MATRIX COEFFICIENTS, COUNTING AND PRIMES FOR ORBITS OF GEOMETRICALLY FINITE GROUPS

AMIR MOHAMMADI AND HEE OH

ABSTRACT. Let $G := \mathrm{SO}(n,1)^\circ$ and $\Gamma < G$ be a geometrically finite Zariski dense subgroup with critical exponent δ bigger than $(n-1)/2$. Under a spectral gap hypothesis on $L^2(\Gamma\backslash G)$, which is always satisfied when $\delta > (n-1)/2$ for $n = 2, 3$ and when $\delta > n-2$ for $n \geq 4$, we obtain an *effective* archimedean counting result for a discrete orbit of Γ in a homogeneous space $H\backslash G$ where H is the trivial group, a symmetric subgroup or a horospherical subgroup. More precisely, we show that for any effectively well-rounded family $\{\mathcal{B}_T \subset H\backslash G\}$ of compact subsets, there exists $\eta > 0$ such that

$$\#[e]\Gamma \cap \mathcal{B}_T = \mathcal{M}(\mathcal{B}_T) + O(\mathcal{M}(\mathcal{B}_T)^{1-\eta})$$

for an explicit measure $\mathcal{M}$ on $H\backslash G$ which depends on Γ. We also apply the affine sieve and describe the distribution of almost primes on orbits of Γ in arithmetic settings.

One of key ingredients in our approach is an effective asymptotic formula for the matrix coefficients of $L^2(\Gamma\backslash G)$ that we prove by combining methods from spectral analysis, harmonic analysis and ergodic theory. We also prove exponential mixing of the frame flows with respect to the Bowen-Margulis-Sullivan measure.

CONTENTS

The authors are supported in part by NSF grants.

1. Introduction

Let $n \geq 2$ and let G be the identity component of the special orthogonal group $\mathrm{SO}(n,1)$. As well known, G can be considered as the group of orientation preserving isometries of the hyperbolic space $\mathbb{H}^n$. A discrete subgroup Γ of G is called *geometrically finite* if the unit neighborhood of its convex core[1] has finite Riemannian volume. As any discrete subgroup admitting a finite sided polyhedron as a fundamental domain in $\mathbb{H}^n$ is geometrically finite, this class of discrete subgroups provides a natural generalization of lattices in G. In particular, for $n = 2$, a discrete subgroup of G is geometrically finite if and only if it is finitely generated.

In the whole introduction, let Γ be a torsion-free geometrically finite, Zariski dense, discrete subgroup of G. We denote by δ the critical exponent of Γ. Note that any discrete subgroup of G with $\delta > (n-2)$ is Zariski dense in G. The main aim of this paper is to obtain *effective* counting results for discrete orbits of Γ in $H\backslash G$, where H is the trivial group, a symmetric subgroup or a horospherical subgroup of G, and to discuss their applications in the affine sieve on Γ-orbits in an arithmetic setting. Our results are formulated under a suitable spectral gap hypothesis for $L^2(\Gamma\backslash G)$ (see Def. 1.1 and 1.3). This hypothesis on Γ is known to be true if the critical exponent δ is strictly bigger than $n-2$. Though we believe that the condition $\delta > (n-1)/2$ should be sufficient to guarantee this hypothesis, it is not yet known in general (see 1.2).

For Γ lattices, i.e., when $\delta = n-1$, both the effective counting and applications to an affine sieve have been extensively studied (see [16], [17], [4], [44], [21], [42],[48], [20], etc. as well as survey articles [51], [49] [36], [37]). Hence our main focus is when Γ is of infinite co-volume in G.

1.1. Effective asymptotic of Matrix coefficients for $L^2(\Gamma\backslash G)$.

We begin by describing an effective asymptotic result on the matrix coefficients for $L^2(\Gamma\backslash G)$, which is a key ingredient in our approach as well as of independent interest. When Γ is not a lattice, a well-known theorem of Howe and Moore [27] implies that for any $\Psi_1, \Psi_2 \in L^2(\Gamma\backslash G)$, the matrix coefficient

$$\langle a\Psi_1, \Psi_2\rangle := \int_{\Gamma\backslash G} \Psi_1(ga)\overline{\Psi_2(g)}dg$$

decays to zero as $a \in G$ tends to infinity (here dg is a G-invariant measure on $\Gamma\backslash G$). Describing the precise asymptotic is much more involved. Fix a Cartan decomposition $G = KAK$ where K is a maximal compact subgroup and A is a one-parameter subgroup of diagonalizable elements. Let M denote the centralizer of A in K. The quotient spaces G/K and G/M can be respectively identified with $\mathbb{H}^n$ and its unit tangent bundle $\mathrm{T}^1(\mathbb{H}^n)$, and

[1]The *convex core* $C_\Gamma \subset \Gamma\backslash\mathbb{H}^n$ of Γ is the image of the minimal convex subset of $\mathbb{H}^n$ which contains all geodesics connecting any two points in the limit set of Γ.

we parameterize elements of $A = \{a_t : t \in \mathbb{R}\}$ so that the right translation action of a_t in G/M corresponds to the geodesic flow on $\mathrm{T}^1(\mathbb{H}^n)$ for time t.

We let $\{m_x : x \in \mathbb{H}^n\}$ and $\{\nu_x : x \in \mathbb{H}^n\}$ be Γ-invariant conformal densities of dimensions $(n-1)$ and δ respectively, unique up to scalings. Each ν_x is a finite measure on the limit set of Γ, called the Patterson-Sullivan measure viewed from x. Let $m^{\mathrm{BMS}}, m^{\mathrm{BR}}, m^{\mathrm{BR}_*}$ and m^{Haar} denote, respectively, the Bowen-Margulis-Sullivan measure, the Burger-Roblin measures for the expanding and the contracting horospherical foliations, and the Liouville-measure on the unit tangent bundle $\mathrm{T}^1(\Gamma\backslash\mathbb{H}^n)$, all defined with respect to the fixed pair of $\{m_x\}$ and $\{\nu_x\}$ (see Def. 2.1). Using the identification $\mathrm{T}^1(\Gamma\backslash\mathbb{H}^n) = \Gamma\backslash G/M$, we may extend these measures to right M-invariant measures on $\Gamma\backslash G$, which we will denote by the same notation and call them the BMS, the BR, the BR_*, the Haar measures for simplicity. We note that for $\delta < n-1$, only the BMS measure has finite mass [52].

In order to formulate a notion of a spectral gap for $L^2(\Gamma\backslash G)$, denote by $\hat{G}$ and $\hat{M}$ the unitary dual of G and M respectively. A representation $(\pi, \mathcal{H}) \in \hat{G}$ is called *tempered* if for any K-finite $v \in \mathcal{H}$, the associated matrix coefficient function $g \mapsto \langle \pi(g)v, v\rangle$ belongs to $L^{2+\epsilon}(G)$ for any $\epsilon > 0$; *non-tempered* otherwise. The non-tempered part of $\hat{G}$ consists of the trivial representation, and complementary series representations $\mathcal{U}(\upsilon, s-n+1)$ parameterized by $\upsilon \in \hat{M}$ and $s \in I_\upsilon$, where $I_\upsilon \subset (\frac{n-1}{2}, n-1)$ is an interval depending on υ. This was obtained by Hirai [26] (see also [30, Prop. 49, 50]). Moreover $\mathcal{U}(\upsilon, s-n+1)$ is spherical (i.e., has a non-zero K-invariant vector) if and only if υ is the trivial representation 1; see discussion in section 3.2.

By the works of Lax-Phillips [40], Patterson [54] and Sullivan [62], if $\delta > \frac{n-1}{2}$, $\mathcal{U}(1, \delta - n + 1)$ occurs as a subrepresentation of $L^2(\Gamma\backslash G)$ with multiplicity one, and $L^2(\Gamma\backslash G)$ possesses *spherical* spectral gap, meaning that there exists $\frac{n-1}{2} < s_0 < \delta$ such that $L^2(\Gamma\backslash G)$ does not weakly contain[2] any *spherical* complementary series representation $\mathcal{U}(1, s-n+1)$, $s \in (s_0, \delta)$. The following notion of a spectral gap concerns both the spherical and non-spherical parts of $L^2(\Gamma\backslash G)$.

Definition 1.1. We say that $L^2(\Gamma\backslash G)$ has a *strong spectral gap* if

(1) $L^2(\Gamma\backslash G)$ does not contain any $\mathcal{U}(\upsilon, \delta - n + 1)$ with $\upsilon \neq 1$;
(2) there exist $\frac{n-1}{2} < s_0(\Gamma) < \delta$ such that $L^2(\Gamma\backslash G)$ does not weakly contain any $\mathcal{U}(\upsilon, s - n + 1)$ with $s \in (s_0(\Gamma), \delta)$ and $\upsilon \in \hat{M}$.

For $\delta \leq \frac{n-1}{2}$, the Laplacian spectrum of $L^2(\Gamma\backslash\mathbb{H}^n)$ is continuous [40]; this implies that there is no spectral gap for $L^2(\Gamma\backslash G)$.

[2]for two unitary representations π and π' of G, π is said to be weakly contained in π' (or π' weakly contains π) if every diagonal matrix coefficient of π can be approximated, uniformly on compact subsets, by convex combinations of diagonal matrix coefficients of π'.

Conjecture 1.2 (Spectral gap conjecture)**.** *If Γ is a geometrically finite and Zariski dense subgroup of G with $\delta > \frac{n-1}{2}$, $L^2(\Gamma\backslash G)$ has a strong spectral gap.*

If $\delta > (n-1)/2$ for $n = 2, 3$, or if $\delta > (n-2)$ for $n \geq 4$, then $L^2(\Gamma\backslash G)$ has a strong spectral gap (Theorem 3.27).

Our main theorems are proved under the following slightly weaker spectral gap property assumption:

Definition 1.3. We say that $L^2(\Gamma\backslash G)$ has a *spectral gap* if there exist $\frac{n-1}{2} < s_0 = s_0(\Gamma) < \delta$ and $n_0 = n_0(\Gamma) \in \mathbb{N}$ such that

(1) the multiplicity of $\mathcal{U}(\upsilon, \delta - n + 1)$ contained in $L^2(\Gamma\backslash G)$ is at most $\dim(\upsilon)^{n_0}$ for any $\upsilon \in \hat{M}$;
(2) $L^2(\Gamma\backslash G)$ does not weakly contain any $\mathcal{U}(\upsilon, s-n+1)$ with $s \in (s_0, \delta)$ and $\upsilon \in \hat{M}$.

The pair $(s_0(\Gamma), n_0(\Gamma))$ will be referred to as the spectral gap data for Γ.

In the rest of the introduction, we impose the following hypothesis on Γ:

$$L^2(\Gamma\backslash G) \text{ has a spectral gap.}$$

Theorem 1.4. *There exist $\eta_0 > 0$ and $\ell \in \mathbb{N}$ (depending only on the spectral gap data for Γ) such that for any real-valued $\Psi_1, \Psi_2 \in C_c^\infty(\Gamma\backslash G)$, as $t \to \infty$,*

$$e^{(n-1-\delta)t} \int_{\Gamma\backslash G} \Psi_1(ga_t)\Psi_2(g) dm^{\mathrm{Haar}}(g) = \frac{m^{\mathrm{BR}}(\Psi_1) \cdot m^{\mathrm{BR}_*}(\Psi_2)}{|m^{\mathrm{BMS}}|} + O(\mathcal{S}_\ell(\Psi_1)\mathcal{S}_\ell(\Psi_2)e^{-\eta_0 t})$$

where $\mathcal{S}_\ell(\Psi_i)$ denotes the ℓ-th L^2-Sobolev norm of Ψ_i for each $i = 1, 2$.

Remark 1.5. We remark that if either Ψ_1 or Ψ_2 is K-invariant, then Theorem 1.4 holds for any Zariski dense Γ with $\delta > \frac{n-1}{2}$ (without the spectral gap hypothesis), as the spherical spectral gap of $L^2(\Gamma\backslash G)$ is sufficient to study the matrix coefficients associated to spherical vectors.

Let $\mathcal{H}_\delta^\dagger$ denote the sum of of all complementary series representations of parameter δ contained in $L^2(\Gamma\backslash G)$, and let P_δ denote the projection operator from $L^2(\Gamma\backslash G)$ to $\mathcal{H}_\delta^\dagger$. By the spectral gap hypothesis on $L^2(\Gamma\backslash G)$, the main work in the proof of Theorem 1.4 is to understand the asymptotic of $\langle a_t P_\delta(\Psi_1), P_\delta(\Psi_2)\rangle$ as $t \to \infty$. Building up on the work of Harish-Chandra on the asymptotic behavior of the Eisenstein integrals (cf. [65], [66]), we first obtain an asymptotic formula for $\langle a_t v, w\rangle$ for all K-finite vectors $v, w \in \mathcal{H}_\delta^\dagger$ (Theorem 3.23). This extension alone does not give the formula of the leading term of $\langle a_t P_\delta(\Psi_1), P_\delta(\Psi_2)\rangle$ in terms of functions Ψ_1 and Ψ_2; however, an ergodic theorem of Roblin [56] and Winter [67] enables us to identify the main term as given in Theorem 1.4.

1.2. **Exponential mixing of frame flows.** Via the identification of the space $\Gamma\backslash G$ with the frame bundle over the hyperbolic manifold $\Gamma\backslash\mathbb{H}^n$, the right translation action of a_t on $\Gamma\backslash G$ corresponds to the frame flow for time t. The BMS measure m^{BMS} on $\Gamma\backslash G$ is known to be mixing for the frame flows ([18], [67]). We deduce the following exponential mixing from Theorem 1.4: for a compact subset Ω of $\Gamma\backslash G$, we denote by $C^\infty(\Omega)$ the set of all smooth functions on $\Gamma\backslash G$ with support contained in Ω.

Theorem 1.6. *There exist $\eta_0 > 0$ and $\ell \in \mathbb{N}$ such that for any compact subset $\Omega \subset \Gamma\backslash G$, and for any $\Psi_1, \Psi_2 \in C^\infty(\Omega)$, as $t \to \infty$,*

$$\int_{\Gamma\backslash G} \Psi_1(ga_t)\Psi_2(g)dm^{\mathrm{BMS}}(g) = \frac{m^{\mathrm{BMS}}(\Psi_1)\cdot m^{\mathrm{BMS}}(\Psi_2)}{|m^{\mathrm{BMS}}|} + O(\mathcal{S}_\ell(\Psi_1)\mathcal{S}_\ell(\Psi_2)e^{-\eta_0 t})$$

where the implied constant depends only on Ω.

For Γ convex co-compact, Theorem 1.6 for Ψ_1 and Ψ_2 M-invariant functions holds for any $\delta > 0$ by Stoyanov [61], based on the approach developed by Dolgopyat [14]; however when Γ has cusps, this theorem seems to be new even for $n = 2$.

1.3. **Effective equidistribution of orthogonal translates of an H-orbit.** When H is a horospherical subgroup or a symmetric subgroup of G, we can relate the asymptotic distribution of orthogonal translates of a closed orbit $\Gamma\backslash\Gamma H$ to the matrix coefficients of $L^2(\Gamma\backslash G)$. We fix a generalized Cartan decomposition $G = HAK$. We parameterize $A = \{a_t\}$ as in section 1.1, and for H horospherical, we will assume that H is the expanding horospherical subgroup for a_t, that is, $H = \{g \in G : a_t g a_{-t} \to e \text{ as } t \to \infty\}$. Let μ_H^{Haar} and μ_H^{PS} be respectively the H-invariant measure on $\Gamma\backslash\Gamma H$ defined with respect to $\{m_x\}$ and the skinning measure on $\Gamma\backslash\Gamma H$ defined with respect to $\{\nu_x\}$, introduced in [52] (cf. (4.2)).

Theorem 1.7. *Suppose that $\Gamma\backslash\Gamma H$ is closed and that $|\mu_H^{\mathrm{PS}}| < \infty$. There exist $\eta_0 > 0$ and $\ell \in \mathbb{N}$ such that for any compact subset $\Omega \subset \Gamma\backslash G$, any $\Psi \in C^\infty(\Omega)$ and any bounded $\phi \in C^\infty(\Gamma \cap H\backslash H)$, as $t \to \infty$,*

$$e^{(n-1-\delta)t}\int_{h\in\Gamma\backslash\Gamma H} \Psi(ha_t)\phi(h)d\mu_H^{\mathrm{Haar}}(h) = \frac{1}{|m^{\mathrm{BMS}}|}\mu_H^{\mathrm{PS}}(\phi)m^{\mathrm{BR}}(\Psi) + O(\mathcal{S}_\ell(\Psi)\cdot S_\ell(\phi)e^{-\eta_0 t})$$

with the implied constant depending only on Ω.

For H horospherical, $|\mu_H^{\mathrm{PS}}| < \infty$ is automatic for $\Gamma\backslash\Gamma H$ closed. For H symmetric (and hence locally isomorphic to $\mathrm{SO}(k,1)\times\mathrm{SO}(n-k)$), the criterion for the finiteness of μ_H^{PS} has been obtained in [52] (see Prop. 4.15); in particular, $|\mu_H^{\mathrm{PS}}| < \infty$ provided $\delta > n - k$.

Letting $Y_\Omega := \{h \in (\Gamma \cap H)\backslash H : ha_t \in \Omega \text{ for some } t > 0\}$, note that

$$\int \Psi(ha_t)\phi(h)d\mu_H^{\mathrm{Haar}} = \int_{Y_\Omega} \Psi(ha_t)\phi(h)d\mu_H^{\mathrm{Haar}}$$

since Ψ is supported in Ω. In the case when μ_H^{PS} is compactly supported, Y_Ω turns out to be a compact subset and in this case, the so-called thickening method ([17], [29]) is sufficient to deduce Theorem 1.7 from Theorem 1.4, using the wave front property introduced in [17] (see [4] for the effective version). The case of μ_H^{PS} not compactly supported is much more intricate to be dealt with. Though we obtain a thick-thin decomposition of Y_Ω with the thick part being compact and control both the Haar measure and the skinning measure of the thin part (Theorem 4.16), the usual method of thickening the thick part does not suffice, as the error term coming from the thin part overtakes the leading term. The main reason for this phenomenon is because we are taking the integral with respect to μ_H^{Haar} as well as multiplying the weight factor $e^{(n-1-\delta)t}$ in the left hand side of Theorem 1.7, whereas the finiteness assumption is made on the skinning measure μ_H^{PS}. However we are able to proceed by comparing the two measures $(a_t)_*\mu_H^{\mathrm{PS}}$ and $(a_t)_*\mu_H^{\mathrm{Haar}}$ via the the transversal intersections of the orbits $\Gamma\backslash\Gamma Ha_t$ with the weak-stable horospherical foliations (see the proof of Theorem 6.9 for more details).

In the special case of $n = 2, 3$ and H horospherical, Theorem 1.7 was proved in [35], [34] and [41] by a different method.

1.4. **Effective counting for a discrete Γ-orbit in $H\backslash G$.** In this subsection, we let H be the trivial group, a horospherical subgroup or a symmetric subgroup, and assume that the orbit $[e]\Gamma$ is discrete in $H\backslash G$. Theorems 1.4 and 1.7 are key ingredients in understanding the asymptotic of the number $\#([e]\Gamma \cap \mathcal{B}_T)$ for a given family $\{\mathcal{B}_T \subset H\backslash G\}$ of growing compact subsets, as observed in [16].

We will first describe a Borel measure $\mathcal{M}_{H\backslash G} = \mathcal{M}^{\Gamma}_{\mathrm{H}\backslash G}$ on $H\backslash G$, depending on Γ, which turns out to describe the distribution of $[e]\Gamma$. Let $o \in \mathbb{H}^n$ be the point fixed by K, $X_0 \in \mathrm{T}^1(\mathbb{H}^n)$ the vector fixed by M, $X_0^+, X_0^- \in \partial(\mathbb{H}^n)$ the forward and the backward endpoints of X_0 by the geodesic flow, respectively and ν_o the Patterson-Sullivan measure on $\partial(\mathbb{H}^n)$ supported on the limit set of Γ, viewed from o. Let dm denote the probability Haar measure of M.

Definition 1.8. For H the trivial subgroup $\{e\}$, define a Borel measure $\mathcal{M}_G = \mathcal{M}^{\Gamma}_G$ on G as follows: for $\psi \in C_c(G)$,

$$\mathcal{M}_G(\psi) := \\ \frac{1}{|m^{\mathrm{BMS}}|}\int_{(K/M)\times A^+\times M\times(M\backslash K)} \psi(k_1a_tmk_2)e^{\delta t}d\nu_o(k_1X_0^+)dtdmd\nu_o(k_2^{-1}X_0^-).$$

Definition 1.9. For H horospherical or symmetric, we have either $G = HA^+K$ or $G = HA^+K \cup HA^-K$ (as a disjoint union except for the identity element) where $A^\pm = \{a_{\pm t} : t \geq 0\}$.

Define a Borel measure $\mathcal{M}_{H\backslash G} = \mathcal{M}^\Gamma_{H\backslash G}$ on $H\backslash G$ as follows: for $\psi \in C_c(H\backslash G)$,

$$\mathcal{M}_{H\backslash G}(\psi) :=$$
$$\begin{cases} \frac{|\mu_H^{\mathrm{PS}}|}{|m^{\mathrm{BMS}}|} \int_{A^+\times M\times(M\backslash K)} \psi([e]a_t mk) e^{\delta t} dt dm d\nu_o(k^{-1}X_0^-) & \text{if } G = HA^+K \\ \sum \frac{|\mu_{H,\pm}^{\mathrm{PS}}|}{|m^{\mathrm{BMS}}|} \int_{A^\pm\times M\times(M\backslash K)} \psi([e]a_{\pm t} mk) e^{\delta t} dt dm d\nu_o(k^{-1}X_0^\mp) & \text{otherwise,} \end{cases}$$

where $\mu^{\mathrm{PS}}_{H,-}$ is the skinning measure on $\Gamma \cap H\backslash H$ in the negative direction, as defined in (6.15).

Definition 1.10. For a family $\{\mathcal{B}_T \subset H\backslash G\}$ of compact subsets with $\mathcal{M}_{H\backslash G}(\mathcal{B}_T)$ tending to infinity as $T \to \infty$, we say that $\{\mathcal{B}_T\}$ is *effectively well-rounded with respect to* Γ if there exists $p > 0$ such that for all small $\epsilon > 0$ and large $T \gg 1$:

$$\mathcal{M}_{H\backslash G}(\mathcal{B}^+_{T,\epsilon} - \mathcal{B}^-_{T,\epsilon}) = O(\epsilon^p \cdot \mathcal{M}_{H\backslash G}(\mathcal{B}_T))$$

where $\mathcal{B}^+_{T,\epsilon} = G_\epsilon \mathcal{B}_T G_\epsilon$ and $\mathcal{B}^-_{T,\epsilon} = \cap_{g_1,g_2\in G_\epsilon} g_1 \mathcal{B}_T g_2$ if $H = \{e\}$; and $\mathcal{B}^+_{T,\epsilon} = \mathcal{B}_T G_\epsilon$ and $\mathcal{B}^-_{T,\epsilon} = \cap_{g\in G_\epsilon} \mathcal{B}_T g$ if H is horospherical or symmetric. Here G_ϵ denotes a symmetric ϵ-neighborhood of e in G with respect to a left invariant Riemannian metric on G.

Since any two left-invariant Riemannian metrics on G are Lipschitz equivalent to each other, the above definition is independent of the choice of a Riemannian metric used in the definition of G_ϵ.

See Propositions 7.11, 7.15 and 7.17 for examples of effectively well-rounded families. For instance, if G acts linearly from the right on a finite dimensional linear space V and H is the stabilizer of $w_0 \in V$, then the family of norm balls $B_T := \{Hg \in H\backslash G : \|w_0 g\| < T\}$ is effectively well-rounded.

If Γ is a lattice in G, then $\mathcal{M}_{H\backslash G}$ is essentially the leading term of the invariant measure in $H\backslash G$ and hence the definition 1.10 is equivalent to the one given in [4], which is an effective version of the well-roundedness condition given in [17]. Under the additional assumption that $H \cap \Gamma$ is a lattice in H, it is known that if $\{\mathcal{B}_T\}$ is effectively well-rounded, then

$$\#([e]\Gamma \cap \mathcal{B}_T) = \mathrm{Vol}(\mathcal{B}_T) + O(\mathrm{Vol}(\mathcal{B}_T)^{1-\eta_0}) \tag{1.11}$$

for some $\eta_0 > 0$, where Vol is computed with respect to a suitably normalized invariant measure on $H\backslash G$ (cf. [16], [17], [44], [20], [4]).

We present a generalization of (1.11). In the next two theorems 1.12 and 1.14, we let $\{\Gamma_d : d \in I\}$ be a family of subgroups of Γ of finite index such that $\Gamma_d \cap H = \Gamma \cap H$. We assume that $\{\Gamma_d : d \in I\}$ has a uniform spectral gap in the sense that $\sup_d s_0(\Gamma_d) < \delta$ and $\sup_d n_0(\Gamma_d) < \infty$.

For our intended application to the affine sieve, we formulate our effective results uniformly for all Γ_d's.

Theorem 1.12. *Let H be the trivial group, a horospherical subgroup or a symmetric subgroup. When H is symmetric, we also assume that $|\mu_H^{\mathrm{PS}}| < \infty$. If $\{\mathcal{B}_T\}$ is effectively well-rounded with respect to Γ, then there exists $\eta_0 > 0$ such that for any $d \in I$ and for any $\gamma_0 \in \Gamma$,*

$$\#([e]\Gamma_d\gamma_0 \cap \mathcal{B}_T) = \tfrac{1}{[\Gamma:\Gamma_d]}\mathcal{M}_{H\backslash G}(\mathcal{B}_T) + O(\mathcal{M}_{H\backslash G}(\mathcal{B}_T)^{1-\eta_0})$$

where $\mathcal{M}_{H\backslash G} = \mathcal{M}^{\Gamma}_{H\backslash G}$ and the implied constant is independent of Γ_d and $\gamma_0 \in \Gamma$.

See Corollaries 7.16 and 7.18 where we have applied Theorem 1.12 to sectors and norm balls.

Remark 1.13. Theorem 1.12 can be used to provide an effective version of circle-counting theorems studied in [35], [41], [53] and [50] (as well as its higher dimensional analogues for sphere packings discussed in [51]).

We also formulate our counting statements for bisectors in the HAK decomposition, motivated by recent applications in [8] and [9]. Let $\tau_1 \in C_c^\infty(H)$ and $\tau_2 \in C^\infty(K)$, and define $\xi_T^{\tau_1,\tau_2} \in C^\infty(G)$ as follows: for $g = hak \in HA^+K$,

$$\xi_T^{\tau_1,\tau_2}(g) = \chi_{A_T^+}(a) \cdot \int_{H\cap M} \tau_1(hm)\tau_2(m^{-1}k)d_{H\cap M}(m)$$

where $\chi_{A_T^+}$ denotes the characteristic function of $A_T^+ = \{a_t : 0 < t < \log T\}$ and $d_{H\cap M}$ is the probability Haar measure of $H \cap M$. Since $hak = h'ak'$ implies that $h = h'm$ and $k = m^{-1}k'$ for some $m \in H \cap M$, $\xi_T^{\tau_1,\tau_2}$ is well-defined.

Theorem 1.14. *There exist $\eta_0 > 0$ and $\ell \in \mathbb{N}$ such that for any compact subset H_0 of H which injects to $\Gamma\backslash G$, any $\tau_1 \in C^\infty(H_0)$, $\tau_2 \in C^\infty(K)$, any $\gamma_0 \in \Gamma$ and any $d \in I$,*

$$\sum_{\gamma\in\Gamma_d\gamma_0} \xi_T^{\tau_1,\tau_2}(\gamma) = \frac{\tilde{\mu}_H^{\mathrm{PS}}(\tau_1)\cdot\nu_o^*(\tau_2)}{\delta\cdot[\Gamma:\Gamma_d]\cdot|m^{\mathrm{BMS}}|}T^\delta + O(\mathcal{S}_\ell(\tau_1)\mathcal{S}_\ell(\tau_2)T^{\delta-\eta_0})$$

where $\nu_o^*(\tau_2) = \int_K\int_M \tau_2(mk)dmd\nu_o(k^{-1}X_0^-)$ and $\tilde{\mu}_H^{\mathrm{PS}}$ is the skinning measure on H with respect to Γ and the implied constant depends only on Γ and H_0.

Theorem 1.14 also holds when τ_1 and τ_2 are characteristic functions of the so-called *admissible* subsets (see Corollary 7.21 and Proposition 7.11).

We remark that unless $H = K$, Corollary 7.16, which is a special case of Theorem 1.12 for sectors in $H\backslash G$, does not follow from Theorem 1.14, as the latter deals only with compactly supported functions τ_1. For $H = K$, Theorem 1.14 was earlier shown in [7] and [64] for $n = 2, 3$ respectively.

Remark 1.15. Non-effective versions of Theorems 1.7, 1.12, and 1.14 were obtained in [52] for a more general class of discrete groups, that is, any non-elementary discrete subgroup admitting finite BMS-measure.

1.5. **Application to Affine sieve.** One of the main applications of Theorem 1.12 can be found in connection with Diophantine problems on orbits of Γ. Let $\mathbf{G}$ be a $\mathbb{Q}$-form of G, that is, $\mathbf{G}$ is a connected algebraic group defined over $\mathbb{Q}$ such that $G = \mathbf{G}(\mathbb{R})^\circ$. Let $\mathbf{G}$ act on an affine space V via a $\mathbb{Q}$-rational representation in a way that $\mathbf{G}(\mathbb{Z})$ preserves $V(\mathbb{Z})$. Fix a non-zero vector $w_0 \in V(\mathbb{Z})$ and denote by $\mathbf{H}$ its stabilizer subgroup and set $H = \mathbf{H}(\mathbb{R})$. We consider one of the following situations: (1) H is a symmetric subgroup of G or the trivial subgroup; (2) $w_0\mathbf{G} \cup \{0\}$ is Zariski closed and H is a compact extension of a horospherical subgroup of G.

In the case (1), $w_0\mathbf{G}$ is automatically Zariski closed by [23]. Set $W := w_0\mathbf{G}$ and $w_0\mathbf{G} \cup \{0\}$ respectively for (1) and (2).

Let Γ be a geometrically finite and Zariski dense subgroup of G with $\delta > \frac{n-1}{2}$, which is contained in $\mathbf{G}(\mathbb{Z})$. If H is symmetric, we assume that $|\mu_H^{\mathrm{PS}}| < \infty$.

For a positive integer d, we denote by Γ_d the congruence subgroup of Γ which consists of $\gamma \in \Gamma$ such that $\gamma \equiv e \bmod d$. For the next two theorems 1.16 and 1.17 we assume that there exists a finite set S of primes that the family $\{\Gamma_d : d \text{ is square-free with no prime factors in } S\}$ has a uniform spectral gap. This property always holds if $\delta > (n-1)/2$ for $n = 2, 3$ and if $\delta > n - 2$ for $n \geq 4$ via the recent works of Bourgain,Gamburd, Sarnak ([6], [5]) and Salehi-Golsefidy and Varju [58] together with the classification of the unitary dual of G (see Theorem 8.2).

Let $F \in \mathbb{Q}[W]$ be an integer-valued polynomial on the orbit $w_0\Gamma$. Salehi-Golsefidy and Sarnak [57], generalizing [6], showed that for some $R > 1$, the set of $\mathbf{x} \in w_0\Gamma$ with $F(\mathbf{x})$ having at most R prime factors is Zariski dense in $w_0\mathbf{G}$. The following are quantitative versions: Letting $F = F_1F_2\cdots F_r$ be a factorization into irreducible polynomials in $\mathbb{Q}[W]$, assume that all F_j's are irreducible in $\mathbb{C}[W]$ and integral on $w_0\Gamma$. Let $\{\mathcal{B}_T \subset w_0G : T \gg 1\}$ be an effectively well-rounded family of subsets with respect to Γ.

Theorem 1.16 (Upper bound for primes)**.** *For all $T \gg 1$,*

$$\{\mathbf{x} \in w_0\Gamma \cap \mathcal{B}_T : F_j(\mathbf{x}) \text{ is prime for } j = 1, \cdots, r\} \ll \frac{\mathcal{M}_{w_0G}(\mathcal{B}_T)}{(\log \mathcal{M}_{w_0G}(\mathcal{B}_T))^r}.$$

Theorem 1.17 (Lower bound for almost primes)**.** *Assume further that $\max_{x\in\mathcal{B}_T} \|x\| \ll \mathcal{M}_{w_0G}(\mathcal{B}_T)^\beta$ for some $\beta > 0$, where $\|\cdot\|$ is any norm on V. Then there exists $R = R(F, w_0\Gamma, \beta) \geq 1$ such that for all $T \gg 1$,*

$$\{\mathbf{x} \in w_0\Gamma \cap \mathcal{B}_T : F(\mathbf{x}) \text{ has at most } R \text{ prime factors}\} \gg \frac{\mathcal{M}_{w_0G}(\mathcal{B}_T)}{(\log \mathcal{M}_{w_0G}(\mathcal{B}_T))^r}.$$

Observe that these theorems provide a description of the asymptotic distribution of almost prime vectors, as $\mathcal{B}_T$ can be taken arbitrarily.

Remark 1.18. In both theorems above, if all $\mathcal{B}_T$ are K-invariant subsets, our hypothesis on the uniform spectral gap for the family $\{\Gamma_d\}$ can be disposed again, as the uniform *spherical* spectral gap property proved in [58] and [6] is sufficient in this case.

