \mathcal{O} \mid \mathrm{nr}(\gamma) = m\}.$$

Let $L > 5$ be a parameter and $\mathcal{P}$ be the set of primes in $[L, 2L]$ that are unramified. We have the pretrace inequality (see [BM16], (2.5))

$$|\mathcal{P}|^2 \cdot |\phi(z)|^2 \ll |\mathcal{P}| \cdot S(\mu^*) + \sum_{\nu=1}^{n} \sum_{l_1, l_2 \in \mathcal{P}} \frac{1}{L^{(n-1)\nu}} \sum_{\gamma \in \mathcal{O}(l_1^\nu l_2^{(n-1)\nu})} |f_\mu(z^{-1}\tilde{\gamma}z)|, \quad (3.3.3)$$

where $\tilde{\gamma} = \gamma/\mathrm{nr}(\gamma)^{1/n} \in \mathrm{SL}_n(\mathbb{R})$. Note that the determinantal divisors in [BM15] do not seem to easily translate into our setting, yet the norm does, as one can easily check using the explicit isomorphism between the classical and adelic Hecke algebras above. Although these additional conditions are very important in the work of Blomer and Maga, we are able to solve the counting problem described below using only the condition on the norm.

Since f_μ has compact support, let $0 < \rho \ll 1$ be such that $f_\mu(g) = 0$ if $d(g, 1) > \rho$, where d is the invariant distance function on $\mathfrak{h}^n$. Using the bound $f_\mu \ll S(\mu^*)$ from (3.3.2), we may obtain an explicit bound for $\phi(z)$ from the pretrace inequality by counting the number of elements $\gamma \in \mathcal{O}(m)$ such that $d(z, \tilde{\gamma}z) < \rho$. We correspondingly define in general

$$\mathcal{O}(m; z, \delta) = \{\gamma \in \mathcal{O} : \mathrm{nr}(\gamma) = m, d(z, \tilde{\gamma}z) < \delta\}.$$

Remark 3.3.3. We note that the compact support of f_μ can be assumed to be small enough in terms of the degree n, i.e. $\rho \ll_n 1$ with an implicit constant of our choice, since we are allowing all implicit constants to depend on n. This follows from a quick inspection of the technique in [BP16].

A more careful use of (3.3.2) gives us a saving in the spectral aspect, i.e. in $S(\mu^*)$, at least if $d(z, \tilde{\gamma}z) > \delta$ for $\delta > 0$ large enough in terms of $S(\mu^*)$. For all other γ appearing in the sum, we trivially bound $f_\mu(z^{-1}\tilde{\gamma}z) \ll S(\mu^*)$ as above. We therefore split the sum using the parameter δ as above and obtain the

inequality

$$|\mathcal{P}|^2 \cdot |\phi(z)|^2 \ll S(\mu^*)\big(|\mathcal{P}| + \sum_{\nu=1}^{n} \sum_{l_1,l_2 \in \mathcal{P}} \frac{1}{L^{(n-1)\nu}} \#\mathcal{O}(l_1^\nu l_2^{(n-1)\nu}; z, \delta)$$
$$+ S(\mu^*)^{\frac{-1}{n(n-1)}} \cdot \delta^{-\frac{1}{2}} \sum_{\nu=1}^{n} \sum_{l_1,l_2 \in \mathcal{P}} \frac{1}{L^{(n-1)\nu}} \#\mathcal{O}(l_1^\nu l_2^{(n-1)\nu}; z, \rho)\big). \quad (3.3.4)$$

Remark 3.3.4. Note that $|\mathcal{P}| \gg L^{1-\varepsilon} \cdot D^{-\varepsilon}$, at least for L large enough. This follows from the prime number theorem and because the number of ramified primes we leave out is bounded by $\tau(D) \ll D^{\varepsilon}$.

To obtain a hybrid bound, we need to count elements in $\mathcal{O}(m; z, \delta)$ uniformly in $\mathrm{disc}(\mathcal{O})$ and $S(\mu^*)$. To deduce anything about the sup-norm of ϕ, the counting must also be done uniformly in z, at least in a fundamental domain for $\mathcal{O}^1$. Though compact, this fundamental domain grows with the volume.

In [Sah20], the problem of uniformity in z was resolved by having a counting argument that only depends on the index of $\mathcal{O}$ inside a maximal order $\mathcal{O}_m$. Saha was then able to conjugate z into a fixed fundamental domain for $\mathcal{O}_m$ and work with a conjugated order with the same index. The implicit constants in the bounds would then possibly depend on the particular choice of fundamental domain for the maximal order.

Our argument does not require a choice of maximal order and the bounds are uniform on the whole generalised upper half plane. This is done by counting using only traces and norms, which are conjugation invariant.

More precisely, the condition $d(z, \tilde{\gamma}z) \ll \rho$ implies that $z^{-1}\tilde{\gamma}z = k + O(\rho)$ for some $k \in \mathrm{SO}(n)$, at least for $\rho \ll 1$ small. Indeed, using the Cartan decomposition, we can write any $g \in \mathrm{SL}_n(\mathbb{R})$ as $g = k_1 \exp(C(g))k_2$, where $k_1, k_2 \in \mathrm{SO}(n)$ and $C(g)$ is diagonal with vanishing trace. Then $d(g, 1) = \|C(g)\|_2$, where we view $C(g)$ as a vector in $\mathbb{R}^n$. The claim follows by writing the exponential as a power series.

On the new condition, as an example, applying the trace directly already provides a bound on $\mathrm{tr}(\tilde{\gamma})$ where any dependence on z is completely removed, noting that orthogonal matrices are bounded. This conjugation invariant approach is used throughout the counting argument. Another example is given in the subsection below, where we derive the so-called convexity bound for the sup-norm automorphic forms in certain cases.

3.3.1 The baseline bound

The benchmark bound for the sup-norm of automorphic forms that one seeks to improve can be usually obtained by using the pretrace formula without the

amplifier. It follows easily from the properties of f_μ that

$$|\phi(z)|^2 \leq \sum_{\gamma \in \mathcal{O}^1} f_\mu(z^{-1}\gamma z) \ll$$
$$S(\mu^*) \cdot \#\{\gamma \in \mathcal{O}^1 : z^{-1}\gamma z = k + O(\rho), \text{ for some } k \in \mathrm{SO}(n)\}.$$

Suppose that n is odd. Then any degree n special orthogonal matrix must have 1 as an eigenvalue, meaning that $\det(k-1) = 0$. Thus, if $\gamma \in \mathcal{O}^1$ and $z^{-1}\gamma z = k + O(\rho)$, then we may subtract the identity matrix and apply the determinant to obtain

$$\mathrm{nr}(\gamma - 1) = \det(k - 1 + O(\rho)) = O_n(\rho).$$

As mentioned in Remark 3.3.3, we can take ρ as small as we wish in terms of n. Since $\gamma - 1 \in \mathcal{O}$, it follows by integrality of the norm that $\mathrm{nr}(\gamma - 1) = 0$. Since A is a division algebra, this implies that $\gamma = 1$. Therefore, there is only one term appearing on the geometric side of the pretrace formula formula and using the inequality (3.3.2) we obtain

$$|\phi(z)| \ll S(\mu^*)^{\frac{1}{2}},$$

where the implied constant depends at most on n. This is the convexity bound (recall also the bound 3.3.1).

This strategy cannot succeed when n is even, if only for the simple observation that -1 lies in $\mathcal{O}^1$. Some ad-hoc arguments involving the classification of motions suffice for the case $n = 2$, but the author was not able to find a general argument for all even n. This is part of the reason why this paper mainly deals with algebras of odd degree, besides quaternion algebras.

3.4 COUNTING IN THE DISCRIMINANT ASPECT

For simplicity, we first describe the counting argument over $\mathbb{Q}$ and adjust it where necessary for quaternion algebras over number fields in the next subsection. From now on, assume that $\deg(A) = p \geq 3$, a prime. Let δ be a positive real number, which we assume to be uniformly bounded, e.g. by 1. As explained in the previous section, we are interested in bounding the cardinality of

$$\mathcal{O}(m; z, \delta) = \{\gamma \in \mathcal{O} : \mathrm{nr}(\gamma) = m, d(z, \tilde{\gamma}z) < \delta\},$$

where $\tilde{\gamma} = \gamma/\mathrm{nr}(\gamma)^{1/p}$. The condition $d(z, \tilde{\gamma}z) < \delta$ is equivalent to $z^{-1}\tilde{\gamma}z = k + O(\delta)$ for some $k \in \mathrm{SO}(p)$.

