SPECTRAL GAP AND EXPONENTIAL MIXING ON GEOMETRICALLY FINITE HYPERBOLIC MANIFOLDS

SAM EDWARDS AND HEE OH

Abstract. Let $\mathcal{M} = \Gamma\backslash\mathbb{H}^{d+1}$ be a geometrically finite hyperbolic manifold with critical exponent exceeding $d/2$. We obtain a precise asymptotic expansion of the matrix coefficients for the geodesic flow in $L^2(\mathrm{T}^1(\mathcal{M}))$, with exponential error term essentially as good as the one given by the spectral gap for the Laplace operator on $L^2(\mathcal{M})$ due to Lax and Phillips. Combined with the work of Bourgain, Gamburd, and Sarnak and its generalization by Golsefidy and Varju on expanders, this implies *uniform* exponential mixing for congruence covers of $\mathcal{M}$ when Γ is a Zariski dense subgroup contained in an arithmetic subgroup of $\mathrm{SO}^\circ(d, 1)$.

Contents

1. Introduction 1
2. Preliminaries 6
3. Complementary series representations and Eisenstein integrals 10
4. Asymptotic expansions of matrix coefficients 18
5. Leading term for Sobolev functions 28
6. Proof of Theorem 1.1 30

References 33

1. Introduction

Let $\mathcal{M} = \Gamma\backslash\mathbb{H}^{d+1}$ be a geometrically finite hyperbolic manifold, where Γ is a discrete subgroup of $G = \mathrm{SO}^\circ(d+1, 1)$. We suppose that the critical exponent δ of Γ is strictly bigger than $d/2$.

Denoting by Δ the negative of the Laplace operator on $\mathcal{M}$, the bottom eigenvalue of Δ on $L^2(\mathcal{M})$ is known to be simple and given as $\lambda_0 = \delta(d-\delta)$ by Patterson [30] and Sullivan [40]. Lax and Phillips [24] showed that there are only finitely many eigenvalues of Δ on $L^2(\mathcal{M})$ in the interval $[\lambda_0, \frac{d^2}{4})$. We denote by λ_1 the smallest eigenvalue of Δ in $(\lambda_0, \frac{d^2}{4})$, and define $s_1 =$

S. E. was supported by postdoctoral scholarship 2017.0391 from the Knut and Alice Wallenberg Foundation and H. O. was supported by NSF grants.

$\frac{d}{2}+\sqrt{(\frac{d}{2})^2-\lambda_1}$. Then s_1 is the unique number in $(\frac{d}{2},\delta)$ such that

$$0<\delta(d-\delta)=\lambda_0<\lambda_1=s_1(d-s_1)<d^2/4.$$

If there is no eigenvalue in the open interval $(\lambda_0,\frac{d^2}{4})$, we set $\lambda_1=d^2/4$, i.e., $s_1=d/2$.

Denote by $\mathcal{G}^t$ the geodesic flow on the unit tangent bundle $\mathrm{T}^1(\mathcal{M})$ and by dx the Liouville measure on $\mathrm{T}^1(\mathcal{M})$; this is a $\mathcal{G}^t$-invariant Borel measure which is infinite when $\delta<d$.

The main result of this paper is that the Lax-Philips spectral gap, that is, $\lambda_1-\lambda_0$, or equivalently $\delta-s_1$, controls *rather precisely* the exponential convergence rate of the correlation function for the geodesic flow $\mathcal{G}^t$ acting on $L^2(\mathrm{T}^1(\mathcal{M}),dx)$:

Theorem 1.1. *Let $\delta>d/2$, and set $\eta:=\min\{\delta-s_1,1\}$. There exists $m>d(d+1)/2$ such that for any $\varepsilon>0$ and functions ψ_1,ψ_2 on $\mathrm{T}^1(\mathcal{M})$ with $\mathcal{S}^m(\psi_1),\mathcal{S}^m(\psi_2)<\infty$, we have, as $t\to+\infty$,*

$$\begin{aligned}e^{(d-\delta)t}\int_{\mathrm{T}^1(\mathcal{M})}\psi_1\big(\mathcal{G}^t(x)\big)\psi_2(x)\,dx=&\frac{1}{m^{\mathrm{BMS}}(\mathrm{T}^1(\mathcal{M}))}m^{\mathrm{BR}}(\psi_1)m^{\mathrm{BR}_*}(\psi_2)\\&+O_\varepsilon\left(e^{(-\eta+\varepsilon)t}\mathcal{S}^m(\psi_1)\mathcal{S}^m(\psi_2)\right),\end{aligned}$$

where

- *m^{BMS} is the Bowen-Margulis-Sullivan measure on $\mathrm{T}^1(\mathcal{M})$;*
- *m^{BR} and m^{BR_*} are, respectively, the unstable and stable Burger-Roblin measures on $\mathrm{T}^1(\mathcal{M})$, which are defined compatibly with the choice of dx and dm^{BMS};*
- *$\mathcal{S}^m(\psi_i)$ denotes the L^2-Sobolev norm of ψ_i of degree m and the implied constant depends only on ε.*

We note that $|m^{\mathrm{BR}}(\psi_i)|<\infty$ when $\mathcal{S}^m(\psi_i)<\infty$ (see Lemma 2.2), and hence the main term above is well-defined. For ψ_i compactly supported, the asymptotic formula (without error term) holds for any $\delta>0$, as was obtained by Roblin [33].

Remark 1.1. (1) When $\mathcal{M}$ has finite volume, i.e., when $\delta=d$, exponential mixing of the geodesic flow is a classical result due to Ratner for $d=1$ and to Moore [28] for general $d\geq 1$. Combining Moore's proof with Shalom's trick [37, proof of Theorem 2.1] yields a rate of mixing $\eta=\min\{d-s_1,1\}$ as in Theorem 1.1 above. This proof makes explicit use of the fact [16] that there are no non-spherical complementary series representations $\mathcal{U}(\upsilon,s)$ with $s>d-1$ (see below for notation). Our proof does not require this statement. However, combining the aforementioned fact with our proof improves the rate of mixing for the geodesic flow on finite-volume hyperbolic manifolds to $\eta=\min\{d-s_1,2\}$. In addition, if there is no non-spherical complementary series representation $\mathcal{U}(\upsilon,s)$ with $s>s_1+1$ appearing in $L^2(\Gamma\backslash G)$, then η can be taken to be $d - s_1$ (see Remark 6.5 for details).

(2) For $d = 1$ and 2, Theorem 1.1 can be deduced from [7] and [43], respectively. For a general $d \geq 1$, the case of $\delta > \max\{d-1, d/2\}$ was obtained in [27] for *some* $\eta > 0$ which is not explicit.

(3) The main novelties of Theorem 1.1 are that it addresses *all* geometrically finite groups with $\delta > d/2$ (even those with cusps), and that it gives an optimal value of η with respect to the dependency on the spectral gap $\delta - s_1$.

(4) The order m of Sobolev norm required is in principle completely computable; it may be chosen independently of Γ, and satisfies $m = O(d^2)$.

The BMS measure m^{BMS} is known to be the unique measure of maximal entropy (which is δ) for the geodesic flow ([40], [31]). Babillot showed that the geodesic flow is mixing with respect to m^{BMS} [2]. Theorem 1.1 is known to imply the following exponential mixing for the BMS measure (see [27, Theorem 1.6] for compactly supported functions, and [18, Theorem 1.9] for general bounded functions):

Theorem 1.2. *There exist $\beta > 0$ (explicitly computable, depending only on η in Theorem 1.1) and $m > d(d+1)/2$ such that for all bounded functions ψ_1, ψ_2 on $\mathrm{T}^1(\mathcal{M})$ supported on the one-neighbourhood of* $\mathrm{supp}(m^{\mathrm{BMS}})$*, we have, as $t \to \infty$,*

$$\int_{\mathrm{T}^1(\mathcal{M})} \psi_1\big(\mathcal{G}^t(x)\big)\psi_2(x)\, dm^{\mathrm{BMS}}(x) = \frac{1}{m^{\mathrm{BMS}}(\mathrm{T}^1(\mathcal{M}))} m^{\mathrm{BMS}}(\psi_1) m^{\mathrm{BMS}}(\psi_2) + O\left(e^{-\beta t}\|\psi_1\|_{C^m}\|\psi_2\|_{C^m}\right)$$

where $\|\psi_i\|_{C^m}$ denotes the C^m-norm of ψ_i.

Remark 1.2. When Γ is convex cocompact, Theorem 1.2, for some $\beta > 0$ (which is not explicit), follows from the work of Stoyanov [38], which is based on symbolic dynamics and Dolgopyat operators (see also [35]). This result in turn implies Theorem 1.1 with implicit $\eta > 0$ (see [29], [35]).

Let Γ be a Zariski dense subgroup of an arithmetic subgroup $G(\mathbb{Z})$ of G. Denote by Γ_q the congruence subgroup of Γ of level q: $\Gamma_q = \{\gamma \in \Gamma : \gamma = e \bmod q\}$. When $\pi^{-1}(\Gamma)$ satisfies the strong approximation property for the spin covering map $\pi : \mathrm{Spin}(d+1, 1) \to G$, the work of Bourgain-Gamburd-Sarnak [6] and its generalization by Golsefidy-Varju [41] on expanders imply that there exists a finite set S of primes such that the family $\mathcal{F} := \{\Gamma_q : q \text{ is square-free with no factors in } S\}$ has a uniform spectral gap in the sense that

$$\inf_{\Gamma_q \in \mathcal{F}} \{\delta - s_1(q)\} > 0, \tag{1.3}$$

where $s_1(q)(d-s_1(q))$ is the second smallest eigenvalue of the Laplacian Δ on $L^2(\Gamma_q\backslash\mathbb{H}^{d+1})$ [1]. This uses the transfer property from the combinatorial spectral gap to the archimedean spectral gap due to [5] (see also [19]). For certain families of subgroups the uniform spectral gap (1.3) may be explicitly bounded from below, see [8, 9, 13, 25, 36].

We then have:

Corollary 1.3. *In Theorems 1.1 and 1.2, the exponents η and β can be chosen uniformly over all congruence covers $\mathrm{T}^1(\Gamma_q\backslash\mathbb{H}^{d+1})$, $\Gamma_q \in \mathcal{F}$.*

When Γ is convex cocompact, Corollary 1.3 was obtained by the second-named author and Winter [29] for $d = 1$ and by Sarkar [34] for a general $d \geq 1$, combining Dolgopyat's methods with the expander theory ([6], [41]).

Theorems 1.1-1.2 and Corollary 1.3 are known to have many immediate applications in number theory and geometry. To name a few, see ([11], [12], [3], [5], [27]) for effective counting and affine sieve, [26] for the prime geodesic theorem, and [18] for shrinking target problems.

Fix $o \in \mathbb{H}^{d+1}$ and $v_o \in \mathrm{T}^1(\mathbb{H}^{d+1})$ based at o. Setting $K = \mathrm{Stab}_G(o) \simeq \mathrm{SO}(d+1)$ and $M = \mathrm{Stab}_G(v_o) \simeq \mathrm{SO}(d)$, we can identify $\mathcal{M} = \Gamma\backslash G/K$ and $\mathrm{T}^1(\mathcal{M}) = \Gamma\backslash G/M$. We let $\{a_t\}$ denote the one-parameter diagonalizable subgroup of G commuting with M whose right translation on $\Gamma\backslash G/M$ corresponds to the geodesic flow $\mathcal{G}^t$ on $\mathrm{T}^1(\mathcal{M})$. As $L^2(\mathrm{T}^1(\mathcal{M}))$ can be identified with the space of M-invariant functions in $L^2(\Gamma\backslash G)$, Theorem 1.1 amounts to understanding the asymptotic behaviour of the matrix coefficients $\langle a_t\psi_1, \psi_2\rangle$ for M-invariant functions $\psi_1, \psi_2 \in L^2(\Gamma\backslash G)$.

The non-tempered part of the unitary dual $\hat{G}$ consists of the trivial representation and the complementary series representations $\mathcal{U}(\upsilon, s)$, parameterized by a representation υ in the unitary dual $\hat{M}$ and a real number $s \in \mathcal{I}_\upsilon$, where $\mathcal{I}_\upsilon \subset (d/2, d)$ is an interval depending on υ. In this parameterization, the complementary series representation $\mathcal{U}(\upsilon, s)$ is spherical if and only if $\upsilon \in \hat{M}$ is trivial, and for υ non-trivial, $\mathcal{I}_\upsilon$ is contained in $(d/2, d-1)$ [16]; this was the main reason for the hypothesis $\delta > d-1$ in [27].

Our main work for the proof of Theorem 1.1 lies in the detailed analysis of the behavior of matrix coefficients of the complementary series representations $\mathcal{U}(\upsilon, s)$. We remark that Harish-Chandra's work ([46], [47]) does not give an asymptotic expansion for the matrix coefficients of $\mathcal{U}(\upsilon, s)$ for all s, as it excludes finitely many (unknown) parameters s. Even for those $\mathcal{U}(\upsilon, s)$ for which Harish-Chandra's expansion is given, it is hard to use his expansion directly, as it relies on various parameters and recursive formulas.

Denoting by $\langle\cdot,\cdot\rangle_{\mathcal{U}(\upsilon,s)}$ the *unitary* inner product, for each $m \in \mathbb{N}$ we define a Sobolev norm $\|\cdot\|_{\mathcal{S}^m(\upsilon,s)}$ on $\mathcal{U}(\upsilon, s)$ as in (2.3), and let $\mathcal{S}^m(\upsilon, s)$ denote the space of vectors in $\mathcal{U}(\upsilon, s)$ with finite $\|\cdot\|_{\mathcal{S}^m(\upsilon,s)}$-norm.

[1]In view of the recent preprint [15] and an upcoming work by He and de Saxcé, the square-free condition in $\mathcal{F}$ may be removed.

Write $\mathcal{U}(\upsilon, s) = \oplus_{\tau\in\hat{K}} \mathcal{U}(\upsilon, s)_\tau$ for the decomposition into different K-types and for $d/2 < s < d$, set

$$\eta_s := \min\{2s - d, 1\} > 0.$$

We show the following *concrete* asymptotic expansion of matrix coefficients with an optimal rate.

Theorem 1.4. *There exists $m \in \mathbb{N}$ such that for any complementary series representation $\mathcal{U}(\upsilon, s)$ containing a non-trivial M-invariant vector, for all $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{S}^m(\upsilon, s)$ and $t \geq 0$,*

$$\begin{aligned}\langle \mathcal{U}(\upsilon, s)(a_t)\boldsymbol{u}, \boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)} =& e^{(s-d)t}\left(\sum_{\tau_1,\tau_2\in\hat{K}} \langle T^{\tau_2}_{\tau_1} C_+(s)\mathsf{P}_{\tau_1}\boldsymbol{u}, \mathsf{P}_{\tau_2}\boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)}\right) \\ &+ O_s\big(e^{(s-d-\eta_s)t}\|\boldsymbol{u}\|_{\mathcal{S}^m(\upsilon,s)}\|\boldsymbol{v}\|_{\mathcal{S}^m(\upsilon,s)}\big)\end{aligned}$$

and the sum

$$\sum_{\tau_1,\tau_2\in\hat{K}} \langle T^{\tau_2}_{\tau_1} C_+(s)\mathsf{P}_{\tau_1}\boldsymbol{u}, \mathsf{P}_{\tau_2}\boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)} \tag{1.4}$$

converges absolutely. Here $T^{\tau_2}_{\tau_1} : \mathcal{U}(\upsilon, s)_{\tau_1} \to \mathcal{U}(\upsilon, s)_{\tau_2}$ is given by (3.5)*, $C_+(s)$ is the Harish-Chandra c-function given in* (4.1)*, and P_τ is the orthogonal projection onto the K-type τ. Moreover, the implied constant is uniformly bounded over s in compact subsets of the interval $\mathcal{I}_\upsilon$.*

One of the key observations made in this paper is that if $\mathcal{U}(\upsilon, s)$ is non-spherical and $\boldsymbol{u} \in \mathcal{U}(\upsilon, s)$ is M-invariant, then the main term (1.4) vanishes (Corollary 3.5): for all $\tau_1, \tau_2 \in \hat{K}$,

$$T^{\tau_2}_{\tau_1} C_+(s)\mathsf{P}_{\tau_1}\boldsymbol{u} = 0.$$

Furthermore, there is additional uniformity with respect to the s-variable when considering only M-invariant vectors.

Corollary 1.5. *There exists $m \in \mathbb{N}$ such that if $\mathcal{U}(\upsilon, s)$ is non-spherical, then for all M-invariant vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{U}(\upsilon, s)$,*

$$|\langle \mathcal{U}(\upsilon, s)(a_t)\boldsymbol{u}, \boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)}| \ll_s e^{(s-d-\eta_s)t}\|\boldsymbol{u}\|_{\mathcal{S}^m(\upsilon,s)}\|\boldsymbol{v}\|_{\mathcal{S}^m(\upsilon,s)}.$$

Moreover, the implied constant is uniformly bounded over s in compact subsets of $(\frac{d}{2}, d)$.

Organization. We start by recalling in Section 2 some key background facts and notation. In Section 3, we define the models of the complementary series that we will work in and the Eisenstein integrals that are key to understanding their matrix coefficients. The main technical work is done in Section 4; establishing the asymptotic expansion and bounds for matrix coefficients in complementary series representations. Section 5 is devoted to extending Roblin's mixing result [33, Theorem 3.4] from compactly supported functions to arbitrary Sobolev functions. Finally, the proof of Theorem 1.1 is given in Section 6, the main part of which consists of decomposing the regular representation of G on $L^2(\Gamma\backslash G)$ into irreducible representations and applying the bounds obtained in Section 4 to each individual component.

2. Preliminaries

Let $d \geq 1$ and $G = \mathrm{SO}^\circ(d+1,1)$ be the group of orientation-preserving isometries of $\mathbb{H}^{d+1}$. Let $\Gamma < G$ be a torsion-free discrete subgroup of G, and let $\mathcal{M} = \Gamma\backslash\mathbb{H}^{d+1}$ be the associated hyperbolic manifold.

2.1. **Structure and subgroups of G.** Let $K = \mathrm{SO}(d+1) < G$ be a maximal compact subgroup, and let $A = \{a_t : t \in \mathbb{R}\}$ be a one-parameter diagonalizable subgroup and $M = \mathrm{SO}(d)$ the centralizer of A in K. We can identify $\mathcal{M}$ with $\Gamma\backslash G/K$ and the unit tangent bundle $\mathrm{T}^1(\mathcal{M})$ with $\Gamma\backslash G/M$ in the way that the geodesic flow on $\mathrm{T}^1(\mathcal{M})$ is given by the right translation action of a_t on $\Gamma\backslash G/M$.

