La croissance maximale des périodes automorphes et intégrales oscillatoires pour les variétés plates maximales

Bart Michels

UNIVERSITÉ PARIS XIII - SORBONNE PARIS NORD
École Doctorale Sciences, Technologies, Santé Galilée

The maximal growth of automorphic periods and oscillatory integrals for maximal flat submanifolds

THÈSE DE DOCTORAT
présentée par

Bart MICHELS
Laboratoire analyse, géométrie et applications

pour l'obtention du grade de
DOCTEUR EN MATHÉMATIQUES

soutenue le 18 juillet 2022 devant le jury d'examen composé de:

Raphaël BEUZART-PLESSIS	Université d'Aix-Marseille	Examinateur
Farrell BRUMLEY	Université Sorbonne Paris Nord	Directeur de thèse
Erez LAPID	Weizmann Institute of Science	Rapporteur
Simon MARSHALL	University of Wisconsin–Madison	Examinateur
Jasmin MATZ	University of Copenhagen	Examinatrice
Farid MOKRANE	Université Sorbonne Paris Nord	Examinateur
Angela PASQUALE	Université de Lorraine	Examinatrice

Abstract

The maximal growth of automorphic periods and oscillatory integrals for maximal flat submanifolds

This dissertation contributes to the analytic theory of automorphic periods and automorphic L-functions.

In the first part we prove an extreme value result for geodesic periods of Hecke-Maass forms on compact arithmetic hyperbolic surfaces, as the Laplacian eigenvalue grows. The proof uses the celebrated amplification method of Iwaniec and Sarnak. We also obtain a theorem on extreme values of Rankin-Selberg L-functions and draw the connection to conjectures on the maximal size of L-functions.

In the second part we prove a spectral aspect extreme value result for maximal flat periods of Hecke-Maass forms, on compact locally symmetric spaces associated to forms of $\mathbf{PGL}_3$. We discuss its significance in the wider context of extreme behavior of automorphic periods. We also prove a mean square asymptotic for maximal flat periods on more general compact locally symmetric spaces.

Our main technical contributions are the following. The first is the analysis of orbital integrals needed to prove asymptotics for a relative trace formula, together with new results on the global geometry of maximal flat submanifolds of symmetric spaces. The second is an exploration of the limits of the amplification method for toric periods on forms of $\mathbf{PGL}_n$.

Keywords: automorphic forms, trace formula, harmonic analysis

Résumé

La croissance maximale des périodes automorphes et intégrales oscillatoires pour les sous-variétés plates maximales

Cette thèse contribue à la théorie analytique des formes automorphes et des fonctions L automorphes.

Dans une première partie nous démontrons un résultat sur les valeurs extrêmes des périodes géodésiques des formes de Hecke-Maass sur les surfaces hyperboliques compactes, quand la valeur propre du Laplacien grandit. La preuve utilise la méthode d'amplification d'Iwaniec et Sarnak. Nous obtenons aussi un résultat sur les valeurs extrêmes des fonctions L de Rankin-Selberg et nous faisons le lien avec les conjectures sur les valeurs extrêmes des fonctions L.

Dans une deuxième partie nous démontrons un résultat sur les valeurs extrêmes des périodes plates maximales des formes de Hecke-Maass, sur les espaces symétriques compacts associés aux formes de $\mathbf{PGL}_3$. Nous discutons de son importance dans le contexte plus large du comportement extrémal des périodes automorphes. Nous obtenons aussi une asymptotique pour des moyennes quadratiques des périodes plates maximales sur des espaces symétriques compacts plus généraux.

Nos principales contributions techniques sont les suivantes. La première est l'analyse des intégrales orbitales utilisée pour obtenir des asymptotiques pour une formule de traces relative, ainsi que des nouveaux résultats sur la géométrie globale des sous-variétés plates maximales des espaces symétriques. La deuxième est une exploration des limites de la méthode d'amplification pour les périodes toriques sur les formes de $\mathbf{PGL}_n$.

Mots clés : formes automorphes, formule des traces, analyse harmonique

Acknowledgements

This dissertation marks the end of a work influenced by conversations with many people, mathematical and non-mathematical alike, whose ideas and feedback may be reflected in subtle ways in this manuscript and make up its unique genome, and for which I am thankful to each and every one of them.

My gratitude goes first and foremost to my advisor, Farrell Brumley, for his guidance to and through the questions that I went on to explore, for sharing his metamathematical and strategic thinking that have shaped my way of practicing mathematics for a lifetime, and not least for his relentless encouragement. Part of the research was carried out at the Hausdorff Research Institute for Mathematics during the summer of 2021, and at the University of Bonn during the 2021-2022 academic year, so that I can only thank these institutions, together with Valentin Blomer for his hospitality and fruitful research ecosystem in Bonn. I am very thankful to Simon Marshall and Jasmin Matz for their careful reading of an earlier version of the manuscript, and to Gergely Harcos and Erez Lapid for taking on the role of reviewers of this thesis.

Finally I must thank my friends, family and soon-to-be family for distracting me sometimes forcefully from what has been a very addictive occupation.

Bonn, June 2022

Contents

Introduction

0.1 The maximal growth of L-functions

A key role in the theory of automorphic forms is played by the automorphic L-functions, which are the subject of major open problems. When π is a unitary automorphic representation of GL_n over $\mathbb{Q}$ with contragredient $\widetilde{\pi}$, we normalize the Godement-Jacquet L-function $L(\pi, s)$ to have functional equation relating $L(\pi, s)$ to $L(\widetilde{\pi}, 1-s)$, and we denote by $C(\pi)$ the analytic conductor of Iwaniec and Sarnak [38], a measure of the complexity of the representation. We may consider families of representations by allowing certain parameters to vary, and we also speak of aspects. A classical example is the family of unramified unitary Hecke characters, leading to vertical shifts of the Riemann zeta function: $L(|\cdot|_\infty^{it}, s) = \zeta(s+it)$ (the t-aspect for ζ), in which case the analytic conductor is of size $1+|t|$.

Many problems in analytic number theory are related to understanding the size of the central value $L(\pi, \frac{1}{2})$ as $C(\pi) \to \infty$ in families. For every reasonable family we can formulate the Lindelöf hypothesis that $L(\pi, \frac{1}{2}) \ll_\epsilon C(\pi)^\epsilon$, the subconvexity problem which asks for a $\delta > 0$ such that $L(\pi, \frac{1}{2}) \ll C(\pi)^{1/4-\delta}$, as well as the following question, which serves as a motivation for the first part of this thesis.

When $\mathcal{F}$ is a reasonable family of L-functions (not necessarily known to be automorphic) for $\mathbf{GL}_n$, how fast does

$$\max_{\substack{L \in \mathcal{F} \\ C(L) \leq x}} |L(\tfrac{1}{2})| \tag{0.1}$$

grow as $x \to \infty$?

Here, we denote $C(L)$ for the analytic conductor of an L-function with known functional equation, as in [16]. In [25], the authors use random models for L-functions to motivate the conjecture that the true growth of (0.1) is

$$\max_{\substack{L \in \mathcal{F} \\ C(L) \leq x}} |L(\tfrac{1}{2})| = \exp\left((C_{\mathcal{F}} + o(1))\sqrt{\log x \cdot \log\log x}\right) \tag{0.2}$$

for some explicit constant $C_{\mathcal{F}}$ that depends on the family. Proofs conditional on GRH [52, 56] or using the resonance method of Soundararajan or Hilberdink [35, 68] typically

yield lower bounds of quality

$$\max_{\substack{L\in\mathcal{F}\\ C(L)\leq x}} |L(\tfrac{1}{2})| \gg \exp\left(C\sqrt{\frac{\log x}{\log\log x}}\right) \tag{0.3}$$

for some $C>0$. Recent progress as well as lower bounds with additional restrictions on the argument of $L(\frac{1}{2})$ can be found in [8], [9], [17].

The following theorem about extreme central values of Rankin-Selberg L-functions is a first example of a spectral aspect family, giving further evidence towards the general conjecture (0.2).

Theorem 1. *Let $N>1$ be a square-free integer and let F be a real quadratic number field of square-free discriminant coprime to N. Assume that N has an even number of prime divisors and that they are all inert in F. Let $(f_j)_{j\geq 1}$ be an orthonormal basis of the space of newforms of level $\Gamma_0(N)$ with eigenvalues $\lambda_j>0$. There exists $C>0$ such that*

$$\max_{|\sqrt{\lambda}_j-\sqrt{\lambda}|\leq C} L(1/2,f_j)L(1/2,f_j\times\omega_F) \geq \exp\left(\sqrt{\frac{\log\lambda}{\log\log\lambda}}\,(1+o(1))\right), \tag{0.4}$$

as $\lambda\to\infty$, where ω_F is the quadratic Dirichlet character associated to the field extension $F/\mathbb{Q}$ by class field theory, and the implicit constant depends on N and F.

We can for example take $N=6$ and $F=\mathbb{Q}(\sqrt{5})$ in Theorem 1. The theorem is proved via a similar statement about geodesic periods on compact arithmetic quotients of the hyperbolic plane $\mathbb{H}$, which is Theorem 2 below, and uses an explicit version of a formula of Waldspurger [75, 59] and a characterization of the image of the Jacquet-Langlands correspondence.

Theorem 2. *Let R be an Eichler order of square-free level in an indefinite quaternion division algebra over $\mathbb{Q}$. Let $\Gamma\subset \mathrm{PSL}_2(\mathbb{R})$ be the corresponding co-compact arithmetic lattice. Let (ϕ_j) be an orthonormal basis of $L^2(\Gamma\backslash\mathbb{H})$ consisting of Laplace–Hecke eigenfunctions with Laplacian eigenvalues $\lambda_j\geq 0$. Let $\ell\subset\Gamma\backslash\mathbb{H}$ be a closed geodesic and denote by $\mathscr{P}_\ell(\phi_j)$ the period of ϕ_j along ℓ. Then there exists $C>0$ such that*

$$\max_{|\sqrt{\lambda}_j-\sqrt{\lambda}|\leq C} \lambda_j^{1/4}\,|\mathscr{P}_\ell(\phi_j)| \geq \exp\left(\frac{1}{2}\sqrt{\frac{\log\lambda}{\log\log\lambda}}\left(1+O\left(\frac{\log\log\log\lambda}{\log\log\lambda}\right)\right)\right) \tag{0.5}$$

as $\lambda\to\infty$, where the implicit constant depends on R and ℓ.

We refer the reader to §1.1 for the notations and definitions used in Theorems 1 and 2.

Remark 0.6. Note the normalizing factor $\lambda_j^{1/4}$ in the left-hand side in (0.5). Its presence can be justified in two ways. One way is using a mean square asymptotic for geodesic periods, and the other is through Waldspurger's formula. We use the notations from the theorem. In [80] it is shown that

$$\sum_{\lambda_j \leq \lambda} |\mathscr{P}_\ell(\phi_j)|^2 = C_\ell \lambda^{1/2} + O(1) \qquad (\lambda \to \infty)\,, \tag{0.7}$$

for some constant $C_\ell > 0$. This implies that $|\mathscr{P}_\ell(\phi_j)| \ll 1$ (the implicit constant depending on ℓ), and we refer to this as the convexity bound. By the Weyl law, the above sum consists of $\asymp \lambda$ terms, so that $\mathscr{P}_\ell(\phi_j)^2 \asymp \lambda^{-1/2}$ 'on average'. This explains why $\lambda_j^{1/4}$ should be the correct normalizing factor in (0.5).

Through a formula of Waldspurger, the Lindelöf hypothesis for certain L-functions leads to the conjecture that $\lambda_j^{1/4}|\mathscr{P}_\ell(\phi_j)| \ll_\epsilon \lambda_j^\epsilon$ [61]. The best known unconditional improvement of the convexity bound is $\lambda_j^{1/4}|\mathscr{P}(\phi_j)| \ll_\epsilon \lambda_j^{1/4-1/24+\epsilon}$ [47]. A geodesic period is 'large' when the normalized period $\lambda_j^{1/4}|\mathscr{P}_\ell(\phi_j)|$ is substantially larger than 1. Thus Theorem 2 can be viewed as producing abnormally large values of the geodesic periods.

Theorem 2 uses the amplification method introduced in the seminal article of Iwaniec and Sarnak [37], and crucially relies on the presence of the Hecke operators, and the fact that many Hecke returns fix a given set (a closed geodesic). The amplification method bears much similarity to the resonance method of Soundararajan in the context of L-functions, and in fact, we will follow the proposal in [50] and refer to it as the (spectral) resonance method. It may be summarized as follows: Construct a "resonator" $R(f) \geq 0$ with the property that a quotient of the form

$$\frac{\sum_j R(f_j)|\mathscr{P}_\ell(f_j)|^2}{\sum_j R(f_j)} \tag{0.8}$$

is large. If this quotient is bigger than a real number $M > 0$, it must be that at least one $|\mathscr{P}_\ell(f_j)|^2$ is bigger than M. When $R(f)$ is defined as a Hecke eigenvalue of f, a sum of the shape of the numerator naturally appears in the relative trace formula, and a sum of the shape of the denominator naturally appears in the trace formula.

The resonance method gives a way to quantify the phenomenon that eigenfunctions have strong concentration properties at sets that are fixed by many symmetries, a phenomenon that is also observed for zonal spherical harmonics and Gaussian beams on round spheres.

The proof of Theorem 2 adapts two main pre-existing ingredients. The first is bounds for archimedean orbital integrals in a relative trace formula, proved by Marshall [47], and the second is the resonator sequence of Milićević [50]. The theorem is an analogue

of a theorem of Milićević [50], replacing CM-points by closed geodesics. The motivation in [37] and [50] is to study the growth of the sup-norms $\|\phi_j\|_\infty$, and the results show in particular that those norms are unbounded as the eigenvalue grows. With the notations as in Theorem 2, let $z \in \Gamma\backslash\mathbb{H}$ be a CM-point. It is shown in [37] that the sequence of point evaluations $|\phi_j(z)|$ is unbounded: there exists a subsequence of (ϕ_j) on which

$$|\phi_j(z)| \gg_z \sqrt{\log\log\lambda_j}\,. \tag{0.9}$$

In [50], the method was optimized and this was strengthened to the existence of a subsequence on which

$$|\phi_j(z)| \gg_z \exp\left(C\sqrt{\frac{\log\lambda_j}{\log\log\lambda_j}}\right) \tag{0.10}$$

for all $C < 1$ (and in fact, with the constant C replaced by 1 plus an error term, as in Theorem 2).

0.2 Heuristics and randomness

In this section, we put the lower bound from Theorem 2 in a different perspective, together with the lower bounds (0.9) and (0.10) on sup norms.

The fact that on arithmetic hyperbolic surfaces the sequence $\|\phi_j\|_\infty$ is unbounded can be explained through random models for L-functions, as remarked above, but can also be explained by random behavior of Laplacian eigenfunctions. We explore how much growth can be explained by randomness of eigenfunctions, and we discuss some recent developments and related ideas.

Unless otherwise stated, let X be a compact Riemannian surface with negative sectional curvature. (Some of the results and conjectures stated will apply to arbitrary curvature or arbitrary dimension, but generality is not our objective.) Let (ϕ_j) be an orthonormal basis of $L^2(X)$ consisting of Laplacian eigenfunctions with eigenvalues λ_k. (If X is a noncompact finite-volume quotient of $\mathbb{H}$, we take it to be an orthonormal basis of the space of Maass cusp forms.) Because X is compact and of negative curvature, its geodesic flow is ergodic, and the *random wave conjecture* of Berry [5] predicts that Laplacian eigenfunctions of large eigenvalue should show Gaussian random behavior. Berry's conjecture has been tested numerically in [3, 4] for certain compact hyperbolic surfaces of genus 2, and in [32] for the modular surface $\mathrm{PSL}_2(\mathbb{Z})\backslash\mathbb{H}$ (which, although non-compact, still has ergodic geodesic flow). In this last article, the authors propose the following mathematical interpretation of Berry's conjecture: equip X with the normalized Riemannian probability measure $\mathrm{Vol}(X)^{-1}d\,\mathrm{Vol}$. Then the sequence of eigenfunctions (ϕ_j), considered as random variables on X, converges in law to the normal distribution with mean 0 and variance $\mathrm{Vol}(X)^{-1}$. The random wave conjecture has been the subject of recent work [1] where it has been interpreted in terms of Benjamini–Schramm convergence

of manifolds. In [1, Theorem 4] it is also shown that this version of the conjecture implies the quantum unique ergodicity conjecture (QUE) of Rudnick and Sarnak [62], which states that the probability measures $\phi_j^2 d\,\mathrm{Vol}$ converge weakly to the uniform measure $\mathrm{Vol}(X)^{-1} d\,\mathrm{Vol}$.

Even though Berry's conjecture remains widely open in any interpretation, it is natural to wonder about the finer statistical properties of Laplacian eigenfunctions. For example, when the random wave conjecture is regarded as an analogue of the central limit theorem, one could ask what the analogue of the law of the iterated logarithm should be. That is, what can be said about the growth of the sup norm $\|\phi_j\|_\infty$ as $j \to \infty$? In order to obtain predictions for this type of question, it is common to study *random wave models*, probability measures on a space of functions. Following [81, §6], we may model an eigenfunction ϕ of large eigenvalue λ by the random quasimode

$$\frac{1}{(\sum_j c_j^2)^{1/2}} \sum_j c_j \phi_j$$

where the c_j are i.i.d. standard Gaussians and the sums run over the set $\{j : \sqrt{\lambda_j} \in [\sqrt{\lambda}, \sqrt{\lambda}+1]\}$. We thus expects the sup norm $\|\phi\|_\infty$ to generically have the same behavior as the sup norm of random Fourier series studied in [39, §6], [63, Chapter IV]. That is, that $\|\phi_j\|_\infty \asymp \sqrt{\log \lambda_j}$ for a subsequence of (ϕ_j) [32, 65]. This is compatible with the sup norm conjecture of Iwaniec-Sarnak [37] and with (0.9). As for (0.10), note that the right-hand side is eventually smaller than any power of λ_j, and eventually larger than any power of $\log \lambda_j$. Such large sup norms, while not in contradiction with the sup norm conjecture nor the random wave conjecture, should be considered exceptional and are presumably specific to the case of arithmetic manifolds. Considering these results, it is somewhat surprising that despite the presence of Hecke operators, numerical evidence indicates that the random wave conjecture does hold for arithmetic hyperbolic surfaces.

We can think of at least two ways to generalize questions about random behavior of the ϕ_j. On the one hand, one can fix a 'thin' set S, such as a finite set or a geodesic segment, equipped with its induced Riemannian metric and volume element μ_S, and ask about the value distribution of the restrictions $\phi_j|_S$. It is clear that any result must be very sensitive to the nature of the set S. We illustrate this with an example. Quantum (unique) ergodic restriction, Q(U)ER, is the question of whether a density one subsequence (resp. the full sequence) of the restrictions $\phi_j^2|_S$ equidistribute. It was first studied in [72]. When X is the modular surface, the sequence (ϕ_j) of Maass cusp forms can be chosen to consist of odd and even cusp forms. The odd ones vanish on the split geodesic $S = [i, i\infty)$, and one expects QUER to hold for the restrictions to S of the even Maass forms, but where the limiting measure is $2\mu_S$ instead of μ_S [78]. We mention also [79], where the analogue of QER for Eisenstein series is found to have a negative answer on certain divergent geodesics on the modular surface, and an analogous negative answer is conjectured for Maass cusp forms. In the other direction, QER does hold for generic

geodesics on a compact hyperbolic surface [21, 73].

The other point of view consists of randomizing the thin set. Given an eigenfunction and a geodesic segment L, we may form the line integral

$$\mathscr{P}_L(\phi_j) := \int_L \phi_j \, .$$

If we fix a real number $l > 0$, the set of geodesic segments of length l is parametrized by the unit tangent bundle S^1X. Every eigenfunction ϕ_j gives rise to a smooth function $\widetilde{\phi}_j$ on S^1X, by integrating over the segment corresponding to a point of S^1X. It is known that the sequence $\|\widetilde{\phi}_j\|_\infty$ is bounded as $\lambda_j \to \infty$ (by a constant depending on X and l) [14]. In particular, when ℓ is a closed geodesic, we recover the convexity bound $|\mathscr{P}_\ell(\phi_j)| \ll 1$, which is sharp when X is the round 2-sphere. Theorem 2 can be viewed as producing large values of the functions $\widetilde{\phi}_j$ for suitable l.

We remark that, in order to obtain heuristics for large values of $\tilde{\phi}_j$, to apply random wave models as in §0.2 would be nonsensical, because Berry's conjecture is fundamentally a statement about the value distribution of eigenfunctions, and does not take into account the relative position of points. That is, the random wave models are not designed to be integrated over sets of measure zero. Instead, it would be interesting to develop a conjecture analogous to Berry's for the value distribution of $\tilde{\phi}_j$ as $\lambda_j \to \infty$. A related type of result is the central limit theorem for geodesic flows [60, 67]: for a fixed smooth function $\phi \in C^\infty(X)$, one determines the limiting distribution of the (normalized) line integral of ϕ along a random geodesic segment of growing length.

0.3 Periods beyond GL_2

Let $\mathbf{G}$ be a semisimple algebraic $\mathbb{Q}$-group and $\mathbf{H}$ a closed algebraic subgroup that we assume to be anisotropic, meaning that the adelic points $\mathbf{H}(\mathbb{A}_\mathbb{Q})$ have compact image $[\mathbf{H}]$ in the automorphic quotient $[\mathbf{G}] = \mathbf{G}(\mathbb{Q})\backslash\mathbf{G}(\mathbb{A}_\mathbb{Q})$. For any automorphic form f on $[\mathbf{G}]$ one may consider the period $\mathscr{P}_\mathbf{H}(f)$ by integrating along a suitable translate of $[\mathbf{H}]$. We may consider families of forms as before, and our focus is on the spectral aspect family of Hecke-Maass forms: Fix a maximal compact subgroup of $\mathbf{G}(\mathbb{R})$ and a level structure, and let $\mathcal{F}$ be the family of spherical cusp forms on the locally symmetric space resulting from these choices.

The second part of this thesis aims to advance our understanding of the following question.

When $\lambda(f)$ denotes the Laplacian eigenvalue of $f \in \mathcal{F}$, how fast does

$$\max_{\substack{f \in \mathcal{F} \\ \lambda(f) \leq x}} |\mathscr{P}_\mathbf{H}(f)| \tag{0.11}$$

grow, if at all, as $x \to \infty$?

To state the question correctly one needs normalizing factors in front of the periods (as in Theorem 2). Further, $\lambda(f)$ is not always the correct measure of complexity for the **H**-period. This leads to difficult questions, which we discuss below. We will be interested in sub-polynomial growth rates, whose significance is best understood in the context of extreme values of L-functions. In fact, question (0.11) is in many ways similar to question (0.1), and this similarity will be our basis for revealing possible arithmetic information about periods.

The similarity between the problems is already visible in the resonance method in [68], where for each family of L-functions under consideration the spectral ingredient is a trace formula: The Fourier inversion formula in the case of ζ, and the Petersson trace formula in the case of L-functions of modular forms. In the proofs of (0.9), (0.10) and Theorem 2, this is replaced by the Selberg trace formula and a relative trace formula, as we explained around (0.8).

Now, while question (0.11) is in some aspects similar to question (0.1), there are important differences. The first is in the problem statement itself: Where for L-functions the correct measure of complexity is the analytic conductor, the situation is much more delicate for periods. The Laplacian eigenvalue $\lambda(f)$ is a naive choice, and the correct replacement is most likely an integral involving an approximate spectral projector around f, which determines the size of mean square asymptotics for periods over $O(1)$ spectral windows (as in Theorem 4). For sufficiently generic spectral parameters, that integral should be of size $\lambda(f)^{(n-r-d)/2}$, with n the dimension of the symmetric space, r its rank, and d the dimension of the projection to the symmetric space of the **H**-orbit underlying the period. But asymptotically evaluating this integral in terms of the spectral parameter of f, even for generic parameters, is a difficult problem.

The second difference is the following. Whereas in the case of L-functions one would expect the resonance method to always produce a nontrivial result of quality (0.3) (provided that we know enough about the family under consideration), this not the case for periods. In fact, the spectral resonance method is sometimes fruitless, and sometimes produces extreme values of periods with power growth. Moreover, a larger variety of techniques exist to prove lower bounds for periods. We give a brief historical account of these facts.

The first example of a different technique is given in [62], which is about discrete periods on certain hyperbolic 3-manifolds. The authors use what is known as a distinction method. It employs a vanishing property that says that only a sparse subsequence of Hecke-Maass forms, lying in the image of a theta lift, have nonzero **H**-period. They are "distinguished" by **H**. This is then contrasted with a mean square asymptotic, which gives the full average of the periods (and which does not see the arithmetic underlying the distinction). If the periods are supported on a sparse subsequence and the average does not know about this, it follows that that the periods must attain large values on the distinguished subsequence. In fact, the sequence is polynomially sparse, and the authors obtain periods with power growth. Note that while the arithmetic properties used are

crucial, the Hecke operators do not play a direct role.

The article [51] characterizes the hyperbolic 3-manifolds to which the proof extends, as being those of Maclachlan-Reid type. Moreover, it gives the first example of the second phenomenon mentioned above: power growth obtained from the resonance method. Further power growth results that use the distinction method include the following settings: discrete periods on hyperbolic n-manifolds with $n \geq 5$ [18], and discrete unitary periods for $\mathbf{GL}_n$ (which includes the case of hyperbolic 2 and 3-manifolds) [43].

A vast generalization of power growth results in the other direction, using the resonance method, is the article [13]. In certain situations, one can obtain exceptional sequences of eigenfunctions with periods of power growth using both methods, and show that they are related. We refer to [13, 51] for a discussion of the relation between the two methods. We do note the following: All $\mathbb{Q}$-groups in these examples are such that $\mathbf{G}(\mathbb{R})$ is not split.

0.4 Sub-polynomial growth

The second part of this thesis is concerned with applications of the resonance method to periods of Hecke-Maass forms in the spectral aspect, in situations that have so far remained gray zones. By "gray zone", we mean the following. A first condition is that no explicit connection to central L-values is known or conjectured. Indeed, when conjecture (0.2) does not apply, without heuristics that come from applications of random matrix theory to L-functions, it becomes an interesting question whether there are sufficiently strong arithmetic reasons for the existence of unusually large periods, and if so, what the maximal growth should be. A second condition is that no power growth is expected, or at least that established distinction techniques or resonance techniques do not produce power growth. In the gray zone it is not clear what period growth, if present at all, should be attributed to: heuristics motivated by random behavior of Laplacian eigenfunctions in negative curvature (Berry's random wave conjecture), random behavior of more exotic underlying artihmetic objects, or neither of these?

In fact, our hope when exploring the gray zone is to reveal arithmetic information: If in some new situation we obtain lower bounds of quality (0.3), this would be a very strong hint that there is an as of yet unknown relation with central L-values. If we obtain periods with power growth, this might indicate the existence of a functorial lift (although it could be attributed to other factors as well; see [13]). If we obtain growth rates smaller than (0.3) or nothing at all, this still leaves the possibility of a relation with L-values further to the right of $\frac{1}{2}$, or at worst in the half-plane of convergence.

An example of such an unexplored situation is that of toric periods on locally symmetric spaces of non-compact type associated to a semisimple group $\mathbf{G}$. Unless $\mathbf{G}$ is isogenous to a product of forms of $\mathbf{PGL}_2$, these fall in the gray zone. When the locally symmetric space X is viewed as a disjoint union of classical locally symmetric spaces, then the period

along a maximal $\mathbb{R}$-split torus corresponds to the integral along a flat submanifold of X, which has the property that the intersection with each component is either maximal flat or empty; see §3.5.4.

To state the main theorems we introduce some minimal notation for spectral parameters. When G is a connected semisimple Lie group with finite center, make a choice of Iwasawa decomposition $G = NAK$. Let $\mathfrak{a} = \mathrm{Lie}(A)$ and define the generic set $(\mathfrak{a}^*)^{\mathrm{gen}} \subset \mathfrak{a}^*$ as the set of elements that are regular and that do not lie in a proper subspace spanned by roots. A locally symmetric space X is assumed to be compatible with the choice of K. When $\mathbf{G}$ is a semisimple group over $\mathbb{Q}$, we may define the above notation with respect to $\mathbf{G}(\mathbb{R})^0$, and we again assume an associated adelic locally symmetric space X to be compatible with K. The spectral parameters of Hecke-Maass forms on X can then be viewed as elements of $(\mathfrak{a}_{\mathbb{C}})^*$. For any additional notation and terminology used in the theorems below, we refer to §2.1, §3.1 and §3.5.

We can now state the following theorem.

Theorem 3. *Let $\mathbf{G}$ be an anisotropic $\mathbb{Q}$-form of $\mathbf{PGL}_3$ with $\mathbf{G}(\mathbb{R})$ noncompact. Let X be an associated adelic locally symmetric space and $(f_j) \in L^2(X)$ an orthonormal basis of Hecke-Maass forms with spectral parameters $\nu_j \in (\mathfrak{a}_{\mathbb{C}})^*$. Let $\mathbf{H} \subset \mathbf{G}$ be a maximal torus with the same $\mathbb{R}$-rank r as $\mathbf{G}$, and denote by $\mathscr{P}_{\mathbf{H}}(f_j)$ the $\mathbf{H}$-period of f_j. Let $D_{\mathfrak{a}^*} \subset (\mathfrak{a}^*)^{\mathrm{gen}}$ be compact. There exist $C > 0$ and $\delta > 0$ such that uniformly for $\nu \in D_{\mathfrak{a}^*}$ and $t \in \mathbb{R}$ we have*

$$\max_{\|\nu_j - t\nu\| \leq C} (1+t)^r \, |\mathscr{P}_{\mathbf{H}}(f_j)|^2 \gg (\log\log(2+t))^{\delta + o(1)} .$$

Moreover, when E is the splitting field of $\mathbf{H}$, we may take $\delta = 6/[E : \mathbb{Q}]$.

In §3.5.2 we give the list of groups to which Theorem 3 can be applied. There we also show that the associated Lie group $\mathbf{G}(\mathbb{R})$ is either $\mathrm{PGL}_3(\mathbb{R})$ or the quasi-split projective unitary group $\mathrm{PU}(2,1)$, and in these cases the $\mathbb{R}$-rank equals 2 or 1 respectively.

At first sight, the double logarithmic rate in Theorem 3 may seem disappointing when compared with results of this type for (forms of) $\mathbf{PGL}_2$ (§0.1), or with results that give polynomial growth of periods (§0.3). But in fact, we have a strong reason to believe that in this case the double logarithmic growth is best possible, up to the exponent δ, and this is simultaneously the answer to the question raised earlier: What must the nontrivial growth in Theorem 3 be attributed to? We believe the reason for this growth is arithmetic, for the following reason. There is an exceptional theta-correspondence between $\mathbf{PGL}_3$ and the exceptional group $\mathbf{G}_2$, which is related to the maximal toric periods in the theorem. In fact, the period $\mathscr{P}_{\mathbf{H}}(f_j)$ should be, up to other arithmetic factors that are equally mysterious, related to a product of L-values

$$L(\pi_{f_j}, 1) L(\widetilde{\pi}_{f_j}, 1) ,$$

where π_{f_j} denotes the representation generated by f_j and $\widetilde{\pi}_{f_j}$ denotes its contragredient. This period relation, while not proven, explains a great deal about the lower bound in Theorem 3. First, it leads us to believe that the lower bound can, perhaps, not simply be attributed to random behaviour of Laplacian eigenfunctions on Riemannian manifolds of nonpositive curvature. But in fact, much more can be said. The period formula explains why the growth rate we obtain should be polynomial in $\log\log C(L(\pi_{f_j}, s))$, the double log of the analytic conductor! Indeed, the double logarithm reminds us of the following result of Levinson [45]: There exists a constant $C > 0$ such that for arbitrarily large $t \in \mathbb{R}$, one has

$$|\zeta(1 + it)| \geq e^{\gamma} \log\log t - C\,.$$

This is certainly not the best known result, but the main term in the right-hand side is conjectured to be optimal [29]. We refer to [2] for the state of the art on the extreme values of $\zeta(1 + it)$, and results with lower order terms. The results and conjectures for $\zeta(1 + it)$ are exemplary for the more general situation. In fact, there is the following conditional statement [6]: When π is a unitary cuspidal tempered automorphic representation for $\mathbf{GL}_n$ whose Godement-Jacquet L-function $L(\pi, s)$ satisfies the generalized Riemann hypothesis, then

$$(\log\log C(\pi))^{-n} \ll |L(\pi, 1)| \ll (\log\log C(\pi))^{n}\,,$$

where the implicit constants depend on n only and we denote $C(\pi)$ for the analytic conductor of $L(\pi, s)$. Moreover, if π is not tempered, one still has similar bounds but with bigger exponents.

Coming back to the interpretation of Theorem 3, we conclude the following. First, ignoring for the sake of the argument the other factors in the period formula mentioned above, it is to be expected that the maximal toric periods exhibit oscillations that are polynomial in $\log\log\lambda_j$. Indeed, this is (at least conjecturally) the largest permitted oscillation of the L-value $L(\pi_{f_j}, 1)L(\widetilde{\pi}_{f_j}, 1)$, and based on what we know about $\zeta(1 + it)$ we would expect this maximal oscillation to be almost realized. Second, if forced to make a conjecture about the maximal growth of the periods in Theorem 3, it would be that the exponent of $\log\log(2+t)$ in the right-hand side can be replaced by $12+o(1)$. (Taking into account the fact the periods appear with a square in the left-hand side.)

About the exponent in the right-hand side, we remark the following: The splitting field E of $\mathbf{H}$ is Galois, and the Galois group embeds naturally as a subgroup of $\mathrm{GL}_2(\mathbb{Z})$ [12, §1.7]. It is well known that finite subgroups of $\mathrm{GL}_2(\mathbb{Z})$ have cardinality at most 12; this is most easily seen by observing that the eigenvalues of a matrix of finite order must be roots of unity. Thus $[E : \mathbb{Q}] \leq 12$, meaning that $\delta = \frac{1}{2}$ is admissible. In the best case, the Galois group is of cardinality 3 and we obtain the exponent 2; still far from the (naive) conjecture that it can be replaced by 12.

The main arithmetic ingredient that goes into the proof of Theorem 3 is the optimization problem of finding the best possible resonator sequence. This is what explains the

restriction to forms of $\mathbf{PGL}_3$. The optimization problem takes as input asymptotics for local p-adic integrals arising in the geometric main term in a relative trace formula, and requires us to construct a suitable Hecke operator. For (forms of) $\mathbf{PGL}_n$ with $n \geq 3$, only for $n = 3$ have we found a suitable winning construction that, when plugged into the optimization problem, yields nonconstant growth. We strongly believe that for $n \geq 4$ no such construction exists, and that the resonance method cannot produce any growth of toric periods for $n \geq 4$. If indeed true, we believe that, roughly speaking, this should be attributed to the heuristic that tori inside $\mathbf{PGL}_3$ are still relatively large, while they are too small inside $\mathbf{PGL}_n$, $n \geq 4$. We refer to §3.7.3 and Remark 3.112 for these negative statements.

0.5 A mean square asymptotic

The main analytic ingredient in the proof of Theorem 3 is a (amplified) mean square asymptotic for maximal flat periods. We require averages over spectral windows of bounded size, and even a non-amplified version requires a considerable amount of work. For this ingredient there is no reason to restrict to the Lie groups associated to groups in Theorem 3, and we prove a more general result, formulated classically in terms of periods along maximal flat submanifolds.

Theorem 4. *Let G be a noncompact connected semisimple Lie group with finite center and rank r, and let X be an associated compact Riemannian locally symmetric space. Assume that at least one of the following holds:*

- *G has rank 1;*
- *$G = \mathrm{SL}_p(\mathbb{R})$ or $\mathrm{SU}(k, p-k)$ with p prime and $0 < k < p$, and X arises from a $\mathbb{Q}$-form of $\mathbf{SL}_p$.*

Let $(f_j) \in L^2(X)$ be an orthonormal basis of Maass forms. Let $\mathscr{F} \subset X$ be a compact maximal flat submanifold and denote by $\mathscr{P}_{\mathscr{F}}(f_j)$ the period of f_j along $\mathscr{F}$. Let $D_{\mathfrak{a}^*} \subset (\mathfrak{a}^*)^{\mathrm{gen}}$ be compact. There exists $C > 0$ such that uniformly for $\nu \in D_{\mathfrak{a}^*}$ and $t \in \mathbb{R}$ we have

$$\sum_{\|\nu_j - t\nu\| \leq C} |\mathscr{P}_{\mathscr{F}}(f_j)|^2 \asymp \beta(t\nu) \cdot (1+t)^{-r},$$

where β denotes the Plancherel density.

The restriction to certain locally symmetric spaces in Theorem 4 comes from our bounds for orbital integrals in the relative trace formula. These are defined in terms of a parameter $g \in G$, and behave differently when g centralizes a proper portion of the group underlying the period. That is, when g lies in the standard Levi subgroup of a semistandard parabolic other than G or the minimal one (see (2.1.4)). We have not been

able to deal with the integrals for such g, which is why we restrict to settings where they do not appear in the relative trace formula.

Theorem 4 tells us that we may view the factor $(1+t)^r$ in the left-hand side in Theorem 3 as a normalizing factor that makes the squared periods of size 1 on average. It is consistent with the mean square asymptotic of Zelditch [80] over eigenvalue intervals, but there is no obvious implication between the two results. The relation to Zelditch's result is analogous to the relation of the spectral parameter Weyl law [19] to the classical Weyl law for compact Riemannian manifolds.

The stationary phase analysis in the proof of Theorem 4 draws inspiration from the proof of bounds for spherical functions in [20] and generalizes work of Marshall [47] for PGL_2 that we use in Chapter 1. Where for PGL_2 one relies on classical facts about the geometry of geodesics in the Poincaré upper half-plane model, the analogues of those facts were not available in the generality needed here and had to be established as well; see §0.6.

The quantification over ν and t in Theorem 3 and Theorem 4 may be visualized as follows. Fix a closed cone in the interior of $(\mathfrak{a}^*)^{\mathrm{gen}}$. Then $t\nu$ tends to ∞ inside the cone, and the maximum is taken over a ball of bounded radius around $t\nu$. Figure 0.1 gives a picture when $\mathfrak{g} = \mathfrak{sl}_3(\mathbb{R})$, and the simple roots are denoted by α and β. When $\mathfrak{g} = \mathfrak{sl}_3(\mathbb{R})$, the condition on tempered spectral parameters to not lie in a proper subspace spanned by roots, is equivalent to not being self-dual, and it also appears in [7].

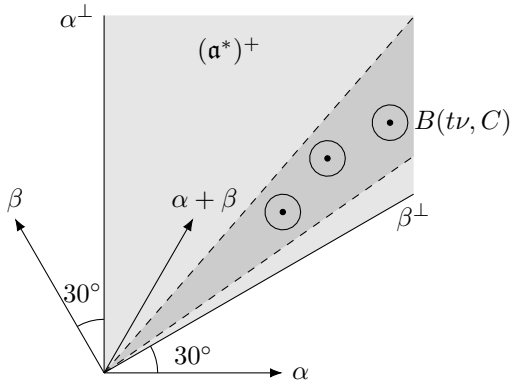

Figure 0.1: A cone in $(\mathfrak{a}^*)^{\mathrm{gen}}$ when $G = \mathrm{SL}_3(\mathbb{R})$, with balls $B(t\nu, C)$ as in Theorem 4.

0.6 The geometry of maximal flat submanifolds

Let G be a connected semisimple Lie group with finite center. In the proof of Theorem 3 and Theorem 4, we are interested in spectral asymptotics in a relative trace formula for

maximal flat submanifolds of an associated locally symmetric space. Let $G = NAK$ be a choice of Iwasawa decomposition. Define $\mathfrak{a} = \mathrm{Lie}(A)$, and recall Harish-Chandra's formula for the spherical function of parameter $i\lambda \in i\mathfrak{a}^*$:

$$\varphi_{i\lambda}(g) = \int_K e^{(\rho+i\lambda)(H(kg))}dk\,, \tag{0.12}$$

where $H : G \to \mathfrak{a}$ is the Iwasawa projection and $\rho \in \mathfrak{a}^*$ the half-sum of positive roots. In the analysis of the relative trace formula we require asymptotics for the integral

$$\int_A \varphi_{i\lambda}(a)b(a)da$$

(with a smooth cutoff $b(a)$) , as well as twisted versions thereof, as λ grows in $\mathfrak{a}^*$. The problems are relative analogues of the problem of bounding spherical functions considered in [20].

In view of (0.12), one is naturally led to study the behavior of the Iwasawa $\mathfrak{a}$-projection along sets of the form kA. More specifically, we are interested in the critical points of

$$\lambda(H(ka))$$

as a function of $a \in A$. The maximal flat submanifolds of the symmetric space G/K are the images of the sets gA. When $G = \mathrm{PSL}_2(\mathbb{R})$, then G/K is the hyperbolic plane, the maximal flats are the geodesics, and the motivating problem of bounding orbital integrals is solved in [47]. It takes as input classical facts about the geometry of geodesics in the hyperbolic plane. For general semisimple G we require analogous results for maximal flat submanifolds, which to our knowledge are not established. We state the results in the next few paragraphs.

Consider the hyperbolic plane $\mathbb{H} = \{x+iy \in \mathbb{C} : y > 0\}$ with its hyperbolic metric. The group $\mathrm{PSL}_2(\mathbb{R})$ acts on $\mathbb{H}$ by orientation-preserving isometries. We make the standard choice of Iwasawa decomposition: Let N be the subgroup of unipotent upper triangular matrices, A the subgroup of diagonal matrices and $K = \mathrm{PSO}_2(\mathbb{R})$. Then $\mathrm{PSL}_2(\mathbb{R}) \cong N\times A\times K$, the diffeomorphism being given by multiplication. The upper half-plane model naturally lends itself to describe the Iwasawa projections geometrically. The element

$$g = \begin{pmatrix} 1 & x \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \sqrt{y} & 0 \\ 0 & 1/\sqrt{y} \end{pmatrix} \in NA$$

sends i to $x + iy$. That is, the real part of gi can be identified with the N-projection of g, and the imaginary part with the A-projection.

By a (maximal) geodesic in $\mathbf{H}$ we mean the 1-dimensional submanifold defined by it, although we will sometimes informally speak of geodesics with an orientation. The geodesics in $\mathbb{H}$ are the semicircles with centers on the horizontal axis, together with the

vertical lines. The action of $\mathrm{PSL}_2(\mathbb{R})$ on i induces a diffeomorphism between A and the vertical geodesic through i. Every geodesic is of the form gAi with $g \in \mathrm{PSL}_2(\mathbb{R})$.

Observe that the real part of every geodesic is a bounded set. We generalize this observation to semisimple Lie groups G, as follows.

Theorem 5. *Let G be a connected semisimple Lie group. For all $g \in G$, the N-projection of gA is a relatively compact set.*

Theorem 5 is proved in §2.2 by diving into the mechanics of the Gram–Schmidt process, and showing that the orthogonalization part can be done with uniformly bounded operations. In fact, we will prove a stronger version with some uniformity, which is Theorem 2.17. The uniform version requires to partition the set of all maximal flats, which is naturally identified with $G/N_G(A)$, into a Zariski open 'generic' set and several lower-dimensional 'exceptional' sets. This partition generalizes the distinction between semicircles and vertical geodesics in the upper half plane $\mathbb{H}$, in which case the semicircles form the generic set. The partition is not dependent on any choice of model for the symmetric space G/K, but inherent to the choice of Iwasawa decomposition of G. Some of the lower-dimensional sets come from semistandard Levi subgroups, and it is no surprise that they are exceptional. But in general there are other exceptional sets, and the partition remains quite mysterious.

The second group of results concerns the Iwasawa A-projection. In the case of $G = \mathrm{PSL}_2(\mathbb{R})$, define the height of a point in $\mathbb{H}$ to be its imaginary part. The heights of the points of a geodesic gAi form a bounded-from-above set precisely when the geodesic is a semicircle. In that case, the height is maximized at a unique critical point, which is the midpoint of the semicircle, and the height tends to 0 at infinity on the geodesic. Such a critical point exists if and only if gAi is not vertical, meaning that $g \notin N \cdot N_G(A)$.

For a general connected semisimple Lie group G, we prove the following.

Theorem 6. *Let $\lambda \in \mathfrak{a}^*$ be an element that is positive with respect to the choice of Iwasawa decomposition, regular, and which does not lie in a proper subspace spanned by roots.*

(i) For all $g \in G$ the "height function"

$$\begin{aligned} h_{\lambda,g} : A &\to \mathbb{R} \\ a &\mapsto \lambda(H(ga)) \end{aligned}$$

has at most one critical point. If it exists, it is non-degenerate and maximizes $h_{\lambda,g}$.

(ii) The set of $g \in G$ for which a given $a \in A$ is a critical point of $h_{\lambda,g}$, is a non-empty smooth submanifold of codimension $\dim(A)$. *The set of $g \in G$ for which $h_{\lambda,g}$ has a critical point, is open.*

We prove Theorem 6 in §2.3. The proof is quite technical and occupies the larger part of this section. Regarding the second part, note that it is by no means obvious that $h_{\lambda,g}$ has a critical point for even a single g, because the domain A is noncompact. The way in which we prove existence is by varying $g \in K$, and realizing the elements g with a given critical point as the minima of a certain smooth function on the compact group K. It is likely that the set of $g \in G$ for which $h_{\lambda,g}$ has a critical point is in fact dense and that this can be proved using a very different argument, which is related to Theorem 5; see Remark 2.54.

Remark 0.13. The critical points in Theorem 6 can be thought of as giving the midpoints of the flat $gA \subset S$. It is in general too much to hope that $H(ga)$ has a critical point as a function of a. That is, the critical points can depend on $\lambda \in \mathfrak{a}^*$. (See Example 2.29.) We will not use the term "midpoint", all the more because we have found no way to generalize the observation that for $\mathrm{PSL}_2(\mathbb{R})$, the critical point corresponds to the center of the semicircle $gAi \subset \mathbb{H}$.

Remark 0.14. In Theorem 6, many things break down when λ is singular, lies in a proper subspace spanned by roots or is nonpositive: the nondegeneracy, uniqueness, and existence of critical points. Regarding non-degeneracy, see Remark 2.46. For non-uniqueness and non-existence, see §2.3.2.

Finally, we turn our attention to the Iwasawa K-projection. The Lie group $\mathrm{PSL}_2(\mathbb{R})$ is naturally identified with the unit tangent bundle $T^1(\mathbb{H})$ via its action on the vertical vector at the base point, $(i, (0, 1))$. The Iwasawa K-projection of an element g corresponds to a choice of direction at the point gi. The elements of the Weyl group $N_K(A)$ correspond to the vertical directions, up and down. As we approach infinity along a geodesic, the tangent line to the geodesic tends to a vertical one.

We generalize this observation as follows.

Theorem 7. *Let G be a connected semisimple Lie group. Let $g \in G$ and $H \in \mathfrak{a}$. Then the K-projection of ge^{tH} tends to $N_K(A)Z_K(H)$ as $t \to +\infty$.*

Theorem 7 is proved by projecting the K-projection of the geodesic flow down to the Lie algebra in a specific way and realizing it as the flow of a vector field. The resulting dynamical system is quite mysterious, but the asymptotic behavior of individual orbits can be well understood. It might seem that Theorem 7 is a statement about individual geodesics rather than maximal flats, but it is possible to formulate statements with uniformity in the variables g and H; see Remark 2.68.

In the end the results on the N- and K-projections were not needed for the stationary phase analysis in §3. But they complete the picture nicely, might be useful elsewhere, and they barely fail to provide alternative proofs of parts of Theorem 6; see Remark 2.54. A number of things remain mysterious, in particular the apparent relation between the partition in Theorem 2.17 on the N-projection, and the dynamical system used in the proof of Theorem 7; see Remark 2.68.

1 Geodesic periods on hyperbolic surfaces

This chapter contains the proofs of Theorem 1 and Theorem 2. It can be read independently from the other sections. The reader may find that the proofs here are quite detailed. This is intentional: Our hope is that it can serve as a helpful tool to read the arguments in the later sections, where we have left out some of the details for the benefit of a cleaner exposition.

The proof of Theorem 2 follows the recipe given by (0.8) and goes by a comparison of trace formulas. In §1.2 we derive both formulas. Here, a novelty is the estimation of the contribution of hyperbolic classes in the Selberg trace formula, which is an unavoidable issue given the nature of the problem; see Remark 1.14. In the relative trace formula, we identify a main term and determine its size. This involves an arithmetic computation (§1.3.1) and an analytic computation (§1.4). Bounding the error terms leads to a Diophantine problem (§1.3.2) and the analytic problem of bounding orbital integrals (§1.5). In §1.6.1, we combine the work and consider an arbitrary amplifier, obtaining an asymptotic estimate for short spectral sums. In §1.6.2 we optimize the amplifier, proving Theorem 2. In §1.6.3 we explain how to adapt the method to prove Theorem 1.

1.1 Notation

1.1.1 Lie groups

Let $G = \mathrm{PSL}_2(\mathbb{R})$ and define the subgroups $K = \mathrm{PSO}_2(\mathbb{R})$, $A = \{\left(\begin{smallmatrix} * & 0 \\ 0 & * \end{smallmatrix}\right)\}$ and $N = \{\left(\begin{smallmatrix} 1 & * \\ 0 & 1 \end{smallmatrix}\right)\}$. Fix parametrizations $a : \mathbb{R} \to A$, $k : \mathbb{R}/2\pi\mathbb{Z} \to K$ defined by

$$a(t) = \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix} \quad , \quad k(\theta) = \begin{pmatrix} \cos(\theta/2) & \sin(\theta/2) \\ -\sin(\theta/2) & \cos(\theta/2) \end{pmatrix}.$$

The action of G on the hyperbolic plane $\mathbb{H}$ induces a map $G \to \mathbb{H} : g \mapsto gi$ and a diffeomorphism $N \times A \cong \mathbb{H}$. Denote by dg the (bi-invariant) Haar measure on $G \cong \mathbb{H} \times K$ that is the product of the hyperbolic measure with the Haar measure dk of mass 1 on K. Fix any Riemannian metric on G, and denote the Riemannian distance on G by $d = d_G$.

We will often use the fact that any two Riemannian distances are locally equivalent, or more generally the following fact: when M is a smooth connected manifold, $N \subset M$ a connected submanifold, each equipped with a Riemannian metric, then for every

compact subset $L \subset N$ and $x, y \in L$ we have $d_N(x,y) \asymp_L d_M(x,y)$. In particular, taking $M = N = G$ and pulling back the metric on G on the left or the right by $g \in G$, we have for every compact $L \subset G$ and $x, y \in L$ that $d(gx, gy) \asymp_{g,L} d(xg, yg) \asymp_{g,L} d(x,y)$. Taking $N = \mathrm{SL}_2(\mathbb{R})$ equipped with any Riemannian metric and $M = M_2(\mathbb{R})$ with the Euclidean metric with respect to the standard basis, we find that for $L \subset \mathrm{SL}_2(\mathbb{R})$ compact and for matrices $x, y \in L$, $d_{\mathrm{SL}_2(\mathbb{R})}(x,y) \asymp_L \|x - y\|_2$. We also have $d_G(x,y) \asymp_L \min(d_{\mathrm{SL}_2(\mathbb{R})}(x,y), d_{\mathrm{SL}_2(\mathbb{R})}(x,-y))$.

1.1.2 Arithmetic quotients

Let B be a quaternion division algebra over $\mathbb{Q}$ that is split at ∞. That is, there exists an isomorphism of $\mathbb{R}$-algebras $\rho : B \otimes_{\mathbb{Q}} \mathbb{R} \xrightarrow{\sim} M_2(\mathbb{R})$. We view B as a subset of $B \otimes_{\mathbb{Q}} \mathbb{R}$ via the natural embedding. Denote the projection $\mathrm{GL}_2^+(\mathbb{R}) \to \mathrm{PSL}_2(\mathbb{R}) = G$ by $g \mapsto \overline{g}$, and define $\overline{\rho}(g) = \overline{\rho(g)}$ for an element $g \in (B \otimes_{\mathbb{Q}} \mathbb{R})^+$ of positive reduced (quaternion) norm. We will not always distinguish between $\eta \in B^+$ and its image $\overline{\rho}(\eta) \in G$. Let $R \subset B$ be a $\mathbb{Z}$-order. For $n \in \mathbb{N}_{>0}$, denote by $R(n) \subset B^+$ the set of elements of reduced norm equal to n, define $R^1 = R(1)$ and define $\Gamma = \overline{\rho}(R^1) \subset G$. It is well known that Γ is a lattice in G, and that the quotient $\Gamma\backslash\mathbb{H}$ is compact (see for example [74, IV: Théorème 1.1.]). We call Γ an arithmetic lattice. We fix any norm $\|\cdot\|$ on $M_2(\mathbb{R}) \cong B \otimes_{\mathbb{Q}} \mathbb{R}$.

1.1.3 Closed geodesics

By a *geodesic* in $\mathbb{H}$ we will mean a maximal geodesic, which we identify with the set of its points, disregarding the parametrization. The image of A in $\mathbb{H}$ is the geodesic joining 0 with ∞. Because G acts transitively on the unit tangent bundle of $\mathbb{H}$, it acts transitively on geodesics. If L is the geodesic that is the image of gA in $\mathbb{H}$, its set-wise stabilizer is the normalizer $N_G(gAg^{-1})$, in which gAg^{-1} has index equal to 2.