To motivate the following lemmata, we recall the tower rule (3.2.1) for division algebras. Especially for prime degree p, this severely restricts the possible dimensions of subalgebras in A. If one can show that the subalgebra

generated by the elements we are counting is proper, then the tower rule drastically reduces the dimension of the counting problem, automatically. In our case, the subalgebra will actually be commutative, which is crucial in our argument. To show properness in the first place, we use a version of the determinant method, for which we need good control over a basis of a vector space.

Lemma 3.4.1. *The $\mathbb{Q}$-algebra generated by $\bigcup_{1\leq m\leq L} \mathcal{O}(m; z, \delta)$ is contained in the $\mathbb{Q}$-vector space spanned by $\bigcup_{1\leq m\leq L^{2p-2}} \mathcal{O}(m; z, (2p-2)\delta)$.*

Proof. By the tower rule, a subalgebra of A is of the form $\mathbb{Q}$, $\mathbb{Q}(x)$, or $\mathbb{Q}(x, y)$, where $x, y \in A$ and $\mathbb{Q}(x, y)$ is the smallest algebra containing both x and y. Indeed, if a subalgebra contains a non-rational element x, then it contains $\mathbb{Q}(x)$, which must have dimension p over $\mathbb{Q}$ (the characteristic polynomial has degree p). If it contains another element not in $\mathbb{Q}(x)$, say y, then it contains $\mathbb{Q}(x, y)$. The tower rule implies now that $A = \mathbb{Q}(x, y)$. The algebra $\mathbb{Q}(x, y)$ is generated as a vector space by monomials of degree at most $2p-2$.

Now if $a_j \in \bigcup_{1\leq m\leq L} \mathcal{O}(m; z, \delta)$ for $j = 1, \ldots, 2p-2$, then the reduced norm of $\prod a_j$ is at most L^{2p-2} and, by the triangle inequality, $d(z, \prod \tilde{a}_j \cdot z) < (2p-2)\delta$. The order structure ensures that $\prod a_j$ lies in $\mathcal{O}$. □

For the next lemma denote $\mathrm{disc}(\mathcal{O}) := D$.

Lemma 3.4.2. *The $\mathbb{Q}$-vector space spanned by $\bigcup_{1\leq m\leq L^{2p-2}} \mathcal{O}(m; z, (2p-2)\delta)$ is proper, i.e. not equal to A, if $L \ll D^{1/4p(p-1)-\varepsilon}$, where the implicit constant depends only on p and δ.*

Proof. Let $\gamma_1, \ldots, \gamma_{p^2}$ be elements of $\bigcup_{1\leq m\leq L^{2p-2}} \mathcal{O}(m; z, (2p-2)\delta)$. In this case $\mathrm{nr}(\gamma_i\gamma_j) \ll L^{4(p-1)}$ and we have that $d(\gamma_i\gamma_j z, z) \leq d(\gamma_i z, z) + d(\gamma_j z, z) < 4(p-1)\delta$, by the triangle inequality. In particular

$$\mathrm{tr}(\gamma_i\gamma_j) \ll_{\delta,p} L^{4(p-1)/p},$$

by applying the trace to the equation $z^{-1}\gamma_i\gamma_j z = \mathrm{nr}(\gamma_i\gamma_j)^{1/p}(k + O_p(\delta))$, with some $k \in \mathrm{SO}(p)$.

Consider now $s = \det(\mathrm{tr}(\gamma_i\gamma_j)_{i,j})$. Recall that D is the generator of the ideal in $\mathbb{Z}$ generated by $\{\det \mathrm{tr}(x_i x_j) \mid x_i \in \mathcal{O}, i = 1, \ldots p^2\}$. Since $\gamma_i \in \mathcal{O}$ for all i, it follows that $D \mid s$.

On the other hand, by using the bound above, we deduce that $s \ll L^{4p(p-1)}$. Thus if $L \ll D^{1/4p(p-1)-\varepsilon}$, then $s = 0$. By the non-degeneracy of the bilinear form given by the reduced trace, it follows that $\gamma_1, \ldots, \gamma_{p^2}$ are *not* linearly independent. □

Thus, if L is small enough, we can assume that we are counting matrices in a proper subalgebra of A, which must be $\mathbb{Q}$ or a field extension $E/\mathbb{Q}$ of degree p. This is where the use of p being a prime is crucial.

We are now counting certain elements in $\mathcal{O}_E$, the ring of integers of E, which certainly includes $\mathcal{O} \cap E$. We do so by counting ideals and units. Since the units in $\mathbb{Z}$ are only ± 1, we can concentrate on the non-trivial extensions, which must have an infinite group of units, at least if $p > 2$. It is important to note that the reduced norm and the reduced trace in A of an element in a subfield $E \subset A$ such that $\dim_{\mathbb{Q}} E = p$ are the same as the number field norm, resp. trace of $E/\mathbb{Q}$ (see [Pie82, Sect. 16.2]).

Lemma 3.4.3. *Let $E/\mathbb{Q}$ be a cyclic extension of degree p that is a subfield of A and let $\mathcal{O}$ be an order of A. The number of units $\xi \in \mathcal{O}^\times \cap E$ such that $d(z, \tilde{\xi}z) \leq \delta$ for a given $z \in \mathfrak{h}^p$ is $\ll p^p(1+\delta)^{p-1}$.*

Proof. Let $\xi \in \mathcal{O}^\times \cap E$. Then $\xi \in \mathcal{O}_E^\times$, where $\mathcal{O}_E$ is the ring of integers of L, by integrality over $\mathbb{Z}$. Thus, $\mathrm{nr}(\xi) = N_{E/\mathbb{Q}}(\xi) = \pm 1$.

Next, the condition $d(z, \tilde{\xi}z) \leq \delta$ is equivalent to $\xi \in zB(\delta)z^{-1}$, where $B(\delta)$ is a union of δ-balls around all elements of $\mathrm{SO}(p)$. Applying the trace, we find that $\mathrm{tr}(\xi) \ll p(1+\delta)$. Since $\mathrm{tr}(\xi) \in \mathbb{Z}$ by integrality, we see that there are $\ll p(1+\delta)$ possibilities for the value of $\mathrm{tr}(\xi)$.

We may apply the same reasoning to ξ^j and derive that there are $\ll p(1+j\delta)$ possibilities for the value of $\mathrm{tr}(\xi^j)$. Indeed,

$$d(z, \tilde{\xi}^j z) \leq d(z, \tilde{\xi}z) + d(\xi z, \tilde{\xi}^j z) = d(z, \tilde{\xi}z) + d(z, \tilde{\xi}^{j-1} z) \leq j\delta,$$

inductively. Note also that $\xi^j \in \mathcal{O}^\times \cap E$ since $\mathcal{O}$ is closed under multiplication.

Now the characteristic polynomial of ξ is

$$X^p - \mathrm{tr}(\xi)X^{p-1} + \frac{1}{2}\left[\mathrm{tr}(\xi)^2 + \mathrm{tr}(\xi^2)\right] X^{p-2} + \ldots \pm \det(\xi).$$

By Newton's identities, each coefficient is determined by the values of $\mathrm{tr}(\xi^j)$ for certain j. By the bounds above, there are only $\ll \prod_{j=1,\ldots p-1} p(1+j\delta) \ll [p(1+\delta)]^{p-1}$ polynomials that are satisfied by a unit ξ as in the statement of the lemma. Since each polynomial can have at most p different roots, the proof is finished. □

Lemma 3.4.4. *Let $E \subset A$ be a field of degree p over $\mathbb{Q}$. Then for a any $z \in \mathfrak{h}^p$ and any positive integer m we have*

$$\mathcal{O}_E(m; z, \delta) \ll_p \tau(m)^{p-1} \cdot (1+\delta)^{p-1}.$$

Proof. Let $\gamma \in \mathcal{O}_E$ with $\mathrm{nr}(\gamma) = N_{E/\mathbb{Q}}(\gamma) = m$. Up to units, there are only $\tau(m)^{p-1}$ elements of $\mathcal{O}_E$ with norm m. Indeed, a principal ideal is determined by its generator up to units and the norm of the ideal is equal to the norm of the generator. Since ideals factorise uniquely into prime factors, we only need to count prime ideals.