We denote by N and $\overline{N}$ the contracting and expanding horospherical subgroups, respectively; i.e.

$$N = \{g \in G : a_{-t} g a_t \to e \text{ as } t \to +\infty\};$$

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

Then $\overline{N} = \omega N \omega^{-1}$, where $\omega \in N_K(A)$ is such that $\omega a \omega^{-1} = a^{-1}$ for all $a \in A$.

We note that N and $\overline{N}$ are both abelian subgroups, isomorphic to $\mathbb{R}^d$ via the logarithm map, which are normalized by AM. Under this isomorphism $N \simeq \mathbb{R}^d$, conjugation by an element $m \in M = \mathrm{SO}(d)$ is an isometry and conjugation by a_t corresponds to scaling by e^t on N, and e^{-t} on $\overline{N}$. As usual, the Lie algebras of G, K, A, N, $\overline{N}$ are denoted by $\mathfrak{g}$, $\mathfrak{k}$, $\mathfrak{a}$, $\mathfrak{n}$ and $\overline{\mathfrak{n}}$, respectively. This gives

$$\mathrm{Ad}(a_t)|_{\mathfrak{n}} = e^t \times \mathrm{Id} \quad \text{and} \quad \mathrm{Ad}(a_t)|_{\overline{\mathfrak{n}}} = e^{-t} \times \mathrm{Id}.$$

We have an Iwasawa decomposition $G = KAN$, where the product map $K \times A \times N \to G$ is a diffeomorphism and the projection to each individual factor is a smooth map [21, Chapter VI.4]. For $g \in G$, we write

$$g = \kappa(g) \exp(H(g)) n_g,$$

where $\kappa(g) \in K$, $H(g) \in \mathfrak{a}$, and $n_g \in N$. Letting $\alpha \in \mathfrak{a}_{\mathbb{C}}^*$ be the unique element such that

$$\alpha(H(a_t)) = t,$$

the map $s \mapsto s \cdot \alpha$ defines an identification $\mathbb{C} \simeq \mathfrak{a}_{\mathbb{C}}^*$. In view of this, we write

$$e^{sH(g)} = e^{s \cdot \alpha(H(g))} \quad \text{for } g \in G \text{ and } s \in \mathbb{C}.$$

2.2. **Various measures on** $\mathrm{T}^1(\mathcal{M})$**.** Let $\Lambda \subset \partial\mathbb{H}^{d+1}$ denote the limit set of Γ. We assume that Γ is non-elementary, or equivalently, Λ has at least 3 points. Throughout the paper, we assume that

$$\Gamma \textit{ is geometrically finite},$$

that is, the unit neighborhood of the convex core of $\mathcal{M}$, given by $\Gamma\backslash\mathrm{hull}(\Lambda)$, has finite Riemannian volume. We let δ denote the critical exponent of Γ, which is known to be equal to the Hausdorff dimension of Λ. We remark that $\delta = d$ if and only if $\partial\mathbb{H}^n = \Lambda$ if and only if Γ is a lattice in G.

We fix $o \in \mathbb{H}^{d+1}$ whose stabilizer is given by K, and $v_o \in \mathrm{T}_o(\mathbb{H}^{d+1})$ the unit vector whose stabilizer is M. Let ν_o be the Patterson-Sullivan measure on Λ which is unique up to a constant multiple; this is characterized by the condition that

$$\gamma^*\nu_o = |\gamma'|^\delta \cdot \nu_o$$

for all $\gamma \in \Gamma$, where $|\gamma'|$ denotes the derivative of γ in the spherical metric on $\partial\mathbb{H}^{d+1}$ with respect to o. We also let m_o denote the K-invariant probability measure on $\partial\mathbb{H}^{d+1}$.

Let $\pi : \mathrm{T}^1(\mathbb{H}^{d+1}) \to \mathbb{H}^{d+1}$ be the base point projection. For $u \in \mathrm{T}^1(\mathbb{H}^{d+1})$, we denote by $u^\pm \in \partial\mathbb{H}^{d+1}$ the forward and the backward endpoints of the geodesic determined by u. Consider the Hopf parameterization of $\mathrm{T}^1(\mathbb{H}^{d+1})$ given by:

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

where β denotes the Busemann function. Using the Hopf coordinates, the Bowen-Margulis-Sullivan measure m^{BMS}, the Liouville measure du, and the Burger-Roblin measure m^{BR} on $\mathrm{T}^1(\mathbb{H}^{d+1})$ are respectively given as follows:

(1) $dm^{\mathrm{BMS}}(u) = e^{\delta\beta_{u^+}(o,\pi(u))}\, e^{\delta\beta_{u^-}(o,\pi(u))}\, d\nu_o(u^+)d\nu_o(u^-)ds;$
(2) $du = dm^{\mathrm{Liouville}}(u) = e^{d\beta_{u^+}(o,\pi(u))}\, e^{d\beta_{u^-}(o,\pi(u))}\, dm_o(u^+)dm_o(u^-)ds;$
(3) $dm^{\mathrm{BR}}(u) = e^{d\beta_{u^+}(o,\pi(u))}\, e^{\delta\beta_{u^-}(o,\pi(u))}\, dm_o(u^+)d\nu_o(u^-)ds.$

The BR measure m^{BR} is a Lebesgue measure on each $\overline{N}$-leaf, so we call it the unstable BR measure. The stable BR-measure m^{BR_*} is defined similarly to m^{BR} by exchanging the roles of u^+ and u^-.

These measures are all left Γ-invariant, and hence induce Borel measures on $\mathrm{T}^1(\mathcal{M}) = \Gamma\backslash \mathrm{T}^1(\mathbb{H}^{d+1})$ for which we use the same notation. We remark that m^{BMS} is a finite measure, invariant under the geodesic flow. We will normalize the Patterson-Sullivan measure ν_o so that $m^{\mathrm{BMS}}(\mathrm{T}^1(\mathcal{M})) = 1$. The other three measures are infinite, unless $\delta = d$.

We will sometimes consider these measures as measures on $\Gamma\backslash G$ by putting: for $\psi \in C_c(\Gamma\backslash G)$, and for $\star = \mathrm{BMS}, \mathrm{Liouville}, \mathrm{BR}$,

$$m^\star(\psi) = m^\star\left(\int_M \psi\, dm\right),$$

where dm denotes the Haar probability measure on M. Note that the Liouville measure, considered as a measure on $\Gamma\backslash G$, is G-invariant; we will denote this by dg.

2.3. **The base eigenfunction ϕ_0.** Throughout the article, we assume

$$\delta > \frac{d}{2},$$

which is a necessary and sufficient condition for the existence of an eigenvalue of Δ on $L^2(\mathcal{M})$. The smallest eigenvalue of Δ on $L^2(\mathcal{M})$ is given by $\lambda_0 = \delta(d-\delta)$, and is known to be simple. We denote by ϕ_0 the the unit eigenfunction in $L^2(\mathcal{M})$ with eigenvalue λ_0, and call it the *base eigenfunction*. Up to a constant multiple, ϕ_0 is given by: for $x \in \mathcal{M}$,

$$\phi_0(x) = \int_{\xi \in \Lambda} e^{-\delta \beta_\xi(o,x)} d\nu_o(\xi).$$

The fact that the base eigenfunction ϕ_0 is square-integrable when $\delta > d/2$ is a key result, which allows the unitary representation theory of G, specifically the right translation action of G on $L^2(\Gamma\backslash G)$, to be used to prove dynamical results for the geodesic flow.

Theorem 2.1 (Lax-Phillips [24])**.** *The intersection of the interval $[0, \frac{d^2}{4})$ with the spectrum of Δ, viewed as an unbounded operator on $L^2(\mathcal{M})$, consists of a finite set of eigenvalues $\{\lambda_i = s_i(d - s_i)\}_{0 \le i \le \ell}$, satisfying*

$$0 < \lambda_0 = \delta(d-\delta) < \lambda_1 \le \ldots \le \lambda_\ell < \frac{d^2}{4}.$$

2.4. **The quasi-regular representation $L^2(\Gamma\backslash G)$.** We denote $L^2(\Gamma\backslash G)$ the space of square-integrable functions on $\Gamma\backslash G$ with respect to dg. The G-invariance of dg gives rise to a unitary representation $\big(\rho, L^2(\Gamma\backslash G)\big)$ of G, where ρ is given by the right translation:

$$[\rho(g)f](x) = f(xg)$$

for all $f \in L^2(\Gamma\backslash G)$, $g \in G$, $x \in \Gamma\backslash G$.

As a number of inner products on different vector spaces will show up throughout the article, we reserve now $\langle \cdot, \cdot \rangle$ to mean the $\rho(G)$-invariant inner product on $L^2(\Gamma\backslash G)$. All other inner products will have some additional notation to distinguish them. The subspace of $\rho(K)$-invariant vectors in $L^2(\Gamma\backslash G)$ is denoted $L^2(\Gamma\backslash G)^K$. Similarly, $L^2(\Gamma\backslash G)^M$ denotes the subspace of $\rho(M)$-invariant vectors. We use $L^2(\Gamma\backslash G)^K$ to construct subrepresentations of $\big(\rho, L^2(\Gamma\backslash G)\big)$ as follows: define

$$L^2(\Gamma\backslash G)_{\mathrm{sph}} := \text{the closure of } \{\rho(g)f \,:\, f \in L^2(\Gamma\backslash G)^K,\, g \in G\}.$$

Then

$$\big(\rho, L^2(\Gamma\backslash G)\big) = \big(\rho, L^2(\Gamma\backslash G)_{\mathrm{sph}}\big) \bigoplus \big(\rho, L^2(\Gamma\backslash G)_{\mathrm{sph}}^{\perp}\big);$$

note that both $L^2(\Gamma\backslash G)_{\mathrm{sph}}$ and $L^2(\Gamma\backslash G)_{\mathrm{sph}}^{\perp}$ are $\rho(G)$-invariant closed subspaces. Viewing the base eigenfunction ϕ_0 as an element of $L^2(\Gamma\backslash G)^K$, we define in a similar manner

$$\mathcal{B}_\delta := \text{the closure of the span of } \{\rho(g)\phi_0 \,:\, g \in G\} \subset L^2(\Gamma\backslash G)_{\mathrm{sph}},$$

and

$$\big(\rho, L^2(\Gamma\backslash G)_{\mathrm{sph}}\big) = (\rho, \mathcal{B}_\delta) \oplus (\rho, \mathcal{W}),$$

$\mathcal{W}$ being the orthogonal complement of $\mathcal{B}_\delta$ in $L^2(\Gamma\backslash G)_{\mathrm{sph}}$. It will be of importance later that $(\rho, \mathcal{B}_\delta)$ is an (irreducible) complementary series representation and that Theorem 2.1 gives a complete understanding of the complementary series representations contained in $(\rho, \mathcal{W})$. A useful fact (cf. [23]) that we make of in Section 5 is that for all $f \in L^2(\Gamma\backslash G)^K$,

$$m^{\mathrm{BR}}(f) = m^{\mathrm{BR}_*}(f) = \langle f, \phi_0 \rangle. \tag{2.1}$$

Finally, the direct integral decomposition of $\big(\rho, L^2(\Gamma\backslash G)\big)$ reads

$$\big(\rho, L^2(\Gamma\backslash G)\big) \cong \int_{\mathsf{Z}}^{\oplus} \big(\pi_\zeta, \mathcal{H}_\zeta\big)\, d\mu_{\mathsf{Z}}(\zeta), \tag{2.2}$$

where $(\pi_\zeta, \mathcal{H}_\zeta)$ is an irreducible unitary representation of G for μ_{Z}-a.e. ζ (cf. [45, Corollary 14.9.5]).

2.5. **Sobolev norms on unitary representations of K.** Given a unitary representation (π, V) of K with invariant inner product $(\cdot,\cdot)_V$, a basis $\{X_j\}$ of $\mathfrak{k}$ and $m \in \mathbb{N}$, we define a Sobolev norm $\|\cdot\|_{\mathcal{S}^m(V)}$ on V by

$$\|\boldsymbol{v}\|^2_{\mathcal{S}^m(V)} := \|\boldsymbol{v}\|_V^2 + \sum_U \|d\pi(U)\boldsymbol{v}\|_V^2 \qquad \text{for any } \boldsymbol{v} \in V, \tag{2.3}$$

where $\|\cdot\|_V$ denotes the corresponding norm and the sum runs over all monomials in $\{X_j\}$ of order up to m.

We set $\mathcal{S}^m(V) := \{\boldsymbol{v} \in V : \|\boldsymbol{v}\|_{\mathcal{S}^m(V)} < \infty\}$. Observe that different choices of the basis $\{X_j\}$ give rise to equivalent norms, and that in the case when V is finite-dimensional, we have $\mathcal{S}^m(V) = V$ for any $m \geq 0$.

Viewing $\big(\rho, L^2(\Gamma\backslash G)\big)$ as a unitary representation of K, given a function $f \in L^2(\Gamma\backslash G)$, we let either $\|f\|_{\mathcal{S}^m(\Gamma\backslash G)}$ or simply $\mathcal{S}^m(f)$ denote the norm $\|f\|_{\mathcal{S}^m(L^2(\Gamma\backslash G))}$ as defined by (2.3) above. We denote by $\mathcal{S}^m(\Gamma\backslash G)$ the space of all functions $f \in L^2(\Gamma\backslash G)$ with $\mathcal{S}^m(f) < \infty$.

For $f \in L^2(\Gamma\backslash G)$, $m^{\mathrm{BR}}(f)$ may be infinite in general when $\delta < d$. However we have the following lemma:

Lemma 2.2. *If $m > \frac{(d+1)d}{4}$, then for any $f \in \mathcal{S}^m(\Gamma\backslash G)$,*

$$m^{\mathrm{BR}}(f),\ m^{\mathrm{BR}_*}(f) \ \ll\ \mathcal{S}^m(f).$$

Proof. Given $f \in C(\Gamma\backslash G) \cap L^2(\Gamma\backslash G)$, define

$$f_K(x) := \max_{k \in K} |f(xk)| \qquad \text{for } x \in \Gamma\backslash G.$$

By construction, f_K is $\rho(K)$-invariant and $|f(x)| \leq f_K(x)$ for all $x \in \Gamma\backslash G$, hence $|m^{\mathrm{BR}}(f)| \leq m^{\mathrm{BR}}(f_K)$. In view of (2.1), the Sobolev embedding theorem on the compact manifold K gives

$$|m^{\mathrm{BR}}(f)| \leq m^{\mathrm{BR}}(f_K) = \langle f_K, \phi_0 \rangle \leq \|f_K\| \ll \|f\|_{\mathcal{S}^m(\Gamma\backslash G)}$$

for all $m > \frac{\dim(K)}{2}$ (cf. [1, Theorem 2.30]). Since $\dim K = \frac{d(d+1)}{2}$, the chain of inequalities above holds for any $m > \frac{(d+1)d}{4}$. The claim for m^{BR_*} follows similarly. □

3. Complementary series representations and Eisenstein integrals

In this section we recall the definition and K-type structure of the complementary series representations of $G = \mathrm{SO}^\circ(d+1, 1)$, as well as the Eisenstein integral representations of matrix coefficients for these representations. We start by reviewing the representation theory of the special orthogonal groups $\mathrm{SO}(n)$.

3.1. **Irreducible representations of K and M.** The primary reference for the first facts listed below is [4, pp. 272-277]. Let $m \geq 1$. The irreducible representations of $\mathrm{SO}(2m)$ are parameterized by m-tuples $\tau = (\tau_1, \tau_2, \ldots, \tau_m) \in \mathbb{Z}^m$ such that

$$\tau_1 \geq \tau_2 \geq \ldots \geq \tau_{m-1} \geq \tau_{m-1} \geq |\tau_m|.$$

Similarly, the irreducible representations of $\mathrm{SO}(2m+1)$ are parameterized by m-tuples $\tau = (\tau_1, \tau_2, \ldots, \tau_m) \in \mathbb{Z}^m$ such that

$$\tau_1 \geq \tau_2 \geq \ldots \geq \tau_{m-1} \geq \tau_m \geq 0.$$

We recall that all irreducible representations of $\mathrm{SO}(2m+1)$ and $\mathrm{SO}(4m)$ are self-dual, and that the dual τ^* of an irreducible representation $\tau = (\tau_1, \tau_2, \ldots, \tau_{2m+1})$ of $\mathrm{SO}(4m+2)$ is given by $\tau^* = (\tau_1, \tau_2, \ldots, -\tau_{2m+1})$; the dual τ^* is defined by $\tau^*(k) = \tau(k^{-1})$, acting on the dual (or complex conjugate) of the underlying vector space.

A key result that we will make repeated use of is the branching law for restrictions of representations of K to M which we recall (cf. [21, Chapter IX.3] or [48, pp. 377-380]). We henceforth call irreducible representations of K or M a K or M-type, respectively. Let τ be a K-type. Then the decomposition of $\tau|_M$ reads

$$\tau|_M = \bigoplus_{\sigma \in \hat{M}} m_{\sigma,\tau} \cdot \sigma,$$

where $m_{\sigma,\tau} = 1$ if $\sigma = (\sigma_1, \ldots, \sigma_{\lfloor \frac{d}{2} \rfloor})$ satisfies the interlacing property

$$\tau_1 \geq \sigma_1 \geq \tau_2 \geq \sigma_2 \geq \ldots \geq \sigma_{m-1} \geq |\tau_m| \qquad \text{for } d = 2m-1,$$

and

$$\tau_1 \geq \sigma_1 \geq \tau_2 \geq \sigma_2 \geq \tau_m \geq |\sigma_m| \qquad \text{for } d = 2m,$$

and $m_{\sigma,\tau} = 0$ otherwise. If $m_{\sigma,\tau} = 1$, we say that τ contains σ, and write $\sigma \subset \tau$. The fact that each M-type occurs at most once in τ will allow us to make repeated use of Schur's lemma in the following manner: if a linear operator on V_τ (V_τ being a vector space on which τ is realized) commutes with $\tau(M)$, then it acts as a scalar on each $\sigma \subset \tau$. From now on we write $\dim(\tau)$ and $\dim(\sigma)$ for $\dim(V_\tau)$ and $\dim(V_\sigma)$, respectively.