If $\Gamma \subset G$ is a discrete subgroup, a geodesic $L = gAi \subset \mathbb{H}$ projects to a smooth curve in the Riemannian orbifold $\Gamma\backslash\mathbb{H}$. It has a periodic image $\ell \subset \Gamma\backslash\mathbb{H}$ precisely when there exists $\gamma \in (\Gamma \cap gAg^{-1}) - \{1\}$ that stabilizes L. We then call ℓ a closed geodesic and denote $\Gamma_L = \mathrm{Stab}_\Gamma(L) \cap gAg^{-1}$. It is a lattice in gAg^{-1} and we have $[\mathrm{Stab}_\Gamma(L) : \Gamma_L] \in \{1, 2\}$. When this index equals 1, ℓ is called a reciprocal geodesic; see [66]. When ℓ is a closed geodesic and $\phi \in C^\infty(\Gamma\backslash\mathbb{H})$, define the period of ϕ along ℓ as the line integral

$$\mathscr{P}_\ell(\phi) := \int_{\Gamma_L\backslash L} \phi\,. \tag{1.1}$$

When Γ is an arithmetic lattice as in §1.1.2, the closed geodesics can be characterized as follows. When $F \subset B$ is a real quadratic number field, the $\mathbb{R}$-algebra $F \otimes_{\mathbb{Q}} \mathbb{R} \subset B \otimes_{\mathbb{Q}} \mathbb{R}$ is isomorphic to $\mathbb{R} \times \mathbb{R}$, hence its image under ρ is conjugate in $M_2(\mathbb{R})$ to the algebra of diagonal matrices. Thus the group $\bar{\rho}((F \otimes_{\mathbb{Q}} \mathbb{R})^1)$ of matrices of determinant ± 1 equals gAg^{-1} for a unique $g \in G/N_G(A)$. One has that gA projects to a closed geodesic in $\Gamma\backslash\mathbb{H}$

and every closed geodesic is obtained exactly once in this way: we obtain a bijection $F \mapsto L_F$ between real quadratic number fields inside B and geodesics in $\mathbb{H}$ that become closed in $\Gamma\backslash\mathbb{H}$. (In particular, whether a geodesic becomes closed in $\Gamma\backslash\mathbb{H}$ depends only on B and ρ.) It induces a bijection between real quadratic $F \subset B$ modulo conjugation by R^1, and closed geodesics $\ell \subset \Gamma\backslash\mathbb{H}$.

When F is such a real quadratic field, we write $R_F = R \cap F$. It is an order in F and one has that $\Gamma_{L_F} = \bar{\rho}(R_F^1)$, whereas $\mathrm{Stab}_\Gamma(L_F)$ equals the normalizer $\overline{\rho}(N_{R^1}(F))$. More generally, one has $\mathrm{Stab}_{B^+}(L_F) = N_{B^+}(F)$. Let $\omega \in B^\times$ be a Skolem-Noether element with respect to F. That is, conjugation by ω leaves F invariant and induces the non-trivial automorphism of F. Then $N_{B^+}(F) = F^+ \sqcup (\omega F)^+$.

Remark 1.2. Closed geodesics are naturally grouped in packets indexed by a class group, and the above bijection gives the geodesic corresponding to the identity of the group. When R is an Eichler order of square-free level, packets can be described without use of the adelic language as follows: let $F \subset B$ be a real quadratic field, and I an invertible fractional R_F-ideal. The right R-ideal IR is principal and generated by an element $a \in B^+$. To I we associate the R^1-conjugacy class of the field $a^{-1}Fa$ (i.e., a closed geodesic). The map obtained in this way factors through the narrow class group of R_F to an injective map.

Remark 1.3. When a closed geodesic is viewed as a subset of $\Gamma\backslash\mathbb{H}$, it may seem natural to define $\mathscr{P}_\ell(\phi_j)$ instead as an integral over $\mathrm{Stab}_\Gamma(L)\backslash L$, resulting in a period that is half as large when L is a reciprocal geodesic. However, the definition (1.1) is closer to the notion of an adelic period, and appears naturally in a period formula, which we now state.

Remark 1.4. Geodesic periods are related to central values of certain Rankin–Selberg L-functions as follows: [59, Theorem 5.4.1] let $N \geq 1$ be a square-free integer and let F be a real quadratic field of square-free discriminant d_F coprime to N. Assume that the number of primes dividing N that are inert in F is even. Let B be a quaternion algebra over $\mathbb{Q}$ that is ramified exactly at the primes dividing N that are inert in F. Fix an embedding $\iota : F \hookrightarrow B$. Let $R \subset B$ be an Eichler order of level N/Δ_B containing the full ring of integers $\iota(\mathcal{O}_F)$. (The freedom in the choice of the Eichler order R is explained in [59, Remark 5.3.3].) Let $\Gamma \subset \mathrm{PSL}_2(\mathbb{R})$ be the corresponding lattice. Let h_F^+ be the narrow class number of F. To F corresponds a packet Λ_{d_F} consisting of h_F^+ closed geodesics in $\Gamma\backslash\mathbb{H}$. Let χ_0 be the trivial character of the narrow class group of F and π_{χ_0} the associated automorphic representation of $\mathrm{GL}_2(\mathbb{A}_\mathbb{Q})$ given by automorphic induction. Let f be a Maass newform of level $\Gamma_0(N)$. Let π_f be the automorphic representation of $\mathrm{GL}_2(\mathbb{A}_\mathbb{Q})$ it generates. Let π_f^{JL} be the corresponding representation of $B^\times(\mathbb{A}_\mathbb{Q})$ given by the Jacquet-Langlands correspondence and take a newvector in π_f^{JL}, which corresponds

to an L^2-normalized Hecke-Laplace eigenfunction ϕ on $\Gamma\backslash\mathbb{H}$. Then

$$\frac{\Lambda(1/2, \pi_f \times \pi_{\chi_0})}{\Lambda(1, \pi_f, \mathrm{Ad})} = 2d_F^{-1/2} \prod_{p|\Delta_B} \frac{p+1}{p-1} \cdot \left| \sum_{\ell \in \Lambda_{d_F}} \mathscr{P}_\ell(\phi) \right|^2 , \tag{1.5}$$

where the Λ's in the left-hand side are completed Rankin–Selberg L-functions. (When f is odd, both sides are 0.)

Note that: (1) the assumption that there is an even number of prime divisors of N that are inert in F implies that the ϵ-factor in the functional equation for $\Lambda(s, \pi_f \times \pi_{\chi_0})$ is 1, so that there is no forced vanishing of the central value; (2) if there is *at least one* such prime divisor, B is a division algebra; (3) if *all* prime divisors of N are inert in F, R is a maximal order. In Corollary 1 we impose all three conditions.

1.1.4 Hecke operators

Assume from now on that R is an Eichler order of square-free level. Let Δ_R be its discriminant and $\Gamma = \overline{\rho}(R^1)$. For $n \in \mathbb{N}_{>0}$, the quotients $R(1)\backslash R(n)$ are finite and we define the Hecke operators T_n, which act on $L^2(\Gamma\backslash\mathbb{H})$, by

$$(T_n f)(z) = \sum_{\eta \in R(1)\backslash R(n)} f(\eta z) . \tag{1.6}$$

The T_n are self-adjoint operators and commute with the Laplacian Δ on smooth functions. Because R is an Eichler order of square-free level, when $(mn, \Delta_R) = 1$ we have the relations

$$T_m T_n = T_n T_m = \sum_{d|m,n} d \cdot T_{mn/d^2} \tag{1.7}$$

(see [23, §III.7]). Note that the Hecke operator T_n is often normalized by multiplying the sum in (1.6) by a factor $1/\sqrt{n}$, which we don't do here.

1.1.5 Convolution operators

In order to extract short spectral sums from the spectral side of the trace formula, we will make use of convolution operators that are approximate spectral projectors, and whose Harish-Chandra transform satisfies certain positivity properties. Their existence is guaranteed by Proposition 1.8.

For $s \in \mathbb{C}$, define the spherical function $\varphi_s \in C^\infty(K\backslash G/K)$ by

$$\varphi_s(g) = \int_K e^{(1/2+s)H(kg)} dk ,$$

where $H : G \to \mathbb{R}$ is defined on A as the inverse of the map $a : \mathbb{R} \to A$ from §1.1.1, and extended using the Iwasawa decomposition to $G = NAK$ by defining $H(nak) := H(a)$.

This way, $H(g)$ is the unique real number for which $g \in Na(H(g))K$. For $a \in A$ we have that

$$\varphi_s(a) = P_{-\frac{1}{2}+s}(\cosh(H(a))),$$

where P stands for the Legendre function (see also (1.33)).

For a function $k \in C_0^\infty(K\backslash G/K)$, define the Harish-Chandra transform $\widehat{k} : \mathbb{C} \to \mathbb{C}$ by

$$\widehat{k}(s) = \int_G k(g)\varphi_s(g)dg\,.$$

It is an entire function of order (at most) 1 and of finite type, whose type remains bounded when the support of k remains bounded. When $k_1, k_2 \in C_0^\infty(K\backslash G/K)$, define the convolution

$$(k_1 * k_2)(g) = \int_G k_1(gh)k_2(h^{-1})dh\,.$$

Proposition 1.8. *There exists a family $(k_\nu)_{\nu\geq 0}$ of bi-K-invariant smooth functions on G satisfying:*

1. *there exists $R > 0$ such that k_ν is supported in the ball $B(1, R)$ for all ν;*
2. *$\widehat{k}_\nu(s) \in \mathbb{R}_{\geq 0}$ when $s \in \mathbb{R} \cup i\mathbb{R}$;*
3. *$\widehat{k}_\nu(ir) \geq 1$ for $|r - \nu| \leq 1$;*
4. *$\widehat{k}_\nu(ir) \ll_N (1 + |\nu - r|)^{-N}$ uniformly for $\nu \geq 0$ and $r \in \mathbb{R}_{\geq 0} \cup [-i/2, i/2]$;*
5. *$k_\nu(1) \asymp \nu$ for ν sufficiently large.*

Proof. Take any real-valued $k \in C_0^\infty(K\backslash G/K)$. The Harish-Chandra transform $\widehat{k}(s)$ is clearly real when $s \in \mathbb{R}$, and in view of the functional equation $\varphi_{-s} = \varphi_s$, also when $s \in i\mathbb{R}$. If we choose k nonnegative and not identically zero, we have $\widehat{k}(s) > 0$ for $s \in \mathbb{R}$; in particular for $s = 0$. Let $k_{(1)} = k * k$. Then $\widehat{k}_{(1)}(s) = \widehat{k}(s)^2 \geq 0$ for $s \in \mathbb{R} \cup i\mathbb{R}$. There exists $\delta > 0$ such that $\widehat{k}_{(1)}(s) > \widehat{k}_{(1)}(0)/2$ for $|s| \leq \delta$. Let $\widehat{k}_{(2)}(s) = 2\widehat{k}_{(1)}(\delta s)/\widehat{k}_{(1)}(0)$. Then $\widehat{k}_{(2)}(s) \geq 1$ for $|s| \leq 1$. Let $k_{(2)}$ be its inverse Harish-Chandra transform. Define

$$\widehat{k}_\nu(s) = \widehat{k}_{(2)}(i\nu + s) + \widehat{k}_{(2)}(i\nu - s)\,.$$

Then $\widehat{k}_\nu$ is entire, even, of exponential type (at most) the exponential type of $\widehat{k}_{(2)}$, so it is the Harish-Chandra transform of a $k_\nu \in C_0^\infty(K\backslash G/K)$ whose support is bounded independently of ν. Conditions (1) and (2) are now satisfied. When $|r - \nu| \leq 1$, we have

$$\begin{aligned}\widehat{k}_\nu(ir) &= \widehat{k}_{(2)}(i\nu + ir) + \widehat{k}_{(2)}(i\nu - ir)\\ &\geq 0 + 1\,,\end{aligned}$$

which implies condition (3). Being the Fourier transform of the Abel transform of $k_{(2)}$, $\widehat{k}_{(2)}$ is of rapid decay on vertical strips of $\mathbb{C}$. Thus when $\mathrm{Re}(r) \geq 0$ and $\mathrm{Im}(r)$ remains bounded,

$$\begin{aligned}\widehat{k}_\nu(ir) &\ll_N (1+|\nu+r|)^{-N} + (1+|\nu-r|)^{-N} \\ &\ll_N (1+|\nu-r|)^{-N}\,,\end{aligned}$$

which is (4). For (5), using the inverse Harish-Chandra transform we have

$$k_\nu(g) = \int_0^\infty \widehat{k}_\nu(ir)\varphi_{ir}(g)\beta(r)dr\,, \tag{1.9}$$

where $\beta(r) = \frac{1}{2\pi}\tanh(\pi r)\cdot r$ is the Plancherel density. When $g = e$, $\varphi_{ir}(g) = 1$ so that

$$k_\nu(1) \geq \int_\nu^{\nu+1} \widehat{k}_\nu(ir)\beta(r)dr \gg \nu$$

on the one hand, and

$$\begin{aligned}k_\nu(1) &\ll \int_0^{2\nu} \widehat{k}_\nu(ir)\nu dr + \int_{2\nu}^\infty \frac{r}{(r-\nu)^3}dr \\ &\ll \nu\cdot\|\widehat{k}_\nu\|_{L^1(i\mathbb{R})} + \nu^{-1} \ll \nu\end{aligned}$$

on the other. □

Throughout, we fix a family (k_ν) as in Proposition 1.8.

1.1.6 Eigenfunctions

Let $(\phi_j)_{j\geq 0}$ be an orthonormal basis of $L^2(\Gamma\backslash\mathbb{H})$ consisting of simultaneous eigenfunctions for Δ and the Hecke operators T_n for $(n, \Delta_R) = 1$, ordered by increasing Laplacian eigenvalue $\lambda_j \geq 0$:

$$(\Delta + \lambda_j)\phi_j = 0\,.$$

Because Δ and the T_n are self-adjoint, we can and will assume that each ϕ_j is real-valued. Write the nth Hecke eigenvalue of ϕ_j as $\widehat{T}_n(\phi_j)$. For each j, write $\lambda_j = \frac{1}{4} + \nu_j^2$, where $\nu_j \in \mathbb{R}_{\geq 0}\cup[0, i/2]$ is called the spectral parameter of ϕ_j, to be thought of as the frequency of the wavefunction ϕ_j. The automorphic kernel $K_\nu(x,y) = \sum_{\gamma\in\Gamma} k_\nu(x^{-1}\gamma y)$ acts as an integral operator on $L^2(\Gamma\backslash\mathbb{H})$, and the K_ν-eigenvalue of ϕ_j is $\widehat{k}_\nu(i\nu_j)$.

1.2 Two trace formulas

As in §1.1, let $\Gamma \subset G$ be the lattice coming from an Eichler order R of square-free level in a quaternion division algebra B, $F \subset B$ a real quadratic field, $L = L_F \subset \mathbb{H}$ the corresponding geodesic and ℓ the corresponding closed geodesic. Denote $\mathscr{P}(\phi_j) = \mathscr{P}_\ell(\phi_j)$. Let $g_0 A g_0^{-1} = A_L$ be the identity component of $\mathrm{Stab}_G(L)$.

We start from the spectral expansion of the automorphization of k_ν:

$$\sum_j \widehat{k}_\nu(i\nu_j)\phi_j(x)\phi_j(y) = \sum_{\gamma\in\Gamma} k_\nu(x^{-1}\gamma y)\,,$$

where the convergence is uniform for x and y in compact sets. Let $n \geq 1$ be an integer coprime to the discriminant Δ_R, and apply T_n to the x-variable to obtain

$$\sum_j \widehat{k}_\nu(i\nu_j)\widehat{T}_n(\phi_j)\phi_j(x)\phi_j(y) = \sum_{\eta\in R(n)/\pm 1} k_\nu(x^{-1}\eta y)\,. \tag{1.10}$$

Setting $x = y$ and integrating over $\Gamma\backslash\mathbb{H}$ we obtain the amplified standard trace formula:

$$\sum_j \widehat{k}_\nu(i\nu_j)\widehat{T}_n(\phi_j) = \int_{\Gamma\backslash\mathbb{H}} \sum_{\eta\in R(n)/\pm 1} k_\nu(x^{-1}\eta x)dx\,. \tag{1.11}$$

On the other hand, to make periods appear, we may integrate (1.10) over $\ell \times \ell$ to get the amplified relative trace formula:

$$\sum_j \widehat{k}_\nu(i\nu_j)\widehat{T}_n(\phi_j)\mathscr{P}(\phi_j)^2 = \int_{\ell\times\ell} \sum_{\eta\in R(n)/\pm 1} k_\nu(x^{-1}\eta y)dxdy\,. \tag{1.12}$$

We now identify a main term and an error term in both trace formulas.

1.2.1 Standard trace formula

Rewrite (1.11) as

$$\begin{aligned}\sum_j \widehat{k}_\nu(i\nu_j)\widehat{T}_n(\phi_j) &= |(R(n)\cap\mathbb{Q})/\pm 1|\cdot \mathrm{Vol}(\Gamma\backslash\mathbb{H})\cdot k_\nu(1)\\ &\quad + \sum_{\eta\in(R(n)-\mathbb{Q})/\pm 1}\int_{\mathscr{F}} k_\nu(x^{-1}\eta x)dx\,,\end{aligned} \tag{1.13}$$

where $\mathscr{F} \subset \mathbb{H}$ is a fundamental domain for the action of Γ. We want to bound the sum in the right-hand side. We follow [37, §2]. The argument changes slightly because our convolution operator k_ν is a different one: (1) the support of k_ν does not decrease with ν, which means that the contribution of hyperbolic elements cannot be ignored. (2) k_ν is not necessarily nonnegative, so we cannot bound the contribution of a single η by the contribution of its conjugacy class.

Remark 1.14. The reason why we (have to) make this choice of convolution operator k_ν, is the following. Our ultimate goal is to prove an inequality of the form

$$\sum_j \widehat{k}_\nu(i\nu_j) w_j \mathscr{P}(\phi_j)^2 \geq C_\nu \cdot \sum_j \widehat{k}_\nu(i\nu_j) w_j \tag{1.15}$$

for suitable nonnegative weights w_j, which will imply (if the right-hand side is nonzero) that at least one eigenfunction ϕ_{j_0} satisfies the bound $\mathscr{P}(\phi_{j_0})^2 \geq C_\nu$. Here, we expect C_ν to be just slightly larger than ν^{-1}. In particular, it decreases with ν (in contrast to the situation in [37, 50]). We want to bound C_ν from below by $\nu_{j_0}^{-1}$. If ν_{j_0} is small compared to ν (less than $\nu^{0.99}$, say), then this is not possible. To ensure that $\nu_{j_0} \gg \nu$, we want to truncate the sum in the left-hand side of (1.15) and keep only the terms with $\nu_{j_0} \asymp \nu$. This is why we have chosen k_ν with the property that it has rapid decay away from ν. As a bonus, we obtain an eigenfunction whose spectral parameter ν_{j_0} lies in an interval of bounded length around ν. As a minus, we lose nonnegativity of k_ν, which means that orbital integrals in the relative trace formula can be negative, and have to be bounded.

For the contributing hyperbolic classes, we will need additional arithmetic information, given by the following lemma.

Lemma 1.16. *Let $n \geq 1$ and $\eta \in R(n)$. Suppose that there exists $\gamma \in R^1$, $x \in \mathscr{F}$ and $\nu \geq 0$ with $x^{-1}\gamma^{-1}\eta\gamma x \in \operatorname{supp} k_\nu$. Let $K = \mathbb{Q}(\eta) \subset B$ be the quadratic subfield generated by η, and let $\mathcal{O} = R \cap K$, which is an order in K. Then there exists a constant $C > 0$, independent of n and η, such that $|\operatorname{Tr} \bar{\rho}(\eta)| \leq C$ and the discriminant $D_\mathcal{O}$ of $\mathcal{O}$ satisfies $|D_\mathcal{O}| \leq Cn\,||\operatorname{Tr} \bar{\rho}(\eta)| - 2|$.*

Proof. Because $K \cap R \cong \gamma K \gamma^{-1} \cap R$, these orders have the same discriminant, and because $\operatorname{Tr}(\gamma^{-1}\eta\gamma) = \operatorname{Tr}(\eta)$ we may assume $\gamma = 1$. Then $\bar{\rho}(\eta)$ belongs to the bounded set $\mathscr{F}(\bigcup_\nu \operatorname{supp} k_\nu)\mathscr{F}^{-1}$. Thus $\frac{1}{\sqrt{n}}\rho(\eta)$ belongs to a bounded set independent of n and ν. By the remarks in §1.1.1 there exists $C' > 0$ with $\|\frac{1}{\sqrt{n}}\eta\| \leq C'$, where $\|\cdot\|$ is the norm on $B \otimes_\mathbb{Q} \mathbb{R}$ we fixed in §1.1.2. Let $\alpha \in \mathcal{O}$ be such that $\alpha^2 = D_\mathcal{O}$. Then $\mathcal{O} \subset \frac{1}{2}\mathbb{Z} + \frac{1}{2}\alpha\mathbb{Z}$. Write $\eta = a + b\alpha$ with $a, b \in \frac{1}{2}\mathbb{Z}$. Then

$$|\operatorname{Tr}(\eta)| = \|2a\| = \|\eta + \bar{\eta}\| \leq 2C'\sqrt{n}$$

so that $|\operatorname{Tr}(\bar{\rho}(\eta))| \leq 2C'$. Because $\eta \notin \mathbb{Q}$ we have $b \neq 0$, so $|b| \geq 1/2$. From $n = N(\eta) = a^2 - D_\mathcal{O} b^2$ we have

$$|D_\mathcal{O}| = \frac{|a^2 - n|}{b^2} \leq 4|a^2 - n| = n(|\operatorname{Tr} \bar{\rho}(\eta)| + 2)\,||\operatorname{Tr} \bar{\rho}(\eta)| - 2|$$

and we can take $C = 2C' + 2$. □

Let s denote the characteristic function of the squares, so that $|(R(n) \cap \mathbb{Q})/\pm 1| = s(n)$.

Proposition 1.17. *There exists an absolute constant $C > 0$ such that*

$$\sum_j \widehat{k}_\nu(i\nu_j)\widehat{T}_n(\phi_j) = s(n)\operatorname{Vol}(\Gamma\backslash\mathbb{H})\cdot k_\nu(1) + O\left(n^{3/2}e^{C\log n/\log\log(n+1)}\right),$$

uniformly in $n \geq 1, \nu \geq 0$.

Proof. We may rewrite the sum in the right-hand side of (1.13) as

$$\frac{1}{2}\sum_{\mathcal{C}}\sum_{\eta\in\mathcal{C}}\int_{\mathscr{F}} k_\nu(x^{-1}\eta x)dx\,, \tag{1.18}$$

where the sum runs over all R^1-conjugacy classes $\mathcal{C} \subset R(n) - \mathbb{Q}$. Fix $\mathcal{C}$ and take $\eta \in \mathcal{C}$. We break into cases based on whether $\overline{\rho}(\eta)$ is parabolic, hyperbolic or elliptic. Because B is a division algebra, $\overline{\rho}(\eta)$ is not parabolic.

When $\overline{\rho}(\eta)$ is hyperbolic, call $P > 1$ the square of its largest eigenvalue. Because η generates a real quadratic quadratic subfield $K \subset B$, the centralizer $Z_\Gamma(\overline{\rho}(\eta))$ is an infinite cyclic group. Let $\eta_0 \in K \cap R^1$ be totally positive and such that $\overline{\rho}(\eta_0)$ generates this centralizer, and let $P_0 > 1$ the square of the largest eigenvalue of $\overline{\rho}(\eta_0)$. By the computation in [36, §10.5] and our choice of k_ν,

$$\begin{aligned}\sum_{\eta\in\mathcal{C}}\int_{\mathscr{F}} k_\nu(x^{-1}\eta x)dx &= \frac{\log P_0}{P^{1/2}-P^{-1/2}}\int_{\mathbb{R}}\widehat{k}_\nu(ir)e^{ir\log P}(2\pi)^{-1}dr\\ &\ll \frac{\log P_0}{P^{1/2}-1}\,.\end{aligned}$$

We have

$$(P^{1/2}-1)^2 \geq \frac{(P^{1/2}-1)^2}{P^{1/2}} = P^{1/2}+P^{-1/2}-2 = |\operatorname{Tr}(\overline{\rho}(\eta))|-2\,.$$

In order to bound P_0, let $\sigma : K \to \mathbb{R}$ be an embedding and observe that $P_0^{1/2}+P_0^{-1/2} = \operatorname{Tr}_{B/\mathbb{Q}}(\eta_0) = \sigma(\eta_0)+\sigma(\eta_0)^{-1}$. Thus $P_0 \in \{\sigma(\eta_0)^2, \sigma(\eta_0)^{-2}\}$, so that $\log P_0 = 2\,|\log\sigma(\eta_0)|$. When $R_\mathcal{O}$ is the regulator of the order $\mathcal{O} = R \cap K$, we have $|\log\sigma(\eta_0)| \in \{R_\mathcal{O}, 2R_\mathcal{O}\}$ according to whether the fundamental unit of $\mathcal{O}$ has norm ± 1. From the proof of Dirichlet's unit theorem, one has that $R_\mathcal{O}$ is bounded up to a constant by a power of the discriminant $D_\mathcal{O}$. Using the class number formula this can be improved to $D_\mathcal{O}^{1/2}\log D_\mathcal{O}\log\log D_\mathcal{O}$. (For non-maximal orders, this was done in [64].) We may now assume that the sum over $\eta \in \mathcal{C}$ is nonzero, in which case $D_\mathcal{O} \ll n(|\operatorname{Tr}\bar{\rho}(\eta)|-2) \ll n$ by Lemma 1.16. Combining this with the bound for $P^{1/2}-1$, we obtain that for $\mathcal{C}$ hyperbolic,

$$\sum_{\eta\in\mathcal{C}}\int_{\mathscr{F}} k_\nu(x^{-1}\eta x)dx \ll \frac{n^{1/2}(|\operatorname{Tr}(\overline{\rho}(\eta))|-2)^{1/2}\log^2 n}{(|\operatorname{Tr}(\overline{\rho}(\eta))|-2)^{1/2}} = n^{1/2}\log^2 n\,. \tag{1.19}$$

When $\overline{\rho}(\eta)$ is elliptic, let $0 < \theta \leq \pi/2$ be such that $\pm e^{i\theta}$ is an eigenvalue. The centralizer $Z_\Gamma(\overline{\rho}(\eta))$ is a finite cyclic group. By the computation in [36, §10.6],

$$\sum_{\eta \in \mathcal{C}} \int_{\mathscr{F}} k_\nu(x^{-1}\eta x)dx = \frac{|Z_\Gamma(\overline{\rho}(\eta))|^{-1}}{\sin\theta} \int_{\mathbb{R}} \widehat{k}_\nu(ir) \frac{\cosh((\pi - 2\theta)r)}{\cosh(\pi r)} \frac{\pi}{2} dr$$
$$\ll \frac{1}{\sin\theta}.$$

Because $\bar{\rho}(\eta)$ is elliptic, $|\mathrm{Tr}(\eta)|^2 < 4n$, so $|\mathrm{Tr}(\eta)|^2 \leq 4n - 1$ as the left-hand side is an integer. We have

$$2 - |\mathrm{Tr}(\overline{\rho}(\eta))| \geq 2 - \sqrt{4 - \frac{1}{n}} = \frac{1}{n\left(2 + \sqrt{4 - \frac{1}{n}}\right)} \geq \frac{1}{4n}$$

so that

$$2\sin^2\theta = 2 - 2\cos^2\theta \geq 2 - 2\cos\theta = 2 - |\mathrm{Tr}(\bar{\rho}(\eta))| \geq (4n)^{-1}.$$

We obtain that, for $\mathcal{C}$ elliptic,

$$\sum_{\eta \in \mathcal{C}} \int_{\mathscr{F}} k_\nu(x^{-1}\eta x)dx \ll n^{1/2}. \tag{1.20}$$

We now count the number of contributing conjugacy classes. For the term in (1.18) corresponding to η to be nonzero, there must exist x in the compact set $\mathscr{F}$ with $x^{-1}\overline{\rho}(\eta)x \in \operatorname{supp} k_\nu$. We may thus restrict the sum (1.18) to $\mathcal{C}$ that contain an element η with $\|\eta\| \leq C'\sqrt{n}$. The number of such conjugacy classes is bounded by the cardinality of the set

$$M = \{\eta \in R(n) : \|\eta\| \leq C'\sqrt{n}\}.$$

This is a counting problem that has arisen many times in the context of sup norms of Maass cusp forms. Letting $(\eta_0, \eta_1, \eta_2, \eta_3)$ be a fixed $\mathbb{Z}$-basis of R and using that $\|\cdot\|$ is equivalent to the sup norm with respect to this basis, we see that $|M| \ll (\sqrt{n})^4$. A more careful treatment gives $|M| \ll n\exp(C\log n/\log\log(n+1))$ for some $C > 0$; see for example [50, §5] or Remark 1.31 below. Multiplying this bound for $|M|$ by the larger one of the bounds (1.19) and (1.20), the claim follows. □

1.2.2 Relative trace formula

Proposition 1.21. *There exists a nonnegative $b \in C_0^\infty(\mathbb{R})$ depending only on Γ and L, such that for every $n \in \mathbb{N}_{>0}$,*

$$\sum_j \widehat{k}_\nu(i\nu_j)\widehat{T}_n(\phi_j)\mathscr{P}(\phi_j)^2 = |N_{R(n)}(F)/R_F^1| \cdot \mathrm{Vol}(\Gamma_L\backslash L) \cdot \int_{\mathbb{R}} k_\nu(a(t))dt$$
$$+ \sum_{\eta \in (R(n) - N_{R(n)}(F))/\pm 1} I(\nu, g_0^{-1}\eta g_0),$$

where for $g \in G$ we define

$$I(\nu, g) = \int_{\mathbb{R}\times\mathbb{R}} b(s)b(t)k_\nu(a(-s)ga(t))dsdt\,. \tag{1.22}$$

Proof. We start from the relative trace formula (1.12). We have that L has unit length parametrization by $g_0a(t)i$ where $a(t) = \exp\left(\begin{smallmatrix} t/2 & 0 \\ 0 & -t/2 \end{smallmatrix}\right)$. Let $\mathscr{F}_L \subset L$ be a fixed fundamental domain for the action of Γ_L. Write

$$\sum_{\eta\in R(n)/\pm 1}\int_{\mathscr{F}_L\times\mathscr{F}_L} k_\nu(x^{-1}\eta y)dxdy = \sum_{\eta\in N_{R(n)}(F)/\pm 1}\int_{\mathscr{F}_L\times\mathscr{F}_L} k_\nu(x^{-1}\eta y)dxdy + \sum_{\eta\in (R(n)-N_{R(n)}(F))/\pm 1}\int_{\mathscr{F}_L\times\mathscr{F}_L} k_\nu(x^{-1}\eta y)dxdy\,.$$

The first term equals, by unfolding the sum in the second variable and then making the change of variables $y \leftarrow \eta^{-1}y$,

$$\begin{aligned}
&\sum_{\eta\in N_{R(n)}(F)/R_F^1}\int_{\mathscr{F}_L\times L} k_\nu(x^{-1}\eta y)dxdy \\
&\qquad = \sum_{\eta\in N_{R(n)}(F)/R_F^1}\int_{\mathscr{F}_L\times L} k_\nu(x^{-1}y)dxdy \\
&\qquad = |N_{R(n)}(F)/R_F^1|\int_{[0,\mathrm{Vol}(\mathscr{F}_L)]\times\mathbb{R}} k_\nu(a(-s+t))dsdt \\
&\qquad = |N_{R(n)}(F)/R_F^1|\cdot\mathrm{Vol}(\Gamma_L\backslash L)\cdot\int_{\mathbb{R}} k_\nu(a(t))dt
\end{aligned}$$

For the second term, we unfold the sum as follows:

$$\begin{aligned}
&\sum_{\delta\in\Gamma_L\backslash\overline{\rho}(R(n)-N_{R(n)}(F))/\Gamma_L}\sum_{\gamma\in\Gamma_L\delta\Gamma_L}\int_{\mathscr{F}_L\times\mathscr{F}_L} k_\nu(x^{-1}\gamma y)dxdy \\
&= \sum_{\delta\in\Gamma_L\backslash\overline{\rho}(R(n)-N_{R(n)}(F))/\Gamma_L}\int_{L\times L} k_\nu(x^{-1}\delta y)dxdy
\end{aligned}$$

Here, we used that for δ as above, the integral over $L\times L$ converges absolutely thanks to the compact support of k_ν, and that the map $\Gamma_L\times\Gamma_L \to G : (\gamma_1,\gamma_2) \mapsto \gamma_1\delta\gamma_2$ is injective. Indeed, the contrary would imply that $\Gamma_L \cap \delta\Gamma_L\delta^{-1} \neq \{1\}$, so in particular $A_L \cap \delta A_L\delta^{-1} \neq \{1\}$. This would imply $A_L = \delta A_L\delta^{-1}$ so that $\delta \in N_G(A_L)$. If $\delta = \overline{\rho}(\eta)$ with $\eta \in R(n)$, this means that $\eta \in N_{R^+}(F)$, a contradiction.

Now we introduce a smooth cutoff function in each of the integrals. Let $b_0 \in C_0^\infty(L)$ be such that $\sum_{\gamma \in \Gamma_L} b_0(\gamma z) = 1$ for all $z \in L$. Then the sum equals

$$\sum_{\delta \in \Gamma_L \backslash \overline{\rho}(R(n) - N_{R(n)}(F))/\Gamma_L} \int_{L \times L} \sum_{\gamma_1, \gamma_2 \in \Gamma_L} b_0(\gamma_1 x) b_0(\gamma_2 y) k_\nu(x^{-1} \delta y) dx dy$$

$$\sum_{\delta \in \Gamma_L \backslash \overline{\rho}(R(n) - N_{R(n)}(F))/\Gamma_L} \int_{L \times L} \sum_{\gamma_1, \gamma_2 \in \Gamma_L} b_0(x) b_0(y) k_\nu(x^{-1} \gamma_1^{-1} \delta \gamma_2 y) dx dy$$

$$= \sum_{\gamma \in (R(n) - N_{R(n)}(F))/\pm 1} \int_{L \times L} b_0(x) b_0(y) k_\nu(x^{-1} \gamma y) dx dy \,,$$

where we made the change of variables $(x, y) \leftarrow (\gamma_1 x, \gamma_2 y)$ and merged the sum over δ with the sum over γ_1, γ_2. We obtain the statement of the proposition, with $b(t) := b_0(g_0 a(t))$. □

For the integral appearing in the main term in Proposition 1.21, we shall prove the following in §1.4.

Proposition 1.23. *We have*

$$\int_{\mathbb{R}} k_\nu(a(t)) dt \asymp 1$$

for ν sufficiently large.

For the integrals appearing in the remaining terms in Proposition 1.21, we have the following.

Proposition 1.24. *Define $I(\nu, g)$ by* (1.22).

1. *There exists $C' > 0$ independent of $\nu \geq 0$ such that $I(\nu, g) = 0$ unless $d(g, 1) \leq C'$.*
2. *For $g \in G$, we have*

$$|I(\nu, g)| \ll (1 + \nu \cdot d(g, N_G(A)))^{-1/2}$$

uniformly in g and $\nu \geq 0$.

The constant C' in (1) and the implicit constant in (2) may depend on ℓ.

Proof. We prove the first assertion and postpone the proof of the second assertion until §1.5. By construction, k_ν is supported on points at bounded distance from e. If $I(\nu, g)$ is nonzero, there exist $t, s \in \operatorname{supp} b$ with $k_\nu(a(-t) g a(s)) \neq 0$. Then g belongs to the set $a(\operatorname{supp} b)(\operatorname{supp} k_\nu) a(-\operatorname{supp} b)$, which is bounded uniformly in ν. □

1.3 Counting Hecke returns

1.3.1 Counting stabilizers

Let $F \subset B$ as before be the real quadratic number field corresponding to the geodesic L. We want to understand the factor $|N_{R(n)}(F)/R_F^1|$, which appears in the main term in the relative trace formula. Denote by $\mathfrak{f}_{R_F}$ the conductor of the order R_F, by P_{R_F} the set of principal ideals of the maximal order $\mathcal{O}_F$ that are generated by an element of R_F, and by $P_{R_F}(n)$ the set of such ideals of norm n.

Lemma 1.25. *When $n \geq 1$ is coprime to $\mathfrak{f}_{R_F}$, we have*

$$|N_{R(n)}(F)/R_F^1| \geq |P_{R_F}(n)|$$

Proof. We may bound the left-hand side from below by $|R_F(n)/R_F^1|$. We show that for $(n, \mathfrak{f}_{R_F}) = 1$, this cardinality equals $|P_{R_F}(n)|$. An element $\eta \in R_F(n)$ determines a principal ideal $\eta R_F \subset R_F$, which determines a principal ideal $\eta\mathcal{O}_F \in P_{R_F}(n)$. The composition of the two maps obtained in this way, is by definition surjective onto $P_{R_F}(n)$. The first map has fibers which are full orbits under multiplication by R_F^1. The second is injective, provided that n is coprime to the conductor $\mathfrak{f}_{R_F}$. □

1.3.2 Bounding approximate stabilizers

In order to control the sum appearing in the error term in Proposition 1.21, we will need an upper bound for the number of elements $\eta \in R(n)$ that are close to stabilizing L without actually stabilizing it, that is, for the cardinality of the sets

$$M(n, \delta) = \left\{\eta \in R(n) : d(\bar{\rho}(\eta), 1) \leq C',\, 0 < d(\bar{\rho}(\eta), N_G(A_L)) \leq \delta\right\},$$

where $C' > 0$ is an arbitrary constant, which we fix throughout this section. This counting problem is similar to, but slightly different from the ones considered in [37, 47] in the context of upper bounds for sup norms and for integrals along geodesic segments, respectively: in the definition of $M(n, \delta)$, we are excluding the $\eta \in R(n)$ that stabilize L, so that the upper bound for $|M(n, \delta)|$ we obtain is smaller than what a direct invocation of [47, Lemma 3.3] would imply. This is necessary, because including those η in the definition of $M(n, \delta)$ would force any upper bound to be at least as large as $|N_{R(n)}(F)/R_F^1|$, while we want the latter quantity to dominate the error term. Our method for bounding $|M(n, \delta)|$ however, uses many ideas that originate in [37].

Lemma 1.26. *There exists an absolute constant $C > 0$ such that when $D \in \mathbb{Z}_{>0}$ is a non-square, $0 < \delta \leq B$ and $n \geq 1$,*

$$\begin{aligned} &\#\left\{(u, v) \in \mathbb{Z}^2 : 0 < \left|u^2 - Dv^2 - n\right| \leq \delta n,\, |u|, |v| \leq Bn\right\} \\ &\quad \ll \delta n e^{C \log n / \log\log(n+1)} \end{aligned}$$

where the implicit constant depends on D and B.

Proof. We want to estimate

$$\begin{aligned}&\#\{(u,v)\in\mathbb{Z}^2 : 0<|u^2-Dv^2-n|\le \delta n,\ |u|,|v|\le Bn\}\\ &=\sum_{0<|m-n|\le\delta n}\#\{u,v : u^2-Dv^2=m,\ |u|,|v|\le Bn\}\end{aligned}$$

Let $K=\mathbb{Q}(\sqrt{D})$. Fix m as in the sum. Any pair (u,v) as above determines an element $z=u+v\sqrt{D}\in K$ of norm m, and if $|\cdot|_1$, $|\cdot|_2$ denote the two Archimedian absolute values of K, we have $|z|_1,|z|_2\le Bn(1+\sqrt{D}))$. We obtain the upper bound

$$\begin{aligned}&\#\{(u,v)\in\mathbb{Z}^2 : u^2-Dv^2=m,\ |u|,|v|\le Bn\}\\ &\le\#\{z\in\mathcal{O}_K : N(z)=m, |z|_1,|z|_2\le Bn(1+\sqrt{D})\}\,.\end{aligned}$$

The latter set maps to the set of integral ideals in $\mathcal{O}_K$ with norm m. There are at most $\tau(m)$ such ideals. We may bound the divisor function $\tau(m)$ by its maximal order $\exp(C\log m/\log\log(m+1))\ll_\delta \exp(C\log n/\log\log(n+1))$ for some $C>0$. The fibers of said map are orbits under multiplication by units of norm 1. Let ε be the fundamental unit of K. The bound on $|z|_1$ and $|z|_2$ implies that the fibers of that map are of size at most

$$\ll\frac{\log(Bn(1+\sqrt{D}))}{\log|\varepsilon|}\ll 1+\log n$$

We obtain the upper bound

$$\begin{aligned}&\sum_{0<|m-n|\le\delta n}\#\{u,v : u^2-Dv^2=m,\ |u|,|v|\le Bn\}\\ &\ll\delta n\cdot\exp(C\log n/\log\log(n+1))\cdot(1+\log n)\,,\end{aligned}$$

as desired. Here we used that the number of terms in the sum is $\ll\delta n$, and not just $\ll\delta n+1$, since we omit the term for $m=n$. □

Lemma 1.27. *There exists an absolute constant $C>0$ such that*

$$\#M(n,\delta)\ll\delta n e^{C\log n/\log\log(n+1)}\,.$$

The implicit constant may depend on ℓ.

Proof. Let $D>0$ be the discriminant of F and take $\alpha\in F$ with $\alpha^2=D$. For $t,u\in\mathbb{R}$ we have $N_{M_2(\mathbb{R})/\mathbb{R}}(t+u\rho(\alpha))=t^2-Du^2$, so that

$$A_L=\bar{\rho}((F\otimes_{\mathbb{Q}}\mathbb{R})^1)=\left\{\overline{t+u\rho(\alpha)} : t^2-Du^2=1\right\}\tag{1.28}$$

Let $\omega \in B^\times$ be the Skolem-Noether element from §1.1.3. Multiplying ω, if necessary, by an element of negative norm in F and by a suitable integer, we may assume that $\omega \in R^+$. Let $E := N_{B/\mathbb{Q}}(\omega) = -\omega^2 \in \mathbb{Z}_{>0}$. Let $S \subset B$ be the order $\mathbb{Z} + \mathbb{Z}\alpha + \mathbb{Z}\omega + \mathbb{Z}\alpha\omega$. Let $f \in \mathbb{N}_{>0}$ be such that $f \cdot R \subset S$. Now let $\eta \in M(n, \delta)$. We can write $\eta = x_0 + x_1\alpha + x_2\omega + x_3\alpha\omega$ with $x_0, x_1, x_2, x_3 \in \frac{1}{f}\mathbb{Z}$, and we have $N_{B/\mathbb{Q}}(\eta) = x_0^2 - Dx_1^2 + Ex_2^2 - DEx_3^2 = n$. All implied constants throughout the proof will be allowed to depend on D, E and f.

Suppose first that $d(\bar{\rho}(\eta), A_L) \leq \delta$. Let C' be the constant from the beginning of the section, with respect to which $M(n, \delta)$ is defined. Because $M(n, \delta) = M(n, C')$ for $\delta \geq C'$, it is no restriction to assume $\delta \leq C'$. By the remarks in §1.1.1, from $d(\bar{\rho}(\eta), A_L) \leq \delta$ and $d(\bar{\rho}(\eta), 1) \leq C'$ it follows that there exists $a \in \mathrm{SL}_2(\mathbb{R})$ with $\overline{a} \in A_L$ and $\|\frac{1}{\sqrt{n}}\rho(\eta) - a\| \ll \delta$. Here, the norm is any fixed norm on $M_2(\mathbb{R})$. Writing $a = t + u\rho(\alpha)$ as in (1.28), it follows that in $B \otimes_{\mathbb{Q}} \mathbb{R}$,

$$\frac{1}{\sqrt{n}}(\eta \otimes 1) = 1 \otimes t + \alpha \otimes u + O(\delta)$$

and comparing coordinates in the basis $(1, \alpha, \omega, \alpha\omega)$, we obtain

$$\begin{aligned} \frac{x_0}{\sqrt{n}} &= t + O(\delta) \\ \frac{x_1}{\sqrt{n}} &= u + O(\delta) \\ \frac{x_2}{\sqrt{n}} &= O(\delta) \\ \frac{x_3}{\sqrt{n}} &= O(\delta)\,. \end{aligned} \tag{1.29}$$

The assumption $d(\bar{\rho}(\eta), 1) \leq C'$ implies that $x_0, x_1, x_2, x_3 \ll \sqrt{n}$, so that

$$1 = t^2 - Du^2 = \frac{x_0^2}{n} - D\frac{x_1^2}{n} + O(\delta)\,.$$

That is, $|x_0^2 - Dx_1^2 - n| \ll \delta n$. Because $d(\bar{\rho}(\eta), N_G(A)) > 0$, we have $\eta \notin F$, so that $(x_2, x_3) \neq (0, 0)$, so that $x_0^2 - Dx_1^2 = n - E(x_2^2 - Dx_3^2) \neq n$. We obtain that the integers (fx_0, fx_1) satisfy

$$0 < |(fx_0)^2 - D(fx_1)^2 - f^2 n| \ll \delta n$$

and $|fx_0|, |fx_1| \ll n$. By Lemma 1.26 and the assumption that $\delta \leq C'$, the number of possible values for (x_0, x_1) is at most $\ll \delta n e^{C \log n / \log\log(n+1)}$ for some $C > 0$. If we fix x_0 and x_1, then x_2 and x_3 satisfy

$$|(fx_2)^2 - D(fx_3)^2| = f^2/E \cdot |n - N_{B/\mathbb{Q}}(x_0 + x_1\alpha)| \ll n \tag{1.30}$$

and $|fx_2|, |fx_3| \ll n$. By counting ideals in $\mathbb{Q}(\sqrt{D})$ in the same way as in the proof of Lemma 1.26, we see that the number of (x_2, x_3) satisfying the equality in (1.30) can be

bounded by $\ll \exp(C \log n / \log\log(n+1))$. This proves that the number of $\eta \in M(n,\delta)$ with $d(\bar{\rho}(\eta), A_L) \leq \delta$ is at most $\ll \delta n e^{2C \log n/\log\log(n+1)}$.

Suppose now that $\eta \in M(n,\delta)$ is such that $d(\overline{\rho}(\eta), N_G(A_L) - A_L) \leq \delta$. We have $\overline{\rho}(\omega) \in N_G(A_L) - A_L$, so we obtain that $d(\overline{\rho}(\omega\eta), A_L) \ll \delta$, $d(\bar{\rho}(\omega\eta), 1) \ll 1$ and $\omega\eta \in R(n \cdot E)$. By the first case (with a different constant C' in the definition of $M(nE,\delta)$), the number of such η is bounded by $\ll \delta n E e^{C \log(nE)/\log\log(nE+1)}$ for some $C > 0$.

Adding the contributions from the two types of η, the claim follows. □

Remark 1.31. 1. It is clear from the proofs of Lemmas 1.26 and 1.27 that when $M'(n,\delta)$ is defined as the set obtained by removing in the definition of $M(n,\delta)$ the condition that $0 < d(\bar{\rho}(\eta), N_G(A))$, then we obtain the upper bound

$$\#M'(n,\delta) \ll (\delta n + 1) e^{C \log n/\log\log(n+1)}$$

for some $C > 0$. In particular,

$$\#\left\{\eta \in R(n) : d(\bar{\rho}(\eta), 1) \leq C'\right\} = M'(n, C') \ll n e^{C \log n/\log\log(n+1)}.$$

2. Note that, as opposed to [47, Lemma 3.3], our bound for $M(n,\delta)$ is not uniform in the geodesic L, which explains why we are able to get a factor δ instead of only $\sqrt{\delta}$. But this improvement will not play a major role when we use this bound in the proof of Lemma 1.47 below. In fact, the larger part of our estimate of the error term in Proposition 1.21 will come from the η with $d(\bar{\rho}(\eta), N_G(A_L)) \asymp 1$, for which the higher power of δ in Lemma 1.27 gives no improvement.

1.4 Stationary phase

We now prove Proposition 1.23. By expanding k_ν in terms of spherical functions, this will reduce to finding an asymptotic estimate for an integral of the form

$$\mathscr{L}(r) = \int_{\mathbb{R}} c(t) \varphi_{ir}(a(t)) dt\,, \tag{1.32}$$

where $c : \mathbb{R} \to [0,1]$ is some smooth and compactly supported function satisfying $c(0) > 0$. We have [36, (1.43)]

$$\varphi_{ir}(a(t)) = \int_{S^1} (\cosh t + x_1 \sinh t)^{ir - \frac{1}{2}} d\mu(x)\,, \tag{1.33}$$

where $S^1 = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$ and the integral is with respect to the Haar measure. Hence

$$\mathscr{L}(r) = \int_{\mathbb{R}} \int_{S^1} c(t) (\cosh t + x_1 \sinh t)^{ir - \frac{1}{2}} d\mu(x) dt\,.$$

By the stationary phase theorem, the asymptotic behavior of this oscillating integral is prescribed by the local properties of the integrand at the critical points of the phase function

$$\begin{aligned}\phi : S^1 \times \mathbb{R} &\to \mathbb{R} \\ (x,t) &\mapsto \log(\cosh t + x_1 \sinh t)\,,\end{aligned}$$

which we now analyze.

Lemma 1.34. *The critical points of ϕ are $((0,\pm 1),0)$, and they are nondegenerate. The Hessian in the (x_1,t)-coordinates at those points is of shape*

$$\mathbf{H} = \begin{pmatrix} 0 & 1 \\ 1 & * \end{pmatrix},$$

which has signature $(p,q) = (1,1)$.

Proof. At the points where $x_2 = 0$ we have $\partial\phi/\partial t = \frac{\sinh t + x_1 \cosh t}{\cosh t + x_1 \sinh t} \neq 0$, because $|\cosh t| > |\sinh t|$. Hence they cannot be critical points. In the open set where $x_2 \neq 0$, we have $\partial\phi/\partial x_1 = \frac{\sinh t}{\cosh t + x_1 \sinh t}$. This is zero only when $t = 0$. We have $\partial\phi/\partial t = \frac{\sinh t + x_1 \cosh t}{\cosh t + x_1 \sinh t}$, which does not vanish when $t = 0$, unless $x_1 = 0$. Thus the only critical points are $((0,\pm 1),0)$.

We have $\phi(x,0) = 0$, so that at the points where $t = 0$ one has $\partial^2\phi/\partial x_1^2 = 0$. At those points we have $\partial\phi/\partial t = x_1$, so that $\partial^2\phi/\partial t\partial x_1 = 1$. This proves that the Hessian at the critical points of ϕ is of shape $\mathbf{H}$. □

Proposition 1.35. *Define $\mathscr{L}(r)$ by (1.32). There exists $C > 0$ such that*

$$\mathscr{L}(r) = Cr^{-1} + O(r^{-2})$$

as $r \to \infty$. The constant C and the implicit constant are allowed to depend on the cutoff function c.

Proof. We use the stationary phase theorem [69, §VIII.2] on the Riemannian manifold $S^1 \times \mathbb{R}$ with the phase function ϕ. Let $\mathbf{H}$, p and q be as in Lemma 1.34, which gives the nature of the critical points of ϕ. Adding the contributions of the two critical points, the stationary phase theorem implies that $\mathscr{L}(r) = Cr^{-1} + O(r^{-2})$, with

$$C = 2 \cdot c(0) \cdot 2\pi \cdot |\det(\mathbf{H})|^{-1/2} \cdot e^{i\pi(p-q)/4} = 4\pi c(0)\,.$$ □

Proof of Proposition 1.23. Let $c : \mathbb{R} \to [0,1]$ be smooth, compactly supported and such that $c(t) = 1$ when $a(t) \in \bigcup_{\nu \geq 0} \operatorname{supp} k_\nu$. Define $\mathscr{L}(r)$ as in (1.32) with respect to this

choice of c. By (1.9) and Fubini,

$$\begin{aligned}\int_{\mathbb{R}} k_\nu(a(t))dt &= \int_{\mathbb{R}} k_\nu(a(t))c(t)dt \\ &= \int_{\mathbb{R}}\int_0^\infty \widehat{k}_\nu(ir)\varphi_{ir}(a(t))\beta(r)c(t)drdt \\ &= \int_0^\infty \widehat{k}_\nu(ir)\beta(r)\mathscr{L}(r)dr\end{aligned}$$

because the compact support of c and the rapid decay of $\widehat{k}_\nu$ make the double integral absolutely convergent. Note that $\mathscr{L}(r)$ is a real number for every $r \in \mathbb{R}$. By Proposition 1.35, there exist $r_0, C > 0$ such that $\beta(r)\mathscr{L}(r) \in [C/2, 2C]$ for $r \geq r_0$. Write

$$\begin{aligned}&\int_0^\infty \widehat{k}_\nu(ir)\beta(r)\mathscr{L}(r)dr \\ &= \int_0^{r_0} \widehat{k}_\nu(ir)\beta(r)\mathscr{L}(r)dr + \int_{r_0}^\infty \widehat{k}_\nu(ir)\beta(r)\mathscr{L}(r)dr\,.\end{aligned}$$

For $\nu > r_0$, the first term is bounded by $r_0(\nu - r_0)^{-1} \cdot \sup_{r\in[0,r_0]} \beta(r)|\mathscr{L}(r)|$, and thus $o(1)$ as $\nu \to \infty$. Because $\beta(r)\mathscr{L}(r) \in [C/2, 2C]$ for $r \geq r_0$, for the second term we have

$$\int_{r_0}^\infty \widehat{k}_\nu(ir)\beta(r)\mathscr{L}(r)dr \asymp \int_{r_0}^\infty \widehat{k}_\nu(ir)dr\,,$$

by using that $\widehat{k}_\nu(ir)$ is nonnegative and not identically zero. The latter integral is bounded from above by

$$\ll \int_{r_0}^\infty (1 + |\nu - r|)^{-2}dr \leq \int_{-\infty}^\infty (1 + |r|)^{-2}dr \ll 1\,,$$

and when $\nu > r_0$ it is bounded from below by

$$\int_\nu^{\nu+1} \widehat{k}_\nu(ir)dr \geq 1\,.$$

Hence $\int_{\mathbb{R}} k_\nu(a(t))dt \asymp \int_{r_0}^\infty \widehat{k}_\nu(ir)dr \asymp 1$. □

1.5 Orbital integrals

We now prove the second part of Proposition 1.24. This is the same as the '$t = 0$' case of [47, Proposition 3.5], with the minor difference that the cutoff function b in the definition of $I(\nu, g)$ can have arbitrary support. We do include the proof here, because

in [47, §7] the argument is only sketched for brevity, and the fact that the upper bound there involves the distance to A instead of $N_G(A)$, appears to be a minor omission.