For instance, Theorems 1.16 and 1.17 can be applied to the norm balls $\mathcal{B}_T = \{\mathbf{x} \in w_0G : \|\mathbf{x}\| < T\}$ and in this case $\mathcal{M}_{w_0G}(\mathcal{B}_T) \asymp T^{\delta/\lambda}$ where λ denotes the log of the largest eigenvalue of a_1 on the $\mathbb{R}$-span of w_0G.

In order to present a concrete example, we consider an integral quadratic form $Q(x_1, \cdots, x_{n+1})$ of signature $(n, 1)$ and for an integer $s \in \mathbb{Z}$, denote by $W_{Q,s}$ the affine quadric given by

$$\{\mathbf{x} : Q(\mathbf{x}) = s\}.$$

As well-known, $W_{Q,s}$ is a one-sheeted hyperboloid if $s > 0$, a two-sheeted hyperboloid if $s < 0$ and a cone if $s = 0$. We will assume that $Q(\mathbf{x}) = s$ has a non-zero integral solution, so pick $w_0 \in W_{Q,s}(\mathbb{Z})$. If $s \neq 0$, the stabilizer subgroup G_{w_0} is symmetric; more precisely, locally isomorphic to $\mathrm{SO}(n-1, 1)$ (if $s > 0$) or $\mathrm{SO}(n)$ (if $s < 0$) and if $s = 0$, G_{w_0} is a compact extension of a horospherical subgroup. By the remark following Theorem 1.7, the skinning measure $\mu^{\mathrm{PS}}_{G_{w_0}}$ is finite if $n \geq 3$. For $n = 2$ and $s > 0$, G_{w_0} is a one-dimensional subgroup consisting of diagonalizable elements, and $\mu^{\mathrm{PS}}_{G_{w_0}}$ is infinite only when the geodesic in $\mathbb{H}^2$ stabilized by G_{w_0} is divergent and goes into a cusp of a fundamental domain of Γ in $\mathbb{H}^2$; in this case, we call w_0 externally Γ-parabolic, following [52]. Therefore the following are special cases of Theorems 1.16 and 1.17:

Corollary 1.19. *Let Γ be a geometrically finite and Zariski dense subgroup of $\mathrm{SO}_Q(\mathbb{Z})$ with $\delta > \frac{n-1}{2}$. In the case when $n = 2$ and $s > 0$, we additionally assume that w_0 is not externally Γ-parabolic. Fixing a K-invariant norm $\|\cdot\|$ on $\mathbb{R}^{n+1}$, we have, for any $1 \leq r \leq n+1$,*

(1) $\{\mathbf{x} \in w_0\Gamma : \|\mathbf{x}\| < T,\ x_j \text{ is prime for all } j = 1, \cdots, r\} \ll \frac{T^\delta}{(\log T)^r}$;
(2) *for some $R > 1$,*

$$\{\mathbf{x} \in w_0\Gamma : \|\mathbf{x}\| < T,\ x_1 \cdots x_r \text{ has at most } R \text{ prime factors}\} \gg \frac{T^\delta}{(\log T)^r}.$$

The upper bound in (1) is sharp up to a multiplicative constant. The lower bound in (2) can also be stated for admissible sectors under the uniform spectral gap hypothesis (cf. Corollary 7.18). Corollary 1.19 was previously obtained in cases when $n = 2, 3$ and $s \leq 0$ ([5], [33], [35],[34], [41]).

To explain how Theorems 1.16 and 1.17 follow from Theorem 1.12, let $\Gamma_{w_0}(d) = \{\gamma \in \Gamma : w_0\gamma = w_0 \bmod (\mathrm{d})\}$ for each square-free integer d. Then $\mathrm{Stab}_{\Gamma_{w_0}(d)}(w_0) = \mathrm{Stab}_\Gamma(w_0)$ and the family $\{\Gamma_{w_0}(d)\}$ admits a uniform spectral gap property as $\Gamma_d < \Gamma_{w_0}(d)$. Hence Theorem 1.12 holds

for the congruence family $\{\Gamma_{w_0}(d) : d \text{ is square-free, with no small primes}\}$, providing a key axiomatic condition in executing the combinatorial sieve (see [28, 6.1-6.4], [25, Theorem 7.4], as well as [6, Sec. 3]). When an explicit uniform spectral gap for $\{\Gamma_d\}$ is known (e.g., [19], [43]), the number $R(F, w_0\Gamma)$ can also be explicitly computed in principle.

The paper is organized as follows. In section 2, we recall the ergodic result of Roblin which gives the leading term of the matrix coefficients for $L^2(\Gamma\backslash G)$. In section 3, we obtain an effective asymptotic expansion for the matrix coefficients of the complementary series representations of G (Theorem 3.23) as well as for those of $L^2(\Gamma\backslash G)$, proving Theorem 1.4. In section 4, we study the reduction theory for the non-wandering component of $\Gamma\backslash\Gamma Ha_t$, describing its thick-thin decomposition; this is needed as $\Gamma\backslash\Gamma H$ has infinite Haar-volume in general. We will see that the non-trivial dynamics of $\Gamma\backslash\Gamma Ha_t$ as $t \to \infty$ can be seen only within a subset of finite PS-measure. In section 5, for ϕ compactly supported, we prove Theorem 1.7 using Theorem 1.4 via thickening. For a general bounded ϕ, Theorem 1.7 is obtained via a careful study of the transversal intersections in section 6. Theorem 1.6 is also proved in section 6. Counting theorems 1.12 and 1.14 are proved in section 7 and Sieve theorems 1.16 and 1.17 are proved in the final section 8.

2. Matrix coefficients in $L^2(\Gamma\backslash G)$ by ergodic methods

Throughout the paper, let G be $\mathrm{SO}(n,1)^\circ = \mathrm{Isom}^+(\mathbb{H}^n)$ for $n \geq 2$, i.e., the group of orientation preserving isometries of $(\mathbb{H}^n, d)$, and $\Gamma < G$ be a non-elementary torsion-free geometrically finite group. Let $\partial(\mathbb{H}^n)$ denote the geometric boundary of $\mathbb{H}^n$. Let $\Lambda(\Gamma) \subset \partial(\mathbb{H}^n)$ denote the limit set of Γ, and δ the critical exponent of Γ, which is known to be equal to the Hausdorff dimension of $\Lambda(\Gamma)$ [63].

A family of measures $\{\mu_x : x \in \mathbb{H}^n\}$ is called a Γ-*invariant conformal density* of dimension $\delta_\mu > 0$ on $\partial(\mathbb{H}^n)$, if each μ_x is a non-zero finite Borel measure on $\partial(\mathbb{H}^n)$ satisfying for any $x, y \in \mathbb{H}^n$, $\xi \in \partial(\mathbb{H}^n)$ and $\gamma \in \Gamma$,

$$\gamma_*\mu_x = \mu_{\gamma x} \quad \text{and} \quad \frac{d\mu_y}{d\mu_x}(\xi) = e^{-\delta_\mu \beta_\xi(y,x)},$$

where $\gamma_*\mu_x(F) = \mu_x(\gamma^{-1}(F))$ for any Borel subset F of $\partial(\mathbb{H}^n)$. Here $\beta_\xi(y,x)$ denotes the Busemann function: $\beta_\xi(y,x) = \lim_{t\to\infty} d(\xi_t, y) - d(\xi_t, x)$ where ξ_t is a geodesic ray tending to ξ as $t \to \infty$.

We denote by $\{\nu_x\}$ the Patterson-Sullivan density, i.e., a Γ-*invariant conformal density* of dimension δ and by $\{m_x : x \in \mathbb{H}^n\}$ a *Lebesgue density*, i.e., a G-invariant conformal density on the boundary $\partial(\mathbb{H}^n)$ of dimension $(n-1)$. Both densities are determined unique up to scalar multiples.

Denote by $\{\mathcal{G}^t : t \in \mathbb{R}\}$ the geodesic flow on $\mathrm{T}^1(\mathbb{H}^n)$. For $u \in \mathrm{T}^1(\mathbb{H}^n)$, we denote by $u^\pm \in \partial(\mathbb{H}^n)$ the forward and the backward endpoints of the geodesic determined by u, i.e., $u^\pm = \lim_{t\to\pm\infty} \mathcal{G}^t(u)$. Fixing $o \in \mathbb{H}^n$, the

map

$$u \mapsto (u^+, u^-, s = \beta_{u^-}(o, \pi(u)))$$

is a homeomorphism between $\mathrm{T}^1(\mathbb{H}^n)$ with

$$(\partial(\mathbb{H}^n) \times \partial(\mathbb{H}^n) - \{(\xi, \xi) : \xi \in \partial(\mathbb{H}^n)\}) \times \mathbb{R}.$$

Using this homeomorphism, we define measures $\tilde{m}^{\mathrm{BMS}}, \tilde{m}^{\mathrm{BR}}, \tilde{m}^{\mathrm{BR}_*}, \tilde{m}^{\mathrm{Haar}}$ on $\mathrm{T}^1(\mathbb{H}^n)$ as follows ([11], [45], [63], [12], [56]):

Definition 2.1. *Set*

(1) $d\tilde{m}^{\mathrm{BMS}}(u) = e^{\delta\beta_{u^+}(o,\pi(u))}\, e^{\delta\beta_{u^-}(o,\pi(u))}\, d\nu_o(u^+) d\nu_o(u^-) ds$;
(2) $d\tilde{m}^{\mathrm{BR}}(u) = e^{(n-1)\beta_{u^+}(o,\pi(u))}\, e^{\delta\beta_{u^-}(o,\pi(u))}\, dm_o(u^+) d\nu_o(u^-) ds$;
(3) $d\tilde{m}^{\mathrm{BR}_*}(u) = e^{\delta\beta_{u^+}(o,\pi(u))}\, e^{(n-1)\beta_{u^-}(o,\pi(u))}\, d\nu_o(u^+) dm_o(u^-) ds$;
(4) $d\tilde{m}^{\mathrm{Haar}}(u) = e^{(n-1)\beta_{u^+}(o,\pi(u))}\, e^{(n-1)\beta_{u^-}(o,\pi(u))}\, dm_o(u^+) dm_o(u^-) ds$.

The conformal properties of $\{\nu_x\}$ and $\{m_x\}$ imply that these definitions are independent of the choice of $o \in \mathbb{H}^n$. We will extend these measures to G; these extensions depend on the choice of $o \in \mathbb{H}^n$ and $X_0 \in \mathrm{T}^1_o(\mathbb{H}^n)$. Let $K := \mathrm{Stab}_G(o)$ and $M := \mathrm{Stab}_G(X_0)$, so that $\mathbb{H}^n \simeq G/K$ and $\mathrm{T}^1(\mathbb{H}^n) \simeq G/M$. Let $A = \{a_t : t \in \mathbb{R}\}$ be the one-parameter subgroup of diagonalizable elements in the centralizer of M in G such that $\mathcal{G}^t(X_0) = [M]a_t = [a_t M]$.

Using the identification $\mathrm{T}^1(\mathbb{H}^n)$ with G/M, we lift the above measures to G, which will be denoted by the same notation by abuse of notation, so that they are all invariant under M from the right.

These measures are all left Γ-invariant, and hence induce locally finite Borel measures on $\Gamma\backslash G$, which we denote by m^{BMS} (the BMS-measure), m^{BR} (the BR-measure), m^{BR_*} (the BR_* measure), m^{Haar} (the Haar measure) by abuse of notation.

Let N^+ and N^- denote the expanding and the contracting horospherical subgroups, i.e.,

$$N^\pm = \{g \in G : a_t g a_{-t} \to e \text{ as } t \to \pm\infty\}.$$

For $g \in G$, define

$$g^\pm := (gM)^\pm \in \partial(\mathbb{H}^n).$$

We note that m^{BMS}, m^{BR}, and m^{BR_*} are invariant under A, N^+ and N^- respectively and their supports are given respectively by $\{g \in \Gamma\backslash G : g^+, g^- \in \Lambda(\Gamma)\}$, $\{g \in \Gamma\backslash G : g^- \in \Lambda(\Gamma)\}$ and $\{g \in \Gamma\backslash G : g^+ \in \Lambda(\Gamma)\}$. The measure m^{Haar} is invariant under both N^+ and N^-, and hence under G, as N^+ and N^- generate G topologically. That is, m^{Haar} is a Haar measure of G.

We consider the action of G on $L^2(\Gamma\backslash G, m^{\mathrm{Haar}})$ by right translations, which gives rise to the unitary action for the inner product:

$$\langle \Psi_1, \Psi_2 \rangle = \int_{\Gamma\backslash G} \Psi_1(g)\overline{\Psi_2(g)} dm^{\mathrm{Haar}}(g).$$

Theorem 2.2. *Let Γ be Zariski dense. For any functions $\Psi_1, \Psi_2 \in C_c(\Gamma\backslash G)$,*

$$\lim_{t\to\infty} e^{(n-1-\delta)t}\langle a_t\Psi_1, \Psi_2\rangle = \frac{m^{\mathrm{BR}}(\Psi_1)\cdot m^{\mathrm{BR}_*}(\Psi_2)}{|m^{\mathrm{BMS}}|}.$$

Proof. Roblin [56] proved this for M-invariant functions Ψ_1 and Ψ_2. His proof is based on the mixing of the geodesic flow on $\mathrm{T}^1(\Gamma\backslash\mathbb{H}^n) = \Gamma\backslash G/M$. For Γ Zariski dense, the mixing of m^{BMS} was extended to the frame flow on $\Gamma\backslash G$, by [67]. Based on this, the proof given in [56] can be repeated verbatim to prove the claim (cf. [67]). □

3. Asymptotic expansion of Matrix coefficients

3.1. **Unitary dual of G.** Let $G = \mathrm{SO}(n,1)^\circ$ for $n \geq 2$ and K a maximal compact subgroup of G. Denoting by $\mathfrak{g}$ and $\mathfrak{k}$ the Lie algebras of G and K respectively, let $\mathfrak{g} = \mathfrak{k} \oplus \mathfrak{p}$ be the corresponding Cartan decomposition of $\mathfrak{g}$. Let $A = \exp(\mathfrak{a})$ where $\mathfrak{a}$ is a maximal abelian subspace of $\mathfrak{p}$ and let M be the centralizer of A in K.

Define the symmetric bi-linear form $\langle\cdot,\cdot\rangle$ on $\mathfrak{g}$ by

$$\langle X, Y\rangle := \frac{1}{2(n-1)}B(X,Y) \tag{3.1}$$

where $B(X,Y) = \mathrm{Tr}(\mathrm{ad}X\mathrm{ad}Y)$ denotes the Killing form for $\mathfrak{g}$. The reason for this normalization is so that the Riemannian metric on $G/K \simeq \mathbb{H}^n$ induced by $\langle\cdot,\cdot\rangle$ has constant curvature -1.

Let $\{X_i\}$ be a basis for $\mathfrak{g}_{\mathbb{C}}$ over $\mathbb{C}$; put $g_{ij} = \langle X_i, X_j\rangle$ and let g^{ij} be the (i,j) entry of the inverse matrix of (g_{ij}). The element

$$\mathcal{C} = \sum g^{ij}X_iX_j$$

is called the Casimir element of $\mathfrak{g}_{\mathbb{C}}$ (with respect to $\langle\cdot,\cdot\rangle$). It is well-known that this definition is independent of the choice of a basis and that $\mathcal{C}$ lies in the center of the universal enveloping algebra $U(\mathfrak{g}_{\mathbb{C}})$ of $\mathfrak{g}_{\mathbb{C}}$.

Denote by $\hat{G}$ the unitary dual, i.e., the set of equivalence classes of irreducible unitary representations of G. A representation $\pi \in \hat{G}$ is said to be *tempered* if for any K-finite vectors v_1, v_2 of π, the matrix coefficient function $g \mapsto \langle\pi(g)v_1, v_2\rangle$ belongs to $L^{2+\epsilon}(G)$ for any $\epsilon > 0$. We describe the non-tempered part of $\hat{G}$ in the next subsection.

3.2. **Standard representations and complementary series.** Let α denote the simple relative root for $(\mathfrak{g},\mathfrak{a})$. The root subspace $\mathfrak{n}$ of α has dimension $n-1$ and hence ρ, the half-sum of all positive roots of $(\mathfrak{g},\mathfrak{a})$ with multiplicities, is given by $\frac{n-1}{2}\alpha$. Set $N = \exp\mathfrak{n}$. By the Iwasawa decomposition, every element $g \in G$ can be written uniquely as $g = kan$ with $k \in K$, $a \in A$ and $n \in N$. We write $\kappa(g) = k$, $\exp H(g) = a$ and $n(g) = n$.

For any $g \in G$ and $k \in K$, we let $\kappa_g(k) = \kappa(gk)$, and $H_g(k) = H(gk)$ so that

$$gk = \kappa_g(k)\exp(H_g(k))n(gk).$$

Given a complex valued linear function λ on $\mathfrak{a}$, we define a G-representation U^λ on $L^2(K)$ by the prescription: for $\phi \in L^2(K)$ and $g \in G$,

$$U^\lambda(g)\phi = e^{(-(\lambda+2\rho)\circ H_{g^{-1}})} \cdot \phi \circ \kappa_{g^{-1}}. \tag{3.2}$$

This is called a standard representation of G (cf. [65, Sec. 5.5]). Observe that the restriction of U^λ to K coincides with the left regular representation of K on $L^2(K)$: $U^\lambda(k_1)f(k) = f(k_1^{-1}k)$. If R denotes the right regular representation of K on $L^2(K)$, then $R(m)U^\lambda(g) = U^\lambda(g)R(m)$ for all $m \in M$. In particular each M-invariant subspace of $L^2(K)$ for the right translation action is a G-invariant subspace of U^λ.

Following [65], for any $\upsilon \in \hat{M}$, we let $\mathbf{Q}(\upsilon)L^2(K)$ denote the isotypic $R(M)$ submodule of $L^2(K)$ of type υ. Choosing a finite dimensional vector space, say, V on which M acts irreducibly via υ, it is shown in [65] that the υ-isotypic space $\mathbf{Q}(\upsilon)L^2(K)$ can be written as a sum of $\dim(\upsilon)$ copies of $\mathcal{U}_\upsilon(\lambda)$ where

$$\mathcal{U}_\upsilon(\lambda) = \left\{ f \in L^2(K,V) : \begin{array}{c} \text{for each } m \in M, \\ f(km) = \upsilon(m)f(k), \text{for almost all } k \in K \end{array} \right\}.$$

If $\lambda \in (\frac{n-1}{2} + i\mathbb{R})\alpha$, then $\mathcal{U}_\upsilon(\lambda)$ is unitary with respect to the inner product $\langle f_1, f_2\rangle = \int_K \langle f_1(k), f_2(k)\rangle_V dk$, and called a unitary principal series representation. These representations are tempered. A representation $\mathcal{U}_\upsilon(\lambda)$ with $\lambda \notin (\frac{n-1}{2} + i\mathbb{R})\alpha$ is called a *complementary series* representation if it is unitarizable. For $\lambda = r\alpha$, we will often omit α for simplicity. For $n = 2$, the complementary series representations of $G = \mathrm{SO}(2,1)^\circ$ are $\mathcal{U}_1(s-1)$ with $1/2 < s < 1$; in particular they are all spherical. For $n \geq 3$, a representation $\upsilon \in \hat{M}$ is specified by its highest weight, which can be regarded as a sequence of $\frac{n-1}{2}$ integers with $j_1 \geq j_2 \geq \cdots \geq |j_{(n-1)/2}|$ if n is odd, and as a sequence of $\frac{n-2}{2}$ integers with $j_1 \geq j_2 \geq \cdots \geq j_{(n-2)/2} \geq 0$ if n is even. In both cases, let $\ell = \ell(\upsilon)$ be the largest index such that $j_\ell \neq 0$ and put $\ell(\upsilon) = 0$ if υ is the trivial representation. Then the complementary series representations are precisely $\mathcal{U}_\upsilon(s-n+1)$, $s \in I_\upsilon := (\frac{n-1}{2}, (n-1)-\ell)$, up to equivalence.

In particular, the spherical complementary series representations are exhausted by $\{\mathcal{U}_1(s-n+1) : (n-1)/2 < s < n-1\}$.

The complementary representation $\mathcal{U}_\upsilon(\lambda)$ contains the minimal K-type, say, σ_υ with multiplicity one.

The classification of $\hat{G}$ says that if $\pi \in \hat{G}$ is non-trivial and non-tempered, then π is (equivalent to) the unique irreducible subquotient of the complementary series representation $\mathcal{U}_\upsilon(s-n+1)$, $s \in I_\upsilon$, containing the K-type σ_υ, which we will denote by $\mathcal{U}(\upsilon, s-n+1)$. This classification was obtained by Hirai [26]; see also [30, Prop. 49 and 50]) and [3].

Note that $\mathcal{U}(\upsilon, s-n+1)$ is spherical if and only if $\mathcal{U}_\upsilon(s-n+1)$ is spherical if and only if $\upsilon = 1$. For convenience, we will call $\mathcal{U}(\upsilon, s-n+1)$ a complementary series representation of parameter (υ, s).

Observe that the non-spherical complementary series representations exist only when $n \geq 4$. For $\frac{n-1}{2} < s < n-1$, we will set $\mathcal{H}_s := \mathcal{U}(1, s-n+1)$, i.e., the spherical complementary series representations of parameter s. Our normalization of the Casimir element $\mathcal{C}$ is so that $\mathcal{C}$ acts on $\mathcal{H}_s$ as the scalar $s(s-n+1)$.

In order to study the matrix coefficients of complementary series representations, we work with the standard representations, which we first relate with Eisenstein integrals.

3.3. **Generalized spherical functions and Eisenstein integrals.** Fix a complex valued linear function λ on $\mathfrak{a}$, and the standard representation U^λ. By the Peter-Weyl theorem, we may decompose the left-regular representation $V = L^2(K)$ as $V = \oplus_{\sigma \in \hat{K}} V_\sigma$, where $V_\sigma = L^2(K;\sigma)$ denotes the isotypic K-submodule of type σ, and $V_\sigma \simeq d_\sigma \cdot \sigma$ where d_σ denotes the dimension of σ.

Set $\Omega_K = 1 + \omega_K = 1 - \sum X_i^2$ where $\{X_i\}$ is an orthonormal basis of $\mathfrak{k}_{\mathbb{C}}$. It belongs to the center of the universal enveloping algebra of $\mathfrak{k}_{\mathbb{C}}$. By Schur's lemma, Ω_K acts on V_σ by a scalar, say, $c(\sigma)$. Since Ω_K acts as a skew-adjoint operator, $c(\sigma)$ is real. Moreover $c(\sigma) \geq 1$, see [65, p. 261], and $\|\Omega_K^\ell v\| = c(\sigma)^\ell \|v\|$ for any smooth vector $v \in V_\sigma$. Furthermore it is shown in [65, Lemma 4.4.2.3] that if ℓ is large enough, then

$$\sum_{\sigma \in \hat{K}} d_\sigma^2 \cdot c(\sigma)^{-\ell} < \infty. \tag{3.3}$$

For $\sigma \in \hat{K}$ and $k \in K$ define

$$\chi_\sigma(k) := d_\sigma \cdot \mathrm{Tr}(\sigma(k))$$

where Tr is the trace operator.

For any continuous representation W of K, and $\sigma \in \hat{K}$, the projection operator from W to the σ-isotypic space $W(\sigma)$ is given as follows:

$$\mathrm{P}_\sigma = \int_K \overline{\chi}_\sigma(k) W(k) dk.$$

For $\upsilon \in \hat{M}$, we write $\upsilon \subset \sigma$ if υ occurs in $\sigma|_M$, and we write $\upsilon \subset \sigma \cap \tau$ if υ occurs both in $\sigma|_M$ and $\tau|_M$. We remark that for $\sigma \in \hat{K}$, given $\upsilon \in \hat{M}$ occurs at most once in $\sigma|_M$ ([15]; [65, p.41]). For $\upsilon \subset \sigma$, we denote by P_υ the projection operator from V_σ to the υ-isotypic subspace $V_\sigma(\upsilon) \simeq d_\sigma \cdot \upsilon$ so that any $w \in V_\sigma$ can be written as $w = \sum_{\upsilon \subset \sigma} \mathrm{P}_\upsilon(w)$. By the theory of representations of compact Lie groups, we have for any $f \in L^2(K;\sigma)$ we have

$$\langle f, \bar{\chi}_\sigma \rangle = f(e).$$

In the rest of this subsection, we fix $\sigma, \tau \in \hat{K}$. Define an M-module homomorphism $T_0 : V_\sigma \to V_\tau$ by

$$T_0(w) = \sum_{\upsilon \subset \sigma \cap \tau} \langle \mathrm{P}_\upsilon(w), \mathrm{P}_\upsilon(\bar{\chi}_\sigma) \rangle \mathrm{P}_\upsilon(\bar{\chi}_\tau).$$

Set $E := \mathrm{Hom}_{\mathbb{C}}(V_\sigma, V_\tau)$. Then E is a (τ,σ)-double representation space, a left τ and right σ-module. We put

$$E_M := \{T \in E : \tau(m)T = T\sigma(m) \text{ for all } m \in M\}.$$

Denote by U^λ_σ and U^λ_τ the representations of K obtained by restricting $U^\lambda|_K$ to the subspace V_σ and V_τ respectively. Define $T_\lambda \in E$ by

$$T_\lambda := \int_M U^\lambda_\tau(m) T_0 U^\lambda_\sigma(m^{-1}) dm$$

where dm denotes the probability Haar measure of M. It is easy to check that $T_\lambda \in E_M$.

An integral of the form $\int_K U^\lambda_\tau(\kappa(ak)) T_\lambda U^\lambda_\sigma(k^{-1}) e^{\lambda(H(ak))} dk$ is called an *Eisenstein integral*.

Clearly, the matrix coefficients of the representation U^λ are understood if we understand $P_\tau U^\lambda(a) P_\sigma$ for all $\tau, \sigma \in \hat{K}$, which can be proved to be an Eisenstein integral:

Theorem 3.4. *For any $a \in A$, we have*

$$\mathrm{P}_\tau U^\lambda(a) \mathrm{P}_\sigma = \int_K U^\lambda_\tau(\kappa(ak)) T_\lambda U^\lambda_\sigma(k^{-1}) e^{\lambda(H(ak))} dk.$$

Proof. For $\bullet \in \hat{K}$ and $\phi \in L^2(K;\bullet)$, we write $U^\lambda_\bullet(k^{-1})\phi = \sum_{\upsilon \subset \bullet} \phi_{k,\upsilon}$, that is, $\phi_{k,\upsilon} = \mathrm{P}_\upsilon(U^\lambda_\bullet(k^{-1})\phi)$. In particular, $\phi(k) = \sum_{\upsilon\subset\bullet} \phi_{k,\upsilon}(e)$ and

$$\phi_{k,\upsilon}(e) = \langle \phi_{k,\upsilon}, \bar{\chi}_\bullet \rangle = \langle U^\lambda_\bullet(k^{-1})\phi, \mathrm{P}_\upsilon(\bar{\chi}_\bullet)\rangle.$$

Let $\varphi \in V_\sigma$ and $\psi \in V_\tau$. For any $g \in G$, we have

$$\begin{aligned}\langle U^\lambda_\tau(\kappa(gk)) T_0 U^\lambda_\sigma(k^{-1})\varphi, \psi\rangle &= \sum_{\upsilon \subset \sigma\cap\tau} \langle U^\lambda_\sigma(k^{-1})\varphi), \mathrm{P}_\upsilon(\bar{\chi}_\sigma)\rangle \langle \mathrm{P}_\upsilon(\bar{\chi}_\tau), U^\lambda_\tau(\kappa(gk))^{-1}\psi\rangle \\ &= \sum_{\upsilon\subset\sigma\cap\tau} \varphi_{k,\upsilon}(e)\overline{\psi_{\kappa(gk),\upsilon}(e)}. \end{aligned} \tag{3.5}$$

On the other hand, we have

$$\begin{aligned}\langle U^\lambda(a)\varphi, \psi\rangle &= \int_K \varphi(\kappa(a^{-1}k)\overline{\psi(k)} e^{-(\lambda+2\rho)H(a^{-1}k)} dk \\ &= \int_K \varphi(k)\overline{\psi(\kappa(ak))} e^{\lambda(H(ak))} dk \\ &= \int_K (\sum_{\upsilon\subset\sigma} \varphi_{k,\upsilon}(e))(\overline{\sum_{\upsilon\subset\tau} \psi_{\kappa(ak),\upsilon}(e)}) e^{\lambda(H(ak))} dk. \end{aligned} \tag{3.6}$$

We now claim that; if $\upsilon_1 \neq \upsilon_2$, then

$$\int_K \varphi_{k,\upsilon_1}(e)\overline{\psi_{\kappa(ak),\upsilon_2}(e)} e^{\lambda(H(ak))} dk = 0$$

To see this, first note that M and a commute, and hence $H(amk) = H(ak)$, and $\kappa(amk) = m\kappa(ak)$. We also note that

$$\varphi_{k,\upsilon_1} \in V_\sigma(\upsilon_1), \text{ and } \psi_{\kappa(ak),\upsilon_2} \in V_\tau(\upsilon_2).$$

Now if $\upsilon_1 \neq \upsilon_2$, then by Schur's orthogonality of matrix coefficients,

$$\int_M \varphi_{k,\upsilon_1}(m^{-1})\overline{\psi_{\kappa(ak),\upsilon_2}(m^{-1})}dm = \int_M \langle U^\lambda_\sigma(m)\varphi_{k,\upsilon_1}, \mathrm{P}_{\upsilon_1}(\bar{\chi}_\sigma)\rangle\overline{\langle U^\lambda_\tau(m)\psi_{k',\upsilon_2}, \mathrm{P}_{\upsilon_2}(\bar{\chi}_\tau)\rangle}dm = 0;$$

we get

$$\begin{aligned}
&\int_K \varphi_{k,\upsilon_1}(e)\overline{\psi_{\kappa(ak),\upsilon_2}(e)}e^{\lambda(H(ak))}dk \\
&= \int_{M\backslash K}(\int_M \varphi_{mk,\upsilon_1}(e)\overline{\psi_{\kappa(amk),\upsilon_2}(e)}e^{\lambda(H(amk))}dm)dk \\
&= \int_{M\backslash K} e^{\lambda(H(ak))}(\int_M \varphi_{k,\upsilon_1}(m^{-1})\overline{\psi_{\kappa(ak),\upsilon_2}(m^{-1})}dm)dk = 0,
\end{aligned}$$

implying the claim.

Therefore, it follows from (3.5) and (3.6) that for any $\varphi \in V_\sigma$ and $\psi \in V_\tau$,

$$\begin{aligned}
\langle \mathrm{P}_\tau U^\lambda(a)\mathrm{P}_\sigma\varphi, \psi\rangle &= \langle U^\lambda(a)\varphi, \psi\rangle \\
&= \int_K \sum_{\upsilon\subset\sigma\cap\tau} \varphi_{k,\upsilon}(e)\overline{\psi_{\kappa(gk),\upsilon}(e)}e^{\lambda(H(ak))}dk \\
&= \int_K \langle U^\lambda_\tau(\kappa(ak))T_0U^\lambda_\sigma(k^{-1})\varphi, \psi\rangle e^{\lambda(H(ak))}dk \\
&= \int_K \langle U^\lambda_\tau(\kappa(ak))T_\lambda U^\lambda_\sigma(k^{-1})\varphi, \psi\rangle e^{\lambda(H(ak))}dk;
\end{aligned}$$

we have used $\kappa(akm) = \kappa(ak)m$ and $H(akm) = H(ak)$ for the last equality.

It follows that

$$P_\tau U^\lambda(a)P_\sigma = \int_K U^\lambda_\tau(\kappa(ak))T_\lambda U^\lambda_\sigma(k^{-1})e^{\lambda(H(ak))}dk.$$

□

For the special case of $\tau = \sigma$, this theorem was proved by Harish-Chandra (see [66, Thm. 6.2.2.4]), where T_0 was taken to be $T_0(w) = (w, \bar{\chi}_\sigma)\bar{\chi}_\sigma$ and $T_\lambda = \int_M U^\lambda_\sigma(m)T_0U^\lambda_\sigma(m^{-1})dm$.

Lemma 3.7. *For any $\lambda \in \mathbb{C}$,*

$$\|T_\lambda\| \leq d^2_\sigma \cdot d^2_\tau$$

where $\|T_\lambda\|$ denotes the operator norm of T_λ.

Proof. Since $\|\chi_\upsilon\| \leq d^2_\upsilon$ for any $\upsilon \in \hat{M}$,

$$\begin{aligned}
\|T_0\|^2 &\leq \sum_{\upsilon\subset\sigma\cap\tau} \|\chi_\upsilon\|^2 \leq \sum_{\upsilon\subset\sigma\cap\tau} d^4_\upsilon \\
&\leq (\sum_{\upsilon\subset\sigma} d^2_\upsilon \cdot (\sum_{\upsilon\subset\tau} d^2_\upsilon) \leq d^2_\sigma \cdot d^2_\tau.
\end{aligned}$$

Since $\|T_\lambda\| \leq \|T_0\| \cdot \|\sigma\| \cdot \|\tau\| = \|T_0\|$, the claim follows. □

3.4. **Harish-Chandra's expansion of Eisenstein integrals.** Fix $\sigma, \tau \in \hat{K}$. Let E and E_M to be as in the previous subsection.