Above each rational prime, there are at most p prime ideals of $\mathcal{O}_E$. Therefore, if $q^v \mid m$ for a prime q, then we need to choose at most p numbers $a_1, \ldots, a_p \in$

$\mathbb{Z}_{\geq 0}$ such that $a_1 + \ldots + a_p = v$ to determine an ideal of norm q^v. The number of such tuples is $v + p - 1$ choose $p - 1$, that is $<< v^{p-1}$. Thus, there are at most

$$\ll \prod_{q^v \| m} v^{p-1} = \tau(m)^{p-1}$$

ideals of norm m.

Now if $\gamma \in \mathcal{O}_E(m; z, \delta)$ and $\xi\gamma \in \mathcal{O}_E(m; z, \delta)$ for some unit $\xi \in \mathcal{O}_E^\times$, then

$$d(z, \tilde{\xi}z) \leq d(z, \tilde{\xi}\tilde{\gamma}z) + d(\tilde{\xi}\tilde{\gamma}z, \xi z) = d(z, \tilde{\xi}\tilde{\gamma}z) + u(z, \tilde{\gamma}z) \leq 2\delta.$$

Thus, we finish the proof by counting such units using Lemma 3.4.3. □

Putting everything together and recalling the divisor bound $\tau(m) \ll m^\varepsilon$ we obtain the following proposition.

Proposition 3.4.5. *Let $m \ll D^{\frac{1}{4p(p-1)}-\varepsilon}$ with implicit constant as in Lemma 3.4.2 and let $z \in \mathfrak{h}^p$. Then $\#\mathcal{O}(m; z, \delta) \ll_p m^\varepsilon$, where the implicit constant depends only on ε, δ, and p.*

3.5 COUNTING IN THE SPECTRAL ASPECT

Let μ be the spectral parameter of ϕ and denote $S \asymp D(\mu^*)$ as in [BM16, (1.2)].

3.5.1 The counting argument for small distances

We are interested in bounding the cardinality of

$$\mathcal{O}(m; z, \delta) = \{\gamma \in \mathcal{O} : \mathrm{nr}(\gamma) = m, d(z, \tilde{\gamma}z) = O(\delta)\},$$

where $\tilde{\gamma} = \gamma/\mathrm{nr}(\gamma)^{1/p}$. The condition $d(z, \tilde{\gamma}z) = O(\delta)$ is equivalent to $z^{-1}\tilde{\gamma}z = k + O(\delta)$ for some $k \in \mathrm{SO}(p)$.

As opposed to the discriminant aspect, we can now gain savings in the pretrace inequality by using δ as a parameter. For $p \geq 3$, instead of showing that the $\mathbb{Q}$-algebra generated by the elements we are counting is a proper algebra and then deducing commutativity, we directly show that it must be a commutative field, at least if δ is small enough.

For the following lemma, assume that $p \geq 3$ is any odd integer, not necessarily prime.

Lemma 3.5.1. *The $\mathbb{Q}$-algebra generated by $\bigcup_{1\leq m\leq L} \mathcal{O}(m; z, \delta)$ is commutative, i.e. a field, if $\delta \ll L^{-2-\varepsilon}$, where the implicit constant depends only on p.*

Proof. Let $\gamma_1, \gamma_2 \in \bigcup_{1\leq m\leq L} \mathcal{O}(m; z, \delta)$. A few applications of the triangle inequality, recalling that $d(\tilde{\gamma}z, z) = d(\tilde{\gamma}^{-1}z, z)$,[3] show that $d(\gamma_1^{-1}\gamma_2^{-1}\gamma_1\gamma_2 z, z) \ll$

[3]For instance, $d(\alpha^{-1}\beta z, z) \leq d(\alpha^{-1}\beta z, \alpha^{-1}z) + d(\alpha^{-1}z, z) = d(\beta z, z) + d(\alpha z, z)$ by invariance of the distance function.

δ. This implies that

$$z^{-1}\gamma_1^{-1}\gamma_2^{-1}\gamma_1\gamma_2 z = k + O(\delta), \tag{3.5.1}$$

since $\mathrm{nr}(\gamma_1^{-1}\gamma_2^{-1}\gamma_1\gamma_2) = 1$.

Recall that we assume $p \geq 3$ and note that $\det(k-1) = 0$ for all $k \in \mathrm{SO}(p)$, since 1 is certainly an eigenvalue of orthogonal matrices in odd degree. Therefore, subtracting 1 from (3.5.1) and taking the determinant gives

$$\begin{aligned}\mathrm{nr}(\gamma_1^{-1}\gamma_2^{-1}\gamma_1\gamma_2 - 1) &= \det(z^{-1}(\gamma_1^{-1}\gamma_2^{-1}\gamma_1\gamma_2 - 1)z) \\ &= \det(k - 1 + O(\delta)) = O_p(\delta).\end{aligned}$$

Multiplying this last equation by $\mathrm{nr}(\gamma_2\gamma_1)$ implies that

$$\mathrm{nr}(\gamma_1\gamma_2 - \gamma_2\gamma_1) = \mathrm{nr}(\gamma_2\gamma_1)\cdot\mathrm{nr}(\gamma_1^{-1}\gamma_2^{-1}\gamma_1\gamma_2 - 1) = O_p(\delta\,\mathrm{nr}(\gamma_1)\,\mathrm{nr}(\gamma_2)). \tag{3.5.2}$$

The commutator of two elements of $\mathcal{O}$ is again in $\mathcal{O}$ and so $\mathrm{nr}(\gamma_1\gamma_2 - \gamma_2\gamma_1) \in \mathbb{Z}$. Thus if $\delta \ll L^{-2-\varepsilon}$, then it follows that $\mathrm{nr}(\gamma_1\gamma_2 - \gamma_2\gamma_1) = 0$, which implies that $\gamma_1\gamma_2 = \gamma_2\gamma_1$ since A is a division algebra. □

The argument above makes crucial use of the fact that special orthogonal matrices in odd degree necessarily have 1 as eigenvalue, which is not the case any more in even degree. To produce an alternative argument for $p = 2$, we shall return to the strategy employed for the discriminant aspect. This is done in Section 3.7.2.

Given the lemma above, we may assume that we are counting in a field $\mathbb{Q} \subset E$. This is done exactly as in Lemma 3.4.4.

Proposition 3.5.2. *Let A have odd degree $p \geq 3$. If $\delta \ll m^{-2-\varepsilon}$, then $\#\mathcal{O}(m; z, \delta) \ll m^{\varepsilon}$.*

3.5.2 The counting argument for large distances

If $\delta = \rho \gg 1$ is as large as the diameter of the support of f_μ (see Section 3.3), we still need to bound the cardinality of $\mathcal{O}(m; z, \delta)$ by a reasonable power of m but with no dependence on the discriminant D. Since the amplifier is constructed only from unramified primes, as explained in Section 3.3, we may assume that m is coprime to the discriminant of A.

Proposition 3.5.3. *We have the bound $\#\mathcal{O}(m; z, \rho) \ll m^{p-1+\varepsilon}$, where the implicit constant depends only on p.*

Proof. Let $\gamma \in \mathcal{O}(m; z, \rho)$ and consider the principal ideal $\gamma\mathcal{O}_m$ in a maximal order $\mathcal{O}_m$ containing $\mathcal{O}$. We have $\mathrm{nr}(\gamma)\mathbb{Z} = \mathrm{nr}(\gamma\mathcal{O}_m)$ and if $\gamma\mathcal{O}_m = \gamma'\mathcal{O}_m$, then $\gamma\xi = \gamma'$ for some unit $\xi \in \mathcal{O}_m^\times$. It therefore suffices to count ideals with norm m and bound the number of units ξ as above.