Throughout, fix a positive constant $C' > 0$, which we will later take to be the constant from Proposition 1.24. The implicit constants in the estimates are allowed to depend on C'.

Expanding k_ν in terms of spherical functions (see (1.9)), we have for $\nu \geq 0$ and $g \in G$,

$$I(\nu, g) = \int_0^\infty J(r, g)\widehat{k}_\nu(ir)\beta(r)dr$$

where

$$J(r, g) = \int_K \int_{\mathbb{R}\times\mathbb{R}} b(s)b(t)e^{(1/2+ir)H(ka(-s)ga(t))}dsdt\,dk$$

and H is as in §1.1.5. We will use the stationary phase method to prove an upper bound for $J(r, g)$. Proposition 1.24 will then follow by integrating this bound against $\widehat{k}_\nu(ir)\beta(r)$.

For each $g \in G$, we have a map $\alpha_g : K \to K$ defined by requiring that

$$kg \in NA\alpha_g(k)\,.$$

It is smooth in (g, k) because the Iwasawa decomposition $G \xrightarrow{\sim} N \times A \times K$ is smooth, and we have $\alpha_{gh} = \alpha_h \circ \alpha_g$. For $k \in K$ and $y, z \in G$ we have that (see for example [47, Lemma 6.2])

$$H(ky^{-1}z) = H(\alpha_{y^{-1}}(k)z) - H(\alpha_{y^{-1}}(k)y)\,.$$

This allows us to separate variables inside the exponential in the definition of $J(r, g)$:

$$\begin{aligned} J(r, g) &= \int_K \int_{\mathbb{R}\times\mathbb{R}} b(s)b(t)e^{(1/2+ir)H(ka(-s)ga(t))}dsdt\,dk \\ &= \frac{1}{2\pi}\int_0^{2\pi} \int_{\mathbb{R}\times\mathbb{R}} b(s)b(t)e^{(1/2+ir)H(k(\sigma)a(-s)ga(t))}dsdt\,d\sigma \\ &= \frac{1}{2\pi}\int_0^{2\pi} \int_{\mathbb{R}\times\mathbb{R}} b'(s,\theta)b(s)b(t)e^{(1/2+ir)(H(k(\theta)ga(t))-H(k(\theta)a(s)))}dsdt\,d\theta\,, \end{aligned}$$

by substituting $k(\theta) = \alpha_{a(-s)}(k(\sigma))$. Here, b' is a nonzero smooth function on $\mathbb{R}\times\mathbb{R}/2\pi\mathbb{Z}$, coming from the change of variables. This oscillating integral has phase function

$$\phi(s, t, \theta, g) := H(k(\theta)ga(t)) - H(k(\theta)a(s))\,. \tag{1.36}$$

Define $c(s, t, \theta, g) = (2\pi)^{-1}b'(s, \theta)b(s)b(t)\exp(\phi(s, t, \theta, g)/2)$ so that

$$J(r, g) = \int_0^{2\pi} \int_{\mathbb{R}\times\mathbb{R}} c(s, t, \theta, g)e^{ir\phi(s,t,\theta,g)}dsdt\,d\theta\,.$$

1.5.1 Critical points of ϕ

We analyze the critical points of ϕ for fixed θ and g. Let $\mathcal{P} = \mathbb{R}/2\pi\mathbb{Z} \times G$ and let $\mathcal{S} \subset \mathcal{P}$ be the closed subset consisting of 'singular' parameters (θ, g) for which one of the geodesics $k(\theta)Ai$ and $k(\theta)gAi$ is vertical. For every g, there are at most four values of θ for which $(\theta, g) \in \mathcal{S}$.

Proposition 1.37. *When $(\theta, g) \in \mathcal{S}$ is fixed, ϕ has no critical points. When $(\theta, g) \in \mathcal{P} - \mathcal{S}$, ϕ has a unique critical point with Hessian given in (s, t)-coordinates by*

$$\begin{pmatrix} \frac{1}{2} & 0 \\ 0 & -\frac{1}{2} \end{pmatrix}.$$

Proof. When g and θ are fixed, we have that (s, t) is a critical point of ϕ if and only if

$$\frac{\partial H(k(\theta)a(s))}{\partial s} = \frac{\partial H(k(\theta)ga(t))}{\partial t} = 0\,.$$

That is, when the geodesics $k(\theta)Ai$ and $k(\theta)gAi$ in $\mathbb{H}$ are not vertical, i.e., are half-circles, and their midpoints are $k(\theta)a(s)i$ resp. $k(\theta)ga(t)i$. (Recall that $H(g) = \log(\mathrm{Im}(gi))$.) It is clear from the geometric interpretation that ϕ has no critical points when $(\theta, g) \in \mathcal{S}$, and has exactly one critical point when $(\theta, g) \in \mathcal{P} - \mathcal{S}$. Moreover, this critical point is then nondegenerate, because the above geodesics (which are half-circles) have nonzero (Euclidean) curvature at their midpoints. Finally, the shape of the Hessian is computed in the proof of [47, Lemma 7.10]. □

When $(\theta, g) \in \mathcal{P} - \mathcal{S}$, let $(\xi_1(\theta), \xi_2(\theta, g))$ be the unique critical point of ϕ. The following lemma says that this point diverges to ∞ as (θ, g) approaches $\mathcal{S}$.

Lemma 1.38. *Let $R > 0$. Suppose $g \in G$ with $d(g, 1) \leq C'$ and $\theta \in \mathbb{R}/2\pi\mathbb{Z}$ are such that $|\xi_1(\theta)|, |\xi_2(\theta, g)| \leq R$. Then there exists $\delta = \delta(R, C') > 0$ such that $d((\theta, g), \mathcal{S}) > \delta$. In particular, the set*

$$\mathcal{P}_0 = \{(\theta, g) \in \mathcal{P} - \mathcal{S} : (\xi_1(\theta), \xi_2(\theta, g), \theta, g) \in \operatorname{supp} c,\ d(g, 1) \leq C'\}$$

is at positive distance from $\mathcal{S}$, and is compact.

Proof. For the sake of contradiction, suppose that $d((\theta, g), \mathcal{S})$ could be arbitrarily small. Then there exists a sequence (θ_n, g_n) for which the point

$$(\theta_n, g_n, \xi_1(\theta_n, g_n), \xi_2(\theta_n, g_n))$$

converges in $\mathbb{R}/2\pi\mathbb{Z} \times G \times \mathbb{R}^2$, with a limit in $\mathcal{S} \times \mathbb{R}^2$. Call (θ, g, x, y) its limit. By continuity of $\partial H(k(\theta)a(s))/\partial s$ and $\partial H(k(\theta)ga(t))/\partial t$, we would have

$$\left.\frac{\partial H(k(\theta)a(s))}{\partial s}\right|_{s=x} = \left.\frac{\partial H(k(\theta)ga(t))}{\partial t}\right|_{t=y} = 0\,,$$

contradicting that one of the geodesics $k(\theta)Ai$, $k(\theta)gAi$ is vertical. It follows that indeed $d(\mathcal{P}_0, \mathcal{S}) > 0$. Because $\mathcal{P}_0$ is bounded in $\mathcal{P}$ and closed in $\mathcal{P} - \mathcal{S}$, it is compact. □

For $(\theta, g) \in \mathcal{P} - \mathcal{S}$, define

$$\psi(\theta, g) = \phi(\xi_1(\theta), \xi_2(\theta, g), \theta, g)\,. \tag{1.39}$$

Define also

$$c_1(\theta, g) = \begin{cases} \pi c(\xi_1(\theta), \xi_2(\theta, g), \theta, g) & : (\theta, g) \in \mathcal{P} - \mathcal{S}\,, \\ 0 & : (\theta, g) \in \mathcal{S}\,. \end{cases}$$

Lemma 1.38 implies that c_1 is smooth on $\mathcal{P}$. Let $\mathcal{P}_0$ be as in Lemma 1.38.

Lemma 1.40. *We have*

$$J(r, g) = r^{-1} \int_0^{2\pi} c_1(\theta, g) e^{ir\psi(\theta,g)} d\theta + O(r^{-2}) \tag{1.41}$$

as $r \to \infty$, where the implicit constant remains bounded as long as $d(g, 1) \le C'$.

Proof. For fixed θ, g we apply the method of stationary phase in the variables s and t: suppose $(\theta, g) \in \mathcal{P} - \mathcal{S}$. Then ϕ has a unique critical point with Hessian determinant $-1/4$. Stationary phase [69, §VIII.2] implies

$$\int_{\mathbb{R}\times\mathbb{R}} c(s, t, \theta, g) e^{ir\phi(s,t,\theta,g)} ds dt = r^{-1} e^{ir\psi(\theta,g)} c_1(\theta, g) + O(r^{-2})\,, \tag{1.42}$$

uniformly for (θ, g) in compact subsets of $\mathcal{P} - \mathcal{S}$. Suppose $(\theta, g) \in \mathcal{P} - \mathcal{P}_0$, so that ϕ has no critical point in the support of c. The absence of critical points implies

$$\int_{\mathbb{R}\times\mathbb{R}} c(s, t, \theta, g) e^{ir\phi(s,t,\theta,g)} ds dt \ll_N r^{-N}\,,$$

uniformly for (θ, g) in compact subsets of $\mathcal{P} - \mathcal{P}_0$. Because $c_1(\theta, g) = 0$ in this case, we see that (1.42) still holds (after assigning any value to $\psi(\theta, g)$ when $(\theta, g) \in \mathcal{S}$). We may now take compact subsets of $\mathcal{P} - \mathcal{S}$ and of $\mathcal{P} - \mathcal{P}_0$ such that the union of the two contains $\mathbb{R}/2\pi\mathbb{Z} \times \{g \in G : d(g, 1) \le C'\}$. Integrating (1.42) with respect to θ then yields the desired estimate. □

1.5.2 Critical points of ψ

In view of the expression (1.41) for $J(r, g)$, we analyze the critical points of ψ, for fixed g. When $(\theta, g) \notin \mathcal{S}$, we have [47, Proposition 7.2, Lemma 7.3] that θ is a critical point of ψ if and only if the geodesics $k(\theta)Ai$ and $k(\theta)gAi$, which by assumption are half-circles, are concentric. A critical point θ is degenerate if and only if these geodesics coincide, that is, if and only if $g \in N_G(A)$, in which case every θ for which $(\theta, g) \notin \mathcal{S}$ is a degenerate critical point.

Remark 1.43. For $g \notin N_G(A)$, the locations of the critical points of $\psi(-, g)$ can be described geometrically: as stated above, θ is a critical point if and only if the half-circles $k(\theta)Ai$ and $k(\theta)gAi$ are concentric, that is, if and only if they admit a common perpendicular vertical geodesic. In particular, the geodesics $k(\theta)Ai$ and $k(\theta)gAi$ must not intersect in $\mathbb{H} \cup \mathbb{P}^1(\mathbb{R})$, that is, the geodesic gAi must have both endpoints in $\mathbb{R}_{>0}$ or in $\mathbb{R}_{<0}$. This condition on g is also sufficient for the existence of a critical point: under this assumption on g, there are exactly two critical points, which can be described as follows. Because Ai and gAi do not intersect in $\mathbb{H} \cup \mathbb{P}^1(\mathbb{R})$, these geodesics lie at positive distance from each other, and this distance is realized in $\mathbb{H}$ by a pair of points on Ai and gAi. Let j be the geodesic carrying the geodesic segment joining those points. As a consequence of Gauss's lemma, j is then perpendicular to both geodesics. Because the sum of the angles of a hyperbolic quadrilateral is less than $360°$, there can be no other geodesic that is perpendicular to both. Thus the critical points are the two values of θ for which $k(\theta)j$ is vertical.

To bound $J(r, g)$, we distinguish two cases, depending on whether g is close to $N_G(A)$ or at positive distance from it. For g away from $N_G(A)$, the absence of nondegenerate critical points of ψ implies the bound below.

Lemma 1.44. *When $d(g, N_G(A)) \geq \delta > 0$ and $d(g, 1) \leq C'$, we have*

$$\int_0^{2\pi} c_1(\theta, g) e^{ir\psi(\theta,g)} d\theta \ll_\delta (1 + r \cdot d(g, N_G(A)))^{-1/2}$$

uniformly in g and $r \geq 0$.

Proof. Because ψ has no degenerate critical points when $g \notin N_G(A)$, we have

$$\max(|\partial\psi/\partial\theta|, |\partial^2\psi/\partial\theta^2|) \gg 1$$

uniformly for g in compact subsets of $G - N_G(A)$ and for $\theta \in \operatorname{supp} c_1(-, g)$. The Van der Corput lemma [69, §VIII.1, Proposition 2, Corollary] implies that

$$\int_0^{2\pi} c_1(\theta, g) e^{ir\psi(\theta,g)} d\theta \ll r^{-1/2}$$

as $r \to \infty$, uniformly for g in compact subsets of $G - N_G(A)$. □

We may now restrict our attention to g that are close to $N_G(A)$. By (1.36) and (1.39),

$$\psi(\theta, g) := H(k(\theta)ga(\xi_2(\theta, g))) - H(k(\theta)a(\xi_1(\theta))) .$$

From the characterization of $k(\theta)ga(\xi_2(\theta, g))i$ as being the midpoint of the geodesic $k(\theta)gAi$, we see that ψ is right $N_G(A)$-invariant in g. Thus in order to estimate the

right-hand side of (1.41), we will first assume that g lies in a small neighborhood of the identity, and then translate the estimate to a small neighborhood of $N_G(A)$.

By explicating the order of vanishing of $\partial^2\psi(\theta,g)/\partial\theta^2$ as $d(g,A)\to 0$, an application of the Van der Corput lemma shows the following bound.

Lemma 1.45. *[47, Corollary 7.6] There is an open neighborhood U of $1\in G$ such that for all $g\in U$ and $b\in C_0^\infty(\mathbb{R}/2\pi\mathbb{Z}-\{\theta:(\theta,g)\in\mathcal{S}\})$ we have*

$$\int_0^{2\pi} b(\theta)e^{ir\psi(\theta,g)}d\theta \ll (1+r\cdot d(g,A))^{-1/2}\,,$$

where the implicit constant remains bounded when $\operatorname{supp}(b)$ *stays at positive distance from the set $\{\theta:(\theta,g)\in\mathcal{S}\}$ and the derivatives of b up to a certain order remain bounded.*

Corollary 1.46. *There exists an open neighborhood V of $N_G(A)$ in G such that when $g\in V$ and $d(g,1)\le C'$,*

$$\int_0^{2\pi} c_1(\theta,g)e^{ir\psi(\theta,g)}d\theta \ll (1+r\cdot d(g,N_G(A)))^{-1/2}$$

uniformly in g and $r\ge 0$.

Proof. Let U be the neighborhood from Lemma 1.45. Take $g_0\in N_G(A)$. When $g\in Ug_0$, then Lemma 1.45 applied to gg_0^{-1} and $b(\theta)=c_1(\theta,g)$ shows that

$$\begin{aligned}\int_0^{2\pi} c_1(\theta,g)e^{ir\psi(\theta,g)} &= \int_0^{2\pi} c_1(\theta,g)e^{ir\psi(\theta,gg_0^{-1})}d\theta\\ &\ll_{g_0} (1+r\cdot d(gg_0^{-1},A))^{-1/2}\\ &\ll_{g_0} (1+r\cdot d(g,N_G(A)))^{-1/2}\end{aligned}$$

uniformly in $g\in Ug_0$. We may now take a fixed subset $N_0\subset N_G(A)$ such that the Ug_0 for $g_0\in N_0$ form a locally finite cover of $N_G(A)$, and let $V=\bigcup_{g_0\in N_0}Ug_0$. □

Proof of Proposition 1.24. Now take C' be the constant from the first part of Proposition 1.24. We may restrict to the $g\in G$ with $d(g,1)\le C'$. The estimates from Lemma 1.44 and Corollary 1.46 may be plugged into Lemma 1.40 to obtain

$$|J(r,g)|\ll r^{-1}(1+r\cdot d(g,N_G(A)))^{-1/2}$$

as $r\to\infty$, uniformly when $d(g,1)\le C'$. On the one hand we have, by estimating $|J(r,g)|\beta(r)\ll 1$, that

$$I(\nu,g)=\int_0^\infty J(r,g)\widehat{k}_\nu(ir)\beta(r)dr\ll\|\widehat{k}_\nu\|_{L^1(i\mathbb{R})}\ll 1\,.$$

On the other hand, when $g \notin N_G(A)$,

$$\begin{aligned} I(\nu, g) &\ll \int_0^{\nu/2} \widehat{k}_\nu(ir)dr + \int_{\nu/2}^{\infty} (r \cdot d(g, N_G(A)))^{-1/2} \widehat{k}_\nu(ir)dr \\ &\ll_N \nu^{-N} + d(g, N_G(A))^{-1/2} \nu^{-1/2} \int_{\nu/2}^{\infty} \widehat{k}_\nu(ir)dr \\ &\ll d(g, N_G(A))^{-1/2} \nu^{-1/2}, \end{aligned}$$

where we have used that $1 \ll d(g,1)^{-1/2} \leq d(g, N_G(A))^{-1/2}$. Combining the two estimates, the claim follows. □

1.6 Proof of Theorem 2

1.6.1 Amplification

Let $(a_n)_{n\geq 1}$ be a sequence of nonnegative real numbers, supported on a finite set of integers $n \geq 1$ that are coprime to the discriminant Δ_R. Let $T = (\sum a_n T_n)^2$, and denote by $\widehat{T}(\phi_j) \geq 0$ the T-eigenvalue of ϕ_j. Define for $\nu \geq 0$,

$$\begin{aligned} Q(\nu, a) &:= \sum_j \widehat{k}_\nu(i\nu_j) \widehat{T}(\phi_j), \\ Q_L(\nu, a) &:= \sum_j \widehat{k}_\nu(i\nu_j) \widehat{T}(\phi_j) \mathscr{P}(\phi_j)^2. \end{aligned}$$

Our aim is to obtain good asymptotic estimates for these spectral sums.

Lemma 1.47. *Define the orbital integral $I(\nu, g)$ by (1.22). There exists an absolute constant $C > 0$, such that for $n \geq 1$ and $\nu \geq 0$ we have*

$$\sum_{\eta \in R(n) - N_{R^+}(F)} |I(\nu, g_0^{-1}\eta g_0)| \ll (1+\nu)^{-1/2} n e^{C \log n / \log\log(n+1)}.$$

The implicit constant is allowed to depend on ℓ.

Proof. By Lemma 1.24, there exists $C' > 0$ independent of ν and n such that only the terms with $d(g_0^{-1}\eta g_0, 1) \leq C'$ have a nonzero contribution. Cover the interval $[0, C']$ with the intervals $I_0 = [0, (1+\nu)^{-1}]$, $I_k = [e^{k-1}(1+\nu)^{-1}, e^k(1+\nu)^{-1}]$ for $1 \leq k \leq \log(C'(1+\nu)) + 1$. Define $M(n, \delta)$ as in §1.3.2, with respect to this value of C'. When $d(g_0^{-1}\eta g_0, N_G(A)) \in I_0$, we apply the bounds

$$\begin{aligned} I(\nu, g_0^{-1}\eta g_0) &\ll 1, \\ \#M(n, (1+\nu)^{-1}) &\ll (1+\nu)^{-1} n e^{C \log n / \log\log(n+1)} \end{aligned}$$

from Proposition 1.24 and Lemma 1.27, which imply that the contribution of all η with $d(g_0^{-1}\eta g_0, N_G(A)) \in I_0$ is bounded by $\ll (1+\nu)^{-1} n e^{C\log n/\log\log(n+1)}$ for some $C > 0$. When $d(g_0^{-1}\eta g_0, N_G(A)) \in I_k$, we apply the bounds

$$I(\nu, g_0^{-1}\eta g_0) \ll e^{-k/2}\,,$$
$$\#M(n, e^k(1+\nu)^{-1}) \ll e^k(1+\nu)^{-1} n e^{C\log n/\log\log(n+1)}\,.$$

Summing over k, the contribution of the η with $d(g_0^{-1}\eta g_0, N_G(A)) \notin I_0$ is bounded by $\ll (1+\nu)^{-1/2} n e^{C\log n/\log\log(n+1)}$. □

Proposition 1.48. *There exists an absolute constant $C > 0$ such that for (a_n) as above and for all $\nu \geq 0$,*

$$Q(\nu, a) = B(a)\operatorname{Vol}(\Gamma\backslash\mathbb{H})k_\nu(1) + O(R(a))\,,$$
$$Q_L(\nu, a) = B_L(a)\operatorname{Vol}(\Gamma_L\backslash L)\int_{\mathbb{R}} k_\nu(a(t))dt + O((1+\nu)^{-1/2}R_L(a))\,,$$

where

$$B(a) := \sum_{m,n} a_m a_n \sum_{d|m,n} d\cdot s(mn/d^2)\,,$$
$$R(a) := \sum_{m,n} a_m a_n \sum_{d|m,n} d\cdot \Big(\frac{mn}{d^2}\Big)^{3/2}\cdot e^{C\log(mn/d^2)/\log\log(1+mn/d^2)}\,,$$
$$B_L(a) := \sum_{m,n} a_m a_n \sum_{d|m,n} d\cdot |N_{R(mn/d^2)}(F)/R_F^1|\,,$$
$$R_L(a) := \sum_{m,n} a_m a_n \sum_{d|m,n} d\cdot \frac{mn}{d^2}\cdot e^{C\log(mn/d^2)/\log\log(1+mn/d^2)}\,.$$

The implicit constants are uniform in (a_n) and ν, but not in ℓ.

Proof. The first asymptotic formula follows from the recurrence relation (1.7) and Proposition 1.17. The second statement follows similarly from Proposition 1.21 and the estimate from Lemma 1.47. □

In order to bound the tails of the spectral sums $Q(\nu, a)$ and $Q_L(\nu, a)$, we shall need the following lemma.

Lemma 1.49. *Let (a_n) be as above. Then for $\nu \geq 0$,*

$$\sum_{|\nu_j-\nu|\leq 1} \widehat{T}(\phi_j) \ll (\nu+1)\cdot B(a) + R(a)\,,$$
$$\sum_{|\nu_j-\nu|\leq 1} \widehat{T}(\phi_j)\mathscr{P}(\phi_j)^2 \ll B_L(a) + (\nu+1)^{-1/2}R_L(a)\,,$$

uniformly in (a_n) and ν, but not in ℓ.

Proof. For the first estimate, we start from Proposition 1.48 and plug in the bound $k_\nu(1) \asymp \nu$, to obtain

$$\sum_j \widehat{k}_\nu(i\nu_j)\widehat{T}(\phi_j) \leq C_0 \cdot (\nu \cdot B(a) + R(a)),$$

for some $C_0 > 0$ and for $\nu \geq \nu_{(0)}$, where $\nu_{(0)} > 0$ depends on the choice of the family (k_ν) given by Proposition 1.8. Here, we used that $B(a), R(a) \geq 0$. The terms in the left-hand side are nonnegative. Because $\widehat{k}_\nu(i\nu_j) \geq 1$ when $|\nu_j - \nu| \leq 1$, the claim follows for $\nu \geq \nu_{(0)}$ by discarding the terms with $|\nu_j - \nu| > 1$.

To treat the case where $\nu \leq \nu_{(0)}$, one may either use Proposition 1.48 and observe that our specific construction of the family (k_ν) satisfies $k_\nu(1) \ll 1$ for $\nu \ll 1$, or use the following argument: for every $\nu \leq \nu_{(0)}$, we may find integers $n_1, n_2 \in [0, \nu_{(0)}]$ such that the set $\overline{B}(\nu, 1) \cap \{\nu_j : j \geq 0\}$ is contained in $\overline{B}(n_1, 1) \cup \overline{B}(n_2, 1)$. Proposition 1.48 applied to k_{n_1} and k_{n_2} gives

$$\begin{aligned}\sum_{|\nu_j - \nu| \leq 1} \widehat{T}(\phi_j) &\leq \sum_j \widehat{k}_{n_1}(\nu_j)\widehat{T}(\phi_j) + \sum_j \widehat{k}_{n_2}(\nu_j)\widehat{T}(\phi_j) \\ &\ll \max_{n \in [0, \nu_{(0)}] \cap \mathbb{Z}} k_n(1) \cdot B(a) + R(a),\end{aligned}$$

as desired. The second estimate in the statement is proven similarly, by using the bound $\int_{\mathbb{R}} k_\nu(a(t)) \ll 1$ from Proposition 1.23. □

Proposition 1.50. *For (a_n) as above, $C \geq 1$ and $\nu > C$,*

$$\begin{aligned}\sum_{|\nu_j - \nu| \leq C} \widehat{k}_\nu(i\nu_j)\widehat{T}(\phi_j) - B(a)\operatorname{Vol}(\Gamma\backslash\mathbb{H})k_\nu(1) \\ \ll C^{-1}\nu B(a) + R(a),\end{aligned}$$

$$\begin{aligned}\sum_{|\nu_j - \nu| \leq C} \widehat{k}_\nu(i\nu_j)\widehat{T}(\phi_j)\mathscr{P}(\phi_j)^2 - B_L(a)\operatorname{Vol}(\Gamma_L\backslash L)\int_{\mathbb{R}} k_\nu(a(t))dt \\ \ll C^{-1}B_L(a) + \nu^{-1/2}R_L(a),\end{aligned}$$

where the implicit constants are uniform in (a_n), C and ν, but not in ℓ.

Proof. This will follow by combining Lemma 1.49 with the rapid decay of $\widehat{k}_\nu$. By Proposition 1.48, for the first estimate it suffices to prove that

$$\sum_{|\nu_j - \nu| > C} \widehat{k}_\nu(i\nu_j)\widehat{T}(\phi_j) \ll C^{-1}\nu B(a) + R(a).$$

When $\nu > C$, the condition $|\nu_j - \nu| > C$ is equivalent to $|\operatorname{Re}(\nu_j) - \nu| > C$. Indeed, when $\nu_j \in \mathbb{R}$ this is trivial, and when $\nu_j \in [-i/2, i/2]$, both conditions are satisfied. Consider

first the sum over the set $\{j : \mathrm{Re}(\nu_j) < \nu - C\}$. We break it up into sums where $\mathrm{Re}(\nu_j)$ belongs to an interval of length 1: for $n \geq 0$, we have

$$\begin{aligned}\sum_{\nu-C-(n+1)\leq \mathrm{Re}(\nu_j)<\nu-C-n} \widehat{k}_\nu(i\nu_j)\widehat{T}(\phi_j) &\ll \frac{1}{(C+n)^2} \sum_{|\nu_j-(\nu-C-n-\frac{1}{2})|\leq 1} \widehat{T}(\phi_j) \\ &\ll \frac{\nu B(a)+R(a)}{(C+n)^2}\,,\end{aligned}$$

where we use the rapid decay of $\widehat{k}_\nu$ and Lemma 1.49. Summing over integers $n \in [0, \nu - C]$, we find that the sum over $\mathrm{Re}(\nu_j) < \nu - C$ is bounded up to a constant by $C^{-1}(\nu B(a) + R(a))$. The sum over $\mathrm{Re}(\nu_j) > \nu + C$ is similarly bounded up to a constant by

$$\begin{aligned}\sum_{n=0}^{\infty} \frac{(\nu+C+n)B(a)+R(a)}{(C+n)^3} &\leq \sum_{n=0}^{\infty} \frac{\nu(C+n)B(a)+R(a)}{(C+n)^3} \\ &\ll C^{-1}(\nu B(a)+R(a))\,.\end{aligned}$$

Combining the two estimates, the first statement follows. For the second statement, it suffices to prove that

$$\sum_{|\nu_j-\nu|>C} \widehat{k}_\nu(i\nu_j)\widehat{T}(\phi_j)\mathscr{P}(\phi_j)^2 \ll C^{-1}B_L(a) + \nu^{-1/2}R_L(a)\,.$$

Using the rapid decay of $\widehat{k}_\nu$ and Lemma 1.49, we find that the sum over the j with $\mathrm{Re}(\nu_j) > \nu + C$ is bounded up to a constant by

$$\sum_{n=0}^{\infty} \frac{B_L(a)+\nu^{-1/2}R_L(a)}{(C+n)^2} \ll C^{-1}(B_L(a)+\nu^{-1/2}R_L(a))\,.$$

The sum over the j with $\mathrm{Re}(\nu_j) < \nu - C$ is bounded up to a constant by

$$\begin{aligned}&\sum_{n\leq \nu-C} \frac{B_L(a)+(|\nu-C-n|+1)^{-1/2}R_L(a)}{(C+n)^2} \\ &\ll C^{-1}B_L(a) + R_L(a)\left(\sum_{C+n\leq\sqrt{\nu}} \frac{\nu^{-1/2}}{(C+n)^2} + \sum_{\sqrt{\nu}\leq C+n} \frac{1}{(C+n)^2}\right) \\ &\ll C^{-1}B_L(a) + \nu^{-1/2}R_L(a)\,.\end{aligned}$$

The second statement follows. □

1.6.2 Optimal resonators

Let $F \subset B$ as before be the field corresponding to the geodesic L, and let $\mathfrak{f}_{R_F}$ be the conductor of R_F. We seek to apply Proposition 1.50 to a sequence (a_n) for which the quotient $B_L(a)/B(a)$ is large, and for which the sums $R(a)$ and $R_L(a)$ are small compared to $B(a)$ resp. $B_L(a)$. A similar optimization problem has been considered in [50] in the context of lower bounds for point evaluations of Maass cusp forms. Recycling some of the results there, we obtain the following in our situation.

Proposition 1.51 (Optimal resonators)**.** *Let $M > 3$ be a real number. There exists a sequence (a_n) of nonnegative real numbers supported on integers $n \leq M$ coprime to Δ_R, such that the following hold:*

1. *we have the lower bound*

$$\frac{B_L(a)}{B(a)} \geq \exp\left(2\sqrt{2}\sqrt{\frac{\log M}{\log\log M}}\left(1 + O\left(\frac{\log\log\log M}{\log\log M}\right)\right)\right); \tag{1.52}$$

2. *there exists a constant $C'' > 0$ independent of M such that*

$$R(a) \ll M^3 e^{C''\log M/\log\log M}, \tag{1.53}$$

$$R_L(a) \ll M^2 e^{C''\log M/\log\log M}; \tag{1.54}$$

3. *$B(a) \geq 1$.*

The implicit constants are allowed to depend on ℓ.

Proof. We first reduce to the precise situation in [50]. When (a_n) is as in §1.6.1, define the sum $B_{R_F}(a)$ by replacing the set $N_{R(mn/d^2)}(F)/R_F^1$ by the set $P_{R_F}(mn/d^2)$ in the definition of $B_L(a)$. When (a_n) is supported on integers coprime to $\mathfrak{f}_{R_F}$, Lemma 1.25 implies $B_L(a) \geq B_{R_F}(a)$, so that $B_L(a)/B(a) \geq B_{R_F}(a)/B(a)$. The latter quantity is the one considered in [50].

When M is sufficiently large, we construct a sequence (a_n) as follows: let $L = \sqrt{2\log M \log\log M}$, and define a multiplicative function f by prescribing its values on prime powers by

$$f(p^n) := \begin{cases} \frac{L}{p\log p} & \text{if } \omega_F(p) = 1,\ p \nmid \Delta_R\mathfrak{f}_{R_F},\ n = 1,\ L^2 < p \leq \exp(\log^2 L), \\ 0 & \text{otherwise.} \end{cases}$$

Define $a_n = f(n)$ when $n \leq M$ and $a_n = 0$ otherwise. The proof of [50, Lemma 5] gives that $B_{R_F}(a)/B(a)$ is as large as the right-hand side in (1.52). (And it is shown that this sequence is optimal, in the sense that for every sequence (a_n) supported on integers

coprime to $\Delta_R \mathfrak{f}_{R_F}$, the ratio $B_{R_F}(a)/B(a)$ is at most as large as the right-hand side in (1.52).)

We have $B(a) \geq a_1^2 = 1$. It remains to prove the bounds (1.53) and (1.54). Because (a_n) is supported on integers $n \leq M$,

$$\begin{aligned} R(a) &\ll M^3 e^{C \log(M^2)/\log\log(M^2)} \sum_{m,n} a_m a_n \sum_{d \mid m,n} \frac{1}{d^2} \\ &\ll M^3 e^{2C \log M/\log\log M} \left(\sum_n a_n \right)^2 . \end{aligned}$$

Using Chebyshev's estimates, we have

$$\begin{aligned} \sum_n a_n \leq \prod_p (1 + f(p)) &\leq \exp\left(\sum_p f(p) \right) \\ &\leq \exp\left(L \cdot \sum_{L^2 < p} \frac{1}{p \log p} \right) \\ &\ll \exp\left(L/\log L \right) \\ &\ll \exp\left(\sqrt{\log M} \right) \end{aligned}$$

which proves (1.53). Similarly, the estimate

$$R_L(a) \ll M^2 e^{2C \log M/\log\log M} \sum_{m,n} a_m a_n \sum_{d \leq M} \frac{1}{d}$$

gives us (1.54). □

With this resonator sequence we are ready to prove Theorem 2.

Proof of Theorem 2. Let $\nu > 0$ be large. Let C'' be the constant from Proposition 1.51, choose any $A > C''/8$ and let $M = \nu^{1/4} e^{-A \log\nu/\log\log\nu}$. Let (a_n) be the corresponding sequence given by Proposition 1.51. We check that $R(a)$ and $R_L(a)$ are small compared to $B(a)$ and $B_L(a)$. From (1.53) we have $R(a) \ll_\epsilon \nu^{3/4+\epsilon}$, so that $R(a) = o(\nu B(a))$. From (1.54) we have

$$\begin{aligned} R_L(a) &\ll \nu^{1/2} e^{-2A \log\nu/\log\log\nu} e^{C'' \log M/\log\log M} \\ &\ll \nu^{1/2} e^{(-2A + C''/4)(\log\nu/\log\log\nu)(1+o(1))} \\ &\ll \nu^{1/2} , \end{aligned}$$

so that $\nu^{-1/2}R_L(a) = o(B_L(a))$. By Proposition 1.8 and Proposition 1.23, there exists $C \geq 1$ such that

$$C^{-1}\nu \leq \frac{1}{2}\operatorname{Vol}(\Gamma\backslash\mathbb{H})k_\nu(1)\,,$$
$$C^{-1} \leq \frac{1}{2}\operatorname{Vol}(\Gamma_L\backslash L)\int_{\mathbb{R}} k_\nu(a(t))dt\,,$$

for ν sufficiently large. Proposition 1.50 applied to the sequence (a_n) and such C gives

$$\sum_{|\nu_j-\nu|\leq C} \widehat{k}_\nu(i\nu_j)\widehat{T}(\phi_j) \asymp B(a)\nu\,,$$
$$\sum_{|\nu_j-\nu|\leq C} \widehat{k}_\nu(i\nu_j)\widehat{T}(\phi_j)\mathscr{P}(\phi_j)^2 \asymp B_L(a)\,.$$

In particular, there must exist at least one $\nu_j \in [\nu - C, \nu + C]$ with the property that $\mathscr{P}(\phi_j)^2 \gg B_L(a)/(\nu B(a))$. That is,

$$(\nu^{1/2}\mathscr{P}(\phi_j))^2 \gg \frac{B_L(a)}{B(a)}$$
$$\gg \exp\left(2\sqrt{2}\sqrt{\frac{\log M}{\log\log M}}\left(1 + O\left(\frac{\log\log\log M}{\log\log M}\right)\right)\right).$$

We have $\log M = \frac{1}{4}\log\nu\cdot(1+O(\log\log\nu/\log\nu))$, $\log\log M = \log\log\nu\cdot(1+O(1/\log\nu))$ and $\log\log\log M \ll \log\log\log\nu$. Using this, we obtain for this particular j,

$$(\nu^{1/2}\mathscr{P}(\phi_j))^2 \gg \exp\left(\sqrt{2}\sqrt{\frac{\log\nu}{\log\log\nu}}\left(1 + O\left(\frac{\log\log\log\nu}{\log\log\nu}\right)\right)\right).$$

Taking square roots and writing $\nu = \sqrt{\lambda - \frac{1}{4}}$, the theorem follows. □

1.6.3 Extreme values of L-functions

We now explain how to deduce Theorem 1 from (the proof of) Theorem 2. There is one additional issue: in general, the right-hand side of the period formula (1.5) involves a sum of periods, rather than a single one.

Proof of Theorem 1. We use the notations from the statement. Let B, R, Γ, Λ_{d_F} be as in Remark 1.4, and associate to f_j the Laplace–Hecke eigenfunction ϕ_j on $\Gamma\backslash\mathbb{H}$. By our assumptions on N and d_F, B is a quaternion division algebra and R a maximal order. Let ϕ_0 be the constant function $\operatorname{Vol}(\Gamma\backslash\mathbb{H})^{-1/2}$. Then $(\phi_j)_{j\geq 0}$ is an orthonormal basis of

$L^2(\Gamma\backslash\mathbb{H})$ as in §1.1.6. In (1.5), take $f = f_j$. The product of the Gamma factors in the left-hand side of (1.5) is of size $\lambda_j^{-1/2}$ by Stirling's formula, and we have the lower bound $L(1, \pi_j, \mathrm{Ad}) \gg 1/\log(1+\lambda_j)$ [28]. We have the factorization

$$L(s, \pi_j \times \pi_{\chi_0}) = L(s, \pi_j) L(s, \pi_j \times \omega_F)\,.$$

Thus it suffices to give a lower bound for the right-hand side of (1.5), of the same quality as in Theorem 2.

For every pair (ℓ, ℓ') of closed geodesics in Λ_{d_F}, choose lifts L, L' and fundamental domains $\mathscr{F}_L, \mathscr{F}_{L'}$ and integrate the amplified pre-trace formula (1.10) over $\mathscr{F}_L \times \mathscr{F}_{L'}$ to obtain

$$\sum_j \widehat{k}_\nu(i\nu_j)\widehat{T}_n(\phi_j)\mathscr{P}_\ell(\phi_j)\mathscr{P}_{\ell'}(\phi_j) = \sum_{\eta \in R(n)/\pm 1} \int_{\mathscr{F}_L \times \mathscr{F}_{L'}} k_\nu(x^{-1}\eta y)dxdy\,.$$

Say $L = g_0 Ai$ and $L' = g_0' Ai$, and call the sum in the right-hand side $S(g_0, g_0')$. Let F, F' be the subfields of B associated to L, L'. In $S(g_0, g_0')$, we isolate a main term, which is the sum over η with $\eta L' = L$, i.e. $\eta F' \eta^{-1} = F$. An unfolding argument as in the proof of Proposition 1.21 shows that the main term equals

$$\#(N_{F,F'}(n)/R^1_{F'})\,\mathrm{Vol}(\Gamma_L\backslash L)\int_{\mathbb{R}} k_\nu(a(t))dt\,,$$

where $N_{F,F'} = \{\eta \in R^+ : \eta F' \eta^{-1} = F\}$. In the error term, we may again introduce nonnegative smooth cutoff functions and write it as

$$\sum_{\eta \in (R(n) - N_{F,F'})/\pm 1} I_{F,F'}(\nu, g_0^{-1}\eta g_0')\,,$$

where $I_{F,F'}(\nu, g_0)$ is defined similarly to (1.22), in terms of the chosen cutoff functions. This error term may be bounded as in Lemma 1.47. Here, the Diophantine problem consists of counting $\eta \in R(n)$ that are at distance $\ll \sqrt{n}$ from 0 and are close to $N_{F,F'}$. Multiplication on the right by a fixed element $\eta_0 \in N_{F',F}$ gives an injection $N_{F,F'}(n) \to N_{R(nN_{B/\mathbb{Q}}(\eta_0))}(F)$, reducing to the counting problem in §1.3.2. The analogue of Proposition 1.24 remains true; the only thing that changes in its proof is the amplitude function in the oscillating integral, to which the proof is insensitive.

Summing over all pairs (ℓ, ℓ') and taking an amplifier (a_n) as in §1.6.1, Proposition 1.50 becomes

$$\begin{aligned}\sum_{|\nu_j - \nu| \le C} \widehat{k}_\nu(i\nu_j)\widehat{T}(\phi_j)\left(\sum_{\ell \in \Lambda_{d_F}} \mathscr{P}_\ell(\phi_j)\right)^2 \\ = \mathrm{Vol}(\Gamma_L\backslash L) B_{d_F}(a)\int_{\mathbb{R}} k_\nu(a(t)) + O(C^{-1}B_{d_F}(a) + R_{d_F}(a))\end{aligned} \tag{1.55}$$

where L is a lift of an arbitrary $\ell \in \Lambda_{d_F}$ (the volume is independent of the choice), $B_{d_F}(a)$ and $R_{d_F}(a)$ are defined similarly to $B_L(a)$ and $R_L(a)$ (see Proposition 1.48), and in particular we have $B_{d_F}(a) \geq B_L(a)$. We conclude by choosing the amplifier as in Proposition 1.51, taking $C > 0$ large enough and comparing (1.55) to the standard trace formula. □

Note that we have placed ourselves in a situation where the order R is maximal, in order to avoid having to worry about oldforms. When R is Eichler of arbitrary level N, it might be possible to extract the contribution of newforms, by writing down the trace formulas for Eichler orders R_d of levels $d \mid N$ and using Möbius-inversion. But the arithmetic of the cardinalities $\#(N_{F,F'}(n)/R_d^1)$ (which appear in the main term in (1.55)) as d varies, is intricate in general.

2 The geometry of maximal flat submanifolds

This chapter contains the proofs of the statements in §0.6, as well as refinements thereof and auxiliary results that will be used in §3. This section is divided into three parts; the results about the N-, A- and K-projections are bundled in §2.2, §2.3 and §2.4 respectively.

2.1 Preliminaries on Lie groups

The notation and lemmas concerning Lie groups in this chapter will be used again in §3.

2.1.1 Lie groups and Lie algebras

Let G be a reductive Lie group in the sense of Harish-Chandra [30]; see also [41, Chapter VII]. Most of the time G will be connected semisimple with finite center and we will then simply say G is semisimple. Let $K \subset G$ be a maximal compact subgroup and θ an involution of G whose fixed point set is K. It induces an involution θ of the Lie algebra $\mathfrak{g}$, whose $+1$ and -1 eigenspaces we denote by $\mathfrak{k}$ and $\mathfrak{p}$ respectively. We denote the exponential of $X \in \mathfrak{g}$ by $\exp(X)$. Define the semisimple part $\mathfrak{g}_{ss} = [\mathfrak{g}, \mathfrak{g}]$. We extend the Killing form on $\mathfrak{g}_{ss}$ to a nondegenerate symmetric bilinear form $\langle \cdot, \cdot \rangle$ on $\mathfrak{g}$ that is positive definite on $\mathfrak{k}$ and negative definite on $\mathfrak{p}$, and with respect to which the center $\mathfrak{z}(\mathfrak{g})$ is orthogonal to $\mathfrak{g}_{ss}$. Define $\langle \cdot, \cdot \rangle_\theta = \langle \cdot, -\theta(\cdot) \rangle$, a positive definite symmetric bilinear form. All statements on $\mathfrak{g}$ involving norms, orthogonality and adjoints will be with respect to $\langle \cdot, \cdot \rangle_\theta$. Let $\mathfrak{a} \subset \mathfrak{p}$ be a maximal abelian subalgebra and $A = \exp(\mathfrak{a})$. The choices of $\mathfrak{a}$ are all conjugate under K. Define $P = \exp(\mathfrak{p})$. The multiplication map $P \times K \to G$ is a diffeomorphism, known as the Cartan decomposition. In particular, K meets all components of G.

2.1.2 Symmetric spaces and maximal flats

References for the following facts about symmetric spaces are [22, 33].

Assume here that G is semisimple. Then θ is a Cartan involution. The quotient $S = G/K$ carries a left-G-invariant Riemannian metric induced by the Killing form on $\mathfrak{p}$. It is a symmetric space of non-compact type, and every such space arises in this way.

The maximal flat submanifolds of S are of the form gAK with $g \in G$. Such g is uniquely determined by the submanifold up to multiplication on the right by $N_G(A)$. When $\dim(A) = 1$, the maximal flats are precisely the geodesics. The rank of G is defined to be $\dim(A)$.

Let $p \in P$. The tangent space $T_{pK}S$ is identified with $\mathfrak{p}$, using left multiplication by p^{-1}. Take $X \in \mathfrak{p}$ with norm 1. The geodesic through $pK \in S$ with tangent vector X has equation $t \mapsto pe^{tX}K$. A geodesic is called regular when a nonzero tangent vector at any (and hence every) point is a regular element of $\mathfrak{g}$. It is called singular otherwise. A geodesic is regular if and only if it lies in a unique maximal flat.

2.1.3 Iwasawa decomposition

Let Σ be the set of restricted roots of $\mathfrak{a}$ in $\mathfrak{g}$. By convention, $0 \notin \Sigma$. We denote by $\mathfrak{g}_\alpha$ the root space of a root $\alpha \in \Sigma$, by $m(\alpha) = \dim(\mathfrak{g}_\alpha)$ its multiplicity and by $H_\alpha \in \mathfrak{a}$ the element corresponding to α under the isomorphism $\mathfrak{a} \cong \mathfrak{a}^*$ given by $\langle \cdot, \cdot \rangle$. Fix a set of positive roots Σ^+ with basis Π. Let $\mathfrak{n} \oplus \mathfrak{a} \oplus \mathfrak{k}$ and $N \times A \times K$ be the corresponding Iwasawa decompositions of $\mathfrak{g}$ and G. Define $M = Z_K(A)$ and $M' = N_K(A)$ and denote by $\mathfrak{m}$ the Lie algebra of M.

Denote the Lie algebra Iwasawa projections by $E_\mathfrak{n}$, $E_\mathfrak{a}$ and $E_\mathfrak{k}$. We have the orthogonal restricted root space decomposition

$$\mathfrak{g} = \mathfrak{a} \oplus \mathfrak{m} \oplus \bigoplus_{\alpha \in \Sigma} \mathfrak{g}_\alpha \,. \tag{2.1}$$

Denote the projection onto $\mathfrak{g}_\alpha$ by R_α.

Denote the Iwasawa projections from G onto N, A and K by n, a and κ. Define the height $H(g) = \log(a(g)) \in \mathfrak{a}$, the logarithm being the Lie logarithm on A.

All choices of the data (K, A, N) are conjugate by an element of G. When $G = \mathrm{GL}_n(\mathbb{R})$ or $\mathrm{SL}_n(\mathbb{R})$ we make all the standard choices: $K = \mathrm{O}_n(\mathbb{R})$ respectively $\mathrm{SO}_n(\mathbb{R})$, A is the connected component of the diagonal subgroup and N is the upper triangular unipotent subgroup.

2.1.4 Centralizers

Denote by $\mathcal{L}$ the set of centralizers in G of subgroups of A. They are the standard Levi subgroups of semistandard parabolic subgroups of G. We will denote such a centralizer typically by L. It is again reductive and inherits all the data as in the beginning of §2.1.1 and §2.1.3 from G in the natural way. We allow G to be reductive because we will occasionally need to apply results to Levis $L \in \mathcal{L}$. When $L \in \mathcal{L}$ with Lie algebra $\mathfrak{l}$, define $\mathfrak{a}^L = \mathfrak{l}_{ss} \cap \mathfrak{a}$ and $\mathfrak{a}_L = \mathfrak{z}(\mathfrak{l}) \cap \mathfrak{a}$. Then $\mathfrak{a} = \mathfrak{a}^L \oplus \mathfrak{a}_L$ orthogonally. The set $\mathcal{L}$ contains A and G, and when G is semisimple we have $\mathfrak{a}^A = \mathfrak{a}_G = 0$ and $\mathfrak{a}_A = \mathfrak{a}^G = \mathfrak{a}$.

Define the positive Weyl chamber $\mathfrak{a}^+ = \{H \in \mathfrak{a} : \forall \alpha \in \Sigma^+ : \alpha(H) > 0\}$ and the regular set

$$\mathfrak{a}^{\mathrm{reg}} = \mathfrak{a} - \bigcup_{L \neq M} \mathfrak{a}_L = \mathfrak{a} - \bigcup_{\alpha \in \Sigma} \ker(\alpha)\,.$$

We have that $H \in \mathfrak{a}^{\mathrm{reg}}$ if and only if its centralizer equals the centralizer of $\mathfrak{a}$. Define also the generic set

$$\mathfrak{a}^{\mathrm{gen}} = \mathfrak{a}^{\mathrm{reg}} - \bigcup_{L \in \mathcal{L} - \{G\}} \mathfrak{a}^L\,.$$

Combined superscripts correspond to intersections: $\mathfrak{a}^{\mathrm{gen},+} = \mathfrak{a}^{\mathrm{gen}} \cap \mathfrak{a}^+$.

We also define $(\mathfrak{a}^*)^{\mathrm{reg}}$, $(\mathfrak{a}^*)^{\mathrm{gen}}$ and $(\mathfrak{a}^*)^+$ to be the corresponding subsets under the isomorphism $\mathfrak{a} \cong \mathfrak{a}^*$ defined by the inner product $\langle \cdot, \cdot \rangle$. When $H \in \mathfrak{a}$ corresponds to $\lambda \in \mathfrak{a}^*$ under this isomorphism, then $H \in \mathfrak{a}^{\mathrm{reg}}$ if and only if λ is not orthogonal to any roots, and $H \in \mathfrak{a}^{\mathrm{gen}}$ if and only if λ is in addition not contained in a proper subspace spanned by roots.

We will frequently use the following lemmas, so we take care to properly reference them.

Lemma 2.2. *Let $g \in G$ and $H \in \mathfrak{a}$. If $\mathrm{Ad}_g(H) \in \mathfrak{a}$, then $g \in M' Z_G(H)$.*

Proof. This is stated in [30, §5, Lemma 1]. When $g \in K$, the statement gives precisely the degree of uniqueness in the KAK decomposition of G, and a proof can be found in [41, Lemma 7.38]. The general case can be reduced to $g \in K$ as follows. Write $g = kp$ in the Cartan decomposition. Then $\mathrm{Ad}_p(H) \in \mathrm{Ad}_{k^{-1}}(\mathfrak{a}) \subset \mathfrak{p}$, and [10, §V.24.C, Proposition 1] implies that $p \in Z_G(H)$. Then $\mathrm{Ad}_k(H) \in \mathfrak{a}$, and the conclusion follows from the $g \in K$ case. □

Lemma 2.3. *Let $g \in G$ and $H \in \mathfrak{a}$. If $\mathrm{Ad}_g(H) \in \mathfrak{m} \oplus \mathfrak{a}$, then $g \in M' Z_G(H)$.*

Proof. If we can show that $\mathrm{Ad}_g(H) \in \mathfrak{a}$, the claim follows from Lemma 2.2.

Consider the adjoint embedding $\mathrm{ad} : \mathfrak{g} \to \mathfrak{sl}(\mathfrak{g})$. Equip $\mathfrak{g}$ with any orthonormal basis compatible with the restricted root space decomposition (2.1). In such a basis, ad sends elements of $\mathfrak{a}$ to diagonal matrices and elements of $\mathfrak{k}$ to antisymmetric matrices.

Write $\mathrm{Ad}_g(H) = X + H'$ with $X \in \mathfrak{m}$ and $H' \in \mathfrak{a}$. In the chosen basis, $\mathrm{ad}_{H'}$ is diagonal with real eigenvalues, and the antisymmetric matrix ad_X is diagonalizable with purely imaginary eigenvalues. Because $[H, X] = 0$, the elements $\mathrm{ad}_{H'}$ and ad_X are simultaneously diagonalizable, so that the eigenvalues of $\mathrm{ad}_{H'} + \mathrm{ad}_X$ are those of $\mathrm{ad}_{H'}$ plus those of ad_X, in a suitable ordering. If these eigenvalues are real, it must be that $X = 0$.

