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 \( L^2(T^1(\mathcal{M})) \), 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-Phillips. Combined with the work of Bourgain, Gamburd and Sarnak and its generalization by Golsefidy-Varju on expanders, this implies uniform exponential mixing for congruence covers of \( \mathcal{M} \) when \( \Gamma \) is a thin subgroup of \( \text{SO}^\circ(d+1,1) \). Our result implies that, with respect to the Bowen-Margulis-Sullivan measure, the geodesic flow on \( T^1(\mathcal{M}) \) is exponentially mixing, uniformly over congruence covers in the case when \( \Gamma \) is a thin subgroup.

Contents

1. Introduction 1
2. Preliminaries 5
3. Complementary series representations and Eisenstein integrals 9
4. Asymptotic expansions of matrix coefficients 16
5. Leading term for Sobolev functions 23
6. Proof of Theorem 1.1 25
References 27

1. Introduction

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

Denoting by \( \Delta \) the negative of the Laplace operator on \( \mathcal{M} \), the bottom eigenvalue of \( \Delta \) on \( L^2(\mathcal{M}) \) is known to be simple and given by \( \lambda_0 = \delta(d-\delta) \) by Patterson [25] and Sullivan [33]. Lax and Phillips [19] showed that there

Date: December 16, 2019.

S. E. was supported by postdoctoral scholarship 2017.0391 from the Knut and Alice Wallenberg Foundation and H. O. was supported by NSF grants.
are only finitely many eigenvalues of $\Delta$ on $L^2(M)$ in the interval $[\lambda_0, \frac{d^2}{4}]$.

We denote by $\lambda_1$ the second smallest eigenvalue of $\Delta$:

$$0 < \lambda_0 := \delta(d - \delta) < \lambda_1 := s_1(d - s_1) \leq \frac{d^2}{4}.$$ 

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

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

The main result of this paper is that the Lax-Phillips spectral gap controls the exponential convergence rate of the correlation function for the geodesic flow $\mathcal{G}^t$ acting on $L^2(T^1(M), dx)$:

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

$$e^{(d - \delta)t} \int_{T^1(M)} \psi_1(\mathcal{G}^t(x)) \psi_2(x) \, dx = \frac{1}{m_{\text{BMS}}(T^1(M))} m^\text{BR}(\psi_1)m^\text{BR}*(\psi_2) + O(\varepsilon(e^{(\eta+\varepsilon)t} S^m(\psi_1)S^m(\psi_2)),$$

where

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

We note that $m^\text{BR}(\psi_i) < \infty$ when $S^m(\psi_i) < \infty$ (see Lemma 2.2), and hence the main term above is well-defined. For $\psi_i$ compactly supported, the asymptotic formula (without error term) holds for any $\delta > 0$, as was obtained by Roblin [28].

**Remark 1.1.**

1. When $\mathcal{M}$ has finite volume, this is a classical result due to Moore [22] (see also [27] for the case of surfaces). Moore’s proof makes explicit use of the fact that there are no non-spherical complementary series representations $\mathcal{U}(v, s)$ 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 $\eta = \min\{2, d - s_1\}$.

2. Theorem 1.1 can be deduced from [7] and [35] for $d = 1$ and 2, respectively. For a general $d \geq 1$, the case of $\delta > \max\{d - 1, d/2\}$ was obtained in [21] for some $\eta > 0$ which is not explicit.
The BMS measure $m^{\text{BMS}}$ is known to be the unique measure of maximal entropy (which is $\delta$) for the geodesic flow ([33], [26]). Babillot showed that the geodesic flow is strong mixing with respect to $m^{\text{BMS}}$ [2]. Theorem 1.1 is known to imply the following exponential mixing for the BMS measure (see [21, Theorem 1.6] for compactly supported functions, and [14, Theorem 1.9] for general bounded functions):

**Theorem 1.2.** There exists $\beta > 0$ (explicitly computable, depending only on $\eta$ in Theorem 1.1 and $m > d(d+1)/2$ such that for all bounded functions $\psi_1, \psi_2$ supported on a one-neighbourhood of $\text{supp}(m^{\text{BMS}})$, we have, as $t \to \infty$,

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

**Remark 1.2.** The main novelty of Theorem 1.2 is that it addresses geometrically finite groups with cusps and the dependence of $\beta$ on the spectral gap $\delta - s_1$ can be made explicit. Indeed, when $\Gamma$ is convex cocompact, Theorem 1.2 for some $\beta > 0$ which is not explicit, follows from the work of Stoyanov [31], which is based on symbolic dynamics and Dolgopyat operators (see also [39]). This result in turn implies Theorem 1.1 with implicit $\eta > 0$ (see [24], [39]).