By the local-global dictionary for ideals, it suffices to count locally. Since m is coprime to the discriminant of A, we only need to count ideals of norm q^e

in $M_p(\mathbb{Z}_q)$ for primes q and positive integers e. This is done for instance as in Lemma 26.4.1 in [Voi21] and is a well-known computation in the theory of zeta functions of algebras. We use the fact that all ideals are principal and employ the theory of elementary divisors to find a generator of the ideal of lower triangular form, where the diagonal is given by $(q^{a_1}, \dots, q^{a_p})$ for non-negative integers $a_1, \dots, a_p$ such that $a_1 + \dots a_p = e$. All entries on the column below q^{a_j} are uniquely defined as elements of $\mathbb{Z}/q^{a_j}\mathbb{Z}$. An easy counting argument thus shows that there are

$$\sum_{a_1+\dots a_p=e} q^{(p-1)a_1+(p-2)a_2+\dots a_{p-1}} \ll q^{e(p-1)(1+\varepsilon)}$$

ideals of norm q^e in $M_p(\mathbb{Z}_q)$.

This shows that there are $\ll m^{p-1}$ ideals of norm m in O_m, which is implicitly also a bound for the number of principal ideals. Now let $\xi \in O_m^\times$ be a unit as in the first paragraph of this proof. By the triangle inequality, we find that $d(\xi z, z) \ll \rho$. Since $\mathrm{nr}(\xi) = 1$, this implies that $z^{-1}\xi z = k + O(\rho)$, for some $k \in \mathrm{SO}(p)$.

Suppose p is odd. If ξ_1, ξ_2 are two such units, then the same reasoning as before (3.5.2) provides the bound

$$\mathrm{nr}(\xi_1\xi_2 - \xi_2\xi_1) = O_p(\rho).$$

As observed in Remark 3.3.3, we may assume that $\mathrm{nr}(\xi_1\xi_2 - \xi_2\xi_1) < 1$. By integrality, this implies that $\xi_1\xi_2 = \xi_2\xi_1$.

Therefore, we may reduce the counting problem to counting units in the maximal order of a field by restricting to the commutative algebra generated by these units. As proved in Lemma 3.4.4, there are $\ll_p 1$ units with the required properties. Putting all bounds together proves the statement. □

3.6 PROOF OF THEOREM 3.1

We are now ready to insert the counting results into the pretrace inequality (3.3.4). Let the degree n be an odd number, at least 3.

3.6.1 The spectral aspect

We take the parameter δ as large as possible to still have control over the counting problem, but also gain a saving in the third term of (3.3.4). Since the largest norm coming up in the pretrace inequality is $l_1^n l_2^{(n-1)n} \ll L^{n^2}$, we must take $\delta \ll L^{-2n^2-\varepsilon}$ to be able to apply Proposition 3.5.2.

With such a choice of δ, the second term in (3.3.4) is bounded as

$$\sum_{\nu=1}^{n} \sum_{l_1,l_2\in\mathcal{P}} \frac{1}{L^{(n-1)\nu}} \#O(l_1^\nu l_2^{(n-1)\nu}; z, \delta) \ll_n |\mathcal{P}|^2 L^{-(n-1)+\varepsilon}. \tag{3.6.1}$$

As already mentioned, to obtain a saving from the third term, we take δ as large as possible, i.e. $\delta \asymp L^{-2n^2-\varepsilon}$. Together with Proposition 3.5.3, the third term in (3.3.4) can be bounded by

$$\begin{aligned} & S(\mu^*)^{\frac{-1}{n(n-1)}} \cdot \delta^{\frac{-1}{2}} \sum_{\nu=1}^{n} \sum_{l_1,l_2 \in \mathcal{P}} \frac{1}{L^{(n-1)\nu}} \#\mathcal{O}(l_1^{\nu} l_2^{(n-1)\nu}; z, \rho) \\ & \ll S(\mu^*)^{\frac{-1}{n(n-1)}} L^{n^2+\varepsilon} |\mathcal{P}|^2 \sum_{\nu=1}^{n} \frac{L^{(n-1)^2\nu}}{L^{(n-1)\nu}} \\ & \ll_n S(\mu^*)^{\frac{-1}{n(n-1)}} |\mathcal{P}|^2 L^{n^2+n(n-1)(n-2)+\varepsilon}. \end{aligned}$$

We introduce the two bounds above into the pretrace inequality and, recalling that $|\mathcal{P}| \gg L^{1-\varepsilon} \cdot D^{-\varepsilon}$, where D is the discriminant, we obtain

$$|\phi(z)|^2 \ll S(\mu^*) \left(L^{-1+\varepsilon} D^{\varepsilon} + L^{-(n-1)+\varepsilon} D^{\varepsilon} + S(\mu^*)^{\frac{-1}{n(n-1)}} L^{n^2+n(n-1)(n-2)+\varepsilon} \right).$$

It is clear that the first term dominates the second. After solving an easy optimisation problem for the first and third terms, we arrive at the bound

$$|\phi(z)|^2 \ll S(\mu^*) \cdot S(\mu^*)^{\frac{-1}{n(n-1)(n^2+n(n-1)(n-2)+1)}+\varepsilon} \cdot D^{\varepsilon} \ll S(\mu^*) \cdot S(\mu^*)^{\frac{-1}{n^4(n-1)}+\varepsilon} \cdot D^{\varepsilon}, \tag{3.6.2}$$

where we weaken the first bound to the last one simply for aesthetic reasons.

3.6.2 The discriminant aspect

Here we assume that $n = p$ is a prime. We cannot gain any saving in the discriminant aspect by using the δ parameter. Therefore, we set $\delta = \rho$ and ignore the last term in (3.3.4). In this case, we take L as large as Proposition 3.4.5 allows. Taking into account the largest norms appearing in (3.3.4), we may choose $L \asymp D^{\frac{1}{4n^3(n-1)}}$. Then the second term in the pretrace inequality can be bounded as in (3.6.1), so that we have

$$|\phi(z)|^2 \ll S(\mu^*)(L^{-1+\varepsilon} + L^{-(n-1)+\varepsilon}) \ll S(\mu^*) \cdot L^{-1+\varepsilon} \ll S(\mu^*) D^{\frac{-1}{4n^3(n-1)}+\varepsilon}. \tag{3.6.3}$$

3.6.3 The hybrid bound

We interpolate between the two bounds (3.6.2) and (3.6.3) simply by multiplying them, so that

$$|\phi(z)| \ll |\phi(z)|^{1/2} \cdot |\phi(z)|^{1/2} \ll S(\mu^*)^{\frac{1}{2}} S(\mu^*)^{\frac{-1}{4n^4(n-1)}+\varepsilon} D^{\frac{-1}{16n^3(n-1)}+\varepsilon}.$$

Recalling (3.3.1), this proves Theorem 3.1.

3.7 QUATERNION ALGEBRAS OVER NUMBER FIELDS

The full classification of cocompact arithmetic subgroups requires us to also consider quaternion algebras over number fields and Templier [Tem10] treats the counting problem in this more general setting (though only in the level aspect and only for Eichler orders). This case is slightly more technical and we treat the problem by applying the same ideas as above more carefully. We first recall the theoretical background.

Let F be a totally real number field of degree n. We denote by $\mathfrak{o}_F$ its ring of integers and by N_F the number field norm of $F/\mathbb{Q}$. Let A be a division quaternion algebra over F and assume that there is an embedding $\sigma_0 \in \hom(F, \mathbb{R})$ such that $A \otimes_{\sigma_0} \mathbb{R} \cong M_2(\mathbb{R})$. For all other embeddings $\sigma_0 \neq \sigma \in \hom(F, \mathbb{R})$ assume that $A \otimes_\sigma \mathbb{R} \cong \mathcal{H}(\mathbb{R})$, where $\mathcal{H}(\mathbb{R})$ is the Hamilton quaternion algebra.