Given $T \in E_M$, $r \in \mathbb{C}$, and $a_t \in A^+$, we investigate an Eisenstein integral:

$$\int_K \tau(\kappa(a_t k)) T_{ir\alpha-\rho} \sigma(k^{-1}) e^{(ir\alpha-\rho)(H(a_t k))} dk.$$

We recall the following fundamental result of Harish-Chandra:

Theorem 3.8. *(Cf.* [66, Theorem 9.1.5.1]*) There exists a subset $\mathcal{O}_{\sigma,\tau}$ of $\mathbb{C}$, whose complement is a locally finite set, such that for any $r \in \mathcal{O}_{\sigma,\tau}$ there exist uniquely determined functions $\mathrm{c}_+(r), \mathrm{c}_-(r) \in \mathrm{Hom}_{\mathbb{C}}(E_M, E_M)$ such that for all $T \in E_M$,*

$$\rho(a_t) \int_K \tau(\kappa(a_t k)) T \sigma(k^{-1}) e^{(ir\alpha-\rho)H(a_t k)} dk \\ = \Phi(r : a_t)\mathrm{c}_+(r)T + \Phi(-r : a_t)\mathrm{c}_-(r)T$$

where Φ is a function on $\mathcal{O}_{\sigma,\tau} \times A^+$ taking values in $\mathrm{Hom}_{\mathbb{C}}(E_M, E_M)$, defined as in (3.12)*.*

Let us note that, fixing T and a_t, the Eisenstein integral on the left hand side of the above is an entire function of r; see [66, Section 9.1.5].

Much of the difficulties lie in the fact that the above formula holds only for $\mathcal{O}_{\sigma,\tau}$ but not for the entire complex plane, as we have no knowledge of which complementary series representations appear in $L^2(\Gamma\backslash G)$. We need to understand the Eisenstein integral $\int_K \tau(\kappa(a_t k)) T_{s\alpha-2\rho} \sigma(k^{-1}) e^{(s\alpha-2\rho)H(a_t k)} dk$ for every $s \in (\frac{n-1}{2}, n-1)$. We won't be able to have as precise as a formula as Theorem 3.8 but will be able to determine a main term with an exponential error term.

We begin by discussing the definition and properties of the functions Φ and $\mathrm{c}_\pm$.

3.4.1. *The function Φ.* As in [66, page 287], we will recursively define rational functions $\{\Gamma_\ell : \ell \in \mathbb{Z}_{\geq 0}\}$ which are holomorphic except at a locally finite subset, say $\mathcal{S}_{\sigma,\tau}$. The subset $\mathcal{O}_{\sigma,\tau}$ in Theorem 3.8 is indeed $\mathbb{C} - \cup_{r \in \mathcal{S}_{\sigma,\tau}} \{\pm r\}$.

More precisely, let $\mathfrak{l}$ be the Lie algebra of the Cartan subalgebra (=the centralizer of A). Let $H_\alpha \in \mathfrak{l}_{\mathbb{C}}$ be such that $B(H, H_\alpha) = \alpha(H)$ for all $H \in \mathfrak{l}_{\mathbb{C}}$.

Let $X_{\pm\alpha} \in \mathfrak{g}_{\mathbb{C}}^{\pm\alpha}$ be chosen so that $[X_\alpha, X_{-\alpha}] = H_\alpha$ and $[H, X_\alpha] = \alpha(H)X_\alpha$. In particular, $B(X_\alpha, X_{-\alpha}) = 1$. Write $X_{\pm\alpha} = Y_{\pm\alpha} + Z_{\pm\alpha}$ where $Y_{\pm\alpha} \in \mathfrak{k}_{\mathbb{C}}$ and $Z_{\pm\alpha} \in \mathfrak{p}_{\mathbb{C}}$.

Letting Ω_M denote the Casimir element of M, given $S \in \mathrm{Hom}_{\mathbb{C}}(E_M, E_M)$, we define $f(S)$ by

$$f(S)T = ST\sigma(\Omega_M), \text{ for all } T \in E_M.$$

We now define $\Gamma_\ell := \Gamma_\ell(ir - \frac{n-1}{2})$'s in $\mathcal{Q}(\mathfrak{a}_{\mathbb{C}}) \otimes \mathrm{Hom}_{\mathbb{C}}(E_M, E_M)$ by the following recursive relation (see [66, p. 286] for the def. of $\mathcal{Q}(\mathfrak{a}_{\mathbb{C}})$): $\Gamma_0 = I$

and

$$\{\ell(2ir-n+1)-\ell(\ell-n+1)-f\}\,\Gamma_\ell = \sum_{j\geq 1}((2ir-n+1)-2(\ell-2j))\Gamma_{\ell-2j}$$
$$+8\sum_{j\geq 1}(2j-1)\tau(Y_\alpha)\sigma(Y_{-\alpha})\Gamma_{\ell-(2j-1)}-8\sum_{j\geq 1} j\,\{\tau(Y_\alpha Y_{-\alpha})+\sigma(Y_\alpha Y_{-\alpha})\}\,\Gamma_{\ell-2j}.$$

The set $\mathcal{O}_{\sigma,\tau}$ consists of r's such that $\{\ell(2ir-n+1)-\ell(\ell-n+1)-f\}$ is invertible so that the recursive definition of the Γ_ℓ's is meaningful.

Lemma 3.9. *Fix any $t_0>0$ and a compact subset $\omega\subset\mathcal{O}_{\sigma,\tau}$. There exist b_ω (depending only on t_0 and ω) and $N_0>1$ (independent of $\sigma,\tau\in\hat{K}$) such that for any $r\in\omega$ and $\ell\in\mathbb{N}$,*

$$\|\Gamma_\ell(ir-\tfrac{n-1}{2})\|\leq b_\omega d_\sigma^{N_0} d_\tau^{N_0} e^{\ell t_0}.$$

Proof. Our proof uses an idea of the proof of [66, Lemmas 9.1.4.3-4]. For $s=ir-\frac{n-1}{2}$, and $T\in\mathrm{Hom}_{\mathbb{C}}(E_M,E_M)$, define

$$\Lambda_\ell(T):=\left(-\ell^2+\ell(2s-n+1)-f\right)T.$$

For q_σ and q_τ which are respectively the highest weights for σ and τ, since $q_\sigma\ll d_\sigma$ with implied constant independent of $\sigma\in\hat{K}$,

$$\begin{aligned}&\max\{\|\tau(Y_\alpha)\sigma(Y_{-\alpha})\|,\|\tau(Y_\alpha Y_{-\alpha})\|,\sigma(Y_\alpha Y_{-\alpha})\|\}\\&\leq c_0\cdot(q_\sigma q_\tau+q_\sigma^2+q_\sigma^2)d_\sigma d_\tau\leq c_0' d_\sigma^3 d_\tau^3\end{aligned}$$

for some $c_0,c_0'>0$ independent of σ and τ. Hence for some $c_1=c_1(\omega)$, for all $r\in\omega$,

$$\|\Gamma_\ell(ir-\tfrac{n-1}{2})\|\leq \ell\cdot\|\Lambda_\ell^{-1}\|\cdot c_1 d_\sigma^3 d_\tau^3\sum_{j<\ell}\|\Gamma_{\ell-j}\|. \tag{3.10}$$

Let N_1 be an integer such that $\ell^2\cdot\|\Lambda_\ell^{-1}\|\cdot(1-e^{-t_0})^{-1}c_1 d_\sigma^3 d_\tau^3\leq N_1$ for all $\ell\geq 1$. Since $\|\Lambda_\ell^{-1}\|\ll\ell^{-2}$ as $\ell\to\infty$ and the coefficients of f depend only on the eigenvalues of Ω_M for those $\upsilon\in\hat{M}$ contained in σ, we can take $N_1=N_1(\omega)$ so that $N_1\leq c_2 d_\sigma^{N_2} d_\tau^{N_2}$ for some $N_2\geq 1$ and $c_2=c_2(\omega)>1$ (independent of σ and τ).

Set

$$M(t_0,\omega):=\max_{\ell\leq N_1, r\in\omega}\|\Gamma_\ell(ir-\tfrac{n-1}{2})\|e^{-\ell t_0}.$$

By (3.10) together with the observation that both $N_1=N_1(\omega)$ and $\max_{\ell\leq N_1}\|\Lambda_\ell^{-1}\|$ are bounded by a polynomial in d_σ and d_τ, we have $M(t_0,\omega)\leq b_\omega d_\sigma^{N_0} d_\tau^{N_0}$ for some $N_0\geq 1$ and $b_\omega>0$.

We shall now show by induction that for all $r\in\omega$ and for all $\ell\geq 1$,

$$\|\Gamma_\ell(ir-\tfrac{n-1}{2})\|\leq M(t_0,\omega)e^{\ell t_0}. \tag{3.11}$$

First note that (3.11) holds for all $\ell \le N_1$ by the definition of $M(t_0,\omega)$. Now for any $N_1 > N$, suppose (3.11) holds for each $\ell < N$. Then

$$\begin{aligned}\|\Gamma_N(ir-\tfrac{n-1}{2})\| &\le N^{-1}(N^2\|\Lambda_N^{-1}\|c_1 d_\sigma^3 d_\tau^3)\sum_{j<N}\|\Gamma_{N-j}(ir-\tfrac{n-1}{2})\|\\ &\le N^{-1}N_1(1-e^{-t_0})M(t_0,\omega)\sum_{j<N}e^{(N-j)t_0}\\ &\le M(t_0,\omega)e^{Nt_0},\end{aligned}$$

finishing the proof. □

Following Warner (Cf. [66, Theorem 9.1.4.1]), we define

$$\Phi(r:a_t) = e^{irt}\sum_{\ell\ge 0}\Gamma_\ell(ir-\tfrac{n-1}{2})e^{-\ell t} \tag{3.12}$$

which converges for all large enough t by Lemma 3.9.

3.4.2. *The function* $\mathrm{c}_\pm$. Let $N^- = \exp(\mathfrak{n}^-)$ be the root subspace corresponding to $-\alpha$, and d_{N^-} denote a Haar measure on N^- normalized so that $\int_{N^-} e^{-2\rho(H(n))}d_{N^-}(n) = 1$.

The following is due to Harish-Chandra; see [66, Thm. 9.1.6.1].

Theorem 3.13. *For $r \in \mathcal{O}_{\sigma,\tau}$ with $\Im(r) < 0$, $\mathrm{c}_+(r)$ is holomorphic and given by*

$$\mathrm{c}_+(r)T = \int_{N^-} T\sigma(\kappa(n)^{-1})e^{-(ir\alpha+\rho)(H(n))}d_{N^-}(n).$$

The integral $\int_{N^-} e^{-(ir\alpha+\rho)H(n)}d_{N^-}(n)$ is absolutely convergent iff $\Im(r) < 0$, shown by Gindikin and Karperlevic ([66, Coro.9.1.6.5]).

Corollary 3.14. *For any $r \in \mathcal{O}_{\sigma,\tau}$ with $\Im(r) < 0$, the operator norm $\|\mathrm{c}_+(r)\|$ is bounded above by $\int_{N^-} e^{(\Im(r)\alpha-\rho)H(n)}d_{N^-}(n)$.*

Proof. Since the operator norm $\|\sigma(k)\|$ is 1 for any $k \in K$, the claim is immediate from Theorem 3.13. □

Proposition 3.15. *Fix a compact subset ω contained in $\mathcal{O}_{\sigma,\tau}\cap\{\Im(r) < 0\}$. There exist $d_1 = d_1(\omega)$ and $N_2 > 1$ such that for any $r \in \omega$, we have*

$$\|\mathrm{c}_\pm(r)T_{\pm ir\alpha-\rho}\| \le d_1\cdot d_\tau^{N_2}d_\sigma^{N_2}.$$

Proof. By the assumption on ω, the integral $\int_{N^-} e^{-(\Re(ir)\alpha+\rho)H(n)}d_{N^-}(n)$ converges uniformly for all $r \in \omega$. Hence the bound for $\mathrm{c}_+(r)T_{ir\alpha-\rho}$ follows from Corollary 3.14 together with Lemma 3.7. To get a bound for $\mathrm{c}_-(r)T_{-ir\alpha-\rho}$, we recall that

$$\begin{aligned}e^{(-ir+\frac{n-1}{2})t}\int_K \tau(\kappa(a_tk))T\sigma(k^{-1})e^{(ir\alpha-\rho)(H(a_tk))}dk&\\ = e^{-irt}\Phi(r:a_t)\mathrm{c}_+(r)T + e^{-irt}\Phi(-r:a_t)\mathrm{c}_-(r)T.&\end{aligned}$$

Then $e^{-irt}\Phi(-r:a_t) = I + \sum_{\ell\geq 1}\Gamma_\ell(-ir-\frac{n-1}{2})e^{-\ell t}$, and applying Lemma 3.9 with $t_0 = 1$, we get

$$\sum_{\ell\geq 1}\|\Gamma_\ell(-ir-\tfrac{n-1}{2})e^{-\ell t}\| \leq b_\omega d_\sigma^{N_0} d_\tau^{N_0}\sum_{\ell\geq 1} e^{\ell(1-t)}.$$

Fix $t_0 > 0$ so that $b_\omega d_\sigma^{N_0} d_\tau^{N_0}\sum_{\ell\geq 1} e^{\ell(1-t_0)} < 1/2$; then $t_0 \gg \log(d_\sigma d_\tau)$. Now $A_r := e^{-irt_0}\Phi(-r:a_{t_0})$ is invertible and for some N_1 and b'_ω,

$$\|A_r^{-1}\| \leq b'_\omega d_\sigma^{N_1} d_\tau^{N_1}. \tag{3.16}$$

Since the map $k \mapsto H(a_{t_0}k)$ is continuous, we have $\int_K |e^{(ir\alpha-\rho)(H(a_{t_0}k))}|dk < d_\omega$ for all $r \in \omega$, and hence

$$\begin{aligned}
&\|\mathrm{c}_-(r)T_{ir\alpha-\rho}\| \\
&\leq \|A_r^{-1}\|\cdot|e^{(-ir+\frac{n-1}{2})t_0}\int_K \tau(\kappa(a_{t_0}k))T_{ir\alpha-\rho}\sigma(k^{-1})e^{(ir\alpha-\rho)(H(a_{t_0}k))}dk| \\
&+ \|A_r^{-1}\|\cdot\|e^{-irt_0}\Phi(r:a_{t_0})\mathrm{c}_+(r)T_{ir\alpha-\rho}\| \\
&\leq \|A_r^{-1}\|\cdot d_\omega\left(\max_{k\in K}\|\tau(\kappa(a_{t_0}k))T_{ir\alpha-\rho}\sigma(k^{-1})\| + \|\mathrm{c}_+(r)T_{ir\alpha-\rho}\|\right).
\end{aligned}$$

Hence the claim on $\|\mathrm{c}_-(r)T_{ir\alpha-\rho}\|$ now follows from (3.16), Lemma 3.7 and the bound for $\|\mathrm{c}_+(r)T_{ir\alpha-\rho}\|$. □

3.5. **Asymptotic expansion of the matrix coefficients of the complementary series.** Fix a parameter $(n-1)/2 < s_0 < (n-1)$, and recall that $2\rho = (n-1)\alpha$. We apply the results of the previous subsections to the standard representation $U^{(s_0-n+1)\alpha} = U^{s_0\alpha-2\rho}$.

The following theorem is a key ingredient of the proof of Theorem 3.30.

Theorem 3.17. *There exist $\eta_0 > 0$ and $N > 1$ such that for any $\sigma, \tau \in \hat{K}$, for all $t > 2$, we have*

$$P_\tau U^{(s_0-n+1)\alpha}(a_t)P_\sigma = e^{(s_0-n+1)t}\mathrm{c}_+(r_{s_0})T_{(s_0-n+1)\alpha} + O(d_\sigma^N\cdot d_\tau^N e^{(s_0-n+1-\eta_0)t}),$$

with the implied constant independent of σ, τ.

Proof. Set $r_s := -i(s-\rho) \in \mathbb{C}$, for all $s \in \mathbb{C}$. In particular, $\Im(r_s) < 0$ for $(n-1)/2 < s < (n-1)$.

Fix $t > 0$, and define $F_t : \mathbb{C} \to E_M$ by

$$F_t(s) := P_\tau U^{s\alpha-2\rho}(a_t)P_\sigma.$$

By Theorem 3.4,

$$F_t(s) = \int_K \tau(\kappa(a_t k))T_{s\alpha-2\rho}\sigma(k^{-1})e^{(s\alpha-2\rho)H(a_t k)}dk.$$

As was remarked following Theorem 3.8, for each fixed $t > 0$, the function $F_t(s)$ is analytic on $\mathbb{C}$. Thus in view of Theorem 3.13, we have, whenever $r_s \in \mathcal{O}_{\sigma,\tau}$ and $\Im(r_s) < 0$,

$$F_t(s) - e^{(s-n+1)t}\mathrm{c}_+(r_s)T_{s\alpha-2\rho} \text{ is analytic.} \tag{3.18}$$

Recall the notation $\mathcal{S}_{\sigma,\tau}$, that is, $\mathcal{O}_{\sigma,\tau} = \mathbb{C} - \cup_{r\in\mathcal{S}_{\sigma,\tau}}\{\pm r\}$, and set $\tilde{\mathcal{S}}_{\sigma,\tau} = \{s : r_s \in \mathcal{S}_{\sigma,\tau}\}$. Define

$$G_t(s) = F_t(s) - e^{(s-n+1)t}\mathrm{c}_+(r_s)T_{s\alpha-2\rho}.$$

Indeed the map $s \mapsto G_t(s)$ is analytic on $\{s : \Im(r_s) < 0\} - \tilde{\mathcal{S}}_{\sigma,\tau}$. Since $\cup_{\sigma',\tau'\in\hat{K}} \pm\tilde{\mathcal{S}}_{\sigma',\tau'}$ is countable, we may choose a small circle ω' of radius at most $1/2$ centered at s_0 such that $\{r_s : s \in \omega'\} \cap \left(\cup_{\sigma',\tau'\in\hat{K}} \pm\mathcal{S}_{\sigma',\tau'}\right) = \emptyset$.

Observe that the intersection of ω' and the real axis is contained in the interval $((n-1)/2, n-1)$. Note that there exists $\eta > 0$ such that for all $s \in \omega'$,

$$(n-1) - s_0 + \eta < \Re(s) < s_0 + 1 - \eta. \tag{3.19}$$

Then $G_t(s)$ is analytic on the disc bounded by ω'. Hence by the maximum principle, we get

$$\|G_t(s_0)\| \le \max_{s\in\omega'} \|G_t(s)\|. \tag{3.20}$$

Since $\omega := \{r_s : s \in \omega'\} \subset \mathcal{O}_{\sigma,\tau}$, Theorems 3.4 and 3.8 imply that for all $s \in \omega'$, we have

$$\begin{aligned} G_t(s) = e^{(s-n+1)t}\Big(&\sum_{\ell\ge1} e^{-\ell t}\Gamma_\ell(ir_s - \tfrac{n-1}{2})\mathrm{c}_+(r_s)T_{s\alpha-2\rho}\Big) \\ &+ e^{-st}\Big(\sum_{\ell\ge0} e^{-\ell t}\Gamma_\ell(-ir_s - \tfrac{n-1}{2})\mathrm{c}_-(r_s)T_{s\alpha-2\rho}\Big). \end{aligned}$$

Fixing any $t_0 > 0$, by Lemma 3.9, there exists $b_0 = b_0(t_0,\omega) > 0$ such that for all $r \in \omega$,

$$\|\Gamma_\ell(ir - \tfrac{n-1}{2})\| \le b_0 \cdot d_\sigma^{N_0} \cdot d_\tau^{N_0} \cdot e^{\ell t_0}. \tag{3.21}$$

By Proposition 3.15, for all $r \in \omega$,

$$\|\mathrm{c}_\pm(r)T_{\pm ir\alpha-\rho}\| \le d_1 \cdot d_\sigma^{N_2} \cdot d_\tau^{N_2}.$$

Let $t > t_0 + 1$ so that $\sum_{\ell\ge0} e^{-\ell(t-t_0)} \le 2$. Then we have for any $t > 0$ and $s \in \omega'$,

$$\begin{aligned} &\|\sum_{\ell\ge1} e^{-\ell t}\Gamma_\ell(ir_s - \tfrac{n-1}{2})\mathrm{c}_+(r_s)T_{s\alpha-2\rho}\| \\ &\le \cdot d_\sigma^{N_0+N_2} d_\tau^{N_0+N_2} \cdot b_{\sigma,\tau} \cdot e^{-t}e^{t_0}\sum_{\ell\ge0} e^{-\ell(t-t_0)} \\ &\le \left(2e^{t_0} \cdot b_0 \cdot d_\sigma^{N_0+N_2} \cdot d_\tau^{N_0+N_2}\right) e^{-t} \end{aligned}$$

and

$$\|\sum_{\ell\ge0} e^{-\ell t}\Gamma_\ell(-ir_s - \tfrac{n-1}{2})\mathrm{c}_-(r_s)T_{s\alpha-2\rho}\| \le 2b_0 \cdot d_\sigma^{N_0+N_2} \cdot d_\tau^{N_0+N_2}. \tag{3.22}$$

We now combine these and the expression for $G_t(s)$, for $s \in \omega'$ and get for all $t > t_0 + 1$,

$$\begin{aligned}&\|G_t(s_0)\|\\&\leq 2b_0(e^{t_0}+1)d_\sigma^{N_0+N_2}\cdot d_\tau^{N_0+N_2}\cdot \max_{s\in\omega'}(e^{(\Re(s)-n)t}+e^{-\Re(s)t})\\&\leq b'\cdot d_\sigma^{N_0+N_2}\cdot d_\tau^{N_0+N_2}e^{(s_0-(n-1)-\eta)t}\end{aligned}$$

where $\eta > 0$ is as in (3.19) and $b' > 0$ is a constant independent of $\sigma, \tau \in \hat{K}$.

Since $P_\tau U^{(s_0-n+1)\alpha}(a_t)P_\sigma = e^{(s_0-n+1)t}\mathrm{c}_+(r_{s_0})T_{(s_0-n+1)\alpha} + G_t(s_0)$, this finishes the proof. □

By Theorem 3.4, Theorem 3.17 implies:

Theorem 3.23. *Let $(n-1)/2 < s_0 < (n-1)$. There exist $\eta_0 > 0$ and $N > 1$ such that for all $t \geq 2$ and for any unit vectors $v_\sigma \in V_\sigma$ and $v_\tau \in V_\tau$,*

$$\begin{aligned}&\langle U^{(s_0-n+1)\alpha}(a_t)(v_\sigma), v_\tau\rangle\\&\qquad = e^{(s_0-n+1)t}\langle \mathrm{c}_+(r_{s_0})T_{(s_0-n+1)\alpha}(v_\sigma), v_\tau\rangle + O(d_\sigma^N d_\tau^N e^{(s_0-n+1-\eta_0)t}),\end{aligned}$$

with the implied constant independent of $\sigma, \tau, v_\sigma, v_\tau$.

3.6. **Decay of matrix coefficients for $L^2(\Gamma\backslash G)$.** Let $\Gamma < G$ be a torsion-free geometrically finite group with $\delta > (n-1)/2$.

By Lax-Phillips [40], Patterson [54] and Sullivan [62], $\mathcal{U}(1, \delta-n+1)$ occurs as a subrepresentation of $L^2(\Gamma\backslash G)$ with multiplicity one, and $L^2(\Gamma\backslash G)$ does not weakly contain any spherical complementary series $\mathcal{U}(1, s-n+1)$ of parameter s strictly bigger than δ. In particular, δ is the maximum s such that $\mathcal{U}(1, s-n+1)$ is weakly contained in $L^2(\Gamma\backslash G)$.

The following proposition then follows from [60, Prop. 3.5] and Theorem 3.23:

Proposition 3.24. *$L^2(\Gamma\backslash G)$ does not weakly contain any complementary series $\mathcal{U}(\upsilon, s-n+1)$ with $\upsilon \in \hat{M}$ and $s > \delta$.*

Definition 3.25 (Spectral Gap)**.** *We say $L^2(\Gamma\backslash G)$ has a spectral gap if the following two conditions hold:*

1. *there exists $n_0 \geq 1$ such that the multiplicity of any complementary series $\mathcal{U}(\upsilon, \delta-n+1)$ of parameter δ occurring in $L^2(\Gamma\backslash G)$ is at most $dim(\upsilon)^{n_0}$ for all $\upsilon \in \hat{M}$;*
2. *there exists $(n-1)/2 < s_0 < \delta$ such that no complementary series with parameter $s_0 < s < \delta$ is weakly contained in $L^2(\Gamma\backslash G)$.*

We set $n_0(\Gamma)$ and $s_0(\Gamma)$ to be the infima of all n_0 and of all s_0 satisfying (1) and (2) respectively. The pair $(n_0(\Gamma), s_0(\Gamma))$ will be referred as the spectral gap data for Γ.

In other words, the spectral gap property of $L^2(\Gamma\backslash G)$ is equivalent to the following decomposition:

$$L^2(\Gamma\backslash G) = \mathcal{H}_\delta^\dagger \oplus \mathcal{W} \tag{3.26}$$

where $\mathcal{H}_\delta^\dagger = \oplus_{\upsilon\in\hat{M}} m(\upsilon)\mathcal{U}(\upsilon, \delta - n + 1)$ with $m(\upsilon) \le \dim(\upsilon)^{n_0}$, and no complementary series representation with parameter $s_0(\Gamma) < s < \delta$ is weakly contained in $\mathcal{W}$.

We recall the strong spectral gap property from Def. 1.1.

Theorem 3.27. *Suppose that $\delta > (n-1)/2$ for $n = 2, 3$ or that $\delta > n-2$ for $n \ge 4$. Then $L^2(\Gamma\backslash G)$ has a strong spectral gap property.*

Proof. By the the classification of the unitary dual $\hat{G}$ explained in the subsection 3.2, any non-spherical complementary series representation is of the form $\mathcal{U}(\upsilon, s-n+1)$ for some $\upsilon \in \hat{M} - \{1\}$ and $s \in (\frac{n-1}{2}, n-2)$ (see [26] and [30]). Together with the aforementioned work of Lax-Phillips on the spherical complementary series representations occurring in $L^2(\Gamma\backslash G)$, this implies the claim. □

For $\Psi \in C^\infty(\Gamma\backslash G)$, $\ell \in \mathbb{N}$ and $1 \le p \le \infty$, we consider the following Sobolev norm:

$$\mathcal{S}_{p,\ell}(\Psi) = \sum \|X(\Psi)\|_p \tag{3.28}$$

where the sum is taken over all monomials X in a fixed basis $\mathcal{B}$ of $\mathfrak{g}$ of order at most ℓ and $\|X(\Psi)\|_p$ denotes the $L^p(\Gamma\backslash G)$-norm of $X(\Psi)$. Since we will be using $\mathcal{S}_{2,\ell}$ most often, we will set $\mathcal{S}_\ell = \mathcal{S}_{2,\ell}$.

For a unitary G-representation space W and a smooth vector $w \in W$, $\mathcal{S}_\ell(w)$ is defined similarly: $\mathcal{S}_\ell(w) = \sum \|X.w\|_2$ where the sum is taken over all monomials X in $\mathcal{B}$ of order at most ℓ.

Proposition 3.29. *Fix $(n-1)/2 < s_0 < (n-1)$. Let W be a unitary representation of G which does not weakly contain any complementary series representation $\mathcal{U}(\upsilon, s-n+1)$ with parameter $s > s_0$ and $\upsilon \in \hat{M}$. Then for any $\epsilon > 0$, there exists $c_\epsilon > 0$ such that for any smooth vectors $w_1, w_2 \in W$ and for any $t > 0$, we have*

$$|\langle a_t w_1, w_2\rangle| \le c_\epsilon \cdot \mathcal{S}_{\ell_0}(w_1) \cdot \mathcal{S}_{\ell_0}(w_2) \cdot e^{(s_0 - n + 1 + \epsilon)t}$$

where $\ell_0 \ge 1$ depends only on G and K.

Proof. This proposition is proved in [35, Proof of Prop. 5.3] for $n = 3$ (based on an earlier idea of [60]), and its proof easily extends to a general $n \ge 2$. □

In the following two theorems, we assume that Γ is Zariski dense in G and that $L^2(\Gamma\backslash G)$ has a spectral gap with the spectral gap data $(s_0(\Gamma), n_0(\Gamma))$.

Theorem 3.30. *There exist $\eta > 0$ (depending only on $s_0(\Gamma)$), and $\ell \in \mathbb{N}$ (depending only on $n_0(\Gamma)$) such that for any real-valued functions $\Psi_1, \Psi_2 \in C_c^\infty(\Gamma\backslash G)$ as $t \to +\infty$,*

$$e^{(n-1-\delta)t}\langle a_t\Psi_1, \Psi_2\rangle = \frac{m^{\mathrm{BR}}(\Psi_1)\cdot m^{\mathrm{BR}_*}(\Psi_2)}{|m^{\mathrm{BMS}}|} + O(e^{-\eta t}\mathcal{S}_\ell(\Psi_1)\mathcal{S}_\ell(\Psi_2)).$$

Proof. As in (3.26), we write

$$L^2(\Gamma\backslash G) = \mathcal{H}_\delta^\dagger \oplus \mathcal{W}$$

where $\mathcal{H}_\delta^\dagger = \oplus_{\upsilon\in\hat{M}} m(\upsilon)\mathcal{U}(\upsilon, (\delta-n+1)\alpha)$ with $m(\upsilon) \le \dim(\upsilon)^{n_0(\Gamma)}$, and no complementary series representation with parameter $s_0(\Gamma) < s$ is weakly contained in $\mathcal{W}$. For simplicity, set $s_0 := s_0(\Gamma)$ and $n_0 := n_0(\Gamma)$. we set $V = \mathcal{H}_\delta^\dagger$ and $V^\perp = \mathcal{W}$. Given $\Psi_1, \Psi_2 \in C_c^\infty(\Gamma\backslash G)$, we write $\Psi_i = \Psi_i' + \Psi_i^\perp$, where Ψ_i' and $\Psi_i^\perp$ are the projections of Ψ_i to $\mathcal{H}_\delta^\dagger$ and $\mathcal{W}$ respectively. Then by Proposition 3.29, there exist $\ell_0 \ge 1$ such that for any $\epsilon > 0$,

$$\langle a_t\Psi_1^\perp, \Psi_2^\perp\rangle = O(\mathcal{S}_{\ell_0}(\Psi_1)\mathcal{S}_{\ell_0}(\Psi_2)e^{(s_0-n+1+\epsilon)t}). \tag{3.31}$$

If $\delta = n-1$ and hence if $\mathcal{H}_\delta^\dagger = \mathbb{C}$, it is easy to see that (3.31) finishes the proof. Now suppose $\delta < n-1$.

For each $\upsilon \in \hat{M}$, the K-representation $\mathcal{U}_\upsilon(\delta-n+1)|_K$ is isomorphic to the induced representation $\mathrm{Ind}_M^K(\upsilon)$ and hence by the Frobenius reciprocity, the multiplicity of σ in $\mathcal{U}_\upsilon(s-n+1)|_K$ is equal to the multiplicity of υ in $\sigma|_M$, which is denoted by $[\sigma : \upsilon]$. Therefore as a K-module,

$$\mathcal{U}(\upsilon, \delta-n+1)|_K = \oplus_{\sigma\in\hat{K}} m_\upsilon(\sigma)\sigma$$

where $m_\upsilon(\sigma) \le [\sigma : \upsilon]$.

As a K-module, we write

$$\begin{aligned}\mathcal{H}_\delta^\dagger &= \oplus_{\upsilon\in\hat{M}} m(\upsilon)\mathcal{U}(\upsilon, (\delta-n+1)\alpha)\\ &= \oplus_{\upsilon\in\hat{M}} m(\upsilon)\left(\oplus_{\sigma\in\hat{K}} m_\upsilon(\sigma)\sigma\right)\\ &= \oplus_{\sigma\in\hat{K}} m(\sigma)\sigma\end{aligned}$$

where $m(\sigma) \le \sum_{\upsilon\in\hat{M}, \upsilon\subset\sigma} m(\upsilon)[\sigma : \upsilon]$. Note that $m(\sigma) \le d_\sigma^{n_0+1}$ for $n_0 = n_0(\Gamma)$.

For each $\sigma \in \hat{K}$, let Θ_σ be an orthonormal basis in the K-isotypic component, say, V_σ, of $\mathcal{H}_\delta^\dagger$, which is formed by taking the union of orthonormal bases of each irreducible component of V_σ. Then $\#\Theta_\sigma \le d_\sigma^{n_0+2}$.

By Theorem 3.23 and our discussion in section 3.2, there exist $\eta_0 > 0$ and $N \in \mathbb{N}$ such that for any $v_\sigma \in \Theta_\sigma$ and $v_\tau \in \Theta_\tau$, we have for all $t \gg 1$,

$$\langle a_t v_\sigma, v_\tau\rangle := c(v_\sigma, v_\tau)e^{(\delta-n+1)t} + O(d_\sigma^N d_\tau^N e^{(\delta-n+1-\eta_0)t}) \tag{3.32}$$

where $c(v_\sigma, v_\tau) = \langle \mathrm{c}_+(r_\delta)T_{(\delta-n+1)\alpha}v_\sigma, v_\tau\rangle$.

As $\Psi_i' = \sum_{\sigma\in\hat{K}}\sum_{v_\sigma\in\Theta_\sigma}\langle\Psi_i, v_\sigma\rangle v_\sigma$, we have for each $t \in \mathbb{R}$,

$$\langle a_t\Psi_1', \Psi_2'\rangle = \sum_{\sigma,\tau\in\hat{K}}\sum_{v_\sigma\in\Theta_\sigma, v_\tau\in\Theta_\tau} \langle\Psi_1, v_\sigma\rangle\cdot\overline{\langle\Psi_2, v_\tau\rangle}\cdot\langle a_t v_\sigma, v_\tau\rangle$$

where the convergence follows from the Cauchy-Schwartz inequality.