We connect the dimensions of K-types with the Sobolev norms introduced in Section 2.5. Firstly, let $(\pi, \mathcal{H})$ be a unitary representation of G. For each $\tau \in \hat{K}$, we define

$$\chi_\tau(k) = \dim(V_\tau) \cdot \operatorname{tr}\big(\tau(k)\big)$$

(the trace being defined with respect to any invariant inner product on any realization of τ) and an operator P_τ by

$$\mathsf{P}_\tau = \int_K \overline{\chi_\tau(k)} \pi(k)\, dk.$$

Note that P_τ is the orthogonal projection onto the space $\mathcal{H}_\tau$, where

$$\mathcal{H}_\tau := \{\boldsymbol{v} \in \mathcal{H} \,:\, \mathsf{P}_\tau \boldsymbol{v} = \boldsymbol{v}\}.$$

This gives rise to a decomposition of $\mathcal{H}$ as the orthogonal direct sum

$$\mathcal{H} = \bigoplus_{\tau \in \hat{K}} \mathcal{H}_\tau.$$

If $(\pi, \mathcal{H})$ is irreducible then each $\mathcal{H}_\tau$ is finite-dimensional. Each $\mathcal{H}_\tau$ has the property that $\pi(k)|\mathcal{H}_\tau \cong \tau(k)$ for all $k \in K$. If $\mathcal{H}_\tau \neq 0$, we say that π contains τ, and write $\tau \subset \pi$. There is a similar decomposition of $\mathcal{H}$ with respect to M-types, and projection operators $\mathsf{P}_\sigma = \int_M \overline{\chi_\sigma(m)} \pi(m)\, dm$ for $\sigma \in \hat{M}$. As usual, $\lceil \beta \rceil$ denotes the smallest integer greater than or equal to β.

Lemma 3.1. *There exists $m_0 \in \mathbb{N}$ depending only on K such that for any unitary representation (π, V) of K, $\alpha > 0$, $m \in \mathbb{N}$, and for all $\boldsymbol{v} \in V$,*

$$\sum_{\tau \subset V} \dim(\tau)^\alpha \|\mathsf{P}_\tau \boldsymbol{v}\|_{\mathcal{S}^m(V)} \ll \|\boldsymbol{v}\|_{\mathcal{S}^{m+m_0\lceil \frac{\alpha+1}{2} \rceil}(V)},$$

where the implied constant depends only on K.

Proof. Though the argument is standard (cf. [46, Chapter 4.4.2] or [20, Lemmas 10.3 and 10.4]), we briefly recount it: using the Harish-Chandra isomorphism and the highest weight theory, we find an element ω_K in the center of the universal enveloping algebra of $\mathfrak{k}$ such that for each $\tau \in \hat{K}$, $d\tau(\omega_K)$ acts as a scalar c_τ on τ, where $|c_\tau| \geq \dim(\tau)^2$. Letting m_0 be the order of ω_K, for any $\ell \in \mathbb{N}$, we have (using the Cauchy-Schwarz inequality),

$$\begin{aligned}\sum_{\tau\subset V} \dim(\tau)^\alpha \|\mathsf{P}_\tau \boldsymbol{v}\|_{\mathcal{S}^m(V)} &= \sum_{\tau\subset V} \frac{\dim(\tau)^\alpha}{|c_\tau|^\ell} \|\mathsf{P}_\tau d\pi(\Omega_K^\ell)\boldsymbol{v}\|_{\mathcal{S}^m(V)} \\ &\le \left(\sum_{\tau\subset V} \dim(\tau)^{2\alpha-4\ell} \sum_{\tau\subset V} \|\mathsf{P}_\tau d\pi(\Omega_K^\ell)\boldsymbol{v}\|^2_{\mathcal{S}^m(V)}\right)^{1/2} \\ &\ll \left(\sum_{\tau\subset V} \dim(\tau)^{2\alpha-4\ell}\right)^{1/2} \|\boldsymbol{v}\|_{\mathcal{S}^{m+\ell m_0}(V)}.\end{aligned}$$

The sum $\sum_\tau \dim(\tau)^{-2}$ is finite by [20, Lemma 13], so choosing $\ell = \lceil \frac{\alpha+1}{2} \rceil$ gives $\sum_{\tau\subset V} \dim(\tau)^{2\alpha-4\ell} < \infty$. □

3.2. **Complementary series representations.** We will now recall Hirai's classification [16] of the non-tempered unitary dual of G and construct the models of the complementary series that we will work with, see [46, Chapter 5.5] (cf. also [27, Sections 3.1-3.3]).

Given $s \in \mathbb{C}$, we define the standard representation U^s of G on $L^2(K)$ by

$$[U^s(g)\boldsymbol{v}](k) = e^{-sH(g^{-1}k)}\boldsymbol{v}\big(\kappa(g^{-1}k)\big) \tag{3.1}$$

for all $\boldsymbol{v} \in L^2(K)$, $g \in G$, and $k \in K$. We let λ and ρ denote the left- and right-regular representations of K on $L^2(K)$ respectively. Observe that $U^s|_K = \lambda$; for this reason, it is practical to always view $L^2(K)$ as the unitary representation $(\lambda, L^2(K))$ of K. The decomposition of $L^2(K)$ into K-types reads

$$L^2(K) = \bigoplus_{\tau\in\hat{K}} L^2(K)_\tau,$$

where each $L^2(K)_\tau$ is isomorphic to $\dim(\tau)$ copies of τ. Given $\upsilon \in \hat{M}$, we define

$$L^2(K:\upsilon) := \left\{\boldsymbol{v} \in L^2(K) \,:\, \int_M \overline{\chi_\upsilon(m)}\rho(m)\boldsymbol{v}\, dm = \boldsymbol{v}\right\}.$$

The fact that $U^s(g)$ and $\rho(m)$ commute for all $g \in G$ and $m \in M$ shows that $\big(U^s, L^2(K:\upsilon)\big)$ is a representation of G; in fact $\big(U^s, L^2(K:\upsilon)\big)$ is isomorphic to $\dim(\upsilon)$ copies of the representation $\mathrm{ind}_{MAN}^G(\upsilon\otimes s\otimes 1)$ (cf. [20, Chapter VII]). Note that $L^2(K:\upsilon)_\tau \subset L^2(K)_\tau$ for all $\tau \in \hat{K}$. From Hirai's classification of the unitary dual of G [16], for each $\upsilon \in \hat{M}$, there exists an interval $\mathcal{I}_\upsilon \subset (\frac{d}{2}, d)$ such that if $s \in \mathcal{I}_\upsilon$, then there exists an irreducible, unitarizable subrepresentation $\mathcal{U}(\upsilon, s)$ of $\big(U^s, L^2(K:\upsilon)\big)$. Furthermore, every non-tempered representation of G may be realized as $\mathcal{U}(\upsilon, s)$ for some $\upsilon \in \hat{M}$ and $s \in \mathcal{I}_\upsilon$ in this way, and each $\tau \in \hat{K}$ that contains υ occurs exactly once in $\mathcal{U}(\upsilon, s)$ (cf. [46, Theorem 5.5.1.5] and [20, Theorem 8.37]).

The inner product that makes $\mathcal{U}(\upsilon, s)$ a unitary representation is denoted by $\langle \cdot, \cdot \rangle_{\mathcal{U}(\upsilon,s)}$. Note that for $g \in G$, and $\boldsymbol{v}, \boldsymbol{u} \in \mathcal{U}(\upsilon, s)$,

$$\langle U^s(g)\boldsymbol{v}, \boldsymbol{u}\rangle_{\mathcal{U}(\upsilon,s)} = \langle \mathcal{U}(\upsilon, s)(g)\boldsymbol{v}, \boldsymbol{u}\rangle_{\mathcal{U}(\upsilon,s)}.$$

We conclude this section by recalling two general facts regarding complementary series representations that let us classify the representation $\mathcal{B}_\delta$ (cf. Section 2.4). Firstly, the spherical complementary series representation are the $\mathcal{U}(1, s)$; i.e. precisely those where the representation υ is trivial. Secondly, the Casimir operator $\mathcal{C}$ of G acts as the scalar $-s(d-s)$ on the smooth vectors of $\mathcal{U}(\upsilon, s)$. Combining these facts with the observation that the restriction of $\mathcal{C}$ to right K-invariant functions on G is the Laplace-Beltrami operator lets one conclude that the representation $\mathcal{B}_\delta$ is isomorphic to $\mathcal{U}(1, \delta)$.

3.3. **Eisenstein integrals.** Here we develop the Eisenstein integrals needed to obtain the desired asymptotic expansion of matrix coefficients (cf. [27, Section 3.3], [47, Chapter 6]). Before starting, we make the following elementary but important observation:

Lemma 3.2. *If $\sigma \in \hat{M} \setminus \{\upsilon^*\}$, then for any $\boldsymbol{v} \in L^2(K : \upsilon)_\sigma$, we have*

$$\boldsymbol{v}(m) = 0 \quad \text{for all } m \in M.$$

Proof. Since $\boldsymbol{v} \in L^2(K : \upsilon)$, we have $\boldsymbol{v} = \int_M \rho(m)\boldsymbol{v}\overline{\chi_\upsilon(m)}\, dm$. Hence for any $m \in M$,

$$\begin{aligned}
\boldsymbol{v}(m) &= \int_M \boldsymbol{v}(mm_1)\overline{\chi_\upsilon(m_1)}\, dm_1 \\
&= \int_M \boldsymbol{v}(m_1 m)\overline{\chi_\upsilon(m^{-1}m_1 m)}\, dm_1 \\
&= \int_M \boldsymbol{v}(m_1^{-1} m)\overline{\chi_\upsilon(m_1^{-1})}\, dm_1 \\
&= \int_M \boldsymbol{v}(m_1^{-1} m)\overline{\chi_{\upsilon^*}(m_1)}\, dm_1 \\
&= [\mathsf{P}_{\upsilon^*}\boldsymbol{v}](m).
\end{aligned}$$

Since $\boldsymbol{v} \in L^2(K : \upsilon)_\sigma$ for $\sigma \in \hat{M} \setminus \{\upsilon^*\}$, the orthogonality of the M-types of $L^2(K)$ gives $\mathsf{P}_{\upsilon^*}\boldsymbol{v} = 0$, yielding the claim. $\square$

Before stating the main result of this section, we introduce some more notation. Firstly, we let $\langle \cdot, \cdot \rangle_K$ denote the usual inner product on $L^2(K)$. The corresponding norm on $L^2(K)$ is denoted $\| \cdot \|_K$, and similarly for the operator norm defined with respect to it.

In the rest of this section, we fix a complementary series representation $\mathcal{U}(\upsilon, s)$, $s \in \mathcal{I}_\upsilon$. For each K-type τ of $\mathcal{U}(\upsilon, s)$ we define a vector $\boldsymbol{\chi}_\tau \in \mathcal{U}(\upsilon, s)_\tau$

by

$$\boldsymbol{\chi}_\tau = \sum_{i=1}^{\dim(\tau)} \overline{\boldsymbol{v}_i(e)}\, \boldsymbol{v}_i, \tag{3.2}$$

where $\{\boldsymbol{v}_i\}_{i=1}^{\dim(\tau)}$ is an orthornormal basis of $\mathcal{U}(\upsilon, s)_\tau$ (recall that τ has multiplicity one in $\mathcal{U}(\upsilon, s)$). Note that $\boldsymbol{\chi}_\tau$ is independent of the choice of basis, and for all $\boldsymbol{v} \in \mathcal{U}(\upsilon, s)_\tau$ and $k \in K$, we have

$$\boldsymbol{v}(k) = \langle \boldsymbol{v}, U^s(k)\boldsymbol{\chi}_\tau \rangle_K. \tag{3.3}$$

Lemma 3.3. *Let τ be a K-type of $\mathcal{U}(\upsilon, s)$.*

(1) *For all $\boldsymbol{v} \in \mathcal{U}(\upsilon, s)_\tau \cap \mathcal{U}(\upsilon, s)_{\upsilon^*}$, we have*

$$\int_M |\boldsymbol{v}(m)|^2\, dm = \frac{\dim(\tau)\|\boldsymbol{v}\|_K^2}{\dim(\upsilon)}.$$

(2) $\int_M |\boldsymbol{\chi}_\tau(m)|^2\, dm = \frac{\dim(\tau)^2}{\dim(\upsilon)}$.

Proof. Let $\{\boldsymbol{v}_j\}$ be an orthonormal basis of $\mathcal{U}(\upsilon, s)_\tau \cap \mathcal{U}(\upsilon, s)_{\upsilon^*}$ with respect to $\langle\cdot,\cdot\rangle_K$. Using (3.2) and the Schur orthogonality relations, we get

$$\begin{aligned}\int_M |\boldsymbol{v}(m)|^2\, dm &= \int_M |\langle \boldsymbol{v}, U^s(m)\boldsymbol{\chi}_\tau\rangle_K|^2\, dm \\ &= \sum_{i,j} \overline{\boldsymbol{v}_i(e)}\boldsymbol{v}_j(e) \int_M \langle U^s(m)\boldsymbol{v}, \boldsymbol{v}_i\rangle_K \overline{\langle U^s(m)\boldsymbol{v}, \boldsymbol{v}_j\rangle_K}\, dm \\ &= \frac{\|\boldsymbol{v}\|_K^2 \sum_i |\boldsymbol{v}_i(e)|^2}{\dim(\upsilon)}.\end{aligned}$$

Complete $\{\boldsymbol{v}_j\}$ into an orthonormal basis $\{\boldsymbol{v}_j\} \cup \{\boldsymbol{w}_i\}$ for all of $\mathcal{U}(\upsilon, s)_\tau$. Note that $\boldsymbol{w}_i(e) = 0$ for all i by Lemma 3.2. Therefore we have

$$\sum_j |\boldsymbol{v}_j(e)|^2 = \sum_j |\boldsymbol{v}_j(e)|^2 + \sum_i |\boldsymbol{w}_i(e)|^2 = \dim(\tau). \tag{3.4}$$

Hence the claim (1) follows. Claim (2) follows from (1). □

For K-types τ_1, τ_2 of $\mathcal{U}(\upsilon, s)$, we define an operator

$$T_{\tau_1}^{\tau_2} : \mathcal{U}(\upsilon, s)_{\tau_1} \to \mathcal{U}(\upsilon, s)_{\tau_2}$$

by

$$T_{\tau_1}^{\tau_2}\boldsymbol{v} = \int_M \boldsymbol{v}(m) U^s(m)\boldsymbol{\chi}_{\tau_2}\, dm \qquad \text{for all } \boldsymbol{v} \in \mathcal{U}(\upsilon, s)_{\tau_1}. \tag{3.5}$$

We have the following interpretation of the Eisenstein integral (a similar formula appears in [27, Theorem 3.4]):

Theorem 3.4. *For any K-types τ_1, τ_2 of $\mathcal{U}(\upsilon, s)$ and $g \in G$, we have*

$$\mathsf{P}_{\tau_2} U^s(g) \mathsf{P}_{\tau_1} = \int_K e^{(s-d)H(gk)} U^s\big(\kappa(gk)\big) T_{\tau_1}^{\tau_2} U^s(k^{-1})\, dk,$$

where $\mathsf{P}_{\tau_2} U^s(g)\mathsf{P}_{\tau_1}$ is viewed as an operator from $\mathcal{U}(\upsilon, s)_{\tau_1}$ to $\mathcal{U}(\upsilon, s)_{\tau_2}$.

Proof. Following [47, Theorem 6.2.2.4, pp. 42-43], the key fact is that for any $g \in G$,

$$dk = e^{d\,H(gk)} d\big(\kappa(gk)\big).$$

Let $\boldsymbol{v} \in \mathcal{U}(\upsilon, s)_{\tau_1}$ and $\boldsymbol{w} \in \mathcal{U}(\upsilon, s)_{\tau_2}$. Then

$$\begin{aligned}\langle U^s(g)\boldsymbol{v}, \boldsymbol{w}\rangle_K &= \int_K e^{-sH(g^{-1}k)} \boldsymbol{v}\big(\kappa(g^{-1}k)\big)\overline{\boldsymbol{w}(k)}\, dk \\ &= \int_K e^{-sH(g^{-1}k)} \boldsymbol{v}\big(\kappa(g^{-1}k)\big)\overline{\boldsymbol{w}(k)}\, e^{d\,H(g^{-1}k)}\, d\big(\kappa(g^{-1}k)\big).\end{aligned}$$

We now carry out the change of variables $\tilde{k} = \kappa(g^{-1}k)$; this gives $k = \kappa(g\tilde{k})$, allowing us to rewrite the above integral as follows

$$\begin{aligned}&= \int_K e^{(d-s)H\left(g^{-1}\kappa(gk)\right)} \boldsymbol{v}(k)\overline{\boldsymbol{w}\big(\kappa(gk)\big)}\; dk \\ &= \int_K e^{(s-d)H(gk)} \boldsymbol{v}(k)\overline{\boldsymbol{w}\big(\kappa(gk)\big)}\, dk,\end{aligned}$$

where we used the identity

$$g^{-1}\kappa(gk) = k\exp(-H(gk))\big(\exp(H(gk)n_{kg}^{-1}\exp(-H(gk)\big)$$

to obtain $H(g^{-1}\kappa(gk)) = -H(gk)$.