This proves that $\mathrm{Ad}_g(H) \in \mathfrak{a}$, and the lemma follows. □

Lemma 2.4. *We have $Z_G(A) = MA$ and $N_G(A) = M'A$.*

Proof. The first statement follows from [41, Proposition 7.25]; the second statement follows by combining it with Lemma 2.2. □

2.1.5 Derivatives

When G is any Lie group with Lie algebra $\mathfrak{g}$ and b is an element of the universal enveloping algebra $U(\mathfrak{g})$, we denote by L_b the corresponding left invariant differential operator on $C^\infty(G)$. When $X \in \mathfrak{g} \subset U(\mathfrak{g})$, by definition

$$(L_X f)(g) = \left.\frac{d}{dt}\right|_{t=0} f(ge^{tX})\,.$$

When $f : M \to N$ is a differentiable map between differentiable manifolds, denote its differential at $m \in M$ by $(Df)_m$. Using left translation we identify all tangent spaces T_gG with $\mathfrak{g}$. When $g \in G$, denote by L_g and R_g the left and right multiplication by g on G. With our convention on tangent spaces, we then have for all $g, h \in G$ that

$$\begin{aligned}(DL_g)_h &= \mathrm{id}\,,\\ (DR_g)_h &= \mathrm{Ad}_{g^{-1}}\ .\end{aligned} \tag{2.5}$$

When $X, Y \in \mathfrak{g}$ we have

$$L_X\,\mathrm{Ad}_g(Y) = \mathrm{Ad}_g([X,Y])\,, \tag{2.6}$$

$$L_X\,\mathrm{Ad}_{g^{-1}}(Y) = -[X, \mathrm{Ad}_{g^{-1}}(Y)]\,. \tag{2.7}$$

Assume now that G is reductive as in the beginning of §2.1.1.

Lemma 2.8. *The differentials of n, a and κ at $g \in G$ are as follows:*

$$\begin{aligned}(Dn)_g &= \mathrm{Ad}_{a(g)} \circ E_{\mathfrak{n}} \circ \mathrm{Ad}_{\kappa(g)}\,,\\ (Da)_g &= E_{\mathfrak{a}} \circ \mathrm{Ad}_{\kappa(g)}\,,\\ (D\kappa)_g &= \mathrm{Ad}_{\kappa(g)^{-1}} \circ E_{\mathfrak{k}} \circ \mathrm{Ad}_{\kappa(g)}\ .\end{aligned}$$

Proof. Write $g = nak$ and take $X \in \mathfrak{g}$. For the N-projection, write

$$\begin{aligned}n(ge^X) &= n \cdot n(ake^X)\\ &= n \cdot n(ake^X k^{-1})\\ &= n \cdot (a \cdot n(e^{\mathrm{Ad}_k(X)}) \cdot a^{-1})\,.\end{aligned}$$

In the last step, we have used that $n(ah) = an(h)a^{-1}$ for $a \in A$ and $h \in G$. Therefore

$$(Dn)_g = (DL_n)_e \circ (D(\mathrm{Ad}_a|_N))_e \circ (Dn)_e \circ \mathrm{Ad}_k\ .$$

The first statement now follows from the fact that $(Dn)_e = E_{\mathfrak{n}}$. The other statements are proved similarly. For the A-projection we write

$$a(ge^X) = a \cdot a(ke^X) = a \cdot a(e^{\mathrm{Ad}_k(X)})$$

and use that $(Da)_e = E_{\mathfrak{a}}$. For the K-projection we write

$$\kappa(ge^X) = \kappa(ke^X) = \kappa(e^{\mathrm{Ad}_k(X)}) \cdot k$$

and use that $(D\kappa)_e = E_{\mathfrak{k}}$. □

The differential of the A-projection is also computed in [46, Lemma 6.1]. Even though the proof there is for $G = \mathrm{SL}_3(\mathbb{R})$, the argument works in general. Compare also [20, Corollary 5.2], but note that the Iwasawa decomposition used there is KAN instead of NAK.

Lemma 2.9. *For $X, Y \in \mathfrak{g}$ and $g \in G$ we have*

$$\begin{aligned}(L_X L_Y a)(g) &= E_{\mathfrak{a}}([E_{\mathfrak{k}}(\mathrm{Ad}_{\kappa(g)}(X)), \mathrm{Ad}_{\kappa(g)}(Y)]) \\ &= E_{\mathfrak{a}}([E_{\mathfrak{k}}(\mathrm{Ad}_{\kappa(g)}(X)), E_{\mathfrak{n}}(\mathrm{Ad}_{\kappa(g)}(Y))]) \\ &= E_{\mathfrak{a}}([\mathrm{Ad}_{\kappa(g)}(X), E_{\mathfrak{n}}(\mathrm{Ad}_{\kappa(g)}(Y))]) \\ &= \sum_{\alpha \in \Sigma^+} \langle \theta R_{-\alpha}(\mathrm{Ad}_{\kappa(g)}(X)), (R_\alpha - \theta R_{-\alpha})(\mathrm{Ad}_{\kappa(g)}(Y)) \rangle_\theta \cdot H_\alpha \,.\end{aligned}$$

Proof. Similar to the proof of [20, Lemma 6.1]. Alternatively, by Lemma 2.8 we find

$$L_Y a(g) = E_{\mathfrak{a}}(\mathrm{Ad}_{\kappa(g)}(Y)) \,.$$

Using the chain rule, (2.6) and Lemma 2.8 to compute $(D\kappa)_g$ we have

$$\begin{aligned}L_X L_Y a(g) &= E_{\mathfrak{a}}(\mathrm{Ad}_{\kappa(g)}([(D\kappa)_g(X), Y])) \\ &= E_{\mathfrak{a}}([E_{\mathfrak{k}}(\mathrm{Ad}_{\kappa(g)}(X)), \mathrm{Ad}_{\kappa(g)}(Y)]) \,.\end{aligned}$$

This is the first equality. The other equalities follow as in [20, Lemma 6.1]. □

2.2 The N-projection and the Gram–Schmidt process

Unless otherwise specified, G is a semisimple Lie group as in §2.1.1. In this section we prove Theorem 5 and the stronger Theorem 2.17.

Recall that the Iwasawa decomposition for $\mathrm{SL}_n(\mathbb{R})$ is nothing but the Gram–Schmidt process: Take $g \in \mathrm{SL}_n(\mathbb{R})$. There is a unique $n \in N$ such that the rows of $n^{-1}g$ are orthogonal for the Euclidean inner product on $\mathbb{R}^n$. There is a unique $a \in A$ such that the rows of $k := a^{-1}n^{-1}g$ have norm 1. The Iwasawa decomposition of g is then nak.

2.2.1 Intro: $\mathrm{SL}_2(\mathbb{R})$

We first prove Theorem 5 when $G = \mathrm{SL}_2(\mathbb{R})$. This is quite trivial but already gives a good idea of what is happening.

Proof of Theorem 5 when $G = \mathrm{SL}_2(\mathbb{R})$. Write $g = \left(\begin{smallmatrix} v \\ w \end{smallmatrix}\right)$, and view $v, w \in \mathbb{R}^2$ as row vectors. Let $y > 0$. Multiplication on the right by $a = \mathrm{diag}(y, y^{-1}) \in A$ corresponds to letting the matrix a act on v and w. The projection $n(ga)$ is the matrix $\left(\begin{smallmatrix} 1 & x \\ 0 & 1 \end{smallmatrix}\right)$ where

$$x = \frac{\langle va, wa \rangle}{\langle wa, wa \rangle}\,.$$

It is the unique matrix $n \in N$ for which the rows of $n^{-1}ga$ are orthogonal. We must show that x is bounded. In the generic case where both coordinates of w in the standard basis are nonzero, the denominator $\langle wa, wa \rangle$ is $\asymp \max(y^2, y^{-2})$, because a acts by y in one coordinate and by y^{-1} in the other. The numerator is always $\ll \max(y^2, y^{-2})$, for the same reason. Therefore in this generic case, x is bounded and the claim follows. When a coordinate of w vanishes, the corresponding term in the denominator $\langle wa, wa \rangle$ is nonzero, but the same is then true in the numerator. So the same bounds hold, with $\max(y^2, y^{-2})$ replaced by y^2 or y^{-2}, depending on which coordinate of w vanishes. The conclusion follows. □

2.2.2 $\mathrm{SL}_n(\mathbb{R})$, exterior powers

For $\mathrm{SL}_n(\mathbb{R})$ the same phenomenon occurs, where if a term in a denominator of a certain fraction vanishes, then so does the corresponding term in the numerator. The denominators will here be norms of orthogonal projections onto the orthocomplement of previous vectors, and they are most conveniently expressed using norms on exterior powers.

When (V, b) is a bilinear space of finite dimension over a field, b can be identified with a linear map $V \to V^*$. It induces a linear map $\bigwedge^k V \to \bigwedge^k V^* \cong (\bigwedge^k V)^*$ for all $k \geq 0$, where the identification $\bigwedge^k V^* \cong (\bigwedge^k V)^*$ comes from the natural pairing between $\bigwedge^k V$ and $\bigwedge^k V^*$. We get a natural bilinear form $\bigwedge^k b$ on $\bigwedge^k V$. Assume now that the field is $\mathbb{R}$, and that b is symmetric positive definite. Then so is $\bigwedge^k b$. We denote all induced bilinear forms by $\langle \cdot, \cdot \rangle$ for convenience. If (e_i) is an orthonormal basis of V, then an induced basis of $\bigwedge^k V$ consisting of elements of the form $e_{j_1} \wedge \cdots \wedge e_{j_k}$ is also orthonormal. As a consequence, if $v, w_1, \ldots, w_k \in V$ with the w_i linearly independent, $W = \mathrm{span}(w_i)$ and $\mathrm{pr}_{W^\perp}$ denotes the orthogonal projection onto $W^\perp$, then

$$\|\mathrm{pr}_{W^\perp}(v)\| = \frac{\left\| v \wedge \left(\bigwedge_{i=1}^k w_i \right) \right\|}{\left\| \bigwedge_{i=1}^k w_i \right\|}\,, \tag{2.10}$$

where the norms are those induced on V, $\bigwedge^{k+1} V$ and $\bigwedge^k V$. This identity may be visualized as follows: The numerator in the right-hand side is the $(k+1)$-volume of the

parallelepiped spanned by the vectors v and w_i, the denominator is the k-volume of the face spanned by the w_i, and left-hand side is the height of the parallelepiped with respect to this face.

More generally, for $v_1, v_2 \in V$ we have

$$\langle \mathrm{pr}_{W^\perp}(v_1), \mathrm{pr}_{W^\perp}(v_2) \rangle = \frac{\left\langle v_1 \wedge \left(\bigwedge_{i=1}^k w_i \right), v_2 \wedge \left(\bigwedge_{i=1}^k w_i \right) \right\rangle}{\left\| \bigwedge_{i=1}^k w_i \right\|^2} . \tag{2.11}$$

We also have the ith coefficient of $\mathrm{pr}_W(v)$ in the basis (w_i) is given by

$$\frac{\left\langle w_1 \wedge \cdots \wedge \widehat{w_i} \wedge v \wedge \cdots \wedge w_k, \bigwedge_{i=1}^k w_i \right\rangle}{\left\| \bigwedge_{i=1}^k w_i \right\|^2} , \tag{2.12}$$

where the hat denotes omission.

Proof of Theorem 5 when $G = \mathrm{SL}_n(\mathbb{R})$. Write g as a column of row vectors:

$$g = (v_1, \ldots, v_n)^T ,$$

with the $v_i \in \mathbb{R}^n$, and view $a \in A$ as acting on $\mathbb{R}^n$ diagonally. Define the subspaces $V_i = \mathrm{span}(v_i, \ldots, v_n)$, so that $V_{n+1} = \{0\}$. Define w_i^a to be the ith row of $n(ga)^{-1}ga$. Equip $\mathbb{R}^n$ with the Euclidean inner product. The first step in the Gram–Schmidt process applied to ga is the recurrence relation

$$\begin{aligned} w_n^a &= v_n a , \\ w_{i-1}^a &= \mathrm{pr}_{(V_i a)^\perp}(v_{i-1}a) \\ &= v_{i-1}a - \sum_{j=i}^n \frac{\langle v_{i-1}a, w_j^a \rangle}{\langle w_j^a, w_j^a \rangle} w_j^a \qquad (i = n, \ldots, 2) . \end{aligned}$$

In the above sum, the coefficient in the term with index j is the $(i-1, j)$th entry of $n(ga)$. We must show that those coefficients are bounded independently of a. For such a coefficient, with $i \le j$, we have

$$\begin{aligned} \frac{\langle v_{i-1}a, w_j^a \rangle}{\langle w_j^a, w_j^a \rangle} &= \frac{\left\langle \mathrm{pr}_{(V_{j+1}a)^\perp}(v_{i-1}a), \mathrm{pr}_{(V_{j+1}a)^\perp}(v_j a) \right\rangle}{\left\| \mathrm{pr}_{(V_{j+1}a)^\perp}(v_j a) \right\|^2} \\ &= \frac{\left\langle v_{i-1}a \wedge \left(\bigwedge_{k=j+1}^n v_k a \right), \bigwedge_{k=j}^n v_k a \right\rangle}{\left\| \bigwedge_{k=j}^n v_k a \right\|^2} , \end{aligned}$$

where we have used (2.11). The action of A on $\bigwedge^{n-j+1}\mathbb{R}^n$ is still diagonal in an induced basis. Fix such a basis and write $a = \mathrm{diag}(a_1, \dots, a_K)$ in that basis. If $c_1, \dots, c_K$ denote the coordinates of $v_{i-1} \wedge \left(\bigwedge_{k=j+1}^{n} v_k\right)$ and $d_1, \dots, d_K$ those of $\bigwedge_{k=j}^{n} v_k$, then by the triangle inequality

$$\begin{aligned} \left| \frac{\langle v_{i-1}a, w_j^a \rangle}{\langle w_j^a, w_j^a \rangle} \right| &= \left| \frac{\sum_{k=1}^K a_k^2 c_k d_k}{\sum_{k=1}^K a_k^2 d_k^2} \right| \\ &\leq \max_{d_k \neq 0} \left| \frac{c_k d_k}{d_k^2} \right| . \end{aligned} \tag{2.13}$$

This bound does not depend on a, so that the entries of $n(ga)$ are bounded. □

2.2.3 Semisimple groups

Now let G be a semisimple Lie group. There is no real reason to assume that G has finite center, and Theorem 5 is insensitive to central extensions in any case, but we do so because our setup in §2.1.1 is under this assumption. Let $\rho : G \to \mathrm{GL}(V)$ be any finite-dimensional representation with discrete kernel, such as the adjoint representation $\mathrm{Ad} : G \to \mathrm{GL}(\mathfrak{g})$, or the standard representation Std if G is already linear. The fact that ρ can be arbitrary is not essential, but it leads to questions about uniformity.

Remark 2.14. Necessarily $\rho(G) \subset \mathrm{SL}(V)$, because $\mathfrak{g} = [\mathfrak{g}, \mathfrak{g}]$ consists of commutators. Note that $\rho(G)$ is automatically closed, by [41, Proposition 7.9] or alternatively by properties of Malcev closure [55, §1.4.2, Theorem 3].

Lemma 2.15. *In a suitable basis of V, the groups $\rho(N)$, $\rho(A)$ and $\rho(K)$ are contained in the standard Iwasawa components of* $\mathrm{GL}(V)$.

Proof. As in the proof of [41, Proposition 7.9], there is a basis of V such that $\rho(K) \subset \mathrm{SO}_n(\mathbb{R})$ and $\rho(P)$ consists of symmetric matrices. Because $A \subset P$ is commutative, by acting on the basis by a suitable element of $\mathrm{SO}_n(\mathbb{R})$ we may assume that $\rho(A)$ is diagonal. Then the basis consists of restricted weight vectors. Sorting them by non-increasing weight ensures that $\rho(N)$ is upper triangular unipotent. □

Equip V with an Iwasawa-compatible basis given by Lemma 2.15 and denote the corresponding Iwasawa decomposition of $\mathrm{SL}(V)$ by $N'A'K'$. Denote the N'-projection on $\mathrm{SL}(V)$ by n'.

Proof of Theorem 5. By the case $G = \mathrm{SL}_n(\mathbb{R})$ proved in §2.2.2, we know that the projection $n'(\rho(gA)) \subset n'(\rho(g)A')$ is relatively compact in N'. It is contained in the closed set $\rho(N)$, and therefore relatively compact in $\rho(N)$. Because ρ has central kernel contained in K, it induces an isomorphism $N \to \rho(N)$. Therefore $n(gA) \cong n'(\rho(gA))$ is relatively compact in N. □

2.2.4 Uniformity

For $G = \mathrm{SL}_2(\mathbb{R})$ it is apparent from the proof in §2.2.1 that uniformity holds in Theorem 5 when the entries of g are bounded and both entries on the second row of $g \in G$ are bounded away from 0. We generalize this and partition any semisimple group G into subsets for which uniformity holds on compact subsets. We briefly describe one such partition here and state a result about a more optimal partition in §2.2.5.

We use the notation from §2.2.3. Let $d = \dim(V)$ and let $\langle \cdot, \cdot \rangle$ be the Euclidean inner product for the chosen basis of V. For $\lambda \in \mathfrak{a}^*$ denote by V_λ the restricted weight space. To prove Theorem 5, instead of reducing to the case of $\mathrm{SL}_n(\mathbb{R})$ as in §2.2.3, one can also directly use the argument in §2.2.2, but restricted to elements $g \in \rho(G)$, and this leads to a statement with uniformity. For every tuple $\mathcal{S} = (\mathcal{S}_1, \ldots, \mathcal{S}_d)$ ($\mathcal{S}$ for support) of sets of integral weights of $\mathfrak{g}$, we may consider the subset $\Omega_\mathcal{S}$ of G that consists of the elements g with the following property: For every $j \in \{1, \ldots, d\}$, the wedge product of every last j rows of $\rho(g)$, as an element of $\bigwedge^j V$, has a nonzero component precisely along the weight spaces $(\bigwedge^j V)_\lambda$ with $\lambda \in \mathcal{S}_j$. It is then apparent that the upper bound analogous to (2.13) is uniform for g in compact subsets of $\Omega_\mathcal{S}$, because the coefficients d_k (which are now norms of weight space projections) are either zero or bounded away from zero.

2.2.5 Canonical partitions

The sets $\Omega_\mathcal{S}$ in §2.2.4 depend on the choice of basis of weight vectors of V. Moreover, they may form a partition of G that is unnecessarily fine for the uniformity statement to be true; this can happen when the weight spaces of V are not 1-dimensional. It is desirable to construct a basis-independent partition, which is what we do now, and which we expect to be the coarsest possible.

We continue to use the notation from §2.2.3 and consider the right action of $\mathrm{GL}(V)$ on V by transposition: $vg := g^T v$. That is, in the chosen basis of V, the rows of $g \in \mathrm{GL}(V)$ are the images of the basis elements under the right action of g. (This awkward definition is an artifact of our decision to write Iwasawa decompositions as NAK rather than KAN.) Note that for $g \in G$ we have $\rho(g)^T = \rho(\theta(g))^{-1}$, because $\rho(A)$ consists of diagonal matrices and $\rho(K) \subset \mathrm{SO}_n(\mathbb{R})$. That is, while the right action of $\mathrm{GL}(V)$ by transposition depends on the choice of basis, the restriction to G depends only on the choice of Iwasawa decomposition of G (and in fact, only on A and K).

We will prove Theorem 2.17 below in a way similar to the argument sketched in §2.2.4, but using a block-by-block rather than a row-by-row orthogonalization process, where we transform a matrix so that a block of rows (corresponding to a weight space) is orthogonal to previous blocks of rows. That the Iwasawa decomposition corresponds to such a process, is the content of the following lemma.

For integral weights λ, μ of $\mathfrak{g}$, we write $\mu < \lambda$ if $\lambda - \mu$ is nonzero and is a nonnegative integer linear combination of positive roots.

Lemma 2.16. *Let $g \in G$ and $n' = n'(\rho(g))$. Then for distinct weights λ, μ of V we have $V_\lambda n'^{-1}\rho(g) \perp V_\mu n'^{-1}\rho(g)$ and $V_\lambda(n'^{-1} - 1) \perp V_\lambda$.*

Proof. We have that $n'^{-1}\rho(g) \in A'K'$. The first statement follows from the facts that $V_\lambda A' = V_\lambda$, that $V_\lambda \perp V_\mu$ and that K' preserves orthogonality. For the second statement, we have for $X \in \mathrm{Lie}(N')$ that $X^T V_\lambda \in \bigoplus_{\mu<\lambda} V_\mu$, and exponentiating gives that $V_\lambda(N' - 1) \in \bigoplus_{\mu<\lambda} V_\mu$. □

Let Λ be the sets of weights of $\mathfrak{a}$ in V, denote by m_λ the multiplicity of $\lambda \in \Lambda$, and define $s_\lambda = \sum_{\mu<\lambda} m_\mu$. For any tuple $\mathcal{S} = (\mathcal{S}_\lambda)_{\lambda\in\Lambda}$ of sets of integral weights of $\mathfrak{g}$, consider the subset $\Omega_\mathcal{S}$ of G that consists of the elements g with the following property: For every $\lambda \in \Lambda$, the line $\bigwedge^{s_\lambda} \bigoplus_{\mu<\lambda} V_\mu\rho(g) \subset \bigwedge^{s_\lambda} V$ has nonzero orthogonal projection on the weight-α spaces $(\bigwedge^{s_\lambda} V)_\alpha$ for $\alpha \in S_\lambda$, and zero projection when $\alpha \notin \mathcal{S}_\lambda$.

Theorem 2.17. *Let $\mathcal{S}$ be any tuple as above, and $D \subset \Omega_\mathcal{S}$ compact. Then $n(DA)$ is relatively compact.*

Before proving Theorem 2.17, we state some basic properties of the sets $\Omega_\mathcal{S}$.

Proposition 2.18. *The sets of the form $\Omega_\mathcal{S}$ (or those that are non-empty) partition G. They are stable on the left by NAM.*

Proof. That the $\Omega_\mathcal{S}$ partition G is clear from their definition. Acting on the left by $\rho(NAM)$ on $g \in \mathrm{SL}(V)$ does not change the lines $\bigwedge^{s_\lambda} \bigoplus_{\mu<\lambda} V_\mu g$, so that the individual conditions defining $\Omega_\mathcal{S}$ are left-invariant under NAM. □

For $\lambda \in \Lambda$, define $\mathcal{S}^0_\lambda$ to be the set of weights α of $\mathfrak{g}$ such that the line

$$\bigwedge^{s_\lambda} \bigoplus_{\mu<\lambda} V_\mu\rho(g) \subset \bigwedge^{s_\lambda} V$$

has nonzero projection onto the weight-α subspace for at least one $g \in G$. Define $\mathcal{S}^0 = (\mathcal{S}^0_\lambda)_{\lambda\in\Lambda}$. Then $\Omega_{\mathcal{S}^0}$ contains "most" elements of G.

Proposition 2.19. *The set $\Omega_{\mathcal{S}^0}$ is open and dense in G.*

Proof. The image $\rho(G)$ is the identity component of the real points of a real algebraic closed subgroup of $\mathrm{SL}(V)$ [55, §3.3.3]. The set $\rho(\Omega_{\mathcal{S}^0})$ is defined by the non-vanishing of finitely many rational functions (corresponding to projections onto weight spaces) and therefore Zariski open in $\rho(G)$. (There are no vanishing conditions when $\mathcal{S} = \mathcal{S}^0$, or rather, they are satisfied on all of $\rho(G)$.) It follows that $\Omega_{\mathcal{S}^0}$ is open and dense in G. □

Example 2.20. Take $G = \mathrm{SL}_2(\mathbb{R})$ and $\rho = \mathrm{Std}$ the standard representation. Then $\Omega_{\mathcal{S}^0} = G - NAM'$, because the only defining condition is that both entries on the bottom row are nonzero. The only other non-empty sets of the form $\Omega_{\mathcal{S}}$ are NAM and $NAM \cdot \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$. The geodesics in G/K corresponding to elements of $\Omega_{\mathcal{S}^0}$ are semicircles. The other $\Omega_{\mathcal{S}}$ give rise to the vertical geodesics, with both orientations. One can check using an explicit computation that $\rho = \mathrm{Ad}$ gives the same partition of $\mathrm{SL}_2(\mathbb{R})$.

Example 2.21. Take $G = \mathrm{SL}_3(\mathbb{R})$ and $\rho = \mathrm{Std}$. The sets $\Omega_{\mathcal{S}}$ are determined by their K-projection, in view of Lemma 2.18. The sets $\kappa(\Omega_S)$ come in all possible dimensions: The 0-dimensional ones, which are the six right cosets of M in M'. The 1-dimensional ones: the three right M'-translates of $G \cap (\mathrm{O}(2) \times \mathrm{O}(1)) - M'$ and the three right M'-translates of $G \cap (\mathrm{O}(1) \times \mathrm{O}(2)) - M'$. The 2-dimensional ones: the three right M'-translates of the product $(G \cap (\mathrm{O}(2) \times \mathrm{O}(1)) - M')(G \cap (\mathrm{O}(1) \times \mathrm{O}(2)) - M')$ and the three right M'-translates of that product with the two factors interchanged. And finally, the dense open set $\kappa(\Omega_{S^0})$.

Remark 2.22. For any G, the set $\kappa(\Omega_{S^0})$ is disjoint from the sets LM' with $L \in \mathcal{L}$ a standard Levi subgroup of a semistandard parabolic; this follows already from the constraint corresponding to the minimal weights λ. But in general it is smaller than just the complement of the sets of the form $K \cap LM'$, as illustrated in Example 2.21.

It remains unclear to us whether the partition of G into the $\Omega_{\mathcal{S}}$ depends on the choice of representation ρ, and if not, whether the partition is the coarsest possible (up to splitting into connected components) for uniformity to hold. It would also be desirable to have a concrete description of the partition in general, as we do have when $G = \mathrm{SL}_3(\mathbb{R})$.

Proof of Theorem 2.17. Let $g \in G$ and $a \in A$. We view $n'(\rho(ga))^{-1}$ as a block matrix: For weights $\lambda \geq \mu$ of V, define

$$\begin{aligned} T_{\lambda,\mu} : V_\lambda &\to V_\mu \\ v &\mapsto \mathrm{pr}_{V_\mu}\left(v \cdot n'(\rho(ga))^{-1}\right) \end{aligned}$$

and define $T_\lambda = \sum_{\mu<\lambda} T_{\lambda,\mu} : V_\lambda \to \bigoplus_{\mu<\lambda} V_\mu$. We must show that each T_λ is a bounded operator, uniformly in $a \in A$ and g in compact subsets of each $\Omega_{\mathcal{S}}$. By the first statement in Lemma 2.16 and the fact that the right action of N does not increase weights, the map T_λ satisfies

$$((1 + T_\lambda)V_\lambda)\rho(ga) \perp \bigoplus_{\mu<\lambda} V_\mu \rho(ga)\,.$$

That is, $(T_\lambda v)\rho(ga) = -\,\mathrm{pr}_{\bigoplus_{\mu<\lambda} V_\mu \rho(ga)}(v\rho(ga))$. (These two identities express the block-by-block orthogonalization process.) To show that T_λ is bounded, we may fix any basis $(b_i)_{1\leq i\leq s_\lambda}$ of $\bigoplus_{\mu<\lambda} V_\mu$ and we must show that for every $v \in V_\lambda$ the coordinates of the orthogonal projection of $v\rho(ga)$ onto $\bigoplus_{\mu<\lambda} V_\mu\rho(ga)$ in the basis $(b_i\rho(ga))$ are bounded.

By (2.12), the ith coordinate is equal to

$$\frac{\langle (b_1 \wedge \cdots \wedge \widehat{b_i} \wedge v \wedge \cdots \wedge b_{s_\lambda})\rho(ga), (b_1 \wedge \cdots \wedge b_{s_\lambda})\rho(ga)\rangle}{\|(b_1 \wedge \cdots \wedge b_{s_\lambda})\rho(ga)\|^2},$$

where the hat denotes omission. We may bound this using an inequality similar to (2.13). Concretely, call d_μ the projection of $(b_1 \wedge \cdots \wedge b_{s_\lambda})\rho(g)$ onto the weight-μ subspace of $\bigwedge^{s_\lambda} V$, and c_μ that of $(b_1 \wedge \cdots \wedge \widehat{b_i} \wedge v \wedge \cdots \wedge b_{s_\lambda})\rho(g)$. Then the fraction above equals

$$\frac{\sum_\mu \mu(a)^2 \langle c_\mu, d_\mu\rangle}{\sum_\mu \mu(a)^2 \|d_\mu\|^2},$$

where we define the weights on A using the exponential map. By the triangle inequality, this is at most

$$\max_{d_\mu \neq 0} \frac{|\langle c_\mu, d_\mu\rangle|}{\|d_\mu\|^2},$$

which gives a uniform upper bound for g in compact subsets of each $\Omega_{\mathcal{S}}$, because the d_μ are then either zero or are bounded away from zero. □

2.3 The A-projection and extreme points

In this section we prove Theorem 6. Unless otherwise stated, G is a semisimple Lie group as in §2.1.1. For $H_0 \in \mathfrak{a}$ and $g \in G$ define

$$\begin{aligned} h_{H_0,g} : A &\to \mathbb{R} \\ a &\mapsto \langle H_0, H(ga)\rangle . \end{aligned}$$

When H_0 corresponds to $\lambda \in \mathfrak{a}^*$ under the isomorphism given by the Killing form, this is precisely the function $h_{\lambda,g}$ in Theorem 6. Recall from §2.1.4 that λ is regular if and only if H_0 is, λ lies in the positive chamber of $\mathfrak{a}^*$ if and only if $H_0 \in \mathfrak{a}^+$, and λ does not lie in a proper subspace spanned by roots if and only if $H_0 \notin \bigcup_{L \in \mathcal{L}-\{G\}} \mathfrak{a}^L$. Thus the condition on λ in Theorem 6 is equivalent to $H_0 \in \mathfrak{a}^{\mathrm{gen},+}$.

The first part of Theorem 6, concerning uniqueness and nondegeneracy, is proved in §2.3.3; see Proposition 2.48. The second part of Theorem 6, concerning existence, is proved in §2.3.4 (although the hard work is done in §2.3.2); see Corollary 2.52 and Proposition 2.53.

2.3.1 Critical points and level sets

Using Lemma 2.8 we find that the differential of $h_{H_0,g}$ at $a \in A$ is

$$\begin{aligned} (Dh_{H_0,g})_a : \mathfrak{a} &\to \mathbb{R} \\ H &\mapsto \langle H_0, \mathrm{Ad}_{\kappa(ga)}(H)\rangle . \end{aligned} \tag{2.23}$$

When G is any reductive group with all associated data as in §2.1.1, and $H_0 \in \mathfrak{a}$, define the set $\mathcal{C}(G, H_0)$ as follows:

$$\mathcal{C}(G, H_0) = \{k \in K : \mathrm{Ad}_{k^{-1}}(H_0) \perp \mathfrak{a}\}\,. \tag{2.24}$$

We allow G to be reductive because we will occasionally need to work with the sets $\mathcal{C}(L, H_0)$ for $L \in \mathcal{L}$.

Now let G again be semisimple. The following two lemmas are clear from (2.23) and (2.24).

Lemma 2.25. *Let $H_0 \in \mathfrak{a}$ and $g \in G$. Then $a \in A$ is a critical point of $h_{H_0,g}$ if and only if $\kappa(ga) \in \mathcal{C}(G, H_0)$.* □

Lemma 2.26. *The set of $k \in K$ for which 1 is a critical point of $h_{H_0,k}$ equals $\mathcal{C}(G, H_0)$. More generally, the set of $k \in K$ for which $a \in A$ is a critical point of $h_{H_0,k}$ equals $\kappa(\mathcal{C}(G, H_0)a^{-1})$.* □

In view of Lemma 2.26, we refer to the sets $\kappa(\mathcal{C}(G, H_0)a^{-1})$ as level sets. This will be fully justified after we show uniqueness of critical points in Proposition 2.48.

By decomposing g in the Iwasawa decomposition we see that

$$h_{H_0,g}(a) = \langle H_0, H(g)\rangle + h_{H_0,\kappa(g)}(a)\,. \tag{2.27}$$

The first term in the right hand side being a mere constant, we see that the critical points of $h_{H_0,g}$ coincide with those of $h_{H_0,\kappa(g)}$. It therefore suffices to understand the critical points of $h_{H_0,k}$ for $k \in K$.

Example 2.28. When $G = \mathrm{PSL}_2(\mathbb{R})$ it is clear from the geometric picture in the introduction that a critical point of $h_{H_0,g}$ is a point $a \in A$ such that the tangent line to gAi at gai is horizontal. This is the content of the above Lemma 2.25 in this case. The set $\mathcal{C}(G, H_0)$ consists of the two elements

$$c_\pm = \begin{pmatrix} \cos(\pi/4) & \pm\sin(\pi/4) \\ \mp\sin(\pi/4) & \cos(\pi/4) \end{pmatrix}$$

corresponding to the two horizontal directions. See Figure 2.1 for a picture.

Example 2.29. Let $G = \mathrm{SL}_3(\mathbb{R})$. Take elements $H_0, H_0' \in \mathfrak{a}$ that are not proportional, and take $k \in K$. We claim that $a \in A$ can never be a critical point of both $h_{H_0,k}$ and $h_{H_0',k}$. Indeed, replacing k by $\kappa(ka)$ we may assume that $a = 1$, and that $k \in \mathcal{C}(G, H_0) \cap \mathcal{C}(G, H_0')$. That is, the symmetric matrices $\mathrm{Ad}_{k^{-1}}(H_0)$ and $\mathrm{Ad}_{k^{-1}}(H_0')$ have zeroes on the diagonal. Because $\mathfrak{a}$ is 2-dimensional the same is then true for any $H_0'' \in \mathfrak{a}$. Take now $H_0'' = H_0^2$. The matrix $\mathrm{Ad}_{k^{-1}}(H_0'') = \mathrm{Ad}_{k^{-1}}(H_0)\,\mathrm{Ad}_{k^{-1}}(H_0)^T$ has zeroes on the diagonal on the one hand, but its trace is the sum of squares of the entries of H_0 on the other hand. This is a contradiction.

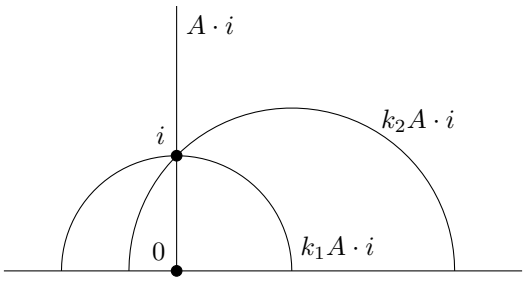

Figure 2.1: Geodesics corresponding to an element $k_1 \in \mathcal{C}(G, H_0)$, and to an element k_2 that does not lie in $\mathcal{C}(G, H_0)$.

2.3.2 Existence of critical points

We want to show that there exists $k \in K$ with the property that $h_{H_0,k}$ has a critical point. Equivalently, in view of Lemma 2.26, that the set $\mathcal{C}(G, H_0)$ is nonempty. Before proving that this is indeed the case for generic $H_0 \in \mathfrak{a}$, we need a negative result.

When $k \in K$ centralizes a nonzero subspace $V \subset \mathfrak{a}$, then $a \mapsto H(ka)$ grows linearly in the directions of V. In particular, for $h_{H_0,k}$ to have a critical point, H_0 must be orthogonal to V. Those critical points behave badly, and this is the reason to impose that $H_0 \notin \bigcup_{L \in \mathcal{L}-\{G\}} \mathfrak{a}^L$ in Theorem 6. We make this more precise in Lemma 2.31.

Lemma 2.30. *When $L \in \mathcal{L}$ and $m \in M'$, we have $\kappa(mL) \subset mL$.*

Proof. It follows from [41, Proposition 7.25, Proposition 7.31] that when $L' \subset G$ is a semistandard Levi subgroup, one has $\kappa(L') \subset L'$. We may apply this to $L' = mLm^{-1}$. □

Proposition 2.31. *When $H_0 \notin \bigcup_{L \in \mathcal{L}-\{G\}} \mathfrak{a}^L$ and $k \in \bigcup_{L \in \mathcal{L}-\{G\}} M'L$, the function $h_{H_0,k}$ has no critical point.*

Proof. Assume $a \in A$ is a critical point of $h_{H_0,k}$, and let $m \in M'$ and $L \in \mathcal{L} - \{G\}$ be such that $k \in mL$. By Lemma 2.30 we also have $\kappa(ka) \in mL$. From (2.23), letting H vary in $\mathfrak{a}_L$, we see that $H_0 \perp \mathrm{Ad}_m(\mathfrak{a}_L) = \mathfrak{a}_{\mathrm{Ad}_m(L)}$. That is, $H_0 \in \mathfrak{a}^{\mathrm{Ad}_m(L)}$, which contradicts our assumption. □

An equivalent formulation of Proposition 2.31 is the following.

Lemma 2.32. *When $H_0 \notin \bigcup_{L \in \mathcal{L}-\{G\}} \mathfrak{a}^L$, the set $\mathcal{C}(G, H_0)$ does not meet $\bigcup_{L \in \mathcal{L}-\{G\}} M'L$.*

Proof. The elements $k \in \mathcal{C}(G, H_0)$ have the property that $h_{H_0,k}$ has a critical point, namely 1. □

When G is reductive, the set $\mathcal{C}(G, H_0)$ is trivially empty when H_0 has a component along $Z(\mathfrak{g})$. That is, when $H_0 \notin \mathfrak{g}_{ss}$. In particular, when L is a semistandard Levi subgroup of G, the set $\mathcal{C}(L, H_0)$ is empty if $H_0 \notin (\mathfrak{a} \cap \mathfrak{l}_{ss}) = \mathfrak{a}^L$. To study the sets $\mathcal{C}(G, H_0)$ with G reductive and $H_0 \in \mathfrak{a} \cap \mathfrak{g}_{ss}$, consider the map

$$\begin{aligned} f_{G,H_0} : K &\to \mathfrak{a} \cap \mathfrak{g}_{ss} \\ k &\mapsto E_{\mathfrak{a}}(\mathrm{Ad}_{k^{-1}}(H_0)) \,. \end{aligned} \tag{2.33}$$

By definition, $\mathcal{C}(G, H_0) = f_{G,H_0}^{-1}(0)$. By (2.7), the differential of f_{G,H_0} at $k \in K$ is given by

$$\begin{aligned} (Df_{G,H_0})_k : \mathfrak{k} &\to \mathfrak{a} \cap \mathfrak{g}_{ss} \\ X &\mapsto -E_{\mathfrak{a}}([X, \mathrm{Ad}_{k^{-1}}(H_0)]) \,. \end{aligned} \tag{2.34}$$

For G semisimple, we want to prove that the sets $\mathcal{C}(G, H_0)$ are generically nonempty. To this end, define

$$\begin{aligned} g_{G,H_0} : K &\to \mathbb{R}_{\geq 0} \\ k &\mapsto \|f_{G,H_0}(k)\|^2 \,. \end{aligned}$$

We have $\mathcal{C}(G, H_0) = g_{G,H_0}^{-1}(0)$. Because g_{G,H_0} is a continuous function on a compact set, it reaches a minimum. Our aim is to show that the minima of g_{G,H_0} satisfy $g_{G,H_0}(k) = 0$, it it will follow that $\mathcal{C}(G, H_0)$ is nonempty.

Using Leibniz's rule and (2.34), the differential of g_{G,H_0} at $k \in K$ is given by

$$\begin{aligned} (Dg_{G,H_0})_k(X) &= 2\langle (Df_{G,H_0})_k(X), f_{G,H_0}(k) \rangle \\ &= -2\langle [X, \mathrm{Ad}_{k^{-1}}(H_0)], f_{G,H_0}(k) \rangle \,. \end{aligned} \tag{2.35}$$

Lemma 2.36. *Let G be reductive and $H_0 \in \mathfrak{a}^{\mathrm{reg}} \cap \mathfrak{g}_{ss}$. Then f_{G,H_0} is a submersion at points $k \notin \bigcup_{L \in \mathcal{L} - \{G\}} M'L$.*

Proof. Suppose $(Df_{G,H_0})_k$ is not surjective for some $k \in K$. Then there exists a nonzero $H \in \mathfrak{a} \cap \mathfrak{g}_{ss}$ such that $\langle [X, \mathrm{Ad}_{k^{-1}}(H_0)], H \rangle = 0$ for all $X \in \mathfrak{k}$. Using associativity of the Killing form, this is equivalent to

$$[\mathrm{Ad}_{k^{-1}}(H_0), H] \perp \mathfrak{k} \,.$$

But $[\mathrm{Ad}_{k^{-1}}(H_0), H] \in [\mathfrak{p}, \mathfrak{p}] \subset \mathfrak{k}$, so that $[\mathrm{Ad}_{k^{-1}}(H_0), H] = 0$. That is, $\mathrm{Ad}_k(H) \in Z_{\mathfrak{p}}(H_0)$. Because H_0 is regular, this implies $\mathrm{Ad}_k(H) \in \mathfrak{a}$ (see [41, Lemma 6.50]), and Lemma 2.2 then implies that $k \in M'L$ with $L = Z_G(H)$. Because $H \notin Z(\mathfrak{g})$, this is a proper semistandard Levi subgroup. This proves the statement. □

Corollary 2.37. *Let G be reductive and $H_0 \in \mathfrak{a}^{\mathrm{gen}} \cap \mathfrak{g}_{ss}$. Then f_{G,H_0} is a submersion at the points of $\mathcal{C}(G, H_0)$.*

Proof. The set $\mathcal{C}(G, H_0)$ does not meet any of the sets $M'L$ by Lemma 2.32, and therefore f_{G,H_0} is a submersion at points of $\mathcal{C}(G, H_0)$, by Lemma 2.36. □

Lemma 2.38. *Let $H_0 \in \mathfrak{a}^{\mathrm{reg}}$. Define $\mathcal{D}_{H_0} = \{k \in K : k \in Z_G(f_{G,H_0}(k))\}$. The following hold:*

(i) The function g_{G,H_0} is right invariant under M'.

(ii) The set of critical points of g_{G,H_0} is $\mathcal{D}_{H_0}M'$.

(iii) Let $k \in \mathcal{D}_{H_0}$ and define $L = Z_G(f_{G,H_0}(k))$. Write $H_0 = H_L + H^L$ with $H_L \in \mathfrak{a}_L$ and $H^L \in \mathfrak{a}^L$. Then $f_{G,H_0}(k) = H_L$ and $k \in \mathcal{C}(L, H^L)$.

Proof. (i) We show that $f_{G,H_0}(km) = \mathrm{Ad}_{m^{-1}}(f_{G,H_0}(k))$ for $m \in M'$. Because the adjoint action of M' on $\mathfrak{a}$ is isometric, it will then follow that g_{G,H_0} is right invariant under M'.

Because M' normalizes $\mathfrak{a}$, it normalizes the orthogonal complement $\mathfrak{a}^\perp$, so that the adjoint action of M' commutes with $E_{\mathfrak{a}}$. This implies

$$\begin{aligned} f_{G,H_0}(km) &= E_{\mathfrak{a}}(\mathrm{Ad}_{m^{-1}k^{-1}}(H_0)) \\ &= \mathrm{Ad}_{m^{-1}}(E_{\mathfrak{a}}(\mathrm{Ad}_{k^{-1}}(H_0))) \\ &= \mathrm{Ad}_{m^{-1}}(f_{G,H_0}(k))\,. \end{aligned}$$

(ii) Let k be a critical point of g. Using (2.35) we have for all $X \in \mathfrak{k}$,

$$0 = (Dg_{G,H_0})_k(X) = -2\langle E_{\mathfrak{a}}([X, \mathrm{Ad}_{k^{-1}}(H_0)]), f_{G,H_0}(k)\rangle\,.$$

That is,

$$\langle \mathfrak{k}, [\mathrm{Ad}_{k^{-1}}(H_0), f_{G,H_0}(k)]\rangle = 0\,.$$

But $[\mathrm{Ad}_{k^{-1}}(H_0), f_{G,H_0}(k)] \in [\mathfrak{p}, \mathfrak{p}] = \mathfrak{k}$, so that this must be zero. Because $H_0 \in \mathfrak{a}^{\mathrm{reg}}$, this implies that $f_{G,H_0}(k) \in \mathrm{Ad}_{k^{-1}}(\mathfrak{a})$. By Lemma 2.2 it follows that $k \in M'Z_G(f_{G,H_0}(k))$. Replacing k by an appropriate right translate under M', this becomes $k \in Z_G(f_{G,H_0}(k))$. That is, $k \in \mathcal{D}_{H_0}$.

(iii) Take $k \in \mathcal{D}_{H_0}$. By definition of L we have $f_{G,H_0}(k) \in \mathfrak{a}_L$. Writing $H_0 = H_L + H^L$ and using that $k \in L$, we have $f_{G,H_0}(k) = E_{\mathfrak{a}}(\mathrm{Ad}_{k^{-1}}(H_0)) = H_L + E_{\mathfrak{a}}(\mathrm{Ad}_{k^{-1}}(H^L))$. If we prove that $E_{\mathfrak{a}}(\mathrm{Ad}_{k^{-1}}(H^L)) = 0$, the two statements follow. On the one hand $f_{G,H_0}(k) \in \mathfrak{a}_L$ implies $E_{\mathfrak{a}}(\mathrm{Ad}_{k^{-1}}(H^L)) = f_{G,H_0}(k) - H_L \in \mathfrak{a}_L$. On the other hand, $H^L \perp \mathfrak{a}_L$ and $k \in L$ implies $\mathrm{Ad}_{k^{-1}}(H^L) \perp \mathfrak{a}_L$, and hence $E_{\mathfrak{a}}(\mathrm{Ad}_{k^{-1}}(H^L)) \perp \mathfrak{a}_L$. It follows that $E_{\mathfrak{a}}(\mathrm{Ad}_{k^{-1}}(H^L)) \in \mathfrak{a}_L \cap \mathfrak{a}^L = \{0\}$. □

Lemma 2.39. *Let $H_0 \in \mathfrak{a}^{\mathrm{reg}}$, $\mathcal{D}_{H_0}$ be as in Lemma 2.38 and $k \in \mathcal{D}_{H_0}$ a critical point of g_{G,H_0}. Define L, H_L and H^L as in the same lemma. Then the Hessian of g_{G,H_0} at k satisfies*

$$\begin{aligned}\frac{1}{2}(\mathrm{Hess}_k\, g_{G,H_0})(X,X) = &- \|[X,H_L]\|^2 - \langle [X,\mathrm{Ad}_{k^{-1}}(H^L)],[X,H_L]\rangle \\ &+ \left\|E_{\mathfrak{a}}([X,\mathrm{Ad}_{k^{-1}}(H^L)])\right\|^2\end{aligned} \tag{2.40}$$

for all $X \in \mathfrak{k}$.

Proof. Starting from (2.35) and using Leibniz's rule and (2.7), we have that the Hessian of g at the critical point k takes the form

$$\begin{aligned}&\mathrm{Hess}_k\, g : \mathfrak{k}\times\mathfrak{k} \to \mathbb{R} \\ &\qquad (X,Y) \mapsto 2\,\langle E_{\mathfrak{a}}([Y,[X,\mathrm{Ad}_{k^{-1}}(H_0)]]), f_{G,H_0}(k)\rangle \\ &\qquad\qquad +2\,\langle E_{\mathfrak{a}}([Y,\mathrm{Ad}_{k^{-1}}(H_0)]), E_{\mathfrak{a}}([X,\mathrm{Ad}_{k^{-1}}(H_0)])\rangle\ .\end{aligned}$$

As $f_{G,H_0}(k) \in \mathfrak{a}$, we may drop the projection $E_{\mathfrak{a}}$ in the first term. Replacing $f_{G,H_0}(k)$ by H_L and using associativity of the Killing form we obtain

$$\begin{aligned}&-2\,\langle [X,\mathrm{Ad}_{k^{-1}}(H_0)],[Y,H_L]\rangle \\ &\quad +2\,\langle E_{\mathfrak{a}}([Y,\mathrm{Ad}_{k^{-1}}(H_0)]), E_{\mathfrak{a}}([X,\mathrm{Ad}_{k^{-1}}(H_0)])\rangle\ .\end{aligned}$$

The associated quadratic form on $\mathfrak{k}$ sends

$$X \mapsto -2\,\langle [X,\mathrm{Ad}_{k^{-1}}(H_0)],[X,H_L]\rangle + 2\,\|E_{\mathfrak{a}}([X,\mathrm{Ad}_{k^{-1}}(H_0)])\|^2\ .$$

Writing $H_0 = H_L + H^L$ and recalling that $k \in L$ centralizes H_L, this becomes

$$\begin{aligned}&-2\,\|[X,H_L]\|^2 - 2\langle [X,\mathrm{Ad}_{k^{-1}}(H^L)],[X,H_L]\rangle \\ &\quad +2\left\|E_{\mathfrak{a}}([X,H_L+\mathrm{Ad}_{k^{-1}}(H^L)])\right\|^2\ .\end{aligned}$$

Because $H_L \in \mathfrak{a}$, associativity of the Killing form implies $E_{\mathfrak{a}}([X,H_L]) = 0$, so that the third term simplifies and we obtain (2.40). □

Lemma 2.41. *Let $H_0 \in \mathfrak{a}^{\mathrm{gen}}$ and $k \in K$ be a critical point of g_{G,H_0} with positive semidefinite Hessian. Then $g_{G,H_0}(k) = 0$.*

Proof. The proof is by contradiction. Assuming that $f_{G,H_0}(k) \neq 0$, we will construct a direction in which the Hessian of g_{G,H_0} is negative definite at k.

Let $\mathcal{D}_{H_0}$ be as in Lemma 2.38. By Lemma 2.38 and right invariance of g under M', we may assume that $k \in \mathcal{D}_{H_0}$. Let $L = Z_G(f_{G,H_0}(k))$ and let H_L and H^L be as in Lemma 2.38, so that $f_{G,H_0}(k) = H_L$ and $k \in \mathcal{C}(L,H^L)$. Suppose that $H_L \neq 0$. We will

construct $X \in \mathfrak{k}$ for which the third term in (2.40) is zero, and for which the other terms are nonpositive and not both zero.

Suppose first that $H^L \notin \mathfrak{a}^{\mathrm{reg}}$. Then there exists a nonzero element $X \in \mathfrak{k}$ with $[X, \mathrm{Ad}_{k^{-1}}(H^L)] = 0$. Indeed, if $\alpha \in \Sigma$ is such that $\alpha(H^L) = 0$, take a nonzero $X' \in (\mathfrak{g}_\alpha + \mathfrak{g}_{-\alpha}) \cap \mathfrak{k}$. Then $[X', H^L] = 0$, and we can take $X = \mathrm{Ad}_{k^{-1}}(X')$.

With this choice of X, the second and third terms in (2.40) vanish. We have $[X, H_L] = [X, \mathrm{Ad}_{k^{-1}}(H_0)] \neq 0$ by regularity of H_0, so that

$$(\mathrm{Hess}_k\, g_{G,H_0})(X,X) = -2 \cdot \|[X, H_L]\|^2 < 0\,.$$

This is a contradiction.

Suppose now that $H^L \in \mathfrak{a}^{\mathrm{reg}}$. We first show that there exists $X \in \mathfrak{k}$ for which the second term in (2.40) is strictly negative. After that, we will modify X in such a way that the second term stays the same and such that the third becomes zero.

The second term in (2.40) equals

$$-\langle [\mathrm{Ad}_k(X), H^L], [\mathrm{Ad}_k(X), H_L] \rangle\,.$$

Write $\mathrm{Ad}_k(X)$ in the restricted root space decomposition (2.1) as

$$\mathrm{Ad}_k(X) = \sum_{\alpha \in \Sigma} X_\alpha\,.$$

Then the above equals

$$-\sum_{\alpha \in \Sigma} \alpha(H^L)\alpha(H_L)\|X_\alpha\|_\theta^2\,. \tag{2.42}$$

Because $H_L \neq 0$ by assumption, there exists $\beta \in \Sigma$ such that $\beta(H_L) \neq 0$. Because $H^L \in \mathfrak{a}^{\mathrm{reg}}$, we have $\beta(H^L)\beta(H_L) \neq 0$. Using [41, Corollary 2.24] we have

$$0 = \langle H^L, H_L \rangle = \sum_{\alpha \in \Sigma} \alpha(H^L)\alpha(H_L)\,,$$

so that there is at least one strictly positive term in this sum. Let $\alpha \in \Sigma$ be such that $\alpha(H^L)\alpha(H_L) > 0$. Take now a nonzero $X \in \mathfrak{k} \cap (\mathfrak{g}_\alpha + \mathfrak{g}_{-\alpha})$, then (2.42) is strictly negative. That is, the second term in (2.40) is strictly negative.