Therefore, by (3.32),

$$\begin{aligned}&\langle a_t\Psi_1', \Psi_2'\rangle\\ &= \left(\sum_{\sigma,\tau\in\hat{K}}\sum_{v_\sigma\in\Theta_\sigma, v_\tau\in\Theta_\tau}\langle\Psi_1, v_\sigma\rangle\overline{\langle\Psi_2, v_\tau\rangle}c(v_\sigma, v_\tau)\right)e^{(\delta-n+1)t}\\ &+ \sum_{\sigma,\tau\in\hat{K}}\sum_{v_\sigma\in\Theta_\sigma, v_\tau\in\Theta_\tau} d_\sigma^N d_\tau^N\langle\Psi_1, v_\sigma\rangle\overline{\langle\Psi_2, v_\tau\rangle}O(e^{(\delta-n+1-\eta_0)t})).\end{aligned}$$

Set

$$\mathcal{E}(\Psi_1,\Psi_2) := \left(\sum_{\sigma,\tau\in\hat{K}}\sum_{v_\sigma\in\Theta_\sigma, v_\tau\in\Theta_\tau}\langle\Psi_1, v_\sigma\rangle\overline{\langle\Psi_2, v_\tau\rangle}c(v_\sigma, v_\tau)\right).$$

By (3.3), there exists $\ell \geq \ell_0$ (depending only on n_0) such that

$$\sum_{\sigma,\tau\in\hat{K}} d_\sigma^{N+n_0+2} d_\tau^{N+n_0+2} c(\sigma)^{-\ell} c(\tau)^{-\ell} < \infty \tag{3.33}$$

where $c(\sigma)$ is as in (3.3). Since for any unit vector $v \in V_\sigma$,

$$|\langle\Psi, v\rangle| \ll c(\sigma)^{-\ell}\mathcal{S}_\ell(\Psi),$$

we now deduce that

$$\langle a_t\Psi_1', \Psi_2'\rangle = \mathcal{E}(\Psi_1,\Psi_2)e^{(\delta-n+1)t} + O(e^{(\delta-n+1-\eta_0)t}\mathcal{S}_\ell(\Psi_1)\mathcal{S}_\ell(\Psi_2)).$$

Hence, together with (3.31), it implies that there exists $\eta > 0$ such that

$$\langle a_t\Psi_1, \Psi_2\rangle = \mathcal{E}(\Psi_1,\Psi_2)e^{(\delta-n+1)t} + O(e^{(\delta-n+1-\eta)t}\mathcal{S}_\ell(\Psi_1)\mathcal{S}_\ell(\Psi_2)).$$

On the other hand, by Theorem 2.2,

$$\lim_{t\to\infty} e^{(n-1-\delta)t}\langle a_t\Psi_1, \Psi_2\rangle = \frac{m^{\mathrm{BR}}(\Psi_1)\cdot m^{\mathrm{BR}_*}(\Psi_2)}{|m^{\mathrm{BMS}}|}.$$

It follows that the infinite sum $\mathcal{E}(\Psi_1, \Psi_2)$ converges and

$$\mathcal{E}(\Psi_1,\Psi_2) = \frac{m^{\mathrm{BR}}(\Psi_1)\cdot m^{\mathrm{BR}_*}(\Psi_2)}{|m^{\mathrm{BMS}}|}.$$

This finishes the proof. □

As $\langle a_{-t}\Psi_1, \Psi_2\rangle = \langle a_t\Psi_2, \Psi_1\rangle$ for Ψ_i's real-valued, we deduce the following from Theorem 3.30:

Corollary 3.34. *There exist $\eta > 0$ and $\ell \in \mathbb{N}$ such that, as $t \to +\infty$,*

$$e^{(n-1-\delta)t}\langle a_{-t}\Psi_1, \Psi_2\rangle = \frac{m^{\mathrm{BR}_*}(\Psi_1)\cdot m^{\mathrm{BR}}(\Psi_2)}{|m^{\mathrm{BMS}}|} + O(e^{-\eta t}\mathcal{S}_\ell(\Psi_1)\mathcal{S}_\ell(\Psi_2)).$$

4. Non-wandering component of $\Gamma\backslash\Gamma H a_t$ as $t \to \infty$

4.1. **Basic setup.** Let H be either a symmetric subgroup or a horospherical subgroup of G. For the rest of the paper, we will set $K, M, A = \{a_t\}$ in each case as follows. If H is symmetric, that is, H is equal to the group of σ-fixed points for some involution σ of G, up to conjugation and commensurability, H is $\mathrm{SO}(k,1)\times\mathrm{SO}(n-k)$ for some $1 \le k \le n-1$. Let θ be a Cartan involution of G which commutes with σ and set K to be the maximal compact subgroup fixed by θ. Let $G = K \exp \mathfrak{p}$ be the Cartan decomposition and write $\mathfrak{g}$ as a direct sum of $d\sigma$ eigenspaces: $\mathfrak{g} = \mathfrak{h} \oplus \mathfrak{q}$ where $\mathfrak{h}$ is the Lie algebra of H and $\mathfrak{q}$ is the -1 eigenspace for $d\sigma$. Let $\mathfrak{a} \subset \mathfrak{p} \cap \mathfrak{q}$ be a maximal abelian subspace and set $A = \exp \mathfrak{a} = \{a_t := \exp(tY_0) : t \in \mathbb{R}\}$ where Y_0 is a norm one element in $\mathfrak{a}$ with respect to the Riemannian metric induced by $\langle , \rangle$ defined in (3.1). Let M be the centralizer of A in K.

If H is a horospherical subgroup of G, we let $A = \{a_t\}$ be a one-parameter subgroup of diagonalizable elements so that H is the expanding horospherical subgroup for a_t. Letting M be the maximal compact subgroup in the centralizer of A, we may assume that the right translation action of a_t corresponds to the geodesic flow on $\mathrm{T}^1(\mathbb{H}^n) = G/M$. Let K be the stabilizer of the base point of the vector in $\mathrm{T}^1(\mathbb{H}^n)$ corresponding to M.

In both cases, let $o \in \mathbb{H}^n$ and $X_0 \in \mathrm{T}^1(\mathbb{H}^n)$ be points stabilized by K and M respectively. Let N^+ and N^- be the expanding and contracting horospherical subgroups of G with respect to a_t, respectively.

4.2. **Measures on gH constructed from conformal densities.** Set $P := MAN^-$, which is the stabilizer of X_0^+. Via the visual map $g \mapsto g^+$, we have $G/P \simeq \partial(\mathbb{H}^n)$. Since $G/P \simeq K/M$, we may consider the visual map as a map from G to K/M. In both cases, the restriction of the visual map to H induces a diffeomorphism from $H/H \cap M$ to its image inside K/M.

Letting $\{\mu_x : x \in G/K\}$ be a Γ-invariant conformal density of dimension δ_μ, we will define an $H\cap M$-invariant measure $\tilde{\mu}_{gH}$ on each $g \in G/H$. Setting $\bar{H} = H/(H \cap M)$, first consider the measure on $g\bar{H}$:

$$d\tilde{\mu}_{g\bar{H}}(gh) = e^{\delta_\mu \beta_{(gh)^+}(o,gh)} d\mu_o((gh)^+).$$

We denote by $\tilde{\mu}_{gH}$ the $H \cap M$-invariant extension of this measure on gH, that is, for $f \in C_c(gH)$,

$$\int f(gh)\, d\tilde{\mu}_{gH}(gh) = \int_{g\bar{H}} \int_{H\cap M} f(ghm) d_{H\cap M}(m) d\tilde{\mu}_{g\bar{H}}(gh)$$

where $d_{H\cap M}(m)$ is the probability Haar measure on $H \cap M$.

By the Γ-invariant conformality of $\{\mu_x\}$, this definition is independent of $o \in \mathbb{H}^n$ and $\tilde{\mu}_{gH}$ is invariant under Γ and hence if $\Gamma\backslash\Gamma gH$ is closed, $\tilde{\mu}_{gH}$ induces a locally finite Borel measure μ_{gH} on $\Gamma\backslash\Gamma gH$.

Recall the Lebesgue density $\{m_x\}$ of dimension $n-1$ and the Patterson-Sullivan density $\{\nu_x\}$ of dimension δ. We normalize them so that $|m_o| =$

$|\nu_o| = 1$. We set

$$\tilde{\mu}_{gH}^{\mathrm{Haar}} = \tilde{m}_{gH} \quad \text{and} \quad \tilde{\mu}_{gH}^{\mathrm{PS}} = \tilde{\nu}_{gH},$$

and for a closed orbit $\Gamma\backslash\Gamma gH$, we denote by μ_{gH}^{Haar} and μ_{gH}^{PS} the measures on $\Gamma\backslash\Gamma gH$ induced by them respectively.

Lemma 4.1. *For each $g \in G$, $d\tilde{\mu}_{gH}^{\mathrm{Haar}}(gh) = d\tilde{\mu}_H^{\mathrm{Haar}}(h)$ and $dh := d\tilde{\mu}_H^{\mathrm{Haar}}(h)$ is a Haar measure on H.*

Proof. As m_o is G-invariant, we have

$$dm_o((gh)^+) = dm_{g^{-1}(o)}(h^+) = e^{(n-1)\beta_{h^+}(o,g^{-1}(o))} dm_o(h^+).$$

Since

$$\beta_{h^+}(o, g^{-1}(o)) + \beta_{(gh)^+}(o, gh) = \beta_{h^+}(o, g^{-1}(o)) + \beta_{h^+}(g^{-1}(o), h) = \beta_{h^+}(o, h),$$

we have

$$d\tilde{\mu}_{gH}^{\mathrm{Haar}}(gh) = e^{(n-1)\beta_{(gh)^+}(o,gh)} d\mu_o((gh)^+) = e^{(n-1)\beta_{h^+}(o,h)} d\mu_o(h^+) = d\tilde{\mu}_H^{\mathrm{Haar}}(h)$$

proving the first claim. The first claim shows that $d\tilde{\mu}_{\bar{H}}$ is left H-invariant. Since $d_{H\cap M}$ is an $H\cap M$-invariant measure, the product measure $d\tilde{\mu}_H(hm)^{\mathrm{Haar}} = d\tilde{\mu}_{\bar{H}}(h) d_{H\cap M}(m)$ is a Haar measure of H. □

4.3. Let Γ be a torsion-free, non-elementary, geometrically finite subgroup of G. For any given compact subset Ω of $\Gamma\backslash G$, the goal of the rest of this section is to describe the set

$$\{h \in \Gamma\backslash\Gamma H : ha_t \in \Omega \text{ for some } t > 0\}.$$

For H horospherical, this turns out to be a compact subset. For H symmetric, we will obtain a thick-thin decomposition, and give estimates of the size of thin parts with respect to the measures μ_H^{PS} and μ_H^{Haar} (Theorem 4.16).

An element $\gamma \in \Gamma$ is called *parabolic* if there exists a unique fixed point of γ in $\partial(\mathbb{H}^n)$, and an element $\xi \in \Lambda(\Gamma)$ is called a parabolic fixed point if it is fixed by a parabolic element of Γ. Let $\Lambda_p(\Gamma) \subset \Lambda(\Gamma)$ denote the set of all parabolic fixed points of Γ. Since Γ is geometrically finite, each parabolic fixed point ξ is bounded, i.e., $\mathrm{Stab}_\Gamma(\xi)$ acts co-compactly on $\Lambda(\Gamma) - \{\xi\}$. Recall the notation $g^+ = g(X_0)^+$ and $g^- = g(X_0)^-$.

Consider the upper half space model for $\mathbb{H}^n$: $\mathbb{H}^n = \{(x, y) : x \in \mathbb{R}^{n-1}, y \in \mathbb{R}_{>0}\}$. We set $\mathbb{R}^n_+ := \{(x, y) : x \in \mathbb{R}^{n-1}, y \in \mathbb{R}_{>0}\}$, so that $\partial(\mathbb{R}^n_+) = \{(x, 0) : x \in \mathbb{R}^{n-1}\}$. Suppose that ∞ is a parabolic fixed point for Γ. Set $\Gamma_\infty := \mathrm{Stab}_\Gamma(\infty)$ and let $\Gamma'(\infty)$ be a normal abelian subgroup of Γ_∞ which is of finite index; this exists by a theorem of Biberbach. Let L be a minimal Γ_∞-invariant subspace in $\mathbb{R}^{n-1}$. By Prop. 2.2.6 in [10], $\Gamma'(\infty)$ acts as translations and cocompactly on L. We note that L may not be unique, but any two such are Euclidean-parallel.

The notation d_{Euc} and $\|\cdot\|$ denote the Euclidean distance and the Euclidean norm in $\mathbb{R}^n$ respectively. Following Bowditch [10], we write for each $r > 0$:

$$C(L, r) := \{x \in \mathbb{R}^n_+ \cup \partial(\mathbb{R}^n_+) : d_{\mathrm{Euc}}(x, L) \geq r\}. \tag{4.2}$$

Each $C(L,r)$ is Γ_∞-invariant and called a standard parabolic region (or an extended horoball) associated to $\xi = \infty$.

Theorem 4.3. [10, Prop. 4.4] *For any $\epsilon_0 > 0$, there exists $r_0 > 0$ such that for any $r \geq r_0$,*

(1) *$\gamma C(L,r) = C(L,r)$ if $\gamma \in \Gamma_\infty$;*
(2) *if $\gamma \in \Gamma - \Gamma_\infty$, $d_{Euc}(C(L,r), \gamma C(L,r)) > \epsilon_0$.*

Corollary 4.4. *Suppose that ∞ is a bounded parabolic fixed point for Γ. Then for any sufficiently large r, the natural projection map*

$$\Gamma_\infty \backslash (C(L,r) \cap \mathbb{H}^n) \to \Gamma \backslash \mathbb{H}^n$$

is injective and proper.

Proof. We fix $\epsilon_0 > 0$, and let $r_0 > 0$ be as in the above theorem 4.3. Let $r > r_0$, and set $C_\infty = C(L,r) \cap \mathbb{H}^n$ for simplicity. The injectivity is immediate from Theorem 4.3(2). Since C_∞ is closed in $\mathbb{H}^n$, so is γC_∞ for all $\gamma \in \Gamma$. Hence to prove the properness of the map, it is sufficient to show that if F is a compact subset of $\mathbb{H}^n$, then F intersects at most finitely many distinct γC_∞'s. Now suppose there exists an infinite sequence $\{\gamma_i \in \Gamma\}$ such that $\gamma_i \Gamma_\infty$'s are all distinct from each other and $F \cap \gamma_i C_\infty \neq \emptyset$. Choosing $y_i \in F \cap \gamma_i C_\infty$, by Theorem 4.3(2), we have $d(y_i, y_j) \geq \epsilon_0$ for all $i \neq j$, which contradicts the assumption that F is compact. □

4.4. H horospherical.

Theorem 4.5. *Let $H = N$ be a horospherical subgroup. Suppose that $\Gamma \backslash \Gamma NM$ is closed in $\Gamma \backslash G$. For any compact subset Ω of $\Gamma \backslash G$, the set $\Gamma \backslash \Gamma NM \cap \Omega A$ is relatively compact.*

Proof. We may assume without loss of generality that the horosphere NK/K in $G/K \simeq \mathbb{H}^n$ is based at ∞. Note that $\Gamma_\infty \subset NM$ and that the closedness of $\Gamma \backslash \Gamma NM$ implies that $\Gamma_\infty \backslash NM \to \Gamma \backslash G$ is a proper map.

Therefore, if the claim does not hold, there exists a sequence $n_i \in NM$ which is unbounded modulo Γ_∞ such that $\gamma_i n_i a_{t_i} \to x$ for some $t_i \in \mathbb{R}$, $\gamma_i \in \Gamma$ and $x \in G$.

It follows that, passing to a subsequence, $n_i a_{t_i}(o) \to \infty$ and $d(n_i a_{t_i}, \gamma_i^{-1} x) \to 0$ as $i \to \infty$. Therefore $\gamma_i^{-1} x(o) \to \infty$ and hence $\infty \in \Lambda(\Gamma)$.

Since the image of the horosphere $N(o)$ in $\Gamma \backslash \mathbb{H}^n = \Gamma \backslash G/K$ is closed, it follows that ∞ is a bounded parabolic fixed point for Γ by [13]. Therefore Γ_∞ acts co-compactly on an r neighborhood of a minimal Γ_∞-invariant subspace L in $\partial(\mathbb{H}^n) - \{\infty\} = \mathbb{R}^{n-1}$ for some $r > 0$. Write $n_i a_{t_i}(o) = (x_i, y_i) \in \mathbb{R}^{n-1} \times \mathbb{R}_{>0}$. Since $\{n_i\}$ is unbounded modulo Γ_∞, after passing to a subsequence if necessary, we have $d_{\mathrm{Euc}}(x_i, L) \to \infty$. It follows that for any r, $(x_i, y_i) \in C(L,r)$ for all large i. Since n_i is unbounded modulo Γ_∞, we get $n_i a_{t_i} = (x_i, y_i)$ is unbounded in $\Gamma_\infty \backslash C(L,r)$. Thus by Corollary 4.4, $n_i a_{t_i}$ must be unbounded modulo Γ, which is a contradiction. □

4.5. H **symmetric.** We now consider the case when H is symmetric. Then $H(o) = H/H \cap K$ is a totally geodesic submanifold in $G(o) = G/K = \mathbb{H}^n$. We denote by π the canonical projection from G to $G/K = \mathbb{H}^n$. We set $\tilde{S} = H(o)$.

Fixing a compact subset Ω of $\Gamma\backslash G$, define

$$H_\Omega := \{h \in H : \Gamma\backslash\Gamma h a_t \in \Omega \text{ for some } t > 0\}$$

and set $\tilde{S}_\Omega = H_\Omega(o)$.

Lemma 4.6. *Let $\xi \in \partial(\tilde{S})$. If $\xi \notin \Lambda(\Gamma)$, then there exists a neighborhood U of ξ in $\overline{\mathbb{H}^n}$ such that $U \cap \tilde{S}_\Omega = \emptyset$.*

Proof. Let Ω_0 be a compact subset of G such that $\Omega = \Gamma\backslash\Gamma\Omega_0$. If the claim does not hold, then there exist $h_n \in H$, $\gamma_n \in \Gamma$ and $t_n > 0$ such that $\gamma_n h_n a_{t_n} \in \Omega_0$ and $h_n(o) \to \xi$. Note that $\{h_n a_t(o) : t > 0\}$ denotes the (half) geodesic emanating from $\pi(h_n)$ and orthogonal to $\tilde{S}$. Since $h_n(o)$ converges to $\xi \in \partial\mathbb{H}^n$, it follows that $h_n a_{t_n}(o) \to \xi$.

Now since Ω_0 is compact, by passing to a subsequence in necessary, we may assume $\gamma_n h_n a_{t_n} \to x$. As G acts by isometries on $\mathbb{H}^n$, we get $\gamma_n^{-1}(x(o)) \to \xi$. This implies $\xi \in \Lambda(\Gamma)$ which is a contradiction. □

Fix a Dirichlet domain $\mathcal{D}$ for $H \cap \Gamma$ in $\tilde{S}$ and set

$$\mathcal{D}_\Omega = \mathcal{D} \cap \tilde{S}_\Omega. \tag{4.7}$$

Corollary 4.8. *Assume that the orbit $\Gamma\backslash\Gamma H$ is closed in $\Gamma\backslash G$. There exist a compact subset Y_0 of $\mathcal{D}$ and a finite subset $\{\xi_1, \dots \xi_m\} \subset \Lambda_p(\Gamma) \cap \partial(\tilde{S})$ such that*

$$\mathcal{D}_\Omega \subset Y_0 \cup (\cup_{i=1}^m U(\xi_i))$$

where $U(\xi_i)$ is a neighborhood of ξ_i in $\overline{\mathbb{H}^n}$. In particular if $\Lambda_p(\Gamma) \cap \partial(\tilde{S}) = \emptyset$, then $\mathcal{D}_\Omega$ is relatively compact.

Proof. For each $\xi \in \partial(\tilde{S}) \cap \partial(\mathcal{D})$, let $U(\xi)$ be a neighborhood of ξ in $\overline{\mathbb{H}^n}$. When $\xi \notin \Lambda(\Gamma)$, we may assume by Lemma 4.6 that $U(\xi) \cap \tilde{S}_\Omega = \emptyset$. By the compactness of $\partial(\tilde{S}) \cap \partial(\mathcal{D})$, there exists a finite covering $\cup_{i\in I} U(\xi_i)$. Set $Y_0 := \mathcal{D} - \cup_{i\in I} U(\xi_i)$, which is a compact subset. Now $\mathcal{D}_\Omega - Y_0 \subset \cup_{i\in I, \xi_i \in \Lambda(\Gamma)} U(\xi_i)$ by the choice of $U(\xi_i)$'s. On the other hand, by [52, Proposition 5.1], we have $\Lambda(\Gamma) \cap \partial\mathcal{D} \subset \Lambda_p(\Gamma)$. Hence the claim follows. □

In the rest of this subsection, we fix $\xi \in \Lambda_p(\Gamma) \cap \partial(\tilde{S})$, and investigate $\mathcal{D}_\Omega \cap U(\xi)$. We consider the upper half space model for $\mathbb{H}^n$ and assume that $\xi = \infty$. In particular, $\tilde{S}$ is a vertical plane. Let Γ_∞, $\Gamma'(\infty)$, L and $C(L, r)$ be as in the subsection 4.3. Without loss of generality, we assume $0 \in L$. We consider the orthogonal decomposition $\mathbb{R}^{n-1} = L \oplus L^\perp$ and let $P_{L^\perp} : \mathbb{R}^{n-1} \to L^\perp$ denote the orthogonal projection map.

Lemma 4.9. *There exists $R_0 > 0$ such that for any $h \in H_\Omega$, we have $\|P_{L^\perp}(h^+)\| \le R_0$.*

Proof. Let Ω_0 be a compact subset of G such that $\Omega = \Gamma\backslash\Gamma\Omega_0$. Then by Corollary 4.4, and part (1) of Theorem 4.3, there exists $R_0' > 0$ depending on Ω, such that

$$\text{if } x \in C(L, R_0') \cap \mathbb{H}^n, \text{ then } x \notin \Gamma\Omega_0. \tag{4.10}$$

Suppose now that $h \in H_\Omega$, thus $ha_{t_0}(o) \in \Gamma\Omega_0$ for some $t_0 > 0$. This, in view of (4.10) and the definition of $C(L, R_0')$, implies $d_{\mathrm{Euc}}(ha_{t_0}(o), L) < R_0'$.

As discussed above, $\{ha_t : t > 0\}$ is the geodesic ray emanating from $h(o)$ and orthogonal to $\tilde{S}$ i.e. a Euclidean semicircle orthogonal to the vertical plane. Hence there exists some absolute constant s_0 such that

$$\lim_{t\to\infty} d_{\mathrm{Euc}}(ha_t(o), L) \le d_{\mathrm{Euc}}(ha_{t_0}(o), L) + s_0 \le R_0' + s_0,$$

which implies $\|P_{L^\perp}(h^+)\| \le R_0 := R_0' + s_0$, as we wanted to show. □

For $N \ge 1$, set

$$U_N(\infty) := \{x \in \mathbb{R}^n_+ \cup \partial(\mathbb{R}^n_+) : \|x\|_{\mathrm{Euc}} > N\}. \tag{4.11}$$

Let $\Delta := \Gamma'(\infty) \cap H$ and let p be the difference of the rank of $\Gamma'(\infty)$ and the rank of Δ. Suppose $p \ge 1$. Let $\gamma = (\gamma_1, \dots, \gamma_p)$ be a p-tuple of elements of Γ' such that the subgroup generated by $\gamma \cup \Delta$ has finite index in Γ'. For $\mathbf{k} = (k_1, \cdots, k_p) \in \mathbb{Z}^p$, we write $\gamma^{\mathbf{k}} = \gamma_1^{k_1} \cdots \gamma_p^{k_p}$. The notation $|\mathbf{k}|$ means the maximum norm of $(k_1, \cdots, k_p)$.

The following gives a description of cuspidal neighborhoods of $\mathcal{D}_\Omega$:

Theorem 4.12. *There exist $c_0 \ge 1$ and a compact subset $\mathcal{F}$ of $\mathbb{R}^{n-1}$ such that for all large $N \gg 1$,*

$$\{h^+ \in \mathbb{R}^{n-1} : h \in H, \pi(h) \in \mathcal{D}_\Omega \cap U_{c_0 N}(\infty)\} \subset \cup_{|\mathbf{k}|\ge N} \Delta\gamma^{\mathbf{k}}\mathcal{F}.$$

Proof. In [52, Prop. 5.8], it is shown that for some $c_0 \ge 1$ and a compact subset $\mathcal{F}$ of $\mathbb{R}^{n-1}$,

$$\{h^+ \in \Lambda(\Gamma) : h \in H, \pi(h) \in \mathcal{D} \cap U_{c_0 N}(\infty)\} \subset \cup_{|\mathbf{k}|\ge N} \Delta\gamma^{\mathbf{k}}\mathcal{F} \tag{4.13}$$

for all large $N \gg 1$. However the only property of $h^+ \in \Lambda(\Gamma)$ used in this proof is the fact that $\sup_{h\in H, h^+\in\Lambda(\Gamma)} \|P_{L^\perp}(h^+)\| < \infty$. Since this property holds for the set in concern by Lemma 4.9, the proof of Proposition 5.8 of [52] can be used. □

4.6. Estimates on the size of thin part. For $\xi \in \partial(\mathbb{H}^n)$, let $U_N(\xi)$ be defined to be $g(U_N(\infty))$ where $g \in G$ is such that $\xi = g(\infty)$ and $U_N(\infty)$ is defined as in (4.11).

Proposition 4.14. *Let $\xi \in \partial(\tilde{S}) \cap \Lambda_p(\Gamma)$ and $p_\xi := \mathrm{rank}(\Gamma_\xi) - \mathrm{rank}(\Gamma_\xi \cap H)$. For all $N \gg 1$, we have*

$$\tilde{\mu}_H^{\mathrm{PS}}\{h \in H : \pi(h) \in \mathcal{D}_\Omega \cap U_N(\xi)\} \ll N^{-\delta+p_\xi};$$
$$\tilde{\mu}_H^{\mathrm{Haar}}\{h \in H : \pi(h) \in \mathcal{D}_\Omega \cap U_N(\xi)\} \ll N^{-n+1+p_\xi}$$

with the implied constants independent of N.

Proof. The first claim is shown in [52, Proposition 5.2]. Without loss of generality, we may assume that $\xi = \infty$. By replacing δ by $n-1$ and ν_o by m_o in the proof of [52, Proposition 5.2], we get

$$\int_{h^+ \in \gamma^{\mathbf{k}} \mathcal{F}} e^{(n-1)\beta_{h^+}(o,h)} dm_o(h^+) \asymp |\mathbf{k}|^{-n+1}$$

where the notation γ, $\mathbf{k}$ and $\mathcal{F}$ are as in Theorem 4.12 and $f(\mathbf{k}) \asymp g(\mathbf{k})$ means that the ratio of $f(\mathbf{k})$ and $g(\mathbf{k})$ lies in between two bounded constants independent of $\mathbf{k}$.

Hence by Proposition 4.12,

$$\tilde{\mu}_H^{\mathrm{Haar}}\{h \in H : \pi(h) \in \mathcal{D}_\Omega \cap U_N(\infty)\} \ll \sum_{\mathbf{k} \in \mathbb{Z}^{p_\infty}, |\mathbf{k}| \geq N} |\mathbf{k}|^{-n+1} \ll N^{-n+1+p_\infty}.$$

□

Recall the notion of the parabolic-corank of Γ with respect to H, introduced in [52]:

$$\text{Pb-corank}_H(\Gamma) := \max_{\xi \in \Lambda_p(\Gamma) \cap \partial(\tilde{S})} (\mathrm{rank}(\Gamma_\xi) - \mathrm{rank}(\Gamma_\xi \cap H)).$$

The following is shown in [52, Thm. 1.14]:

Proposition 4.15. *We have*

- $\text{Pb-corank}_H(\Gamma) = 0$ *if and only if the support of* μ_H^{PS} *is compact;*
- $\text{Pb-corank}_H(\Gamma) < \delta$ *if and only if* μ_H^{PS} *is finite.*

It is also shown in [52, Lem. 6.2] that $\text{Pb-corank}_H(\Gamma)$ is bounded above by $n - \dim(H/(H \cap K))$. Therefore if H is locally isomorphic to $\mathrm{SO}(k,1) \times \mathrm{SO}(n-k)$ and $\delta > n-k$, then μ_H^{PS} is finite.

For $h \in \Gamma\backslash G$, we denote by r_h the injectivity radius, that is, the map $g \mapsto hg$ is injective on the set $d(g,e) \leq r_h$. By Corollary 4.8, (4.13), Proposition 4.14, and by the structure of the support of μ_H^{PS} obtained in [52], we have the following:

Theorem 4.16. *Suppose that* $\Gamma\backslash\Gamma H$ *is closed. For any compact subset* Ω *of* $\Gamma\backslash G$*, there exists an open subset* $Y_\Omega \subset \Gamma\backslash\Gamma H$ *containing the union* $\mathrm{supp}(\mu_H^{\mathrm{PS}}) \cup \{h \in \Gamma\backslash\Gamma H : ha_t \in \Omega \text{ for some } t > 0\}$ *and satisfying the following properties:*

1. *if* $\text{Pb-corank}_H(\Gamma) = 0$*,* Y_Ω *is relatively compact;*
2. *if* $\text{Pb-corank}_H(\Gamma) \geq 1$*, then the following hold:*
   - (a) $Y_\epsilon := \{h \in Y_\Omega : r_h > \epsilon\}$ *is relatively compact;*
   - (b) *there exist* $\xi_1, \cdots, \xi_m \in \Lambda_p(\Gamma) \cap \partial(\tilde{S})$ *and* $c_1 > 0$ *such that for all small* $\epsilon > 0$*,* $Y_\Omega - Y_\epsilon \subset \cup_{i=1}^m U_{c_1\epsilon^{-1}}(\xi_i)$;
   - (c) *for all small* $\epsilon > 0$*,*

$$\mu_H^{\mathrm{PS}}(Y_\Omega - Y_\epsilon) \ll \epsilon^{\delta - p_0} \quad \text{and} \quad \mu_H^{\mathrm{Haar}}(Y_\Omega - Y_\epsilon) \ll \epsilon^{n-1-p_0}$$

for $p_0 := \text{Pb-corank}_H(\Gamma)$.

5. Translates of a compact piece of $\Gamma\backslash\Gamma H$ via thickening

Let Γ be a non-elementary geometrically finite subgroup of G. Let H be either symmetric or horospherical, and let $A = \{a_t\}$, M, K, $N^{\pm}$, o, X_0 be as in the subsection 4.1.

5.1. Decomposition of measures. Set $P := MAN^-$, which is the stabilizer of X_0^+. The measure

$$dn_0 = e^{(n-1)\beta_{n_0^-}(o,n_0)} dm_o(n_0^-)$$

can be seen to be a Haar measure on N^- by a similar argument as in Lemma 4.1. Then for $p = n_0 a_t m \in N^- AM$,

$$dp := dn_0 dt dm$$

is a right invariant measure on P where dm is the probability Haar measure of M and dt is the Lebesgue measure on $\mathbb{R}$.

For $g \in G$, consider the measure on gP given by

$$d\nu_{gP}(gp) = e^{\delta t} d\nu_o((gp)^-) dt \ \text{ for } \ t = \beta_{(gp)^{-1}}(o, gp). \tag{5.1}$$

For $\Psi \in C_c(G)$, we have:

$$\tilde{m}^{\mathrm{Haar}}(\Psi) = \int_{gP} \int_N \Psi(gpn) dn \, dp; \tag{5.2}$$

$$\tilde{m}^{\mathrm{BR}}(\Psi) = \int_{gP} \int_N \Psi(gpn) d\tilde{\mu}^{\mathrm{Haar}}_{gpN}(gpn) d\nu_{gP}(gp); \tag{5.3}$$

$$\tilde{m}^{\mathrm{BMS}}(\Psi) = \int_{gP} \int_N \Psi(gpn) d\tilde{\mu}^{\mathrm{PS}}_{gpN}(gpn) d\nu_{gP}(gp). \tag{5.4}$$

5.2. Approximations of Ψ. We fix a left invariant metric d on G, which is right $H \cap M$-invariant and which descends to the hyperbolic metric on $\mathbb{H}^n = G/K$. For a subset S of G and $\epsilon > 0$, S_ϵ denotes the ϵ-neighborhood of e in S: $S_\epsilon = \{g \in S : d(g, e) \le \epsilon\}$.

We fix a compact subset Ω of $\Gamma\backslash G$. Let $r_0 := r_\Omega$ denote the infimum of the injectivity radius over all $x \in \Omega$. That is, for all $x \in \Omega$, the map $g \mapsto xg$ in injective on the set $\{g \in G : d(g, e) < r_\Omega\}$.

We fix a function $\kappa_\Omega \in C^\infty(\Gamma\backslash G)$ such that $0 \le \kappa_\Omega \le 1$, $\kappa_\Omega(x) = 1$ for all x in the $\frac{r_0}{2}$-neighborhood of Ω and and $\kappa_\Omega(x) = 0$ for x outside the r_0-neighborhood of Ω.