By (3.3), we get

$$\begin{aligned}\boldsymbol{v}(k)\overline{\boldsymbol{w}\big(\kappa(gk)\big)} =& \langle \boldsymbol{v}, U^s(k)\boldsymbol{\chi}_{\tau_1}\rangle_K \cdot \overline{\langle \boldsymbol{w}, U^s\big(\kappa(gk)\big)\boldsymbol{\chi}_{\tau_2}\rangle_K} \\ =& \langle U^s(k^{-1})\boldsymbol{v}, \boldsymbol{\chi}_{\tau_1}\rangle_K \cdot \langle U^s\big(\kappa(gk)\big)\boldsymbol{\chi}_{\tau_2}, \boldsymbol{w}\rangle_K \\ =& \Big\langle \langle U^s(k^{-1})\boldsymbol{v}, \boldsymbol{\chi}_{\tau_1}\rangle \cdot U^s\big(\kappa(gk)\big)\boldsymbol{\chi}_{\tau_2}, \boldsymbol{w}\Big\rangle_K.\end{aligned}$$

Hence

$$\begin{aligned}\langle U^s(g)\boldsymbol{v}, \boldsymbol{w}\rangle_K &= \int_K e^{(s-d)H(gk)} \Big\langle \langle U^s(k^{-1})\boldsymbol{v}, \boldsymbol{\chi}_{\tau_1}\rangle \cdot U^s\big(\kappa(gk)\big)\boldsymbol{\chi}_{\tau_2}, \boldsymbol{w}\Big\rangle_K\, dk \\ &= \Big\langle \int_K e^{(s-d)H(gk)} \langle U^s(k^{-1})\boldsymbol{v}, \boldsymbol{\chi}_{\tau_1}\rangle \cdot U^s\big(\kappa(gk)\big)\boldsymbol{\chi}_{\tau_2}\, dk, \boldsymbol{w}\Big\rangle_K.\end{aligned}$$

We define $T_0 : \mathcal{U}(\upsilon, s)_{\tau_1} \to \mathcal{U}(\upsilon, s)_{\tau_2}$ by

$$T_0\boldsymbol{v} = \langle \boldsymbol{v}, \boldsymbol{\chi}_{\tau_1}\rangle_K \cdot \boldsymbol{\chi}_{\tau_2} = \boldsymbol{v}(e)\boldsymbol{\chi}_{\tau_2}.$$

Then

$$\mathsf{P}_{\tau_2}U^s(g)\mathsf{P}_{\tau_1} = \int_K e^{(s-d)H(gk)} U^s\big(\kappa(gk)\big)T_0U^s(k^{-1})\, dk. \tag{3.6}$$

Using (3.1) and (3.5), we observe that

$$T_{\tau_1}^{\tau_2} = \int_M U^s(m)T_0U^s(m^{-1})\, dm.$$

Writing $g = k_1 a_t k_2$, we then have

$$\begin{aligned}\mathsf{P}_{\tau_2} U^s(g) \mathsf{P}_{\tau_1} &= U^s(k_1) \mathsf{P}_{\tau_2} U^s(a_t) \mathsf{P}_{\tau_1} U^s(k_2) \\ &= \int_M U^s(k_1) U^s(m) \mathsf{P}_{\tau_2} U^s(a_t) \mathsf{P}_{\tau_1} U^s(m^{-1}) U^s(k_2)\, dm,\end{aligned}$$

where we used the facts that $M = Z_K(A)$ and $M \subset K$ to commute $U^s(m)$ past $U^s(a_t)$ and P_{τ_i}. Now using (3.6) gives

$$\begin{aligned}&= U^s(k_1) \left(\int_M \int_K e^{(s-d)H(a_t k)} U^s\big(m\kappa(a_t k)\big) T_0 U^s(k^{-1} m^{-1})\, dk\, dm \right) U^s(k_2) \\ &= \int_K e^{(s-d)H(a_t k)} U^s\big(k_1 \kappa(a_t k)\big) \left(\int_M U^s(m) T_0 U^s(m^{-1})\, dm \right) U^s(k^{-1} k_2)\, dk.\end{aligned}$$

Finally, using the identities $k_1 \kappa(a_t k) = \kappa(k_1 a_t k)$ and $H(a_t k) = H(k_1 a_t k)$ together with the change of variables $k = k_2 \tilde{k}$ gives

$$\mathsf{P}_{\tau_2} U^s(g) \mathsf{P}_{\tau_1} = \int_K e^{(s-d)H(g\tilde{k})} U^s\big(\kappa(g\tilde{k})\big) T_{\tau_1}^{\tau_2} U^s(\tilde{k}^{-1})\, d\tilde{k},$$

proving the first identity claimed in the theorem. The second identity follows from the fact that $\langle U^s(m^{-1})\boldsymbol{v}, \boldsymbol{\chi}_{\tau_1} \rangle_K = \boldsymbol{v}(m)$ for all $\boldsymbol{v} \in \mathcal{U}(\upsilon, s)_{\tau_1}$ and $m \in M$. $\square$

3.4. **On the operator $T_{\tau_1}^{\tau_2}$.** Using Lemma 3.2, we deduce:

Proposition 3.5. *Let τ_1, τ_2 be K-types of $\mathcal{U}(\upsilon, s)$. If $\boldsymbol{v} \in \mathcal{U}(\upsilon, s)_{\tau_1}$ is orthogonal to $\mathcal{U}(\upsilon, s)_{\upsilon^*}$, then*

$$T_{\tau_1}^{\tau_2} \boldsymbol{v} = 0.$$

Consequently, $T_{\tau_1}^{\tau_2} \mathcal{U}(\upsilon, s)_{\tau_1} \subset \mathcal{U}(\upsilon, s)_{\tau_2} \cap \mathcal{U}(\upsilon, s)_{\upsilon^*}$.

Proof. We have $T_{\tau_1}^{\tau_2} \boldsymbol{v} = \int_M \boldsymbol{v}(m) U^s(m) \boldsymbol{\chi}_{\tau_2}\, dm$. By Lemma 3.2, if $\boldsymbol{v}$ is orthogonal to $\mathcal{U}(\upsilon, s)_{\upsilon^*}$, then $\boldsymbol{v}(m) = 0$ for all $m \in M$ and thus $T_{\tau_1}^{\tau_2} \boldsymbol{v} = 0$. A direct computation using (3.5) shows that $T_{\tau_1}^{\tau_2}$ commutes with $U^s(m)$ for all $m \in M$, so Schur's lemma then gives that if $T_{\tau_1}^{\tau_2} \boldsymbol{v}$ is non-zero, it must be contained in $\mathcal{U}(\upsilon, s)_{\tau_2} \cap \mathcal{U}(\upsilon, s)_{\upsilon^*}$. $\square$

If $\boldsymbol{v}$ is M-invariant and $\upsilon \in \hat{M}$ is non-trivial, then υ^* is non-trivial, and hence $\boldsymbol{v}$ is orthogonal to all $\mathcal{U}(\upsilon, s)_{\upsilon^*}$. Therefore we deduce the following corollary:

Corollary 3.6. *If $\upsilon \in \hat{M}$ is non-trivial, then for any M-invariant $\boldsymbol{v} \in \mathcal{U}(\upsilon, s)_{\tau_1}$,*

$$T_{\tau_1}^{\tau_2} \boldsymbol{v} = 0.$$

Lemma 3.7. *For any K-types τ_1, τ_2 of $\mathcal{U}(\upsilon, s)$, we have*

$$(T_{\tau_1}^{\tau_2})^* = T_{\tau_2}^{\tau_1},$$

where the adjoint is defined with respect to $\langle \cdot, \cdot \rangle_K$.

Proof. For any $\boldsymbol{u} \in \mathcal{U}(\upsilon, s)_{\tau_1}$ and $\boldsymbol{v} \in \mathcal{U}(\upsilon, s)_{\tau_2}$, we have

$$\begin{aligned}\langle T_{\tau_1}^{\tau_2}\boldsymbol{u}, \boldsymbol{v}\rangle_K &= \int_M \boldsymbol{u}(m)\langle U^s(m)\boldsymbol{\chi}_{\tau_2}, \boldsymbol{v}\rangle_K \, dm \\ &= \int_M \langle \boldsymbol{u}, U^s(m)\boldsymbol{\chi}_{\tau_1}\rangle_K \langle U^s(m)\boldsymbol{\chi}_{\tau_2}, \boldsymbol{v}\rangle_K \, dm \\ &= \left\langle \boldsymbol{u}, \int_M \langle U^s(m^{-1})\boldsymbol{v}, \boldsymbol{\chi}_{\tau_2}\rangle_K \cdot U^s(m)\boldsymbol{\chi}_{\tau_1} \, dm \right\rangle \\ &= \langle \boldsymbol{u}, T_{\tau_2}^{\tau_1}\boldsymbol{v}\rangle_K.\end{aligned}$$

□

Corollary 3.8. *For any K-types τ_1, τ_2 of $\mathcal{U}(\upsilon, s)$,*

$$\|T_{\tau_1}^{\tau_2}\|_K \le \frac{\sqrt{\dim(\tau_1)\dim(\tau_2)}}{\dim(\upsilon)}.$$

Proof. For any unit vector $\boldsymbol{v} \in \mathcal{U}(\upsilon, s)_{\tau_1} \cap \mathcal{U}(\upsilon, s)_{\upsilon^*}$, we have

$$\begin{aligned}\langle T_{\tau_1}^{\tau_2}\boldsymbol{v}, T_{\tau_1}^{\tau_2}\boldsymbol{v}\rangle_K &= \int_M \overline{\boldsymbol{v}(m)}\langle T_{\tau_1}^{\tau_2}\boldsymbol{v}, U^s(m)\boldsymbol{\chi}_{\tau_2}\rangle_K \, dm \\ &= \int_M \int_M \overline{\boldsymbol{v}(m_1)}\boldsymbol{v}(m_2)\langle U^s(m_2)\boldsymbol{\chi}_{\tau_2}, U^s(m_1)\boldsymbol{\chi}_{\tau_2}\rangle_K \, dm_2 \, dm_1 \\ &= \int_M \int_M \overline{\langle \boldsymbol{v}, U^s(m_1)\boldsymbol{\chi}_{\tau_1}\rangle_K}\langle \boldsymbol{v}, U^s(m_2)\boldsymbol{\chi}_{\tau_1}\rangle_K \boldsymbol{\chi}_{\tau_2}(m_2^{-1}m_1) \, dm_2 \, dm_1 \\ &= \int_M \left(\int_M \overline{\langle \boldsymbol{v}, U^s(m_2m_3)\boldsymbol{\chi}_{\tau_1}\rangle_K}\langle \boldsymbol{v}, U^s(m_2)\boldsymbol{\chi}_{\tau_1}\rangle_K \, dm_2\right) \boldsymbol{\chi}_{\tau_2}(m_3) \, dm_3.\end{aligned}$$

Since $\boldsymbol{v}$ and $\boldsymbol{\chi}_{\tau_1}$ are both in $\mathcal{U}(\upsilon, s)_{\tau_1} \cap \mathcal{U}(\upsilon, s)_{\upsilon^*}$, the Schur orthogonality relations for M give

$$\begin{aligned}&\int_M \overline{\langle \boldsymbol{v}, U^s(m_2m_3)\boldsymbol{\chi}_{\tau_1}\rangle_K}\langle \boldsymbol{v}, U^s(m_2)\boldsymbol{\chi}_{\tau_1}\rangle_K \, dm_2 \\ &\qquad = \frac{\|\boldsymbol{v}\|^2}{\dim(\upsilon)}\overline{\langle \boldsymbol{\chi}_{\tau_1}, U^s(m_3)\boldsymbol{\chi}_{\tau_1}\rangle} = \frac{\overline{\boldsymbol{\chi}_{\tau_1}(m_3)}}{\dim(\upsilon)}.\end{aligned}$$

We thus have

$$\begin{aligned}\langle T_{\tau_1}^{\tau_2}\boldsymbol{v}, T_{\tau_1}^{\tau_2}\boldsymbol{v}\rangle_K &= \frac{1}{\dim(\upsilon)}\int_M \overline{\boldsymbol{\chi}_{\tau_1}(m)}\boldsymbol{\chi}_{\tau_2}(m) \, dm \\ &\le \frac{1}{\dim(\upsilon)}\sqrt{\int_M |\boldsymbol{\chi}_{\tau_1}(m)|^2 \, dm \int_M |\boldsymbol{\chi}_{\tau_2}(m)|^2 \, dm} \\ &= \frac{\dim(\tau_1)\dim(\tau_2)}{\dim(\upsilon)^2}\end{aligned}$$

by Lemma 3.3 (2). Combining this estimate with Proposition 3.5, the claim follows. □

Corollary 3.9. *For any K-type τ of $\mathcal{U}(\upsilon, s)$, we have*

$$T_\tau^\tau = \frac{\dim(\tau)}{\dim(\upsilon)} \mathsf{P}_{\upsilon^*}.$$

Proof. Following the proof of Corollary 3.8, we obtain

$$\|T_\tau^\tau\|_K^2 = \frac{1}{\dim(\upsilon)} \int_M |\boldsymbol{\chi}_\tau(m)|^2 \, dm = \frac{\dim(\tau)^2}{\dim(\upsilon)^2}.$$

So by Proposition 3.5, we have

$$T_\tau^\tau = c \cdot \mathsf{P}_{\upsilon^*},$$

where c is one of $\pm\frac{\dim(\tau)}{\dim(\upsilon)}$. Since

$$c = \langle T_\tau^\tau \boldsymbol{v}, \boldsymbol{v} \rangle_K = \int_M |\boldsymbol{v}(m)|^2 \, dm \geq 0$$

for any unit vector $\boldsymbol{v} \in \mathcal{U}(\upsilon, s)_\tau \cap \mathcal{U}(\upsilon, s)_{\upsilon^*}$, we get $c = \frac{\dim(\tau)}{\dim(\upsilon)}$. □

4. Asymptotic expansions of matrix coefficients

We fix a complementary series representation $\mathcal{U}(\upsilon, s)$ of G for some $\upsilon \in \hat{M}$ and $s \in \mathcal{I}_\upsilon$. The main goal of this section is to obtain effective expansions of matrix coefficients for $\mathcal{U}(\upsilon, s)$. More precisely, we will write a matrix coefficient $\langle U^s(a_t)\boldsymbol{v}, \boldsymbol{u} \rangle_{\mathcal{U}(\upsilon,s)}$ as a main term that decays like $e^{(s-d)t}$ as $t \to \infty$ and an error term that decays exponentially faster depending only on $2s - d > 0$. To do this, we first work out the asymptotics of matrix coefficients with respect to $\langle \cdot, \cdot \rangle_K$ and then use explicit formulas for intertwining operators to convert our results into statements for $\langle \cdot, \cdot \rangle_{\mathcal{U}(\upsilon,s)}$.

4.1. **Matrix coefficients with respect to $\langle \cdot, \cdot \rangle_K$.** We start by proving a bound on how far elements of K move vectors in representations of K. We let dist_K denote the bi-invariant Riemannian metric on K induced from the negative of the Killing form on $\mathfrak{k}$, and let $|\cdot|_K$ denote the corresponding norm on $\mathfrak{k}$.

Lemma 4.1. *Let (π, V) be a finite-dimensional unitary representation of K with invariant inner product $(\cdot, \cdot)_V$. Then for all $\boldsymbol{v} \in V$ and $k \in K$, we have*

$$\|\pi(k)\boldsymbol{v} - \boldsymbol{v}\|_V \ll \mathrm{dist}_K(k, e) \, \|\boldsymbol{v}\|_{\mathcal{S}^1(V)},$$

where the implied constant depends solely on the choice of basis defining $\|\cdot\|_{\mathcal{S}^m(V)}$.

Proof. Since dist_K is bi-invariant, there exists $J \in \mathfrak{k}$ with $|J|_K = 1$ such that the $k = \exp(\mathrm{dist}_K(k, e)J)$. This gives

$$\begin{aligned}
\pi(k)\boldsymbol{v} - \boldsymbol{v} &= \int_0^{\mathrm{dist}_K(k,e)} \frac{d}{dt} \pi\big(\exp(tJ)\big)\boldsymbol{v} \, dt \\
&= \int_0^{\mathrm{dist}_K(k,e)} \pi(\exp(tJ)) d\pi\big(\mathrm{Ad}(\exp(-tJ))J\big)\boldsymbol{v} \, dt.
\end{aligned}$$

Since π is unitary, $|J|_K = 1$, and K is compact, we get

$$\big\|\pi(\exp(tJ))d\pi\big(\mathrm{Ad}(\exp(-tJ))J\big)\boldsymbol{v}\big\|_V \le \max_{\substack{X\in\mathfrak{k}\\ |X|_K=1}} \|d\pi(X)\boldsymbol{v}\|_V \ll \|\boldsymbol{v}\|_{\mathcal{S}^1(V)}.$$

□

Definition 4.1. The Harish-Chandra c-function $C_+(s) : \mathcal{U}(\upsilon, s) \to \mathcal{U}(\upsilon, s)$ is defined as follows:

$$C_+(s) = \int_{\overline{N}} U^s\big(\kappa(\overline{n})^{-1}\big)e^{-sH(\overline{n})}\,d\overline{n}. \tag{4.1}$$

Since U^s is unitary on $\mathcal{U}(\upsilon, s)$ and $\int_{\overline{N}} e^{-sH(\overline{n})}\,d\overline{n} < \infty$ for all $s > \frac{d}{2}$ (cf. [20, Proposition 7.6]), $C_+(s)$ is a well-defined bounded operator on $\mathcal{U}(\upsilon, s)$. Since $C_+(s)$ is defined using only the restriction of U^s to K, it preserves the K-types of $\mathcal{U}(\upsilon, s)$.

Lemma 4.2. *$C_+(s)$ preserves the M-types of $\mathcal{U}(\upsilon, s)$.*

Proof. Note that M normalizes $\overline{N}$, $d\overline{n} = d(m\overline{n}m^{-1})$ and $H(mgm^{-1}) = H(g)$ for all $m \in M$, $\overline{n} \in \overline{N}$, and $g \in G$. Therefore for all $m \in M$,

$$\begin{aligned} C_+(s)U^s(m) &= \int_{\overline{N}} U^s\big(\kappa(\overline{n})^{-1}m\big)e^{-sH(\overline{n})}\,d\overline{n} \\ &= \int_{\overline{N}} U^s\big(m\kappa(m^{-1}\overline{n}m)^{-1}\big)e^{-sH(\overline{n})}\,d\overline{n} = U^s(m)C_+(s); \end{aligned}$$

hence the claim follows. □

For $d/2 < s < d$, set

$$\eta_s = \min\{2s - d, 1\} > 0. \tag{4.2}$$

We remark that the following theorem was shown in [27, Theorem 3.23] for *some* η_s, based on Harish-Chandra's expansion formula and the maximum modulus principle. We give a more direct argument, with the explicit η_s given in (4.2). Let $\mathcal{S}^1(K)$ denote the Sobolev norm $\mathcal{S}^1(L^2(K))$ defined in Section 2.5.