Following the terminology coined by Sarnak, we say $\Gamma < G$ is a thin subgroup if $\Gamma$ is a Zariski dense subgroup of an arithmetic subgroup $G(\mathbb{Z})$ of $G$ such that $\hat{\Gamma} := \pi^{-1}(\Gamma)$ satisfies the strong approximation property for the spin covering map $\pi : \text{Spin}(d + 1, 1) \to G$. We denote by $\Gamma_q$ the congruence subgroup of $\Gamma$: $\Gamma_q = \{\gamma \in \Gamma : \gamma = e \mod q\}$. The work of Bourgain-Gamburd-Sarnak [6] and its generalization by Golsefidy and Varju [34] 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

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

where $s_1(q)(d-s_1(q))$ is the second smallest eigenvalue of $\Delta$ on $L^2(\Gamma(q)\backslash \mathbb{H}^{d+1})$. This uses the transfer property from the combinatorial spectral gap to the archimedean spectral gap due to [5] (see also [15]).

We therefore have:

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

When $\Gamma$ is convex cocompact, Corollary 1.3 was obtained by [24] for $d = 1$ and by Sarkar [29] for a general $d \geq 1$, combining Dolgopyat’s methods with the expander theory.

Theorems 1.1 and 1.2 and Corollary 1.3 are known to have many immediate applications in number theory and geometry. To name a few, see [9],
complementary series representations in $L^2$-parameters involved, allows us to control the contribution of non-spherical $M$-geodesic theorem, and [14] for shrinking target problems.

Fix $o \in \mathbb{H}^{d+1}$ and $v_o \in T_o(\mathbb{H}^{d+1})$. Setting $K = \text{Stab}_G(o) \simeq \text{SO}(d+1)$ and $M = \text{Stab}_G(v_o) \simeq \text{SO}(d)$, we can identify $\mathcal{M} = \Gamma \backslash G / K$ and $T^1(\mathcal{M}) = \Gamma \backslash G / M$. 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 $G^t$ on $T^1(\mathcal{M})$. As $L^2(T^1(\mathcal{M}))$ 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 $\psi_i \in L^2(\Gamma \backslash G)$.

The non-tempered part of the unitary dual $\hat{G}$ consists of the trivial representation and the complementary series representations $\mathcal{U}(v, s)$, parametrized by $v \in \hat{M}$ and $s \in \mathcal{I}_v$ where $\mathcal{I}_v \subset (d/2, d)$ is an interval depending on $v$. In this parametrization $\mathcal{U}(v, s)$ is spherical if and only if $v \in \hat{M}$ is trivial, and for $v$ non-trivial, $\mathcal{I}_v$ is contained in $(d/2, d - 1)$ [12]; this was the main reason for the hypothesis $\delta > d - 1$ in [21].

Our main work lies in the detailed analysis of the behavior of matrix coefficients of the complementary series representations $\mathcal{U}(v, s)$. We remark that Harish-Chandra’s work (37, 38) does not give an asymptotic expansion for the matrix coefficients of $\mathcal{U}(v, s)$ for all $s$, as it excludes finitely many (unknown) parameters $s$. Even for those $\mathcal{U}(v, s)$ 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.

We write $\mathcal{U}(v, s) = \bigoplus_{\tau \in \hat{K}} \mathcal{U}(v, s)_{\tau}$ for the decomposition into different $K$-types. Denoting by $\langle \cdot, \cdot \rangle_{\mathcal{U}(v, s)}$ the unitary inner product, we show the following concrete asymptotic expansion with an essentially optimal rate:

**Theorem 1.4.** There exists $m \in \mathbb{N}$ such that for any complementary series representation $\mathcal{U}(v, s)$ with a non-trivial $M$-invariant vector, any $K$-types $\tau_1, \tau_2 \in \hat{K}$, and $u \in \mathcal{U}(v, s)_{\tau_1}, v \in \mathcal{U}(v, s)_{\tau_2}$, we have, for any $t > 0$,

$$
\langle U^s(a_t) u, v \rangle_{\mathcal{U}(v, s)} = e^{(s-d)t} \langle T^{\tau_2}_{\tau_1} C_+(s) u, v \rangle_{\mathcal{U}(v, s)} + O_s(e^{(s-d-\eta_s)t} \|u\|_{\mathcal{S}^m(v, s)} \|v\|_{\mathcal{S}^m(v, s)}) ,
$$

where $\eta_s := \min(1, 2s - d)$, $T^{\tau_2}_{\tau_1} : \mathcal{U}(v, s)_{\tau_1} \rightarrow \mathcal{U}(v, s)_{\tau_2}$ is given by (3.4), $C_+(s)$ is the Harish-Chandra c-function given in (4.1), and the implied constant is uniformly bounded over $s$ in compact subsets of $(\frac{d}{2}, d)$.

Our key observation is that if $v \in \hat{M}$ is non-trivial and $u \in \mathcal{U}(v, s)$ is $M$-invariant, then the main term (1.3) vanishes:

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

This observation, combined with a more explicit control over the various parameters involved, allows us to control the contribution of non-spherical complementary series representations in $L^2(\Gamma \backslash G)^M$: 

[10], [3], [5], [21]) for effective counting and affine sieve, [20] for the prime geodesic theorem, and [14] for shrinking target problems.
Theorem 1.5. There exists \( m \in \mathbb{N} \) such that if \( U(v, s) \) is non-spherical, then for all \( M \)-invariant vectors \( u, v \in S^m(v, s) \),

\[
|\langle U^s(a_t)u, v \rangle_{U(v, s)}| \ll_s e^{(s-d-\eta)s} \|u\|_{S^m(v, s)} \|v\|_{S^m(v, s)},
\]

where 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 [28, 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 \( L^2(\Gamma \setminus G) \) into irreducible representations and applying the bounds obtained in Section 4 to each individual component.

2. Preliminaries

Let \( d \geq 1 \) and \( G = \text{SO}^0(d+1, 1) \) 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 \( \mathcal{M} = \Gamma \setminus \mathbb{H}^{d+1} \) be the associated hyperbolic manifold.

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

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

\[
N = \{g \in G : a_{-t}ga_t \to e \text{ as } t \to +\infty\};
\]

\[
\overline{N} = \{g \in G : a_tga_{-t} \to e \text{ as } t \to +\infty\}.
\]

Then \( \overline{N} = \omega N \omega^{-1} \), where \( \omega \in N_K(A) \) is such that \( \omega \omega^{-1} = a^{-1} \) 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 \( N \simeq \mathbb{R}^d \), the conjugation by an element \( m \in M = \text{SO}(d) \) 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 \( \mathfrak{g}, \mathfrak{k}, \mathfrak{a}, \mathfrak{n} \) and \( \mathfrak{p} \), respectively. This gives

\[
\text{Ad}_{a_t}|_n = e^t \times \text{Id} \quad \text{and} \quad \text{Ad}_{a_t}|_{\overline{n}} = e^{-t} \times \text{Id}
\]
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 \cite[Chapter VI.4]{17}. For $g \in G$, we write
\[ g = \kappa(g) \exp(H(g)) n_g, \]
where $\kappa(g) \in K$, $H(g) \in \text{Lie}(A) = a$, and $n_g \in N$. Since $a$ is one-dimensional, we may identify $a^\ast$ with $\mathbb{C}$ as follows: let $\alpha \in a^\ast$ be the unique element such that
\[ \alpha(H(a_t)) = t. \]
Then $s \in \mathbb{C}$ is viewed as an element of $a^\ast_{\mathbb{C}}$ via the map $s \mapsto s \cdot \alpha$. In particular, we write
\[ e^{sH}(g) = e^{s \cdot \alpha(H(g))} \]
for all $g \in G$ and $s \in \mathbb{C}$.