Fix $\Psi \in C^\infty(\Omega)$. For all small $\epsilon > 0$, set

$$\Psi_\epsilon^+(x) = \sup_{g \in G_\epsilon} \Psi(xg) \quad \text{and} \quad \Psi_\epsilon^-(x) = \inf_{g \in G_\epsilon} \Psi(xg). \tag{5.5}$$

For each $0 < \epsilon \le r_\Omega$, $x \in \Gamma\backslash G$ and $g \in G_\epsilon$, we have

$$\Psi_\epsilon^-(x) \le \Psi(xg) \le \Psi_\epsilon^+(x) \tag{5.6}$$

and

$$|\Psi_\epsilon^\pm(x) - \Psi(x)| \le c_1 \epsilon \mathcal{S}_{\infty,1}(\Psi) \kappa_\Omega(x)$$

for some absolute constant $c_1 > 0$.

For $\bullet = \mathrm{Haar}, \mathrm{BR}, \mathrm{BR}_*$ or BMS, we define

$$A_\Psi^\bullet = \mathcal{S}_{\infty,1}(\Psi) \cdot m^\bullet(\mathrm{supp}(\Psi)).$$

Define for each $g \in G$,

$$\phi_0(g) = |\nu_{g(o)}|.$$

Then ϕ_0 is left Γ-invariant and right K-invariant, and hence induces a smooth function in $C^\infty(\Gamma\backslash G)^K = C^\infty(\mathbb{H}^n)$. Moreover ϕ_0 is an eigenfunction of the Laplacian with eigenvalue $\delta(n-1-\delta)$ [63].

Lemma 5.7. *For a compact subset Ω of $\Gamma\backslash G$,*

(1) $m^{\mathrm{BR}}(\Omega) \ll \sup_{x\in\Omega} \phi_0(x) \cdot m^{\mathrm{Haar}}(\Omega K)$;
(2) $m^{\mathrm{BR}_*}(\Omega) \ll \sup_{x\in\Omega} \phi_0(x) \cdot m^{\mathrm{Haar}}(\Omega K)$;
(3) $m^{\mathrm{BMS}}(\Omega) \ll \sup_{x\in\Omega} \phi_0(x)^2 \cdot m^{\mathrm{Haar}}(\Omega K)$.

Proof. The first two claims follow since for any K-invariant function ψ in $\Gamma\backslash G$, $m^{\mathrm{BR}_*}(\psi) = m^{\mathrm{BR}}(\psi) = \int_{\Gamma\backslash G} \psi(g)\phi_0(g) dm^{\mathrm{Haar}}(g)$. The third one follows from the smearing argument of Sullivan, see [63, Proof of Prop. 5]. □

On the other hand, there exists $\ell \in \mathbb{N}$ such that for all $\Psi \in C^\infty(\Omega)$, $\mathcal{S}_{\infty,1}(\Psi) \ll \mathcal{S}_\ell(\Psi)$ [1]. Hence it follows from Lemma 5.7 that there exists $\ell \in \mathbb{N}$ such that for all $\Psi \in C^\infty(\Omega)$, any $\bullet = \mathrm{Haar}, \mathrm{BR}, \mathrm{BR}_*$ or BMS, and any $0 < \epsilon < r_\Omega$,

$$A_\Psi^\bullet \ll \mathcal{S}_{\infty,1}(\Psi) \cdot m^{\mathrm{Haar}}(\mathrm{supp}(\Psi)) \ll \mathcal{S}_\ell(\Psi) \quad \text{and} \quad \mathcal{S}_\ell(\Psi_\epsilon^\pm) \ll \mathcal{S}_\ell(\Psi) \tag{5.8}$$

where the implied constants depend only on Ω.

5.3. Thickening of a compact piece of yH. For the rest of this section, fix $y \in \Gamma\backslash G$ and $H_0 \subset H$ be a compact subset such that the map $h \mapsto yh$ is injective on H_0. Fix $0 < \epsilon_0 < r_\Omega$ which is smaller than the injectivity radius of yH_0.

Fix non-negative functions $\Psi \in C^\infty(\Omega)$ and $\phi \in C^\infty(yH_0)$. Let $M' \subset M$ be a smooth cross section for $H \cap M$ in M and set $P' := M'AN^-$. As $hp = h'p'$ implies $h = h'm$ and $p = m^{-1}p'$ for $m \in H \cap M$, it follows that the product map $H \times P' \to G$ is a diffeomorphism onto its image, which is a Zariski open neighborhood of e. Let dp' be a smooth measure on P' such that $dp = d_{H\cap M} m dp'$ for $p = mp'$. For $0 < \epsilon < \epsilon_0$, let $\rho_\epsilon \in C^\infty(P'_\epsilon)$ be a non-negative function such that $\int \rho_\epsilon dp' = 1$, and we define $\Phi_\epsilon \in C_c^\infty(\Gamma\backslash G)$ by

$$\Phi_\epsilon(g) = \begin{cases} \phi(yh)\rho_\epsilon(p) & \text{if } g = yhp \in yH_0P'_\epsilon \\ 0 & \text{otherwise.} \end{cases} \tag{5.9}$$

Lemma 5.10. *For all $0 < \epsilon < \epsilon_0$ and $t > 0$,*

$$\int_{\Gamma\backslash G} \Psi_\epsilon^-(ga_t)\Phi_\epsilon(g)dg \le \int_{h\in H_0} \Psi(yha_t)\phi(yh)dh \le \int_{\Gamma\backslash G} \Psi_\epsilon^+(ga_t)\Phi_\epsilon(g)dg.$$

Proof. For all $p \in P'_\epsilon$, $h \in H_0$ and $t > 0$, $yhpa_t = yha_t(a_{-t}pa_t) \in yha_tP_\epsilon$ and hence

$$\int_{h\in H_0} \Psi(yha_t)\phi(yh)dh \le \int_{h\in H_0} \Psi^+_\epsilon(yhpa_t)\phi(yh)dh.$$

Integrating against ρ_ϵ, we have

$$\begin{aligned}&\int_{h\in H_0} \Psi(yha_t)\phi(yh)dh\\ \le &\int_{yhp\in yH_0P'_\epsilon} \Psi^+_\epsilon(yhpa_t)\phi(yh)\rho_\epsilon(p)dhdp\\ = &\int_{\Gamma\backslash G} \Psi^+_\epsilon(ga_t)\Phi_\epsilon(g)dg.\end{aligned}$$

The other direction is proved similarly. □

Lemma 5.11. *For all $0 < \epsilon < \epsilon_0$,*

$$m^{\mathrm{BR}_*}(\Phi_\epsilon) = (1 + O(\epsilon))\mu^{\mathrm{PS}}_{yH}(\phi).$$

Proof. Choose $g_y \in G$ so that $y = \Gamma\backslash\Gamma g_y$ and set $\tilde\phi(g_yh) := \phi(yh)$ and $\tilde\Phi_\epsilon(g_yhp) := \tilde\phi(g_yh)\rho_\epsilon(p)$ for $hp \in H_0P'_\epsilon$ and zero otherwise. As $0 < \epsilon < \epsilon_0$, we have $m^{\mathrm{BR}_*}(\Phi_\epsilon) = \tilde m^{\mathrm{BR}_*}(\tilde\Phi_\epsilon)$ and $\mu^{\mathrm{PS}}_{yH}(\phi) = \tilde\mu^{\mathrm{PS}}_{yH}(\tilde\phi)$. By the definition, we have

$$\tilde m^{\mathrm{BR}_*}(\tilde\Phi_\epsilon) = \int_{g\in G/M}\int_M \tilde\Phi_\epsilon(gm)dm\; e^{\delta\beta_{g^+}(o,g)}\; e^{(n-1)\beta_{g^-}(o,g)}\; d\nu_o(g^+)dm_o(g^-)ds$$

where $s = \beta_{g^-}(o, g)$. For simplicity, we set $g_y = y \in G$ by abuse of notation. For $g = yhp \in H_0P'_\epsilon$, as $|\beta_{g^+}(yh, g)| \le d(yh, yhp) = d(e, p) \le \epsilon$, we have $e^{\delta\beta_{g^+}(yh,g)} = 1 + O(\epsilon)$. Since $g^+ = (yh)^+$, we have

$$e^{\delta\beta_{g^+}(o,g)}d\nu_o(g^+) = (1+O(\epsilon))e^{\delta\beta_{(yh)^+}(o,yh)}d\nu_o((yh)^+) = (1+O(\epsilon))d\tilde\mu^{\mathrm{PS}}_{yH}(yh).$$

On the other hand, as $\{m_x\}$ is G-invariant,

$$dm_o(g^-) = dm_{(yh)^{-1}(o)}(p^-) = e^{(n-1)\beta_{p^-}(o,(yh)^{-1}(o))}dm_o(p^-).$$

Since $p^- = n_0^-$ for $p = n_0a_tm$, we have

$$\begin{aligned}&\beta_{g^-}(o, g) + \beta_{p^-}(o, (yh)^{-1}(o))\\ &= \beta_{p^-}((yh)^{-1}(o), p) + \beta_{p^-}(o, (yh)^{-1}(o))\\ &= \beta_{p^-}(o, p)\\ &= \beta_{n_0^-}(o, n_0a_t) = \beta_{X_0^-}(o, a_t) + \beta_{n_0^-}(o, n_0)\\ &= -t + \beta_{n_0^-}(o, n_0).\end{aligned}$$

As $n_0 a_t m \in P_\epsilon$, we have $e^{-(n-1)t} = 1 + O(\epsilon)$ and hence

$$\begin{aligned} & e^{(n-1)\beta_{g^-}(o,g)} dm_o(g^-) ds dm \\ &= e^{(n-1)(\beta_{g^-}(o,g)+\beta_{p^-}(o,(yh)^{-1}(o))} dm_o(p^-) ds dm \\ &= e^{-(n-1)t} e^{(n-1)\beta_{n_0^-}(o,n_0)} dm_o(n_0^-) dt dm \\ &= e^{-(n-1)t} dn_0 dt dm = (1+O(\epsilon)) dp. \end{aligned}$$

Since $dp = d_{H\cap M}(m) dp'$ for $p = mp'$, for $\overline{\phi}(yh) := \int_{H\cap M} \tilde{\phi}(yhm) d_{H\cap M}(m)$, we have

$$\begin{aligned} \tilde{m}^{\mathrm{BR}_*}(\tilde{\Phi}_\epsilon) &= (1+O(\epsilon)) \int_{P'_\epsilon} \int_{yh \in yH_0/(H\cap M)} \overline{\phi}(yh) \rho_\epsilon(p') d\tilde{\mu}^{\mathrm{PS}}_{yH}(yh) dp' \\ &= (1+O(\epsilon)) \tilde{\mu}^{\mathrm{PS}}_{yH}(\tilde{\phi}). \end{aligned}$$

□

Corollary 5.12. *There exists $\ell \in \mathbb{N}$ such that for any $\phi \in C^\infty(yH_0)$, $\mu^{\mathrm{PS}}_H(\phi) \ll \mathcal{S}_\ell(\phi)$ where the implied constant depends only on the compact subset yH_0.*

Proof. By Lemmas 5.11, 5.7 and (5.8), there exists $\ell \in \mathbb{N}$ such that

$$\mu^{\mathrm{PS}}_H(\phi) \ll m^{\mathrm{BR}_*}(\Phi_{\epsilon_0}) \ll \mathcal{S}_\ell(\Phi_\ell) \ll \mathcal{S}_\ell(\phi)\mathcal{S}_\ell(\rho_{\epsilon_0}) \ll \mathcal{S}_\ell(\phi).$$

where the implied constants depending only on ϵ_0 and yH_0. □

Theorem 5.13. *Suppose that Γ is Zariski dense in G and that $L^2(\Gamma\backslash G)$ has a spectral gap. Then there exist $\eta_0 > 0$ and $\ell \geq 1$ such that for any $\Psi \in C^\infty(\Omega)$ and $\phi \in C^\infty(yH_0)$, we have*

$$\begin{aligned} & e^{(n-1-\delta)t} \int_{yh\in yH} \Psi(yha_t)\phi(yh) dh \\ & \qquad = \frac{1}{|m^{\mathrm{BMS}}|} m^{\mathrm{BR}}(\Psi) \tilde{\mu}^{\mathrm{PS}}_{yH}(\phi) + e^{-\eta_0 t} O(\mathcal{S}_\ell(\Psi)\mathcal{S}_\ell(\phi)), \end{aligned}$$

with the implied constant depending on Ω and yH_0.

Proof. It suffices to prove the claim for Ψ and ϕ non-negative. Let $\ell \geq 1$ be bigger than those ℓ's in Theorem 3.30, (5.8) and Corollary 5.12. Let $q_\ell > 0$ (depending only on the dimension of P') be such that $\mathcal{S}_\ell(\rho_\epsilon) = O(\epsilon^{-q_\ell})$, so that

$$\mathcal{S}_\ell(\Phi_\epsilon) \ll \mathcal{S}_\ell(\phi)\mathcal{S}_\ell(\rho_\epsilon) \ll \mathcal{S}_\ell(\phi)\epsilon^{-q_\ell}.$$

Note that $\mathcal{S}_\ell(\Psi^\pm_\epsilon) \ll \mathcal{S}_\ell(\Psi)$ and that $m^{\mathrm{BR}}(\Psi^\pm_\epsilon) = m^{\mathrm{BR}}(\Psi) + O(\epsilon A^{\mathrm{BR}}_\Psi)$. By Lemma 5.10,

$$\langle a_t \Psi^-_\epsilon, \Phi_\epsilon \rangle \leq \int_{yh\in yH} \Psi(yha_t)\phi(yh) dh \leq \langle a_t \Psi^+_\epsilon, \Phi_\epsilon \rangle.$$

By Lemma 5.11 and Theorem 3.30, there exists $\eta > 0$ such that

$$\begin{aligned} & e^{(n-1-\delta)t}\langle a_t\Psi_\epsilon^\pm, \Phi_\epsilon\rangle \\ &= \tfrac{1}{|m^{\mathrm{BMS}}|} m^{\mathrm{BR}}(\Psi_\epsilon^\pm) m^{\mathrm{BR}_*}(\Phi_\epsilon) + e^{-\eta t} O(\mathcal{S}_\ell(\Psi)\mathcal{S}_\ell(\phi)\epsilon^{-q_\ell}) \\ &= \tfrac{1}{|m^{\mathrm{BMS}}|} m^{\mathrm{BR}}(\Psi)\tilde{\mu}_{yH}^{\mathrm{PS}}(\phi) + O(\epsilon A_\Psi^{\mathrm{BR}}\tilde{\mu}_{yH}^{\mathrm{PS}}(\phi)) + e^{-\eta t} O(\mathcal{S}_\ell(\Psi)\mathcal{S}_\ell(\phi)\epsilon^{-q_\ell}). \end{aligned}$$

By taking $\epsilon = e^{-\eta t/(1+q_\ell)}$ and $\eta_0 = \eta/(1+q_\ell)$, we obtain that

$$\begin{aligned} & e^{(n-1-\delta)t} \int_{yh\in yH} \Psi(yha_t)\phi(yh)dh \\ &= \tfrac{1}{|m^{\mathrm{BMS}}|} m^{\mathrm{BR}}(\Psi)\tilde{\mu}_{yH}^{\mathrm{PS}}(\phi) + e^{-\eta_0 t} O(A_\Psi^{\mathrm{BR}}\tilde{\mu}_{yH}^{\mathrm{PS}}(\phi) + \mathcal{S}_\ell(\Psi)\mathcal{S}_\ell(\phi)). \end{aligned}$$

By (5.8) and Corollary 5.12, this proves the theorem. □

We remark that we don't need to assume yH is closed in the above theorem, as ϕ is assumed to be compactly supported.

When H is horospherical or symmetric with $\text{Pb-corank}_H(\Gamma) = 0$, Theorem 1.7 is a special case of Theorem 5.13 by Theorem 4.5 and Theorem 4.15.

6. Distribution of $\Gamma\backslash\Gamma Ha_t$ and Transversal intersections

Let $\Gamma, H, A = \{a_t\}$, $P = MAN^-$, etc be as in the last section 5. We set $N = N^+$. Let $\{\mu_x\}$ be a Γ-invariant conformal density of dimension $\delta_\mu > 0$ and let $\tilde{\mu}_{gH}$ and $\tilde{\mu}_{gN}$ be the measures on gH and gN respectively defined with respect to $\{\mu_x\}$.

6.1. Transversal intersections. Fix $x \in \Gamma\backslash G$. Let $\epsilon_0 > 0$ be the injectivity radius at x. In particular, the product map $P_{\epsilon_0} \times N_{\epsilon_0} \to \Gamma\backslash G$ given by $(p, n) \mapsto xpn$ is injective. For any $\epsilon \le \epsilon_0$ we set $B_\epsilon := P_\epsilon N_\epsilon$.

For some $c_1 > 1$, we have $N_{c_1^{-1}\epsilon}P_{c_1^{-1}\epsilon} \subset B_\epsilon := P_\epsilon N_\epsilon \subset N_{c_1\epsilon}P_{c_1\epsilon}$ for all $\epsilon > 0$. Therefore, in the arguments below, we will frequently identify B_ϵ with $N_\epsilon P_\epsilon$, up to a fixed Lipschitz constant.

In the next lemma, let $\Psi \in C_c^\infty(xB_{\epsilon_0})^{H\cap M}$ and $\phi \in C_c^\infty(yH)^{H\cap M}$. For $0 < \epsilon \le \epsilon_0$, define $\psi_\epsilon^\pm \in C^\infty(xP)$ by

$$\psi_\epsilon^\pm(xp) = \int_{xpN} \Psi_\epsilon^\pm(xpn) d\mu_{xpN}(xpn)$$

where $\Psi_\epsilon^\pm$ are as given in (5.5).

Define $\phi_\epsilon^\pm \in C_c^\infty(yH)$ by

$$\phi_\epsilon^+(yh) = \sup_{h'\in H_\epsilon} \phi(yhh') \quad \text{and} \quad \phi_\epsilon^-(yh) = \inf_{h'\in H_\epsilon} \phi(yhh'). \tag{6.1}$$

Since the metric d on G is assumed to be left G-invariant and right $H \cap M$-invariant, we have $mH_\epsilon m^{-1} = H_\epsilon$ and $mN_\epsilon m^{-1} = N_\epsilon$. Therefore the functions $\psi_\epsilon^\pm$ and $\phi_\epsilon^\pm$ are $H \cap M$-invariant.

The following lemma is analogous to Corollary 2.14 in [52]; however we are here working in $\Gamma\backslash G$ rather than in $\mathrm{T}^1(\Gamma\backslash\mathbb{H}^n)$ as opposed to [52]. Let

$$P_x(t) := \{p \in P_{\epsilon_0}/(H\cap M) : \mathrm{supp}(\phi)a_t \cap xpN_{\epsilon_0}(H\cap M) \neq \emptyset\}.$$

Lemma 6.2. *For any $0 < \epsilon \ll \epsilon_0$, we have*

$$(1 - c\epsilon) \sum_{p \in P_x(t)} \phi^-_{ce^{-t}\epsilon_0}(xpa_{-t})\psi^-_{c\epsilon}(xp) \le e^{\delta_\mu t} \int_{yH} \Psi(yha_t)\phi(yh)d\mu_{yH}(yh)$$
$$\le (1 + c\epsilon) \sum_{p \in P_x(t)} \phi^+_{ce^{-t}\epsilon_0}(xpa_{-t})\psi^+_{c\epsilon}(xp),$$

where $c > 0$ is an absolute constant, depending only on the injectivity radii of $\mathrm{supp}(\phi)$ *and* $\mathrm{supp}(\Psi)$.

Proof. By considering a smooth partition of unity for the support of ϕ, it suffices to prove the lemma, assuming $\mathrm{supp}(\phi) \subset yN_\epsilon P_\epsilon \cap yH \subset yB_\epsilon$. Fix $g, g' \in G$ so that $y = \Gamma g$ and $x = \Gamma g'$. Then for $\bar{H} = H/H \cap M$,

$$\begin{aligned}
&\int_{yH} \Psi(yha_t)\phi(yh)d\mu_{yH}(yh)\\
&= \sum_{\gamma \in (\Gamma \cap gHg^{-1})\backslash \Gamma} \int_{\gamma gH} \Psi(yha_t)\phi(yh)d\tilde{\mu}_{\gamma gH}(\gamma gh)\\
&= \sum_{\gamma \in (\Gamma \cap gHg^{-1})\backslash \Gamma} \int_{\gamma g\bar{H}} \int_{H\cap M} \Psi(yha_tm)\ \phi(yhm)\ dmd\tilde{\mu}_{\gamma g\bar{H}}(\gamma gh)\\
&= \sum_{\gamma \in (\Gamma \cap gHg^{-1})\backslash \Gamma} \int_{\gamma g\bar{H}} \Psi(yha_t)\phi(yh)\ d\tilde{\mu}_{\gamma g\bar{H}}(\gamma gh)
\end{aligned}$$

as Ψ and ϕ are $H \cap M$-invariant and dm is the probability Haar measure of $H \cap M$.

Suppose $yh \in \mathrm{supp}(\phi) \cap yH$, and write $h = n_h p_h$ where $n_h \in N_\epsilon$ and $p_h \in P_\epsilon$. As $h^+ = n_h^+$ and $d(h, n_h) = O(\epsilon)$, we have that for any $\gamma \in \Gamma$

$$\frac{d\tilde{\mu}_{\gamma g\bar{H}}(\gamma gh)}{d\tilde{\mu}_{\gamma gN}(\gamma gn_h)} = 1 + O(\epsilon). \tag{6.3}$$

Let $\gamma \in (\Gamma \cap gHg^{-1})\backslash\Gamma$. If $\gamma gha_t = g'p_{h,t}n_{h,t} \in g'P_{\epsilon_0}N_{\epsilon_0}$, then we claim that

$$e^{\delta_\mu t}\frac{d\tilde{\mu}_{\gamma gN}(\gamma gn_h)}{d\tilde{\mu}_{g'p_{h,t}N}(g'p_{h,t}n_{h,t})} = O(e^\epsilon). \tag{6.4}$$

Note that $\gamma gha_t = g'p_{h,t}n_{h,t}$ implies $\gamma gn_h a_t = g'p_{h,t}n_{h,t}(a_t^{-1}p_h a_t)$. Hence $\xi := (\gamma gn_h)^+ = (g'p_{h,t}n_{h,t})^+$, and for $p'_{h,t} := (a_t^{-1}p_h a_t) \in P_\epsilon$,

$$\begin{aligned}
&\beta_\xi(o, \gamma gn_h) = \beta_\xi(o, g'p_{h,t}n_{h,t}) + \beta_\xi(g'p_{h,t}n_{h,t}, g'p_{h,t}n_{h,t}p'_{h,t})\\
&+ \beta_\xi(g'p_{h,t}n_{h,t}p'_{h,t}, g'p_{h,t}n_{h,t}p'_{h,t}a_{-t}) = \beta_\xi(o, g'p_{h,t}n_{h,t}) + O(\epsilon) - t,
\end{aligned}$$

proving the claim (6.4).

Note that xB_{ϵ_0} is the disjoint union $\cup_{p\in P_{\epsilon_0}} xpN_{\epsilon_0}$. Since $n_{h,t} \in N_{\epsilon_0}$ and $\gamma gh = g'p_{h,t}a_{-t}(a_t n_{h,t}a_{-t})$ with $a_t n_{h,t}a_{-t} \in N_{e^{-t}\epsilon_0}$, in view of (6.3) and (6.4),

we have

$$\begin{aligned} & e^{\delta_\mu t} \int_{\gamma g \bar{H}} \Psi(yha_t)\phi(yh) d\tilde{\mu}_{\gamma g \bar{H}}(\gamma g h) \\ &= (1+O(\epsilon)) \sum_p \phi^+_{ce^{-t}\epsilon_0}(xpa_{-t}) \cdot \int_{g'pN} \Psi^+_{c\epsilon}(xpn) d\tilde{\mu}_{g'pN}(g'pn) \\ &= (1+O(\epsilon)) \sum_p \phi^+_{ce^{-t}\epsilon_0}(xpa_{-t}) \cdot \psi^+_{c\epsilon}(xp) \end{aligned}$$

where the both sums are taken over the set of $p \in P_{\epsilon_0}/(H \cap M)$ such that $\gamma g H_\epsilon a_t \cap g' p N_{\epsilon_0}(H \cap M) \neq \emptyset$ and $c > 0$ is an absolute constant.

Summing over $\gamma \in (\Gamma \cap gHg^{-1})\backslash\Gamma$, we obtain one side of the inequality and the other side follows if one argues similarly using $\Psi^-_{c\epsilon}$. □

By a similar argument, we can prove the following: In the following lemma, let $\phi \in C_c(yH)^{H\cap M}$ and $\psi \in C^\infty(xP_{\epsilon_0})^{H\cap M}$, and assume that $\mu_{xpN}(xpN_{\epsilon_0}) > 0$ for all $p \in P_{\epsilon_0}$, so that the function $\Psi \in C^\infty(xB_{\epsilon_0})$ can be defined by

$$\Psi(xpn) = \frac{1}{\mu_{xpN}(xpN_{\epsilon_0})}\psi(xp)$$

for each $pn \in P_{\epsilon_0}N_{\epsilon_0}$.

Lemma 6.5. *There exists $c > 1$ such that for all small $0 < \epsilon \le \epsilon_0$,*

$$\begin{aligned} (1-c\epsilon) \int_{yH} \Psi^-_{c\epsilon}(yha_t)\phi^-_{ce^{-t}\epsilon_0}(yh) d\mu_{yH}(yh) &\le e^{-\delta_\mu t} \sum_{p \in P_x(t)} \psi(xp)\phi(xpa_{-t}) \\ &\le (1+c\epsilon) \int_{yH} \Psi^+_{c\epsilon}(yha_t)\phi^+_{ce^{-t}\epsilon_0}(yh) d\mu_{yH}(yh) \end{aligned}$$

Similarly to the definitions of $A^\bullet_\Psi$, we define for $\phi \in C(yH)$ and $\psi \in C(xP_{\epsilon_0})$,

$$A^{\mathrm{PS}}_\phi := \mathcal{S}_{\infty,1}(\phi) \cdot \mu^{\mathrm{PS}}_{yH}(\mathrm{supp}(\phi)), \quad A^\nu_\psi := \mathcal{S}_{\infty,1}(\psi) \cdot \nu_{xP}(\mathrm{supp}(\psi))$$

where ν_{xP} is defined as in (5.1).

By a similar argument as in (5.8), we have $A^{\mathrm{PS}}_\phi \ll \mathcal{S}_\ell(\phi)$ and $A^\nu_\psi \le \mathcal{S}_\ell(\psi)$ for some $\ell \in \mathbb{N}$.

Lemma 6.6. *Let $\psi \in C(xP_{\epsilon_0})^{H\cap M}$. For $\Psi \in C^\infty(xB_{\epsilon_0})^{H\cap M}$ given by $\Psi(xpn) = \frac{1}{\mu^{\mathrm{Haar}}_{xpN}(xpN_{\epsilon_0})}\psi(xp)$, we have*

$$m^{\mathrm{BR}}(\Psi) = \nu_{xP}(\psi) \quad \text{and} \quad A^{\mathrm{BR}}_\Psi \ll A^\nu_\psi.$$

Proof. For $g = xpn$, we have $g^- = (xp)^-$ and $\beta_{(xp)^-}(o, xpn) = \beta_{(xp)^-}(o, xp)$. Based on this, the claims follow from the definition. The second claim follows from $m^{\mathrm{BR}}(\mathrm{supp}(\Psi)) = \nu_{xP}(\mathrm{supp}(\psi))$ and $\mathcal{S}_{\infty,1}(\Psi) \ll_{\epsilon_0} \mathcal{S}_{\infty,1}(\psi)$. □

In the rest of this section, we assume that

$$\Gamma \text{ is Zariski dense and } L^2(\Gamma\backslash G) \text{ has a spectral gap.}$$

Theorem 6.7. *There exist $\beta > 0$ and $\ell \in \mathbb{N}$ such that for any $0 < \epsilon \ll \epsilon_0$ and any $\psi \in C^\infty(xP_{\epsilon_0})^{H\cap M}$, and $\phi \in C_c^\infty(yH)^{H\cap M}$, we have*

$$e^{-\delta t}\sum_{p\in P_x(t)} \psi(xp)\phi(xpa_{-t}) = \frac{1}{|m^{\mathrm{BMS}}|}\nu_{xP}(\psi)\mu_{yH}^{\mathrm{PS}}(\phi) + e^{-\beta t}O(\mathcal{S}_\ell(\psi)\mathcal{S}_\ell(\phi)),$$

where $P_x(t) := \{p \in P_{\epsilon_0}/(H\cap M) : \mathrm{supp}(\phi)a_t \cap xpN_{\epsilon_0}(H\cap M) \neq \emptyset\}$ and the implied constant depends only on the injectivity radii of $\mathrm{supp}(\psi)$ *and* $\mathrm{supp}(\phi)$.

Proof. Define

$$\Psi_\epsilon^\pm(xpn) = \frac{1}{\mu_{xpN}^{\mathrm{Haar}}(xpN_\epsilon)}\psi_\epsilon^\pm(xp).$$

Then $m^{\mathrm{BR}}(\Psi_\epsilon^\pm) = \nu_{xP}(\psi_\epsilon^\pm)$ by Lemma 6.6.

We take ℓ big enough to satisfy Theorem 5.13, Corollary 5.12 and that $A_\Psi^{\mathrm{BR}} \ll A_\psi^\nu \ll \mathcal{S}_\ell(\psi)$ and $A_\phi^{\mathrm{PS}} \ll \mathcal{S}_\ell(\phi)$.

By Theorem 5.13, for some $\eta_0 > 0$,

$$\begin{aligned}
&e^{(n-1-\delta)t}\int_{yH}\Psi_\epsilon^\pm(yha_t)\phi_{e^{-t}\epsilon_0}^\pm(yh)dh\\
&= \tfrac{1}{|m^{\mathrm{BMS}}|}m^{\mathrm{BR}}(\Psi_\epsilon^\pm)\mu_{yH}^{\mathrm{PS}}(\phi_{e^{-t}\epsilon_0}^\pm) + e^{-\eta_0 t}O(\mathcal{S}_\ell(\Psi_\epsilon^\pm)\mathcal{S}_\ell(\phi_{e^{-t}\epsilon_0}^\pm))\\
&= m^{\mathrm{BR}}(\Psi)\mu_{yH}^{\mathrm{PS}}(\phi) + O((\epsilon+e^{-t})A_\Psi^{\mathrm{BR}}A_\phi^{\mathrm{PS}}) + e^{-\eta_0 t}O(\mathcal{S}_\ell(\Psi)\mathcal{S}_\ell(\phi))\\
&= \nu_{xP}(\psi)\mu_{yH}^{\mathrm{PS}}(\phi) + O((e^{-\eta_0 t}+\epsilon)\mathcal{S}_\ell(\psi)\mathcal{S}_\ell(\phi)).
\end{aligned}$$

Therefore the claim now follows by applying Lemma 6.5 for $d\mu_{yH}(yh) = dh$ and $\delta_\mu = n-1$ with $\beta = \eta_0/2$ and $\epsilon = e^{-\eta_0 t/2}$. □

Using Theorem 6.7, we now prove the following theorem, which is analogous to Theorem 5.13 with dh replaced by $d\mu_{yH}^{\mathrm{PS}}(yh)$. Translates of $d\mu_{yH}^{\mathrm{Haar}}$ and $d\mu_{yH}^{\mathrm{PS}}$ on yH are closely related as their transversals are essentially the same. More precisely, Theorem 6.7 provides a link between translates of these two measures.

Theorem 6.8. *There exist $\beta > 0$ and $\ell \in \mathbb{N}$ such that for any $\Psi \in C^\infty(xB_{\epsilon_0})^{H\cap M}$ and $\phi \in C_c^\infty(yH)^{H\cap M}$,*

$$\int_{yH}\Psi(yha_t)\phi(yh)d\mu_{yH}^{\mathrm{PS}}(yh) = \frac{1}{|m^{\mathrm{BMS}}|}m^{\mathrm{BMS}}(\Psi)\mu_{yH}^{\mathrm{PS}}(\phi) + O(e^{-\beta t}\mathcal{S}_\ell(\Psi)\mathcal{S}_\ell(\phi)).$$

Proof. Define $\psi \in C^\infty(xP_{\epsilon_0})^{H\cap M}$ by

$$\psi(xp) = \int_{xpN_{\epsilon_0}}\Psi(xpn)d\mu_{xpN}^{\mathrm{PS}}(xpn).$$

We apply Theorem 6.7 and Lemma 6.2 for the Patterson-Sullivan density $\{\mu_x\}$ and with this ψ. It follows from the definition of ψ (see (5.4)) that $\nu_{xP}(\psi) = m^{\mathrm{BMS}}(\Psi)$ and $A_\psi^\nu \ll \mathcal{S}_\ell(\Psi)$ for some $\ell \geq 1$. We take ℓ large enough to satisfy Theorem 6.7.