Theorem 4.3. *Let τ_1, τ_2 be K-types of $\mathcal{U}(\upsilon, s)$. For all $\boldsymbol{u} \in \mathcal{U}(\upsilon, s)_{\tau_1}$ and $\boldsymbol{v} \in \mathcal{U}(\upsilon, s)_{\tau_2}$, we have for any $t \ge 0$,*

$$\begin{aligned} \langle U^s(a_t)\boldsymbol{u}, \boldsymbol{v}\rangle_K = e^{(s-d)t}&\langle T^{\tau_2}_{\tau_1} C_+(s)\boldsymbol{u}, \boldsymbol{v}\rangle_K \\ &+ O_s\left(e^{(s-d-\eta_s)t}\|T^{\tau_2}_{\tau_1}\|_K\|\boldsymbol{u}\|_K\|\boldsymbol{v}\|_{\mathcal{S}^1(K)}\right), \end{aligned}$$

where the implied constant is uniformly bounded over s in compact subsets of $(\frac{d}{2}, d)$.

Proof. Applying Theorem 3.4, we have

$$\langle U^s(a_t)\boldsymbol{u}, \boldsymbol{v}\rangle_K = \int_K e^{(s-d)H(a_tk)}\big\langle U^s\big(\kappa(a_tk)\big)T^{\tau_2}_{\tau_1}U^s(k^{-1})\boldsymbol{u}, \boldsymbol{v}\big\rangle_K\,dk.$$

Since the function $k \mapsto e^{(s-d)H(a_t k)}\langle U^s\big(\kappa(a_t k)\big)T_{\tau_1}^{\tau_2}U^s(k^{-1})\boldsymbol{u},\boldsymbol{v}\rangle_{\mathcal{U}(v,s)}$ is right M-invariant, we may use the integration formula [20, Consequence 3, p. 147] to obtain

$$\begin{aligned}&\langle U^s(a_t)\boldsymbol{u},\boldsymbol{v}\rangle_K\\ &\quad= \int_{\overline{N}} e^{(s-d)H(a_t\kappa(\overline{n}))}\big\langle U^s\big(\kappa(a_t\kappa(\overline{n}))\big)T_{\tau_1}^{\tau_2}U^s(\kappa(\overline{n})^{-1})\boldsymbol{u},\boldsymbol{v}\big\rangle_K e^{-d\,H(\overline{n})}\,d\overline{n}.\end{aligned}$$

The identities

$$\begin{aligned}H\big(a_t\kappa(\overline{n})\big) &= H(a_t\overline{n}a_{-t}) + H(a_t) - H(\overline{n}) \quad \text{ and}\\ \kappa\big(a_t\kappa(\overline{n})\big) &= \kappa(a_t\overline{n}a_{-t})\end{aligned}$$

then give that the previous integral is equal to

$$e^{(s-d)t}\int_{\overline{N}}\big\langle T_{\tau_1}^{\tau_2}U^s(\kappa(\overline{n})^{-1})\boldsymbol{u},U^s\big(\kappa(a_t\overline{n}a_{-t})^{-1}\big)\boldsymbol{v}\big\rangle_K e^{(s-d)H(a_t\overline{n}a_{-t})-s\,H(\overline{n})}\,d\overline{n}.$$

We now use the identification of $\overline{N}$ with $\mathbb{R}^d$ to again rewrite:

$$e^{(s-d)t}\int_{\mathbb{R}^d}\big\langle T_{\tau_1}^{\tau_2}U^s(\kappa(\overline{n}_{\boldsymbol{x}})^{-1})\boldsymbol{u},U^s\big(\kappa(\overline{n}_{e^{-t}\boldsymbol{x}})^{-1}\big)\boldsymbol{v}\big\rangle_K e^{(s-d)H(\overline{n}_{e^{-t}\boldsymbol{x}})-s\,H(\overline{n}_{\boldsymbol{x}})}\,d\boldsymbol{x}; \tag{4.3}$$

note that the integral is absolutely convergent due to $s > \frac{d}{2}$. Using the fact that $e^{H(\overline{n}_{\boldsymbol{x}})} = 1 + \|\boldsymbol{x}\|^2$ (cf. [21, p. 486 and p. 564]) gives

$$\begin{aligned}&\langle U^s(a_t)\boldsymbol{u},\boldsymbol{v}\rangle_K = e^{(s-d)t}\langle T_{\tau_1}^{\tau_2}C_+(s)\boldsymbol{u},\boldsymbol{v}\rangle_K +\\ &e^{(s-d)t}\int_{\mathbb{R}^d}\big\langle T_{\tau_1}^{\tau_2}U^s(\kappa(\overline{n}_{\boldsymbol{x}})^{-1})\boldsymbol{u},(1+\|e^{-t}\boldsymbol{x}\|^2)^{s-d}U^s\big(\kappa(\overline{n}_{e^{-t}\boldsymbol{x}})^{-1}\big)\boldsymbol{v}-\boldsymbol{v}\big\rangle_K\tfrac{d\boldsymbol{x}}{(1+\|\boldsymbol{x}\|^2)^s}.\end{aligned}$$

Changing to spherical coordinates, we let $\boldsymbol{x} = (r,\boldsymbol{\theta})$, $r \geq 0$, $\boldsymbol{\theta} \in \mathbb{S}^{d-1}$ and set

$$\boldsymbol{w}(r,\boldsymbol{\theta}) = (1+r^2)^{s-d}U^s\big(\kappa(\overline{n}_{(r,\boldsymbol{\theta})})^{-1}\big)\boldsymbol{v}-\boldsymbol{v}.$$

Using this, we deduce

$$\begin{aligned}&\langle U^s(a_t)\boldsymbol{u},\boldsymbol{v}\rangle_K = e^{(s-d)t}\langle T_{\tau_1}^{\tau_2}C_+(s)\boldsymbol{u},\boldsymbol{v}\rangle_K +\\ &O\left(e^{(s-d)t}\int_0^\infty\int_{\mathbb{S}^{d-1}}\big\langle T_{\tau_1}^{\tau_2}U^s(\kappa(\overline{n}_{(r,\boldsymbol{\theta})})^{-1})\boldsymbol{u},\boldsymbol{w}(e^{-t}r,\boldsymbol{\theta})\big\rangle_K\,dm(\boldsymbol{\theta})\,\tfrac{r^{d-1}}{(1+r^2)^s}\,dr\right),\end{aligned}$$

where $dm(\boldsymbol{\theta})$ denotes the spherical measure.

Since the map $\overline{n} \mapsto \kappa(\overline{n})$ is smooth, we get

$$d_K\big(\kappa(\overline{n}_{(r,\boldsymbol{\theta})}),e\big) \ll r \qquad \text{for all } r > 0, \boldsymbol{\theta} \in \mathbb{S}^{d-1}.$$

Using $(1+r^2)^{s-d} = 1 - O_s(r)$ (with the implied constant depending continuously on s) and Lemma 4.1, we have

$$\|\boldsymbol{w}(r,\theta)\|_K \ll_s \min\{1,r\}\|\boldsymbol{v}\|_{\mathcal{S}^1(K)}.$$

This gives

$$\langle U^s(a_t)\boldsymbol{u},\boldsymbol{v}\rangle_K = e^{(s-d)t}\langle T_{\tau_1}^{\tau_2} C_+(s)\boldsymbol{u},\boldsymbol{v}\rangle_K \\ + O_s\left(\|T_{\tau_1}^{\tau_2}\|_K\|\boldsymbol{u}\|_K\|\boldsymbol{v}\|_{\mathcal{S}^1(K)} e^{(s-d)t}\int_0^\infty \min\{1,e^{-t}r\}(1+r^2)^{-s}r^{d-1}\,dr\right).$$

The proof is completed by writing the integral $\int_0^\infty$ as $\int_0^1+\int_1^{e^t}+\int_{e^t}^\infty$ to obtain

$$\begin{aligned}&\int_0^\infty \min\{1,e^{-t}r\}(1+r^2)^{-s}r^{d-1}\,dr\\ &\le e^{-t}+e^{-t}\int_1^{e^t} r\,\cdot\, r^{-2s}\,\cdot\, r^{d-1}\,dr+\int_{e^t}^\infty r^{-2s}\,\cdot r^{d-1}\,dr\\ &\ll_s\ e^{-t}+e^{(d-2s)t}.\end{aligned}$$

□

4.2. **The invariant inner product on $\mathcal{U}(\upsilon,s)$.** The intertwining operator $\mathcal{A}(\upsilon,s)$ on $\mathcal{U}(\upsilon,s)$ is defined so that

$$\langle\boldsymbol{u},\boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)}=\langle\boldsymbol{u},\mathcal{A}(\upsilon,s)\boldsymbol{v}\rangle_K$$

for all K-finite vectors $\boldsymbol{u},\boldsymbol{v}\in\mathcal{U}(\upsilon,s)$. The key intertwining property of $\mathcal{A}(\upsilon,s)$ reads (cf. [22, Lemmas 22 and 23])

$$\mathcal{A}(\upsilon,s)U^s(g)=U^{d-s}(g)\mathcal{A}(\upsilon,s)\qquad\text{for all } g\in G. \tag{4.4}$$

In particular, $\mathcal{A}(\upsilon,s)$ commutes with $U^s(k)$ for all $k\in K$. Since each K-type occurs at most once in $\mathcal{U}(\upsilon,s)$, by Schur's lemma, $\mathcal{A}(\upsilon,s)$ acts as a scalar $a(\upsilon,s,\tau)$ on each K-type τ of $\mathcal{U}(\upsilon,s)$:

$$\mathcal{A}(\upsilon,s)=\sum_{\tau\supset\upsilon}a(\upsilon,s,\tau)\mathsf{P}_\tau. \tag{4.5}$$

The positive definiteness of the inner product $\langle\cdot,\cdot\rangle_{\mathcal{U}(\upsilon,s)}$ implies that for all $\tau\in\hat{K}$ contained in $\mathcal{U}(\upsilon,s)$, we have

$$a(\upsilon,s,\tau)>0.$$

Recalling the parameterization of K and M types given in Section 3.1, we now assume that $\mathcal{U}(\upsilon,s)$ has a non-trivial M-invariant vector. There is then a K-type $\sigma=(\sigma_1,\sigma_2,\ldots)$ of $\mathcal{U}(\upsilon,s)$ that contains the trivial representation of M. Thus $(\sigma_1,\sigma_2,\ldots)$ satisfies the interlacing relation with the trivial representation $(0,0,0,\ldots,0)$ of M:

$$\sigma_1\ge 0\ge\sigma_2\ge 0\ge\sigma_3\ge\ldots.$$

From this, we conclude that $\sigma=(\sigma_1,0,0,\ldots,0)$. Now writing $\upsilon=(\upsilon_1,\upsilon_2,\ldots)$, the classification of K-types of $\mathcal{U}(\upsilon,s)$ ensures that υ is a subrepresentation of σ, and so $(\upsilon_1,\upsilon_2,\ldots)$ must satisfy the interlacing relation with $(\sigma_1,0,0,0,\ldots)$. We therefore see that $\upsilon=(\upsilon_1,0,0,0,\ldots)$. For notational convenience we will simply write $\upsilon=(\upsilon,0,0,0,\ldots)\in\mathbb{Z}^{\lfloor\frac{d}{2}\rfloor}$, where

$\upsilon \geq 0$. Note that if $d = 1$ or 2, then by the classification of the unitary dual of $\mathrm{SO}(2,1)$ and $\mathrm{SO}(3,1)$, $\upsilon = 0$ [16]. Combining the fact that each K-type of $\mathcal{U}(\upsilon, s)$ must have υ as a subrepresentation with the interlacing relation gives that all K-types τ of $\mathcal{U}(\upsilon, s)$ may be parameterized as vectors $\tau = (t_1, t_2, 0, \ldots, 0) \in \mathbb{Z}^{\lfloor \frac{d+1}{2} \rfloor}$ with $t_1 \geq \upsilon \geq t_2 \geq 0$ if $d > 3$, $t_1 \geq \upsilon \geq |t_2|$ if $d = 3$, $t_1 \geq 0$ if $d = 2$, and just $t_1 \in \mathbb{Z}$ if $d = 1$. We will thus write $\tau = (t_1, t_2)$ and $a(\upsilon, s, \tau) = a(\upsilon, s, t_1, t_2)$, with $\upsilon = t_2 = 0$ for $d = 1, 2$. Again using the classification of the unitary dual of G, observe that if $s \geq d-1$, then $\upsilon = t_2 = 0$. Finally, we let $\Omega_K \in Z(\mathfrak{k}_{\mathbb{C}})$ denote the Casimir operator of K.

The following lemma is the main technical result needed to establish the bounds on the quotients $\frac{a(\upsilon,s,\tau_1)}{a(\upsilon,s,\tau_2)}$ given in Propositions 4.5 and 4.6.

Lemma 4.4. *Assume that $\mathcal{U}(\upsilon, s)$ has a non-trivial M-invariant vector. Then for any K-types $\tau_1 = (t_1, t_2)$ and $\tau_2 = (t_3, t_4)$ of $\mathcal{U}(\upsilon, s)$, if $s < d-1$,*

$$\frac{a(\upsilon, s, \tau_2)}{a(\upsilon, s, \tau_1)} = \frac{\Gamma(d-s+t_3)}{\Gamma(s+t_3)} \cdot \frac{\Gamma(d-s+t_4-1)}{\Gamma(s+t_4-1)} \cdot \frac{\Gamma(s+t_1)}{\Gamma(d-s+t_1)} \cdot \frac{\Gamma(s+t_2-1)}{\Gamma(d-s+t_2-1)},$$

and if $s \geq d-1$,

$$\frac{a(\upsilon, s, \tau_2)}{a(\upsilon, s, \tau_1)} = \frac{\Gamma(d-s+t_3)}{\Gamma(s+t_3)} \cdot \frac{\Gamma(s+t_1)}{\Gamma(d-s+t_1)}.$$

Proof. The claimed formula follows from a recursion formula similar to that for the Harish-Chandra c-function given in [10]. We start by letting $H \in \mathfrak{a}$ be such that $a_t = \exp(tH)$, and defining an inner product $\langle \cdot, \cdot \rangle_{\mathfrak{g}}$ on $\mathfrak{g}_{\mathbb{C}}$ by

$$\langle X, Y \rangle_{\mathfrak{g}} := c \cdot (-B(X, \theta Y)),$$

where $B(\cdot,\cdot)$ denotes the Killing form on $\mathfrak{g}$, θ is the Cartan involution defining K, and $c \in \mathbb{R}_{>0}$ is chosen so that $\langle H, H \rangle_{\mathfrak{g}} = 1$. Denote the -1-eigenspace of θ by $\mathfrak{p}$. Then for all $X \in \mathfrak{p}_{\mathbb{C}}$, $\boldsymbol{v} \in C^\infty(K)$, and $k \in K$, we have

$$\begin{aligned}[dU^s(X)\boldsymbol{v}](k) &= (s - \tfrac{d}{2})\langle \mathrm{Ad}_{k^{-1}} X, H \rangle_{\mathfrak{g}} \boldsymbol{v}(k) \\ &\quad + \tfrac{1}{2}\Big(d\rho(\Omega_K)\big\{\langle \mathrm{Ad}_{k^{-1}} X, H \rangle_{\mathfrak{g}} \boldsymbol{v}(k)\big\} - \langle \mathrm{Ad}_{k^{-1}} X, H \rangle_{\mathfrak{g}} [d\rho(\Omega_K)\boldsymbol{v}](k)\Big),\end{aligned} \tag{4.6}$$

where ρ denotes right-translation (cf. [42, Lemma 1] and [10, Lemma 3.2]). Let $\tau = (r_1, r_2)$ be an arbitrary K-type of $\mathcal{U}(\upsilon, s)$. By e.g. [42, Lemma 2] (cf. also [21, Proposition 5.28]), for any vector $\boldsymbol{v} \in \mathcal{U}(\upsilon, s)_\tau$,

$$d\rho(\Omega_K)\boldsymbol{v} = dU^s(\Omega_K)\boldsymbol{v} = (r_1^2 + r_2^2 + (d-1)r_1 + (d-3)r_2)\boldsymbol{v}. \tag{4.7}$$

We now denote the orthogonal projection onto τ by $\mathsf{P}_{(r_1,r_2)}$. Combining the above expression for $d\rho(\Omega_K)\boldsymbol{v}$ with (4.6) gives

$$\begin{aligned}
&[\mathsf{P}_{(r_1+1,r_2)} dU^s(X)\boldsymbol{v}](k) \\
&= \Big(s - \tfrac{d}{2} + \tfrac{(r_1+1)^2 + r_2^2 + (d-1)(r_1+1) + (d-3)r_2}{2} - \tfrac{r_1^2 + r_2^2 + (d-1)r_1 + (d-3)r_2}{2}\Big) \\
&\qquad\qquad\qquad \times \mathsf{P}_{(r_1+1,r_2)} \left\{ \langle \mathrm{Ad}_{k^{-1}} X, H\rangle_{\mathfrak{g}} \boldsymbol{v}(k)\right\} \\
&= (s + r_1)\, \mathsf{P}_{(r_1+1,r_2)} \left\{ \langle \mathrm{Ad}_{k^{-1}} X, H\rangle_{\mathfrak{g}} \boldsymbol{v}(k)\right\},
\end{aligned} \tag{4.8}$$

and similarly

$$[\mathsf{P}_{(r_1,r_2+1)} dU^s(X)\boldsymbol{v}](k) = (s + r_2 - 1)\, \mathsf{P}_{(r_1,r_2+1)} \left\{ \langle \mathrm{Ad}_{k^{-1}} X, H\rangle_{\mathfrak{g}} \boldsymbol{v}(k)\right\}. \tag{4.9}$$

From (4.4), we obtain

$$[\mathsf{P}_{(r_1,r_2)} dU^s(X)\mathcal{A}(\upsilon, s)\boldsymbol{v}](k) = [\mathsf{P}_{(r_1,r_2)} \mathcal{A}(\upsilon, s) dU^{d-s}(X)\boldsymbol{v}](k).$$

Combining this with (4.5) and (4.8) and (4.9), respectively, gives

$$\begin{aligned}
&a(\upsilon, s, r_1, r_2)(s + r_1)\mathsf{P}_{(r_1+1,r_2)} \left\{ \langle \mathrm{Ad}_{k^{-1}} X, H\rangle_{\mathfrak{g}} \boldsymbol{v}(k)\right\} \\
&\qquad = a(\upsilon, s, r_1 + 1, r_2)(d - s + r_1)\mathsf{P}_{(r_1+1,r_2)} \left\{ \langle \mathrm{Ad}_{k^{-1}} X, H\rangle_{\mathfrak{g}} \boldsymbol{v}(k)\right\},
\end{aligned}$$

and

$$\begin{aligned}
&a(\upsilon, s, r_1, r_2)(s + r_2 - 1)\mathsf{P}_{(r_1,r_2+1)} \left\{ \langle \mathrm{Ad}_{k^{-1}} X, H\rangle_{\mathfrak{g}} \boldsymbol{v}(k)\right\} \\
&\qquad = a(\upsilon, s, r_1, r_2 + 1)(d - s + r_2 - 1)\mathsf{P}_{(r_1,r_2+1)} \left\{ \langle \mathrm{Ad}_{k^{-1}} X, H\rangle_{\mathfrak{g}} \boldsymbol{v}(k)\right\}.
\end{aligned}$$