Now let $\mathcal{O}$ be an $\mathfrak{o}_F$-order and let $\mathfrak{D} = \mathrm{disc}(\mathcal{O}) \subset \mathfrak{o}_F$ be its discriminant. By abuse of notation, we also denote a generator of the discriminant ideal by $\mathfrak{D}$. Let

$$D := |N_F(\mathfrak{D})|.$$

We may view A as embedded in $A_\infty \cong M_2(\mathbb{R}) \times \mathcal{H}(\mathbb{R})^{n-1}$. We use φ_0 to denote the projection onto the first component $M_2(\mathbb{R})$ and φ_i, $i = 1, \dots, n-1$, to denote the projections onto the Hamiltonian components.

Note that $\sigma_i(\mathrm{tr}(\gamma)) = \mathrm{tr}(\varphi_i(\gamma))$ for φ_i the projection onto the σ_i component. The trace on the left hand side refers to the quaternion trace and on the right hand side it refers to the usual matrix trace. Similarly, $\sigma_i(\mathrm{nr}(\gamma)) = \mathrm{nr}(\varphi_i(\gamma))$.

3.7.1 *Remarks on the amplifier*

We consider automorphic forms on $\mathfrak{h}^2$ invariant under the arithmetic group $\mathcal{O}^\times$, viewed as a subgroup of $\mathrm{PGL}_2(\mathbb{R})$. [4] The same consideration on the pretrace formula as in Section 3.3 (without the amplifier) apply to this case as well.

Next, Hecke theory over number fields is best understood adelically. Given our classical motivation and for the sake of brevity, we prefer not to introduce the general formalism and additional notation, since it is not essential for the main argument of this paper. We refer to the detailed description of the classical and adelic theory given in Section 2 of [Shi78] in the related case of Hilbert modular forms.

Instead, we note as in [Tem10], Section 5.5, that the action of a subalgebra of the Hecke algebra suffices for our purposes. Namely, to amplify the pretrace

[4] This is a slightly different subgroup than in the rest of the paper, where we take norm 1 units. It is a technical assumption due to the fact that the units $\mathfrak{o}_F^\times$ (or even the totally positive units, depending on the setup) might not all be squares. This implies that the quotient $\mathcal{O}^1 \backslash \mathrm{SL}_2(\mathbb{R})$ might be larger than $\mathbb{R}^\times \mathcal{O}^\times \backslash \mathrm{GL}_2(\mathbb{R})$. The latter is the more natural one from the point of view of automorphic forms and Hecke theory and is also used by Templier (see (5.8) in [Tem10]). Nevertheless, the difference consists merely of a character on $\mathfrak{o}_F^\times / \mathfrak{o}_F^{\times 2}$.

formula, we use only Hecke operators that are associated to principal ideals of $\mathfrak{o}_F$, whose action is explained for instance in Templier's article. By Chebotarev's density theorem (see Theorem 13.2 in [Neu92]), these ideals make up a positive proportion of all prime ideals of F, its numerical value depending only on F. As before, we assume that $\mathcal{O}$ is locally norm-maximal for simplicity and recall Remark 3.3.2.

Define

$$\mathcal{O}(m;z,\delta) = \{\gamma \in \mathcal{O} : |N_F(\mathrm{nr}(\gamma))| = m, d(z,\varphi_0(\tilde{\gamma})z) = O(\delta)\}.$$

Using the notation of Section 3.3 and following Sections 6.5 and 6.6 in [Tem10], we obtain a version of (3.3.4). There is a certain sequence y_m supported on positive integers less than L^4 such that

$$L^{2-\varepsilon}D^{-\varepsilon}|\phi(z)|^2 \ll_F S(\mu^*)\cdot\left(\sum_{m\ll L^4}\frac{y_m}{\sqrt{m}}\#\mathcal{O}(m;z,\delta) + (S(\mu^*)\delta)^{-\frac{1}{2}}\sum_{m\ll L^4}\frac{y_m}{\sqrt{m}}\#\mathcal{O}(m;z,\rho)\right). \quad (3.7.1)$$

Furthermore, $y_m \ll 1$ and $\sum_m y_m \ll L^2$, as in (6.17) of [Tem10].

As before, we are now faced with a counting problem. Note that we may count elements of $\mathcal{O}(m;z,\delta)$ modulo units $\mathfrak{o}_F^\times$, since $F \subset \mathbb{R}$ is the centre of A. This remark is much more useful in the number field case than over the rationals, given that $\mathfrak{o}_F^\times$ is generally infinite.

3.7.2 The counting argument

We now follow the argument in Section 3.4. We shall therefore show that the algebra generated by the elements we are counting is a proper algebra and thus a field by the tower rule.

Recall that by Lemma 3.4.1, which is independent of the ground field, it suffices to show that the $\mathbb{Q}$-*vector space* spanned by $\bigcup_{1\le m\le L^2}\mathcal{O}(m;z,2\delta)$ is proper. The following is an adapted version of Lemma 3.4.2. By chance though, this lemma now gives bounds in terms of the parameter δ as well. We exploit the fact that SO(2) can only span a two dimensional vector space. This behaviour is not generic since orthogonal matrices in degree larger than 2 can span the entire algebra of matrices over $\mathbb{R}$.

Lemma 3.7.1 (Lemma 3.4.2 revisited)**.** *The F-vector space spanned by*

$$\bigcup_{1\le m\le L^2}\mathcal{O}(m;z,2\delta)$$

is proper, i.e. not equal to A, if $\delta \ll D^{1-\varepsilon}L^{-6-\varepsilon}$, where the implied constant depends only on ε.

Proof. Let $\gamma_1, \gamma_2, \gamma_3, \gamma_4 \in \bigcup_{1\leq m\leq L^2} \mathcal{O}(m; z, 2\delta)$. We wish to show that these elements are linearly dependent and we may assume without loss of generality that $\gamma_1 = 1$. Since the reduced trace gives a *non-degenerate* bilinear form, it suffices to show that $s := \det(\mathrm{tr}(\gamma_i, \gamma_j))_{i,j} = 0$.

By assumption, we have $\sigma_0(\mathrm{nr}(\gamma))^{-1/2} \cdot z^{-1}\varphi_0(\gamma_i)z = k_i + O(\delta)$, for some $k_i \in \mathrm{SO}(2)$, and thus

$$\frac{1}{\sigma_0(\mathrm{nr}(\gamma_i)\,\mathrm{nr}(\gamma_j))^{1/2}} z^{-1}\varphi_0(\gamma_i\gamma_j)z = k_i k_j + O(\delta).$$

Therefore,

$$\frac{1}{\prod_i \sigma_0(\mathrm{nr}(\gamma_i))}\sigma_0 \det(\mathrm{tr}(\gamma_i\gamma_j)_{i,j}) = \det(\mathrm{tr}(k_i k_j)_{i,j}) + O(\delta).$$

Note that SO(2) spans only a 2-dimensional vector space, which is easily seen using the standard parametrisation. It follows that $\det(\mathrm{tr}(k_i k_j)_{i,j}) = 0$, since we are considering the 4-dimensional matrix space. Therefore,

$$\sigma_0 \det(\mathrm{tr}(\gamma_i\gamma_j)_{i,j}) \ll \delta \prod_i |\sigma_0\,\mathrm{nr}(\gamma_i)|,$$

recalling that $\gamma_1 = 1$.

If $\sigma \neq \sigma_0$, then $\sigma\,\mathrm{tr}(\gamma_i\gamma_j) \leq 2|\sigma\,\mathrm{nr}(\gamma_i\gamma_j)|^{1/2}$. Indeed, for the corresponding projections $\varphi_i, i \neq 0$, $\varphi_i(\xi)$ is an element of the real Hamilton quaternion algebra. For an arbitrary such element $a + ib + jc + kd$ in the usual notation with $a, b, c, d \in \mathbb{R}$, its trace is equal to $2a$ and its norm is $a^2 + b^2 + c^2 + d^2$, whence the inequality. Therefore, we deduce that

$$\sigma \det(\mathrm{tr}(\gamma_i\gamma_j)_{i,j}) \ll \delta \prod_i |\sigma\,\mathrm{nr}(\gamma_i)|.$$

By the inequalities above, we have that

$$N_F(s) \ll \delta \prod_i |N_F(\gamma_i)| \ll \delta L^6.$$

Recall that $\mathfrak{D}$ is the generator of the ideal in $\mathfrak{o}_F$ generated by $\{\det \mathrm{tr}(x_i x_j) \mid x_i \in \mathcal{O}, i = 1, \ldots 4\}$. Since $\gamma_i \in \mathcal{O}$ for all i, it follows that $s = \mathfrak{D} \cdot x$ for some $x \in \mathfrak{o}_F$ and therefore $D \mid N_F(s)$.