Observe that when $Y \in \mathfrak{k} \cap \mathfrak{l}$, adding Y to X does not change the first two terms of (2.40), because $[Y, H_L] = 0$. Thus in order to make the third term in (2.40) zero and obtain a contradiction, it suffices to find $Y \in \mathfrak{k} \cap \mathfrak{l}$ with

$$E_{\mathfrak{a}}([Y, \mathrm{Ad}_{k^{-1}}(H^L)]) = E_{\mathfrak{a}}([X, \mathrm{Ad}_{k^{-1}}(H^L)])\,,$$

and it will follow that

$$(\mathrm{Hess}_k\, g_{G,H_0})(X-Y, X-Y) = -2\,\|[X, H_L]\|^2 - 2\langle [X, \mathrm{Ad}_{k^{-1}}(H^L)], [X, H_L] \rangle < 0\,.$$

Let $L' \subset L$ be the smallest semistandard Levi subgroup with the property that $k \in L'M'$, say $k = \ell' m$. From $k \in \mathcal{C}(L, H^L)$ it follows that $\mathrm{Ad}_{k^{-1}}(H^L) \perp \mathfrak{a}$, and therefore $H^L \in \mathfrak{a}^{L'}$. We also have that $[\mathfrak{k}, \mathrm{Ad}_{k^{-1}}(H^L)] \perp \mathrm{Ad}_{k^{-1}}(\mathfrak{a}) \supset \mathrm{Ad}_m^{-1}(\mathfrak{a}_{L'})$, so that $E_{\mathfrak{a}}([\mathfrak{k}, \mathrm{Ad}_{k^{-1}}(H^L)]) \subset \mathrm{Ad}_m^{-1}(\mathfrak{a}^{L'})$. Hence it suffices to show that the map

$$\begin{aligned} \mathfrak{k} \cap \mathfrak{l} &\to \mathrm{Ad}_m^{-1}(\mathfrak{a}^{L'}) \\ Y &\mapsto E_{\mathfrak{a}}([Y, \mathrm{Ad}_{k^{-1}}(H^L)]) \end{aligned} \tag{2.43}$$

is surjective. Its restriction to $\mathfrak{k} \cap \mathfrak{l}'$ is precisely $-\mathrm{Ad}_m^{-1} \circ (Df_{L', H^L})_{\ell'}$ (compare (2.34)). We seek to apply Lemma 2.36 to L'.

By assumption we have $H^L \in \mathfrak{a}^{\mathrm{reg}}$, so that $H^L \in (\mathfrak{a}^{L'})^{\mathrm{reg}}$. By minimality of L', we have $\ell' \notin \bigcup_{L'' \subsetneq L'} M'L''$. Therefore Lemma 2.36 shows that $(Df_{L', H^L})_{\ell'}$ is surjective, so that the map (2.43) is surjective and the conclusion follows. □

Corollary 2.44. *When $H_0 \in \mathfrak{a}^{\mathrm{gen}}$ we have $\mathcal{C}(G, H_0) \neq \varnothing$.*

Proof. The continuous function $g_{G,H_0} : K \to \mathbb{R}_{\geq 0}$ attains a minimum, which must be 0 by Lemma 2.41. □

2.3.3 Uniqueness of critical points

Lemma 2.45. *Let $H_0 \in \mathfrak{a}^+$ and $k \in K - \bigcup_{L \in \mathcal{L} - \{G\}} M'L$. Then the quadratic form on $\mathfrak{a}$ defined by*

$$H \mapsto \langle H_0, [\mathrm{Ad}_k(H), E_{\mathfrak{n}}(\mathrm{Ad}_k(H))] \rangle$$

is negative definite. In particular, it is nondegenerate.

Proof. Take a nonzero $H \in \mathfrak{a}$, and write the element $\mathrm{Ad}_k(H) \in \mathfrak{p}$ in the restricted root space decomposition (2.1) as $\sum_{\alpha \in \Sigma \cup \{0\}} X_\alpha$. Then $\theta X_{-\alpha} = -X_\alpha$. By Lemma 2.9, the quadratic form evaluated at H equals

$$\sum_{\alpha \in \Sigma^+} \langle -X_\alpha, 2X_\alpha \rangle_\theta \langle H_\alpha, H_0 \rangle = -2 \sum_{\alpha \in \Sigma^+} \|X_\alpha\|_\theta^2 \cdot \alpha(H_0)\,,$$

which is nonpositive because H_0 lies in the positive Weyl chamber. Therefore the quadratic form is negative semidefinite. The fact that $k \notin \bigcup_{L \in \mathcal{L} - \{G\}} M'L$ implies $\mathrm{Ad}_k(H) \notin \mathfrak{a}$ (Lemma 2.2). So at least one X_α is nonzero, and the statement follows. □

Remark 2.46. When H_0 lies in a mixed Weyl chamber, the quadratic form on $\mathfrak{a}$ in Lemma 2.45 can sometimes be degenerate. This happens already for $G = \mathrm{SL}_3(\mathbb{R})$, even for completely generic values of H_0, and there appears to be no structure in the bad pairs (H_0, k).

Proposition 2.47. *Let $H_0 \in \mathfrak{a}^+$ and $k \in K - \bigcup_{L \in \mathcal{L} - \{G\}} M'L$. Then the Hessian of $h_{H_0,k}$ is everywhere negative definite.*

Proof. Take $a \in A$ and define $k_1 = \kappa(ka)$. From Lemma 2.30 it follows that $k_1 \notin \bigcup_{L \in \mathcal{L}-\{G\}} M'L$. Using Lemma 2.9 we have for $H \in \mathfrak{a}$ that

$$\begin{aligned}(\mathrm{Hess}_a\, h_{H_0,k})(H) &= \langle H_0, [\mathrm{Ad}_{k_1}(H), E_{\mathfrak{n}}(\mathrm{Ad}_{k_1}(H))]\rangle \\ &= \langle [H_0, \mathrm{Ad}_{k_1}(H)], E_{\mathfrak{n}}(\mathrm{Ad}_{k_1}(H))\rangle\,.\end{aligned}$$

By Lemma 2.45, this quadratic form is negative definite. □

Proposition 2.48. *When $H_0 \in \mathfrak{a}^{\mathrm{gen},+}$ and $k \in K$, the function $h_{H_0,k}$ has at most one critical point, which if it exists, is nondegenerate and maximizes $h_{H_0,k}$.*

Proof. Assume that $h_{H_0,k}$ has a critical point. By Proposition 2.31, the existence of a critical point implies that $k \notin \bigcup_{L \in \mathcal{L}-\{G\}} M'L$. The Hessian of $h_{H_0,k}$ is then everywhere nondegenerate by Proposition 2.47. In particular, its critical points are nondegenerate. Moreover, Proposition 2.47 implies that the Hessian of $h_{H_0,k}$ is everywhere negative definite, so that there can be no other critical point and any critical point maximizes $h_{H_0,k}$. □

2.3.4 Structure of the level sets

We have all the necessary information to describe the set $\mathcal{C}(G, H_0)$.

Proposition 2.49. *When $H_0 \in \mathfrak{a}^{\mathrm{gen}}$, the set $\mathcal{C}(G, H_0)$ is a smooth manifold of dimension $\dim K - \dim A$ that varies smoothly with H_0.*

Proof. By Corollary 2.44, this set is nonempty, and by Lemma 2.37, f_{G,H_0} is a submersion at the points of $\mathcal{C}(G, H_0)$. Thus $\mathcal{C}(G, H_0)$ is a smooth manifold of dimension $\dim K - \dim A$, and from the local normal form for submersions [44, Theorem 4.12] it follows that it admits local parametrizations that depend smoothly on H_0. □

Lemma 2.50. *When $H_0 \in \mathfrak{a}$, the set $\mathcal{C}(G, H_0)$ is invariant on both sides by M.*

Proof. This is clear from the definition (2.24), using Ad_G-invariance of the Killing form. □

Remark 2.51. Many arguments simplify when $\mathcal{C}(G, H_0)$ is but a finite union of cosets of M in K, effectively reducing arguments to the case of $G = \mathrm{SL}_2(\mathbb{R})$. However, the set $\mathcal{C}(G, H_0)$ is usually much bigger; see Proposition 2.57.

Corollary 2.52. *When $H_0 \in \mathfrak{a}^{\mathrm{gen}}$ and $a \in A$, the set of $g \in G$ for which a is a critical point of $h_{H_0,g}$ is a smooth submanifold of codimension $\dim(A)$. It is stable on the left by NAM and on the right by M.*

Proof. From (2.27) we saw that $h_{H_0,g}$ and $h_{H_0,\kappa(g)}$ have the same critical points. Therefore by Lemma 2.26 the set in question is equal to $NA \cdot \kappa(\mathcal{C}(G, H_0)a^{-1})$, which has codimension $\dim(A)$ in $G = NAK$ by Proposition 2.49. It is stable on the left by NA. It is stable on the left by M because M normalizes NA, which implies that $\kappa(\cdot)$ commutes with left multiplication by M, and $\mathcal{C}(G, H_0)$ is stable on the left by M by Lemma 2.50. Similarly, invariance on the right follows from Lemma 2.50, using that $\kappa(\cdot)$ commutes with right multiplication by K. □

When $H_0 \in \mathfrak{a}$, define $\mathcal{R}_{H_0} \subset G$ to be the set of elements g for which the function $h_{H_0,g}$ has a critical point.

Proposition 2.53. *When $H_0 \in \mathfrak{a}^{\mathrm{gen},+}$, the set $\mathcal{R}_{H_0} \subset G$ is open and stable on the left by NAM.*

Proof. Take $g_0 \in \mathcal{R}_{H_0}$. By Proposition 2.48, the critical point of h_{H_0,g_0} is nondegenerate, call it ξ. We may reformulate this by saying that the map

$$\begin{aligned} A &\to \mathfrak{a}^* \\ a &\mapsto (Dh_{H_0,g_0})_a \end{aligned}$$

has invertible differential at the level set above 0, which is the singleton $\{\xi\}$. By the implicit function theorem applied to this smooth map with parameter $g \in G$, it follows that it has a zero for all g in a neighborhood of g_0, which is to say that $\mathcal{R}_{H_0}$ is open in G. The stability under NAM follows from Corollary 2.52. □

Remark 2.54. Let $\rho : G \to \mathrm{SL}(V)$ be any representation with finite kernel, and let $\Omega_{\mathcal{S}^0} \subset G$ be the set defined in §2.2.5. It is open and dense by Proposition 2.19. It is reasonable to expect that for $g \in \Omega_{\mathcal{S}^0}$ and $\lambda \in (\mathfrak{a}^*)^+$ we have $\lambda(H(ga)) \to -\infty$ as $a \to \infty$ in A. This would then imply that when $H_0 \in \mathfrak{a}^+$ the function $h_{H_0,g}$ has a critical point for all $g \in \Omega_{S^0}$, in particular, for all g in an open dense subset of G that does not depend on H_0. In fact, when $H_0 \in \mathfrak{a}^{\mathrm{gen},+}$ it is reasonable to expect that $\mathcal{R}_{H_0} = \Omega_{\mathcal{S}^0}$.

The statement about the limit and the corollary that $\mathcal{R}_{H_0} \supset \Omega_{\mathcal{S}^0}$ are certainly true when $G = \mathrm{SL}_n(\mathbb{R})$ and $\rho = \mathrm{Std}$. Then the entries of $H(ga)$ can be expressed in terms of quotients of sums of subdeterminants (using (2.10)), and when $g \in \Omega_{\mathcal{S}^0}$ those entries are bounded from above and below by constants times the entries of a, but arranged in decreasing order. The general case would require a more careful analysis of the weights of the exterior powers of V. This would in particular give an alternative proof of Corollary 2.44, whose only proof we have now is very technical and little intuitive.

2.3.5 Some dimension bounds

Lemma 2.55. *Let $\mathfrak{g}$ be a complex simple Lie algebra of rank r, $\mathfrak{h}$ a Cartan subalgebra with roots Σ and a choice of simple roots Π. Let $\alpha \in \Pi$. Then there exist at least r roots $\beta \in \Sigma$ with the property that $\beta \geq \alpha$.*

Proof. We will show by induction the following statement: for every $1 \leq m \leq r$, there exists a set of simple roots S with $|S| = m$ whose span contains m linearly independent roots $\beta \geq \alpha$. For $m = 1$ we take $S = \{\alpha\}$. Assume we have found S of cardinality $m < r$. Because $\mathfrak{g}$ is simple, Σ is irreducible, so there exists a simple root $\gamma \in \Pi - S$ which is not orthogonal to all roots in S. Thus there exists a positive root $\beta \in \mathrm{span}(S)$ with $\beta \geq \alpha$ which is not orthogonal to γ. Because $\langle \beta, \gamma \rangle_\theta \leq 0$ by [41, Lemma 2.51], we must have $\langle \beta, \gamma \rangle_\theta < 0$. By [41, Proposition 2.48] this implies that $\beta + \gamma$ is a root. We may now take $S' = S \cup \{\gamma\}$, whose span contains m linearly independent roots $\geq \alpha$ in $\mathrm{span}(S)$ together with the root $\beta + \gamma \geq \alpha$. This completes the induction. □

Lemma 2.56. *Let $\mathfrak{g}$ be a real simple Lie algebra with Cartan decomposition $\mathfrak{k} \oplus \mathfrak{p}$, maximal abelian subalgebra $\mathfrak{a} \subset \mathfrak{p}$ with restricted roots $\Sigma \subset \mathfrak{a}^*$ and choice of positive roots Σ^+. The following are equivalent:*

1. *$\mathfrak{g}$ is compact or isomorphic to $\mathfrak{sl}_2(\mathbb{R})$.*
2. *$\sum_{\alpha \in \Sigma^+} m(\alpha) = \dim(\mathfrak{a})$.*

Proof. Write $r = \dim(\mathfrak{a})$. Note that Σ spans $\mathfrak{a}^*$ so that we always have

$$\sum_{\alpha \in \Sigma^+} m(\alpha) \geq \#\Sigma^+ \geq r\,.$$

It is clear that equality holds if $\mathfrak{g}$ is compact or isomorphic to $\mathfrak{sl}_2(\mathbb{R})$.

Now let $\mathfrak{g}$ be any simple Lie algebra. We will show that equality only holds for the examples stated. There are two cases, coming from the classification of simple real Lie algebras.

Suppose first that the complexification $\mathfrak{g}_\mathbb{C}$ is not simple. By [41, Theorem 6.94] $\mathfrak{g}$ is the restriction of scalars of a simple complex Lie algebra. Let $\mathfrak{h} \supset \mathfrak{a}$ be a Cartan subalgebra. It follows as in [41, §VI.11, p.425] that all roots of $\mathfrak{h}$ have nonzero restriction to $\mathfrak{a}$ (and all restrictions are different) and all restricted roots have multiplicity 2. It follows that

$$\sum_{\alpha \in \Sigma^+} m(\alpha) = 2\#\Sigma^+ \geq 2r \geq r + 1\,,$$

because $r \geq 1$, meaning that equality does not hold in this case.

Suppose now that the complexification $\mathfrak{g}_\mathbb{C}$ is simple. Let Π be a system of simple restricted roots of $\mathfrak{a}$. Let θ be a Cartan involution corresponding to $\mathfrak{k} \oplus \mathfrak{p}$ and $\mathfrak{h} \subset \mathfrak{g}$ a θ-stable Cartan subalgebra containing $\mathfrak{a}$. We may choose an ordering on $\mathfrak{h}_\mathbb{C}^*$ such that the restrictions of positive roots of $\mathfrak{h}_\mathbb{C}$ to $\mathfrak{a}$ are either zero or positive. As in [41, §VI.12, Problem 7], all simple $\alpha \in \Pi$ are restrictions of simple roots of $\mathfrak{h}_\mathbb{C}$. Let $\alpha_1, \ldots, \alpha_r$ be simple lifts to $\mathfrak{h}_\mathbb{C}$ of the simple restricted roots of $\mathfrak{a}$. If $\mathfrak{g}$ is noncompact, then $r \geq 1$. Because $\mathfrak{g}_\mathbb{C}$ is simple, by Lemma 2.55 there are at least $\dim_\mathbb{C}(\mathfrak{h}_\mathbb{C}) - 1$ positive roots β of

$\mathfrak{h}_{\mathbb{C}}$ with $\beta > \alpha_1$. The restrictions of all these roots are positive. Though the restrictions may coincide, they give the following lower bound for multiplicities:

$$\sum_{\alpha\in\Sigma^+} m(\alpha) \geq r + \dim_{\mathbb{C}}(\mathfrak{h}_{\mathbb{C}}) - 1\,.$$

if this is equal to r, then $\dim_{\mathbb{C}}(\mathfrak{h}_{\mathbb{C}}) = 1$. That is, $\mathfrak{g}_{\mathbb{C}} \cong \mathfrak{sl}_2(\mathbb{C})$. Because we are assuming $\mathfrak{g}$ is noncompact, it must be the unique noncompact form of $\mathfrak{sl}_2(\mathbb{C})$, which is $\mathfrak{sl}_2(\mathbb{R})$. □

Proposition 2.57. *Let G be a semisimple Lie group. The following are equivalent:*

1. $\dim(K) - \dim(A) = \dim(M)$.
2. *All simple factors of $\mathfrak{g}$ are either compact or $\mathfrak{sl}_2(\mathbb{R})$.*

Note that we always have $\dim(K) \geq \dim(M) + \dim(A)$.

Proof. The difference $\dim(K) - \dim(M)$ is equal to $\sum_{\alpha\in\Sigma^+} m(\alpha)$. If this is an equality, then we must have the same equality for all simple factors of $\mathfrak{g}$. By Lemma 2.56, this is equivalent to saying that all simple factors of $\mathfrak{g}$ are either compact or $\mathfrak{sl}_2(\mathbb{R})$. □

2.3.6 Proofs specific to $\mathrm{SL}_3(\mathbb{R})$

Some of the lemmas that go into the proof of Proposition 2.49 admit more direct and computational proofs when $G = \mathrm{SL}_3(\mathbb{R})$. We include one of those here, because it features a rather curious inequality.

Direct proof of Corollary 2.44 when $G = \mathrm{SL}(3,\mathbb{R})$. We use the standard choice of Iwasawa decomposition. Write $H_0 = \mathrm{diag}(a,b,c)$. Proving that the set $\mathcal{C}(G,H_0)$ is nonempty is equivalent to showing that there exists a symmetric matrix $X = \mathrm{Ad}_{k^{-1}}(H_0)$ with zeroes on the diagonal, which is isospectral with H_0. Write $X = \left(\begin{smallmatrix} 0 & x & y \\ x & 0 & z \\ y & z & 0 \end{smallmatrix}\right)$. Then H_0 and X have the same characteristic polynomial when

$$\begin{cases} x^2 + y^2 + z^2 = -(ab+bc+ca) = \frac{1}{2}(a^2+b^2+c^2) \\ 2xyz = abc\,. \end{cases} \tag{2.58}$$

Because the first equation does not see the signs of x, y, z, squaring the second does not affect solvability. We may now view (2.58) as prescribing the arithmetic and geometric mean of x^2, y^2, z^2, namely $\frac{1}{6}(a^2+b^2+c^2)$ respectively $(\frac{1}{4}a^2b^2c^2)^{1/3}$. It is well known that a necessary and sufficient condition for this to have a solution in nonnegative numbers x^2, y^2, z^2, is that the arithmetic mean is at least the geometric mean, meaning that

$$\frac{1}{6}(a^2+b^2+c^2) \geq \left(\frac{1}{4}a^2b^2c^2\right)^{1/3}\,.$$

That this condition is satisfied is Lemma 2.59 below, and it does not require the hypothesis that $H_0 \in \mathfrak{a}^{\mathrm{gen}}$. □

Lemma 2.59 (AM–$\sqrt[3]{2}$GM)**.** *Let $a, b, c \in \mathbb{R}$ with $a + b + c = 0$. Then*

$$\left(\frac{a^2 + b^2 + c^2}{3}\right)^3 \geq 2 \cdot a^2 b^2 c^2 \,.$$

The proof of Lemma 2.59 is an amusing exercise. Equality holds if and only if (a, b, c) is proportional to $(1, 1, -2)$ or a permutation thereof. With the notations as in Corollary 2.44, this corresponds to the case where H_0 lies in $\mathfrak{a}_L$ for some semistandard Levi subgroup $L \in \mathcal{L} - \{G\}$.

2.4 The K-projection and the geodesic flow

In this section we prove Theorem 7. For the possibility of uniformity, see Remark 2.68. For $H \in \mathfrak{a}$ and $k \in K$, define a map

$$\begin{aligned} p_{H,k} : \mathbb{R} &\to \mathfrak{p} \\ t &\mapsto \mathrm{Ad}_{\kappa(ke^{tH})}(H) \,. \end{aligned} \tag{2.60}$$

Using (2.6) and Lemma 2.8 to differentiate κ, we find that

$$p'_{H,k}(t) = [E_{\mathfrak{k}}(p_{H,k}(t)), p_{H,k}(t)] \,.$$

Therefore $p_{H,k}(t)$ is a solution to the homogeneous quadratic differential equation in $\mathfrak{p}$ given by

$$X' = [E_{\mathfrak{k}} X, X] \tag{2.61}$$

with initial value $\mathrm{Ad}_k(H)$. Its flow

$$\begin{aligned} p : \mathbb{R} \times \mathfrak{p} &\to \mathfrak{p} \\ (t, X) &\mapsto p_t(X) \,. \end{aligned} \tag{2.62}$$

is given explicitly by $p_t(X) = p_{H,k}(t)$ when $X = \mathrm{Ad}_k(H)$, and is in particular defined for all $t \in \mathbb{R}$. We will prove Theorem 7 by gathering enough information about the dynamical system (2.62).

Example 2.63. Let $G = \mathrm{PSL}_2(\mathbb{R})$ and identify $\mathfrak{p}$ with $\mathbb{R}^2$ via an isometry that sends $\mathfrak{a}$ to the horizontal axis $\mathbb{R} \times \{0\}$. It is clear from (2.61) that the points of $\mathfrak{a}$ are stationary, and it is clear from (2.60) that the norm of $X \in \mathfrak{p}$ is invariant under the flow. It is then not hard to see that the phase portrait of the dynamical system (2.62) is as follows: the points of $\mathfrak{a}^+$ are unstable equilibra, the points of $\mathfrak{a}^-$ are stable equilibria, and apart from the equilibrium at 0 every other orbit is heteroclinic and describes a Euclidean half circle with endpoints on the horizontal axis, starting at $\mathfrak{a}^+$ and ending at $\mathfrak{a}^-$.

It is clear from (2.61) that all elements of $\mathfrak{a}$ are equilibria. The key in the proof of Theorem 7 is the existence of functions that are monotonic along orbits. For a root $\alpha \in \Sigma$, let $H_\alpha \in \mathfrak{a}$ be as in §2.1.3. The set $\{H_\alpha : \alpha \in \Pi\}$ is a basis of $\mathfrak{a}$. When $\alpha \in \Pi$ and $H \in \mathfrak{a}$, define $c_\alpha(H)$ to be the αth coordinate of H in this basis. More generally, define $c_\alpha(X) = c_\alpha(E_\mathfrak{a}(X))$ for $X \in \mathfrak{g}$.

Lemma 2.64. *Let $X \in \mathfrak{p} - \mathfrak{a}$. Then $c_\alpha([E_\mathfrak{k}X, X]) \leq 0$ for all $\alpha \in \Pi$, and at least one is strictly negative. That is, $E_\mathfrak{a}[E_\mathfrak{k}X, X]$ is nonzero and lies in the closure of the negative Weyl chamber.*

Proof. Write $X = \sum_{\alpha \in \Sigma \cup \{0\}} X_\alpha$ in the restricted root space decomposition (2.1). Using that $X_{-\alpha} = -\theta(X_\alpha)$ and $E_\mathfrak{k}X = \sum_{\alpha \in \Sigma^+}(X_{-\alpha} - X_\alpha)$, we find that

$$\begin{aligned} E_\mathfrak{a}[E_\mathfrak{k}X, X] &= \sum_{\alpha \in \Sigma^+} [X_{-\alpha} - X_\alpha, X_{-\alpha} + X_\alpha] \\ &= -2 \sum_{\alpha \in \Sigma^+} [X_\alpha, X_{-\alpha}] \\ &= -2 \sum_{\alpha \in \Sigma^+} \langle X_\alpha, X_{-\alpha} \rangle \cdot H_\alpha \\ &= -2 \sum_{\alpha \in \Sigma^+} \|X_\alpha\|_\theta^2 \cdot H_\alpha \\ &= -2 \sum_{\alpha \in \Pi} \sum_{\beta \geq \alpha} \|X_\beta\|_\theta^2 \cdot H_\alpha \end{aligned}$$

where in the third equality we have used [41, §II, Lemma 2.18]. Therefore the coefficients of $E_\mathfrak{a}[E_\mathfrak{k}X, X]$ in the basis $\{H_\alpha : \alpha \in \Pi\}$ are nonnegative, and because $X \notin \mathfrak{a}$ at least one is nonzero. (This is essentially the same proof as that of Lemma 2.45.) □

Lemma 2.65. *Let $X \in \mathfrak{p} - \mathfrak{a}$ and $t > 0$. Then $c_\alpha(p_t(X)) \leq c_\alpha(X)$ for all $\alpha \in \Pi$, and at least one inequality is strict.*

Proof. By Lemma 2.64, every $c_\alpha(p_t(X))$ is monotonically decreasing, and near every $t \in \mathbb{R}$ at least one is strictly decreasing. □

Lemma 2.66. *The only non-wandering points for the dynamical system* (2.62) *are those of $\mathfrak{a}$.*

Proof. Let $X \in \mathfrak{p} - \mathfrak{a}$. By Lemma 2.65 there exists $\alpha \in \Pi$ for which $c_\alpha(p_1(X)) < c_\alpha(X)$. By continuity of the flow, there exists a neighborhood U of X such that $c_\alpha(p_1(U)) < c_\alpha(U)$. Because c_α is decreasing along orbits, we have $p_t(U) \cap U = \varnothing$ for all $t \geq 1$. Thus X is wandering. □

Lemma 2.67. *Let $X \in \mathfrak{p}$. There exists $H \in \mathfrak{a}$ such that $\lim_{t \to \infty} p_t(X) = H$.*

Proof. Write $X = \mathrm{Ad}_k(H)$ with $k \in K$ and $H \in \mathfrak{a}$. Consider its ω-limit $\omega(X)$. By Lemma 2.66 we have $\omega(X) \subset \mathfrak{a}$. In view of the explicit expression for $p_t(X)$, we have $\omega(X) \subset \mathrm{Ad}_K(H)$ by continuity. Thus $\omega(X) \subset \mathfrak{a} \cap \mathrm{Ad}_K(H)$. By Lemma 2.2, this intersection is equal to the Weyl group orbit of H. In particular, it is discrete. Because $\omega(X)$ is connected, it must consist of a single point of $\mathfrak{a}$. □

Proof of Theorem 7. The K-projection of ge^{tH} does not depend on the triangular part of g, so we may assume that $g \in K$. Conjugating H by $k \in K$ gives a diffeomorphism

$$K/Z_K(H) \to \mathrm{Ad}_K(H)\,.$$

Take $k \in G$. When $t \to +\infty$, the image of $\kappa(ke^{tH})$ under this map tends to an element $\mathrm{Ad}_m(H) \in \mathrm{Ad}_{M'}(H)$ by Lemma 2.67. Therefore $\kappa(ke^{tH})$ tends to the point $mZ_K(H)$ in the quotient $K/Z_K(H)$. Equivalently, $\kappa(ke^{tH})$ tends to the set $mZ_K(H)$ in K. □

Remark 2.68. In general the convergence in Lemma 2.67 is not uniform. Already for $G = \mathrm{SL}_3(\mathbb{R})$ there are heteroclinic orbits that have nearby orbits with totally different limit points, even some that emanate from equilibria in the positive Weyl chamber. Some of these orbits, that come in one-dimensional families, can be seen to lie in Levi subalgebras (Lie algebras of $L \in \mathcal{L}$). But not all of them do and certainly not those that come in two-dimensional families. In fact, it seems likely that these badly behaved orbits correspond exactly to the partition of K into the sets $\kappa(\Omega_{\mathcal{S}})$ defined in §2.2.5; compare Example 2.21.

Further evidence for this is the following observation. Let $\rho : G \to \mathrm{SL}(V)$ be any representation with finite kernel, and let $\Omega_{\mathcal{S}^0}$ be the set defined in §2.2.5. As explained in Remark 2.54 we expect that for $g \in \Omega_{\mathcal{S}^0}$ and $\lambda \in (\mathfrak{a}^*)^+$ we have $\lambda(H(ga)) \to -\infty$ as $a \to \infty$ in A, and we sketched a proof of this when $G = \mathrm{SL}_n(\mathbb{R})$ and $\rho = \mathrm{Std}$. Moreover, when $\lambda \in \Sigma^+$ is merely a positive root, it is still reasonable to expect that $\lambda(H(ga)) \to -\infty$ as $a \to \infty$ in A uniformly along regular directions; this is also apparent for $G = \mathrm{SL}_n(\mathbb{R})$. Now writing $ga = n'a'k'$, we have for $H \in \mathfrak{a}$ that $\mathrm{Ad}_{k'}(H) = \mathrm{Ad}_{a'^{-1}n'^{-1}g}(H)$. When $g \in \Omega_{\mathcal{S}^0}$ stays in a compact set, then so does n', by Theorem 2.17, and if $\lambda(a') \to -\infty$ for all positive roots λ, then the positive root space components of $\mathrm{Ad}_{a'^{-1}n'^{-1}g}(H)$ tend to zero. On the other hand this equals $\mathrm{Ad}_{k'}(H)$, which lives in a bounded subset of $\mathfrak{p}$, so that the negative root space components approach zero as well, meaning that $\mathrm{Ad}_{k'}(H) \to \mathfrak{a}$ and therefore $k' \to M'Z_K(H)$. Taking $H \in \mathfrak{a}$ regular, we would obtain that $k' \to M'$, uniformly for g in compact subsets of $\Omega_{\mathcal{S}^0}$ and $a \to \infty$ inside regular cones of A.

3 Toric periods on semisimple groups

This chapter contains the proofs of Theorem 3 and Theorem 4. Section §3.2 contains the skeleton of the proof of Theorem 4. It relies on asymptotics and bounds for orbital integrals, which are proved using stationary phase analysis in §3.3 and §3.4. The analysis relies heavily on results from §2.3. The remainder of the chapter is concerned with Theorem 3. The proof is given in §3.6. The main arithmetic input, which is the construction of a suitable Hecke operator, is contained in §3.7.

3.1 Preliminaries on symmetric spaces

We will use the notation from §2.1 for Lie groups and symmetric spaces. So let G be reductive as in §2.1.1, and we will assume throughout that the identity component G^0 is semisimple. (Recall that a semisimple Lie group for us is by definition connected. That is, we follow [41].) The Lie group G will always be connected in §3.2 and in the analytic sections following it. It will be again allowed to be disconnected from §3.5 onwards.

We equip G with any Riemannian metric and the associated distance function $d(\cdot,\cdot)$. We will care only about distances up to constant factors, so we do not need to impose any invariance properties of $d(\cdot,\cdot)$.

3.1.1 Measures and convolution

We fix any Haar measure dg on G and let dk be the Haar measure on K for which $\mathrm{Vol}(K) = 1$. We fix the Haar measure on A that coincides with the measure induced from the left-invariant Riemannian metric on the symmetric space G/K on the submanifold $A \subset G/K$. If H is a subgroup between A and MA, this then defines a unique Haar measure on H.

The universal enveloping algebra $U(\mathfrak{g})$ acts on $C^\infty(G)$ as in §2.1.5. The convolution algebra $C_c^\infty(K\backslash G/K)$ acts on $C^\infty(G/K)$ by right translation, and it is commutative [34, Theorem IV.3.1].

3.1.2 Maass forms

Let G be semisimple and $\Gamma \subset G$ a co-compact and torsion-free lattice. The quotient $X = \Gamma\backslash G/K$ is a compact locally symmetric space, and $C^\infty(\Gamma\backslash G)$ carries an action

of $U(\mathfrak{g})$. The algebra $Z(U(\mathfrak{g}))$ is commutative, and we may find an orthonormal basis $(f_j)_{j\geq 0}$ of $L^2(X)$ consisting of simultaneous eigenfunctions for $Z(U(\mathfrak{g}))$, the Maass forms.

For $\nu \in (\mathfrak{a}^*)_{\mathbb{C}}$, the complexified dual of $\mathfrak{a}$, let $\varphi_\nu : G \to \mathbb{C}$ be the spherical function of parameter ν [34, Chapter IV]. It is given explicitly by Harish-Chandra's integral representation as

$$\varphi_\nu(g) = \int_K \exp\left((\rho + i\nu)(H(kg))\right) dk\,, \tag{3.1}$$

where $H : G \to \mathfrak{a}$ is the Iwasawa projection and $\rho \in \mathfrak{a}^*$ the half-sum of positive roots (§2.1.1) [34, Theorem IV.4.3]. We say f_j has spectral parameter $\nu_j \in (\mathfrak{a}^*)_{\mathbb{C}}$ if it has the same $Z(U(\mathfrak{g}))$-eigenvalues as φ_{ν_j}; such ν_j is uniquely determined up to the action of the Weyl group.

The f_j's are eigenfunctions for the algebra $C_c^\infty(K\backslash G/K)$. For any compactly supported smooth kernel $k \in C_c^\infty(K\backslash G/K)$, define the Harish-Chandra transform

$$\begin{aligned} \widehat{k} : (\mathfrak{a}^*)_{\mathbb{C}} &\to \mathbb{C} \\ \nu &\mapsto \int_G k(g)\varphi_{-\nu}(g) dg\,. \end{aligned} \tag{3.2}$$

Then $k * f_j = \widehat{k}(\nu_j) f_j$ (the convolution being on G). We recover k from $\widehat{k}$ by

$$k(g) = \frac{1}{|W|} \int_{\mathfrak{a}^*} \varphi_\nu(g)\widehat{k}(\nu)\beta(\nu) d\nu \tag{3.3}$$

where $\beta(\nu)$ is the Plancherel density [34, §IV.7], provided that the Haar measure on G in (3.2) is appropriately normalized (see [34, Exercise IV.C.4]).

3.1.3 Periods

When $\mathscr{F} \subset X$ is a compact maximal flat submanifold, we may lift it to a maximal flat submanifold of the symmetric space G/K. The lift is the image of a subset of G of the form gA. Such g is uniquely determined by the choice of the lift, up to multiplication on the right by $N_G(A)$. Because $\mathscr{F}$ is compact, Γ intersects gAg^{-1} in a lattice. We define the period of f_j along $\mathscr{F}$ to be the integral

$$\mathscr{P}_{\mathscr{F}}(f_j) = \int_{(\Gamma \cap gAg^{-1})\backslash gA} f_j\,. \tag{3.4}$$

Remark 3.5. The definition (3.4) does not depend on the choice of g or even on the choice of Iwasawa A-component A. But do note that the inclusion $gA \subset G$ induces a closed embedding of $N_\Gamma(gAg^{-1})\backslash gA$ in $\Gamma\backslash G/K$ with image $\mathscr{F}$, and not (always) of the quotient of gA by $\Gamma \cap gAg^{-1}$. That is, one may also consider the integral $\int_{\mathscr{F}} f_j$ with respect to the induced metric on the submanifold $\mathscr{F} \subset G/K$. By our choice of Haar measure on A, the latter integral equals $\mathscr{P}_{\mathscr{F}}(f_j)$ divided by the finite number $|N_\Gamma(gAg^{-1})/(\Gamma \cap gAg^{-1})|$.

3.2 Mean square asymptotics for maximal flat periods

This section contains the skeleton of the proof of Theorem 4. We set up a relative pre-trace formula in §3.2.1 and prove the theorem in §3.2.3, using the analytic results from the later sections §3.3 and §3.4. For most of this section, G can be any semisimple Lie group and X any associated compact locally symmetric space. Only in §3.2.3 will we use that G and X as in Theorem 4.

3.2.1 Classical setup

Let $k \in C_c^\infty(K\backslash G/K)$ be a compactly supported smooth kernel. The automorphic kernel

$$k^{\mathrm{aut}}(x,y) = \sum_{\gamma\in\Gamma} k(x^{-1}\gamma y)$$

acts on $L^2(\Gamma\backslash G)$, and its image lies in the space of right K-invariant functions. If this operator is positive on K-invariants (which will always be the case for us), then it has spectral expansion

$$\sum_{\gamma\in\Gamma} k(x^{-1}\gamma y) = \sum_j \widehat{k}(\nu_j) f_j(x)\overline{f_j(y)}\,, \tag{3.6}$$

and both sums are uniformly convergent on $X\times X$. When Γ is torsion free and preserves orientations of the Riemannian manifold G/K, this is essentially a theorem of Mercer [49, p. 445] (see also [76, Satz VI.4.2.]). For the general case, see for example [70].

As in §3.1.3, let $g \in G$ be such that gA projects to $\mathscr{F}$ in the quotient $\Gamma\backslash G/K$. Define $H = gAg^{-1}$ and $\Gamma_H = \Gamma\cap H$.

Lemma 3.7 (Partitions of unity)**.** *There exists a nonnegative $b \in C_c^\infty(A)$ satisfying $\sum_{\gamma\in\Gamma_H} b(g^{-1}\gamma g a) = 1$ for all $a \in A$.*

Proof. We may find $b_0 \in C_c^\infty(\mathbb{R})$ with the property that $\sum_{n\in\mathbb{Z}} b_0(x+n) = 1$ for all $x\in\mathbb{R}$. Because $\log(\Gamma_H)$ is a lattice of full rank in $\log(H) = \mathrm{Lie}(H)$, out of b_0 we may construct $b \in C_c^\infty(A)$ as in the statement. □

The group $N_G(H)/H$ is compact, which implies that the discrete subgroup $N_\Gamma(H)/\Gamma_H$ of $N_G(H)/H$ is finite.

Lemma 3.8. *There exists a nonngative $b \in C_c^\infty(A)$ with the property that for all $k \in C_c^\infty(K\backslash G/K)$ we have*

$$\sum_j \widehat{k}(\nu_j)|\mathscr{P}_{\mathscr{F}}(f_j)|^2 = \mathrm{Vol}(\Gamma_H\backslash H)\cdot|N_\Gamma(H)/\Gamma_H|\int_A k(a)da + \sum_{\gamma\in\Gamma-N_\Gamma(H)}\int_{A\times A} b(a_1)b(a_2)k(a_1^{-1}g^{-1}\gamma g a_2)da_1da_2\,.$$

Proof. This follows from fairly standard manipulations. Take b to be a cutoff function given by Lemma 3.7. Integrating (3.6) over $(\Gamma_H\backslash H)^2$ and using (3.4) gives

$$\begin{aligned}\sum_j \widehat{k}(\nu_j)|\mathscr{P}_{\mathscr{F}}(f_j)|^2 &= \int_{(\Gamma_H\backslash H)^2} \sum_{\gamma\in\Gamma} k(x^{-1}\gamma y)dxdy \\ &= \int_{((g^{-1}\Gamma_H g)\backslash A)^2} \sum_{\gamma\in\Gamma} k(a_1^{-1}g^{-1}\gamma g a_2)da_1da_2\,.\end{aligned}$$

We split the γ-sum into sums over the disjoint subsets $N_\Gamma(H)$ and $\Gamma - N_\Gamma(H)$. Because these sets are stable on both sides under Γ_H we may distribute the integral over the two terms. Concretely, the integral is the sum of the "diagonal term"

$$\int_{((g^{-1}\Gamma_H g)\backslash A)^2} \sum_{\gamma\in N_\Gamma(H)} k(a_1^{-1}g^{-1}\gamma g a_2)da_1da_2 \tag{3.9}$$

and an "off-diagonal term" where the sum goes over $\Gamma - N_\Gamma(H)$. For (3.9), unfolding in a_2 gives

$$= \int_{((g^{-1}\Gamma_H g)\backslash A)\times A} \sum_{\gamma\in N_\Gamma(H)/\Gamma_H} k(a_1^{-1}g^{-1}\gamma g a_2)da_1da_2\,.$$

The γ-sum is now finite. We may change the order of summation and make for fixed γ and a_1 the change of variables in the a_2-integral that makes the argument of k equal to a_2, keeping in mind that $g^{-1}N_G(H)g = N_G(A) \subset N_K(A)A$ and that k is bi-K-invariant. This then gives the main term in the statement.

For the off-diagonal term, inserting the partitions of unity from Lemma 3.7, making a change of variables in the γ-sum and unfolding gives that the integral equals

$$\begin{aligned}&= \int_{(g^{-1}\Gamma_H g\backslash A)^2} \sum_{\gamma_1,\gamma_2\in\Gamma_H} b(g^{-1}\gamma_1 g a_1)b(g^{-1}\gamma_2 g a_2) \\ &\qquad \cdot \sum_{\gamma\in\Gamma-N_\Gamma(H)} k(a_1^{-1}g^{-1}\gamma_1^{-1}\gamma\gamma_2 g a_2)da_1da_2 \\ &= \int_{A\times A} b(a_1)b(a_2) \sum_{\gamma\in\Gamma-N_\Gamma(H)} k(a_1^{-1}g^{-1}\gamma g a_2)da_1da_2\,.\end{aligned}$$

Note that this $A\times A\times\Gamma$-integral is indeed absolutely convergent, even compactly supported thanks to the compact support of b and k. Applying Fubini gives us the sum in the statement. □

3.2.2 Test functions

To prove Theorem 4 we make a choice of $k \in C_c^\infty(K\backslash G/K)$ to filter out $O(1)$ sums on the spectral side of the pre-trace formula (3.6). For all spectral parameters ν_j we either have $\nu \in \mathfrak{a}^*$ or $\mathrm{Re}(\nu)$ is singular and $\|\mathrm{Im}(\nu)\| \leq \|\rho\|$ (see [34, §IV.8, Theorem 8.2] for the bound on $\|\mathrm{Im}(\nu)\|$ and [40, Theorem 16.6] for the statement about singularity).

Lemma 3.10. *Let $R > 0$. We may find for every $\nu \in \mathfrak{a}^*$ a $k_\nu \in C_c^\infty(K\backslash G/K)$ with the following properties:*

1. *$\widehat{k}_\nu(\mu) \geq 0$ for $\mu \in \mathfrak{a}^*$;*
2. *$\widehat{k}_\nu(\mu) \geq 1$ for $\mu \in \mathfrak{a}^*$ with $\|\mu - \nu\| \leq R$;*
3. *$\widehat{k}_\nu(\mu) \ll_N (1 + \|\nu - \mu\|)^{-N}$ uniformly in ν and $\mu \in (\mathfrak{a}_\mathbb{C})^*$ with $\|\mathrm{Im}(\mu)\| \leq R$;*
4. *k_ν has support bounded independently of ν;*
5. *If G is simple modulo its center, $k_\nu(g) \ll \beta(\nu)(1 + \|\nu\| d(g, K))^{-1/2}$ uniformly in ν and g.*

Proof. Such k_ν can be constructed using the Paley-Wiener theorem of Gangolli [26, Theorem 3.5]; see for example [48, §2.2] for a construction, and [13, §4.1] for a proof of the last condition. □

We will apply Lemma 3.10 with any fixed $R > \|\rho\|$ in order to get decay on the spectrum of $L^2(X)$.

Remark 3.11. The last condition in Lemma 3.10 is only used in the proof of Theorem 3. Specifically, in Proposition 3.88.

3.2.3 Diagonal and off-diagonal estimates

Assume now that G and the locally symmetric space are as in Theorem 4. The theorem will be proven by estimating the different terms that appear in the right-hand side in Lemma 3.8. Using the inversion formula (3.3) it will suffice to bound the analogous terms with the test function k replaced by the spherical function. The results in question are proven in §3.3 and §3.4.

Lemma 3.12. *When $\gamma \in \Gamma$ is such that $g_\infty^{-1}\gamma g_\infty \notin M'A$, then $g_\infty^{-1}\gamma g_\infty \notin \bigcup_{L \in \mathcal{L} - \{G\}} M'L$.*

Proof. When G has rank 1 there is nothing to prove because $\mathcal{L} = \{MA, G\}$. Assume now that $G = \mathrm{SL}_p(\mathbb{R})$ with p prime, and that Γ comes from a $\mathbb{Q}$-form $\mathbf{G}$ of the algebraic group $\mathbf{SL}_p$. By the discussion in §3.5.4, the maximal flat submanifold comes from a maximal torus $\mathbf{H} \subset \mathbf{G}$. If $g_\infty^{-1}\gamma g_\infty \notin M'A$ but it lies in some $M'L$, then the subvariety of $\mathbf{H}$ that is sent to $\mathbf{H}$ by the conjugacy action of γ is a proper subtorus defined over $\mathbb{Q}$. This is

not possible because p is prime, in fact, the algebraic group $\mathbf{G}$ has no nontrivial proper closed connected subgroups other than its maximal tori [27, Proposition 4.1, Corollary 4.2]. □

Remark 3.13. In general, even when a locally symmetric space $X = \Gamma\backslash G/K$ is compact, it can happen that $g^{-1}\Gamma g$ intersects the groups $L \in \mathcal{L} - \{G, MA\}$ (see §2.1.4) nontrivially. When for example $G = \mathrm{PGL}_4(\mathbb{R})$ and Γ is an arithmetic lattice coming from a degree 4 division algebra D over $\mathbb{Q}$, then $\mathscr{F}$ corresponds to a quartic totally real field $F \subset D$. Because F is quartic, it has a quadratic subfield E. It gives rise to a 1-dimensional subgroup $H' \subset H$ and the centralizer $Z_D(E)$ gives rise to a hyperbolic plane $\mathcal{H} \subset G/K$ and an infinite discrete subgroup of $\Gamma \cap Z_G(H')$ that acts co-compactly on $\mathcal{H}$.

Lemma 3.14. *Let $D_{\mathfrak{a}^*} \subset (\mathfrak{a}^*)^{\mathrm{gen}}$ be compact and let k_ν for $\nu \in \mathfrak{a}^*$ be as in Lemma 3.10. Uniformly for $\nu \in D_{\mathfrak{a}*}$ and $t \geq 1$ we have*

$$\sum_j \widehat{k}_{t\nu}(\nu_j) \, |\mathscr{P}_{\mathscr{F}}(f_j)|^2 \asymp \beta(t\nu) \cdot (1+t)^{-r} \, .$$

Proof. With the analytic results from §3.3 and §3.4, this follows from standard properties of the Plancherel density and our assumptions on k_ν. We apply Lemma 3.8 to $k = k_{t\nu}$. Let $b_0 \in C_c^\infty(K\backslash G/K)$ with the property that $b_0(g) = 1$ when $g \in \bigcup_{\nu \in \mathfrak{a}^*} \mathrm{supp}(k_\nu)$; this exists by Lemma 3.10. Using (3.3) and Fubini the first term in the right-hand side of Lemma 3.8 can be written as

$$\mathrm{Vol}(\mathscr{F}) \cdot |N_\Gamma(H)/\Gamma_H| \int_{\mathfrak{a}^*} \widehat{k}_{t\nu}(\mu)\beta(\mu) \int_A b_0(a)\varphi_\mu(a) da d\mu \, .$$

The contribution of $\mu \in \mathfrak{a}^*$ with $\|\mu - t\nu\| \geq \sqrt{t}$ (say) can be bounded trivially using the rapid decay of $\widehat{k}_{t\nu}$ (Lemma 3.10), the polynomial growth of $\beta(\mu)$ [34, §IV.7, Proposition 7.2] and the bound $\varphi_\mu(g) \ll 1$ which follows for example from the fact that φ_μ is positive definite [34, Exercise IV.B.9]. For the remaining $\mu \in \mathfrak{a}^*$ we have $\mu \in (\mathfrak{a}^*)^{\mathrm{gen}}$ when t is sufficiently large, and we may apply the asymptotic from Proposition 3.15 to the inner integral. Note that for such μ we have $\beta(\mu) \asymp \beta(t\nu)$ by the almost-polynomial behavior of β [19, Lemma 3.11]. Finally, using the positivity of $\widehat{k}_{t\nu}$ and the lower bound from Lemma 3.10 we get that the double integral is $\asymp \beta(t\nu)(1+t)^{-r}$.

The other terms in Lemma 3.8 are finite in number by the support condition on $k_{t\nu}$. By our assumptions on G and X, Lemma 3.12 says that $g^{-1}\gamma g \notin \bigcup_{L \in \mathcal{L}-\{G\}} M'L$ for such terms. Using Proposition 3.28 and entirely similar arguments as for the diagonal term, one shows that they are $\ll \beta(t\nu)(1+t)^{-r-\delta}$ for some $\delta > 0$. □

Proof of Theorem 4. We wish to replace the weight $\widehat{k}_{t\nu}$ in Lemma 3.14 by a sharp cut-off. This can be done by applying the lemma to various $t\nu \in \mathfrak{a}^*$ just like we did in Proposition 1.50; see for example [13, Lemma 4.5]. □

3.3 Archimedean model integrals

The main result of this section is the following proposition, whose proof is in §3.3.4. The notations for Lie groups and Lie algebras are as in §2.1.1.

Proposition 3.15. *Let G be a semisimple Lie group and $b \in C_c^\infty(A)$ with $b(1) > 0$. Let $D_{\mathfrak{a}^*} \subset (\mathfrak{a}^*)^{\mathrm{gen}}$ be a compact set. Then*

$$\int_A \varphi_{it\nu}(a) b(a) da \asymp (1+t)^{-r},$$

uniformly for $t \in \mathbb{R}$ and $\nu \in D_{\mathfrak{a}^*}$.

3.3.1 Setup

We fix $b \in C_c^\infty(A)$ with $b(1) > 0$ and we will not indicate the dependence on b in our notations. Take $\nu \in \mathfrak{a}^*$ and let $H_0 \in \mathfrak{a}$ be the corresponding element under the isomorphism given by the Killing form. Define

$$I(H_0) = \int_A \varphi_{i\nu}(a) b(a) da\,.$$

Inserting Harish-Chandra's formula for the spherical function (3.1) yields the oscillatory integral

$$I(H_0) = \int_A \int_K \exp\left(i\phi_{H_0}(a,k)\right) b'(a,k) dk da\,, \tag{3.16}$$

with phase function

$$\phi_{H_0}(a,k) = \langle H_0, H(ka)\rangle \tag{3.17}$$

and with an amplitude function $b' \in C_c^\infty(A \times K)$ satisfying $b'(1,k) = b(1) > 0$. We will determine the critical points of ϕ_{H_0} and obtain Proposition 3.15 as an application of the stationary phase method.

3.3.2 Structure of the critical set

By Lemma 2.8 we have for $H \in \mathfrak{a}$ that

$$(D\phi_{H_0}(\cdot,k))_a(H) = \langle H_0, \mathrm{Ad}_{\kappa(ka)}(H)\rangle\,. \tag{3.18}$$

Recall from (2.24) the set

$$\mathcal{C}(G, H_0) = \{k \in K : \mathrm{Ad}_{k^{-1}}(H_0) \perp \mathfrak{a}\}\,.$$

Lemma 3.19. *When $H_0 \in \mathfrak{a}^{\mathrm{gen}}$, the set of critical points of ϕ_{H_0} is $\{1\} \times \mathcal{C}(G, H_0)$.*

Proof. Assume (a,k) is a critical point of ϕ_{H_0}. By [20, Proposition 5.4], criticality in k is equivalent to

$$k \in Z_K(H_0)M'Z_K(a) = M'Z_K(a)\,,$$

where we have used that $H_0 \in \mathfrak{a}^{\mathrm{reg}}$, so that $Z_K(H_0) = M$. By (3.18) and the definition (2.24), criticality in a is equivalent to $\kappa(ka) \in \mathcal{C}(G,H_0)$. Because $k \in M'Z_K(a)$ we have $\kappa(ka) = k$, so that $k \in \mathcal{C}(G,H_0)$. It remains to show that $a = 1$. If not, we would have $k \in M'L$ with $L = Z_G(a) \in \mathcal{L}$ a Levi subgroup different from G. This contradicts the fact that $\mathcal{C}(G,H_0) \cap M'L = \varnothing$ (Lemma 2.32). □

Remark 3.20. When more generally $H_0 \in \mathfrak{a}^{\mathrm{reg}}$, by being more careful in the proof of Lemma 3.19 one shows that the critical set of ϕ_{H_0} is

$$\bigcup_{L \in \mathcal{L}} \mathrm{Ad}_{M'}(A_L) \times \mathcal{C}(L,H_0)\,,$$

where the sets $\mathcal{C}(L,H_0)$ are defined analogously by

$$\mathcal{C}(L,H_0) = \{k \in K \cap L : \mathrm{Ad}_{k^{-1}}(H_0) \perp \mathfrak{a}\}\,,$$

and only the sets $\mathcal{C}(L,H_0)$ with $H_0 \in \mathfrak{a}^L$ are nonempty.

Recall from (2.33) the map (now dropping the dependence on G in the notation)

$$\begin{aligned} f_{H_0} : K &\to \mathfrak{a} \\ k &\mapsto E_{\mathfrak{a}}(\mathrm{Ad}_{k^{-1}}(H_0))\,, \end{aligned}$$

which has the properties that $\mathcal{C}(G,H_0) = f_{H_0}^{-1}(0)$ and

$$(D\phi_{H_0}(\cdot,k))_a(H) = \langle f_{H_0}(\kappa(ka)), H\rangle\,. \tag{3.21}$$

Remark 3.22. The set $\mathcal{C}(G,H_0)$ has the geometric interpretation that it consists of the $k \in K$ for which the maximal flat $kA \subset G/K$ is orthogonal to H_0 at the base point $k \cdot 1$. The geometric picture will come more fully into its own right in §3.4.

The main properties of $\mathcal{C}(G,H_0)$ are established in §2.