Let β be as in Theorem 6.7 and let $\epsilon = e^{-\beta t}$. Now by Lemma 6.2, we get

$$\int_{yH} \Psi(yha_t)\phi(yh)d\mu_{yH}^{\mathrm{PS}}(yh) = (1+O(\epsilon))e^{-\delta t} \sum_{p\in P_x(t)} \phi^{\pm}_{e^{-t}\epsilon_0}(xpa_{-t})\psi^{\pm}_{\epsilon}(xp).$$

By Theorem 6.7,

$$\begin{aligned} & e^{-\delta t} \sum_{p\in P_x(t)} \phi^{\pm}_{e^{-t}\epsilon_0}(xpa_{-t})\psi^{\pm}_{\epsilon}(xp) \\ &= \tfrac{1}{|m^{\mathrm{BMS}}|}\nu_{xP}(\psi^{\pm}_{\epsilon})\mu^{\mathrm{PS}}_{yH}(\phi^{\pm}_{e^{-t}\epsilon_0}) + e^{-\beta t}O(\mathcal{S}_\ell(\psi)\mathcal{S}_\ell(\phi)) \\ &= \tfrac{1}{|m^{\mathrm{BMS}}|}\nu_{xP}(\psi)\mu^{\mathrm{PS}}_{yH}(\phi) + O(\epsilon + e^{-\beta t})(\mathcal{S}_\ell(\psi)\mathcal{S}_\ell(\phi)). \end{aligned}$$

Since $\nu_{xP}(\psi) = m^{\mathrm{BMS}}(\Psi)$ and $\mathcal{S}_\ell(\psi), A^{\nu}_{\psi} \ll \mathcal{S}_\ell(\Psi)$, this finishes the proof. □

6.2. Effective equidistribution of $\Gamma\backslash\Gamma Ha_t$. We now extend Theorems 5.13 and 6.8 to bounded functions $\phi \in C^\infty((\Gamma\cap H)\backslash H)$ which are not necessarily compactly supported. Hence the goal is to establish the following: set $\Gamma_H := \Gamma \cap H$.

Theorem 6.9. *Suppose that $\Gamma\backslash\Gamma H$ is closed and that $|\mu_H^{\mathrm{PS}}| < \infty$. There exist $\beta > 0$ and $\ell \geq 1$ such that for any compact subset $\Omega \subset \Gamma\backslash G$, for any $\Psi \in C^\infty(\Omega)$ and any bounded $\phi \in C^\infty(\Gamma_H\backslash H)$, we have, as $t \to +\infty$,*

$$e^{(n-1-\delta)t}\int_{h\in\Gamma_H\backslash H}\Psi(ha_t)\phi(h)dh = \frac{\mu_H^{\mathrm{PS}}(\phi)}{|m^{\mathrm{BMS}}|}m^{\mathrm{BR}}(\Psi)+O(e^{-\beta t}\mathcal{S}_\ell(\Psi)\mathcal{S}_\ell(\phi))$$

where the implied constant depends only on Ω.

We first prove the following which is an analogous version of Theorem 6.9 for μ_H^{PS}:

Theorem 6.10. *Suppose that $\Gamma\backslash\Gamma H$ is closed and that $|\mu_H^{\mathrm{PS}}| < \infty$. There exist $\beta_0 > 0$ and $\ell \geq 1$ such that for any compact subset $\Omega \subset \Gamma\backslash G$, for any $\Psi \in C^\infty(\Omega)^{H\cap M}$ and for any bounded $\phi \in C^\infty((\Gamma\cap H)\backslash H)^{H\cap M}$, we have*

$$\int_{h\in\Gamma_H\backslash H}\Psi(ha_t)\phi(h)d\mu_H^{\mathrm{PS}}(h) = \frac{\mu_H^{\mathrm{PS}}(\phi)}{|m^{\mathrm{BMS}}|}m^{\mathrm{BMS}}(\Psi) + O(e^{-\beta_0 t}\mathcal{S}_\ell(\phi)\mathcal{S}_\ell(\Psi)).$$

Proof. Fix $\ell \in \mathbb{N}$ large enough to satisfy Theorem 6.8, $A^{\mathrm{PS}}_{\phi} \ll \mathcal{S}_\ell(\phi)$ and $A^{\mathrm{BMS}}_{\Psi} \ll \mathcal{S}_\ell(\Psi)$. If H is horospherical, set $Y_\Omega = \{h \in \Gamma\backslash\Gamma HM : ha_t \in \Omega \text{ for some } t \in \mathbb{R}\}$; if H is symmetric, let Y_Ω and Y_ϵ be as in Theorem 4.15 and set $p_0 := \text{Pb-corank}_H(\Gamma)$. For $\epsilon > 0$, we choose $\tau_\epsilon \in C^\infty(Y_\Omega)$ which is an $H\cap M$-invariant smooth approximation of the set Y_ϵ; $0 \leq \tau_\epsilon \leq 1$, $\tau_\epsilon(x) = 1$ for $x \in Y_\epsilon$ and $\tau_\epsilon(x) = 0$ for $x \notin Y_{\epsilon/2}$; we refer to [4] for the construction of such τ_ϵ. Let $q_\ell \gg 1$ be such that $\mathcal{S}_\ell(\tau_\epsilon) = O(\epsilon^{-q_\ell})$. By the definition of Y_Ω, we may write the integral $\int_{h\in\Gamma_H\backslash H}\Psi(ha_t)\phi(h)d\mu_H^{\mathrm{PS}}(h)$ as the sum

$$\int_{\Gamma_H\backslash H}\Psi(ha_t)(\phi\cdot\tau_\epsilon)(h)d\mu_H^{\mathrm{PS}}(h) + \int_{Y_\Omega}\Psi(ha_t)(\phi-\phi\cdot\tau_\epsilon)(h)d\mu_H^{\mathrm{PS}}(h).$$

Note that by (3) of Theorem 4.16, we have $\mu_H^{\mathrm{PS}}(Y_\Omega - Y_\epsilon) \ll \epsilon^{\delta-p_0}$ and hence $\mu_H^{\mathrm{PS}}(\phi - \phi\cdot\tau_\epsilon) \ll A_\phi^{\mathrm{PS}}\cdot\epsilon^{\delta-p_0} \ll \mathcal{S}_\ell(\phi)\epsilon^{\delta-p_0}$. Now by Theorem 6.8,

$$\begin{aligned}
&\int_{\Gamma\backslash\Gamma H} \Psi(ha_t)(\phi\cdot\tau_\epsilon)(h)d\mu_H^{\mathrm{PS}}(h)\\
&= \frac{\mu_H^{\mathrm{PS}}(\phi\cdot\tau_\epsilon)}{|m^{\mathrm{BMS}}|}m^{\mathrm{BMS}}(\Psi) + O(\epsilon^{-q_\ell}e^{-\beta t}\mathcal{S}_\ell(\phi)\mathcal{S}_\ell(\Psi))\\
&= \frac{\mu_H^{\mathrm{PS}}(\phi)}{|m^{\mathrm{BMS}}|}m^{\mathrm{BMS}}(\Psi) + O(A_\phi^{\mathrm{PS}}A_\Psi^{\mathrm{BMS}}\epsilon^{\delta-p_0}) + O(\epsilon^{-q_\ell}e^{-\beta t}\mathcal{S}_\ell(\phi)\mathcal{S}_\ell(\Psi))\\
&= \frac{\mu_H^{\mathrm{PS}}(\phi)}{|m^{\mathrm{BMS}}|}m^{\mathrm{BMS}}(\Psi) + O((\epsilon^{\delta-p_0}+\epsilon^{-q_\ell}e^{-\beta t})\mathcal{S}_\ell(\phi)\mathcal{S}_\ell(\Psi)).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\int_{Y_\Omega} \Psi(ha_t)(\phi-\phi\cdot\tau_\epsilon)(h)d\mu_H^{\mathrm{PS}}(h)\\
&\qquad\ll \mathcal{S}_{\infty,1}(\Psi)\mathcal{S}_{\infty,1}(\phi)\mu_H^{\mathrm{PS}}(Y_\Omega - Y_\epsilon) \ll \mathcal{S}_\ell(\Psi)\mathcal{S}_\ell(\phi)\epsilon^{\delta-p_0}.
\end{aligned}$$

Hence by combining these two estimates, and taking $\epsilon = e^{-\beta/(\delta-p_0+q_\ell)}$ and $\beta_0 := e^{-\beta(\delta-p_0)/(\delta-p_0+q_\ell)}$, we obtain

$$\int_{h\in\Gamma_H\backslash H} \Psi(ha_t)\phi(h)d\mu_H^{\mathrm{PS}}(h) = \frac{\mu_H^{\mathrm{PS}}(\phi)}{|m^{\mathrm{BMS}}|}m^{\mathrm{BMS}}(\Psi) + O(e^{-\beta_0 t}\mathcal{S}_\ell(\phi)\mathcal{S}_\ell(\Psi)).$$

□

Proof of Theorem 6.9. We will divide the integration region into three different regions: compact part, thin part, intermediate range. The compact part is the region where we get the main term using Theorem 5.13. The thin region can be controlled using Theorem 4.16. However there is an intermediate range where we need some control. This is in some sense the main technical difference from the case where Γ is a lattice. We control the contribution from this range, using results proved in this section in particular by relating this integral to summation over the "transversal"; see Lemmas 6.2 and 6.5.

We use the notation from the proof of Theorem 6.10. In particular, if H is horospherical, set $Y_\Omega = \{h \in \Gamma\backslash\Gamma HM : ha_t \in \Omega \text{ for some } t\in\mathbb{R}\}$; if H is symmetric, let Y_Ω and Y_ϵ be as in Theorem 4.15 and set $p_0 := \text{Pb-corank}_H(\Gamma)$. Let $0 < \epsilon_1 < \epsilon_0$. Here, we regard Y_{ϵ_0} as a thick part, $Y_{\epsilon_1} - Y_{\epsilon_0}$ as an intermediate range and $\Gamma_H\backslash H - Y_{\epsilon_1}$ as a thin part.

As above we choose $\tau_{\epsilon_0} \in C^\infty(Y)$ which is an $H\cap M$-invariant smooth approximation of Y_{ϵ_0} and recall that $\mu_H^{\mathrm{PS}}(Y_\Omega - Y_{\epsilon_0}) \ll \epsilon_0^{\delta-p_0}$ by (1) of Proposition 4.14.

We may write

$$\int_{h\in\Gamma_H\backslash H}\Psi(ha_t)\phi(h)dh$$
$$=\int_{\Gamma_H\backslash H}\Psi(ha_t)(\phi\cdot\tau_{\epsilon_0})(h)dh+\int_{Y_\Omega}\Psi(ha_t)(\phi-\phi\cdot\tau_{\epsilon_0})(h)dh.$$

Then by Theorem 5.13 with $\eta_0>0$ therein and Theorem 4.16, we get the asymptotic for the thick part:

$$e^{(n-1-\delta)t}\int_{\Gamma_H\backslash H}\Psi(ha_t)(\phi\cdot\tau_{\epsilon_0})(h)dh$$
$$=\frac{\mu_H^{\mathrm{PS}}(\phi)}{|m^{\mathrm{BMS}}|}m^{\mathrm{BR}}(\Psi)+(\epsilon_0^{\delta-p_0}+e^{-\eta_0 t}\epsilon_0^{-q_\ell})O(\mathcal{S}_\ell(\phi)\mathcal{S}_\ell(\Psi)).\quad(6.11)$$

On the other hand, by Theorem 4.16, we have, for $\mathcal{T}_{\epsilon_1}:=\tau_{\epsilon_1}-\tau_{\epsilon_0}$,

$$\int_{Y_\Omega}\Psi(ha_t)(\phi-\phi\cdot\tau_{\epsilon_0})(h)dh$$
$$\ll\mathcal{S}_{\infty,1}(\phi)\left(\int_{Y_\Omega}\Psi(ha_t)\mathcal{T}_{\epsilon_1}(h)dh+\int_{Y_\Omega-Y_{\epsilon_1}}\Psi(ha_t)dh\right).\quad(6.12)$$

Set $\overline{\Psi}(x):=\int_{H\cap M}\Psi(xm)dm$. Applying Lemma 6.2 for the Haar measure $d\mu_H^{\mathrm{Haar}}=dh$ and Lemma 6.5 for the PS measure μ_H^{PS}, and for the function $\mathcal{T}:=\mathcal{T}_{\epsilon_1}$, with the notation as in the proof of Theorem 6.7, we get the following estimate of the integral over the intermediate range $Y_{\epsilon_1}-Y_{\epsilon_0}$:

$$e^{(n-1-\delta)t}\int_{Y_\Omega}\Psi(ha_t)\mathcal{T}_{\epsilon_1}(h)dh$$
$$\ll e^{(n-1-\delta)t}\int_{Y_\Omega}\overline{\Psi}(ha_t)\mathcal{T}_{\epsilon_1}(h)dh$$
$$\ll e^{-\delta t}\sum_{p\in P_e(t)}\overline{\psi}_{\epsilon_1}^+(xp)\mathcal{T}^+_{e^{-t}\epsilon_1}(xpa_{-t})$$
$$\ll\int_{\Gamma\backslash\Gamma H}\overline{\Psi}^+_{\epsilon_1}(ha_t)\mathcal{T}^+_{e^{-t}\epsilon_1}(h)d\mu_H^{\mathrm{PS}}(h)$$
$$\ll\mathcal{S}_{\infty,1}(\overline{\Psi})\mu_H^{\mathrm{PS}}(Y_{\epsilon_1}-Y_{\epsilon_0})\ll\mathcal{S}_\ell(\Psi)\epsilon_0^{\delta-p_0}.\quad(6.13)$$

Using Theorem 4.16, we also get the following estimate of the integral over the thin part, which is the complement of Y_{ϵ_1}:

$$e^{(n-1-\delta)t}\int_{Y_\Omega-Y_{\epsilon_1}}\Psi(ha_t)dh\le\mathcal{S}_\ell(\Psi)e^{(n-1-\delta)t}\epsilon_1^{n-1+p_0}.\quad(6.14)$$

Therefore by (6.11), (6.12),(6.13), and (6.14),

$$e^{(n-1-\delta)t}\int_{h\in\Gamma_H\backslash H}\Psi(ha_t)\phi(h)dh$$
$$=\frac{\mu_H^{\mathrm{PS}}(\phi)}{|m^{\mathrm{BMS}}|}m^{\mathrm{BR}}(\Psi)+O(\epsilon_0^{\delta-p_0}+e^{-\eta_0 t}\epsilon_0^{-q_\ell}+e^{(n-1-\delta)t}\epsilon_1^{n-1-p_0})\mathcal{S}_\ell(\phi)\mathcal{S}_\ell(\Psi).$$

Recalling $\delta > p_0$, take ϵ_0 and ϵ_1 by $\epsilon_0 = e^{-\eta_0 t/(\delta-p_0+q_\ell)}$ and $\epsilon_1^{n-1-p_0} = \epsilon_0^{\delta-p_0}e^{(\delta-n+1)t}$. We may assume that $\epsilon_1 < \epsilon_0$ by taking ℓ and hence q_ℓ big enough. Finally, we obtain the claim with $\beta := \eta_0(\delta-p_0)/(\delta-p_0+q_\ell)$. □

We can also prove an analogue of Theorem 6.9 with a_t replaced by a_{-t}, by following a similar argument step by step but using Corollary 3.34 in place of Theorem 3.30. Consider the $H\cap M$-invariant measure $\mu_{H,-}^{\mathrm{PS}}$ on $\Gamma_H\backslash H$ induced by the measure $e^{\delta\beta_{h^-}(o,h)}d\nu_o(h^-)$ on $\bar{H} = H/(H\cap M)$:

$$d\mu_{H,-}^{\mathrm{PS}}(hm) = e^{\delta\beta_{h^-}(o,h)}d\nu_o(h^-)d_{H\cap M}(m). \tag{6.15}$$

Theorem 6.16. *Suppose that $|\mu_{H,-}^{\mathrm{PS}}| < \infty$. There exist $\beta > 0$ and $\ell \geq 1$ such that for any compact subset Ω in $\Gamma\backslash G$, any $\Psi \in C^\infty(\Omega)$ and any bounded $\phi \in C^\infty(\Gamma_H\backslash H)$, we have, as $t\to+\infty$,*

$$e^{(n-1-\delta)t}\int_{h\in\Gamma_H\backslash H}\Psi(ha_{-t})\phi(h)dh = \frac{\mu_{H,-}^{\mathrm{PS}}(\phi)}{|m^{\mathrm{BMS}}|}m^{\mathrm{BR}_*}(\Psi)+O(e^{-\beta t}\mathcal{S}_\ell(\Psi)\mathcal{S}_\ell(\phi))$$

where the implied constant depends only on Ω.

6.3. Effective mixing of the BMS measure. In this subsection we prove an effective mixing for the BMS measure:

Theorem 6.17. *There exist $\beta > 0$ and $\ell \in \mathbb{N}$ such that for any compact subset $\Omega \subset \Gamma\backslash G$, and for any $\Psi, \Phi \in C^\infty(\Omega)$,*

$$\int_{\Gamma\backslash G}\Psi(ga_t)\Phi(g)dm^{\mathrm{BMS}}(g) = \tfrac{1}{|m^{\mathrm{BMS}}|}m^{\mathrm{BMS}}(\Psi)m^{\mathrm{BMS}}(\Phi)+O(e^{-\beta t}\mathcal{S}_\ell(\Psi)\mathcal{S}_\ell(\Phi))$$

with the implied constant depending only on Ω.

Proof. Using a smooth partition of unity for Ω, it suffices to prove the claim for $\Phi \in C_c(xB_{\epsilon_0})$ for $x\in\Omega$, $B_{\epsilon_0} = P_{\epsilon_0}N_{\epsilon_0}$ and $\epsilon_0 > 0$ smaller than the injectivity radius of Ω.

By Theorem 6.8 with $H = N$ and for each $p \in P_{\epsilon_0}$,

$$\int_{xpN_{\epsilon_0}}\Psi(xpna_t)\Phi(xpn)d\mu_{xpN}^{\mathrm{PS}}(xpn)$$
$$=\tfrac{1}{|m^{\mathrm{BMS}}|}m^{\mathrm{BMS}}(\Psi)\mu_{xpN}^{\mathrm{PS}}(\Phi|_{xpN_{\epsilon_0}})+e^{-\beta t}O(\mathcal{S}_\ell(\Psi)\mathcal{S}_\ell(\Phi|_{xpN_{\epsilon_0}}))$$

for some $\beta > 0$ and $\ell \in \mathbb{N}$. As

$$\int_{xP_{\epsilon_0}}\mu_{xpN}^{\mathrm{PS}}(\Phi|_{xpN_{\epsilon_0}})d\nu_{xP}(xp) = m^{\mathrm{BMS}}(\Phi),$$

we have

$$\int_{xB_{\epsilon_0}} \Psi(ga_t)\Phi(g)dm^{\mathrm{BMS}}(g)$$
$$= \int_{xp\in xP_{\epsilon_0}} \int_{xpN_{\epsilon_0}} \Psi(xpna_t)\Phi(xpn)d\mu^{\mathrm{PS}}_{xpN}(xpn)d\nu_{xP}(xp)$$
$$= \tfrac{1}{|m^{\mathrm{BMS}}|} m^{\mathrm{BMS}}(\Psi)m^{\mathrm{BMS}}(\Phi) + O(e^{-\beta t})\mathcal{S}_\ell(\Psi)\cdot \int_{xP_{\epsilon_0}} \mathcal{S}_\ell(\Phi|_{xpN_{\epsilon_0}})d\nu_{xP}(xp).$$

Since

$$\int_{xP_{\epsilon_0}} \mathcal{S}_\ell(\Phi|_{xpN_{\epsilon_0}})d\nu_{xP}(xp) \ll \mathcal{S}_\ell(\Phi)m^{\mathrm{BR}}(\operatorname{supp}\Phi) \ll_\Omega S_\ell(\Phi),$$

this finishes the proof. □

7. Effective uniform counting

7.1. The case when H is symmetric or horospherical. Let $G, H, A = \{a_t : t \in \mathbb{R}\}$, K, etc be as in the section 5. Let Γ be a Zariski dense and geometrically finite group with $\delta > (n-1)/2$. Suppose that $[e]\Gamma$ is discrete in $H\backslash G$, equivalently, $\Gamma\backslash\Gamma H$ is closed in $\Gamma\backslash G$ and that $|\mu^{\mathrm{PS}}_H| < \infty$.

In this section, we will obtain effective counting results from Theorem 6.9 with ϕ being the constant function 1 on $(\Gamma\cap H)\backslash H$.

Definition 7.1 (Uniform spectral gap)**.** *A family of subgroups $\{\Gamma_i < \Gamma : i \in I\}$ of finite index is said to have a uniform spectral gap property if*

$$\sup_{i\in I} s_0(\Gamma_i) < \delta \quad \text{and} \quad \sup_{i\in I} n_0(\Gamma_i) < \infty.$$

where $s_0(\Gamma_i)$ and $n_0(\Gamma_i)$ are defined as in (1.3)*.*

The pair $(\sup_{i\in I} s_0(\Gamma_i), \sup_{i\in I} n_0(\Gamma_i))$ will be referred to as the uniform spectral gap data for the family $\{\Gamma_i : i \in I\}$.

As we need to keep track of the main term when varying Γ over its subgroups of finite index for our intended applications to affine sieve, we consider the following situation: let $\Gamma_0 < \Gamma$ be a subgroup of finite index with $\Gamma_0 \cap H = \Gamma \cap H$ and fix $\gamma_0 \in \Gamma$. Throughout this section, we assume that both Γ and Γ_0 have spectral gaps; hence $\{\Gamma, \Gamma_0\}$ is assumed to have a uniform spectral gap. By Theorem 3.27, this assumption is automatic if $\delta > (n-1)/2$ for $n = 2, 3$ and if $\delta > n-2$ for $n \geq 4$.

For a family $\mathcal{B}_T \subset H\backslash G$ of compact subsets, we would like to investigate $\#[e]\Gamma_0\gamma_0 \cap \mathcal{B}_T$.

Define a function $F_T := F_{\mathcal{B}_T}$ on $\Gamma_0\backslash G$ by

$$F_T(g) := \sum_{\gamma\in H\cap\Gamma\backslash\Gamma_0} \chi_{\mathcal{B}_T}([e]\gamma g)$$

where $\chi_{\mathcal{B}_T}$ denotes the characteristic function of $\mathcal{B}_T$. Note that

$$F_T(\gamma_0) = \#[e]\Gamma_0\gamma_0 \cap \mathcal{B}_T.$$

Denote by $\{\nu_x\}$ the Patterson-Sullivan density for Γ normalized so that $|\nu_o| = 1$. Clearly, $\{\nu_x\}$ is the unique PS density for Γ_0 with $|\nu_o| = 1$. Recall the Lebesgue density $\{m_x\}$ with $|m_o| = 1$.

Therefore if $\tilde{m}^{\mathrm{BMS}}$, $\tilde{m}^{\mathrm{BR}}$ and $\tilde{m}^{\mathrm{Haar}}$ are the BMS measure, the BR measure, the Haar measure on G, the corresponding measures $m^{\mathrm{BMS}}_{\Gamma_0}$, $m^{\mathrm{BR}}_{\Gamma_0}$ and $m^{\mathrm{Haar}}_{\Gamma_0}$ on $\Gamma_0\backslash G$ are naturally induced from them. In particular, for each $\bullet = \mathrm{BMS}, \mathrm{BR}, \mathrm{Haar}$, $|m^{\bullet}_{\Gamma_0}| = [\Gamma : \Gamma_0] \cdot |m^{\bullet}|$. Since $H \cap \Gamma = H \cap \Gamma_0$, we have $|\mu^{\mathrm{PS}}_H| = |\mu^{\mathrm{PS}}_{\Gamma_0,H}|$ and $|\mu^{\mathrm{PS}}_{H,-}| = |\mu^{\mathrm{PS}}_{\Gamma_0,H,-}|$.

7.2. Weak-convergence of counting function. Fix $\psi \in C_c^\infty(G)$. For $k \in K$ and $\gamma \in \Gamma$, define $\psi^k, \psi_\gamma \in C_c^\infty(G)$ by $\psi^k(g) = \psi(gk)$ and $\psi_\gamma(g) = \psi(\gamma g)$. Also define $\Psi, \Psi_\gamma \in C_c^\infty(\Gamma_0\backslash G)$ by

$$\Psi(g) := \sum_{\gamma' \in \Gamma_0} \psi(\gamma' g) \quad \text{and} \quad \Psi_\gamma(g) := \sum_{\gamma' \in \Gamma_0} \psi(\gamma\gamma' g).$$

For a function f on K, define a function $\psi *_K f$, or simply $\psi * f$, on G by

$$\psi * f(g) = \int_{k \in K} \psi(gk) f(k)\, dk.$$

For a subset $\mathcal{B} \subset H\backslash G$, define a function $f^{\pm}_{\mathcal{B}}$ on K by

$$f^{\pm}_{\mathcal{B}}(k) = \int_{a_{\pm t} \in \mathcal{B}k^{-1} \cap [e]A^{\pm}} e^{\delta t} dt.$$

We adopt the notation $\tilde{m}^{\mathrm{BR}}_+ = \tilde{m}^{\mathrm{BR}}$ and $\tilde{m}^{\mathrm{BR}}_- = \tilde{m}^{\mathrm{BR}_*}$ below. Recall that m^{BMS} means m^{BMS}_Γ in the whole section.

Proposition 7.2. *There exist $\beta_1 > 0$ and $\ell \geq 1$ (depending only on the uniform spectral gap data of Γ and Γ_0) such that for any $T \gg 1$, and any $\gamma \in \Gamma$, the pairing $\langle F_T, \Psi_\gamma \rangle$ in $\Gamma_0\backslash G$ is given by*

$$\begin{cases} \frac{|\mu^{\mathrm{PS}}_H|}{[\Gamma:\Gamma_0]\cdot|m^{\mathrm{BMS}}|}\tilde{m}^{\mathrm{BR}}(\psi * f^+_{\mathcal{B}_T}) + O(\max_{a_t \in \mathcal{B}_T} e^{(\delta-\beta_1)t} \cdot \mathcal{S}_\ell(\psi)) & \text{if } G = HA^+K \\ \sum \frac{|\mu^{\mathrm{PS}}_{H^\pm}|}{[\Gamma:\Gamma_0]\cdot|m^{\mathrm{BMS}}|}\tilde{m}^{\mathrm{BR}}_\pm(\psi * f^\pm_{\mathcal{B}_T}) + O(\max_{a_t \in \mathcal{B}_T} e^{(\delta-\beta_1)|t|} \cdot \mathcal{S}_\ell(\psi)) & \text{otherwise.} \end{cases}$$

Proof. For the Haar measure $d\tilde{m}^{\mathrm{Haar}}(g) = dg$, we may write $dg = \rho(t) dh dt dk$ where $g = ha_t k$ and $\rho(t) = e^{(n-1)|t|}(1 + O(e^{-\alpha_1|t|}))$ for some $\alpha_1 > 0$ (cf. [52]).

Setting $\kappa^{\pm}(\Gamma_0) := \frac{|\mu^{\mathrm{PS}}_{\Gamma_0,H,\pm}|}{|m^{\mathrm{BMS}}_{\Gamma_0}|}$, we have $\kappa^{\pm}(\Gamma_0) = \frac{1}{[\Gamma:\Gamma_0]}\kappa^{\pm}(\Gamma)$. We will only prove the claim for the case $G = HA^+K$, as the other case can be deduced in a similar fashion, based on Theorem 6.16. We apply Theorem 6.9 and obtain:

$$\begin{aligned}\langle F_T, \Psi_\gamma\rangle &= \int_{[e]a_tk\in\mathcal{B}_T}\left(\int_{(H\cap\Gamma)\backslash H}\Psi_\gamma(ha_tk)dh\right)\rho(t)dtdk\\ &= \kappa^+(\Gamma_0)\int_{[e]a_tk\in\mathcal{B}_T} e^{(\delta-n+1)t}m^{\mathrm{BR}}_{\Gamma_0}(\Psi^k_\gamma)\rho(t)dtdk\\ &+ \int_{[e]a_tk\in\mathcal{B}_T} e^{(\delta-n+1-\beta)t}\rho(t)dtdk\cdot O(\mathcal{S}_\ell(\Psi))\\ &= \kappa^+(\Gamma_0)\int_{[e]a_tk\in\mathcal{B}_T} e^{\delta t}\tilde{m}^{\mathrm{BR}}(\psi^k_\gamma)dtdk + O(\max_{a_t\in\mathcal{B}_T} e^{(\delta-\beta_1)t}\cdot\mathcal{S}_\ell(\psi))\end{aligned}$$

where $\beta_1 = \min\{\beta, \alpha_1\}$.

By the left Γ-invariance of $\tilde{m}^{\mathrm{BR}}$, we have $\tilde{m}^{\mathrm{BR}}(\psi^k_\gamma) = \tilde{m}^{\mathrm{BR}}(\psi^k)$. Hence

$$\int_{[e]a_tk\in\mathcal{B}_T} e^{\delta t}\tilde{m}^{\mathrm{BR}}(\psi^k_\gamma)dtdk = \int_{k\in K}\int_{a_t\in\mathcal{B}_Tk^{-1}} e^{\delta t}\tilde{m}^{\mathrm{BR}}(\psi^k)dtdk = \tilde{m}^{\mathrm{BR}}(\psi * f^+_{\mathcal{B}_T}).$$

This finishes the proof. □

7.3. Counting and the measure $\mathcal{M}_{H\backslash G}$. We denote by $X_0 \in \mathrm{T}^1(\mathbb{H}^n)$ the vector fixed by M. In the rest of this section, we define the measures $d\nu^\pm_o(k)$ on K as follows: for $f \in C(K)$,

$$\int_K f(k)d\nu^\pm_o(k) = \int_{M\backslash K}\int_M f(km)dmd\nu_o(kX^\pm_0) \tag{7.3}$$

where dm is the probability Haar measure of M.

Define a measure $\mathcal{M}_{H\backslash G} = \mathcal{M}^\Gamma_{H\backslash G}$ on $H\backslash G$: for $\phi \in C_c(H\backslash G)$,

$$\mathcal{M}_{H\backslash G}(\phi) = \tag{7.4}$$

$$\begin{cases}\frac{|\mu^{\mathrm{PS}}_H|}{|m^{\mathrm{BMS}}|}\int_{a_tk\in A^+K}\phi(a_tk)e^{\delta t}dtd\nu^-_o(k^{-1}) & \text{if } G = HA^+K\\ \sum\frac{|\mu^{\mathrm{PS}}_{H,\pm}|}{|m^{\mathrm{BMS}}|}\int_{a_{\pm t}k\in A^\pm K}\phi(a_{\pm t}k)e^{\delta t}dtd\nu^\pm_o(k^{-1}) & \text{otherwise.}\end{cases} \tag{7.5}$$

Observe that the measure $\mathcal{M}_{H\backslash G}$ depends on Γ but is independent of the normalization of the PS-density.

Theorem 7.6. *If $\{\mathcal{B}_T \subset H\backslash G\}$ is effectively well-rounded with respect to Γ (see Def. 1.10), then there exists $\eta_0 > 0$ (depending only on a uniform spectral gap data for Γ and Γ_0) such that for any $\gamma_0 \in \Gamma$*

$$\#([e]\Gamma_0\gamma_0\cap\mathcal{B}_T) = \tfrac{1}{[\Gamma:\Gamma_0]}\mathcal{M}_{H\backslash G}(\mathcal{B}_T) + O(\mathcal{M}_{H\backslash G}(\mathcal{B}_T)^{1-\eta_0})$$

with the implied constant independent of Γ_0 and $\gamma_0 \in \Gamma$.

Proof. Let $\psi^\epsilon \in C^\infty(G)$ be an ϵ-smooth approximation of e: $0 \le \psi^\epsilon \le 1$, $\mathrm{supp}(\psi^\epsilon) \subset G_\epsilon$ and $\int\psi^\epsilon dg = 1$. Set $\mathcal{B}^+_{T,\epsilon} := \mathcal{B}_TG_\epsilon$ and $\mathcal{B}^-_{T,\epsilon} := \cap_{g\in G_\epsilon}\mathcal{B}_Tg$. Then

$$\langle F_{\mathcal{B}^-_{T,\epsilon}}, \Psi^\epsilon_{\gamma_0^{-1}}\rangle \le F_T(\gamma_0) \le \langle F_{\mathcal{B}^+_{T,\epsilon}}, \Psi^\epsilon_{\gamma_0^{-1}}\rangle.$$

Again, we will provide a proof only for the case $G = HA^+K$; the other case can be done similarly, based on Proposition 7.2. By Proposition 7.2, for $\kappa^+(\Gamma_0) := \frac{|\mu^{\mathrm{PS}}_{\Gamma_0,H}|}{|m^{\mathrm{BMS}}_{\Gamma_0}|}$,

$$\langle F_{\mathcal{B}^{\pm}_{T,\epsilon}}, \Psi^{\epsilon}_{\gamma_0^{-1}}\rangle = \kappa^+(\Gamma_0)\tilde{m}^{\mathrm{BR}}(\psi^{\epsilon} * f_{\mathcal{B}^{\pm}_{T,\epsilon}}) + O(\max_{a_t\in\mathcal{B}_T} e^{(\delta-\beta_1)t}\epsilon^{-q_\ell}).$$

where q_ℓ is so that $\mathcal{S}_\ell(\psi^\epsilon) = O(\epsilon^{-q_\ell})$. For $g = a_r n k' \in ANK$, define $H(g) = r$ and $\kappa(g) = k'$.