In conclusion, if $\delta \ll D^{-1-\varepsilon}L^{6-\varepsilon}$, then $s = 0$. By the non-degeneracy of the bilinear form given by the reduced trace, it follows that $\gamma_1, \ldots, \gamma_{p^2}$ are *not* linearly independent. □

Remark 3.7.2. We remark that for $n \geq 3$ the $n \times n$ orthogonal matrices span the full space of real matrices. Thus, an application of the same proof as above in higher rank is bound to fail since the determinant $\det(\mathrm{tr}(k_i k_j)_{i,j})$ might be non-zero.

Next, Lemma 3.4.4 goes through with the same proof if we can control the action of units. This cannot be done directly as in Lemma 3.4.3 since $N_{E/F}(\xi) \in \mathfrak{o}_F^\times$ for any unit $\xi \in O_E^\times$, and the group $\mathfrak{o}_F^\times$ is infinite for $F \neq \mathbb{Q}$. We can balance this out by recalling that we only need to count ξ up to units in $\mathfrak{o}_F^\times$, since these act trivially on the upper half plane.

Lemma 3.7.3 (Lemma 3.4.3 revisited)**.** *Let E/F be an extension of degree 2 that is a subfield of A and let O be an order of A. The number of units $\xi \in O^\times \cap E$ up to multiplication by units in $\mathfrak{o}_F$, such that $d(z, \varphi_0(\xi)z) \leq \delta$ for some $z \in \mathfrak{h}^2$, is $\ll_F (1+\delta)^2$.*

Proof. We begin by investigating the quantity $(\mathrm{tr}_{E/F}\,\xi)^2/N_{E/F}\xi$ and proving that it can only take finitely many values. For any embedding $\sigma \neq \sigma_0$, we have

$$\sigma\left(\frac{(\mathrm{tr}_{E/F}\,\xi)^2}{N_{E/F}\xi}\right) \in [0,4],$$

and the condition $d(z, \varphi_0(\xi)z) \leq \delta$ implies that

$$\sigma_0\left(\frac{(\mathrm{tr}_{E/F}\,\xi)^2}{N_{E/F}\xi}\right) \ll (1+\delta)^2,$$

as we have seen already in the proof of Lemma 3.7.1.

Since $\xi \in O_E^\times$, the maximal order, the quantity $(\mathrm{tr}_{E/F}\,\xi)^2/N_{E/F}\xi$ must lie in $\mathfrak{o}_F$. Recall that the image of $\mathfrak{o}_F$ inside $\mathbb{R}^n$ under all embeddings is a discrete lattice. Since the image of $(\mathrm{tr}_{E/F}\,\xi)^2/N_{E/F}\xi$ is bounded, it follows that the number of possibilities for the value of this quantity is bounded by $(1+\delta)^2$, up to a constant depending on F.

For the last step, recall Dirichlet's unit theorem, stating that $\mathfrak{o}_F^\times$ is a finitely generated group. This implies that $\mathfrak{o}_F^\times/(\mathfrak{o}_F^\times)^2$ is finite. Now if $\kappa \in \mathfrak{o}_F^\times$, then $N(\kappa\xi) = \kappa^2 N(\xi)$. Thus, if we are only counting $\xi \in \mathfrak{o}_F^\times \backslash O_E^\times$, then the value of $N(\xi)$ can only lie in $\mathfrak{o}_F^\times/(\mathfrak{o}_F^\times)^2$.

Nevertheless, we have

$$\frac{(\mathrm{tr}_{E/F}(\kappa\xi))^2}{N_{E/F}(\kappa\xi)} = \frac{(\mathrm{tr}_{E/F}\,\xi)^2}{N_{E/F}\xi}.$$

Since there are only finitely many possibilities for this quantity and finitely many for $N_{E/F}(\xi)$, if follows that there are only finitely many possibilities for $\mathrm{tr}_{E/F}\,\xi$. Finally, ξ is determined up to the action of $\mathrm{Gal}(E/F)$ (which has order 2) by its minimal polynomial. This polynomial is determined by the trace and norm of ξ. The lemma now follows by bookkeeping. □

Remark 3.7.4. We remark that Lemma 6.4, part (i), of [Tem10], having the same ultimate goal as Lemma 3.7.3 in this paper, might not hold in general. Indeed,

in the proof, the condition $u(z, \varphi_0(\xi)z) \leq \delta$ is said to be equivalent to

$$\frac{\varphi_0(\xi)}{\sigma_0(N_A(\xi))^{1/2}} \in zB(\delta)z^{-1},$$

for a single δ-ball $B(\delta)$ around the identity. This cannot be true in general. For instance if $z = i$, corresponding to the identity matrix, and $\xi = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, then $u(z, \varphi_0(\xi)z) = 0$, yet ξ lies far from the identity.

Our proof uses the idea of Lemma 6.3 in [Tem10] and couples it with an application of Dirichlet's unit theorem, so as to apply the argument using characteristic polynomials, as in the higher degree case.

Having adapted the lemmata above, we now complete the argument as in the Section 3.4 and obtain the following counting result.

Proposition 3.7.5. *Let $\delta \ll D^{1-\varepsilon}m^{-6-\varepsilon}$ and let $z \in \mathfrak{h}^2$. Then $\#\mathcal{O}(m; z, \delta) \ll m^{\varepsilon}$, where the implicit constant depends only on ε and F.*

Compared to the counting results in Sections 3.4 and 3.5, Proposition 3.7.5 is conceptually more satisfying, since it seemingly blends together the spectral and discriminant aspect. Nevertheless, we still need a counting results for large δ.

Proposition 3.7.6. *We have the bound $\#\mathcal{O}(m; z, \rho) \ll m^{1+\varepsilon}$.*

Proof. The proof is essentially the same as the proof of Proposition 3.5.3. Yet again, additional care must be taken when counting units at the end. For this, we apply the same argument as in Lemma 3.7.1, which simplifies since L is now equal to 1. In particular, if $\rho \ll D$, then the relevant units generate a proper subalgebra, which must be a field in this case. Since D is an integer and ρ is small enough, the condition is met. To conclude the proof, we apply Lemma 3.7.3. □

3.7.3 Proof of Theorem 3.2

Considering the range of m in the pretrace inequality (3.7.1), in order to use Proposition 3.7.5, we take $\delta \asymp D^{1-\varepsilon}L^{-24-\varepsilon}$. Then, applying Proposition 3.7.6, we have

$$\begin{aligned}
|\phi(z)|^2 &\ll_F D^{\varepsilon}L^{-2+\varepsilon}S(\mu^*)\left(\sum_{m\ll L^4} y_m m^{-1/2+\varepsilon} + (S(\mu^*)D^{1-\varepsilon}L^{-24-\varepsilon})^{-\frac{1}{2}}\sum_{m\ll L^4} y_m m^{1/2+\varepsilon}\right) \\
&\ll S(\mu^*)L^{-2+\varepsilon}D^{\varepsilon}\left(\left(\sum_m y_m^2\right)^{1/2}\left(\sum_m \frac{1}{m}\right)^{1/2} + (S(\mu^*)D)^{-\frac{1}{2}}L^{12}\cdot L^2\sum_m y_m\right) \\
&\ll S(\mu^*)L^{\varepsilon}D^{\varepsilon}\left(L^{-1} + (S(\mu^*)D)^{-\frac{1}{2}}L^{14}\right),
\end{aligned}$$

where we make use of the properties of the sequence y_m. Optimising by setting the two terms in the last factor equal, we set $L = (S(\mu^*)D)^{\frac{1}{30}}$ and obtain a saving of L^{-1}. This implies that

$$|\phi(z)|^2 \ll_F S(\mu^*) \cdot (S(\mu^*)D)^{-\frac{1}{30}+\varepsilon}.$$

This proves Theorem 3.2.