Lemma 3.23. *When $H_0 \in \mathfrak{a}^{\mathrm{gen}}$, the function f_{H_0} is a submersion at the points of $\mathcal{C}(G,H_0)$. The set $\mathcal{C}(G,H_0)$ is a non-empty smooth submanifold of K of codimension* $\dim(A)$ *that varies smoothly with $H_0 \in \mathfrak{a}^{\mathrm{gen}}$.*

Proof. That f_{H_0} is a submersion at $\mathcal{C}(G,H_0)$ is Corollary 2.37. This second sentence is Proposition 2.49. □

3.3.3 Stationary phase

Recall the map $f_{H_0} : K \to \mathfrak{a}$ defined in (2.33).

Lemma 3.24. *For $k \in \mathcal{C}(G, H_0)$ we have*

$$T_k(\mathcal{C}(G, H_0)) = \ker((Df_{H_0})_k) = \{X \in \mathfrak{k} : \mathrm{Ad}_k(X) \perp [H_0, \mathrm{Ad}_k(\mathfrak{a})]\}\,.$$

Proof. The first statement holds because the defining map f_{H_0} for $\mathcal{C}(G, H_0)$ is a submersion on $\mathcal{C}(G, H_0)$ (Lemma 3.23). Indeed, it is a general fact that when a submanifold is the level set of a submersion, then its tangent spaces are the kernels of the differential of the defining map [44, Proposition 5.38]. The differential may be computed using (2.7):

$$(Df_{H_0})_k(X) = -E_{\mathfrak{a}}([X, \mathrm{Ad}_k^{-1}(H_0)])\,.$$

Therefore $X \in \ker((Df_{H_0})_k)$ if and only if $[X, \mathrm{Ad}_k^{-1}(H_0)] \perp \mathfrak{a}$. Using associativity and Ad_k-invariance of the Killing form, this is equivalent to $\mathrm{Ad}_k(X) \perp [H_0, \mathrm{Ad}_k(\mathfrak{a})]$. □

The first equality in Lemma 3.24 will be used in the following lemma. The second will be used in the proof of Lemma 3.66.

When $M \subset N$ are Riemannian manifolds and $m \in M$, a symmetric bilinear form on T_mN is called transversely nondegenerate to M if its radical is contained in T_mM. Recall the phase function ϕ_{H_0} defined by (3.17).

Lemma 3.25. *When $H_0 \in \mathfrak{a}^{\mathrm{gen}}$, the Hessian of ϕ_{H_0} is transversely nondegenerate to the critical set of ϕ_{H_0}. Moreover, its signature on the critical set equals $(n_0, n_+, n_-) = (\dim(K) - r, r, r)$, where $r = \dim(A)$.*

Proof. The Hessian of ϕ_{H_0} at a critical point (a, k) is given simply by

$$(X, Y) \mapsto L_X L_Y \phi_{H_0}(a, k)\,.$$

By Lemma 3.19 the critical set of ϕ_{H_0} is $\{1\} \times \mathcal{C}(G, H_0)$. Let $(1, k)$ be a critical point of ϕ_{H_0}. Let $L : \mathfrak{a} \oplus \mathfrak{k} \to \mathfrak{a} \oplus \mathfrak{k}$ be the unique and self-adjoint linear map such that

$$(\mathrm{Hess}_{(1,k)}\, \phi_{H_0})(X, Y) = \langle LX, Y \rangle_\theta$$

for all $X, Y \in \mathfrak{a} \oplus \mathfrak{k}$. We must show that $\ker L \subset \{0\} \oplus T_k\mathcal{C}(G, H_0)$. Write

$$L = \begin{pmatrix} L_{\mathfrak{a}\mathfrak{a}} & L_{\mathfrak{a}\mathfrak{k}} \\ L_{\mathfrak{a}\mathfrak{k}}^* & L_{\mathfrak{k}\mathfrak{k}} \end{pmatrix}$$

relative to the natural decomposition of $\mathfrak{a} \oplus \mathfrak{k}$. Because $\phi_{H_0}(1, k) = 0$ for all $k \in K$, it is clear that $L_{\mathfrak{k}\mathfrak{k}} = 0$. To compute $L_{\mathfrak{a}\mathfrak{k}} : \mathfrak{k} \to \mathfrak{a}$, let $X \in \mathfrak{k}$ and $H \in \mathfrak{a}$. From (3.21) we have $L_H \phi_{H_0}(1, k) = \langle f_{H_0}(k), H \rangle$. Therefore

$$\begin{aligned} \mathrm{Hess}_{(1,k)}\, \phi_{H_0}(X, H) &= L_{XH} \phi_{H_0}(a, k) \\ &= L_X \langle f_{H_0}(k), H \rangle \\ &= \langle (Df_{H_0})_k(X), H \rangle\,, \end{aligned}$$

so that $L_{\mathfrak{a}\mathfrak{k}} = (Df_{H_0})_k$. From Lemma 3.23 it follows that $L_{\mathfrak{a}\mathfrak{k}}$ is surjective, so that the adjoint $L^*_{\mathfrak{a}\mathfrak{k}}$ is injective. This implies

$$\ker L = \{0\} \oplus \ker L_{\mathfrak{a}\mathfrak{k}} = \{0\} \oplus \ker((Df_{H_0})_k) = \{0\} \oplus T_k\mathcal{C}(G, H_0)\,,$$

where the last equality is Lemma 3.24. This proves the first part of the statement.

We can compute the signature at a critical point $(1, k)$ as follows. Let V be a complement of $T_k\mathcal{C}(G, H_0)$ in $\mathfrak{k}$. Relative to the decomposition $\mathfrak{a} \oplus V \oplus T_k\mathcal{C}(G, H_0)$, the map L has the form

$$L = \begin{pmatrix} L_{\mathfrak{a}\mathfrak{a}} & L_{\mathfrak{a}V} & 0 \\ L^*_{\mathfrak{a}V} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

with $L_{\mathfrak{a}V}$ invertible, because $\ker(L_{\mathfrak{a}k}) = T_k\mathcal{C}(G, H_0)$. A self-adjoint map of this form has signature $(n_0, n_+, n_-) = (\dim(\mathcal{C}(G, H_0)), r, r)$. Indeed, to show this we must show that a real symmetric $2r \times 2r$ block matrix of the form

$$M = \begin{pmatrix} A & B \\ B^T & 0 \end{pmatrix}$$

with B invertible, has signature (r, r). Doing as if the blocks were numbers, define

$$\Lambda_\pm = \frac{1}{2}(A \pm \sqrt{A^2 + 4BB^T})\,,$$

where the square root is the symmetric positive definite one. Because "square roots are monotone on symmetric positive definite matrices", Λ_+ is positive definite and Λ_- is negative definite. Now one can either write down explicit matrices to show that M is congruent to $\mathrm{diag}(\Lambda_+, \Lambda_-)$, or observe that if $(v_i^\pm)$ are bases of eigenvectors for $\Lambda_\pm$, then the vectors $(\Lambda_+ v_i^+, B^T v_i^+)$ and $(\Lambda_- v_i^-, B^T v_i^-)$ form a basis of eigenvectors for M, with the same eigenvalues as those of $\Lambda_\pm$. □

We are ready to prove Proposition 3.15. We use the stationary phase theorem with critical manifolds of positive dimension, which we now recall.

Theorem 3.26. *[15, Théorème 4.1] Let M be a smooth Riemannian manifold, $b \in C_c^\infty(M)$ and $\phi \in C^\infty(X)$ real-valued. Assume that the set of critical points of ϕ intersects the support of b in a smooth closed submanifold $W \subset M$ of codimension d, and that the Hessian of ϕ is transversely nondegenerate to W. Denote for $w \in W$ by $\mathrm{Hess}_{w,\perp}\phi$ the restriction of the Hessian to the orthogonal complement of T_wW. Assume that ϕ is constant on W and that $\mathrm{Hess}_\perp \phi$ has constant signature (n_+, n_-) on W. Then as $t \to +\infty$,*

$$\int_M e^{it\phi(x)} b(x)dx = \left(\frac{2\pi}{t}\right)^{d/2} e^{itf(W) + \pi i(n_+ - n_-)/4} \cdot \int_W \frac{b(w)}{|\det(\mathrm{Hess}_{w,\perp}\phi)|^{1/2}} dw + O(t^{-d/2-1})\,,$$

where the integral over W is with respect to the induced metric.

3.3.4 Proof of Proposition 3.15

Proof of Proposition 3.15. Take $\nu_0 \in (\mathfrak{a}^*)^{\mathrm{gen}}$ and let $H_0 \in \mathfrak{a}^{\mathrm{gen}}$ correspond to it via the identification $\mathfrak{a} \cong \mathfrak{a}^*$ given by the Killing form. Via the reduction in the beginning of §3.3 we must bound the oscillatory integrals $I(tH_0)$ given by (3.16):

$$I(tH_0) = \int_A \int_K \exp\left(it\phi_{H_0}(a,k)\right) b'(a,k) dk da\,.$$

By Lemma 3.25 the Hessian of ϕ_{H_0} has signature $(n_+, n_-) = (r, r)$ transversely to its critical set, which has codimension $2\dim(A) = 2r$ by Lemma 3.19 and Lemma 3.23, and ϕ_{H_0} takes the value 0 there because $H(K) = 0$. The stationary phase theorem implies

$$I(tH_0) = t^{-r} \cdot (2\pi)^r e^{\pi i(n_+ - n_-)/4} \int_{\mathcal{C}(G,H_0)} \frac{b'(1,k)}{\sqrt{|\det(\mathbf{H}_k)|}} dk + O(t^{-r-1})$$

as $t \to +\infty$, where $\mathbf{H}_k = \mathrm{Hess}_{(1,k),\perp}\, \phi_{H_0}$ is the Hessian of ϕ_{H_0} restricted to the orthogonal complement of the tangent space $T_{(1,k)}(\{1\} \times \mathcal{C}(G, H_0))$, the determinant is taken in an orthonormal basis and the integral is with respect to the induced metric. Moreover, this bound is uniform for H_0 in compact subsets of $\mathfrak{a}^{\mathrm{gen}}$ because of the smooth dependence in Lemma 3.23. We have $b'(1,k) > 0$ (see the beginning of §3.3), so that the constant in the main term is strictly positive. □

Remark 3.27. A possible way to remove the condition $\nu \in \mathfrak{a}^{\mathrm{gen}}$ in Proposition 3.15 (or at least to replace $\mathfrak{a}^{\mathrm{gen}}$ by $\mathfrak{a}^{\mathrm{reg}}$) would be to use the more precise statement about the critical set in Remark 3.20 and to resolve the singularities of f_{H_0} and lift the integration in (3.16) to a blowup of $\mathfrak{a} \times A \times K$ (the $\mathfrak{a}$-factor corresponding to H_0).

3.4 Bounds for orbital integrals

In this section we prove the following proposition. The notations for Lie groups and Lie algebras are as in §2.1.1.

Proposition 3.28. *Let G be semisimple and $b \in C_c^\infty(A \times A)$. Let $D_{\mathfrak{a}^*} \subset (\mathfrak{a}^*)^{\mathrm{gen}}$ be compact and $D_G \subset G$ be compact. Then there exist $\delta > 0$ and $N > 0$ such that*

$$\int_{A\times A} \varphi_{it\nu}(a_1^{-1} g a_2) b(a_1, a_2) da_1 da_2$$
$$\ll (1+t)^{-r} \cdot \left(1 + t \cdot d\left(g, \bigcup_{L \in \mathcal{L} - \{G\}} M'L\right)^N\right)^{-\delta},$$

uniformly for $t \in \mathbb{R}$, $\nu \in D_{\mathfrak{a}^*}$ and $g \in D_G$.

Remark 3.29. The uniformity in g in the above result is not needed in the proof of Theorem 4, because only finitely many $\gamma \in \Gamma$ contribute to the sum in Lemma 3.8. We will need this dependence in the proof of Theorem 3.

3.4.1 Setup and phase functions

We fix $b \in C_c^\infty(A \times A)$ and we will not incorporate it in the notations. Take $\nu \in \mathfrak{a}^*$ and let $H_0 \in \mathfrak{a}$ be the corresponding element under the isomorphism given by the Killing form. For $g \in G$ define

$$J(H_0, g) = \int_{A \times A} \varphi_{i\nu}(a_1^{-1} g a_2) b(a_1, a_2) da_1 da_2 \,. \tag{3.30}$$

By invariance of $\varphi_{i\nu}$ under the action of the Weyl group on ν, we have for $e \in M'$ that

$$J(H_0, g) = J(\mathrm{Ad}_e(H_0), g) \,. \tag{3.31}$$

It is therefore no restriction to assume that H_0 lies in a Weyl chamber of our choice.

Inserting Harish-Chandra's formula for the spherical function (3.1) yields the oscillatory integral

$$J(H_0, g) = \int_{A \times A} \int_K \exp\left(i \widetilde{\phi}_{H_0, g}(a_1, a_2, k)\right) b_g(a_1, a_2, k) dk da_1 da_2 \,,$$

with phase function

$$\widetilde{\phi}_{H_0, g}(a_1, a_2, k) = \langle H_0, H(k a_1^{-1} g a_2) \rangle \tag{3.32}$$

and with amplitude $b_g \in C_c^\infty(A \times A \times K)$ depending smoothly on g and with support bounded independently of g, which incorporates the real exponential factor in (3.1). Following [47], we now bring this in a form that makes the A-derivatives more manageable.

For $h \in G$ define the map

$$\begin{aligned} \Theta_h : K &\to K \\ k &\mapsto \kappa(kh) \,. \end{aligned} \tag{3.33}$$

By smoothness of the Iwasawa decomposition it is a smooth map, with smooth inverse $\Theta_{h^{-1}}$, and therefore a diffeomorphism. For $k \in K$ and $y, z \in G$ we have that (see [47, Lemma 6.2])

$$H(k y^{-1} z) = H(\Theta_{y^{-1}}(k) z) - H(\Theta_{y^{-1}}(k) y) \,.$$

Applying this with $y = a_1$ and $z = g a_2$ and making the change of variables $k \leftarrow \Theta_{a_1}(k)$ gives

$$J(H_0, g) = \int_K \int_{A \times A} \exp(i \phi_{H_0, g}(a_1, a_2, k)) b'_g(a_1, a_2, k) da_1 da_2 dk \tag{3.34}$$

with phase function

$$\begin{aligned}\phi_{H_0,g}(a_1,a_2,k) &:= \widetilde{\phi}_{H_0,g}(a_1,a_2,\Theta_{a_1}(k)) \\ &= \langle H_0, H(kga_2)\rangle - \langle H_0, H(ka_1)\rangle\end{aligned} \tag{3.35}$$

and some amplitude $b'_g \in C_c^\infty(A\times A\times K)$ depending smoothly on g.

Remark 3.36. The expression (3.32) will be useful when computing derivatives of $\phi_{H_0,g}$ with respect to k, and (3.35) will be useful when computing derivatives with respect to a_1 and a_2.

The expression (3.35) separates the variables a_1 and a_2. Our strategy, inspired by [47], is to first apply the stationary phase theorem in the variables a_1 and a_2, leaving us with an oscillatory integral over K, and then to apply the van der Corput lemma to this integral.

3.4.2 Extremal points on maximal flats

In view of equation (3.35), we are naturally led to study the critical points of the 'height' functions

$$\begin{aligned}h_{H_0,g} : A &\to \mathbb{R} \\ a &\mapsto \langle H_0, H(ga)\rangle\end{aligned}$$

with $g\in G$, which allow us to write

$$\phi_{H_0,g}(a_1,a_2,k) = h_{H_0,kg}(a_2) - h_{H_0,k}(a_1)\,. \tag{3.37}$$

The critical points of $h_{H_0,g}$ are studied in §2. Many of the results concerning them require that $H_0\in\mathfrak{a}^{\mathrm{gen},+}$, so that this assumption is propagated throughout most of the analysis in this section. We summarize the results as follows. Recall the set $\mathcal{C}(G,H_0)\subset K$ defined in (2.24).

Lemma 3.38. *Let $H_0\in\mathfrak{a}^{\mathrm{gen},+}$. Then $h_{H_0,g}$ has at most one critical point. A critical point $a\in A$ is characterized by the condition that $\kappa(ga)\in\mathcal{C}(G,H_0)$. The Hessian at a critical point is negative definite and (as a quadratic form) given by*

$$\begin{aligned}\mathfrak{a} &\to \mathbb{R} \\ H &\mapsto \langle [H_0,\mathrm{Ad}_c(H)], E_{\mathfrak{n}}(\mathrm{Ad}_c(H))\rangle\ ,\end{aligned}$$

where $c=\kappa(ga)$.

Proof. The uniqueness and the fact that the Hessian is negative definite are contained in Theorem 6. The characterization is Lemma 2.25. The Hessian is computed in Proposition 2.47 when $g\in K$, but it does not depend on the triangular part of g by (2.27). □

We require some additional facts about the dependence of the critical point of $h_{H_0,g}$ on the parameters H_0 and g. Take $H_0 \in \mathfrak{a}^{\mathrm{gen},+}$. Define $\mathcal{R}_{H_0} \subset G$ to be the set of elements g for which the function $h_{H_0,g}$ has a critical point. Define also

$$\mathcal{R} = \bigcup_{H_0 \in \mathfrak{a}^{\mathrm{gen},+}} \{H_0\} \times \mathcal{R}_{H_0} \,. \tag{3.39}$$

When $g \in \mathcal{R}_{H_0}$, by Lemma 3.38 there is a unique critical point of $h_{H_0,g}$. Define the function

$$\xi_{H_0} : \mathcal{R}_{H_0} \to A \tag{3.40}$$

that sends g to the critical point of $h_{H_0,g}$, and define

$$\begin{aligned} \xi : \mathcal{R} &\to A \\ (H_0, g) &\mapsto \xi_{H_0}(g) \,. \end{aligned}$$

Define also the $\mathcal{C}$-projection

$$\begin{aligned} c_{H_0} : \mathcal{R}_{H_0} &\to \mathcal{C}(G, H_0) \\ g &\mapsto \kappa(g\xi_{H_0}(g)) \end{aligned}$$

which is guaranteed to take values in $\mathcal{C}(G, H_0)$ by Lemma 3.38.

Lemma 3.41. *The set $\mathcal{R} \subset \mathfrak{a} \times G$ is open, and $\xi_{H_0}(g)$ and $c_{H_0}(g)$ are real analytic in $(H_0, g) \in \mathcal{R}$.*

Proof. Take $(H_0, g) \in \mathcal{R}$. By Lemma 3.38 the critical points of $h_{H_0,g}$ are nondegenerate, which we may reformulate by saying that the map

$$\begin{aligned} A &\to \mathfrak{a}^* \\ a &\mapsto (Dh_{H_0,g})_a \end{aligned}$$

has invertible differential at the level set above 0, which is the singleton $\{\xi_{H_0}(g)\}$. By the implicit function theorem applied to this real analytic map with parameter $(H_0, g) \in \mathfrak{a}^{\mathrm{gen},+} \times G$, it follows that $\mathcal{R}$ is open in $\mathfrak{a}^{\mathrm{gen},+} \times G$, and that $\xi_{H_0}(g)$ is real analytic in (H_0, g). Consequently, $c_{H_0}(g)$ is also real analytic in (H_0, g). □

The following coordinate system for $\mathcal{R}_{H_0} \cap K$ will be useful when dealing with expressions involving the ξ_{H_0}.

Lemma 3.42. *Let $H_0 \in \mathfrak{a}^{\mathrm{gen},+}$. The map*

$$\begin{aligned} \mathcal{C}(G, H_0) \times A &\to \mathcal{R}_{H_0} \cap K \\ (c, a) &\mapsto \kappa(ca^{-1}) \end{aligned}$$

is a real analytic isomorphism whose inverse is

$$k \mapsto (c_{H_0}(k), \xi_{H_0}(k)) \,.$$

Proof. The two maps are real analytic by Lemma 3.41, and the fact that they are mutual inverses follows from the definitions of $\xi_{H_0}(k)$ and $c_{H_0}(k)$. □

Example 3.43. When $G = \mathrm{PSL}(2,\mathbb{R})$ with the standard choice of Iwasawa decomposition, Lemma 3.42 can be visualized as follows. Identify G with the unit tangent bundle of $\mathbb{H}$ and K-projections with unit tangent vectors. We have $\mathcal{R}_{H_0} \cap K = K - M'$, which has two connected components, corresponding to the directions pointing east or west. (The north and south directions are excluded as they come from M'.) Accordingly, the set $\mathcal{C}(G, H_0)$ has two points:

$$c_\pm = \begin{pmatrix} \cos(\pi/4) & \pm\sin(\pi/4) \\ \mp\sin(\pi/4) & \cos(\pi/4) \end{pmatrix}.$$

When a runs through A, the direction of c_+a runs through all directions pointing east, and the direction of c_-a runs trough all directions pointing west.

3.4.3 Reduction to an integral over K

We seek to apply the stationary phase theorem to evaluate the inner integral in (3.34). The main result of this subsection is Proposition 3.47. Given the results of §3.4.2, for which most of the work was done in §2, the proof is parallel to that of Lemma 1.40 when $G = \mathrm{PSL}_2(\mathbb{R})$. Some care must be taken to obtain uniformity in H_0.

Define the "parameter space"

$$\mathcal{P} = \mathfrak{a}^{\mathrm{gen},+} \times G \times K$$

and let $\mathcal{R}' \subset \mathcal{P}$ be the set of parameters (H_0, g, k) for which the function $\phi_{H_0,g}(\cdot,\cdot,k)$ has a critical point. Denote by $\mathcal{R}'_{H_0,g} \subset K$ the fiber of $\mathcal{R}'$ above (H_0, g).

Lemma 3.44. *The set $\mathcal{R}'$ is open in $\mathcal{P}$. When $(H_0, g, k) \in \mathcal{R}'$, the function $\phi_{H_0,g}(\cdot,\cdot,k)$ has a unique critical point (a_1, a_2) given by $(\xi_{H_0}(k), \xi_{H_0}(kg))$.*

Proof. Recall from (3.37) that

$$\phi_{H_0,g}(a_1, a_2, k) = h_{H_0,kg}(a_2) - h_{H_0,k}(a_1),$$

so that (a_1, a_2) is a critical point of $\phi_{H_0,g}(\cdot,\cdot,k)$ if and only if a_1 is a critical point of $h_{H_0,k}$ and a_2 is a critical point of $h_{H_0,kg}$. Lemma 3.38 gives the uniqueness, and the last part is the definition of ξ_{H_0} (see (3.40)). We show that $\mathcal{R}'$ is open. Let $\mathcal{R}$ be as in (3.39); it is open in $\mathfrak{a} \times G$ by Lemma 3.41, and it suffices to note that $\mathcal{R}'$ is the preimage of $\mathcal{R} \times \mathcal{R}$ under the continuous map

$$\begin{aligned} \mathcal{P} &\to (\mathfrak{a} \times G) \times (\mathfrak{a} \times G) \\ (H_0, g, k) &\mapsto ((H_0, k), (H_0, kg)). \end{aligned}$$

□

Define a function on K by

$$\psi_{H_0,g}(k) = \begin{cases} \phi_{H_0,g}(\xi_{H_0}(k), \xi_{H_0}(kg), k) & \text{when } k \in \mathcal{R}'_{H_0,g}\,, \\ 0 & \text{otherwise.} \end{cases} \tag{3.45}$$

When $(H_0, g, k) \in \mathcal{R}'$, denote by $d_{H_0,g}(k)$ the Hessian determinant of the function $\phi_{H_0,g}(\cdot,\cdot,k)$ (for fixed k) at its unique critical point. Let b'_g be as in (3.34) and define

$$b''_{H_0,g}(k) = \begin{cases} b'_g(\xi_{H_0}(k), \xi_{H_0}(kg), k) \frac{(2\pi)^r}{\sqrt{|d_{H_0,g}(k)|}} & \text{when } k \in \mathcal{R}'_{H_0,g}\,, \\ 0 & \text{otherwise,} \end{cases}$$

and call $b'' : \mathcal{P} \to \mathbb{C}$ the corresponding function of (H_0, g, k).

Lemma 3.46. *The function b'' is smooth and* $\operatorname{supp}(b'') \subset \mathcal{R}'$.

Proof. Roughly speaking, this is because $\xi_{H_0}(g)$ diverges as (H_0, g) approaches the boundary of $\mathcal{R}$, and the argument is just like in Lemma 1.38. We must only show that every point of $\mathcal{P} - \mathcal{R}'$ has a neighborhood on which b'' is zero. Suppose not, then there exists a sequence $(H_{0,n}, g_n, k_n) \in \mathcal{R}'$ on which b'' is nonzero and which converges to a point (H_0, g, k) of $\mathcal{P} - \mathcal{R}'$. Because $b''(H_{0,n}, g_n, k_n) \neq 0$ and b'_g has support bounded independently of g (even independent of g altogether) the sequences $\xi_{H_{0,n}}(k_n)$ and $\xi_{H_{0,n}}(k_n g_n)$ are bounded. We may then extract a subsequence on which $\xi_{H_{0,n}}(k_n)$ and $\xi_{H_{0,n}}(k_n g_n)$ converge, say to $\xi_1, \xi_2 \in A$. By continuity we then have

$$(Dh_{H_0,k})_{\xi_1} = \lim_{n\to\infty} (Dh_{H_{0,n},k_n})_{\xi_{H_{0,n}}(k_n)} = 0\,,$$

and similarly for ξ_2, so that $(H_0, g, k) \in \mathcal{R}'$. This is a contradiction. □

Proposition 3.47. *Let $D_{\mathfrak{a}} \subset \mathfrak{a}^{\mathrm{gen},+}$ be compact and $D_G \subset G$ be compact. Define $J(H_0, g)$ by (3.30). Then*

$$J(tH_0, g) = t^{-r} \int_K e^{it\psi_{H_0,g}(k)} b''_{H_0,g}(k) dk + O(t^{-r-1})$$

as $t \to +\infty$, uniformly for $H_0 \in D_{\mathfrak{a}}$ and $g \in D_G$.

Proof. We prove a uniform asymptotic for the double A-integral in (3.34), which we then integrate over K. Let $\mathcal{P}_0 = \operatorname{supp}(b'') \cap (D_{\mathfrak{a}} \times D_G \times K)$. It is closed in $\mathcal{P}$ and contained in $\mathcal{R}'$ by Lemma 3.46. We distinguish two cases depending on where (H_0, g, k) lies. When $(H_0, g, k) \in \mathcal{R}'$, the phase function $\phi_{H_0,g}(\cdot,\cdot,k)$ has a unique and nondegenerate critical point $(\xi_{H_0}(k), \xi_{H_0}(kg))$, where the Hessian has signature (r, r) by (3.37) and Lemma 3.38. The stationary phase theorem [69, §VIII.2] implies that

$$\begin{aligned} \int_{A\times A} e^{it\phi_{H_0,g}(a_1,a_2,k)} b'_{H_0,g}(a_1, a_2, k) da_1 da_2 = t^{-r} e^{it\psi_{H_0,g}(k)} b''_{H_0,g}(k) \\ + O(t^{-r-1}) \end{aligned} \tag{3.48}$$

as $t \to +\infty$, uniformly for (H_0, g, k) in compact subsets of $\mathcal{R}'$. When $(H_0, g, k) \in \mathcal{P} - \mathcal{P}_0$, the phase function $\phi_{H_0,g}(\cdot, \cdot, k)$ has no critical points in the support of $b'_{H_0,g}$, and the Van der Corput lemma [69, §VIII.2] implies that

$$\int_{A \times A} e^{it\phi_{H_0,g}(a_1,a_2,k)} b'_{H_0,g}(a_1, a_2, k) da_1 da_2 \ll_N t^{-N}$$

as $t \to +\infty$, uniformly for (H_0, g, k) in compact subsets of $\mathcal{P} - \mathcal{P}_0$. Because b'' is by definition zero on $\mathcal{P} - \mathcal{P}_0$, the estimate (3.48) also holds in this case, uniformly in compact subsets of $\mathcal{P} - \mathcal{P}_0$. Because $\mathcal{P} - \mathcal{P}_0$ and $\mathcal{R}'$ are open in $\mathcal{P}$ (Lemma 3.44) and they cover $\mathcal{P}$, we may find compact subsets of $\mathcal{R}'$ and $\mathcal{P} - \mathcal{P}_0$ that cover the compact set $\mathcal{P} \cap (D_{\mathfrak{a}} \times D_G \times K)$. Therefore (3.48) holds uniformly for all $(H_0, g, k) \in \mathcal{P} \cap (D_{\mathfrak{a}} \times D_G \times K)$. The statement follows then by integrating (3.48) over K. □

3.4.4 Critical points of $\psi_{H_0,g}$

This subsection is concerned with the critical points of $\psi_{H_0,g}$ (defined in (3.45)) when g lies in the dense open set $G - \bigcup_{L \in \mathcal{L} - \{G\}} M'L$. The main result is Proposition 3.49, which will imply Proposition 3.28.

Proposition 3.49. *Let $H_0 \in \mathfrak{a}^{\mathrm{gen},+}$ and $g \in G - \bigcup_{L \in \mathcal{L} - \{G\}} M'L$. Then $\psi_{H_0,g}$ is nowhere locally constant in $\mathcal{R}'_{H_0,g}$.*

The proof of Proposition 3.49 is at the end of this subsection.

Remark 3.50. The proof differs from the proof in [47] for $\mathrm{PSL}_2(\mathbb{R})$. Namely, we do not prove that the Hessian of $\psi_{H_0,g}$ is nondegenerate at critical points. In fact, to obtain a useful expression for the Hessian of $\psi_{H_0,g}$ in order to generalize to proof for $\mathrm{PSL}_2(\mathbb{R})$, we must know something about the derivatives of the ξ_{H_0} (see (3.40)), which are determined implicitly by their image under an injective linear map $\mathfrak{a} \to \mathfrak{k}$ involving the Hessian of $h_{H_0,g}$. When $\dim(\mathfrak{a}) = \dim(\mathfrak{k}) = 1$, all of this can be made relatively explicit, but in general it seems very hard to say anything about the $D\xi_{H_0}$. As a replacement for this computation, we will use the change of variables from Lemma 3.42, through which one of the occurrences of ξ_{H_0} simplifies greatly. The second step in the proof is to eliminate the other occurrence of $\xi_{H_0}(g) \in A$, by using the fact that its adjoint action on $\mathfrak{a}$ is trivial; see Lemma 3.62.

We first deduce Proposition 3.28 using very general principles.

Proof of Proposition 3.28 assuming Proposition 3.49. By (3.31) it is no restriction to assume that $H_0 \in \mathfrak{a}^+$. Proposition 3.47 then reduces the statement to showing that for some $N > 0$ and $\delta > 0$,

$$\int_K e^{it\psi_{H_0,g}(k)} b''_{H_0,g}(k) dk \ll \left(1 + t \cdot d\left(g, \bigcup_{L \in \mathcal{L} - \{G\}} M'L\right)^N\right)^{-\delta}, \tag{3.51}$$

with uniformity in H_0. When g lies at positive distance from $\bigcup_{L\in\mathcal{L}-\{G\}} M'L$, this follows directly from the Van der Corput lemma [69, §VIII.2, Proposition 5], because by Proposition 3.49 there is no $k \in \mathrm{supp}(b''_{H_0,g})$ at which $\psi_{H_0,g}$ is somewhere constant, meaning that some higher k-derivative is bounded away from zero around every $k \in \mathrm{supp}(b''_{H_0,g})$.

To get control over the dependence on g, consider for every $n \geq 1$ the real analytic map

$$\begin{aligned} D_n : \mathcal{R}' &\to (\mathfrak{k}^*)^{\otimes n} \\ (H_0, g, k) &\mapsto ((D^{(n)}\psi_{H_0,g})_k \end{aligned}$$

where we denote $D^{(n)}$ for the higher derivatives. Call $Z \subset \mathcal{R}'$ the joint zero locus of the D_n.

Proposition 3.49 says that

$$Z \subset \mathfrak{a} \times \left(\bigcup_{L\in\mathcal{L}-\{G\}} M'L \right) \times K\,.$$

Locally around a fixed (H_0, g, k) we may embed $\mathcal{R}'$ as a real submanifold of a complex manifold Ω on which $\psi_{H_0,g}(k)$ is complex analytic in all three variables. The maps D_n are derivatives of $\psi_{H_0,g}(k)$ and therefore also extend analytically to Ω. By [77, §3.9 Theorem 9 C], their joint zero locus is locally the zero locus of only finitely many D_n, say of $D_1, \ldots, D_J$, on a neighborhood $U \subset \mathcal{R}'$ of (H_0, g, k). Consider now the Taylor polynomial map

$$T_J : U \to (\mathfrak{k}^*)^{\otimes 1} \oplus \cdots \oplus (\mathfrak{k}^*)^{\otimes J}$$

obtained by pairing $D_1, \ldots, D_J$, and equip the right-hand side with a fixed norm. By Lojasiewicz's inequality, for every compact subset $V \subset U$ there exists $N > 0$ such that for $(H'_0, g', k') \in V$ we have

$$\left\| T_J(H'_0, g', k') \right\| \gg d\left((H_0, g, k), Z\right)^N \gg d\left(g, \bigcup_{L\in\mathcal{L}-\{G\}} M'L \right)^N .$$

Because such a bound holds locally in $\mathcal{R}'$, it holds on compact subsets of $\mathcal{R}'$ but with possibly bigger values of J and N. This gives a lower bound for a k-derivative order at most J for $\psi_{H_0,h}$. To apply this, take $D \subset \mathcal{R}'$ to be any compact neighborhood of $\mathrm{supp}(b'') \cap (D_{\mathfrak{a}} \times D_G \times K)$. This is possible by Lemma 3.46 and because $\mathcal{R}'$ is open in $\mathcal{P}$. The bound (3.51) then follows from the Van der Corput lemma [69, §VIII.2, Proposition 5], with the corresponding value of N and with $\delta = 1/J$. □

The following lemma is an immediate application of the chain rule.

Lemma 3.52. *Let $H_0 \in \mathfrak{a}^{\mathrm{gen},+}$ and $g \in G$. When $k \in \mathcal{R}'_{H_0,g}$, we have that k is a critical point of $\psi_{H_0,g}$ if and only if $(\xi_{H_0}(k), \xi_{H_0}(kg), k)$ is a critical point of $\phi_{H_0,g}$.* □

For $g \in G$, define $\Theta_g : K \to K$ by (3.33).

Lemma 3.53. *Let $H_0 \in \mathfrak{a}^{\mathrm{reg}}$ and $g \in G$. A point (a_1, a_2, k) is a critical point of $\phi_{H_0,g}$ if and only if*

$$\kappa(ka_1) \in \mathcal{C}(G, H_0)\,, \tag{3.54}$$
$$\kappa(kga_2) \in \mathcal{C}(G, H_0)\,, \tag{3.55}$$
$$n(\Theta_{a_1}(k)a_1^{-1}ga_2) = 1\,. \tag{3.56}$$

Proof. Let (a_1, a_2, k) be a critical point of $\phi_{H_0,g}$. By (3.37) and Lemma 3.38, criticality in a_1 and a_2 is equivalent to (3.54) and (3.55). In view of (3.35), k is a critical point of $\phi_{H_0,g}(a_1, a_2, \cdot)$ if and only if $\Theta_{a_1}(k)$ is a critical point of $\widetilde{\phi}_{H_0,g}(a_1, a_2, \cdot)$. Using the expression (3.32), by [20, Lemma 5.3] this in turn is equivalent to

$$\Theta_{a_1}(k)a_1^{-1}ga_2 \in Z_N(H_0)AK = AK\,,$$

where we have used that $H_0 \in \mathfrak{a}^{\mathrm{reg}}$. This is (3.56). □

Remark 3.57. The condition (3.56) can also be written as $n(ka_1) = n(kga_2)$. Indeed, writing $ka_1 = n'a'k'$, condition (3.56) says that

$$k'a_1^{-1}ga_2 = a'^{-1}n'^{-1}kga_2 \in AK\,.$$

We may write this as $kga_2 \in n'AK$, which is what we claimed. This is analogous to the criticality condition in §1.5.2, where it can be interpreted geometrically as saying that two geodesics in the upper half plane are concentric.

At this point it is helpful to make the change of variables given by the diffeomorphism in Proposition 3.42. Define therefore

$$\begin{aligned} \widetilde{\psi}_{H_0,g} : \mathcal{C}(G, H_0) \times A &\to \mathbb{R} \\ (c, a) &\mapsto \psi_{H_0,g}(\kappa(ca^{-1}))\,. \end{aligned} \tag{3.58}$$

For $H_0 \in \mathfrak{a}$ and $g \in G$, define an open subset of $\mathcal{C}(G, H_0) \times A$ by

$$\mathcal{R}''_{H_0,g} = \{(c, a) \in \mathcal{C}(G, H_0) \times A : (H_0, g, \kappa(ca^{-1})) \in \mathcal{R}'\}\,.$$

Thanks to this change of variables, $\widetilde{\psi}_{H_0,g}$ has a more manageable expression. Using that $\xi_{H_0}(\kappa(ca^{-1})) = a$ and using the definitions (3.58), (3.45), (3.35), (3.32) we have for

$(c,a) \in \mathcal{R}''_{H_0,g}$ that

$$\begin{aligned}\widetilde{\psi}_{H_0,g}(c,a) &= \psi_{H_0,g}(\kappa(ca^{-1})) \\ &= \phi_{H_0,g}(a, \xi_{H_0}(\kappa(ca^{-1})g), \kappa(ca^{-1})) \\ &= \widetilde{\phi}_{H_0,g}(a, \xi_{H_0}(\kappa(ca^{-1})g), c) \\ &= \langle H_0, H(ca^{-1}g\xi_{H_0}(\kappa(ca^{-1})g))\rangle \\ &= \langle H_0, H(ca^{-1}g\xi_{H_0}(ca^{-1}g))\rangle\,.\end{aligned}$$

In the last equality we have used that ξ_{H_0} is invariant under left multiplication by NA. For $(c,a) \in \mathcal{R}''_{H_0,g}$ define

$$\gamma_{H_0,g}(c,a) = ca^{-1}g\xi_{H_0}(ca^{-1}g)\,, \tag{3.59}$$

so that $\widetilde{\psi}_{H_0,g}(c,a) = \langle H_0, H(\gamma_{H_0,g}(c,a))\rangle$. By Lemma 3.38 and the definition of ξ_{H_0} (3.40) we have for all $(c,a) \in \mathcal{R}''_{H_0,g}$ that

$$\kappa(\gamma_{H_0,g}(c,a)) \in \mathcal{C}(G,H_0)\,. \tag{3.60}$$

Lemma 3.61. *Let $H_0 \in \mathfrak{a}^{\mathrm{gen},+}$ and $g \in G$. Then $(c,a) \in \mathcal{R}''_{H_0,g}$ is a critical point of $\widetilde{\psi}_{H_0,g}$ if and only if*

$$n(\gamma_{H_0,g}(c,a)) = 1\,.$$

Proof. A point $(c,a) \in \mathcal{R}''_{H_0,g}$ is a critical point of $\widetilde{\psi}_{H_0,g}$ if and only if $\kappa(ca^{-1})$ is a critical point of $\psi_{H_0,g}$. Using Lemma 3.52 and Lemma 3.53, this is seen to be equivalent to $n(\gamma_{H_0,g}(c,a)) = 1$. □

Lemma 3.62. *If $\widetilde{\psi}_{H_0,g}$ is critical at $(c,a) \in \mathcal{R}''_{H_0,g}$, then*

$$\langle H_0, \mathrm{Ad}_{ca^{-1}g}(\mathfrak{a})\rangle = 0\,.$$

Proof. By equation (3.60) and Lemma 3.61 we have that $\gamma_{H_0,g}(c,a) \in A\mathcal{C}(G,H_0)$, so that

$$ca^{-1}g \in A\mathcal{C}(G,H_0)A\,.$$

For $k \in \mathcal{C}(G,H_0)$ we have by definition that $\langle H_0, \mathrm{Ad}_k(\mathfrak{a})\rangle = 0$ (see (2.24)). By Ad_G-invariance of the Killing form, this gives

$$\langle H_0, \mathrm{Ad}_{ca^{-1}g}(\mathfrak{a})\rangle = 0\,,$$

as claimed. □

Remark 3.63. 1. In Lemma 3.62 we write $\langle H_0, \cdot\rangle$ and not $H_0 \perp \cdot$, because $\mathrm{Ad}_{ca^{-1}g}(\mathfrak{a})$ is not in $\mathfrak{p}$ in general.

2. Lemma 3.62 gets rid of the (complicated) function $\xi_{H_0,g}$ by considering the adjoint action on $\mathfrak{a}$. This makes differentiating much easier, at the cost of a (possible) loss of information about $\gamma_{H_0,g}(c,a)$.

Lemma 3.64. *If $\widetilde{\psi}_{H_0,g}$ is constant around a point $(c,a) \in \mathcal{R}''_{H_0,g}$, then*

$$[H_0, \mathrm{Ad}_c(T_c\mathcal{C}(G,H_0))] \perp E_{\mathfrak{p}}(\mathrm{Ad}_{cAg}(\mathfrak{a}))\,,$$

where $E_{\mathfrak{p}} = \frac{1}{2}(\mathrm{id} - \theta)$ denotes the orthogonal projection onto $\mathfrak{p}$.

Proof. By the locally constant hypothesis implies, the conclusion of Lemma 3.62 is true for all nearby c and nearby a. In fact, by analytic continuation it is valid for all $a' \in A$. Differentiating it with respect to c gives

$$\langle H_0, [\mathrm{Ad}_c(T_c\mathcal{C}(G,H_0)), \mathrm{Ad}_{ca'g}(\mathfrak{a})]\rangle = 0\,,$$

for all $a' \in A$. Now we use associativity of the Killing form:

$$\langle [H_0, \mathrm{Ad}_c(T_c\mathcal{C}(G,H_0))], \mathrm{Ad}_{ca'g}(\mathfrak{a})\rangle = 0\,.$$

And finally, we use that the left member lies in $[\mathfrak{p},\mathfrak{k}] = \mathfrak{p}$. □

Lemma 3.65. *For $H_0 \in \mathfrak{a}^{\mathrm{gen}}$ we have that*

$$\mathfrak{m} + \mathrm{Ad}_c(\mathfrak{m}) \subset \mathrm{Ad}_c(T_c\mathcal{C}(G,H_0))\,.$$

Proof. This is easily seen from Lemma 3.24. Associativity of the Killing form gives that $\mathfrak{m} \perp [H_0,\mathfrak{g}]$, which implies that $\mathfrak{m} \subset \mathrm{Ad}_c(T_c\mathcal{C}(G,H_0))$. Associativity also gives that $\mathrm{Ad}_c(\mathfrak{m}) \perp [\mathrm{Ad}_c(\mathfrak{a}),\mathfrak{g}]$, which implies $\mathrm{Ad}_c(\mathfrak{m}) \subset \mathrm{Ad}_c(T_c\mathcal{C}(G,H_0))$. □

Define $\mathfrak{k}^{\perp\mathfrak{m}} = \mathfrak{k} \cap \bigoplus_\alpha \mathfrak{g}_\alpha$ and $\mathfrak{p}^{\perp\mathfrak{a}} = \mathfrak{p} \cap \bigoplus_\alpha \mathfrak{g}_\alpha$. They are the "root space parts" of $\mathfrak{k}$ and $\mathfrak{p}$ respectively. Define $E_{\mathfrak{k}^{\perp\mathfrak{m}}}$ and $E_{\mathfrak{p}^{\perp\mathfrak{a}}}$ to be the orthogonal projections onto these spaces.

The following is nothing but a useful reformulation of Lemma 3.24.

Lemma 3.66. *Let $H_0 \in \mathfrak{a}^{\mathrm{gen}}$ and $c \in \mathcal{C}(G,H_0)$. The orthogonal complement of the subspace $[H_0, \mathrm{Ad}_c(T_c\mathcal{C}(G,H_0))]$ in $\mathfrak{p}$ equals $\mathfrak{a} \oplus \mathrm{Ad}_c(\mathfrak{a})$.*

Proof. Call V the orthogonal complement of $[H_0, \mathrm{Ad}_c(T_c\mathcal{C}(G,H_0))]$ in $\mathfrak{p}$. Certainly $\mathfrak{a} \subset V$, because $\mathfrak{a} \perp [\mathfrak{a},\mathfrak{k}]$. By Lemma 3.24, the space

$$\mathrm{Ad}_c(T_c\mathcal{C}(G,H_0)) \subset \mathfrak{k}$$

is the orthogonal complement of $[H_0, \mathrm{Ad}_c(\mathfrak{a})]$. It follows that $\mathrm{Ad}_c(\mathfrak{a}) \subset V$. Therefore

$$\mathfrak{a} \oplus \mathrm{Ad}_c(\mathfrak{a}) \subset V\,.$$

That this is an equality, follows from dimension considerations. Consider the projection $E_{\mathfrak{k}^{\perp\mathfrak{m}}}(\mathrm{Ad}_c(T_c\mathcal{C}(G,H_0)))$ to $\mathfrak{k}^{\perp\mathfrak{m}}$. Because $\mathrm{Ad}_c(T_c\mathcal{C}(G,H_0))$ has codimension $\dim(\mathfrak{a})$ in $\mathfrak{k}$ by Lemma 2.49, and because $\mathrm{Ad}_c(T_c\mathcal{C}(G,H_0)) \supset \mathfrak{m}$ by Lemma 3.65, this projection has codimension $\dim(\mathfrak{a})$ in $\mathfrak{k}^{\perp\mathfrak{m}}$. Now the action of $[H_0,\cdot]$ on $\mathfrak{k}^{\perp\mathfrak{m}}$ gives a bijection with $\mathfrak{p}^{\perp\mathfrak{a}}$ because H_0 is regular. Therefore $[H_0, \mathrm{Ad}_c(T_c\mathcal{C}(G,H_0))]$ has codimension $\dim(\mathfrak{a})$ in $\mathfrak{p}^{\perp\mathfrak{a}}$, and the orthogonal complement in this space must be $E_{\mathfrak{p}^{\perp\mathfrak{a}}}(\mathrm{Ad}_c(\mathfrak{a}))$, which has the correct dimension because $\mathrm{Ad}_c(\mathfrak{a})$ intersects trivially with $\mathfrak{a}$ by Lemma 2.32 and Lemma 2.2. □

The idea of the argument is the following. Lemma 3.66 gives the precise orthogonal complement of $[H_0, \mathrm{Ad}_c(T_c\mathcal{C}(G,H_0))]$ in $\mathfrak{p}$, and Lemma 3.64 also gives elements in $\mathfrak{p}$ orthogonal to it, under the assumption that $\widetilde{\psi}_{H_0,g}$ is constant around a point $(c,a) \in \mathcal{R}''_{H_0,g}$. As soon as the latter lemma gives one new element, we have a contradiction. We may also formulate this in the following way.

Lemma 3.67. *If $\widetilde{\psi}_{H_0,g}$ is constant around a point $(c,a) \in \mathcal{R}''_{H_0,g}$, then*

$$E_{\mathfrak{p}}(\mathrm{Ad}_{cAg}(\mathfrak{a})) \subset \mathfrak{a} \oplus \mathrm{Ad}_c(\mathfrak{a})\,.$$

Proof. Immediate from Lemma 3.64 and Lemma 3.66. □

Lemma 3.68. *If $\widetilde{\psi}_{H_0,g}$ is constant around a point $(c,a) \in \mathcal{R}''_{H_0,g}$, then*

$$E_{\mathfrak{p}^{\perp\mathfrak{a}}}(\mathrm{Ad}_{Ag}(\mathfrak{a})) \subset E_{\mathfrak{p}^{\perp\mathfrak{a}}}(\mathrm{Ad}_{c^{-1}}(\mathfrak{a}))\,.$$

Proof. We use Lemma 3.67 and apply $\mathrm{Ad}_{c^{-1}}$, which gives that

$$E_{\mathfrak{p}}(\mathrm{Ad}_{Ag}(\mathfrak{a})) \subset \mathrm{Ad}_{c^{-1}}(\mathfrak{a}) \oplus \mathfrak{a}\,.$$

Here we have used that Ad_K commutes with $E_{\mathfrak{p}}$. Projecting further to $\mathfrak{p}^{\perp\mathfrak{a}}$ we may discard the factor $\mathfrak{a}$ in the right-hand side, and the claim follows. □

If we can show that the left-hand side in Lemma 3.68 is of dimension $\dim(\mathfrak{a})$, then the inclusion becomes an equality; in particular, an inclusion in the other direction. The following lemmas show that it is indeed of that dimension, provided that g is generic.

Lemma 3.69. *Let $X = (X_\alpha) \in \bigoplus_{\alpha\in\Sigma} \mathfrak{g}_\alpha$. The span of $\mathrm{Ad}_A(X)$ is $\bigoplus_{\alpha\in\Sigma} \mathbb{R}X_\alpha$.*

Proof. Clearly $\mathrm{Ad}_A(X) \subset \bigoplus_{\alpha\in\Sigma} \mathbb{R}X_\alpha$.

We must show that the vectors $v_a = (\alpha(a))_{\alpha\in\Sigma} \in \mathbb{R}^\Sigma$ with $a \in A$ span the entire space $\mathbb{R}^\Sigma$. This is equivalent to the statement that the roots of A are linearly independent, and it is a general fact that the characters of an abelian group are linearly independent. More precisely, suppose they don't span everything, then they lie on a hyperplane and there exist $c_\alpha \in \mathbb{R}$, not all zero such that

$$\forall a \in A : \sum_{\alpha\in\Sigma} c_\alpha \alpha(a) = 0\,.$$

This is saying that the $\alpha : A \to \mathbb{R}^\times$ are linearly dependent. □

Lemma 3.70. *If $g \notin \bigcup_{L\in\mathcal{L}-\{G\}} M'L$, then*

$$\dim(\mathrm{span}(E_{\mathfrak{p}^{\perp\mathfrak{a}}}(\mathrm{Ad}_{Ag}(\mathfrak{a})))) \geq \dim(\mathfrak{a})\,.$$

Proof. Let V be the span of $\mathrm{Ad}_{Ag}(\mathfrak{a})$. By Lemma 3.69, it is of the form $V_0 \oplus \bigoplus_{\alpha\in\Sigma} V_\alpha$ with $V_0 \subset \mathfrak{a}+\mathfrak{m}$ and $V_\alpha \subset \mathfrak{g}_\alpha$. Let $S = \{\alpha : V_\alpha \neq 0\}$. We claim that S spans $\mathfrak{a}^*$. Suppose not, then there exists a nonzero $H \in \bigcap_{\alpha\in S}\ker(\alpha)$, meaning that $[H,V]=0$. Then $[H,\mathrm{Ad}_g(\mathfrak{a})]=0$, or equivalently, $[\mathrm{Ad}_{g^{-1}}(H),\mathfrak{a}]=0$. Thus $\mathrm{Ad}_{g^{-1}}(H)\in\mathfrak{m}\oplus\mathfrak{a}$. By Lemma 2.3 this implies that $g\in Z_G(H)M'$, contradicting our hypothesis on g. Therefore S spans $\mathfrak{a}^*$.

In particular, there are at least $\dim(\mathfrak{a})$ positive roots $\alpha\in\Sigma^+$ with $V_\alpha\neq 0$, and in particular with $E_{\mathfrak{p}^{\perp\mathfrak{a}}}(V_\alpha)\neq 0$. Because the projections $E_{\mathfrak{p}^{\perp\mathfrak{a}}}(\mathfrak{g}_\alpha)$ for different $\alpha\in\Sigma^+$ are orthogonal, the projection $E_{\mathfrak{p}^{\perp\mathfrak{a}}}(V)$ has dimension at least $\dim(\mathfrak{a})$. □

Lemma 3.71. *Suppose $\widetilde{\psi}_{H_0,g}$ is constant around a point $(c,a)\in\mathcal{R}''_{H_0,g}$, and suppose $g\notin\bigcup_{L\in\mathcal{L}-\{G\}}M'L$. Then*

$$\mathrm{span}(E_{\mathfrak{p}^{\perp\mathfrak{a}}}(\mathrm{Ad}_{Ag}(\mathfrak{a}))) = E_{\mathfrak{p}^{\perp\mathfrak{a}}}(\mathrm{Ad}_{c^{-1}}(\mathfrak{a}))\,.$$

Proof. Lemma 3.68 says that the left-hand side is contained in the right-hand side. The left-hand side is of dimension at least $\dim(\mathfrak{a})$ by Lemma 3.70, and the right-hand side is of dimension $\dim(\mathfrak{a})$ because $\mathrm{Ad}_c(\mathfrak{a})$ intersects trivially with $\mathfrak{a}$ by Lemma 2.32 and Lemma 2.2. Therefore, equality must hold. □

Proof of Proposition 3.49. Assume $\psi_{H_0,g}$ is constant around some point $k\in\mathcal{R}'_{H_0,g}$. Then $\widetilde{\psi}_{H_0,g}$ (defined in (3.58)) is constant around some point $(c,a)\in\mathcal{R}''_{H_0,g}$.