Now, using the strong wave front property for ANK decomposition [24], and the definition 1.10, there exists $c > 1$ such that for any $g \in G_\epsilon$ and $T \gg 1$,

$$f_{\mathcal{B}^-_{T,c\epsilon}}(\kappa(k^{-1})) \le f_{\mathcal{B}_T}(\kappa(k^{-1}g)) \le f_{\mathcal{B}^+_{T,c\epsilon}}(\kappa(k^{-1})).$$

We use the formula for $\tilde{m}^{\mathrm{BR}}$ (cf. [52]):

$$d\tilde{m}^{\mathrm{BR}}(ka_rn) = e^{-\delta r} dn dr d\nu_o^-(k)$$

and deduce

$$\begin{aligned}
&\kappa^+(\Gamma)\tilde{m}^{\mathrm{BR}}(\psi^\epsilon * f_{\mathcal{B}^+_{T,\epsilon}})\\
&= \kappa^+(\Gamma)\int_{k'\in K}\int_{KAN} \psi^\epsilon(ka_rnk') f_{\mathcal{B}^+_{T,\epsilon}}(k') e^{-\delta r} dk' dn dr d\nu_o^-(k)\\
&= \kappa^+(\Gamma)\int_{k\in K}\int_G \psi^\epsilon(kg) f_{\mathcal{B}^+_{T,\epsilon}}(\kappa(g)) e^{(-\delta+(n-1))H(g)} dg d\nu_o^-(k)\\
&= \kappa^+(\Gamma)\int_{k\in K}\int_G \psi^\epsilon(g) f_{\mathcal{B}^+_{T,\epsilon}}(\kappa(k^{-1}g)) e^{(-\delta+(n-1))H(k^{-1}g)} dg d\nu_o^-(k)\\
&\le (1+O(\epsilon))\kappa^+(\Gamma)\int_{k\in K}\int_G \psi^\epsilon(g) f_{\mathcal{B}^+_{T,c\epsilon}}(k^{-1}) dg d\nu_o^-(k)\\
&= (1+O(\epsilon))\mathcal{M}_{H\backslash G}(\mathcal{B}^+_{T,c\epsilon}) = (1+O(\epsilon^p))\mathcal{M}_{H\backslash G}(\mathcal{B}_T) \qquad (7.7)
\end{aligned}$$

since $\int \psi^\epsilon dg = 1$ and $\kappa^+(\Gamma)\int_{k\in K} f_{\mathcal{B}_T}(k^{-1}) d\nu_o^-(k) = \mathcal{M}_{H\backslash G}(\mathcal{B}_T)$.

Similarly,

$$\kappa^+(\Gamma)\tilde{m}^{\mathrm{BR}}(\psi^\epsilon * f_{\mathcal{B}^-_{T,c\epsilon}}) = (1+O(\epsilon^p))\mathcal{M}_{H\backslash G}(\mathcal{B}_T).$$

Since $\max_{a_t\in\mathcal{B}_T} e^{(\delta-\beta_1)t} \ll \mathcal{M}_{H\backslash G}(\mathcal{B}_T)^{1-\eta}$ for some $\eta > 0$,

$$\#(\Gamma_0\gamma_0\cap\mathcal{B}_T) = \tfrac{1}{[\Gamma:\Gamma_0]}\mathcal{M}_{H\backslash G}(\mathcal{B}_T) + O(\epsilon^p\mathcal{M}_{H\backslash G}(\mathcal{B}_T)) + O(\epsilon^{-q_\ell}\mathcal{M}_{H\backslash G}(\mathcal{B}_T)^{1-\eta}).$$

Hence by taking $\epsilon = \mathcal{M}_{H\backslash G}(\mathcal{B}_T)^{-\eta/(p+q_\ell)}$ and $\eta_0 = -p\eta/(p+q_\ell)$, we complete the proof. □

7.4. Effectively well-rounded families of $H\backslash G$.

7.4.1. *Sectors.* For $\omega \subset K$, we consider the following sector in $H\backslash G$:

$$S_T(\omega) := [e]\{a_t : 0 \le t \le \log T\}\omega.$$

In this subsection, we show that the family of sectors $\{S_T(\omega) : T \gg 1\}$ is effectively well-rounded provided ω is admissible in the following sense:

Definition 7.8. We will call a Borel subset $\omega \subset K$ with $\nu_o^-(\omega^{-1}) > 0$ *admissible* if there exists $0 < p \le 1$ such that for all small $\epsilon > 0$,

$$\nu_o((\omega^{-1}K_\epsilon - \cap_{k\in K_\epsilon}\omega^{-1}k)) \ll \epsilon^p \tag{7.9}$$

with the implied constant depending only on ω.

Lemma 7.10. *Let $\omega \subset K$ be a Borel subset. If $\nu_o^-(\omega^{-1}) > 0$ and $\partial(\omega^{-1}X_0^-)\cap \Lambda(\Gamma) = \emptyset$, then ω is admissible.*

Proof. As $\partial(\omega^{-1}X_0^-)$ and $\Lambda(\Gamma)$ are compact subsets, we can find $\epsilon_0 > 0$ such that the ϵ_0-neighborhood of $\partial(\omega^{-1}X_0^-)$ is disjoint from $\Lambda(\Gamma)$. Hence we can find $\epsilon_1 > 0$ such that $\partial(\omega^{-1})K_{\epsilon_1}X_0^-$ is disjoint from $\Lambda(\Gamma)$; so $\nu_o(\partial(\omega^{-1})K_{\epsilon_1}X_0^-) = 0$. □

Proposition 7.11. *Let $\kappa_o := \max_{\xi\in\Lambda_p(\Gamma)} \text{rank }(\xi)$. If*

$$\delta > \max\{n-2, \frac{n-2+\kappa_0}{2}\},$$

then any Borel subset $\omega \subset K$ such that $\nu_o^-(\omega^{-1}) > 0$ and $\partial(\omega^{-1})$ is piecewise smooth is admissible.

Proof. Let $s_\xi = \{\xi_t : t \in [0,\infty)\}$ be a geodesic ray emanating from o toward ξ and let $b(\xi_t) \in \overline{\mathbb{H}}^n$ be the Euclidean ball centered at ξ whose boundary is orthogonal to s_ξ at ξ_t. Then by Sullivan [63], there exists a Γ-invariant collection of pairwise disjoint horoballs $\{\mathcal{H}_\xi : \xi \in \Lambda_p(\Gamma)\}$ for which the following holds: there exists a constant $c > 1$ such that for any $\xi \in \Lambda(\Gamma)$ and for any $t > 0$,

$$c^{-1}e^{-\delta t}e^{d(\xi_t,\Gamma(o))(k(\xi_t)-\delta)} \le \nu_o(b(\xi_t)) \le ce^{-\delta t}e^{d(\xi_t,\Gamma(o))(k(\xi_t)-\delta)}$$

where $k(\xi_t)$ is the rank of ξ' if $\xi_t \in \mathcal{H}_{\xi'}$ for some $\xi' \in \Lambda_p(\Gamma)$ and δ otherwise.

Therefore, using $0 \le d(\xi_t, \Gamma(o)) \le t$, it follows that for any $\xi \in \Lambda(\Gamma)$ and $t > 1$,

$$\nu_o(b(\xi_t)) \ll \begin{cases} e^{(-2\delta+k(\xi_t))t} & \text{if } k(\xi_t) \ge \delta \\ e^{-\delta t} & \text{otherwise.} \end{cases} \tag{7.12}$$

By standard computations in hyperbolic geometry, there exists $c_0 > 1$ such that $B(\xi, c_0^{-1}e^{-t}) \subset b(\xi_t) \subset B(\xi, c_0e^{-t})$ where $B(\xi, r)$ denotes the Euclidean ball in $\partial(\mathbb{H}^n)$ of radius r. Hence it follows from (7.12) that if we set $\kappa_0 := \max_{\xi'\in\Lambda_p(\Gamma)} \text{rank }(\xi')$, then for all small $\epsilon > 0$ and $\xi \in \Lambda(\Gamma)$,

$$\nu_o(B(\xi,\epsilon)) \ll \epsilon^\delta + \epsilon^{2\delta-\kappa_0}.$$

Clearly, this inequality holds for all $\xi \in \partial(\mathbb{H}^n)$, as the support of ν_o is equal to $\Lambda(\Gamma)$.

Now if $\partial(\omega^{-1})$ is a piece-wise smooth subset of K, we can cover its ϵ-neighborhood by $O(\epsilon^{1-d_K})$ number of ϵ-balls, where d_K is the dimension of K.

Since for any $k \in K$,

$$\nu_o^+(B(k,\epsilon)) \ll \epsilon^{d_M} \cdot \nu_o(B(k(X_0^+),\epsilon)) \ll \epsilon^{\delta+d_M} + \epsilon^{2\delta-\kappa_0+d_M};$$

where d_M is the dimension of M, we obtain that the ν_o measure of an ϵ-neighborhood of $\partial(\omega^{-1})$ is at most of order

$$\epsilon^{\delta+d_M-d_K+1} + \epsilon^{2\delta-\kappa_0+d_M-d_K+1} = \epsilon^{\delta-n+2} + \epsilon^{2\delta-\kappa_0-n+2}.$$

Hence ω is admissible if δ is bigger than both $(n-2)$ and $\frac{n-2+\kappa_0}{2}$. □

Corollary 7.13. *If $\delta > n-2$ and rank $(\xi) < \delta$ for all $\xi \in \Lambda_p(\Gamma)$, then any Borel subset $\omega \subset K$ such that $\nu_o^-(\omega^{-1}) > 0$ and $\partial(\omega^{-1})$ is piece-wise smooth is admissible.*

The following strong wave front property of HAK decomposition is a crucial ingredient in proving an effective well-roundedness of a given family:

Lemma 7.14 (Strong wave front property)**.** [24, Theorem 4.1] *There exists $c > 1$ and $\epsilon_0 > 0$ such that for any $0 < \epsilon < \epsilon_0$ and for any $g = ha_tk \in HA^+K$ with $t > 1$,*

$$gG_\epsilon \subset (hH_{c\epsilon})\,(a_tA_{c\epsilon})\,(kK_{c\epsilon})$$

where $H_{c\epsilon} = H \cap G_{c\epsilon}$ and $A_{c\epsilon}$ and $K_{c\epsilon}$ are defined similarly.

Proposition 7.15. *Let $\omega \subset K$ be an admissible subset. Then the family $\{S_T(\omega) : T \gg 1\}$ is effectively well-rounded and*

$$\mathcal{M}_{H\backslash G}(S_T(\omega)) = \tfrac{|\mu_H^{\mathrm{PS}}|\cdot\nu_o^-(\omega^{-1})}{\delta\cdot|m^{\mathrm{BMS}}|}(T^\delta - 1).$$

Proof. We compute

$$\begin{aligned}\mathcal{M}_{H\backslash G}(S_T(\omega)) &= \tfrac{|\mu_H^{\mathrm{PS}}|}{|m^{\mathrm{BMS}}|}\int_{t=0}^{\log T} e^{\delta t}dt \int_{k\in\omega} d\nu_o^-(k^{-1}) \\ &= \tfrac{|\mu_H^{\mathrm{PS}}|\cdot\nu_o^-(\omega^{-1})}{\delta\cdot|m^{\mathrm{BMS}}|}(T^\delta - 1).\end{aligned}$$

By Lemma 7.14, there exists $c \geq 1$ such that for all $T > 1$ and $\epsilon > 0$

$$S_T(\omega)G_\epsilon \subset [e]\{a_t : \log(1-c\epsilon) \leq t \leq \log(1+c\epsilon)T\}\omega_{c\epsilon}^+$$

where $\omega_{c\epsilon}^+ = \omega K_{c\epsilon}$ and $K_{c\epsilon}$ is a $c\epsilon$-neighborhood of e in K. Hence with $p > 0$ given in (7.9),

$$\begin{aligned}\mathcal{M}_{H\backslash G}(S_T(\omega)G_\epsilon) &\ll \tfrac{|\mu_H^{\mathrm{PS}}|\cdot\nu_o^-((\omega_{c\epsilon}^+)^{-1})}{\delta\cdot|m^{\mathrm{BMS}}|}(1+c\epsilon)^\delta T^\delta \\ &\ll \tfrac{|\mu_H^{\mathrm{PS}}|\cdot(1+O(\epsilon^p))\nu_o^-(\omega^{-1})}{\delta\cdot|m^{\mathrm{BMS}}|}(1+c\epsilon)^\delta T^\delta \\ &= (1+O(\epsilon^p))\mathcal{M}_{H\backslash G}(S_T(\omega)).\end{aligned}$$

Similarly, we can show that

$$\mathcal{M}_{H\backslash G}(\cap_{g\in G_\epsilon} S_T(\omega)g) = (1+O(\epsilon^p))\mathcal{M}_{H\backslash G}(S_T(\omega)).$$

Hence the family $\{S_T(\omega)\}$ is an effectively well-rounded family for Γ. $\square$

Therefore we deduce from Theorem 7.6:

Corollary 7.16. *Let $\omega \subset K$ be an admissible subset. Then there exists $\eta_0 > 0$ (depending only on a uniform spectral gap data for Γ and Γ_0) such that for any $\gamma_0 \in \Gamma$*

$$\#([e]\Gamma_0\gamma_0 \cap S_T(\omega)) = \tfrac{|\mu_H^{\mathrm{PS}}|\cdot \nu_o^-(\omega^{-1})}{[\Gamma:\Gamma_0]\cdot|m^{\mathrm{BMS}}|\cdot\delta} T^\delta + O(T^{\delta-\eta_0}).$$

7.4.2. *Counting in norm-balls.* Let V be a finite dimensional vector space on which G acts linearly from the right and let $w_0 \in V$. We assume that $w_0\Gamma$ is discrete and that $H := G_{w_0}$ is either a symmetric subgroup or a horospherical subgroup. We let $A = \{a_t\}, K, M$ be as in section 5.3. Let $\lambda \in \mathbb{N}$ be the log of the largest eigenvalue of a_1 on the $\mathbb{R}$-span of w_0G, and set

$$w_0^{\pm\lambda} := \lim_{t\to\infty} e^{-\lambda t} w_0 a_{\pm t}.$$

Fixing a norm $\|\cdot\|$ on V, let $B_T := \{v \in w_0G : \|v\| \le T\}$.

Proposition 7.17. *For any admissible $\omega \subset K$, the family $\{B_T \cap w_0A^\pm\omega\}$ is effectively well-rounded. In particular, $\{B_T\}$ is effectively well-rounded.*

We also compute that for some $0 < \eta < \delta/\lambda$,

$$\mathcal{M}_{H\backslash G}(B_T \cap w_0A^\pm\omega) = \tfrac{|\mu_{H,\pm}^{\mathrm{PS}}|}{\delta\cdot|m^{\mathrm{BMS}}|} \cdot \int_\omega \|w_0^{\pm\lambda}k\|^{-\delta/\lambda} d\nu_o^\pm(k^{-1}) \cdot T^{\delta/\lambda} + O(T^\eta).$$

Proof. By the definition of λ and w_0^λ, it follows that $w_0a_tk = e^{\lambda t}w_0^\lambda k + O(e^{\lambda_1 t})$ for some $\lambda_1 < \lambda$. Noting that $\|w_0a_tk\| \le T$ implies that $e^{\lambda t} = O(T)$ and $e^{\lambda_1 t} = O(T^{\lambda_1/\lambda})$, we have

$$\begin{aligned}
&\mathcal{M}_{H\backslash G}(B_T \cap w_0A^+\omega)\\
&= \tfrac{|\mu_H^{\mathrm{PS}}|}{|m^{\mathrm{BMS}}|} \int_{k\in\omega} \int_{\|w_0a_tk\|\le T} e^{\delta t} dt d\nu_o^-(k^{-1})\\
&= \tfrac{|\mu_H^{\mathrm{PS}}|}{|m^{\mathrm{BMS}}|} \int_{k\in\omega} \int_{e^{\lambda t}\le \|w_0^\lambda k\|^{-1}T + O(T^{\lambda_1/\lambda})} e^{\delta t} dt d\nu_o^-(k^{-1})\\
&= \tfrac{|\mu_H^{\mathrm{PS}}|}{|m^{\mathrm{BMS}}|\cdot\delta} T^{\delta/\lambda} \int_{k\in\omega} \|w_0^\lambda k\|^{-\delta/\lambda} d\nu_o^-(k^{-1}) + O(T^\eta)
\end{aligned}$$

for some $\eta < \delta/\lambda$. The claim about $\mathcal{M}_{H\backslash G}(B_T \cap w_0A^-\omega)$ can be proven similarly. To show the effective well-roundedness, we first note that by Lemma 7.14, for some $c > 1$, we have

$$(B_T \cap w_0A^+\omega)G_\epsilon \subset B_{(1+c\epsilon)T} \cap w_0A^+\omega^+_{c\epsilon}.$$

Therefore, using the admissibility of ω, and with p given in (7.9), we deduce

$$\begin{aligned}
&\mathcal{M}_{H\backslash G}((B_T \cap w_0 A^+\omega)G_\epsilon - (B_T \cap w_0 A^+\omega)) \\
&\ll \int_{k\in\omega^+_{c\epsilon}-\omega} \int_{\|w_0 a_t k\|\le (1+c\epsilon)T} e^{\delta t} dt d\nu_o^-(k^{-1}) \\
&+ \int_{k\in\omega^+_{c\epsilon}} \int_{T\le\|w_0 a_t k\|\le (1+c\epsilon)T} e^{\delta t} dt d\nu_o^-(k^{-1}) \\
&\ll \epsilon^p \cdot T^{\delta/\lambda} + ((1+c\epsilon)T)^{\delta/\lambda} - T^{\delta/\lambda}) \\
&\ll \epsilon^p T^{\delta/\lambda} \ll \epsilon^p \mathcal{M}_{H\backslash G}(B_T \cap w_0 A^+\omega).
\end{aligned}$$

Similarly we can show that

$$\mathcal{M}_{H\backslash G}((B_T \cap w_0 A^+\omega) - \cap_{g\in G_\epsilon}(B_T \cap w_0 A^+\omega)g) \ll \epsilon^p \mathcal{M}_{H\backslash G}(B_T \cap w_0 A^+\omega).$$

This finishes the proof for the effective well-roundedness of $\{B_T \cap w_0 A^+\omega\}$. The claims about $\{B_T \cap w_0 A^-\omega\}$ can be shown in a similar fashion. □

Put

$$\Xi_{w_0}(\Gamma) := \begin{cases} \frac{|\mu_H^{\mathrm{PS}}|}{\delta\cdot|m^{\mathrm{BMS}}|} \cdot \int_K \|w_0^\lambda k\|^{-\delta/\lambda} d\nu_o^-(k^{-1}) & \text{if } G = HA^+K \\ \sum \frac{|\mu_{H^\pm}^{\mathrm{PS}}|}{\delta\cdot|m^{\mathrm{BMS}}|} \cdot \int_K \|w_0^{\pm\lambda} k\|^{-\delta/\lambda} d\nu_o^\mp(k^{-1}) & \text{otherwise.} \end{cases}$$

We deduce the following from Proposition 7.17 and Theorem 7.6:

Corollary 7.18. (1) *For any admissible $\omega \subset K$, there exists $\eta_0 > 0$ such that for any $\gamma_0 \in \Gamma$,*

$$\begin{aligned}
&\#\{v \in w_0\Gamma_0\gamma_0 \cap w_0 A^+\omega : \|v\| \le T\} \\
&\qquad = \tfrac{|\mu_H^{\mathrm{PS}}|}{\delta\cdot[\Gamma:\Gamma_0]\cdot|m^{\mathrm{BMS}}|} \cdot \int_\omega \|w_0^\lambda k\|^{-\delta/\lambda} d\nu_o^-(k^{-1}) T^{\delta/\lambda} + O(T^{\delta/\lambda-\eta_0}).
\end{aligned}$$

(2) *There exists $\eta_0 > 0$ (depending only on a uniform spectral gap data for Γ and Γ_0) such that for any $\gamma_0 \in \Gamma$,*

$$\#\{v \in w_0\Gamma_0\gamma_0 : \|v\| \le T\} = \tfrac{1}{[\Gamma:\Gamma_0]}\Xi_{w_0}(\Gamma) T^{\delta/\lambda} + O(T^{\delta/\lambda-\eta_0}).$$

7.5. The case when H is trivial. In this subsection, we will prove the following theorem directly from the asymptotic of the matrix coefficient functions in Theorem 3.30.

Recall from the introduction the following Borel measure $\mathcal{M}_G = \mathcal{M}_G^\Gamma$ on G: for $\psi \in C_c(G)$,

$$\mathcal{M}_G(\psi) = \tfrac{1}{|m^{\mathrm{BMS}}|} \int_{k_1 a_t k_2 \in KA^+K} \psi(k_1 a_t k_2) e^{\delta t} d\nu_o^+(k_1) dt d\nu_o^-(k_2^{-1}).$$

Theorem 7.19. *Let $\Gamma_0 < \Gamma$ be a subgroup of finite index. If $\{\mathcal{B}_T \subset G\}$ is effectively well-rounded with respect to Γ (see Def. 1.10), then there exists*

$\eta_0 > 0$ (depending only on a uniform spectral gap data for Γ and Γ_0) such that for any $\gamma_0 \in \Gamma$

$$\#(\Gamma_0\gamma_0 \cap \mathcal{B}_T) = \tfrac{1}{[\Gamma:\Gamma_0]}\mathcal{M}_G(\mathcal{B}_T) + O(\mathcal{M}_G(\mathcal{B}_T)^{1-\eta_0})$$

with the implied constant independent of Γ_0 and $\gamma_0 \in \Gamma$.

Consider the following function on $\Gamma_0\backslash G \times \Gamma_0\backslash G$: for a compact subset $\mathcal{B} \subset G$,

$$F_{\mathcal{B}}(g,h) := \sum_{\gamma\in\Gamma_0} \chi_{\mathcal{B}}(g^{-1}\gamma h)$$

where $\chi_{\mathcal{B}}$ is the characteristic function of $\mathcal{B}$. We set $F_T := F_{\mathcal{B}_T}$ for simplicity. Observe that $F_T(e,\gamma_0) = \#(\Gamma_0\gamma_0 \cap \mathcal{B}_T)$. Let $\mathcal{B}^{\pm}_{T,\epsilon}$ be as in the definition 1.10 and let $\phi^\epsilon \in C^\infty(G)$ and $\Phi^\epsilon \in C^\infty(\Gamma_0\backslash G)$ be as in the proof of Theorem 7.6. We then have

$$\langle F_{\mathcal{B}^-_{T,\epsilon}}, \Phi^\epsilon \otimes \Phi^\epsilon_{\gamma_0^{-1}}\rangle \le F_T(e,\gamma_0) \le \langle F_{\mathcal{B}^+_{T,\epsilon}}, \Phi^\epsilon \otimes \Phi^\epsilon_{\gamma_0^{-1}}\rangle.$$

Note that for $\Psi_1, \Psi_2 \in C_c(\Gamma_0\backslash G)$

$$\langle F_T, \Psi_1 \otimes \Psi_2\rangle_{\Gamma_0\backslash G\times\Gamma_0\backslash G} = \int_{g\in\mathcal{B}_T} \langle \Psi_1, g.\Psi_2\rangle_{L^2(\Gamma_0\backslash G)}\, dm^{\mathrm{Haar}}(g).$$

For a Borel subset $\mathcal{B}$ of G, consider a function $f_{\mathcal{B}}$ on $K \times K$ given by

$$f_{\mathcal{B}}(k_1,k_2) = \int_{a_t \in k_1^{-1}\mathcal{B}k_2^{-1}\cap A^+} e^{\delta t} dt,$$

and define a function on $G \times G$ by

$$((\psi^\epsilon \otimes \psi^\epsilon) * f_{\mathcal{B}})(g,h) = \int_{K\times K} \psi^\epsilon(gk_1^{-1})\psi^\epsilon(hk_2) f_{\mathcal{B}}(k_1,k_2) dk_1 dk_2.$$

We deduce by applying Theorem 1.4 and using the left Γ-invariance of the measures $\tilde{m}^{\mathrm{BR}}$ and $\tilde{m}^{\mathrm{BR}}$ that for some $\eta', \eta > 0$,

$$\begin{aligned}
&\langle F_{\mathcal{B}}, \Psi^\epsilon \otimes \Psi^\epsilon_{\gamma_0^{-1}}\rangle_{\Gamma_0\backslash G\times\Gamma_0\backslash G}\\
&= \int_{x\in\mathcal{B}} \int_{\Gamma_0\backslash G} \Psi^\epsilon(g)\Psi^\epsilon_{\gamma_0^{-1}}(gx) dm^{\mathrm{Haar}}_{\Gamma_0}(g) dx\\
&= \int_{k_1 a_t k_2 \in \mathcal{B}} \left(\int_{\Gamma_0\backslash G} \Psi^\epsilon(gk_1^{-1})\Psi^\epsilon_{\gamma_0^{-1}}(ga_t k_2) dm^{\mathrm{Haar}}_{\Gamma_0}(g)\right) e^{(n-1)t}(1+O(e^{-\eta' t})) dt dk_1 dk_2\\
&= \frac{1}{|m^{\mathrm{BMS}}_{\Gamma_0}|} \int_{k_1 a_t k_2\in\mathcal{B}} e^{\delta t}(1+O(e^{-\eta t})) m^{\mathrm{BR}}_{\Gamma_0}(k_2\Psi^\epsilon_{\gamma_0^{-1}}) m^{\mathrm{BR}}_{*,\Gamma_0}(k_1^{-1}\Psi^\epsilon) dt dk_1 dk_2\\
&= \frac{1}{|m^{\mathrm{BMS}}_{\Gamma_0}|} \int_{k_1 a_t k_2\in\mathcal{B}} e^{\delta t}(1+O(e^{-\eta t})) \tilde{m}^{\mathrm{BR}}(k_2\psi^\epsilon) \tilde{m}^{\mathrm{BR}}_*(k_1^{-1}\psi^\epsilon) dt dk_1 dk_2.
\end{aligned}$$

Therefore

$$\begin{aligned}&\langle F_{\mathcal{B}^{\pm}_{T,\epsilon}}, \Phi^{\epsilon} \otimes \Phi^{\epsilon}_{\gamma_0^{-1}} \rangle_{\Gamma_0\backslash G\times\Gamma_0\backslash G} \\ &= \tfrac{1}{|m^{\mathrm{BMS}}_{\Gamma_0}|}(\tilde{m}^{\mathrm{BR}_*}\otimes\tilde{m}^{\mathrm{BR}})((\psi^{\epsilon}\otimes\psi^{\epsilon})*f_{\mathcal{B}^{\pm}_{T,\epsilon}}) + O(\max_{a_t\in\mathcal{B}_T} e^{(\delta-\eta)t}\epsilon^{q_\ell}).\end{aligned} \tag{7.20}$$

Recall

$$dm^{\mathrm{BR}}(ka_rn^+) = e^{-\delta r}dn^+drd\nu_o^-(k) \quad \text{for } ka_rn^+ \in KAN^+;$$

$$dm^{\mathrm{BR}_*}(ka_rn^-) = e^{\delta r}dn^-drd\nu_o^+(k) \quad \text{for } ka_rn^- \in KAN^-$$

and $dg = d\tilde{m}^{\mathrm{Haar}}(a_rn^{\pm}k) = drdn^{\pm}dk$.

For $x \in G$, let $\kappa^{\pm}(x)$ denote the K-component of x in $AN^{\pm}K$ decomposition and let $H^{\pm}(x)$ be uniquely given by the requirement $x \in e^{H^{\pm}(x)}N^{\pm}K$.

We obtain

$$(\tilde{m}^{\mathrm{BR}_*}\otimes\tilde{m}^{\mathrm{BR}})((\psi^{\epsilon}\otimes\psi^{\epsilon})*f_{\mathcal{B}}) =$$
$$\int_{K\times K}\int_{G\times G}\psi^{\epsilon}(g_1k_1^{-1})\psi^{\epsilon}(h_1k_2)f_{\mathcal{B}}(k_1,k_2)d\tilde{m}^{\mathrm{BR}_*}(g_1)d\tilde{m}^{\mathrm{BR}}(h_1)dk_1dk_2 =$$
$$\int_{K\times K}\int_{G\times G}\psi^{\epsilon}(kg)\psi^{\epsilon}(k_0h)f_{\mathcal{B}}(\kappa^-(g)^{-1},\kappa^+(h))e^{(\delta-n+1)(H^-(g)-H^+(h))}$$
$$dgdhd\nu_o^-(k_0)d\nu_o^+(k) =$$
$$\int_{K\times K}\int_{G\times G}\psi^{\epsilon}(g)\psi^{\epsilon}(h)f_{\mathcal{B}}(\kappa^-(gk^{-1})^{-1},\kappa^+(hk_0^{-1}))e^{(\delta-n+1)(H^-(gk^{-1})-H^+(hk_0^{-1}))}$$
$$dgdhd\nu_o^-(k_0)d\nu_o^+(k);$$

first replacing k_1 with k_1^{-1}, substituting $g_1 = ka_rn \in KAN^+$ and $h_1 = k_0a_{r_0}n_0 \in KAN^-$ and again substituting $a_rnk_1 = g$ and $a_{r_0}n_0k_2 = h$.

Therefore, using the strong wave front property for $AN^{\pm}K$ decompositions [24] and the assumption that $\int\psi^{\epsilon}dg = 1$, we have, for some $p > 0$,

$$\begin{aligned}&(\tilde{m}^{\mathrm{BR}_*}\otimes\tilde{m}^{\mathrm{BR}})((\psi^{\epsilon}\otimes\psi^{\epsilon})*f_{\mathcal{B}^{\pm}_{T,\epsilon}}) \\ &= (1+O(\epsilon^p))\int_{K\times K} f_{\mathcal{B}_T}(k,k_0^{-1})d\nu_o^+(k)d\nu_o^-(k_0) \\ &= (1+O(\epsilon^p))\int_{ka_tk_0^{-1}\in\mathcal{B}_T} e^{\delta t}d\nu_o^+(k)d\nu_o^-(k_0) \\ &= (1+O(\epsilon^p))\int_{ka_tk_0\in\mathcal{B}_T} e^{\delta t}d\nu_o^+(k)d\nu_o^-(k_0^{-1}) \\ &= (1+O(\epsilon^p))\mathcal{M}_G(\mathcal{B}_T)\end{aligned}$$

with $\mathcal{M}_G$ defined as in Definition 1.8. Since $|m^{\mathrm{BMS}}_{\Gamma_0}| = |m^{\mathrm{BMS}}_{\Gamma}|\cdot[\Gamma:\Gamma_0]$, putting the above together, we get

$$F_T(e,\gamma_0) = \frac{1}{[\Gamma:\Gamma_0]}\mathcal{M}_G(\mathcal{B}_T) + O(\mathcal{M}_G(\mathcal{B}_T)^{1-\eta_0})$$

for some $\eta_0 > 0$ depending only on a uniform spectral gap data of Γ and Γ_0. This proves Theorem 7.19.

Corollary 7.21. *Let $\omega_1, \omega_2 \subset K$ be Borel subsets in K such that ω_1^{-1} and ω_2 are admissible in the sense of (7.8). Set $S_T(\omega_1, \omega_2) := \omega_1\{a_t : 0 < t < \log T\}\omega_2$. Then the family $\{S_T(\omega_1, \omega_2) : T \gg 1\}$ is effectively well-rounded with respect to Γ, and for some $\eta_0 > 0$,*

$$\#(\Gamma_0\gamma_0 \cap S_T(\omega_1, \omega_2)) = \frac{\nu_o^+(\omega_1) \cdot \nu_o^-(\omega_2^{-1})}{\delta \cdot |m^{\mathrm{BMS}}| \cdot [\Gamma : \Gamma_0]} T^\delta + O(T^{\delta - \eta_0})$$

with the implied constant independent of Γ_0 and $\gamma_0 \in \Gamma$.

Using Proposition 7.14 for $H = K$, we can prove the effective well-roundedness of $\{S_T(\omega_1, \omega_2) : T \gg 1\}$ with respect to Γ in a similar fashion to the proof of Proposition 7.15. Hence Corollary 7.21 follows from Theorem 7.19; we refer to Lemma 7.10 and Proposition 7.11 for admissible subsets of K.

7.6. Counting in bisectors of HA^+K coordinates. We state a counting result for bisectors in HA^+K coordinates.

Let $\tau_1 \in C_c^\infty(H)$ with its support being injective to $\Gamma\backslash G$ and $\tau_2 \in C^\infty(K)$, and define $\xi_T \in C^\infty(G)$ as follows: for $g = hak \in HA^+K$,

$$\xi_T(g) = \chi_{A_T^+}(a) \cdot \int_{H \cap M} \tau_1(hm)\tau_2(m^{-1}k)dm$$

where $\chi_{A_T^+}$ denotes the characteristic function of $A_T^+ = \{a_t : 0 < t < \log T\}$ for $T > 1$. Since if $hak = h'ak'$, then $h = h'm$ and $k = m^{-1}k'$ for some $m \in H \cap M$, the above function is well-defined.

Theorem 7.22. *Let $\Gamma_0 < \Gamma$ be a subgroup of finite index. There exist $\eta_0 > 0$ (depending only on a uniform spectral gap data for Γ and Γ_0) and $\ell \in \mathbb{N}$ such that for any $\gamma_0 \in \Gamma$,*

$$\sum_{\gamma \in \Gamma_0} \xi_T(\gamma\gamma_0) = \frac{\tilde{\mu}_H^{\mathrm{PS}}(\tau_1) \cdot \nu_o^*(\tau_2)}{\delta \cdot |m^{\mathrm{BMS}}| \cdot [\Gamma : \Gamma_0]} T^\delta + O(T^{\delta - \eta_0} \mathcal{S}_\ell(\tau_1)\mathcal{S}_\ell(\tau_2))$$

where $\nu_o^*(\tau_2) := \int_K \tau_2(k) d\nu_o^-(k^{-1})$.