3.8 REMARKS ON THE CASE OF COMPOSITE DEGREE

It is not clear how to generalise the counting arguments, particularly those in the discriminant aspect (Section 3.4), to the case of algebras of general degree $n \in \mathbb{N}_{\geq 2}$. The main issue is the existence of non-commutative proper subalgebras in general. Yet the ideas in this paper suffice for proving the partial results given in Theorem 3.3 for special types of orders, which we sketch in this section.

More precisely, let N be a positive integer and $\mathcal{O}_0(N)$ be an order of the division algebra A over $\mathbb{Q}$ of degree n such that, at all unramified primes p, its completion is of the form

$$\mathcal{O}_0(N)_p = \left\{\gamma \in M_n(\mathbb{Z}_p) \mid \text{last row of } \gamma \equiv (0, \dots, 0, *) \bmod N\mathbb{Z}_p\right\},$$

up to conjugation. These orders are interesting from the point of view of newform theory. If $n = 2$, then these are precisely the Eichler orders, which have generally received much attention in the theory of automorphic forms. In higher degree, these orders form a (proper) subset of the orders that are the intersection of two maximal orders. It is important for the proof of Theorem 3.3 to note that the level of an order of type $\mathcal{O}_0(N)$ is N^{n-1}.

Now let n be odd and set $\mathcal{O} := \mathcal{O}_0(N)$ for simplicity. We can then make the same observation as in the proof of Lemma 3.5.1, bound (3.5.2). Namely, if $\gamma_1, \gamma_2 \in \bigcup_{1\leq m\leq L} \mathcal{O}(m; z, \delta)$, then

$$\mathrm{nr}(\gamma_1\gamma_2 - \gamma_2\gamma_1) \ll \delta L^2.$$

The advantage of working with the family of $\mathcal{O}_0(N)$ is that $N \mid \mathrm{nr}(\gamma_1\gamma_2 - \gamma_2\gamma_1)$. Indeed, an easy computation shows that all commutators of $\mathcal{O}_0(N)_p$, where p is unramified, have last row congruent to the zero vector modulo $N\mathbb{Z}_p$. The claim follows since the norm can be computed locally.

The remarks above imply that $N \ll \delta L^2$ or $\mathrm{nr}(\gamma_1\gamma_2 - \gamma_2\gamma_1) = 0$. Thus, if $\delta \ll N^{1-\varepsilon}L^{-2-\varepsilon}$, then the algebra generated by $\bigcup_{1\leq m\leq L} \mathcal{O}(m; z, \delta)$ is commutative. Therefore, the same counting strategy employed in the rest of this article (counting in commutative fields) would give a strong bound in this case as well.

We apply the bounds to the amplifier as follows. Let $\delta \asymp N^{1-\varepsilon}L^{-2-\varepsilon}$. Then by (3.3.4) we have

$$|\mathcal{P}|^2 \cdot |\phi(z)|^2 \ll_n S(\mu^*)(LN)^{\varepsilon}\left(|\mathcal{P}| + |\mathcal{P}|^2 L^{-(n-1)} + S(\mu^*)^{\frac{-1}{n(n-1)}} \cdot N^{-\frac{1}{2}} L|\mathcal{P}|^2 \frac{L^{(n-1)^2 n}}{L^{(n-1)n}}\right).$$

These are essentially the same computations as in Section 3.6. We deduce in the same way that

$$|\phi(z)|^2 \ll S(\mu^*)(LN)^{\varepsilon}(L^{-1} + S(\mu^*)^{\frac{-1}{n(n-1)}} \cdot N^{-\frac{1}{2}} L^{1+n(n-1)(n-2)}).$$

We may simplify one of the exponents of L by noting that $1 + n(n-1)(n-2) \leq n^3 - 1$ (for $n \geq 2$). We then find an optimal value of $L \asymp S(\mu^*)^{1/n^4(n-1)} N^{1/2n^3}$, so that

$$|\phi(z)|^2 \ll S(\mu^*)^{1-\frac{1}{n^4(n-1)}+\varepsilon} \cdot N^{-\frac{1}{2n^3}+\varepsilon}.$$

This proves Theorem 3.3.

The sup-norm bound we obtain is unfortunately not uniform in the full volume aspect, since we did not include the discriminant of A in the bounds. It is possible to include ramified primes p where A_p is a division algebra, since then p divides the norm of commutators. Nevertheless, in the composite degree case considered here there is also the possibility of A_p being a more general matrix algebra over a division algebra, which we are not able to treat using this approach. It would certainly be interesting to at least extend these bounds to include the full discriminant, but even more so to find a more flexible argument to treat arbitrary orders.

For example, the argument does not apply to the larger family of generalised Eichler orders, which we define to be intersections of two maximal orders. One example in degree 4, which is also a hereditary order, has the form

$$\mathcal{O}_p = \begin{pmatrix} \mathbb{Z}_p & \mathbb{Z}_p & \mathbb{Z}_p & \mathbb{Z}_p \\ \mathbb{Z}_p & \mathbb{Z}_p & \mathbb{Z}_p & \mathbb{Z}_p \\ p\mathbb{Z}_p & p\mathbb{Z}_p & \mathbb{Z}_p & \mathbb{Z}_p \\ p\mathbb{Z}_p & p\mathbb{Z}_p & \mathbb{Z}_p & \mathbb{Z}_p \end{pmatrix},$$

at a prime p. In this case, the norm of a commutator need not be divisible by p.

Bibliography

[AL70] A. O. L. Atkin and J. Lehner. 'Hecke operators on $\Gamma_0(m)$'. *Math. Ann.* 185 (1970), pp. 134–160.

[Ass a] E. Assing. 'On sup-norm bounds part I: ramified Maaß newforms over number fields'. *J. Eur. Math. Soc.* (to appear).

[AU95] A. Abbes and E. Ullmo. 'Comparaison des métriques d'Arakelov et de Poincaré sur $X_0(N)$'. *Duke Math. J.* 80.2 (1995), pp. 295–307.

[BH10] V. Blomer and R. Holowinsky. 'Bounding sup-norms of cusp forms of large level'. *Invent. Math.* 179.3 (2010), pp. 645–681.

[BH97] C. J. Bushnell and G. Henniart. 'An upper bound on conductors for pairs'. *J. Number Theory* 65.2 (1997), pp. 183–196.

[BHM16] V. Blomer, G. Harcos and D. Milićević. 'Bounds for eigenforms on arithmetic hyperbolic 3-manifolds'. English. *Duke Mathematical Journal* 165.4 (2016), pp. 625–659.

[BHM20] V. Blomer, G. Harcos and P. Maga. 'Analytic properties of spherical cusp forms on GL(n)'. *J. Anal. Math.* 140.2 (2020), pp. 483–510.

[Blo+20] V. Blomer, G. Harcos, P. Maga and D. Milićević. 'The sup-norm problem for GL(2) over number fields'. *J. Eur. Math. Soc. (JEMS)* 22.1 (2020), pp. 1–53.

[BM13] V. Blomer and P. Michel. 'Hybrid bounds for automorphic forms on ellipsoids over number fields'. *J. Inst. Math. Jussieu* 12.4 (2013), pp. 727–758.

[BM15] V. Blomer and P. Maga. 'The sup-norm problem for PGL(4)'. *Int. Math. Res. Not. IMRN* 14 (2015), pp. 5311–5332.

[BM16] V. Blomer and P. Maga. 'Subconvexity for sup-norms of cusp forms on PGL(n)'. *Selecta Math. (N.S.)* 22.3 (2016), pp. 1269–1287.

[BM66] A. Borel and G. D. Mostow, eds. *Proceedings of Symposia in Pure Mathematics. Vol. IX: Algebraic groups and discontinuous subgroups*. American Mathematical Society, Providence, RI, 1966, pp. vii+426.

[Bor19] A. Borel. *Introduction to arithmetic groups*. Vol. 73. University Lecture Series. Translated from the 1969 French original [MR0244260] by Lam Laurent Pham, Edited and with a preface by Dave Witte Morris. American Mathematical Society, Providence, RI, 2019, pp. xii+118.

[Bor66] A. Borel. 'Density and maximality of arithmetic subgroups'. *J. Reine Angew. Math.* 224 (1966), pp. 78–89.