Because $\mathrm{Ad}_{c^{-1}}(H_0)\perp\mathfrak{a}$ (by definition of $\mathcal{C}(G,H_0)$), Lemma 3.71 gives that

$$\mathrm{Ad}_{c^{-1}}(H_0)\in\mathrm{span}(E_{\mathfrak{p}^{\perp\mathfrak{a}}}(\mathrm{Ad}_{Ag}(\mathfrak{a})))\,.$$

On the other hand, by the hypothesis of being locally constant and by analytic continuation in a, Lemma 3.62 says that

$$\langle \mathrm{Ad}_{c^{-1}}(H_0), \mathrm{Ad}_{Ag}(\mathfrak{a})\rangle = 0\,,$$

so that

$$\mathrm{Ad}_{c^{-1}}(H_0)\perp\mathrm{span}(E_{\mathfrak{p}^{\perp\mathfrak{a}}}(\mathrm{Ad}_{Ag}(\mathfrak{a})))\,.$$

Therefore $\mathrm{Ad}_{c^{-1}}(H_0)=0$, which is a contradiction. □

Remark 3.72. When $g\in M'L$ with $L\in\mathcal{L}$ chosen to be minimal, an argument analogous to that used to prove Lemma 3.70 shows that

$$\dim(\mathrm{span}(E_{\mathfrak{p}^{\perp\mathfrak{a}}}(\mathrm{Ad}_{Ag}(\mathfrak{a}))))\geq\dim(\mathfrak{a}^L)\,.$$

Moreover, the span in the left-hand side is a direct sum of subspaces of the spaces $\mathfrak{p} \cap (\mathfrak{g}_\alpha + \mathfrak{g}_{-\alpha})$. If $\widetilde{\psi}_{H_0,g}$ is constant around a point $(c, a) \in \mathcal{R}''_{H_0,g}$, this gives a very strong restriction on the space $E_{\mathfrak{p}^{\perp_\mathfrak{a}}}(\mathrm{Ad}_{c^{-1}}(\mathfrak{a}))$ (which contains this span), but we have not been able to obtain a contradiction from this, even when exploiting the fact that the inclusion is true for all nearby $c' \in \mathcal{C}(G, H_0)$.

3.5 Preliminaries on algebraic groups

3.5.1 Algebraic groups

Let $\mathbf{G}/\mathbb{Q}$ be a linear connected anisotropic algebraic group which is almost simple over $\mathbb{R}$. Let $\mathbf{H} \subset \mathbf{G}$ be a maximal torus. The base fields of groups are indicated by subscripts if they are not clear from the context. We use the same subscript notation for base change and when E/F is a finite field extension we denote $\mathrm{Res}_{E/F}$ for Weil restriction.

Fix a closed embedding $\rho : \mathbf{G} \to \mathbf{SL}_d$ for some $d > 0$. Equipping $\mathbf{SL}_d$ with the standard schematic structure and taking the schematic closures of $\rho(\mathbf{G})$ and $\rho(\mathbf{H})$ we obtain integral models that we continue to denote by $\mathbf{G}$ and $\mathbf{H}$.

Remark 3.73. Eventually $\mathbf{G}$ and $\mathbf{H}$ will be as in Theorem 3. That is, $\mathbf{G}$ an anisotropic form of $\mathbf{PGL}_3$ and $\mathbf{H}$ maximal split over $\mathbb{R}$. But we allow some generality in order to be able to make remarks about other groups, and because for most results there is no reason to be restrictive. The condition that $\mathbf{G}$ is almost simple over $\mathbb{R}$ will only be used to apply a Weyl law-type result from [13] (see Proposition 3.88) but is of course automatic for forms of $\mathbf{PGL}_{3,\overline{\mathbb{Q}}}$.

3.5.2 Forms of $\mathbf{PGL}_3$

By a form of a group $\mathbf{G}$ over a field E we shall mean a form of the base change to the algebraic closure, $\mathbf{G}_{\overline{E}}$. The following proposition gives a concrete list of all groups $\mathbf{G}$ that are allowed in Theorem 3. Proposition 3.75 below says what the corresponding Lie groups are.

Proposition 3.74. *The $\mathbb{Q}$-forms $\mathbf{G}$ of $\mathbf{PGL}_3$, and the anisotropic ones among them, are given by the following constructions.*

1. *(Inner forms) Let $D/\mathbb{Q}$ be a central simple algebra of degree 3 and define $\mathbf{G} = \mathbf{GL}_{1,D}/\mathbb{G}_m$. Then $\mathbf{G}$ is anisotropic if and only if D is a division algebra.*

2. *(Outer forms) Let $F/\mathbb{Q}$ be a quadratic field, (V, q) a 3-dimensional nondegenerate Hermitian space over F and define $\mathbf{G} = \mathbf{PU}(V)$. Then $\mathbf{G}$ is anisotropic if and only if (V, q) is.*

3. *(Outer forms) Let $F/\mathbb{Q}$ be a quadratic field, D/F be a central division algebra of degree 3 over F with an anti-involution $\tau \in \mathrm{End}_{\mathbb{Q}}(D)$ that acts nontrivially on F, (V,q) a 1-dimensional Hermitian space over D with $q \neq 0$, and define $\mathbf{G} = \mathbf{PU}(V)$. Concretely, there is a nonzero $d \in D$ fixed by τ such that $\mathbf{G}(\mathbb{Q}) = \{g \in D^\times : gd\tau(g) = d\}/F^1$, where $F^1 \subset F^\times$ denotes the norm 1 subgroup. Then $\mathbf{G}$ is anisotropic.*

Proof. That this list is exhaustive can be proven as in [53, §18.4], or can be deduced from [53, Theorem 18.4.1] by using that $\mathbf{PGL}_{n,\overline{\mathbb{Q}}}$ and $\mathbf{SL}_{n,\overline{\mathbb{Q}}}$ have the same automorphism groups. The statements about anisotropy can be seen as follows:

1. By Wedderburn's theorem, D is either a division algebra or isomorphic to $M_3(\mathbb{Q})$. When D is a matrix algebra, clearly $\mathbf{G}$ is split. Conversely, let D be a division algebra, $\mathbf{T} \subset \mathbf{G}$ be a maximal torus and χ a character of $\mathbf{T}$. We may lift $\mathbf{T}$ to a maximal torus $\mathbf{T}' \subset \mathbf{GL}_{1,D}$ [12, Proposition 10.6]. It is well known that such tori are obtained from the multiplicative groups of étale subalgebras of D. Because D is division, $\mathbf{T}' = \mathrm{Res}_{F/\mathbb{Q}} \mathbb{G}_{m,F}$ for a field extension $F/\mathbb{Q}$, and it follows that χ is a power of the norm character. Because χ is trivial on $\mathbb{G}_{m,\mathbb{Q}}$, it is trivial altogether. Therefore $\mathbf{T}$ is anisotropic.

2. If q represents 0, then V contains a hyperbolic plane H, and we find a rank 1 split torus in $\mathbf{U}(V)$, necessarily noncentral, that acts diagonally on H in an isotropic basis and acts trivially on $H^\perp$. Conversely, if $\mathbf{T} \subset \mathbf{G}$ is a split torus then we may lift it to a torus of $\mathbf{U}(V)$ by [12, Proposition 10.6], which contains a nontrivial split subtorus $\mathbf{T}'$ by [10, §II.8.14]. A vector $v \in V$ in a weight space with respect to $\mathbf{T}'$ for a nonzero weight then satisfies $q(v) = 0$. (Such a vector exists because the split torus $\mathbf{T}' \subset \mathrm{Res}_{F/\mathbb{Q}} \mathbf{GL}(V) \subset \mathbf{GL}_6$ is diagonalizable.)

3. This can be proven as in the second case: A split torus in $\mathbf{G}$ would give rise to a nonzero vector $v \in V \cong D$ satisfying $vd\tau(v) = 0$. Because D is a division algebra, this is impossible. □

Proposition 3.75. *Let $\mathbf{G}$ be as in Proposition 3.74. With the same numbering, its group of real points is as follows:*

1. *In the first case, $\mathbf{G}(\mathbb{R}) \cong \mathrm{PGL}_3(\mathbb{R})$.*

2. *In the second case: If F is real quadratic we have $\mathbf{G}(\mathbb{R}) \cong \mathrm{PGL}_3(\mathbb{R})$. If F is imaginary quadratic we have $\mathbf{G}(\mathbb{R}) \cong \mathrm{PU}(3)$ or $\mathrm{PU}(2,1)$ depending on the signature of (V,q) over $\mathbb{R}$.*

3. *In the third case, $\mathbf{G}(\mathbb{R}) \cong \mathrm{PGL}_3(\mathbb{R})$ when F is real quadratic and $\mathbf{G}(\mathbb{R}) \cong \mathrm{PU}(3)$ or $\mathrm{PU}(2,1)$ otherwise.*

Proof. 1. By a classical theorem of Frobenius there is no degree 3 division algebra over $\mathbb{R}$, so that $D \otimes_{\mathbb{Q}} \mathbb{R} \cong M_3(\mathbb{R})$ by Wedderburn's theorem. Therefore $\mathbf{G}(\mathbb{R}) \cong \mathrm{PGL}_3(\mathbb{R})$.

2. When F is real quadratic we have $F \otimes_{\mathbb{Q}} \mathbb{R} \cong \mathbb{R} \times \mathbb{R}$, and the induced automorphism permutes the two factors. Using this isomorphism we have $\mathbf{GL}(V)_{\mathbb{R}} \cong \mathrm{GL}_3(\mathbb{R}) \times \mathrm{GL}_3(\mathbb{R})$. The hermitian matrix defining q takes the form (A, A^T) under this isomorphism, and $A \in \mathrm{GL}_3(\mathbb{R})$ can be assumed diagonal. Now $\mathbf{U}(V)(\mathbb{R}) \cong \{(B,C) \in \mathrm{GL}_3(\mathbb{R})^2 : BAC^T = CA^TB^T = A\}$. It is clear that for every $B \in \mathrm{GL}_3(\mathbb{R})$ there is a unique element $(B,C) \in \mathbf{U}(V)(\mathbb{R})$, and this gives an isomorphism $\mathbf{U}(V)(\mathbb{R}) \cong \mathrm{GL}_3(\mathbb{R})$. The conclusion follows.

   When F is imaginary we have $F \otimes_{\mathbb{Q}} \mathbb{R} \cong \mathbb{C}$ and it follows that $\mathbf{U}(V)(\mathbb{R})$ is a classical unitary group, necessarily of signature $(3,0)$ or $(2,1)$.

3. This can be shown using similar arguments as in the first and second case. When F is real quadratic, using the isomorphism $F \otimes_{\mathbb{Q}} \mathbb{R} \cong \mathbb{R} \times \mathbb{R}$ we have $D \otimes_{\mathbb{Q}} \mathbb{R} \cong M_3(\mathbb{R})^2$. Call σ the anti-involution of $M_3(\mathbb{R})^2$ that swaps the two factors and transposes them. Then $\tau \circ \sigma$ is an automorphism of $M_3(\mathbb{R})^2$. (In fact, it preserves the center $\mathbb{R} \times \mathbb{R}$, therefore preserves both factors and is inner by Skolem-Noether.) Applying an automorphism of $M_3(\mathbb{R})^3$ we may assume that $\sigma = \tau$. A similar argument as in the second case now gives an isomorphism $\mathbf{U}(V)(\mathbb{R}) \cong \mathrm{GL}_3(\mathbb{R})$.

   When F is imaginary we have $F \otimes_{\mathbb{Q}} \mathbb{R} \cong \mathbb{C}$ and $D \otimes_{\mathbb{Q}} \mathbb{R} \cong M_3(\mathbb{C})$. By Skolem-Noether, we may assume that the automorphism induced by τ is conjugate transpose, subsequently that d is diagonal, and it follows that $\mathbf{U}(V)(\mathbb{R})$ is a classical unitary group. □

We could state a result similar to Proposition 3.75 with $\mathbb{R}$ replaced by any $\mathbb{Q}_p$: We would still distinguish cases based on whether p is split or inert in F, together with the ramified case. The main difference is that there exist degree 3 division algebras over $\mathbb{Q}_p$ [57, §17.10] (and exactly two of them), which would be reflected in the classification. However, the only result that we will require is the following.

Lemma 3.76. *Let* $\mathbf{G}$ *be a* $\mathbb{Q}$*-form of* $\mathbf{PGL}_3$ *and suppose* p *is a prime such that* $\mathbf{G}_{\mathbb{Q}_p}$ *is split. Then* $\mathbf{G}_{\mathbb{Q}_p} \cong \mathbf{PGL}_{3,\mathbb{Q}_p}$.

Proof. Both $\mathbf{G}_{\mathbb{Q}_p}$ and $\mathbf{PGL}_{3,\mathbb{Q}_p}$ are split forms of $\mathbf{PGL}_{3,\overline{\mathbb{Q}_p}}$. Because a reductive group has only one split form over any field [12, Théorème 2.13], they are isomorphic. □

3.5.3 Locally symmetric spaces

Because $\mathbf{G}$ is connected, the Lie group $G = \mathbf{G}(\mathbb{R})$ is as in §2.1.1, and the identity component G^0 is semisimple. (Recall that we require a semisimple group by definition to

be connected.) Let $K_\infty \subset G$ be a maximal compact subgroup. To G we associate data as in §2.1 with respect to the maximal compact K_∞: The Cartan involution is denoted by θ, the Iwasawa decomposition by NAK_∞ and the restricted root system by $(\mathfrak{a}^*, \Sigma)$. (In §3.5.7 we shall use different notation for the roots of a maximal split torus over an algebraic closure, but that will be only relevant to define Hecke algebras.) We let $d(\cdot,\cdot)$ be a distance function as in §3.1.

Remark 3.77. Contrary to [13] we do not, and cannot, assume that the involution θ is defined over $\mathbb{Q}$. Indeed, when $\mathbf{G}$ arises from a degree 3 division algebra over $\mathbb{Q}$ (see §3.5.2), such θ would have to come from an anti-involution of the division algebra, and that does not exist [42, Corollary 2.8]. This remark will only play a role in Proposition 3.88.

Denote by $\mathbb{A}_\mathbb{Q}$ the ring of adeles over $\mathbb{Q}$ and by $\mathbb{A}_f$ the finite adeles. When S is any set of places of $\mathbb{Q}$, define $\mathbb{A}_S$ to be the restricted product $\prod_{v\in S} \mathbb{A}_\mathbb{Q}$. Equip the groups $\mathbf{G}(\mathbb{Q}_p)$, $\mathbf{G}(\mathbb{A}_\mathbb{Q})$, $\mathbf{G}(\mathbb{A}_S)$ with their natural topologies, and similarly for $\mathbf{H}$. Let $\rho : \mathbf{G} \to \mathbf{SL}_d$ and the integral models be as in §3.5.1. Choose compact subgroups $K_p \subset \mathbf{G}(\mathbb{Q}_p)$ for all primes p with the following properties:

- $\rho(K_p) \subset \mathbf{SL}_d(\mathbb{Z}_p)$;
- the compact subgroup
$$K_f := \prod_p K_p$$
is open in $\mathbf{G}(\mathbb{A}_f)$;
- the center $\mathbf{Z}(\mathbf{G})(\mathbb{Q})$ intersects K_f trivially;

The first two conditions can be satisfied for example by taking $K_p = \mathbf{G}(\mathbb{Z}_p)$, and the third can be obtained by replacing any single K_p by a smaller subgroup. Define $K = K_\infty \times K_f \subset \mathbf{G}(\mathbb{A}_\mathbb{Q})$, and for any set S of places of $\mathbb{Q}$ define $K_S = \prod_{v\in S} K_v$. Define the automorphic spaces

$$[\mathbf{G}] = \mathbf{G}(\mathbb{Q})\backslash\mathbf{G}(\mathbb{A}_\mathbb{Q}) \quad \text{and} \quad [\mathbf{H}] = \mathbf{H}(\mathbb{Q})\backslash\mathbf{H}(\mathbb{A}_\mathbb{Q}) \subset [\mathbf{G}]$$

and the locally symmetric space

$$X = [\mathbf{G}]/K\,.$$

It is a finite disjoint union of locally symmetric spaces as in §3.1.2 [58, Theorem 5.1] and is compact because $\mathbf{G}$ is anisotropic [11]. Explicitly, let $(g_i)_{i\in I} \subset \mathbf{G}(\mathbb{A}_f)$ be a system of representatives for the double quotient $[\mathbf{G}]/(G \times K_f)$ and define $\Gamma_i = \mathbf{G}(\mathbb{Q}) \cap g_i K_f g_i^{-1}$. Then

$$[\mathbf{G}] = \bigsqcup_{i\in I} \Gamma_i\backslash(G \times g_i K_f) \tag{3.78}$$

and X is the disjoint union of the compact locally symmetric spaces

$$X_i = \Gamma_i \backslash G / K_\infty \,. \tag{3.79}$$

(Here we view $\mathbf{G}(\mathbb{Q})$ inside $\mathbf{G}(\mathbb{A}_f)$ when taking the intersection with $g_i K_f g_i^{-1}$, inside $\mathbf{G}(A_\mathbb{Q})$ in (3.78) and inside G in (3.79). Notice that the lattices Γ_i do not depend on the choice of representatives g_i.

3.5.4 Compact torus orbits

We explain now how to relate tori to maximal flat submanifolds of X. Let A as in §3.5.3 be the group given by the choice of Iwasawa decomposition.

Assume $\mathbf{H} \subset \mathbf{G}$ has maximal split rank over $\mathbb{R}$. Then there exists $g_\infty \in G$ such that $\mathbf{H}(\mathbb{R})$ contains $g_\infty A g_\infty^{-1}$, so that $\mathbf{H}(\mathbb{R}) \subset Z_G(g_\infty A g_\infty^{-1}) = g_\infty M A g_\infty^{-1}$. To $\mathbf{H}$ we then associate the compact maximal flat submanifold $\mathscr{F} \subset X$ that is the image of $\mathbf{H}(\mathbb{A}_\mathbb{Q}) g_\infty$. Relative to the decomposition (3.79), it is the disjoint union of some (but not necessarily all) of the maximal flats $g_\infty A \subset X_i$. More precisely, $\mathscr{F}$ intersects X_i nontrivially if and only if $\mathbf{H}(\mathbb{A}_f)$ intersects $g_i K_f$ nontrivially. Note that $\mathscr{F}$ does not depend on the choice of g_∞, because g_∞ is determined up to multiplication on the right by $N_G(A) = M'A$. Likewise, the definition does not depend on the choice of A, which can be seen by using that two choices of A are conjugate by an element of K_∞.

Conversely, one can show that every compact maximal flat in X_i is obtained in this way. Assume $\mathscr{F} \subset X_i$ is a compact maximal flat submanifold. As in §2.1.2 there exists $g_\infty \in G$ such that $\mathscr{F}$ is the image of $g_\infty A$ in X_i. The lattice Γ_i intersects $g_\infty A g_\infty^{-1}$ in a lattice, and define $\mathbf{T}$ to be the Zariski closure of $\Gamma_i \cap g_\infty A g_\infty^{-1}$. Then $\mathbf{T}$ is a torus that is maximal split over $\mathbb{R}$, and we may take $\mathbf{H} \subset \mathbf{G}$ to be any maximal torus containing $\mathbf{T}$.

Remark 3.80. The correspondence between tori and maximal flats that we outlined above, is slightly awkward because X can have different components and because $\mathbf{H}$ (in our notation) is always a maximal torus, not a maximal $\mathbb{R}$-split torus. When X is connected, one does obtain a clean bijection between compact maximal flats $\mathscr{F} \subset X$ and maximal $\mathbb{R}$-split tori $\mathbf{T} \subset \mathbf{G}$. Compare also [24, §2].

3.5.5 Integration

For every prime p we choose the Haar measures $d\mu_{\mathbf{G},p}$ and $d\mu_{\mathbf{H},p}$ on $\mathbf{G}(\mathbb{Q}_p)$ and $\mathbf{H}(\mathbb{Q}_p)$ respectively, for which K_p and $K_p \cap \mathbf{H}(\mathbb{Q}_p)$ have volume 1. We choose Haar measures on Lie groups as in §3.1.1. The abelian Lie group $\mathbf{H}(\mathbb{R})$ factors as a conjugate of A times a compact group, and we equip it with the measure that is the product of the measure coming from A with the volume 1 measure on the compact group. We form the product measures $d\mu_\mathbf{G} = \prod_v d\mu_{\mathbf{G},v}$ on $\mathbf{G}(\mathbb{A}_\mathbb{Q})$ and likewise the product measure $d\mu_\mathbf{H}$ on $\mathbf{H}(\mathbb{A}_\mathbb{Q})$. When S is any set of places of $\mathbb{Q}$, we similarly define measures $\mu_{\mathbf{G},S}$ and $\mu_{\mathbf{H},S}$ on $\mathbf{G}(\mathbb{A}_S)$ and $\mathbf{H}(\mathbb{A}_S)$ respectively.

3.5.6 Hecke algebras

A smooth function on $\mathbf{G}(\mathbb{Q}_p)$ is defined to be a locally constant one. When S is a set of places of $\mathbb{Q}$, a compactly supported Schwartz-Bruhat function on $\mathbf{G}(\mathbb{A}_S)$ is a finite linear combination of functions of the form

$$\prod_{v \in S} f_v \,,$$

where $f_v \in C_c^\infty(\mathbf{G}(\mathbb{Q}_v))$ and $f_p = 1_{K_p}$ for almost all primes $p \in S$. We denote by $C_c^\infty(\mathbf{G}(\mathbb{A}_S))$ the set of such (complex-valued) functions. If $f \in C_c^\infty(\mathbf{G}(\mathbb{A}_S))$, we define an operator $\pi(f)$ on $L^2([\mathbf{G}])$ by the rule

$$(\pi(f)\phi)(x) = \int_{\mathbf{G}(\mathbb{A}_S)} \phi(xg) f(g) d\mu_{\mathbf{G},S}(g) \,.$$

If we define $f^*(g) = \overline{f(g^{-1})}$, then $\pi(f)$ and $\pi(f^*)$ are adjoints. For every set S of places of $\mathbb{Q}$, define the algebra $\mathcal{H}_S$ of compactly supported smooth bi-K_S invariant functions on $\mathbf{G}(\mathbb{A}_S)$, with convolution given by

$$(f_1 * f_2)(g) = \int_{\mathbf{G}(\mathbb{A}_S)} f(gh^{-1}) f_2(h) d\mu_{\mathbf{G},S}(h) \,.$$

The algebras $\mathcal{H}_S$ act on $L^2(X)$ by the same rule as above. When p is a prime, the element $1_{K_p} \in \mathcal{H}_p$ is the identity.

By [71, §3.9], there exists an integer $D > 0$ with the property that for $p \nmid D$, K_p is the compact subgroup associated with a hyperspecial point of the building of $\mathbf{G}_{\mathbb{Q}_p}$. By [71, §3.3.3] this implies that the Hecke algebra $\mathcal{H}_p$ is commutative when $p \nmid D$.

Define $\mathcal{H}_f$ to be the algebra $\mathcal{H}_S$ when S consists of all finite places that do not divide D, and $\mathcal{H} = \mathcal{H}_\infty \otimes \mathcal{H}_f$. It is commutative by the above fact for finite places, and by [34, Theorem IV.3.1] for the infinite place.

3.5.7 Truncated Hecke algebras

We will use the notion of truncated Hecke algebras that is also used in [13], which makes it more convenient to formulate bounds for orbital integrals. While we will mostly be interested in primes at which all data is split, we do provide the general definition. Let $\mathbf{T}$ be any maximal split torus of $\mathbf{G}_{\overline{\mathbb{Q}}}$, $X^*(\mathbf{T})$ and $X_*(\mathbf{T})$ be the groups of characters and cocharacters of $\mathbf{T}$. Let Δ the set of roots of $\mathbf{T}$ in $\mathbf{G}_{\overline{\mathbb{Q}}}$, Δ^+ be a choice of positive roots and W the Weyl group with respect to $\mathbf{T}$. Let $\rho \in X^*(\mathbf{T}) \otimes_{\mathbb{Z}} \mathbb{Q}$ be the half-sum of positive roots and denote the natural paring $X^*(\mathbf{T}) \times X_*(\mathbf{T}) \to \mathbb{Z}$ by $\langle \cdot, \cdot \rangle$. Define a norm on $X_*(\mathbf{T})$ by

$$\|\mu\| = \max_{w \in W} \langle w\mu, \rho \rangle \in \frac{1}{2}\mathbb{Z} \,.$$

Now let $p \nmid D$ be a prime. As noted above there is a maximal split torus $\mathbf{A} \subset \mathbf{G}_{\mathbb{Q}_p}$ such that K_p corresponds to a hyperspecial point of the apartment of $\mathbf{A}$, and by the Cartan decomposition [71, §3.3.3] the double cosets $K_p \backslash \mathbf{G}(\mathbb{Q}_p)/K_p$ are represented uniquely by the elements $\mu(p)$ when $\mu \in X_*(\mathbf{A})$ runs through the cocharacters in the closure of the positive chamber. Conjugating $\mathbf{A}$ to $\mathbf{T}$ over $\overline{\mathbb{Q}_p}$ yields a norm $\|\cdot\|$ on $X_*(\mathbf{A})$ induced from the one on $X_*(\mathbf{T})$, which does not depend on the choice of $\mathbf{A}$ [13, §2.6]. For $\kappa \geq 0$ we define now the "truncated algebra" (which is not an algebra)

$$\mathcal{H}_p^{\leq \kappa} = \mathrm{span}_{\mathbb{C}}\{1_{K_p \mu(p) K_p} : \mu \in X_*(\mathbf{A}), \|\mu\| \leq \kappa\}\,.$$

When S is a set of primes not dividing D, we define $\mathcal{H}_S^{\leq \kappa}$ to be the restricted tensor product $\bigotimes_{p \in S} \mathcal{H}_p^{\leq \kappa} \subset \mathcal{H}_S$.

Finally, we will use yet another notion of truncated Hecke algebra to accomodate the Hecke operators that we will use to prove Theorem 3. When S is a set of primes not dividing D, $\kappa \geq 0$ and $M \geq 1$, define the truncated algebra

$$\mathcal{H}_{S,M}^{\leq \kappa} = \bigoplus_{\substack{n \leq M \\ \text{squarefree}}} \bigotimes_{\substack{p \in S \\ p \mid n}} \mathcal{H}_p^{\leq \kappa}\,.$$

3.5.8 Hecke-Maass forms

The group G acts on $[\mathbf{G}]$ by translation on the right, and this induces an action of the universal enveloping algebra $U(\mathfrak{g})$ on $C^\infty([\mathbf{G}])$, which using the decomposition (3.78) we may view as given by left-invariant differential operators on G (§2.1.5). By a Hecke-Maass form on X we mean a smooth function on X that is a joint eigenfunction for the center $Z(U(\mathfrak{g}))$ and for the Hecke algebra $\mathcal{H}_f$ from §3.5.6.

Let $(f_j)_{j \geq 0}$ be an orthonormal basis of $L^2(X)$ consisting of (complex-valued) Hecke-Maass forms. It may be obtained by decomposing $L^2(X)$ into eigenspaces for the Laplace-Beltrami operator, and finding in every eigenspace a basis of eigenfunctions for the commutative algebras $Z(U(\mathfrak{g}))$ and $\mathcal{H}_f$. When $k \in \mathcal{H}$, $\mathcal{H}_\infty$ or $\mathcal{H}_f$, denote by $\widehat{k}(f_j)$ the eigenvalue of f_j under k, and define the spectral parameter ν_j as in §3.1.2.

Let $g_\infty \in G$ be as in §3.5.4 with the property that $\mathbf{H}(\mathbb{R})$ contains $g_\infty A g_\infty^{-1}$. Define the $\mathbf{H}$-period of f_j by

$$\mathscr{P}_{\mathbf{H}}(f_j) = \int_{[\mathbf{H}]} f_j(h g_\infty) d\mu_{\mathbf{H}}(h)\,, \tag{3.81}$$

where the automorphic quotient $[\mathbf{H}]$ is as in §3.5.3. As in §3.5.4, this definition does not depend on the choice of A or g_∞.

3.6 Extreme values of toric periods

In this section we prove Theorem 3. We set up a relative pre-trace formula in §3.6.1 and prove the theorem in §3.6.3. Our notations for locally symmetric spaces, operator

algebras and Hecke-Maass forms are as in §3.5.

3.6.1 Adelic setup

Let $\mathbf{G}/\mathbb{Q}$ be semisimple anisotropic as in §3.5.1, and let $\mathbf{H} \subset \mathbf{G}$ be a maximal torus of maximal split rank over $\mathbb{R}$.

The pre-trace formula for $\mathbf{G}$ states that for any $k \in \mathcal{H}$ (with archimedean component satisfying a positivity condition as in §3.2.1) one has

$$\sum_j \widehat{k}(f_j) f_j(x) \overline{f_j(y)} = \sum_{\gamma \in \mathbf{G}(\mathbb{Q})} k(x^{-1}\gamma y)\,,$$

uniformly for $x, y \in [\mathbf{G}]$. As in §3.5.4 let $g_\infty \in \mathbf{G}(\mathbb{R})$ be such that $\mathbf{H}(\mathbb{R}) \supset g_\infty A g_\infty^{-1}$, define the $\mathbf{H}$-periods $\mathscr{P}_{\mathbf{H}}$ as in (3.81) and define $K_{\mathbf{H},f} = K_f \cap \mathbf{H}(\mathbb{A}_f)$.

We will require the following analogue of Lemma 3.7.

Lemma 3.82 (Partitions of unity)**.** *There exists a nonnegative Schwartz-Bruhat function $b \in C_c^\infty(\mathbf{H}(\mathbb{A}_\mathbb{Q})/K_{\mathbf{H},f})$ such that $\sum_{\gamma \in \mathbf{H}(\mathbb{Q})} b(\gamma h) = 1$ for all $h \in \mathbf{H}(\mathbb{A}_\mathbb{Q})$.*

Proof. The quotient $\mathbf{H}(\mathbb{A}_\mathbb{Q})/(\mathbf{H}(\mathbb{Q})\mathbf{H}(\mathbb{R})K_{\mathbf{H},f})$ is finite by [58, Theorem 5.1]. Take coset representatives $h_1, \dots, h_N \in \mathbf{H}(\mathbb{A}_f)$ and define $b_f \in C_c^\infty(\mathbf{H}(\mathbb{A}_f))$ by

$$b_f = \sum_{j=1}^N 1_{h_j K_{\mathbf{H},f}}\,.$$

The group $\mathbf{H}(\mathbb{R})$ factors uniquely as $V \times T$ with V a Euclidean space and T compact (possibly disconnected). The discrete subgroup $\Gamma_{\mathbf{H}} := \mathbf{H}(\mathbb{Q}) \cap K_{\mathbf{H},f}$ is a lattice in $\mathbf{H}(\mathbb{R})$ and therefore $\Gamma_V := \Gamma_{\mathbf{H}} \cap V$ has finite index in $\Gamma_{\mathbf{H}}$. We may construct $b_V \in C_c^\infty(V)$ as in Lemma 3.7 relative to the lattice $\Gamma_V \subset V$ and define $b_T(h) = [\Gamma_{\mathbf{H}} : \Gamma_V]^{-1}$ for $h \in T$. Define $b_\infty \in C_c^\infty(\mathbf{H}(\mathbb{R}))$ by

$$b_\infty(h) = b_V(h_V) \cdot b_T(h_T)\,,$$

where we denote $h = (h_V, h_T) \in V \times T \cong \mathbf{H}(\mathbb{R})$. Define now $b \in C_c^\infty(\mathbf{H}(\mathbb{A}_\mathbb{Q}))$ by

$$b = b_\infty b_f\,.$$

That this b satisfies the requirements can be checked by writing

$$\sum_{\gamma \in \mathbf{H}(\mathbb{Q})} b(\gamma h) = \sum_{\gamma \in \mathbf{H}(\mathbb{Q})/(\mathbf{H}(\mathbb{Q}) \cap K_{H,f})} \sum_{\mu \in \mathbf{H}(\mathbb{Q}) \cap K_{H,f}} b(\gamma \mu h)\,. \qquad \square$$

Remark 3.83. Lemma 3.82 is a special case of a much more general statement about continuous functions on locally compact homogeneous spaces; see [54, §III.4].

The Weyl group $N_{\mathbf{G}}(\mathbf{H})/\mathbf{H}$ is finite [10, §IV.11.19] and therefore so is its group of rational points.

Lemma 3.84. *Let $k \in C_c^\infty(K\backslash\mathbf{G}(\mathbb{A}_\mathbb{Q})/K)$ be a Schwartz-Bruhat function. There exists $b \in C_c^\infty(\mathbf{H}(\mathbb{A}_\mathbb{Q}))$ such that*

$$\begin{aligned}
&\sum_j \widehat{k}(f_j)|\mathscr{P}_{\mathbf{H}}(f_j)|^2 \\
&\quad = \mathrm{Vol}([\mathbf{H}]) \sum_{\gamma \in N_{\mathbf{G}}(\mathbf{H})(\mathbb{Q})/\mathbf{H}(\mathbb{Q})} \int_{\mathbf{H}(\mathbb{A}_\mathbb{Q})} k(g_\infty^{-1}\gamma h g_\infty) d\mu_{\mathbf{H}}(h) \\
&\qquad + \sum_{\gamma \in \mathbf{G}(\mathbb{Q}) - N_{\mathbf{G}}(\mathbf{H})(\mathbb{Q})} \int_{\mathbf{H}(\mathbb{A}_\mathbb{Q})^2} b(h_1)b(h_2)k(g_\infty^{-1}h_1^{-1}\gamma h_2 g_\infty) d\mu_{\mathbf{H}}(h_1) d\mu_{\mathbf{H}}(h_2)\,.
\end{aligned}$$

Proof. Using Lemma 3.82, this is entirely analogous to the proof of Lemma 3.8 by an unfolding argument and introduction of cutoff functions. □

Remark 3.85. The integrals in the diagonal term in Lemma 3.84 can depend on (the finite part of) γ, because it is not always true that $N_{\mathbf{G}}(\mathbf{H})(\mathbb{Q}) \subset \mathbf{H}(\mathbb{Q}) \cdot K_f$. When G is $\mathbf{PGL}_2$ or an inner form, this inclusion is related to the notion of "reciprocal geodesics" in [66].

3.6.2 Comparison of trace formulas

We now specialize to the groups $\mathbf{G}$ in Theorem 3, although $\mathbf{G}$ can still be as general as in Theorem 4 in this subsection. We will prove Theorem 3 by comparing asymptotics for an amplified trace formula and an amplified relative trace formula.

Let $\rho : \mathbf{G} \to \mathbf{SL}_d$ be as in §3.5.1, and when p is prime define $\|g\|_p$ for $g \in \mathbf{G}(\mathbb{Q}_p)$ as the maximal absolute value of the entries of $\rho(g)$. When $g \in \mathbf{G}(\mathbb{Q})$ we have for almost all primes p that $\rho(g) \in \mathbf{SL}_d(\mathbb{Z}_p)$ and consequently $\|g\|_p = 1$. We may therefore define $\|g\|_f = \prod_p \|g\|_p$.

Lemma 3.86. *There exists $A > 0$ such that when $\gamma \in \mathbf{G}(\mathbb{Q})$ is such that $g_\infty^{-1}\gamma g_\infty \notin \bigcup_{L \in \mathcal{L}-\{G\}} M'L$, then*

$$d\left(g_\infty^{-1}\gamma g_\infty, \bigcup_{L \in \mathcal{L}-\{G\}} M'L\right) \gg \|\gamma\|_f^{-A}\,.$$

Proof. This can be shown as in [13, Lemma 5.1] as a consequence of the product formula for global fields. The difference is that we must deal with varieties that are not defined over our base field $\mathbb{Q}$.

It suffices to prove a bound as in the statement for every individual Levi $L \in \mathcal{L} - \{G\}$. Fix such L.

Let $E \subset \mathbb{R}$ be a number field such that the extension of scalars $\mathbf{H}_E$ has the same split rank as $\mathbf{H}_{\mathbb{R}}$. Let $\mathbf{T} \subset \mathbf{H}_E$ be the maximal split subtorus, defined over E. Because all roots of $\mathbf{H}_{\mathbb{R}}$ are defined over E, we may find an element $h \in \mathbf{T}(E)$ whose centralizer in $\mathbf{G}$ has real points $g_\infty L g_\infty^{-1}$. Consider now the subvariety $\mathbf{V}$ of $\mathbf{G}$ defined over E by the equation $\mathrm{Ad}_g(h) \in \mathbf{H}$. Replacing h by a finite power if necessary, we may assume that $h \in \mathbf{T}(\mathbb{R})^0 = g_\infty A g_\infty^{-1}$. Lemma 2.2 then says that $\mathbf{V}(\mathbb{R}) = g_\infty M' L g_\infty^{-1}$. Thus by assumption, $\gamma \notin \mathbf{V}(E)$.

Consider now the embedding $\rho : \mathbf{G} \to \mathbf{SL}_d \subset \mathbf{A}_{\mathbb{Q}}^{d^2}$, and let $p_1, \dots, p_k$ be polynomials defined over E whose joint zero locus is $\rho(\mathbf{V})$. Expanding their coefficients in a basis for $E/\mathbb{Q}$, we may find polynomials $q_1, \dots, q_t$ defined over $\mathbb{Q}$ whose rational joint zero locus is $\rho(\mathbf{V}(E) \cap \mathbf{G}(\mathbb{Q}))$. There exists $A > 0$ such that for all q_i and all $\gamma \in \mathbf{G}(\mathbb{Q})$ we have a bound of the form

$$\prod_p |p_i(\rho(\gamma))|_p \ll \|\gamma\|_f^A \,.$$

Indeed, $A > 0$ can be chosen to be $\max_i \deg(q_i)$, simply because a polynomial of degree s can only increase denominators as much as raising them to the power s.

Now assume $\gamma \in \mathbf{G}(\mathbb{Q})$ does not lie in $\mathbf{V}(E)$. Then there exists q_i such that $q_i(\rho(\gamma)) \neq 0$. By the product formula for $\mathbb{Q}^\times$, this means that

$$\prod_v |q_i(\rho(\gamma))|_v = 1 \,,$$

where the product now runs over all absolute values of $\mathbb{Q}$. Using our bound for the product over the finite primes, this gives

$$|q_i(\rho(\gamma))|_\infty \gg \|\gamma\|_f^{-A} \,.$$

By smoothness of the action of $\mathbf{G}(\mathbb{R})$ on h, the left-hand side is bounded from above by a constant times $d(\gamma, \mathbf{V}(\mathbb{R}))$. This shows that

$$d(\gamma, g_\infty M' L g_\infty^{-1}) \gg \|\gamma\|_f^{-A} \,,$$

and the lemma follows. □

When S is a finite set of primes, define $q_S = \prod_{p \in S} p$. Let the integer D be as in §3.5.6.

Proposition 3.87. *Let $D_{\mathfrak{a}^*} \subset (\mathfrak{a}^*)^{\mathrm{gen}}$ be compact. let $\kappa \geq 0$. There exist $A, B, \delta > 0$ such that the following holds. Let S a finite set of primes not dividing D and $k_f \in \mathcal{H}_S^{\leq \kappa}$. Let $\nu \in D_{\mathfrak{a}^*}$, $t \geq 0$ and $k_{t\nu} \in \mathcal{H}_\infty$ as in Lemma 3.10. Define $k = k_{t\nu} \otimes k_f$. Then*

$$\sum_j \widehat{k}(f_j) |\mathscr{P}_{\mathbf{H}}(f_j)|^2 = \mathrm{Vol}([\mathbf{H}]) \cdot \sum_{\gamma \in N_{\mathbf{G}}(\mathbf{H})(\mathbb{Q})/\mathbf{H}(\mathbb{Q})} \int_{\mathbf{H}(\mathbb{A}_{\mathbb{Q}})} k(g_\infty^{-1} \gamma h g_\infty) d\mu_{\mathbf{H}}(h)$$
$$+ O\left(\beta(t\nu)(1+t)^{-r}(1 + q_S^{-A} t)^{-\delta} q_S^B \|k_f\|_\infty\right) \,,$$

uniformly in ν and t. If $M \geq 1$ and $k_f \in \mathcal{H}_{S,M}^{\leq \kappa}$ then the same result holds with q_S replaced by M.

Proof. We start from Lemma 3.84 and must bound the off-diagonal terms. The fixed function b is a finite sum of factorizable Schwartz-Bruhat functions, so up to introducing a finite sum, we may replace the factors $b(h_1)$ and $b(h_2)$ by factorizable Schwartz-Bruhat functions $b_1(h_1)$ and $b_2(h_2)$. (In fact, we only require that they factorize as an infinite times a finite part, and this is already the case for b as constructed in Lemma 3.82.) For fixed γ, the double integral now factorizes as the integral

$$\int_{\mathbf{H}(\mathbb{R})^2} b_1(h_1) b_2(h_2) k_{t\nu}(g_\infty^{-1} h_1^{-1} \gamma h_2 g_\infty) d\mu_{\mathbf{H},\mathbb{R}}(h_1) d\mu_{\mathbf{H},\mathbb{R}}(h_2)$$

times an integral involving b_1, b_2 and k_f over a fixed compact subset of $\mathbf{H}(\mathbb{A}_f)$. The latter can be bounded trivially by

$$\ll \|k_f\|_\infty \cdot \mathrm{Vol}(\mathrm{supp}(b_1)) \cdot \mathrm{Vol}(\mathrm{supp}(b_2)) \cdot \mathrm{Vol}(\mathrm{supp}(k_f)) \,.$$

The first two volumes are bounded, and the third can be bounded by a power of q_S, as in [13, Lemma 4.4].

The archimedean integral equals

$$\int_{A \times A} b_1(g_\infty a_1 g_\infty^{-1}) b_2(g_\infty a_2 g_\infty^{-1}) k_{t\nu}(a_1 g_\infty^{-1} \gamma g_\infty a_2) da_1 da_2 \,,$$

where we have written $\mathbf{H}(\mathbb{R})$ as a product of $g_\infty A g_\infty^{-1}$ and a compact torus, where the compact torus necessarily lies in $g_\infty M g_\infty^{-1}$ (see §3.5.4) and we may omit it from the integration by our choice of measure on $\mathbf{H}(\mathbb{R})$ (see §3.5.5). By our choice of G and X, we have that

$$g_\infty^{-1} \gamma g_\infty \notin \bigcup_{L \in \mathcal{L} - \{G\}} M' L$$

for $\gamma \in \mathbf{G}(\mathbb{Q})$ as above. Using the inversion formula (3.3), the rapid decay of $\widehat{k}_{t\nu}$ and Proposition 3.28, this means that the archimedean integral above is bounded by

$$\ll \beta(t\nu)(1+t)^{-r} \left(1 + d\left(g_\infty^{-1} \gamma g_\infty, \bigcup_{L \in \mathcal{L} - \{G\}} M' L \right)^{-N} \cdot t \right)^{-\delta}$$

for certain $\delta, N > 0$. By Lemma 3.86 we may bound the distance from below by $\|\gamma\|_f^{-A}$ for some $A > 0$, and as in [13, Lemma 4.4] the latter is again bounded from below by $q_S^{-A'}$ for some $A' > 0$. It remains to show that the number of γ that contribute to the sum is at most polynomial in q_S. Indeed, for such γ we have that the rational element

$\rho(\gamma) \in \mathbf{SL}_d(\mathbb{Q})$ is bounded in the archimedean sense by the support conditions on b_1 and b_2, and its denominators are at most of size $q_S^{A'}$ by the same argument we just used. This proves the desired bound for the off-diagonal terms.

When $k_f \in \mathcal{H}_{S,M}^{\leq\kappa}$, it is a sum of at most M elements $k_{f,n} \in \mathcal{H}_{S_n}^{\leq\kappa}$, with S_n a set of primes with $q_{S_n} \leq M$. Moreover, we may assume that the $k_{f,n}$ have disjoint supports. The result then follows by summing the different bounds for each of the $k_{f,n}$, and replacing B by $B+1$. □

We will require the following estimates for the trace formula.

Proposition 3.88. *There are constants $\eta, A, B > 0$ and an integer multiple D' of D such that the following holds. Let $\kappa \geq 0$ and S a finite set of primes not dividing D', and $k_f \in \mathcal{H}_S^{\leq\kappa}$. Let $D_{\mathfrak{a}^*} \subset \mathfrak{a}^* - \{0\}$ be compact, $\nu \in D_{\mathfrak{a}^*}$, $t \geq 1$ and $k_{t\nu} \in \mathcal{H}_\infty$ as in Lemma 3.10. Define $k = k_{t\nu} \otimes k_f$. Then*

$$\sum_j \widehat{k}(f_j) = \mathrm{Vol}(X) \cdot k(1) + O\left(\beta(t\nu)(1+t)^{-\eta} q_S^{A\kappa+B} \|k_f\|_\infty\right),$$

uniformly in ν and t. If $M \geq 1$ and $k_f \in \mathcal{H}_{S,M}^{\leq\kappa}$ then the same result holds with q_S replaced by M.

Proof. The first statement follows from [13, Theorem 3.1]. Note that our Cartan involution is not assumed to be defined over $\mathbb{Q}$ (see Remark 3.77). This is compensated for by our assumption that $\mathbf{G}$ is almost simple over $\mathbb{R}$, which we can use to replace the argument in [13, Lemma 6.3] needed to show that the centralizer of G^0 in $\mathbf{G}(\mathbb{Q})$ is reduced to the center.

The statement about $k_f \in \mathcal{H}_{S,M}^{\leq\kappa}$ is shown as for Proposition 3.87. □

Remark 3.89. The estimate [13, Theorem 3.1] is stronger than what is needed here, because we do not need uniformity with respect to a variable compact subgroup K (the level aspect) nor do we need asymptotics for spectral parameters that can be singular. But it is the only directly quotable Weyl law-type result that we have found, that covers anisotropic groups and adelic test functions. Note also that the exponents A and B are ineffective, but they will not play any role.

We will apply Propositions 3.87 and 3.88 always with the following positivity assumption on k_f:

$$\widehat{k}_f(f_j) \geq 0 \text{ for all } f_j,\ k_f(1) \geq 1 \text{ and } k_f(g) \geq 0 \text{ for } g \in G(\mathbb{A}_f). \tag{3.90}$$

The following Proposition combines the two trace formulae and reduces Theorem 3 to finding a suitable test function k_f.

Proposition 3.91. *Let $D_{\mathfrak{a}^*} \subset (\mathfrak{a}^*)^{\mathrm{gen}}$ be compact. Let $\kappa \geq 0$. There exist $\delta, C > 0$ and an integer multiple D' of D such that the following holds. Let $\nu \in D_{\mathfrak{a}^*}$, S be a finite set of primes not dividing D', $t, M \geq 1$ and $k_f \in \mathcal{H}_{S,M}^{\leq k}$ satisfying (3.90). If $M \leq (1+t)^\delta$ and $\|k_f\|_\infty \leq (1+t)^\delta$, then*

$$\frac{\sum_{\|\nu_j - t\nu\| \leq C} \widehat{k}_f(f_j) |\mathscr{P}_H(f_j)|^2}{\sum_{\|\nu_j - t\nu\| \leq C} \widehat{k}_f(f_j)} \gg (1+t)^{-r} \cdot \frac{\int_{H(\mathbb{A}_f)} k_f}{k_f(1)},$$

uniformly in ν, t, S, M, k_f.

Proof. We first bound the main term in Proposition 3.87 from below. The integrals factorize. The archimedean component can be bounded from below by $\beta(t\nu)(1+t)^{-r}$. Indeed, using the inversion formula (3.3) it reduces to a similar integral involving the spherical function, which is seen to be independent of γ by making a change of variables, and we conclude using Proposition 3.15. For the finite component we may use the positivity of k_f and bound the integral trivially from below by $k_f(1)\operatorname{Vol}(\mathbf{H}(\mathbb{A}_f) \cap K_f) \geq 1$. Given this lower bound for the main term and given the quality of the error term in Proposition 3.87, a truncation argument as in Proposition 1.50 or [13, Lemma 4.5] shows the asymptotic formula

$$\begin{aligned} &\sum_{\|\nu_j - t\nu\| \leq C} \widehat{k}_f(f_j) \, |\mathscr{P}_H(f_j)|^2 \\ &\qquad \asymp \operatorname{Vol}([\mathbf{H}]) \sum_{\gamma \in N_{\mathbf{G}}(\mathbf{H})(\mathbb{Q})/\mathbf{H}(\mathbb{Q})} \int_{\mathbf{H}(\mathbb{A}_\mathbb{Q})} k(g_\infty^{-1} \gamma h g_\infty) d\mu_{\mathbf{H}}(h), \end{aligned}$$

provided that C is large enough and that M and $\|k_f\|_\infty$ are bounded by a sufficiently small power of t so that the error terms are controlled. Similarly, the inversion formula (3.3) gives that $k_{t\nu}(1) \asymp \beta(t\nu)$ and a truncation argument applied to Proposition 3.88 gives that

$$\sum_{\|\nu_j - t\nu\| \leq C} \widehat{k}_f(f_j) \asymp \operatorname{Vol}(X) k(1),$$

provided that C is large enough and that M and $\|k_f\|_\infty$ are bounded by a sufficiently small power of t.

The statement now follows by taking the quotient of those two asymptotics, using again the archimedean asymptotics, and in the case of the relative trace formula, using positivity of k_f to discard the terms corresponding to γ that are not the identity. □

3.6.3 Amplification

The construction of amplifiers is our only reason to restrict to groups $\mathbf{G}$ as in Theorem 3. The theorem follows by applying Proposition 3.91 with k_f as given by the following proposition, and with M a sufficiently small power of t.

Proposition 3.92. *Let G be a $\mathbb{Q}$-form of $\mathbf{PGL}_3$ and let δ be as stated in Theorem 3. Let $M \geq 2$ and S be the set of primes less than M that do not divide D'. There exists $k_f \in \mathcal{H}_{S,M}^{\leq 4}$ satisfying (3.90), such that $\|k_f\|_\infty \leq M^A$ for some $A > 0$ and*

$$\frac{\int_{\mathbf{H}(\mathbb{A}_f)} k_f}{k_f(1)} \gg (\log\log M)^{\delta + o(1)} .$$

We prove Proposition 3.92 in §3.7. We will also prove a result that states that the lower bound in Proposition 3.92 is optimal in a certain sense, although the exact result is not as strong as the optimality statement in [50], and we do not formally exclude the existence of amplifiers of similar quality for, say, forms of $\mathbf{PGL}_n$. See Remark 3.112.

3.7 Construction of amplifiers

In this section we prove the existence of the amplifier in Proposition 3.92 (in §3.7.2) . We also prove an optimality result in Proposition 3.111 (in §3.7.3).

3.7.1 Preliminary computations

We begin with some preliminary computations of integrals on p-adic groups. It is convenient to re-introduce some notation in order to make the key computations more self-contained. Let $n \geq 2$, let p be a prime number and denote $G_p = \mathrm{PGL}_n(\mathbb{Q}_p)$, $K_p = \mathrm{PGL}_n(\mathbb{Z}_p)$ and define $H_p \subset G_p$ to be the diagonal torus. Most of the time we will specialize to $n = 3$, but we allow some generality to be able to make remarks about other n, and because certain statements will be proven for arbitrary n. We fix Haar measures on G_p and H_p such that the compact open subgroups K_p and $H_p \cap K_p$ have volume 1. Define $\mathcal{H}_p = C_c^\infty(K_p \backslash G_p / K_p)$, a commutative algebra with convolution defined by

$$(k_1 * k_2)(g) = \int_{G_p} k_1(h) k_2(h^{-1} g) dh .$$

The adjoint of $k \in \mathcal{H}_p$ is defined by $k^*(g) = \overline{k(g^{-1})}$. This defines an algebra involution on $\mathcal{H}_p$.

The question we seek to answer is the following: When $k_p \in \mathcal{H}_p$ is a function of the form $k_p' * (k_p')^*$ with $k_p' \in \mathcal{H}_p$, how big can $\int_{H_p} k_p$ be relative to $k_p(1)$?

Such k_p' is a finite linear combination of the basic double coset kernels, which are defined as follows. Let $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{Z}^n$ and denote $\mu_{\mathbf{a}}(p) = \mathrm{diag}(p^{a_1}, \ldots, p^{a_n})$. One may view $\mu_{\mathbf{a}}$ as a cocharacter of H_p. Define the function

$$\tau(\mathbf{a}, p) = 1_{K_p \mu_{\mathbf{a}}(p) K_p} .$$

We have $\tau(\mathbf{a}, p)^* = \tau(-\mathbf{a}, p)$, and $\tau(\mathbf{0}, p)$ is the identity of $C_c^\infty(K_p \backslash G_p / K_p)$. The function $\tau(\mathbf{a}, p)$ depends only on the entries of $\mathbf{a}$ up to permutation and up to translation by a

common element in $\mathbb{Z}$. We may therefore always reduce to the case where $a_1 \geq a_2 \geq \ldots \geq a_n = 0$. The degree $\deg(\mu_{\mathbf{a}}(p))$ is defined as the cardinality $\#(K_p \mu_{\mathbf{a}}(p) K_p / K_p)$. With our choice of Haar measures we have

$$\deg(\mu_{\mathbf{a}}(p)) = \|\tau(\mathbf{a}, p)\|_2^2 = (\tau(\mathbf{a}, p) * \tau(-\mathbf{a}, p))(1)\,. \tag{3.93}$$

We will need bounds for integrals of convolutions of these basic kernels. Specifically, the quantity we will be interested in is the off-diagonal contribution to the integral $\int_{H_p} \tau(\mathbf{a}_1, p) * \tau(\mathbf{a}_2, p)$, when $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{Z}^n$. That is, we are interested in the integral over $H_p - (H_p \cap K_p)$, or what is the same, the integral over H_p minus $(\tau(\mathbf{a}_1, p) * \tau(\mathbf{a}_2, p))(1)$.