The decomposition of $(\mathrm{Ad}, \mathfrak{p}_{\mathbb{C}}) \otimes (U^s, \mathcal{U}(\upsilon, s)_\tau)$ into irreducible representations of K ensures the existence of $X, Y \in \mathfrak{p}_{\mathbb{C}}$ and $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathcal{U}(\upsilon, s)_\tau$ such that $\mathsf{P}_{(r_1+1,r_2)} \left\{ \langle \mathrm{Ad}_{k^{-1}} X, H\rangle_{\mathfrak{g}} \boldsymbol{v}_1(k)\right\} \not\equiv 0$, and $\mathsf{P}_{(r_1,r_2+1)} \left\{ \langle \mathrm{Ad}_{k^{-1}} Y, H\rangle_{\mathfrak{g}} \boldsymbol{v}_2(k)\right\} \not\equiv 0$ if $(r_1 + 1, r_2)$, and $(r_1, r_2 + 1)$ are K-types of $\mathcal{U}(\upsilon, s)$, cf., e.g., [42, Lemma 3] and [44, Theorem 3.4.12]. This gives

$$(s + r_1) \cdot a(\upsilon, s, r_1 + 1, r_2) = (d - s + r_1) \cdot a(\upsilon, s, r_1, r_2)$$

and

$$(s + r_2 - 1) \cdot a(\upsilon, s, r_1, r_2 + 1) = (d - s + r_2 - 1) \cdot a(\upsilon, s, r_1, r_2)$$

(cf. [10, (6.1) and (6.7)]). Note that if one of the four factors that are multiplied with $a(\upsilon, s, \ldots)$'s is zero, then the representation given by that choice of υ and s is not in the unitary dual of G. In particular, if $s \geq d - 1$, then $\mathcal{U}(\upsilon, s)$ is spherical, hence $\upsilon = r_2 = 0$. These two recursion formulas imply that if $\upsilon \neq 0$,

$$a(\upsilon, s, r_1, r_2) = \tfrac{\Gamma(d-s+r_1)}{\Gamma(s+r_1)} \cdot \tfrac{\Gamma(d-s+r_2-1)}{\Gamma(s+r_2-1)} \cdot \tfrac{\Gamma(s+\upsilon)}{\Gamma(d-s+\upsilon)} \cdot \tfrac{\Gamma(s+\upsilon-1)}{\Gamma(d-s+\upsilon-1)} \cdot a(\upsilon, s, \upsilon, \upsilon),$$

and if $\upsilon = 0$,

$$a(0, s, r_1, 0) = \tfrac{\Gamma(d-s+r_1)}{\Gamma(s+r_1)} \cdot \tfrac{\Gamma(s)}{\Gamma(d-s)} \cdot a(0, s, 0, 0).$$

The formulas claimed in the proposition then follow. $\square$

By Schur's lemma, Ω_K acts on any realization of a K-type τ by the same scalar, which we denote $\tau(\Omega_K)$.

Proposition 4.5. *Assume that $\mathcal{U}(\upsilon, s)$ has a non-trivial M-invariant vector. Then for any K-types τ_1, τ_2 of $\mathcal{U}(\upsilon, s)$,*

$$\frac{a(\upsilon, s, \tau_2)}{a(\upsilon, s, \tau_1)} \ll_s \left(1 + \tau_1(\Omega_K)^d\right)\left(1 + \tau_2(\Omega_K)^d\right),$$

and the implied constant is uniformly bounded over s in compact subsets of $\mathcal{I}_\upsilon$.

Proof. The key result needed in the proof is a consequence of [17, Theorem 1]: for all $t \geq 0$,

$$\frac{\Gamma(s+t)}{\Gamma(d-s+t)} \asymp_s 1 + t^{2s-d}, \tag{4.10}$$

where the implied constant is uniformly bounded over s in compact subsets of $(\frac{d}{2}, d)$. We now write $\tau_1 = (t_1, t_2)$, $\tau_2 = (t_3, t_4)$.

Assuming first that $d > 3$ and $\upsilon > 0$, we then have $t_i \geq 0$ for all i. Since $\upsilon > 0$, $s < d - 1$, and so rewriting Lemma 4.4 gives

$$\frac{a(\upsilon,s,\tau_2)}{a(\upsilon,s,\tau_1)} = \frac{\Gamma(d-s+t_3)}{\Gamma(s+t_3)} \cdot \frac{\Gamma(d-s+t_4)}{\Gamma(s+t_4)} \cdot \frac{\Gamma(s+t_1)}{\Gamma(d-s+t_1)} \cdot \frac{\Gamma(s+t_2)}{\Gamma(d-s+t_2)} \cdot \frac{s+t_4-1}{d-s+t_4-1} \cdot \frac{d-s+t_2-1}{s+t_2-1},$$

and so (after recalling the formula (4.7) for $\tau_1(\Omega_K)$ and $\tau_2(\Omega_K)$),

$$\begin{aligned}\frac{a(\upsilon,s,\tau_2)}{a(\upsilon,s,\tau_1)} &\asymp_s (1 + t_1^{2s-d})(1 + t_2^{2s-d})(1 + t_3^{2s-d})(1 + t_3^{2s-d}) \\ &\ll \left(1 + \tau_1(\Omega_K)^d\right)\left(1 + \tau_2(\Omega_K)^d\right),\end{aligned}$$

with the implied constant uniformly bounded over $(\frac{d}{2}, d-1) = \mathcal{I}_\upsilon$.

In the case $d \geq 2$ and $\upsilon = 0$ (and so $t_2 = t_4 = 0$), Lemma 4.4 gives

$$\frac{a(\upsilon,s,\tau_2)}{a(\upsilon,s,\tau_1)} = \frac{\Gamma(d-s+t_3)}{\Gamma(s+t_3)} \cdot \frac{\Gamma(s+t_1)}{\Gamma(d-s+t_1)},$$

so by a direct application of (4.7) and (4.10),

$$\frac{a(\upsilon,s,\tau_2)}{a(\upsilon,s,\tau_1)} \ll_s (1 + t_1^{2s-d})(1 + t_3^{2s-d}) \ll \left(1 + \tau_1(\Omega_K)^{\frac{d}{2}}\right)\left(1 + \tau_2(\Omega_K)^{\frac{d}{2}}\right), \tag{4.11}$$

with the implied constant uniformly bounded over s in compact subsets of $(\frac{d}{2}, d)$.

For the remaining cases, $d = 1$ and $d = 3$ with $\upsilon > 0$, negative values of the t_i can appear in the formulas given in Lemma 4.4. It thus remains to bound quotients of the form $\frac{\Gamma(s+t)}{\Gamma(d-s+t)}$, where t is a negative integer and $s \in (\frac{1}{2}, 1)$ if $d = 1$, and $s \in (\frac{3}{2}, 2)$ if $d = 3$. By the reflection formula, in both cases,

$$\frac{\Gamma(s+t)}{\Gamma(d-s+t)} = \frac{\Gamma(s-|t|)}{\Gamma(d-s-|t|)} = \frac{\Gamma(|t|+1+s-d)}{\Gamma(|t|+1-s)}. \tag{4.12}$$

If $d = 1$, we then have

$$\frac{\Gamma(s+t)}{\Gamma(1-s+t)} = \frac{\Gamma(|t|+s)}{\Gamma(|t|+1-s)},$$

and so (4.10) gives

$$\frac{a(\upsilon,s,\tau_2)}{a(\upsilon,s,\tau_1)} = \frac{\Gamma(1-s+|t_3|)}{\Gamma(s+|t_3|)} \cdot \frac{\Gamma(s+|t_1|)}{\Gamma(1-s+|t_1|)} \tag{4.13}$$
$$\ll_s (1+|t_1|^{2s-1})(1+|t_2|^{2s-1}) \ll \big(1+\tau_1(\Omega_K)^{\frac{1}{2}}\big)\big(1+\tau_2(\Omega_K)^{\frac{1}{2}}\big).$$

If $d=3$, using (4.12), we have

$$\frac{\Gamma(s+t)}{\Gamma(3-s+t)} = \frac{\Gamma(|t|+s-2)}{\Gamma(|t|+2-s)} = \frac{(|t|+1-s)(|t|+2-s)}{(s+|t|-1)(s+|t|-2)} \cdot \frac{\Gamma(s+|t|)}{\Gamma(3-s+|t|)}.$$

Now using $\frac{(|t|+1-s)(|t|+2-s)}{(s+|t|-1)(s+|t|-2)} \asymp_s 1$, where the implied constants are uniformly bounded over s in compact subsets of $(\frac{3}{2},2)$, together with (4.10) as previously completes the proof. ☐

For K-types that contain M-invariant vectors, we have the following strengthening of the bound in Proposition 4.5:

Proposition 4.6. *Let τ_1,τ_2 be K-types of $\mathcal{U}(\upsilon,s)$ containing non-trivial M-invariant vectors. Then*

$$\frac{a(\upsilon,s,\tau_2)}{a(\upsilon,s,\tau_1)} \ll_s \big(1+\tau_1(\Omega_K)^{\frac{d}{2}}\big)\big(1+\tau_2(\Omega_K)^{\frac{d}{2}}\big),$$

and the implied constant is uniformly bounded over s in compact subsets of $(\frac{d}{2},d)$.

Proof. For $d\geq 3$, we observe that if a K-type $\tau=(r_1,r_2)$ contains an M-invariant vector, the interlacing relation implies $r_2=0$, hence for t_1,t_2 as in the statement of the lemma, we have $\tau_1=(t_1,0)$ and $\tau_2=(t_2,0)$. Lemma 4.4 in this case reads

$$\frac{a(\upsilon,s,\tau_2)}{a(\upsilon,s,\tau_1)} = \frac{\Gamma(d-s+t_2)}{\Gamma(s+t_2)} \cdot \frac{\Gamma(s+t_1)}{\Gamma(d-s+t_1)}.$$

The lemma then follows from (4.11) for $d\geq 2$ and (4.13) for $d=1$. ☐

Remark 4.14. The main point of Proposition 4.6 is that the implied constant remains bounded even as s approaches $d-1$ in the case $\upsilon>0$. This allows us to use the bound uniformly over complementary series representations appearing in the direct integral decomposition of $L^2(\Gamma\backslash G)$.

The bounds from Propositions 4.5 and 4.6 allow us to restate Theorem 4.3 in terms of $\langle\cdot,\cdot\rangle_{\mathcal{U}(\upsilon,s)}$. This can then be applied to the matrix coefficients of irreducible unitary representations weakly contained in $L^2(\Gamma\backslash G)$. Retaining the notation of Theorem 4.3, we have:

Proposition 4.7. *There exists $m\in\mathbb{N}$ such that for any $\mathcal{U}(\upsilon,s)$ with a non-trivial M-invariant vector, for all $\boldsymbol{u}\in\mathcal{U}(\upsilon,s)_{\tau_1}, \boldsymbol{v}\in\mathcal{U}(\upsilon,s)_{\tau_2}$, and $t\geq 0$,*

$$\begin{aligned}\langle U^s(a_t)\boldsymbol{u},\boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)} =& e^{(s-d)t}\langle T_{\tau_1}^{\tau_2}C_+(s)\boldsymbol{u},\boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)} \\ &+ O_s\big(e^{(s-d-\eta_s)t}\|\boldsymbol{u}\|_{\mathcal{S}^m(\upsilon,s)}\|\boldsymbol{v}\|_{\mathcal{S}^m(\upsilon,s)}\big),\end{aligned}$$

where the implied constant is uniformly bounded over s in compact subsets of $\mathcal{I}_\upsilon$. Furthermore, if the K-types τ_1 and τ_2 both contain non-trivial M-invariant vectors, the the implied constant is uniformly bounded over s in compact subsets of $(\frac{d}{2}, d)$.

Proof. Since $\boldsymbol{v} \in \mathcal{U}(\upsilon, s)_{\tau_2}$, using the expression for $\langle \cdot, \cdot \rangle_{\mathcal{U}(\upsilon,s)}|_{\mathcal{U}(\upsilon,s)_{\tau_2}}$, Theorem 4.3 gives

$$\begin{aligned}
&\langle U^s(a_t)\boldsymbol{u}, \boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)} = a(\upsilon, s, \tau_2)\langle U^s(a_t)\boldsymbol{u}, \boldsymbol{v}\rangle_K \\
&= e^{(s-d)t} a(\upsilon, s, \tau_2)\langle T_{\tau_1}^{\tau_2} C_+(s)\boldsymbol{u}, \boldsymbol{v}\rangle_K \\
&\qquad + O_s\left(e^{(s-d-\eta_s)t} a(\upsilon, s, \tau_2)\|T_{\tau_1}^{\tau_2}\|_K \|\boldsymbol{u}\|_K \|\boldsymbol{v}\|_{\mathcal{S}^1(K)}\right) \\
&= e^{(s-d)t}\langle T_{\tau_1}^{\tau_2} C_+(s)\boldsymbol{u}, \boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)} \\
&\qquad + O_s\left(e^{(s-d-\eta_s)t}\sqrt{\tfrac{a(\upsilon,s,\tau_2)}{a(\upsilon,s,\tau_1)}}\|T_{\tau_1}^{\tau_2}\|_K \|\boldsymbol{u}\|_{\mathcal{U}(\upsilon,s)} \|\boldsymbol{v}\|_{\mathcal{S}^1(\upsilon,s)}\right).
\end{aligned}$$

By Proposition 4.5, or Proposition 4.6 if τ_1 and τ_2 both have M-invariant vectors,

$$\begin{aligned}
&\sqrt{\tfrac{a(\upsilon,s,\tau_2)}{a(\upsilon,s,\tau_1)}}\|\boldsymbol{u}\|_{\mathcal{U}(\upsilon,s)}\|\boldsymbol{v}\|_{\mathcal{S}^1(\upsilon,s)} \\
&\qquad \ll_s \sqrt{(1+\tau_1(\Omega_K)^d)(1+\tau_2(\Omega_K)^d)}\|\boldsymbol{u}\|_{\mathcal{U}(\upsilon,s)}\|\boldsymbol{v}\|_{\mathcal{S}^1(\upsilon,s)} \\
&\qquad \ll \|(1+dU^s(\Omega_K^{\lceil d/2\rceil}))\boldsymbol{u}\|_{\mathcal{U}(\upsilon,s)}\|(1+dU^s(\Omega_K^{\lceil d/2\rceil}))\boldsymbol{v}\|_{\mathcal{S}^1(\upsilon,s)} \\
&\qquad \ll \|\boldsymbol{u}\|_{\mathcal{S}^{d+1}(\upsilon,s)}\|\boldsymbol{v}\|_{\mathcal{S}^{d+2}(\upsilon,s)},
\end{aligned}$$

with the implied constant uniformly bounded over s in compact subsets of $\mathcal{I}_\upsilon$, or $(\frac{d}{2}, d)$, respectively. Now Corollary 3.8 and Lemma 3.1 give

$$\begin{aligned}
\|T_{\tau_1}^{\tau_2}\|_K\|\boldsymbol{u}\|_{\mathcal{S}^{d+1}(\upsilon,s)}\|\boldsymbol{v}\|_{\mathcal{S}^{d+2}(\upsilon,s)} &\le \sqrt{\dim(\tau_1)\dim(\tau_2)}\|\boldsymbol{u}\|_{\mathcal{S}^{d+1}(\upsilon,s)}\|\boldsymbol{v}\|_{\mathcal{S}^{d+2}(\upsilon,s)} \\
&\ll \|\boldsymbol{u}\|_{\mathcal{S}^m(\upsilon,s)}\|\boldsymbol{v}\|_{\mathcal{S}^m(\upsilon,s)}
\end{aligned}$$

for some $m \in \mathbb{N}$, completing the proof. $\square$

Remark 4.15. Both Proposition 4.5 and Proposition 4.7 are expected to hold for all complementary series $\mathcal{U}(\upsilon, s)$. We have proved them only for those representations with M-invariant vectors since in this case slight simplifications occur in the proof of Lemma 4.4. Note that if Proposition 4.5 were proved for all complementary series representations, Proposition 4.7 would also follow automatically for all complementary series.

Theorem 4.8. *There exists $m \in \mathbb{N}$ such that for any complementary series representation $\mathcal{U}(\upsilon, s)$ containing a non-trivial M-invariant vector, for all $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{S}^m(\upsilon, s)$ and $t \ge 0$,*

$$\begin{aligned}
\langle U^s(a_t)\boldsymbol{u}, \boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)} = &e^{(s-d)t}\left(\sum_{\tau_1,\tau_2\in\hat{K}}\langle T_{\tau_1}^{\tau_2} C_+(s)\mathsf{P}_{\tau_1}\boldsymbol{u}, \mathsf{P}_{\tau_2}\boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)}\right) \\
&+ O_s\big(e^{(s-d-\eta_s)t}\|\boldsymbol{u}\|_{\mathcal{S}^m(\upsilon,s)}\|\boldsymbol{v}\|_{\mathcal{S}^m(\upsilon,s)}\big),
\end{aligned}$$

and the sum

$$\sum_{\tau_1,\tau_2\in\hat{K}} \langle T_{\tau_1}^{\tau_2} C_+(s)\mathsf{P}_{\tau_1}\boldsymbol{u}, \mathsf{P}_{\tau_2}\boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)} \tag{4.16}$$

converges absolutely.