[BP16] V. Blomer and A. Pohl. 'The sup-norm problem on the Siegel modular space of rank two'. *Amer. J. Math.* 138.4 (2016), pp. 999–1027.

[Bru06] F. Brumley. 'Effective multiplicity one on GL_N and narrow zero-free regions for Rankin-Selberg L-functions'. *Amer. J. Math.* 128.6 (2006), pp. 1455–1474.

[BT20] F. Brumley and N. Templier. 'Large values of cusp forms on GL_n'. *Selecta Math. (N.S.)* 26.4 (2020), Paper No. 63, 71.

[Bum97] D. Bump. *Automorphic forms and representations*. Vol. 55. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, 1997, pp. xiv+574.

[Cas97] J. W. S. Cassels. *An introduction to the geometry of numbers.* German. Repr. of the 1971 ed. Class. Math. Berlin: Springer, 1997.

[Coh08] P. M. Cohn. *Skew fields. Theory of general division rings.* English. Paperback reprint of the hardback edition 1995. Vol. 57. Encycl. Math. Appl. Cambridge University Press, 2008.

[Eve19] J.-H. Evertse. 'Mahler's Work on the Geometry of Numbers'. *Documenta Mathematica* Extra Volume Mahler Selecta (2019), pp. 29–43.

[Gil20] N. Gillman. 'Explicit subconvexity savings for sup-norms of cusp forms on $\mathrm{PGL}_n(\mathbb{R})$'. *J. Number Theory* 206 (2020), pp. 46–61.

[Gol06] D. Goldfeld. *Automorphic forms and L-functions for the group* $\mathrm{GL}(n, \mathbb{R})$. Vol. 99. Cambridge Studies in Advanced Mathematics. With an appendix by Kevin A. Broughan. Cambridge University Press, Cambridge, 2006, pp. xiv+493.

[HB02] D. R. Heath-Brown. 'The density of rational points on curves and surfaces'. *Ann. of Math. (2)* 155.2 (2002), pp. 553–595.

[Hel84] S. Helgason. *Groups and geometric analysis*. Vol. 113. Pure and Applied Mathematics. Integral geometry, invariant differential operators, and spherical functions. Academic Press, 1984, pp. xix+654.

[HS20] Y. Hu and A. Saha. 'Sup-norms of eigenfunctions in the level aspect for compact arithmetic surfaces, II: newforms and subconvexity'. *Compos. Math.* 156.11 (2020), pp. 2368–2398.

[HT12] G. Harcos and N. Templier. 'On the sup-norm of Maass cusp forms of large level: II'. *Int. Math. Res. Not. IMRN* 20 (2012), pp. 4764–4774.

[HT13] G. Harcos and N. Templier. 'On the sup-norm of Maass cusp forms of large level. III'. *Math. Ann.* 356.1 (2013), pp. 209–216.

[Hu18] Y. Hu. *Sup norm on* PGL_n *in depth aspect*. 2018.

[Hua19] B. Huang. 'Sup-norm and nodal domains of dihedral Maass forms'. *Comm. Math. Phys.* 371.3 (2019), pp. 1261–1282.

[IK04] H. Iwaniec and E. Kowalski. *Analytic number theory*. Vol. 53. American Mathematical Society Colloquium Publications. American Mathematical Society, Providence, RI, 2004, pp. xii+615.

[IS95] H. Iwaniec and P. Sarnak. 'L^∞ norms of eigenfunctions of arithmetic surfaces'. *Ann. of Math. (2)* 141.2 (1995), pp. 301–320.

[Kle00] E. Kleinert. *Units in skew fields*. Vol. 186. Progress in Mathematics. Birkhäuser Verlag, Basel, 2000, pp. viii+80.

[KM88] M. A. Kenku and F. Momose. 'Automorphism groups of the modular curves $X_0(N)$'. *Compositio Math.* 65.1 (1988), pp. 51–80.

[KNS22] I. Khayutin, P. D. Nelson and R. S. Steiner. *Theta functions, fourth moments of eigenforms, and the sup-norm problem II*. 2022.

[Lap13] E. Lapid. 'On the Harish-Chandra Schwartz space of $G(F)\backslash G(\mathbb{A})$'. *Automorphic representations and L-functions*. Vol. 22. Tata Inst. Fundam. Res. Stud. Math. With an appendix by Farrell Brumley. Tata Inst. Fund. Res., Mumbai, 2013, pp. 335–377.

[Mah55] K. Mahler. 'On compound convex bodies. I'. *Proc. London Math. Soc. (3)* 5 (1955), pp. 358–379.

[Mar14] S. Marshall. *Upper bounds for Maass forms on semisimple groups*. 2014.

[Mar16] S. Marshall. 'Local bounds for L^p norms of Maass forms in the level aspect'. *Algebra Number Theory* 10.4 (2016), pp. 803–812.

[McD78] B. R. McDonald. 'Automorphisms of $\mathrm{GL}_n(R)$'. *Trans. Amer. Math. Soc.* 246 (1978), pp. 155–171.

[Miy89] T. Miyake. *Modular forms*. Translated from the Japanese by Yoshitaka Maeda. Springer-Verlag, Berlin, 1989, pp. x+335.

[Mor15] D. W. Morris. *Introduction to arithmetic groups*. Deductive Press, [place of publication not identified], 2015, pp. xii+475.

[Neu92] J. Neukirch. *Algebraische Zahlentheorie*. Springer-Verlag, Berlin, 1992, pp. xiii+595.

[New72] M. Newman. *Integral matrices*. Vol. Vol. 45. Pure and Applied Mathematics. Academic Press, New York-London, 1972, pp. xvii+224.

[Pie82] R. S. Pierce. *Associative algebras*. Vol. 9. Studies in the History of Modern Science. Graduate Texts in Mathematics, 88. Springer-Verlag, New York-Berlin, 1982, pp. xii+436.

[Rei75] I. Reiner. *Maximal orders*. Vol. No. 5. London Mathematical Society Monographs. Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York, 1975, pp. xii+395.

[Sah17] A. Saha. 'Hybrid sup-norm bounds for Maass newforms of powerful level'. *Algebra Number Theory* 11.5 (2017), pp. 1009–1045.

[Sah20] A. Saha. 'Sup-norms of eigenfunctions in the level aspect for compact arithmetic surfaces'. *Math. Ann.* 376.1-2 (2020), pp. 609–644.

[Sar04] P. Sarnak. 'Letter to Morawetz'. Scanned letter. 2004.

[Sch67] A. Schinzel. 'Reducibility of polynomials of the form $f(x) - g(y)$'. *Colloq. Math.* 18 (1967), pp. 213–218.

[Shi78] G. Shimura. 'The special values of the zeta functions associated with Hilbert modular forms'. *Duke Math. J.* 45.3 (1978), pp. 637–679.

[SV19] L. Silberman and A. Venkatesh. 'Entropy bounds and quantum unique ergodicity for Hecke eigenfunctions on division algebras'. *Probabilistic methods in geometry, topology and spectral theory*. Vol. 739. Contemp. Math. Amer. Math. Soc., [Providence], RI, [2019] ©2019, pp. 171–197.

[Tem10] N. Templier. 'On the sup-norm of Maass cusp forms of large level'. *Selecta Math. (N.S.)* 16.3 (2010), pp. 501–531.

[Tom23] R. Toma. 'Hybrid bounds for the sup-norm of automorphic forms in higher rank'. *Trans. Amer. Math. Soc.* 376.8 (2023), pp. 5573–5600.

[Tom24] R. Toma. *The sup-norm problem for newforms of large level on* PGL(n). 2024. arXiv: `2401.02741 [math.NT]`.

[Ven06] A. Venkatesh. 'Large sieve inequalities for GL(n)-forms in the conductor aspect'. *Adv. Math.* 200.2 (2006), pp. 336–356.

[Voi21] J. Voight. *Quaternion algebras*. Vol. 288. Graduate Texts in Mathematics. Springer, Cham, 2021, pp. xxiii+885.

[You18] M. P. Young. 'A note on the sup norm of Eisenstein series'. *Q. J. Math.* 69.4 (2018), pp. 1151–1161.