The following is an example computation of off-diagonal contributions when $n = 2$, which to an extent can be seen geometrically in the Bruhat-Tits tree.

Example 3.94. Let $n = 2$ and define $\mathbf{a} = (1, 0)$ and $\mathbf{b} = (0, 0)$. Then $\deg(\mu_{\mathbf{a}}(p)) = p+1$, $\tau(\mathbf{a}, p)$ is self-adjoint and

$$\left(\int_{H_p} \tau(\mathbf{a}, p) * \tau(\mathbf{a}, p)\right) - \deg(\mu_{\mathbf{a}}(p)) = (p+3) - (p+1) = 2\,,$$
$$\int_{H_p} \tau(\mathbf{a}, p) * \tau(\mathbf{b}, p) = 2\,.$$

That $\deg(\mu_{\mathbf{a}}(p)) = p + 1$, the cardinality of a radius 1 ball in the Bruhat-Tits tree, is a classical counting problem of double cosets. That $\tau(\mathbf{a}, p)$ is self-adjoint holds because $\mathbf{a}$ and $-\mathbf{a}$ are $\mathbb{Z}$-translates. We look at the second integral first. It equals 2 because it is the intersection of a radius 1 ball in the Bruhat-Tits tree with an apartment containing the center of the ball. We can also show this more explicitly as follows. We must count how many left cosets of K_p are contained in $K_p \operatorname{diag}(p, 1) K_p$ and intersect H_p. We may lift this problem to $\mathrm{GL}_2(\mathbb{Q}_p)$ and ask how many diagonal matrices lie in $\mathrm{GL}_2(\mathbb{Z}_p) \operatorname{diag}(p, 1) \mathrm{GL}_2(\mathbb{Z}_p)$. Multiplication on the left or the right by $\mathrm{GL}_2(\mathbb{Z}_p)$ does not change the invariant factors of a matrix, so that such a diagonal matrix must have invariant factors $(p, 1)$. It is clear that this gives exactly two matrices, and the statement follows.

The first integral can be computed using a similar argument, after using the Hecke relation $\tau(\mathbf{a}, p) * \tau(\mathbf{a}, p) = \tau(2\mathbf{a}, p) + (p+1)\tau(\mathbf{0}, p)$.

Lemma 3.95. *When $\mathbf{a}, \mathbf{b} \in \mathbb{Z}^n$ are decreasing tuples with $a_n = b_n = 0$, we have $(\tau(\mathbf{a}, p) * \tau(-\mathbf{b}, p))(1) = \delta_{\mathbf{a},\mathbf{b}} \deg(\mu_{\mathbf{a}}(p))$.*

Proof. When $\mathbf{a} = \mathbf{b}$ this is (3.93). When $\mathbf{a} \neq \mathbf{b}$ the double cosets represented by $\mu_{\mathbf{a}}(p)$ and $\mu_{\mathbf{b}}(p)$ are distinct, either by the Cartan decomposition [71, §3.3.3] or by an argument using invariant factors. They are therefore disjoint, so that

$$(\tau(\mathbf{a}, p) * \tau(-\mathbf{b}, p))(1) = \langle \tau(\mathbf{a}, p), \tau(\mathbf{b}, p) \rangle_{L^2(G_p)} = 0\,. \qquad \square$$

To do explicit computations, we translate integrations into counting problems involving lattices in $\mathbb{Q}_p^n$, which are the natural generalization of points in the Bruhat-Tits tree. By a lattice in $\mathbb{Q}_p^n$ we mean a finitely generated $\mathbb{Z}_p$-submodule of rank n. Denote by $\mathscr{R}$ the set of lattices in $\mathbb{Q}_p^n$, and by $\overline{\mathscr{R}}$ the set of homothety classes of lattices. The group $\mathrm{GL}_n(\mathbb{Q}_p)$ acts on $\mathscr{R}$ and the group G_p acts on $\overline{\mathscr{R}}$. Denote $L_0 = \mathbb{Z}_p^n$ and $\overline{L_0}$ its homothety class; they are our base points. The stabilizers of L_0 and $\overline{L_0}$ in $\mathrm{GL}_n(\mathbb{Q}_p)$ and G_p are $\mathrm{GL}_n(\mathbb{Z}_p)$ and K_p, respectively. Acting on the base point yields bijections $\mathrm{GL}_n(\mathbb{Q}_p)/\mathrm{GL}_n(\mathbb{Z}_p) \cong \mathscr{R}$ and $G_p/K_p \cong \overline{\mathscr{R}}$.

Lemma 3.96. *Let $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{Z}^n$ with $a_1 \geq a_2 \geq \ldots \geq a_n = 0$. There is a natural bijection between the following sets:*

- *The lattices $L \subset L_0$ for which L_0/L has invariant factors $p^{a_1}, \ldots, p^{a_n}$;*
- *The left coset space $K_p \mu_{\mathbf{a}}(p) K_p / K_p$,*

which is given as follows: To a lattice L one associates the homothety class $\overline{L} \in \overline{\mathscr{R}}$, which is identified with a left coset of K_p through the bijection $G_p/K_p \cong \overline{\mathscr{R}}$.

Proof. That the map is well-defined (meaning, that it lands in $K_p \mu_{\mathbf{a}}(p) K_p/K_p$) is the following fact: When M is a free module over a PID with a submodule N, then there exists a basis (e_i) of M that is adapted to N, meaning that N has a basis consisting of scalar multiples of the e_i. This fact shows that every lattice $L \subset L_0$ such that L_0/L has invariant factors as given, is of the form $k_p \operatorname{diag}(a_1, \ldots, a_n) L_0$ with $k_p \in \mathrm{GL}_n(\mathbb{Z}_p)$. That the map is surjective is trivial, because the lattice $k_p \operatorname{diag}(a_1, \ldots, a_n) L_0$ is sent to the corresponding left coset in $K_p \mu_{\mathbf{a}}(p) K_p / K_p$. □

Denote by $(e_1, \ldots, e_n)$ the standard basis of L_0. We call $L \in \mathscr{R}$ an adapted lattice if it has a basis of the form $(b_1 e_1, \ldots, b_n e_n)$ with the $b_i \in \mathbb{Q}_p^\times$.

Lemma 3.97. *Let $\mathbf{a} = (a_1, \ldots, a_n) \in \mathbb{Z}^n$ with $a_1 \geq a_2 \geq \ldots \geq a_n = 0$. The bijection from Lemma 3.96 restricts to a bijection between the following sets:*

- *The set of adapted lattices $L \subset L_0$ for which L_0/L has invariant factors $p^{a_1}, \ldots, p^{a_n}$.*
- *$(H_p \cap K_p \mu_{\mathbf{a}}(p) K_p)/(H_p \cap K_p)$.*

Proof. This can be proven as for Lemma 3.96. □

Lemma 3.98. *Let $\mathbf{a}, \mathbf{b} \in \mathbb{N}^n$ be decreasing tuples with $a_n = b_n = 0$. Then $\int_{H_p} \tau(\mathbf{a}, p) * \tau(-\mathbf{b}, p)$ counts the pairs of lattices (L_1, L) with the following properties:*

- *L_1 is adapted and $L \subset L_0 \cap L_1$.*
- *The invariant factors of L_0/L are given by $\mathbf{a}$ and those of L_1/L by $\mathbf{b}$.*

Proof. We have

$$\int_{H_p} \tau(\mathbf{a},p) * \tau(-\mathbf{b},p) = \int_{H_p/(H_p\cap K_p)} \int_{G_p/K_p} \tau(\mathbf{a},p)(g)\tau(-\mathbf{b},p)(g^{-1}h)dgdh\,.$$

By our convention on Haar measures, the right hand side is a sum over cosets $h(H_p \cap K_p)$ and gK_p. Take such cosets for which the integrand is nonzero (and thus equal to 1). Then $g\overline{L}_0$ is represented by a unique lattice $L \subset L_0$ for which L/L_0 has invariant factors given by $\mathbf{a}$, and this L does not depend on the representative g. There is a unique lift $h_1 \in \mathrm{GL}_n(\mathbb{Q}_p)$ such that $h_1^{-1}L \subset L_0$ has invariant factors given by $\mathbf{b}$, and this again does not depend on the representative h. Define then $L_1 = h_1 L_0$. The pair (L_1, L) satisfies the conditions of the statement. Conversely, to such a pair we can associated unique cosets $h(H_p \cap K_p)$ and gK_p on which the integrand is nonzero. □

3.7.2 Lower bounds

We prove Proposition 3.92. Our construction of an amplifier k_f is inspired by the construction in [50], which was used in Chapter 1. When translated into the adelic language, the amplifier from [50] takes the form $k_f = \omega * \omega^*$, with

$$\omega = \sum_{\substack{n\le M\\ \text{squarefree}}} \prod_{p|n} c_p\omega_p, \tag{3.99}$$

for ω_p elementary Hecke operators, which in the notation from §3.7.1 equal $\tau((1,0),p)$, and with $c_p > 0$ parameters to be optimized. The equalities in Example 3.94, which are about $\tau((1,0),p)$, appear in [50], not explicitly but through global versions thereof that are formulated in terms of divisor functions, in the proof of [50, Lemma 5].

One can think of the Hecke operator ω in (3.99) as being the formal expansion of the product $\prod(1 + c_p\omega_p)$, truncated to only include terms corresponding to sets of primes with product less than M.

Our aim is to optimize the choice of ω_p and of c_p for forms of $\mathbf{PGL}_3$. We continue to use the notation from §3.7.1 but will soon after switch to a global setup. The following lemma is the key in the construction of an amplifier.

Lemma 3.100. *Let $n = 3$ and $\mathbf{a} = (1,0,0)$ or $(1,1,0)$. Then $\deg(\mu_{\mathbf{a}}(p)) = p^2 + p + 1$, and we have*

$$\begin{aligned}
&\int_{H_p} \tau(\mathbf{a},p) * \tau(\mathbf{a},p) = 3(p+2)\,,\\
&\left(\int_{H_p} \tau(\mathbf{a},p) * \tau(-\mathbf{a},p)\right) - \deg(\mu_{\mathfrak{a}}(p)) = 6\,,\\
&\int_{H_p} \tau(\mathbf{a},p) = 3\,.
\end{aligned}$$

Proof. Using that adjugation is an algebra involution and that H_p is unimodular, it suffices to prove each identity for either $\mathbf{a} = (1,0,0)$ or $(1,1,0)$, because $\tau(\mathbf{a},p)$ and $\tau(-\mathbf{a},p)$ are adjoints. When $L_1, L_2 \in \mathscr{R}$ are lattices, define their generalized index by

$$[L_1 : L_2] = \frac{[L_1 : L_1 \cap L_2]}{[L_2 : L_1 \cap L_2]}\,.$$

It satisfies the usual transitivity relation.

For the degree, choose $\mathbf{a} = (1,0,0)$. We must count lattices $L \subset L_0$ of index p. Such L contain pL_0 and are determined by their quotient modulo pL_0. We must then count index p subgroups in $(\mathbb{Z}/p\mathbb{Z})^3$, or equivalently lines in $\mathbb{F}_p^3$, of which there are p^2+p+1.

For the first integral, choose $\mathbf{a} = (1,1,0)$. We must count pairs of lattices (L_1, L) satsifying the conditions in Lemma 3.98, with $\mathbf{b} = (1,0,0)$. Thus $L \subset L_0$ is a lattice with $L_0/L \cong (\mathbb{Z}/p\mathbb{Z})^2$ and $[L_1 : L] = p$. Because L_1 is adapted, contains $L \supset p\mathbb{Z}_p$ and $[L_0 : L_1] = [L_0 : L][L : L_1] = p$, it is clear that the only possibilities for L_1 are the following:

- $p^{-1}\mathbb{Z}_p \oplus p\mathbb{Z}_p \oplus p\mathbb{Z}_p$ and
- $\mathbb{Z}_p \oplus \mathbb{Z}_p \oplus p\mathbb{Z}_p$,

up to permutations of the factors. (This list would have been longer if we had switched $\mathbf{a}$ and $\mathbf{b}$; what helps us here is that $\mathbf{b}$ has small entries.) It suffices to count pairs (L_1, L) where L_1 is one of these two lattices, and multiply the result by 3. In the first case, the condition $L \subset L_0 \cap L_1$ forces $L \subset \mathbb{Z}_p \oplus p\mathbb{Z}_p \oplus p\mathbb{Z}_p$, and by considering indices we must have equality. This L is indeed a solution. In the second case, note that L is determined by its quotient modulo pL_0, and that $L/pL_0 \subset L_1/pL_0 \cong (\mathbb{Z}/p\mathbb{Z})^2$ has index p. There are $p+1$ such subgroups, and the corresponding L satisfy the conditions. This proves the first statement.

For the second integral, take instead $\mathbf{a} = (1,0,0)$ and apply Lemma 3.98. Using as before the observation that $L \supset p\mathbb{Z}_p$ and using that $[L_0 : L_1] = 1$, the possibilities for L_1 are now:

- L_0,
- $p^{-1}\mathbb{Z}_p \oplus \mathbb{Z}_p \oplus p\mathbb{Z}_p$ and
- $p^{-2}\mathbb{Z}_p \oplus p\mathbb{Z}_p \oplus p\mathbb{Z}_p$,

up to permutation of the factors. When $L_1 = L_0$ we obtain $\deg(\mu_{\mathfrak{a}}(p))$ pairs (L_0, L). In the second case we use that $L \subset L_0 \cap L_1$ and observe that equality must occur, giving one solution L. In the third case the condition $L \subset L_0 \cap L_1$ forces $L \subset \mathbb{Z}_p \oplus p\mathbb{Z}_p \oplus p\mathbb{Z}_p$, and this is impossible. The second statement follows, taking into account permutations.

The computation of the last integral also follows from Lemma 3.98, with $\mathbf{b} = \mathbf{0}$. We must count the adapted lattices of index p in L_0, and there are 3 of them. □

We now return to the notation from the previous subsections, and let $\mathbf{G}$ be a form of $\mathbf{PGL}_3$. Let E be the minimal splitting field of $\mathbf{H}$ and choose an isomorphism $\sigma : \mathbf{G}_E \to \mathbf{PGL}_{3,E}$ that sends $\mathbf{H}$ to the diagonal torus. Define $\mathcal{P}_{\text{good}}$ to be the set of primes with the following properties:

1. $p \nmid D'$, with D' as in Proposition 3.91;
2. p splits in E; equivalently, E embeds in $\mathbb{Q}_p$;
3. $\sigma(K_p) = \mathbf{PGL}_3(\mathbb{Z}_p)$.

The condition that p splits in E is equivalent to saying that $\mathbf{H}$ is split over $\mathbb{Q}_p$. In particular we have for $p \in \mathcal{P}_{\text{good}}$ that $G_{\mathbb{Q}_p}$ is split, and therefore isomorphic to $\mathrm{PGL}_{3,\mathbb{Q}_p}$, by Lemma 3.76. This is also apparent from the fact that we may extend σ from E to $\mathbb{Q}_p$.

Lemma 3.101. *We have*

$$\sum_{\substack{p \in \mathcal{P}_{\text{good}} \\ p \leq x}} \log p = \frac{x}{[E:\mathbb{Q}]} + o(x)\,,$$

where E is as above.

Proof. The first condition defining $\mathcal{P}_{\text{good}}$ does not influence the asymptotic, and modulo the second, neither does the third. Meaning, the statement is that the set of primes that split in E has natural density $1/[E:\mathbb{Q}]$. Because the splitting field E is Galois, this follows from Chebotarev's density theorem with natural density [31, Theorem 4]. □

Proof of Proposition 3.92. Let $M \geq 2$. Fix a real number $c > 0$, which we will optimize later. Let $c_1 > 0$ be a real number that we will later assume to be sufficiently small. Let $M_1 = c_1 \log M$. When p is a prime, define

$$a_p = \begin{cases} \frac{c}{p} & \text{when } p \in \mathcal{P}_{\text{good}} \text{ and } p \leq M_1\,, \\ 0 & \text{otherwise.} \end{cases}$$

For $p \in \mathcal{P}_{\text{good}}$ we may define the elementary Hecke operators $\tau(\mathfrak{a}, p) \in \mathcal{H}_p$ as in §3.7.1, through the isomorphism σ with $\mathbf{PGL}_3(\mathbb{Q}_p)$, where $\mathcal{H}_p = C_c^\infty(K_p \backslash \mathbf{G}(\mathbb{Q}_p)/K_p)$. Define $\mathbf{a} = (1,0,0)$ and for $p \in \mathcal{P}_{\text{good}}$ define

$$\omega_p = a_p \tau(\mathfrak{a}, p) + a_p \tau(\mathfrak{a}, p)^* \in \mathcal{H}_p^{\leq 2}\,.$$

Define

$$\omega = \sum_{\substack{n \leq M \\ \text{squarefree}}} \prod_{p \mid n} \omega_p \in \mathcal{H}^{\leq 2}_{\mathcal{P}_{\text{good}}, M}\,, \tag{3.102}$$

and finally

$$k_f = \omega * \omega^* = \sum_{\substack{n,m \le M \\ \text{squarefree}}} \prod_{\substack{p|n \\ q|m}} \omega_p * \omega_q^* = \sum_{\substack{n \le M \\ \text{squarefree}}} \prod_{p|n} (\omega_p + \omega_p^* + \omega_p * \omega_p^*)\,, \tag{3.103}$$

where the second equality holds by grouping pairs (n, m) with the same least common multiple. (In fact $\omega_p^* = \omega_p$.) Clearly $k_f \in \mathcal{H}_{S,M}^{\le 4}$.

It is clear that k_f satisfies (3.90). Indeed, it is a self-convolution and therefore has nonnegative eigenvalues. It takes nonnegative values because the a_p are nonnegative, and $k_f(1) \ge 1$ thanks to the term for $n = m = 1$.

We have that $\|k_f\|_\infty \ll M^A$ for some $A > 0$. This can be seen either by expanding the convolution $\omega_p \omega_p^*$ in terms of elementary Hecke operators, or using the same arguments as used in [13, Lemma 4.4].

It remains to show the lower bound in Proposition 3.92. Because $\omega_p(1) = \omega_p^*(1) = 0$, we have

$$k_f(1) = \sum_{\substack{n \le M \\ \text{squarefree}}} \prod_{p|n} \|\omega_p\|_2^2\,.$$

To prove the lower bound, we begin by trivially estimating

$$k_f(1) \le \prod_{p \le M_1} \left(1 + \|\omega_p\|_2^2\right) \tag{3.104}$$

by completing the sum over n to all square-free integers. To compute $\int_{H(\mathbb{A}_f)} k_f$ we integrate (3.103) and use the computations from Lemma 3.100, which give

$$\int_{H(\mathbb{A}_f)} k_f = \sum_{\substack{n \le M \\ \text{squarefree}}} \prod_{p|n} \left(12a_p + (6p + 24)a_p^2 + \|\omega_p\|_2^2\right)\,,$$

where we use that $\|\omega_p\|_2^2 = 2a_p^2 \deg(\mu_p)$. The term $\|\omega_p\|_2^2$ corresponds to the diagonal contribution, and we will want the other terms to be large relative to this. Let $\alpha > 0$. We complete the sum over n to a full product, which introduces an error term that we

estimate using Rankin's trick by introducing a factor $(n/M)^\alpha$ in the resulting error terms.

$$\begin{aligned}\int_{H(\mathbb{A}_f)} k_f &= \prod_{p\le M_1}(1+12a_p+(6p+24)a_p^2+\|\omega_p\|_2) \\ &\quad +O\left(M^{-\alpha}\sum_{\substack{n>M\\ \text{squarefree}}}\prod_{p\mid n}p^\alpha(12a_p+(6p+24)a_p^2+\|\omega_p\|_2^2)\right) \\ &= \prod_{p\le M_1}(1+12a_p+(6p+24)a_p^2+\|\omega_p\|_2^2) \\ &\quad +O\left(M^{-\alpha}\prod_{p\le M_1}\left(1+p^\alpha\left(12a_p+(6p+24)a_p^2+\|\omega_p\|_2^2\right)\right)\right)\end{aligned} \tag{3.105}$$

Applying the inequality $\frac{1+x}{1+y}\le 1+(x-y)\le \exp(x-y)$ (for $x\ge y\ge 0$), the ratio of the error term to the main term in (3.105) is at most

$$M^{-\alpha}\exp\left(\sum_{p\le M_1}(p^\alpha-1)(12a_p+(6p+24)a_p^2+\|\omega_p\|_2^2)\right). \tag{3.106}$$

We want this to be strictly less than 1. By the mean value theorem, $p^\alpha-1\le \alpha p^\alpha\log p$. Using the bound $a_p\ll 1/p$, we can thus bound (3.106) by

$$\exp\left(-\alpha\log M+O\left(\sum_{p\le M_1}\alpha p^\alpha\log p\right)\right).$$

Now choose $\alpha=1/\log M_1$, so that by Chebyshev's estimates this is at most

$$\exp\left(-\alpha(\log M+O(M_1))\right).$$

By choosing c_1 sufficiently small, $M_1=c_1\log M$ is small enough for the above expression to be at most $1/2$ (say).

It remains to find a lower bound for the ratio of the main term in (3.105) to the

right-hand side of (3.104). Using the bound $1 + x \geq \exp(x + O(x^2))$, we have

$$\begin{aligned}\frac{\int_{H(\mathbb{A}_f)} k_f}{k_f(1)} &\gg \prod_{p \leq M_1} \frac{1 + 12a_p + (6p+24)a_p^2 + \|\omega_p\|_2^2}{1 + \|\omega_p\|_2^2} \\ &\gg \exp\left(\sum_{p \leq M_1} \left(\frac{12a_p + (6p+24)a_p^2}{1 + \|\omega_p\|_2^2} + O(1/p^2)\right)\right) \\ &\gg \exp\left(\sum_{\substack{p \leq M_1 \\ p \in \mathcal{P}_{\mathrm{good}}}} \left(\frac{12c + 6c^2}{1 + 2c^2} \cdot \frac{1}{p} + O(1/p^2)\right)\right) \\ &\gg \exp\left(\frac{1}{[E:\mathbb{Q}]} \cdot \frac{12c + 6c^2}{1 + 2c^2} (\log\log M_1)(1 + o(1))\right).\end{aligned}$$

In the last step we have used Lemma 3.101 together with Abel's summation formula to deduce a Mertens type result from the PNT-type result. The rational function in c that appears is maximal for $c = 1$, where it takes the value 6. If we recall that $M_1 = c_1 \log M$, this is the lower bound stated in Proposition 3.92. □

3.7.3 Upper bounds

We prove that the construction in §3.7.2 is in a sense optimal, by giving an upper bound for the quotient

$$\frac{\int_{H(\mathbb{A}_f)} k_f}{k_f(1)}$$

modulo certain restrictions on k_f. The main result is Proposition 3.111. When $p \in \mathcal{P}_{\mathrm{good}}$ is a prime, we may identify $\mathbf{G}_{\mathbb{Q}_p}$ with $\mathbf{PGL}_{3,\mathbb{Q}_p}$ using the isomorphism σ. To simplify the notation, we will not keep track of the set $\mathcal{P}_{\mathrm{good}}$ here, and state the global bounds instead for the group $\mathbf{PGL}_3$, for which we define truncated Hecke algebras in the same way as in §3.5.7. In fact, we will give statements that are valid more generally for $\mathbf{PGL}_n$ with $n \geq 3$. In either case, $\mathbf{H}$ denotes the diagonal torus.

We will first prove the following key bound. While it is valid for all $n \geq 3$, it is likely not the strongest possible result, see Remark 3.112. We use the local notation from §3.7.1.

Proposition 3.107. *Let $n \geq 3$ and let $\mathbf{a}, \mathbf{b} \in \mathbb{N}^n$ be decreasing tuples with $a_n = b_n = 0$. Then*

$$\left(\int_{H_p} \tau(\mathbf{a}, p) * \tau(-\mathbf{b}, p)\right) - \delta_{\mathbf{a},\mathbf{b}} \deg(\mu_{\mathbf{a}}(p)) \ll \frac{1}{p} \deg(\mu_{\mathbf{a}}(p))^{1/2} \deg(\mu_{\mathbf{b}}(p))^{1/2},$$

where the implicit constant is allowed to depend on $\mathbf{a}$ and $\mathbf{b}$ but not on p.

The proof of Proposition 3.107 uses various arguments, one of which is the following trivial bound.

Lemma 3.108. *The conclusion of Proposition 3.107 holds when* $\deg(\mu_{\mathbf{b}}(p)) \geq p^2 \cdot \deg(\mu_{\mathbf{a}}(p))$.

Proof. We may bound the integral using Lemma 3.98, by counting pairs (L_1, L) as in the statement of the lemma. First, because $\mathbf{a}$ and $\mathbf{b}$ may be assumed fixed, the number of possibilities for L_1 is bounded. To be precise, the inclusions $L_1 \supset L \supset p^{a_1}L_0$ and $p^{b_1}L_1 \subset L \subset L_0$ imply that there are at most $(a_1 + b_1 + 1)^n$ possibilities for L_1. Second, the number of possibilities for L is trivially bounded by $\deg(\mu_{\mathbf{a}}(p))$. Therefore the left-hand side in Proposition 3.107 (even without the subtraction) is bounded up to a constant by

$$\deg(\mu_{\mathbf{a}}(p)) \leq \frac{1}{p} \deg(\mu_{\mathbf{a}}(p))^{1/2} \deg(\mu_{\mathbf{b}}(p))^{1/2} . \qquad \square$$

To prove Proposition 3.107, it suffices to prove an upper bound of the form

$$\left(\int_{H_p} \tau(\mathbf{a}, p) * \tau(-\mathbf{b}, p) \right) - \delta_{\mathbf{a},\mathbf{b}} \deg(\mu_{\mathbf{a}}(p)) \ll \frac{1}{p^2} \cdot \max\left(\deg(\mu_{\mathbf{a}}(p)), \deg(\mu_{\mathbf{b}}(p))\right) .$$

Indeed, by symmetry and by Lemma 3.108 we must only consider the situation where

$$\deg(\mu_{\mathbf{a}}(p)) \leq \deg(\mu_{\mathbf{b}}(p)) \leq p^2 \deg(\mu_{\mathbf{a}}(p)) ,$$

in which case the above bound is at least as strong as what is needed.

The other type of argument we will use is the following. Let G be any (abstract) group and $f : X \to Y$ a G-equivariant map between finite transitive G-sets. Then the preimages $f^{-1}(y)$ have the same cardinality. In particular, when $S \subset X$, a bound of the form $|f(S)| \leq \delta |Y|$ implies $|S| \leq \delta |X|$. We will apply this principle with $G = \mathrm{GL}_n(\mathbb{Z}/p^a\mathbb{Z})$, X a set of subgroups of $(\mathbb{Z}/p^a\mathbb{Z})^n$, and f reduction mod p.

Lemma 3.109. *Let* $\mathbf{a} \in \mathbb{N}^n$ *be a decreasing tuple with* $a_n = 0$. *Let* $1 \leq m \leq n$ *and consider the subgroups* $L \subset (\mathbb{Z}/p^{a_1}\mathbb{Z})^n$ *with the following properties:*

- *The quotient* $(\mathbb{Z}/p^{a_1}\mathbb{Z})^n/L$ *has invariant factors given by* $\mathbf{a}$.
- $L \subset (p\mathbb{Z}/p^{a_1}\mathbb{Z})^m \oplus (\mathbb{Z}/p^{a_1}\mathbb{Z})^{n-m}$.

Let $d = \#\{i : a_i = 0\}$. *Then the number of such subgroups is bounded up to a constant by* $p^{-md} \deg(\mu_{\mathbf{a}}(p))$, *where the constant does not depend on* $\mathbf{a}$ *nor* p.

Proof. The group $G = \mathrm{GL}_n(\mathbb{Z}/p^{a_1}\mathbb{Z})$ acts transitively on the subgroups L whose quotient has invariant factors given by $\mathbf{a}$. There are precisely $\deg(\mu_{\mathbf{a}}(p))$ of those. The reduction mod p of all such L is a subspace of $\mathbb{F}_p^n$ of dimension d. The group G acts transitively

on subspaces of given dimension in a way compatible with reduction mod p. There are $\asymp p^{d(n-d)}$ such subspaces. On the other hand, the subgroups L as in the statement have reduction mod p lying in a fixed $(n-m)$-dimensional subspace. When $n-m<d$ there is nothing to prove. When $n-m\geq d$, the number of such subspaces is $\asymp p^{d(n-m-d)}$. We conclude that the number of subgroups we want to count, is bounded up to a constant by

$$\frac{p^{d(n-m-d)}}{p^{d(n-d)}}\deg(\mu_{\mathbf{a}}(p))=p^{-md}\deg(\mu_{\mathbf{a}}(p))\,. \qquad \square$$

Finally, we will use the perfect pairing on $(\mathbb{Z}/p^a\mathbb{Z})^n$, which provides a notion of duality. Specifically, denote by $\langle\cdot,\cdot\rangle$ the component-wise pairing in the standard basis. When $L\subset(\mathbb{Z}/p^a\mathbb{Z})^n$, define $L^*=\{x:\langle x,L\rangle=0\}$. Then $(L^*)^*=L$, duality reverses inclusions and if $(\mathbb{Z}/p^a\mathbb{Z})^n/L$ has invariant factors (p^{a_i}) then $(\mathbb{Z}/p^a\mathbb{Z})^n/L$ has invariant factors (p^{a-a_i}). For an explicit example, the dual of $\bigoplus p^{a-a_i}\mathbb{Z}/p^a\mathbb{Z}$ is $\bigoplus p^{a_i}\mathbb{Z}/p^a\mathbb{Z}$.

We have the following dual version of Lemma 3.109.

Lemma 3.110. *Let $\mathbf{a}\in\mathbb{N}^n$ be a decreasing tuple with $a_n=0$. Let $1\leq m\leq n$ and consider the subgroups $L\subset(\mathbb{Z}/p^{a_1}\mathbb{Z})^n$ with the following properties:*

- *The quotient $(\mathbb{Z}/p^{a_1}\mathbb{Z})^n/L$ has invariant factors given by $\mathbf{a}$.*
- *$L\supset(p^{a_1-1}\mathbb{Z}/p^{a_1}\mathbb{Z})^m\oplus\{0\}^{n-m}$.*

Let $d=\#\{i:a_i=a_1\}$. Then the number of such subgroups is bounded up to a constant by $p^{-md}\deg(\mu_{\mathbf{a}}(p))$, where the constant does not depend on $\mathbf{a}$ nor p.

Proof. For L as in the statement, we have that L^* satisfies the conditions in Lemma 3.109. The only observation we have to make is that the tuple $\mathbf{a}^*:=(a_1-a_n,\dots,a_1-a_2,a_1-a_1)$ satisfies $\deg(\mu_{\mathbf{a}^*}(p))=\deg(\mu_{\mathbf{a}}(p))$; this is just the statement that $\deg(\mu_{-\mathbf{a}}(p))=\deg(\mu_{\mathbf{a}}(p))$. $\square$

With these three ingredients we are ready to prove the key bound.

Proof of Proposition 3.107. By symmetry we may assume that $\deg(\mu_{\mathbf{a}}(p))\leq\deg(\mu_{\mathbf{b}}(p))$. To bound the integral we use Lemma 3.98. We must bound the number of pairs (L_1,L) with L_1 adapted and different from L_0, and $L\subset L_0\cap L_1$ for which L_0/L has invariant factors given by $\mathbf{a}$ and those of L_1/L are given by $\mathbf{b}$. Write $L_1=\bigoplus p^{t_i}\mathbb{Z}_p$ with the $t_i\in\mathbb{Z}$. As in the proof of Lemma 3.108 there are only finitely many possibilities for L_1, so we may assume L_1 is fixed.

Suppose there exists $t_i>0$. Then up to permutation of factors, L lies in $p\mathbb{Z}_p\bigoplus\mathbb{Z}_p^{n-1}$, so that the subgroup $L/p^{a_1}L_0\subset L_0/p^{a_1}L_0\cong(\mathbb{Z}/p^{a_1}\mathbb{Z})^n$ satisfies the conditions of Lemma 3.109 with $d,m\geq1$. Therefore there are at most (up to a multiplicative constant) $p^{-1}\deg(\mu_{\mathbf{a}}(p))$ possibilities for L in this case, and this bound is good enough because $\deg(\mu_{\mathbf{a}}(p))\leq\deg(\mu_{\mathbf{b}}(p))$.

We may now assume that all $t_i \leq 0$, so that $L_0 \subset L_1$. If two distinct $t_i < 0$, then we may apply Lemma 3.109 again, this time to $L/p^{b_1}L_1 \subset L_1/p^{b_1}L_1$ and with $m \geq 2$, $d \geq 1$. We find that there are at most $p^{-2}\deg(\mu_{\mathbf{b}}(p))$ possibilities for L in this case, and this bound is good enough as remarked below Lemma 3.108.

We may now assume that a single $t_i < 0$ and all others are 0. If $b_1 \leq a_1$, then L contains $p^{a_1+t_i}e_i$ and in particular $p^{a_1-1}e_i$. We may then apply Lemma 3.110 to $L/p^{a_1}L_0 \subset L_0/p^{a_1}L_0$, with $m, d \geq 1$ and conclude that there are at most $p^{-1}\deg(\mu_{\mathbf{a}}(p))$ possibilities for L. If $b_1 > a_1$, then L_0 contains all the elements $p^{b_1-1}e_j$ with $j \neq i$. Because $n \geq 3$, there are at least two of these. We may now apply Lemma 3.110 to $L/p^{b_1}L_1 \subset L_1/p^{b_1}L_1$ with $m \geq 2$ and $d \geq 1$, and conclude that there are at most (up to a constant) $p^{-2}\deg(\mu_{\mathbf{b}}(p))$ possibilities for L in this case. Again, this is sufficient by the comment below Lemma 3.108. □

Let $\mathcal{P}$ denote the set of all primes.

Proposition 3.111. *Let $n \geq 3$, $\kappa > 0$ and $M \geq 3$. For every nonzero $\omega \in \mathcal{H}^{\leq\kappa}_{\mathcal{P},M}$ the convolution operator $k_f = \omega * \omega^*$ satisfies*

$$\frac{\int_{\mathbf{H}(\mathbb{A}_f)} k_f}{k_f(1)} \ll (\log\log M)^C\,,$$

for a constant C that is allowed to depend on n and κ.

Proof. We may find a finite family of squarefree integers $(n_i)_{i\in I} \in [0, M]$ and for every prime $p \mid n_i$ a scalar multiple of elementary Hecke operator $\omega_{i,p} \in \mathcal{H}^{\leq\kappa}_p$, such that $\omega = \sum_{i\in I}\prod_{p|n_i}\omega_{i,p}$. Moreover, we may assume that the $\prod_{p|n_i}\omega_{i,p}$ have disjoint supports. If X_κ denotes the set of cocharacters of $\mathbf{H}_{\overline{\mathbb{Q}}}$ of norm at most κ, then $|X_\kappa|$ is bounded. If $\omega(m)$ denotes the number of prime divisors of m, then every $m \leq M$ occurs at most $|X_\kappa|^{\omega(m)}$ times as an integer n_i in the family.

Let C be the largest implicit constant in Proposition 3.107 when $\mathbf{a}$ and $\mathbf{b}$ run trough the tuples with $\mu_{\mathbf{a}}, \mu_{\mathbf{b}} \in X_\kappa$.

When $p \nmid n_i$, define $\omega_{i,p} = 1_{K_p}$. We then have

$$k_f = \sum_{i,j\in I}\prod_{\substack{p|n_i\\ q|n_j}}\omega_{i,p} * \omega^*_{j,q} = \sum_{i,j\in I}\prod_{p|n_in_j}\omega_{i,p} * \omega^*_{j,p}\,.$$

The disjointness of the supports of the $\prod_{p|n_i}\omega_{i,p}$ implies

$$k_f(1) = \sum_{i\in I}\prod_{p|n_i}\|\omega_{i,p}\|_2^2\,.$$

Using Proposition 3.107 we have

$$\begin{aligned}\int_{\mathbf{H}(\mathbb{A}_f)} k_f &= \sum_{i,j\in I}\prod_{p\mid n_in_j}\int_{H_p}(\omega_{i,p}*\omega_{j,p}^*)\\ &\le \sum_{i,j\in I}\prod_{p\mid n_in_j} Cp^{-1}\|\omega_{i,p}\|_2\|\omega_{j,p}\|_2\\ &= \sum_{i,j\in I}\operatorname{lcm}(n_i,n_j)^{-1}C^{\omega(\operatorname{lcm}(n_in_j))}\prod_{p\mid n_in_j}\|\omega_{i,p}\|_2\|\omega_{j,p}\|_2\\ &\le \sum_{d\le M}\frac{d}{C^{\omega(d)}}\left(\sum_{d\mid n_i}\frac{C^{\omega(n_i)}}{n_i}\prod_{p\mid n_i}\|\omega_{i,p}\|_2\right)^2,\end{aligned}$$

where in the last equality we have written $\operatorname{lcm}(n_in_j) = n_in_j/d$ with $d = \gcd(n_i,n_j)$, and then extended the sum to run over all $d \mid n_i, n_j$. By Cauchy–Schwarz, this is bounded by

$$\sum_{d\le M}\frac{d}{C^{\omega(d)}}\left(\sum_{d\mid n_i}\frac{C^{2\omega(n_i)}}{n_i^2}\right)\left(\sum_{d\mid n_i}\prod_{p\mid n_i}\|\omega_{i,p}\|_2^2\right).$$

It follows that

$$\frac{\int_{\mathbf{H}(\mathbb{A}_f)} k_f}{k_f(1)} \le \sup_{\substack{n\le M\\ \square\text{-free}}}\sum_{\substack{d,n_i\\ d\mid n,n_i}}\frac{d}{C^{\omega(d)}}\frac{C^{2\omega(n_i)}}{n_i^2}.$$

Because every m occurs at most $|X_\kappa|^{\omega(m)}$ times in the family $(n_i)_{i\in I}$, this is at most

$$\begin{aligned}&\ll_{\kappa,\epsilon}\sup_{\substack{n\le M\\ \square\text{-free}}}\sum_{\substack{d\mid n,m\\ m\ \square\text{-free}}}\frac{d}{C^{\omega(d)}}\frac{(C^2|X_\kappa|)^{\omega(m)}}{m^2}\\ &= \sup_{\substack{n\le M\\ \square\text{-free}}}\sum_{d\mid n}\frac{(C|X_\kappa|)^{\omega(d)}}{d}\sum_{m\ \square\text{-free}}\frac{(C^2|X_\kappa|)^{\omega(m)}}{m^2}\\ &\ll \sup_{\substack{n\le M\\ \square\text{-free}}}\sum_{d\mid n}\frac{(C|X_\kappa|)^{\omega(d)}}{d},\end{aligned}$$

because the sum over m is convergent. The latter expression is largest when n has the smallest possible prime factors. So take $n = \prod_{p\le x} p$ for some $x > 1$. Then $n \le M$

implies $x \ll \log M$ by Chebyshev's estimate, and we have

$$\begin{aligned}\sum_{d|n} \frac{(C|X_\kappa|)^{\omega(d)}}{d} &= \prod_{p\leq x} \left(1 + \frac{C|X_\kappa|}{p}\right) \\ &\ll (\log x)^{C|X_\kappa|} \\ &\ll (\log\log M)^{C|X_\kappa|},\end{aligned}$$

where we have used the Mertens' theorem in the first estimate. □

Remark 3.112. It is likely that a stronger version of Proposition 3.107 is still true. Namely, when $n \geq 4$ we expect that a a similar bound holds with the power of p in the right-hand side replaced by $p^{-3/2}$. Moreover, when $n = 3$ we expect that this can also be shown, except in the situation of Lemma 3.100. We have partial proofs of these statements, which use mostly the same arguments as in the proof of Proposition 3.107, and contains lots of casework. However a few cases remain, where we notably require bounds for specific Hall polynomials. We hope to settle this stronger version in the near future.

The stronger version would imply that the upper bound in Proposition 3.111 can be replaced by 1 (no growth at all) for $n \geq 4$. Indeed, the exponent 3/2 is then propagated throughout the proof until the very last lines, where we may then use that $\prod_{p\leq X} \left(1 + p^{-3/2}\right)$ is bounded.

Remark 3.113. When $n = 3$, Proposition 3.107 is not as strong as we would like in different ways. It would be desirable to remove or make explicit the dependence on κ and to make the power of $\log\log M$ match with the exponent in Proposition 3.92. Such bounds should follow from the stronger version of Proposition 3.107 that we expect to hold.

Bibliography

[1] M. Abért, N. Bergeron, and E. Le Masson. Eigenfunctions and random waves in the Benjamini–Schramm limit. preprint at `https://arxiv.org/abs/1810.05601`, 2018.

[2] C. Aistleitner, K. Mahatab, and M. Munsch. Extreme Values of the Riemann Zeta Function on the 1-Line. *Int. Math. Res. Not*, 2019(22):6924–6932, November 2019.

[3] R. Aurich and F. Steiner. Exact theory for the quantum eigenstates of the Hadamard–Gutzwiller model. *Phys. D*, 48(2–3):445–470, March 1991.

[4] R. Aurich and F. Steiner. Statistical properties of highly excited quantum eigenstates of a strongly chaotic system. *Phys. D*, 64(1–3):185–214, April 1993.

[5] M. V. Berry. Regular and irregular semiclassical wavefunctions. *J. Phys. A*, 10(12):2083–2091, 1977.

[6] V. Blomer and F. Brumley. Simultaneous equidistribution of toric periods and fractional moments of L-functions. preprint at `https://arxiv.org/abs/2009.07093`.

[7] V. Blomer and J. Buttcane. On the subconvexity problem for L-functions on GL(3). *Ann. Sci. Éc. Norm. Supér.*, 53(6):1441–1500, 2020.

[8] V. Blomer, É. Fouvry, E. Kowalski, Ph. Michel, D. Milićević, and W. Sawin. *The second moment theory of families of L-functions*, pages 93–119. Memoirs of the American Mathematical Society. American Mathematical Society, to appear.

[9] A. Bondarenko and K. Seip. Large greatest common divisor sums and extreme values of the Riemann zeta function. *Duke Math. J.*, 166(9):1685–1701, 2017.

[10] A. Borel. *Linear algebraic groups*, volume 126 of *Graduate Texts in Mathematics*. Springer-Verlag, second enlarged edition, 1991.

[11] A. Borel and Harish-Chandra. Arithmetic subgroups of algebraic groups. *Ann. of Math.*, 75(3):485–535, 1962.

[12] A. Borel and J. Tits. *Groupes réductifs*. Publications mathématiques de l'I.H.É.S., 1965.

[13] F. Brumley and S. Marshall. Lower bounds for Maass forms on semisimple groups. *Compos. Math.*, 156(5):959–1003, 2020.

[14] X. Chen and C. D. Sogge. On integrals of eigenfunctions over geodesics. *Proc. Amer. Math. Soc.*, 143(1):151–161, 2015.

[15] Y. Colin de Verdière. Spectre du laplacien et longueur des géodésiques périodiques. II. *Compos. Math.*, 27(2):159–184, 1973.

[16] J. B. Conrey, D. W. Farmer, J. P. Keating, M. O. Rubinstein, and N. C. Snaith. Integral moments of L-functions. *Proc. London Math. Soc.*, 91(1):33–104, 2005.

[17] R. de la Bretèche and G. Tenenbaum. Gál sums and applications (in French). *Proc. Lond. Math. Soc. (3)*, 119(3):104–134, 2019.

[18] H. Donnelly. Exceptional sequences of eigenfunctions for hyperbolic manifolds. *Proc. Amer. Math. Soc.*, 135(5):1551–1555, 2007.

[19] J. Duistermaat, J. Kolk, and V. Varadarajan. Spectra of compact locally symmetric manifolds of negative curvature. *Invent. Math.*, 52:27–93, 1979.

[20] J. Duistermaat, J. Kolk, and V. Varadarajan. Functions, flows and oscillatory integrals on flag manifolds and conjugacy classes in real semisimple Lie groups. *Compos. Math.*, 49(3):309–398, 1983.

[21] S. Dyatlov and M. Zworski. Quantum ergodicity for restrictions to hypersurfaces. *Nonlinearity*, 26(1):35–52, 2013.

[22] P. B. Eberlein. *Geometry of Nonpositively Curved Manifolds*. Chicago Lectures in Mathematics. The University of Chicago Press, 1996.

[23] M. Eichler. *Lectures on modular correspondences*, volume 9 of *Lectures on Mathematics and Physics: Mathematics*. Tata Institute of Fundamental Research, Mumbai, India, reissued edition, 1965.

[24] M. Einsiedler, E. Lindenstrauss, Ph. Michel, and A. Venkatesh. Distribution of periodic torus orbits on homogeneous spaces. *Duke Math. J.*, 148(1):119–174, 2009.

[25] D. W. Farmer, S. M. Gonek, and C. P. Hughes. The maximum size of L-functions. *J. Reine Angew. Math.*, 609:215–236, August 2007.

[26] R. Gangolli. On the Plancherel formula and the Paley-Wiener theorem for spherical functions on semisimple Lie groups. *Ann. of Math.*, 93(1):150–165, 1971.

[27] S. Garibaldi and Ph. Gille. Algebraic groups with few subgroups. *J. London Math. Soc.*, 80(2):405–430, 2009.

[28] D. Goldfeld, J. Hoffstein, and D. Lieman. Appendix: An effective zero-free region. *Ann. of Math. (2)*, 140(1):177–181, July 1994.

[29] A. Granville and K. Soundararajan. Extreme values of $|\zeta(1+it)|$. In *The Riemann zeta functionand related themes: papers in honour of Professor K. Ramachandra*, volume 2 of *Ramanujan Math. Soc. Lect. Notes Ser.*, pages 65–80, 2006.

[30] Harish-Chandra. Harmonic analysis on real reductive groups I: the theory of the constant term. *J. Funct. Anal.*, 19:104–204, 1975.

[31] H. Heilbronn. Zeta-Functions and L-Functions. In J. W. W. Cassels and A. Fröhlich, editors, *Algebraic Number Theory*. Academic Press, 1967.

[32] D. A. Hejhal and B. N. Rackner. On the topography of Maass waveforms for PSL(2,Z). *Exp. Math.*, 1(4):275–305, 1992.

[33] S. Helgason. *Differential geometry, Lie groups and symmetric spaces*. Academic Press, 1978.

[34] S. Helgason. *Groups and geometric analysis*. Academic Press, 1984.

[35] T. Hilberdink. An arithmetical mapping and applications to ω-results for the Riemann zeta function. *Acta Arith.*, 139(4):341–367, 2009.

[36] H. Iwaniec. *Spectral methods of automorphic forms*, volume 53 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island, second edition, 2002.

[37] H. Iwaniec and P. Sarnak. L^∞ norms of eigenfunctions on arithmetic surfaces. *Ann. of Math. (2)*, 141:301–320, 1995.

[38] H. Iwaniec and P. Sarnak. Perspectives on the analytic theory of L-functions. *Geom. Funct. Anal.*, Special volume:705–741, 2000.

[39] J.-P. Kahane. *Some random series of functions*, volume 5 of *Cambridge studies in advanced mathematics*. Cambridge University Press, Cambridge, United Kingdom, second edition, 1985.

[40] A. W. Knapp. *Representation Theory of Semisimple Group*. Princeton University Press, 1989.

[41] A. W. Knapp. *Lie Groups Beyond an Introduction*. Birkhauser, 2 edition, 2002.

[42] M.-A. Knus, A. Merkujev, M. Rost, and J.-P. Tignol. *The Book of Involutions*. Number 44 in Colloquium publications. American Mathematical Society, 1998.

[43] E. Lapid and O. Offen. Compact unitary periods. *Compos. Math.*, 143:323–338, 2007.

[44] J. M. Lee. *Introduction to Smooth Manifolds*. Number 218 in Graduate Texts in Mathematics. Springer, second edition, 2013.

[45] N. Levinson. Ω-theorems for the Riemann zeta-function. *Acta Arith.*, 20:317–330, 1972.

[46] S. Marshall. Restrictions of SL_3 Maass forms to maximal flat subspaces. *Int. Math. Res. Not.*, 2015(16):6988–7015, 2015.

[47] S. Marshall. Geodesic restrictions of arithmetic eigenfunctions. *Duke Math. J.*, 165(3):463–508, 2016.

[48] S. Marshall. L^p norms of higher rank eigenfunctions and bounds for spherical functions. *J. Eur. Math. Soc.*, 18(7):1437–1493, 2016.

[49] J. Mercer. Functions of positive and negative type, and their connection the theory of integral equations. *Phil. Trans. Roy. Soc. London (A)*, 209:415–446, 1909.

[50] D. Milićević. Large values of eigenfunctions on arithmetic hyperbolic surfaces. *Duke Math. J.*, 155(2):365–401, 2010.

[51] D. Milićević. Large values of eigenfunctions on arithmetic hyperbolic 3-manifolds. *Geom. Funct. Anal.*, 21(6):1375–1418, December 2011.

[52] H. L. Montgomery. Extreme values of the Riemann zeta function. *Comment. Math. Helv.*, 52:511–518, 1977.

[53] D. W. Morris. *Introduction to arithmetic groups*. Deductive Press, 2015.

[54] L. Nachbin. *The Haar Integral*. D. Van Nostrand Company, 1965.

[55] A. L. Onishchik and E. B. Vinberg. *Lie groups and algebraic groups*. Springer Series in Soviet Mathematics. Springer-Verlag, 1980.

[56] Ł. Pańkowski and J. Steuding. Extreme values of L-functions from the Selberg-class. *Int. J. Number Theory*, 9(5):1113–1124, 2013.

[57] R. S. Pierce. *Associative Algebras*. Number 88 in Graduate Texts in Mathematics. Springer-Verlag, 1982.

[58] V. Platonov and A. Rapinchuk. *Algebraic Groups and Number Theory*. Academic Press, 1994.

[59] A. A. Popa. Central values of Rankin L-series over real quadratic fields. *Compos. Math.*, 142(4):811–866, July 2006.

[60] M. Ratner. The central limit theorem for geodesic flows on n-dimensional manifolds of negative curvature. *Israel J. Math.*, 16:181–197, 1973.

[61] A. Reznikov. A uniform bound for geodesic periods of eigenfunctions on hyperbolic surfaces. *Forum Math.*, 27(3):1569–1590, 2015.

[62] Z. Rudnick and P. Sarnak. The behaviour of eigenstates of arithmetic hyperbolic manifolds. *Commun. Math. Phys.*, 161:195–213, 1994.

[63] R. Salem and A. Zygmund. Some properties of trigonometric series whose terms have random signs. *Acta Math.*, 91:245–301, 1954.

[64] J. W. Sands. Generalization of a theorem of Siegel. *Acta Arith.*, 58(1):47–56, 1991.

[65] P. Sarnak. Arithmetic quantum chaos. In *The Schur lectures*, volume 8 of *Israel mathematical conference proceedings*, Providence, Rhode Island, 1995. American Mathematical Society.

[66] P. Sarnak. Reciprocal geodesics. *Clay Math. Proc.*, 7:217–237, 2007.

[67] Ya. G. Sinai. The central limit theorem for geodesic flows on manifolds of constant negative curvature. *Proc. USSR Acad. Sci.*, 133(6):1303–1306, 1960. in Russian.

[68] K. Soundararajan. Extreme values of zeta and L-functions. *Math. Ann.*, 342:467–486, 2008.

[69] E. M. Stein. *Harmonic analysis: real-variable methods, orthogonality, and oscillatory integrals.* Princeton Mathematical Series. Princeton University Press, 1993.

[70] T. Tamagawa. On Selberg's trace formula. *J. Fac. Sci. Univ. Tokyo*, pages 363–386, 1960.

[71] J. Tits. Reductive groups over local fields. In *Automorphic forms, representations, and L-functions*, 1979.

[72] J. A. Toth and S. Zelditch. Quantum ergodic restriction theorems. I: interior hypersurfaces in domains with ergodic billiards. *Ann. Henri Poincaré*, 13(4):599–670, May 2012.

[73] J. A. Toth and S. Zelditch. Quantum ergodic restriction theorems: manifolds without boundary. *Geom. Funct. Anal.*, 23(2):715–775, April 2013.

[74] M. Vignèras. *Arithmetic of quaternion algebras (in French)*, page 104. Number 800 in Lecture Notes in Mathematics. Springer-Verlag, Heidelberg, Germany, 1980.

[75] J.-L. Waldspurger. On the values of certain automorphic L-functions and their center of symmetry (in French). *Compos. Math.*, 54:173–242, 1985.

[76] D. Werner. *Funktionalanalysis*. Springer Spektrum, 8. auflage edition, 2018.

[77] H. Whitney. *Complex analytic varieties*. Addison-Wesley, 1972.

[78] M. Young. The quantum unique ergodicity conjecture for thin sets. *Adv. Math.*, 286:958–1016, January 2016.

[79] M. Young. Equidistribution of Eisenstein series on geodesic segments. *Adv. Math.*, 340:1166–1218, 2018.

[80] S. Zelditch. Kuznecov sum formulae and Szegő limit formulae on manifolds. *Comm. Partial Differential Equations.*, 17(1–2):221–260, 1992.

[81] S. Zelditch. Quantum ergodicity and mixing of eigenfunctions. In *Encyclopedia of mathematical physics*, pages 183–196, Cambridge, Massachusetts, 2006. Academic Press.