Proof. Since smooth vectors are dense in $\mathcal{S}^m(\upsilon,s)$ for all $m\in\mathbb{N}$, and both sides of the inequality are continuous with respect to $\|\cdot\|_{\mathcal{S}^m(\upsilon,s)}$, we start by assuming that $\boldsymbol{u}$ and $\boldsymbol{v}$ are smooth, and decompose them according to the K-types of $\mathcal{U}(\upsilon,s)$:

$$\boldsymbol{u} = \sum_{\tau_1\subset\mathcal{U}(\upsilon,s)} \boldsymbol{u}_{\tau_1}, \qquad \boldsymbol{v} = \sum_{\tau_2\subset\mathcal{U}(\upsilon,s)} \boldsymbol{v}_{\tau_2},$$

where $\boldsymbol{u}_{\tau_1} = \mathsf{P}_{\tau_1}\boldsymbol{u}$ and $\boldsymbol{u}_{\tau_2} = \mathsf{P}_{\tau_2}\boldsymbol{v}$. By [46, Theorem 4.4.2.1],

$$\langle U^s(a_t)\boldsymbol{u},\boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)} = \sum_{\tau_1,\tau_2}\langle U^s(a_t)\boldsymbol{u}_{\tau_1},\boldsymbol{v}_{\tau_2}\rangle_{\mathcal{U}(\upsilon,s)},$$

with the sum converging absolutely. Applying Proposition 4.7 gives

$$\begin{aligned}\langle U^s(a_t)\boldsymbol{u}_{\tau_1},\boldsymbol{v}_{\tau_2}\rangle_{\mathcal{U}(\upsilon,s)} = e^{(s-d)t}&\langle T_{\tau_1}^{\tau_2}C_+(s)\boldsymbol{u}_{\tau_1},\boldsymbol{v}_{\tau_2}\rangle_{\mathcal{U}(\upsilon,s)} \\ &+ O_s\big(e^{(s-d-\eta_s)t}\|\boldsymbol{u}_{\tau_1}\|_{\mathcal{S}^{m'}(\upsilon,s)}\|\boldsymbol{v}_{\tau_2}\|_{\mathcal{S}^{m'}(\upsilon,s)}\big)\end{aligned} \tag{4.17}$$

for some $m'\in\mathbb{N}$. Hence

$$\begin{aligned}\langle U^s(a_t)\boldsymbol{u},\boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)} &= e^{(s-d)t}\left(\sum_{\tau_1,\tau_2}\langle T_{\tau_1}^{\tau_2}C_+(s)\mathsf{P}_{\tau_1}\boldsymbol{u},\mathsf{P}_{\tau_2}\boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)}\right) \\ &\quad + O_s\left(e^{(s-d-\eta_s)t}\left(\sum_{\tau}\|\boldsymbol{u}_\tau\|_{\mathcal{S}^{m'}(\upsilon,s)}\right)\left(\sum_{\tau}\|\boldsymbol{v}_\tau\|_{\mathcal{S}^{m'}(\upsilon,s)}\right)\right).\end{aligned}$$

By Lemma 3.1, there exists $m\ge m'$ (depending only on K) large enough such that

$$\sum_{\tau\in\mathcal{U}(\upsilon,s)}\|\boldsymbol{u}_\tau\|_{\mathcal{S}^{m'}(\upsilon,s)} \ll \|\boldsymbol{u}\|_{\mathcal{S}^m(\upsilon,s)} \quad \text{and} \quad \sum_{\tau\in\mathcal{U}(\upsilon,s)}\|\boldsymbol{v}_\tau\|_{\mathcal{S}^{m'}(\upsilon,s)} \ll \|\boldsymbol{v}\|_{\mathcal{S}^m(\upsilon,s)}$$

where the implied constants depend only on K.

It remains to prove that the sum $\sum_{\tau_1,\tau_2}\langle T_{\tau_1}^{\tau_2}C_+(s)\mathsf{P}_{\tau_1}\boldsymbol{u},\mathsf{P}_{\tau_2}\boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)}$ converges absolutely. Looking at an individual summand, we have

$$\begin{aligned}|\langle T_{\tau_1}^{\tau_2}C_+(s)\boldsymbol{u}_{\tau_1},\boldsymbol{v}_{\tau_2}\rangle_{\mathcal{U}(\upsilon,s)}| &= a(\upsilon,s,\tau_2)|\langle T_{\tau_1}^{\tau_2}C_+(s)\boldsymbol{u}_{\tau_1},\boldsymbol{v}_{\tau_2}\rangle_K| \\ &\le a(\upsilon,s,\tau_2)\|T_{\tau_1}^{\tau_2}\|_K\cdot\|C_+(s)\boldsymbol{u}_{\tau_1}\|_K\cdot\|\boldsymbol{v}_{\tau_2}\|_K\end{aligned}$$

Since $U^s|_K$ is unitary on $L^2(K)$, from the definition of $C_+(s)$ (see Theorem 4.3), we get

$$\|C_+(s)\boldsymbol{u}_{\tau_1}\|_K \ll_s \|\boldsymbol{u}_{\tau_1}\|_K,$$

giving

$$|\langle T_{\tau_1}^{\tau_2} C_+(s)\boldsymbol{u}_{\tau_1}, \boldsymbol{v}_{\tau_2}\rangle_{\mathcal{U}(\upsilon,s)}| \ll_s a(\upsilon,s,\tau_2)\|T_{\tau_1}^{\tau_2}\|_K \cdot \|\boldsymbol{u}_{\tau_1}\|_K \cdot \|\boldsymbol{v}_{\tau_2}\|_K$$
$$\leq \sqrt{\tfrac{a(\upsilon,s,\tau_2)}{a(\upsilon,s,\tau_1)}}\|T_{\tau_1}^{\tau_2}\|_K \cdot \|\boldsymbol{u}_{\tau_1}\|_{\mathcal{U}(\upsilon,s)} \cdot \|\boldsymbol{v}_{\tau_2}\|_{\mathcal{U}(\upsilon,s)}.$$

This expression is now bounded using Proposition 4.5, Corollary 3.8, and Lemma 3.1 as in the proof of Proposition 4.7. Lemma 3.1 then gives the desired convergence of the sum. □

Theorem 4.9. *There exists $m \in \mathbb{N}$ such that for any non-spherical complementary series representation $\mathcal{U}(\upsilon,s)$, for all M-invariant vectors $\boldsymbol{u}$, $\boldsymbol{v} \in \mathcal{S}^m(\upsilon,s)$, and $t \geq 0$, we have*

$$|\langle U^s(a_t)\boldsymbol{u}, \boldsymbol{v}\rangle_{\mathcal{U}(\upsilon,s)}| \ll_s e^{(s-d-\eta_s)t}\|\boldsymbol{u}\|_{\mathcal{S}^m(\upsilon,s)}\|\boldsymbol{v}\|_{\mathcal{S}^m(\upsilon,s)},$$

where the implied constant is uniformly bounded over s in compact subsets of $(\frac{d}{2}, d)$.

Proof. Note that since $\boldsymbol{u}$, $\boldsymbol{v}$ are both M-invariant and $M \subset K$, $\mathsf{P}_\tau \boldsymbol{u}$ and $\mathsf{P}_\tau \boldsymbol{v}$ are as well for any K-type τ of $\mathcal{U}(\upsilon,s)$. Proposition 4.7 then gives that the implied constant in Theorem 4.8 is uniformly bounded over s in compact subsets of $(\frac{d}{2}, d)$. Thus, in order to prove the theorem, it suffices to show that for any K-types τ_1, τ_2 of $\mathcal{U}(\upsilon,s)$ and an arbitrary M-invariant vector $\boldsymbol{w}$ of $\mathcal{U}(\upsilon,s)$,

$$T_{\tau_1}^{\tau_2} C_+(s)\mathsf{P}_{\tau_1}\boldsymbol{w} = 0. \tag{4.18}$$

Since $C_+(s)$ preserves the M-types of each K-type by Lemma 4.2 and $\boldsymbol{w}$ is M-invariant (hence $\mathsf{P}_{\tau_1}\boldsymbol{w}$ is as well), $C_+(s)\mathsf{P}_{\tau_1}\boldsymbol{w}$ is M-invariant. Therefore (4.18) follows from Corollary 3.6. □

5. Leading term for Sobolev functions

In this section, we will extend Roblin's result on mixing of the geodesic flow for continuous functions with compact support to general functions in a Sobolev space of sufficiently high order.

Recall the following theorem of Roblin [33, Theorem 3.4]:

Theorem 5.1. *For all M-invariant $f_1, f_2 \in C_c(\Gamma\backslash G)$,*

$$\lim_{t\to+\infty} e^{(d-\delta)t}\langle \rho(a_t)f_1, f_2\rangle = m^{\mathrm{BR}}(f_1)\, m^{\mathrm{BR}_*}(f_2).$$

We note that $m^{\mathrm{BR}}(f)$ may be infinite for a general function $f \in L^2(\Gamma\backslash G)$, and hence Theorem 5.1 does not generalize to arbitrary L^2-functions.

In order to extend this theorem to Sobolev functions (which are not necessarily compactly supported), we first prove the following bound on matrix coefficients of $L^2(\Gamma\backslash G)$:

Lemma 5.2. *There exists $m \in \mathbb{N}$ such that for all $f_1, f_2 \in \mathcal{S}^m(\Gamma\backslash G)$ and $t \geq 0$,*

$$|\langle \rho(a_t)f_1, f_2\rangle| \ll_\Gamma e^{(\delta-d)t}\mathcal{S}^m(f_1)\mathcal{S}^m(f_2).$$

Proof. We start by assuming that f_1 and f_2 are $\rho(K)$-invariant. Using the notation of Section 2.4, the decomposition of the functions according to $L^2(\Gamma\backslash G)_{\mathrm{sph}} = \mathcal{B}_\delta \oplus \mathcal{W}$ reads

$$f_i = \langle f_i, \phi_0\rangle\phi_0 + f_i', \qquad i = 1, 2,$$

where $f_i' \in \mathcal{W}$ are K-invariant, and $\phi_0 \in \mathcal{B}_\delta$ is the base-eigenfunction in $L^2(\Gamma\backslash\mathbb{H}^{d+1}) = L^2(\Gamma\backslash G)^K$ of unit norm.

Since $\mathcal{B}_\delta$ is a complementary series representation $\mathcal{U}(1,\delta)$ (cf. Sections 2.4 and 3.2), it follows from Theorem 4.8 that for all $t \geq 0$,

$$|\langle\rho(a_t)\phi_0, \phi_0\rangle| \ll e^{(\delta-d)t}.$$

As a consequence of Theorem 2.1, f_1' and f_2' are orthogonal to any complementary series representation $\mathcal{U}(\upsilon, s)$ with $s_1 < s \leq \delta$. It follows that for any $\varepsilon > 0$, for any K-invariant $f_1, f_2 \in L^2(\Gamma\backslash G)$ and $t > 0$,

$$|\langle\rho(a_t)f_1', f_2'\rangle| \ll_\varepsilon e^{(s_1-d+\varepsilon)t}\|f_1'\|\|f_2'\|$$

(see [37, Theorem 2.1, 2], [23, Proposition 5.3]).

Therefore, choosing $0 < \varepsilon < \delta - s_1$ gives

$$|\langle\rho(a_t)f_1, f_2\rangle| \ll e^{(\delta-d)t}\big(|\langle f_1, \phi_0\rangle||\langle f_2, \phi_0\rangle| + \|f_1'\|\|f_2'\|\big) \leq e^{(\delta-d)t}\|f_1\|\|f_2\|.$$

In view of [37, Proposition 2.5], this bound extends for all K-*finite* functions f_1, f_2 in $L^2(\Gamma\backslash G)$:

$$|\langle\rho(a_t)f_1, f_2\rangle| \ll e^{(\delta-d)t}\|f_1\|\|f_2\|\sqrt{\dim(\rho(K)f_1)\,\dim(\rho(K)f_2)}.$$

To pass to Sobolev functions, we first observe that

$$\dim(\rho(K)\mathsf{P}_\tau f) \leq \dim(\tau)^2$$

for all $f \in L^2(\Gamma\backslash G)$ (cf. [20, p. 206 (1)]). Then for all $f_1, f_2 \in \mathcal{S}^m(\Gamma\backslash G)$,

$$\begin{aligned} |\langle\rho(a_t)f_1, f_2\rangle| &\leq e^{(\delta-d)t}\sum_{\tau_1,\tau_2\in\hat{K}}\dim(\tau_1)\dim(\tau_2)\|\mathsf{P}_{\tau_1}f_1\|\|\mathsf{P}_{\tau_2}f_2\| \\ &\ll e^{(\delta-d)t}\|f_1\|_{\mathcal{S}^m(\Gamma\backslash G)}\|f_2\|_{\mathcal{S}^m(\Gamma\backslash G)} \end{aligned}$$

for $m \in \mathbb{N}$ large enough, by Lemma 3.1. ☐

Proposition 5.3. *There exists $m \in \mathbb{N}$ such that for all M-invariant $f_1, f_2 \in \mathcal{S}^m(\Gamma\backslash G)$,*

$$\lim_{t\to+\infty} e^{(d-\delta)t}\langle\rho(a_t)f_1, f_2\rangle = m^{\mathrm{BR}}(f_1)\,m^{\mathrm{BR}_*}(f_2).$$

Proof. Let $m > d(d+1)/4$ satisfy the conclusion of Lemma 5.2. For simplicity, we write $\|f\|_{\mathcal{S}^m} = \|f\|_{\mathcal{S}^m(\Gamma\backslash G)}$. Using the density of $C_c^\infty(\Gamma\backslash G)^M$ in $\mathcal{S}^m(\Gamma\backslash G)^M$, given $\varepsilon > 0$, there exist $f_1^\varepsilon, f_2^\varepsilon \in C_c^\infty(\Gamma\backslash G)$ such that

$$\|f_i - f_i^\varepsilon\|_{\mathcal{S}^m} \leq \varepsilon \qquad i = 1, 2.$$

We then write, using Lemma 5.2,

$$\begin{aligned}
&e^{(d-\delta)t}\langle\rho(a_t)f_1,f_2\rangle\\
&= e^{(d-\delta)t}\Big(\langle\rho(a_t)f_1^\varepsilon,f_2^\varepsilon\rangle+\langle\rho(a_t)(f_1-f_1^\varepsilon),f_2\rangle+\langle\rho(a_t)f_1^\varepsilon,f_2-f_2^\varepsilon\rangle\Big)\\
&= e^{(d-\delta)t}\langle\rho(a_t)f_1^\varepsilon,f_2^\varepsilon\rangle+O\big(\varepsilon(\|f_1\|_{\mathcal{S}^m}+\|f_2\|_{\mathcal{S}^m})\big).
\end{aligned}$$

By Theorem 5.1, we have

$$\lim_{t\to\infty} e^{(d-\delta)t}\langle\rho(a_t)f_1^\varepsilon,f_2^\varepsilon\rangle = m^{\mathrm{BR}}(f_1^\varepsilon)\, m^{\mathrm{BR}_*}(f_2^\varepsilon).$$

So

$$\limsup_{t\to\infty} e^{(d-\delta)t}\langle\rho(a_t)f_1,f_2\rangle = m^{\mathrm{BR}}(f_1^\varepsilon)\, m^{\mathrm{BR}_*}(f_2^\varepsilon)+O\big(\varepsilon(\|f_1\|_{\mathcal{S}^m}+\|f_2\|_{\mathcal{S}^m})\big).$$

Since $m > d(d+1)/4$, we may apply Lemma 2.2 to get

$$m^{\mathrm{BR}}(f_1^\varepsilon)\, m^{\mathrm{BR}_*}(f_2^\varepsilon) = m^{\mathrm{BR}}(f_1)\, m^{\mathrm{BR}_*}(f_2)+O\big(\varepsilon(\|f_1\|_{\mathcal{S}^m}+\|f_2\|_{\mathcal{S}^m})\big).$$

Therefore

$$\limsup_{t\to\infty} e^{(d-\delta)t}\langle\rho(a_t)f_1,f_2\rangle = m^{\mathrm{BR}}(f_1)\, m^{\mathrm{BR}_*}(f_2)+O\big(\varepsilon(\|f_1\|_{\mathcal{S}^m}+\|f_2\|_{\mathcal{S}^m})\big).$$

Since $\varepsilon > 0$ was arbitrary, we in fact have

$$\limsup_{t\to\infty} e^{(d-\delta)t}\langle\rho(a_t)f_1,f_2\rangle = m^{\mathrm{BR}}(f_1)\, m^{\mathrm{BR}_*}(f_2).$$

An analogous calculation with lim inf in place of lim sup proves the proposition. □

6. Proof of Theorem 1.1

In this final section, we prove Theorem 1.1. Recall the isomorphism

$$\big(\rho, L^2(\Gamma\backslash G)\big) \cong \int_{\mathsf{Z}}^{\oplus} (\pi_\zeta,\mathcal{H}_\zeta)\, d\mu_{\mathsf{Z}}(\zeta).$$

Fix arbitrary $0 < r < \delta - s_1$. We partition Z as

$$\mathsf{Z} = \mathsf{Z}_r^- \cup \mathsf{Z}_r^+,$$

where

$$\mathsf{Z}_r^- = \Big\{\zeta\in\mathsf{Z} : (\pi_\zeta,\mathcal{H}_\zeta)\cong\mathcal{U}(\upsilon,s),\ \text{where } \upsilon\in\hat{M} \text{ and } s\in[s_1+r,\delta]\Big\}$$

and $\mathsf{Z}_r^+ = \mathsf{Z}\setminus\mathsf{Z}_r^-$ (cf. (2.2)). This partition is then used to decompose $\big(\rho, L^2(\Gamma\backslash G)\big)$ as

$$\big(\rho, L^2(\Gamma\backslash G)\big) = \big(\rho, L^2(\Gamma\backslash G)^-\big)\oplus\big(\rho, L^2(\Gamma\backslash G)^+\big),$$

where

$$\big(\rho, L^2(\Gamma\backslash G)^\pm\big) \cong \int_{\mathsf{Z}_r^\pm}^{\oplus} (\pi_\zeta,\mathcal{H}_\zeta)\, d\mu_{\mathsf{Z}}(\zeta)$$

with the dependency on r being slightly suppressed. Recall that $\mathcal{B}_\delta$ occurs as a subrepresentation of $\big(\rho, L^2(\Gamma\backslash G)^-\big)$ and $\mathcal{B}_\delta\cong\mathcal{U}(1,\delta)$ (cf. Sections 2.4 and

3.2). We may further decompose $\big(\rho, L^2(\Gamma\backslash G)^-\big)$ accordingly: let $L^2(\Gamma\backslash G)^-_0$ be the orthogonal complement of $\mathcal{B}_\delta$ in $L^2(\Gamma\backslash G)^-$. We thus have

$$\big(\rho, L^2(\Gamma\backslash G)\big) = (\rho, \mathcal{B}_\delta) \oplus \big(\rho, L^2(\Gamma\backslash G)^-_0\big) \oplus \big(\rho, L^2(\Gamma\backslash G)^+\big).$$

Note that Theorem 2.1 and the duality between eigenfunctions of the Laplacian on $\mathcal{M} = \Gamma\backslash G/K$ and spherical representations in the decomposition of $L^2(\Gamma\backslash G)$ imply that no spherical representation is weakly contained in $\big(\rho, L^2(\Gamma\backslash G)^-_0\big)$. At the level of functions, we write

$$f_i = f_i^0 + f_i^- + f_i^+, \qquad i = 1, 2$$

where $f_i^0 \in \mathcal{B}_\delta$, $f_i^- \in L^2(\Gamma\backslash G)^-_0$, and $f_i^+ \in L^2(\Gamma\backslash G)^+$.