Proof. Define a function F_T on $\Gamma_0\backslash G$ by

$$F_T(g) = \sum_{\gamma \in \Gamma_0} \xi_T(\gamma g).$$

For any $\psi \in C_c^\infty(G)$, set $\Psi \in C_c^\infty(\Gamma_0\backslash G)$ to be $\Psi(g) = \sum_{\gamma \in \Gamma_0} \psi(\gamma g)$ and then we have:

$$\langle F_T, \Psi \rangle_{\Gamma_0\backslash G} = \int_{k \in K} \tau_2(k) \int_{a_t \in A_T^+} \left(\int_{h \in H} \tau_1(h)\Psi(ha_tk)dh \right) \rho(t)dkdt.$$

As $\Psi \in C(\Gamma_0\backslash G)$, $\mathrm{supp}(\tau_1)$ injects to $\Gamma_0\backslash G$ and $H \cap \Gamma = H \cap \Gamma_0$, we have $\mu^{\mathrm{PS}}_{\Gamma_0,H}(\tau_1) = \tilde{\mu}^{\mathrm{PS}}_H(\tau_1)$ and

$$\int_{h\in H} \tau_1(h)\Psi(ha_tk)dh = \int_{h\in \Gamma_0\backslash \Gamma_0 H} \tau_1(h)\Psi(ha_tk)dh.$$

Therefore, by applying Theorem 5.13 to the inner integral, we obtain $\eta > 0$ and $\ell \in \mathbb{N}$ such that

$$\begin{aligned}
&\langle F_T, \Psi\rangle_{\Gamma_0\backslash G}\\
&= \frac{\tilde{\mu}^{\mathrm{PS}}_H(\tau_1)}{|m^{\mathrm{BMS}}_{\Gamma_0}|}\int_{k\in K}\int_{a_t\in A^+_T} \tau_2(k) m^{\mathrm{BR}}_{\Gamma_0}(\Psi_k)e^{\delta t}dkdt + O(\mathcal{S}_\ell(\tau_1)\mathcal{S}_\ell(\psi)T^{\delta-\eta})\\
&= \frac{\tilde{\mu}^{\mathrm{PS}}_H(\tau_1)\cdot \tilde{m}^{\mathrm{BR}}(\psi * \tau_2)}{\delta\cdot |m^{\mathrm{BMS}}|\cdot[\Gamma:\Gamma_0]}\cdot T^\delta + O(\mathcal{S}_\ell(\tau_1)\mathcal{S}_\ell(\tau_2)\mathcal{S}_\ell(\psi)T^{\delta-\eta}). \qquad (7.23)
\end{aligned}$$

Let $\tau_i^{\epsilon,\pm}$ be ϵ-approximations of τ_i; $\tau_i^{\epsilon,\pm}(x)$ are respectively the supremum and the infimum of τ_i in the ϵ-neighborhood of x. Then for a suitable $\ell \geq 1$, $\tilde{\mu}^{\mathrm{PS}}_H(\tau_1^{\epsilon,+} - \tau_1^{\epsilon,-}) = O(\epsilon\cdot \mathcal{S}_\ell(\tau_1))$, and $\nu_o^-(\tau_2^{\epsilon,+} - \tau_2^{\epsilon,-}) = O(\epsilon\cdot\mathcal{S}_\ell(\tau_2))$.

Let $F_T^{\epsilon,\pm}$ be a function on $\Gamma\backslash G$ defined similarly as F_T, with respect to $\xi_T^{\epsilon,\pm}(hak) = \chi_{A^+_{(1\pm\epsilon)T}}(a)\cdot \int_{H\cap M}\tau_1^{\epsilon,+}(hm)\tau_2^{\epsilon,+}(m^{-1}k)dm$.

As before, let $\psi^\epsilon \in C^\infty(G)$ be an ϵ smooth approximation of e: $0 \leq \psi^\epsilon \leq 1$, $\mathrm{supp}(\psi^\epsilon) \subset G_\epsilon$ and $\int \psi^\epsilon dg = 1$. Let $\Psi^\epsilon_{\gamma_0^{-1}}$ be defined as in the subsection 7.2 with respect to ψ^ϵ. Lemma 7.14 implies that there exists $c > 0$ such that for all $g \in G_{c\epsilon}$,

$$F_T^{\epsilon,-}(\gamma_0 g) \leq F_T(\gamma_0) \leq F_T^{\epsilon,+}(\gamma_0 g)$$

and hence

$$\langle F_T^{\epsilon,-}, \Psi^\epsilon_{\gamma_0^{-1}}\rangle \leq F_T(\gamma_0) \leq \langle F_T^{\epsilon,+}, \Psi^\epsilon_{\gamma_0^{-1}}\rangle. \qquad (7.24)$$

By a similar computation as in the proof of Theorem 7.6 (cf. [52, proof of Prop. 7.5]), we have $\tilde{m}^{\mathrm{BR}}(\psi^\epsilon * \tau_2) = \nu_o^*(\tau_2) + O(\epsilon)\mathcal{S}_\ell(\tau_2)$.

Therefore, for q_ℓ given by $\mathcal{S}_\ell(\psi^\epsilon) = O(\epsilon^{-q_\ell})$, we deduce from (7.23) and (7.24) that, using the left Γ-invariance of the measure $\tilde{m}^{\mathrm{BR}}$,

$$\begin{aligned}
&\delta\cdot|m^{\mathrm{BMS}}|\cdot[\Gamma:\Gamma_0]\cdot F_T(\gamma_0)\\
&= \tilde{\mu}^{\mathrm{PS}}_H(\tau_1)\cdot\tilde{m}^{\mathrm{BR}}(\psi^\epsilon * \tau_2)\cdot T^\delta + O(\mathcal{S}_\ell(\tau_1)S_\ell(\tau_2)\mathcal{S}_\ell(\psi^\epsilon)T^{\delta-\eta})\\
&= \tilde{\mu}^{\mathrm{PS}}_H(\tau_1)\nu_o^*(\tau_2)T^\delta + O(\epsilon T^\delta + \epsilon^{-q_\ell}T^{\delta-\eta})\mathcal{S}_\ell(\tau_1)\mathcal{S}_\ell(\tau_2)\\
&= \tilde{\mu}^{\mathrm{PS}}_H(\tau_1)\nu_o^*(\tau_2)T^\delta + O(T^{\delta-\eta_0})\mathcal{S}_\ell(\tau_1)\mathcal{S}_\ell(\tau_2)
\end{aligned}$$

for some $\eta_0 > 0$, by taking $\epsilon = T^{-\eta/(1+q_\ell)}$. □

Corollary 7.21 as well as its analogues in the HAK decomposition can be deduced easily from Theorem 7.22 by approximation admissible sets by smooth functions.

8. Affine sieve

In this final section, we prove Theorems 1.16 and 1.17. We begin by recalling the combinatorial sieve (see [25, Theorem 7.4]).

Let $\mathsf{A} = \{a_n\}$ be a sequence of non-negative numbers and let B be a finite set of primes. For $z > 1$, let P be the product of primes $P = \prod_{p \notin B, p<z} p$. We set

$$S(\mathsf{A}, P) := \sum_{(n,P)=1} a_n.$$

To estimate $S(\mathsf{A}, P)$, we need to understand how A is distributed along arithmetic progressions. For d square-free, define

$$\mathsf{A}_d := \{a_n \in \mathsf{A} : n \equiv 0(d)\}$$

and set $|\mathsf{A}_d| := \sum_{n \equiv 0(d)} a_n$.

We will use the following combinatorial sieve:

Theorem 8.1. (A_1) *For d square-free with no factors in B, suppose that*

$$|\mathsf{A}_d| = g(d)\mathcal{X} + r_d(\mathsf{A})$$

where g is a function on square-free integers with $0 \le g(p) < 1$, g is multiplicative outside B, i.e., $g(d_1 d_2) = g(d_1)g(d_2)$ if d_1 and d_2 are square-free integers with $(d_1, d_2) = 1$ and $(d_1 d_2, B) = 1$, and for some $c_1 > 0$, $g(p) < 1 - 1/c_1$ for all prime $p \notin B$.

(A_2) A *has level distribution $D(\mathcal{X})$, in the sense that for some $\epsilon > 0$ and $C_\epsilon > 0$,*

$$\sum_{d<D} |r_d(\mathsf{A})| \le C_\epsilon \mathcal{X}^{1-\epsilon}.$$

(A_3) A *has sieve dimension r in the sense that there exists $c_2 > 0$ such that for all $2 \le w \le z$,*

$$-c_2 \le \sum_{(p,B)=1, w \le p \le z} g(p) \log p - r \log \frac{z}{w} \le c_2.$$

Then for $s > 9r$, $z = D^{1/s}$ and $\mathcal{X}$ large enough,

$$S(\mathsf{A}, P) \asymp \frac{\mathcal{X}}{(\log \mathcal{X})^r}.$$

Let $\mathbf{G}$, G $V = \mathbb{C}^m$, Γ, $w_0 \in V(\mathbb{Z})$, etc., be as in Theorem 1.16. We consider the spin cover $\tilde{\mathbf{G}} \to \mathbf{G}$. Noting that the image of $\tilde{\mathbf{G}}(\mathbb{R})$ is precisely $G = \mathbf{G}(\mathbb{R})^\circ$, we replace Γ by its preimage under the spin cover. This does not affect the orbit $w_0\Gamma$ and all our counting statements hold equally. Set $W := w_0\mathbf{G}$ (resp. $w_0\mathbf{G} \cup \{0\}$) if $w_0\mathbf{G}$ (resp. $w_0\mathbf{G} \cup \{0\}$) is Zariski closed,

Let $F \in \mathbb{Q}[W]$ be an integer-valued polynomial on $w_0\Gamma$ and let $F = F_1 \cdots F_r$ where $F_i \in \mathbb{Q}[W]$ are all irreducible also in $\mathbb{C}[W]$ and integral on the orbit $w_0\Gamma$. We may assume without loss of generality that $\gcd\{F(x) : x \in w_0\Gamma\} = 1$, by replacing F by $m^{-1}F$ for $m := \gcd\{F(x) : x \in w_0\Gamma\}$.

Let $\{\mathcal{B}_T \subset w_0 G\}$ be an effectively well-rounded family of subsets with respect to Γ. Set $\mathcal{O} := w_0\Gamma$. For $n \in \mathbb{N}$, $d \in \mathbb{N}$, and $T > 1$, we also set

$$a_n(T) := \#\{x \in \mathcal{O} \cap \mathcal{B}_T : F(x) = n\};$$

$$\Gamma_{w_0}(d) := \{\gamma \in \Gamma : w_0\gamma \equiv w_0 \ (d)\},$$

$$|\mathsf{A}(T)| := \sum_n a_n(T) = \#\mathcal{O} \cap \mathcal{B}_T;$$

$$|\mathsf{A}_d(T)| := \sum_{n \equiv 0(d)} a_n(T) = \#\{x \in \mathcal{O} \cap \mathcal{B}_T : F(x) \equiv 0 \ (d)\}.$$

Let $\Gamma_d := \{\gamma \in \Gamma : \gamma \equiv e \ \ (d)\}$.

Theorem 8.2. *If $\delta > n-2$, then there exists a finite set S of primes such that the family $\{\Gamma_d : d$ is square-free with no factors in $S\}$ has a uniform spectral gap.*

Proof. As $\delta > (n-1)/2$, by [58] and by the transfer property obtained in [6], there exists a finite set S of primes such that the family $L^2(\Gamma_d\backslash\mathbb{H}^n)$ has a uniform spectral gap where d runs over all square-free integers with no prime factors in S, that is, there exists $s_1 < \delta$ such that $L^2(\Gamma_d\backslash G)$ does not contain a spherical complementary series representation of parameter $s_1 < s < \delta$. By Theorem 3.27 and the classification of $\hat{G}$ [30], $L^2(\Gamma_d\backslash G)$ does not contain a non-spherical complementary series representation of parameter $s > (n-2)$.

It follows that $L^2(\Gamma_d\backslash G) = \mathcal{H}_\delta \oplus \mathcal{W}_d$ where $\mathcal{H}_\delta = U(1, (\delta - n + 1)\alpha)$ is the spherical complementary series representation of parameter δ; hence $n_0(\Gamma_d) = 1$ and $\mathcal{W}_d$ does not weakly contain any complementary series representation of parameter $\max(n-2, s_1) < s < \delta$. So $\sup s_0(\Gamma_d) \le \max(n-2, s_1) < \delta$ and $\sup n_0(\Gamma_d) = 1$ where d runs over all square-free integers with no prime factors in S. □

Denote by $\Gamma(d)$ the image of Γ under the reduction map $\tilde{\mathbf{G}} \to \tilde{\mathbf{G}}(\mathbb{Z}/d\mathbb{Z})$ and set $\mathcal{O}_d$ to be the orbit of w_0 in $(\mathbb{Z}/d\mathbb{Z})^m$ under $\Gamma(d)$; so $\#\mathcal{O}_d = [\Gamma : \Gamma_{w_0}(d)]$. We also set

$$\mathcal{O}_F(d) := \{x \in \mathcal{O}_d : F(x) \equiv 0 \ \ (d)\}.$$

Corollary 8.3. *Put $\mathcal{M}_{w_0 G}(\mathcal{B}_T) = \mathcal{X}$. Suppose that for some finite set S of primes, the family $\{\Gamma_d : d$ is square-free with no factors in $S\}$ has a uniform spectral gap. Then there exists $\eta_0 > 0$ such that for any square-free integer d with no factors in S, we have*

$$|\mathsf{A}_d(T)| = g(d)\mathcal{X} + r_d(\mathsf{A})$$

where $g(d) = \frac{\#\mathcal{O}_F(d)}{\#\mathcal{O}_d}$ and $r_d(\mathsf{A}) = \#\mathcal{O}_F(d) \cdot O(\mathcal{X}^{1-\eta_0})$.

Proof. Since $\Gamma_d \subset \Gamma_{w_0}(d)$, the assumption implies that the family $\{\Gamma_{w_0}(d) : d \text{ is square-free with no factors in } S\}$ has a uniform spectral gap. Therefore, Theorem 1.12 on $\#(w_0\Gamma_{w_0}(d)\gamma \cap \mathcal{B}_T)$ implies that for some uniform $\epsilon_0 > 0$,

$$\begin{aligned}|\mathsf{A}_d(T)| &= \sum_{\gamma\in\Gamma_{w_0}(d)\backslash\Gamma, F(w_0\gamma)\equiv 0(d)} \#(w_0\Gamma_{w_0}(d)\gamma\cap\mathcal{B}_T)\\ &= \sum_{\gamma\in\Gamma_{w_0}(d)\backslash\Gamma, F(w_0\gamma)\equiv 0(d)} \left(\tfrac{1}{[\Gamma:\Gamma_{w_0}(d)]}\mathcal{X} + O(\mathcal{X}^{1-\epsilon_0})\right).\end{aligned}$$

Since $\#\mathcal{O}_F(d) = \#\{\gamma \in \Gamma_{w_0}(d)\backslash\Gamma, F(w_0\gamma) \equiv 0 \ \ (d)\}$, the claim follows. □

In the following we verify the sieve axioms (A_1), (A_2) and (A_3) in this set-up. This step is very similar to [48, sec. 4] as we use the same combinatorial sieve and the only difference is that we use the variable $\mathcal{X} = \mathcal{M}_{w_0G}(\mathcal{B}_T)$ instead of T. This is needed for us, as we are working with very general sets $\mathcal{B}_T$; however if $\mathcal{M}_{w_0G}(\mathcal{B}_T) \asymp T^\alpha$ for some $\alpha > 0$, we could also use the parameter T.

Using a theorem of Matthews, Vaserstein and Weisfeiler [46], and enlarging S if necessary, the diagonal embedding of Γ is dense in $\prod_{p\notin S}\tilde{\mathbf{G}}(\mathbb{F}_p)$. The multiplicative property of g on square-free integers with no factors in S follows from this (see [48, proof of Prop. 4.1]).

Letting $W_j = \{x \in W : F_j(x) = 0\}$, W_j is an absolutely irreducible affine variety over $\mathbb{Q}$ of dimension $\dim(W) - 1$ and hence by Noether's theorem, W_j is absolutely irreducible over $\mathbb{F}_p$ for all $p \notin S$, by enlarging S if necessary. We may also assume that $W(\mathbb{F}_p) = w_0\mathbf{G}(\mathbb{F}_p)$ (possibly after adding $\{0\}$) for all $p \notin S$ by Lang's theorem [38]. Using Lang-Weil estimate [39] on $\#W(\mathbb{F}_p)$ and $\#W_j(\mathbb{F}_p)$, we obtain that for $p \notin S$,

$$\#\mathcal{O}_F(p) = r\cdot p^{\dim(W)-1} + O(p^{\dim W - 3/2}) \ \text{ and } \ \#\mathcal{O}_p = p^{\dim W} + O(p^{\dim W - 1/2}).$$

Hence

$$g(p) = r \cdot p^{-1} + O(p^{-3/2})$$

for all $p \notin S$. This implies A_3 (cf. [47, Thm 2.7]), as well as the last claim of A_1.

Moreover this together with Corollary 8.3 imply that

$$r(\mathsf{A}, d) \ll d^{\dim W - 1}\mathcal{X}^{1-\eta_0}.$$

Hence for $D \le \mathcal{X}^{\eta_0/(2\dim W)}$ and $\epsilon_0 = \eta_0/2$,

$$\sum_{d\le D} r(\mathsf{A}, d) \ll D^{\dim W}\mathcal{X}^{1-\eta_0} \le \mathcal{X}^{1-\epsilon_0},$$

providing (A_2). Therefore for any $z = D^{1/s} \le \mathcal{X}^{\eta_0/(2s\dim W)}$ and $s > 9r$, and for all large $\mathcal{X}$, we have

$$S(\mathsf{A}, P) \asymp \frac{\mathcal{X}}{(\log \mathcal{X})^r}. \tag{8.4}$$

Proof of Theorem 1.16. Using arguments in the proof of Corollary 8.3, we first observe (cf. [48, Lem. 4.3]) that there exists $\eta > 0$ such that for any $k \in \mathbb{N}$,

$$\#\{x \in \mathcal{O} \cap \mathcal{B}_T : F_j(x) = k\} \ll \mathcal{X}^{1-\eta}.$$

Fixing $0 < \epsilon_1 < \eta$, it implies that

$$\#\{x \in \mathcal{O} \cap \mathcal{B}_T : |F_j(x)| \le \mathcal{X}^{\epsilon_1}\} \ll \mathcal{X}^{1-\eta+\epsilon_1}. \tag{8.5}$$

Now

$$\#\{x \in \mathcal{O} \cap \mathcal{B}_T : \text{all } F_j(x) \text{ prime}\} \le \sum_{j=1}^{r} \#\{x \in \mathcal{O} \cap \mathcal{B}_T : |F_j(x)| \le \mathcal{X}^{\epsilon_1}\}$$
$$+ \#\{x \in \mathcal{O} \cap \mathcal{B}_T : |F_j(x)| \ge \mathcal{X}^{\epsilon_1} \text{ for all } 1 \le j \le r \text{ and all } F_j(x) \text{ prime}\}.$$

Now for $z \le \mathcal{X}^{\eta_0/(2s \dim W)}$ such that $P = \prod_{p<z} p \ll \mathcal{X}^{\epsilon_1}$, we have

$$\{x \in \mathcal{O} \cap \mathcal{B}_T : |F_j(x)| \ge \mathcal{X}^{\epsilon_1} \text{ for all } 1 \le j \le r \text{ and all } F_j(x) \text{ prime}\}$$
$$\subset \{x \in \mathcal{O} \cap \mathcal{B}_T : (F_j(x), P) = 1\}$$

and the cardinality of the latter set is $S(\mathsf{A}, P)$ according to our definition of a_n's.

Therefore, we obtain the desired upper bound:

$$\#\{x \in \mathcal{O} \cap \mathcal{B}_T : \text{all } F_j(x) \text{ prime}\} \ll \mathcal{X}^{1-\eta+\epsilon_1} + \frac{\mathcal{X}}{(\log \mathcal{X})^r} \ll \frac{\mathcal{X}}{(\log \mathcal{X})^r}.$$

Proof of Theorem 1.17. By the assumption, for some $\beta > 0$,

$$\max_{x \in \mathcal{B}_T} \|x\| \ll \mathcal{M}_{w_0 G}(\mathcal{B}_T)^{\beta} = \mathcal{X}^{\beta}. \tag{8.6}$$

It follows that

$$\max_{x \in \mathcal{B}_T} |F(x)| \ll \mathcal{M}_{w_0 G}(\mathcal{B}_T)^{\beta \deg(F)} = \mathcal{X}^{\beta \deg(F)}. \tag{8.7}$$

Then for $z = \mathcal{X}^{\eta_0/(2s \dim W)}$ and $P = \prod_{p<z, p \notin S} p$, $R = \frac{\beta \cdot \deg(F) 2s \dim W}{\eta_0}$, we have

$$\{x \in \mathcal{O} \cap \mathcal{B}_T : (F(x), P) = 1\} \subset$$
$$\{x \in \mathcal{O} \cap \mathcal{B}_T : F(x) \text{ has at most } R \text{ prime factors}\},$$

since all prime factors of $F(x)$ has to be at least the size of z if $(F(x), P) = 1$ and $|F(x)| \ll \mathcal{X}^{\beta \deg(F)}$ if $x \in \mathcal{B}_T$. Since $S(\mathsf{A}, P) = \#\{x \in \mathcal{O} \cap \mathcal{B}_T : (F(x), P) = 1\}$, we get the desired lower bound $\frac{\mathcal{X}}{(\log \mathcal{X})^r}$ from (8.4).

Remark 8.8. When Γ is an arithmetic subgroup of a simply connected semisimple algebraic $\mathbb{Q}$-group G, and H is a symmetric subgroup, the analogue of Theorem 1.12 has been obtained in [4], assuming that $H \cap \Gamma$ is a lattice in H. Strictly speaking, [4, Theorem 1.3] is stated only for a fixed group Γ; however it is clear from its proof that the statement also holds uniformly over its congruence subgroups with the correct main term, as in Theorem 1.12. Based on this, one can use the combinatorial sieve 8.1 to

obtain analogues of Theorems 1.16 and 1.17, as it was done for a group variety in [48]. Theorem 1.17 on lower bound for Γ arithmetic was obtained in [21] further assuming that $H \cap \Gamma$ is co-compact in H.

Acknowledgment: We are grateful to Gregg Zuckerman for pointing out an important mistake regarding the spectrum of $L^2(\Gamma\backslash G)$ in an earlier version of this paper. We would like to thank Peter Sarnak for helpful comments and Sigurdur Helgason for the reference on the work of Harish-Chandra. We also thank Dale Winter for helpful conversations on related topics.

References

[1] T. Aubin. Nonlinear analysis on manifolds. *GM 252 Springer*, 1982.
[2] Martine Babillot. On the mixing property for hyperbolic systems. *Israel J. Math.*, 129:61–76, 2002.
[3] M. W. Baldoni Silva and D. Barbasch The unitary spectrum for real rank one groups. *Inventiones Math.*, 72:27–55, 1983.
[4] Yves Benoist and Hee Oh. Effective equidistribution of S-integral points on symmetric varieties. *Annales de L'Institut Fourier*, Vol 62 (2012), 1889-1942.
[5] Jean Bourgain, Alex Gamburd, and Peter Sarnak. Generalization of Selberg's 3/16 theorem and Affine sieve. *Acta Math.* 207, (2011) 255-290.
[6] Jean Bourgain, Alex Gamburd, and Peter Sarnak. Affine linear sieve, expanders, and sum-product. *Inventiones* 179, (2010) 559-644.
[7] Jean Bourgain, Alex Kontorovich, and Peter Sarnak. Sector estimates for Hyperbolic isometries. *GAFA* 20, 1175–1200, 2010.
[8] Jean Bourgain and Alex Kontorovich. On representations of integers in thin subgroups of $\mathrm{SL}(2,\mathbb{Z})$. *GAFA* 20, 1144–1174, 2010.
[9] Jean Bourgain and Alex Kontorovich. On the local-global conjecture for integral Apollonian gaskets. *Inventiones*, Vol 196, 2014, 589-650
[10] B. H. Bowditch. Geometrical finiteness for hyperbolic groups. *J. Funct. Anal.*, 113(2):245–317, 1993.
[11] Rufus Bowen. Periodic points and measures for Axiom A diffeomorphisms. *Trans. Amer. Math. Soc.*, 154:377–397, 1971.
[12] Marc Burger. Horocycle flow on geometrically finite surfaces. *Duke Math. J.*, 61(3):779–803, 1990.
[13] F. Dal'bo. Topologie du feuilletage fortement stable. *Ann. Inst. Fourier (Grenoble)*, 50(3):981–993, 2000.
[14] Dmitry Dolgopyat. On decay of correlations in Anosov flows. *Ann. Math*, 147:357–390, 1998.
[15] J. Diximer. Sur les représentations de certains groupes orthogonaux. *C. R. Acad. Sc. Paris*, vol 89 (1960), 3263-3265.
[16] W. Duke, Z. Rudnick, and P. Sarnak. Density of integer points on affine homogeneous varieties. *Duke Math. J.*, 71(1):143–179, 1993.
[17] Alex Eskin and C. T. McMullen. Mixing, counting, and equidistribution in Lie groups. *Duke Math. J.*, 71(1):181–209, 1993.
[18] L. Flaminio and R. Spatzier. Geometrically finite groups, Patterson-Sullivan measures and Ratner's theorem. *Inventiones*, 99 (1990), 601-626.
[19] Alex Gamburd. On the spectral gap for infinite index congruence subgroups of $\mathrm{SL}_2(\mathbb{Z})$. *Israel J. Math*, 127 (2002), 157-200.
[20] Alex Gorodnik and Amos Nevo. The ergodic theory of Lattice subgroups. *Annals of Math Studies.* Vol 172, 2009.

[21] Alex Gorodnik and Amos Nevo. Lifting, Restricting and Sifting integral points on affine homogeneous varieties. *To appear in Compositio Math.* arXiv:1009.5217.
[22] Alex Gorodnik and Hee Oh. Orbits of discrete subgroups on a symmetric space and the Furstenberg boundary. *Duke Math. J.*, 139(3):483-525, 2007.
[23] Alex Gorodnik, Hee Oh and Nimish Shah. Integral points on symmetric varieties and Satake compactifications *American J. Math.*, 131:1-57, 2009.
[24] Alex Gorodnik, Hee Oh and Nimish Shah. Strong wavefront lemma and counting lattice points in sectors. *Israel J. Math.*, 176:419-444, 2010.
[25] H. Halberstam and H. Richert. Sieve methods. *Academic Press.,* (1974) 167-242.
[26] Takeshi Hirai. On irreducible representations of the Lorentz group of n-th order. *Proc. Japan. Acad.*, Vol 38., 258-262, 1962.
[27] Roger Howe and Calvin Moore. Asymptotic properties of unitary representations. *J. Funct. Anal.*, 72-96, 1979.
[28] Henryk Iwaniec and Emmanuel Kowalski. *Analytic number theory*, volume 53 of *American Mathematical Society Colloquium Publications*. American Mathematical Society, Providence, RI, 2004.
[29] D. Y. Kleinbock and G. A. Margulis. Bounded orbits of nonquasiunipotent flows on homogeneous spaces. In *Sinaĭ's Moscow Seminar on Dynamical Systems*, volume 171 of *Amer. Math. Soc. Transl. Ser. 2*, pages 141-172. Amer. Math. Soc., Providence, RI, 1996.
[30] A. Knapp and E. Stein. Intertwining operators for semisimple groups *Annals of Math.*, 489-578, 1971.
[31] Inkang Kim. Counting, Mixing and Equidistribution of horospheres in geometrically finite rank one locally symmetric manifolds. *To appear in J. Reine Angew. Math.*, arXiv:1103.5003.
[32] A. Knapp. Representation theory of semisimple Lie groups. *Princeton University press*.
[33] Alex Kontorovich. The hyperbolic lattice point count in infinite volume with applications to sieves. *Duke Math. J.*, 149(1):1-36, 2009.
[34] Alex Kontorovich and Hee Oh. Almost prime Pythagorean triples in thin orbits. *J. Reine Angew. Math.* Vol 667, 89-131, 2012.
[35] Alex Kontorovich and Hee Oh. Apollonian circle packings and closed horospheres on hyperbolic 3-manifolds. *Journal of AMS.*, Vol 24 (2011), 603-648.
[36] E. Kowalski. Sieve in expansion. *Séminaire Bourbaki* arXiv:1012.2793.
[37] E. Kowlaski. Sieve in discrete groups, especially sparse. *In "Thin groups and superstrong approximation", edited by Breuillard and Oh*, Math. Sci.Res.Inst.Publ. 61, Cambridge Univ. Press. Cambridge 2014
[38] S. Lang. Algebraic groups over finite fields. *American J. Math.* 78 (1956) 555-563.
[39] S. Lang and A. Weil. Number of points of varieties in finite fields. *American J. Math.* 76 (1954) 819-827.
[40] Peter D. Lax and Ralph S. Phillips. The asymptotic distribution of lattice points in Euclidean and non-Euclidean spaces. *J. Funct. Anal.*, 46(3):280-350, 1982.
[41] Min Lee and Hee Oh. Effective count for Apollonian circle packings and closed horospheres. *GAFA.,* Vol 23 (2013), 580-621.
[42] J. Liu and P. Sarnak. Integral points on quadrics in three variables whose coordinates have few prime factors. *Israel J. Math.,* 178 (2010) 393-426.
[43] M. Magee. Quantitative spectral gap for thin groups of hyperbolic isometries. *JEMS*, Vol 17 (2015), 151-187
[44] F. Maucourant. Homogeneous asymptotic limits of Haar measures of semisimple Lie groups and their lattices. *Duke Math. J.*. vol 136 (2) 1007, 357-399.
[45] Gregory Margulis. On some aspects of theory of Anosov systems. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2004.

[46] C. Matthews, L. Vaserstein and B. Weisfeiler. Congruence properties of Zariski dense subgroups. *Proc. London Math. Soc.* 48, 1984, 514-532.
[47] H. Montgomery and R. Vaughan. Multiplicative number theory. I. Classical theory. *Camb. Studies in Advanced Math.* 97, Camb, Univ. Press. 2007.
[48] Amos Nevo and Peter Sarnak. Prime and Almost prime integral points on principal homogeneous spaces. *Acta Math.*, 205 (2010), 361-402
[49] Hee Oh. Orbital counting via mixing and unipotent flows *Homogeneous flows, Moduli spaces and Arithmetic.* Clay Math. Proceedings (2010), Vol 10, 339-375.
[50] Hee Oh. Dynamics on geometrically finite hyperbolic manifolds with applications to Apollonian circle packings and beyond. *Proceedings of ICM.* 2010, Vol III, 1308-1331.
[51] Hee Oh. Harmonic analysis, Ergodic theory and Counting for thin groups. *In "Thin groups and superstrong approximation", edited by Breuillard and Oh*, Math. Sci.Res.Inst.Publ. 61, Cambridge Univ. Press. Cambridge 2014
[52] Hee Oh and Nimish Shah. Equidistribution and counting for orbits of geometrically finite hyperbolic groups. *Journal of AMS.* Vol 26 (2013), 511-562.
[53] Hee Oh and Nimish Shah. The asymptotic distribution of circles in the orbits of Kleinian groups. *Inventiones*, Vol 187, 1-35, 2012.
[54] S.J. Patterson. The limit set of a Fuchsian group. *Acta Mathematica*, 136:241–273, 1976.
[55] J. Parkkonen and F. Paulin. Counting common perpendicular arcs in negative curvature. *Preprint*, arXiv:1305.1332.
[56] Thomas Roblin. Ergodicité et équidistribution en courbure négative. *Mém. Soc. Math. Fr. (N.S.)*, (95):vi+96, 2003.
[57] A. Salehi Golsefidy and P. Sarnak. Affine Sieve. *Journal of AMS*, Vol 26, (2013), 1085-1105
[58] A. Salehi Golsefidy and P. Varjú. Expansion in perfect groups. *GAFA.* Vol 22 (2012), 1832-1891.
[59] Barbara Schapira. Equidistribution of the horocycles of a geometrically finite surface. *Int. Math. Res. Not.*, (40):2447-2471, 2005.
[60] Yehuda Shalom. Rigidity, unitary representations of semisimple groups, and fundamental groups of manifolds with rank one transformation group. *Ann. of Math. (2)*, 152(1):113-182, 2000.
[61] L. Stoyanov. Spectra of Ruelle transfer operators for axiom A flows. *Nonlinearity*, 24 (2011), 1089-1120.
[62] Dennis Sullivan. The density at infinity of a discrete group of hyperbolic motions. *Inst. Hautes Études Sci. Publ. Math.*, (50):171-202, 1979.
[63] Dennis Sullivan. Entropy, Hausdorff measures old and new, and limit sets of geometrically finite Kleinian groups. *Acta Math.*, 153(3-4):259-277, 1984.
[64] Ilya Vinogradov. Effective bisector estimate with applications to Apollonian circle packings. *IMRN,* Vol. 2014, No 12, 3217-3262
[65] Garth Warner. Harmonic Analysis on semisimple Lie groups I. *Mém. Soc. Math. Fr. (N.S.)*, (95):vi+96, 2003.
[66] Garth Warner. Harmonic Analysis on semisimple Lie groups II. *Mém. Soc. Math. Fr. (N.S.)*, (95):vi+96, 2003.
[67] Dale Winter. *Mixing of frame flow for rank one locally symmetric spaces and measure classification. To appear in Israel Journal of Math*, arXiv 1403.2425.

Department of Mathematics, The University of Texas at Austin, Austin, TX 78750
E-mail address: amir@math.utexas.edu

Mathematics department, Yale university, New Haven, CT 06520 and Korea Institute for Advanced Study, Seoul, Korea
E-mail address: hee.oh@yale.edu