The matrix coefficients we are interested in now decompose as

$$\langle \rho(a_t) f_1, f_2\rangle = \langle \rho(a_t) f_1^0, f_2^0\rangle + \langle \rho(a_t) f_1^-, f_2^-\rangle + \langle \rho(a_t) f_1^+, f_2^+\rangle. \tag{6.1}$$

We deal with the three summands in turn: by construction, $\big(\rho, L^2(\Gamma\backslash G)^+\big)$ does not weakly contain any $\mathcal{U}(\upsilon, s)$ with $s > s_1 + r$ (this also uses the fact $\big(\rho, L^2(\Gamma\backslash G)\big)$ does not weakly contain any $\mathcal{U}(\upsilon, s)$ with $s > \delta$; cf. [27, Proposition 3.23]).

We assume that m is large enough so that all the results of the previous sections hold for $f_1, f_2 \in \mathcal{S}^m(\Gamma\backslash G)$. By [23, Proposition 5.3] (cf. also [27, Proposition 3.29] or [37, Theorem 2.1] combined with the argument from the proof of Lemma 5.2), we have that for all $\xi > 0$,

$$\begin{aligned} |\langle \rho(a_t) f_1^+, f_2^+\rangle| &\ll_\xi e^{(s_1 - d + r + \xi)t} \|f_1^+\|_{\mathcal{S}^m(\Gamma\backslash G)} \|f_2^+\|_{\mathcal{S}^m(\Gamma\backslash G)} \\ &\le e^{(s_1 - d + r + \xi)t} \|f_1\|_{\mathcal{S}^m(\Gamma\backslash G)} \|f_2\|_{\mathcal{S}^m(\Gamma\backslash G)}. \end{aligned} \tag{6.2}$$

We will now use Theorem 4.9 to bound $\langle \rho(a_t) f_1^-, f_2^-\rangle$. Let $\widetilde{\mathsf{Z}}_r^- \subset \mathsf{Z}_r^-$ be such that $\big(\rho, L^2(\Gamma\backslash G)^-_0\big) \cong \int^{\oplus}_{\widetilde{\mathsf{Z}}_r^-} (\pi_\zeta, \mathcal{H}_\zeta)\, d\mu_{\mathsf{Z}}(\zeta)$. The corresponding decomposition of the functions f_1^-, f_2^- reads $f_i^- = \int_{\widetilde{\mathsf{Z}}_\varepsilon^-} (f_i^-)_\zeta \, d\mu_{\mathsf{Z}}(\zeta)$ $(i = 1, 2)$. Since f_i is M-invariant, so is f_i^- and μ_{Z}-a. e. $(f_i^-)_\zeta$. The matrix coefficient $\langle \rho(a_t) f_1^-, f_2^-\rangle$ is now written as

$$\langle \rho(a_t) f_1^-, f_2^-\rangle = \int_{\widetilde{\mathsf{Z}}_\varepsilon^-} \langle \pi_\zeta(a_t)(f_1^-)_\zeta, (f_2^-)_\zeta\rangle_{\mathcal{H}_\zeta} \, d\mu_{\mathsf{Z}}(\zeta).$$

Each $(\pi_\zeta, \mathcal{H}_\zeta)$ is isomorphic to some $\mathcal{U}(\upsilon, s)$ with υ *non-trivial* and $s \in [s_1 + r, \delta]$. Setting

$$\lambda(\delta, r) = \max_{s \in [s_1 + r, \delta]} s - d - \eta_s$$

where $\eta_s = \min(2s-d, 1)$, we apply Theorem 4.9 to each $\langle \pi_\zeta(a_t)(f_1^-)_\zeta, (f_2^-)_\zeta \rangle_{\mathcal{H}_\zeta}$ and obtain

$$\begin{aligned}
&|\langle \rho(a_t)f_1^-, f_2^- \rangle| \ll_{r,\delta} e^{\lambda(\delta,r)t} \int_{\widetilde{\mathsf{Z}}_r^-} \|(f_1^-)_\zeta\|_{\mathcal{S}^m(\mathcal{H}_\zeta)} \|(f_2^-)_\zeta\|_{\mathcal{S}^m(\mathcal{H}_\zeta)}\, d\mu_{\mathsf{Z}}(\zeta) \\
\le & e^{\lambda(\delta,r)t} \sqrt{\int_{\widetilde{\mathsf{Z}}_r^-} \|(f_1^-)_\zeta\|^2_{\mathcal{S}^m(\mathcal{H}_\zeta)}\, d\mu_{\mathsf{Z}}(\zeta)} \sqrt{\int_{\widetilde{\mathsf{Z}}_r^-} \|(f_2^-)_\zeta\|^2_{\mathcal{S}^m(\mathcal{H}_\zeta)}\, d\mu_{\mathsf{Z}}(\zeta)} \\
= & e^{\lambda(\delta,r)t} \|f_1^-\|_{\mathcal{S}^m(\Gamma\backslash G)} \|f_2^-\|_{\mathcal{S}^m(\Gamma\backslash G)} \\
\le & e^{\lambda(\delta,r)t} \|f_1\|_{\mathcal{S}^m(\Gamma\backslash G)} \|f_2\|_{\mathcal{S}^m(\Gamma\backslash G)},
\end{aligned} \tag{6.3}$$

where the implied constant remains uniformly bounded as $r \to 0$. The remaining term in the left-hand side is $\langle \rho(a_t)f_1^0, f_2^0 \rangle$. Using the fact that $(\rho, \mathcal{B}_\delta)$ is isomorphic to $\mathcal{U}(1,\delta)$, applying Proposition 4.8 gives

$$\begin{aligned}
\langle \rho(a_t)f_1^0, f_2^0 \rangle &= \Phi(f_1^0, f_2^0)\, e^{(\delta-d)t} + O_\delta\left(e^{(\delta-d-\eta_\delta)t} \|f_1^0\|_{\mathcal{S}^m(\Gamma\backslash G)} \|f_2^0\|_{\mathcal{S}^m(\Gamma\backslash G)}\right) \\
&= \Phi(f_1^0, f_2^0)\, e^{(\delta-d)t} + O_\delta\left(e^{(\delta-d-\eta_\delta)t} \|f_1\|_{\mathcal{S}^m(\Gamma\backslash G)} \|f_2\|_{\mathcal{S}^m(\Gamma\backslash G)}\right),
\end{aligned} \tag{6.4}$$

where $\Phi(f_1^0, f_2^0)$ is given by the sum (4.16) under the aforementioned isomorphism.

Note that

$$\begin{aligned}
\beta &:= \min\left\{\eta_\delta, \delta - s_1 - r - \xi, \delta - d - \lambda(\delta, r)\right\} \\
&= \min\left\{1, 2\delta - d, \delta - s_1 - r - \xi, \min_{s\in[s_1+r,\delta]} \left(\delta - s + \min\{1, 2s-d\}\right)\right\} \\
&\ge \min\left\{1, 2\delta - d, \delta - s_1 - r, \min_{s\in[s_1+r,\delta]} \left(\delta - s + \min\{1, 2s-d\}\right)\right\} - \xi \\
&= \min\{1, 2\delta - d, \delta - s_1 - r, \delta + s_1 + r - d\} - \xi \\
&= \min\{1, \delta - s_1 - r, \delta + s_1 + r - d\} - \xi \\
&= \min\{1, \delta - s_1 - r\} - \xi \\
&\ge \min\{1, \delta - s_1\} - (\xi + r).
\end{aligned}$$

Entering (6.2), (6.3), and (6.4) into (6.1) gives

$$e^{(d-\delta)t}\langle \rho(a_t)f_1, f_2 \rangle = \Phi(f_1^0, f_2^0) + O_{r,\xi}\big(e^{-\beta t} \|f_1\|_{\mathcal{S}^m(\Gamma\backslash G)} \|f_2\|_{\mathcal{S}^m(\Gamma\backslash G)}\big).$$

Writing $\eta = \min(1, \delta - s_1)$ and $\varepsilon = \xi + r$, we thus have

$$e^{(d-\delta)t}\langle \rho(a_t)f_1, f_2 \rangle = \Phi(f_1^0, f_2^0) + O_\varepsilon\big(e^{-(\eta-\varepsilon)t} \|f_1\|_{\mathcal{S}^m(\Gamma\backslash G)} \|f_2\|_{\mathcal{S}^m(\Gamma\backslash G)}\big).$$

Choosing $0 < \varepsilon < \eta$ gives

$$\lim_{t\to\infty} e^{(d-\delta)t}\langle \rho(a_t)f_1, f_2 \rangle = \Phi(f_1^0, f_2^0),$$

so $\Phi(f_1^0, f_2^0) = m^{\mathrm{BR}}(f_1)\, m^{\mathrm{BR}_*}(f_2)$ by Proposition 5.3, completing the proof of Theorem 1.1.

Remark 6.5. In the case when Γ is a lattice, i.e., when $\delta = d$, the above proof leads to the claim in Remark 1.1 (1) as follows. Firstly, we have $\langle \rho(a_t) f_1^0, f_2^0 \rangle = \int f_1 \, dx \cdot \int f_2 \, dx$ for any $t \in \mathbb{R}$. As before, we use (6.2) to bound $\langle \rho(a_t) f_1^+, f_2^+ \rangle$. Now, since complementary series representations $\mathcal{U}(\upsilon, s)$ with *non-trivial* υ exist only for $s \le d-1$, the constant $\lambda(d,r)$ appearing in the bound (6.3) for $|\langle \rho(a_t) f_1^-, f_2^- \rangle|$ is at most $\alpha := \max\{s - d - \eta_s : s \in [s_1 + r, d-1]\}$. We note that $\alpha < -2 + r$, and if, in addition, $\{s > s_1 + 1 : \mathcal{U}(\upsilon, s) \text{ is weakly contained in } L^2(\Gamma \backslash G)\} = \emptyset$, then one can take $\lambda(d,r) < s_1 - d + r$. Hence combining these bounds, we obtain Remark 1.1 (1).

References

[1] T. Aubin. *Nonlinear Analysis on Manifolds*, Grundlehren Math. Wiss. 252, Springer (1982).

[2] M. Babillot. *On the mixing property for hyperbolic systems*, Israel J. Math. 129 (2002), 61-76.

[3] Y. Benoist and H. Oh. *Effective equidistribution of S-integral points on symmetric varieties*, Ann. Inst. Fourier 62 (2012), 1889-1942.

[4] T. Bröcker and T. tom Dieck. *Representations of compact Lie groups*, GTM 98, Springer (1985).

[5] J. Bourgain, A. Gamburd, and P. Sarnak. *Generalization of Selberg's* 3/16 *theorem and Affine sieve*, Acta Math. 20 (2011), 255-290.

[6] J. Bourgain, A. Gamburd, and P. Sarnak. *Affine linear sieve, expanders, and sum-product*, Inventiones 179 (2010), 559-644.

[7] J. Bourgain, A. Kontorovich, and P. Sarnak. *Sector estimates for hyperbolic isometries*, Geom. Funct. Anal. 20 (2010), 1175-1200.

[8] N. Bergeron and L. Clozel. *Quelques conséquences des travaux dArthur pour le spectre et la topologie des variétés hyperboliques.* Invent. Math. 192 (2013), 505-532.

[9] M. Burger and P. Sarnak. *Ramanujan duals II.* Invent. Math. 106 (1991), 1-11.

[10] M. Eguchi, S. Koizumi, and M. Mamiuda. *The expressions of the Harish-Chandra C-functions of semisimple Lie groups Spin(n,1), SU(n,1)*, J. Math. Soc. Japan 51 (1999), 955-985.

[11] W. Duke, Z. Rudnick, and P. Sarnak. *Density of integer points on affine homogeneous varieties*, Duke Math. J. 71 (1993), 143-179.

[12] A. Eskin and C. T. McMullen. *Mixing, counting, and equidistribution in Lie groups*, Duke Math. J. 71 (1993), 181-209.

[13] A. Gamburd. *On the spectral gap for infinite index "congruence" subgroups of* $\mathrm{SL}_2(\mathbb{Z})$, Isr. J. Math. 123 (2002), 157-200.

[14] R. Goodman and N. Wallach. *Symmetry, representations, and invariants*, GTM 255, Springer (2009).

[15] W. He and N. de Saxcé. *Linear random walks on the torus*, preprint: arXiv:1910.1342.

[16] T. Hirai. *On irreducible representations of the Lorentz group of n-th order*, Proc. Japan Acad. 38 (1962), 258-262.

[17] J. Kečkić and P. Vasić. *Some inequalities for the gamma function*, Pub. Inst. Math. 11 (1971), 107-114.

[18] D. Kelmer and H. Oh. *Exponential mixing and shrinking targets for geodesic flow on geometrically finite hyperbolic manifolds*, arXiv:1812.05251, to appear in JMD.

[19] I. Kim. *Counting, Mixing and Equidistribution of horospheres in geometrically finite rank one locally symmetric manifolds*, J. Reine Angew. Math. 704 (2015), 85-133.

[20] A. Knapp. *Representation Theory of semisimple Lie Groups*, Princeton University Press (1986).

[21] A. Knapp. *Lie Groups Beyond an Introduction* Second Edition, Birkhäuser, 2002.

[22] A. Knapp and E. Stein *Intertwining Operators for Semisimple Groups*, Ann. of Math., Second Series, Vol. 93 (1971), 489-578.

[23] A. Kontorovich and H. Oh. *Apollonian circle packings and closed horospheres on hyperbolic 3-manifolds*, J. Amer. Math. Soc. 24 (2011), 603-648.

[24] P. Lax and R. Phillips. *The asymptotic distribution of lattice points in Euclidean and non-Euclidean spaces*, J. Funct. Anal. 46 (1982), 280-350.

[25] M. Magge. *Quantitative spectral gap for thin groups of hyperbolic isometries*, JEMS, 17 (2015), 151-187

[26] G. Margulis, A. Mohammadi and H. Oh. *Closed geodesics and holonomies for Kleinian manifolds*, Geom. Funct. Anal. 24 (2014), 1608-1636.

[27] A. Mohammadi and H. Oh. *Matrix coefficients, counting and primes for orbits of geometrically finite groups*, J. Eur. Math. Soc. (JEMS) 17 (2015), 837-897.

[28] C. Moore. *Exponential decay of correlation coefficients for geodesic flows*, in: Group Representations, Ergodic Theory, Operator Algebras, and Mathematical Physics (eds C. C. Moore), MSRI Pub. Vol 6, Springer (1987).

[29] H. Oh and D. Winter. *Uniform exponential mixing and resonance free regions for convex co-compact congruence subgroups of* $\mathrm{SL}_2(\mathbb{Z})$, J. Amer. Math. Soc. 29 (2016), 1069-1115.

[30] S. J. Patterson. *The limit set of a Fuchsian group*, Acta Math. 136 (1976), 241-273.

[31] J.-P. Otal and M. Peigné. *Principe variationnel et groupes Kleiniens*, Duke Math. J. 125 (2004), 15-44.

[32] M. Ratner. *The rate of mixing for geodesic and horocycle flows*, Ergod. Th. & Dynam. Sys. 7 (1987), 267-288.

[33] T. Roblin. *Ergodicité et équidistribution en courbure négative*, Mém. Soc. Math. France (N.S.) 95 (2003), vi+96 pp.

[34] P. Sarkar. *Generalization of Selberg's* $3/16$ *theorem for convex cocompact thin subgroups of* $\mathrm{SO}(n,1)$, preprint, arXiv:2006.07787.

[35] P. Sarkar and D. Winter. *Exponential mixing of frame flows for convex cocompact hyperbolic manifolds*, preprint: arXiv:2004.14551.

[36] P. Sarnak. *Notes on the Generalized Ramanujan Conjectures.* Clay Math. Proceedings 4 (2005).

[37] Y. Shalom. *Rigidity, unitary representations of semisimple groups, and fundamental groups of manifolds with rank one transformation group*, Ann. of Math. 152 (2000), 113-182.

[38] L. Stoyanov. *Spectra of Ruelle transfer operators for axiom A flows*, Nonlinearity 24 (2011), 1089-1120.

[39] D. Sullivan. *Entropy, Hausdorff measures old and new, and limit sets of geometrically finite Kleinian groups*, Acta Math. 153 (1984), 259-277.

[40] D. Sullivan. *The density at infinity of a discrete group of hyperbolic motions*, Inst. Hautes Études Sci. Publ. Math. 50 (1979), 171-202.

[41] A. Salehi-Golsefidy and P. Varjú. *Expansion in perfect groups*, Geom. Funct. Anal. 22 (2012), 1832-1891.

[42] E. Thieleker. *On the quasi-simple irreducible representations of the Lorentz groups*, Trans. Amer. Math. Soc. 179 (1973), 465505.

[43] I. Vinogradov. *Effective bisector estimate with applications to Apollonian circle packings*, Int. Math. Res. Not. IMRN 12 (2014), 3217-3262.

[44] N. Wallach. *Real Reductive Groups I*, Academic Press (1988).

[45] N. Wallach. *Real Reductive Groups II*, Academic Press (1992).

[46] G. Warner. *Harmonic Analysis on Semi-Simple Lie Groups I*, Grundlehren Math. Wiss. 188, Springer (1972).

[47] G. Warner. *Harmonic Analysis on Semi-Simple Lie Groups II*, Grundlehren Math. Wiss. 189, Springer (1972).

[48] D. Želobenko. *Compact Lie groups and their representations*, Translations of Mathematical Monographs, Vol. 40, American Mathematical Society, Providence, R.I. (1973).

Department of Mathematics, Yale University, New Haven, CT 06520
E-mail address: samuel.edwards@yale.edu

Department of Mathematics, Yale University, New Haven, CT 06520 and Korea Institute for Advanced Study, Seoul, Korea
E-mail address: hee.oh@yale.edu