2.2. Various measures on $T^1(\mathcal{M})$. Let $\Lambda \subset \partial \mathbb{H}^{d+1}$ denote the limit set of $\Gamma$. Throughout the paper, we assume that
\[ \Gamma \text{ is geometrically finite}, \]
that is, the unit neighborhood of the convex core of $\mathcal{M}$, given by $\Gamma/\text{hull}(\Lambda)$ has finite volume. We let $\delta$ denote the critical exponent of $\Gamma$, which is known to be equal to the Hausdorff dimension of $\Lambda$. We remark that $\delta = d$ if and only if $\Gamma$ is a lattice in $G$.

We fix $o \in \mathbb{H}^{d+1}$ whose stabilizer is given by $K$, and $\nu_o \in T_o(\mathbb{H}^{d+1})$ whose stabilizer is $M$. Let $\nu_o$ be the Patterson-Sullivan measure on $\Lambda$ 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 $|\gamma'|$ is the derivative of $\gamma$ in the spherical metric on $\partial \mathbb{H}^{d+1}$ with respect to $o$. We also let $m_o$ be the $K$-invariant probability measure on $\partial \mathbb{H}^{d+1}$.

Let $\pi : T^1(\mathbb{H}^{d+1}) \to \mathbb{H}^{d+1}$ be the base point projection. For $u \in T^1(\mathbb{H}^{d+1})$, 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-parametrization of $T^1(\mathbb{H}^{d+1})$ given by:
\[ u \mapsto (u^+, u^-, s = \beta_{u-}(o, \pi(u))), \]
where $\beta$ denotes the Busemann function. Using the Hopf coordinates, the BMS measure $m_{\text{BMS}}$, the Liouville measure $du$, and the BR measure $m_{\text{BR}}$ on $T^1(\mathbb{H}^{d+1})$ are given as follows:

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

The BR measure $m_{\text{BR}}$ is a Lebesgue measure on each $N$-leaf, so we call it the unstable BR measure. The stable BR-measure $m_{\text{BR}}^\ast$ is defined similarly to $m_{\text{BR}}$ by exchanging the roles of $u^+$ and $u^-$. These measures are all left $\Gamma$-invariant, and hence induce Borel measures on $T^1(\mathcal{M}) = \Gamma \setminus T^1(\mathbb{H}^{d+1})$ for which we use the same notation. We remark that $m_{\text{BMS}}$ is a finite measure, invariant under the geodesic flow. We will
normalize the Patterson-Sullivan measure $\nu_o$ so that $m_{\text{BMS}}^o(\Gamma^1(M)) = 1$. The other three measures are infinite, unless $\delta = d$.

We will sometimes consider these measures as measures on $\Gamma \backslash G$ by putting:

$$m^o(\psi) = m^o(\int_M \psi dm).$$

Note that the Liouville measure considered as a measure on $\Gamma \backslash G$ is $G$-invariant, which we will denote by $dg$.

2.3. The base eigenfunction $\phi_0$. Throughout the article we assume

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

As mentioned in the introduction, $\lambda_0 = \delta(d - \delta)$ is the smallest eigenvalue of $\Delta$ on $L^2(M)$, which is known to be simple. We denote by $\phi_0$ the the unit eigenfunction in $L^2(M)$ with eigenvalue $\lambda_0$, and call it the base eigenfunction. Up to a constant multiple, $\phi_0$ is a positive function given by:

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

The fact that, for $\delta > d/2$, the base eigenfunction $\phi_0$ is square-integrable 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 [19]). The intersection of the interval $[0, \frac{d^2}{4}]$ with the spectrum of $\Delta$, viewed as an unbounded operator on $L^2(M)$, consists of a finite set of eigenvalues $\{\lambda_i = s_i(d - s_i)\}_{0 \leq i \leq \ell}$, satisfying

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

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 $(\rho, L^2(\Gamma \backslash G))$ of $G$, where $\rho$ is right-translation, i.e.

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

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, and corresponding notation is used for any subspace of $L^2(\Gamma \backslash G)$. We use $L^2(\Gamma \backslash G)^\text{sph}$ to construct subrepresentations of $(\rho, L^2(\Gamma \backslash G))$ as follows: define

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

Then

$$(\rho, L^2(\Gamma \backslash G)) = (\rho, L^2(\Gamma \backslash G)^\text{sph}) \oplus (\rho, L^2(\Gamma \backslash G)^{\perp\text{sph}});$$
note that both $L^2(\Gamma \backslash G)\text{sph}$ and $L^2(\Gamma \backslash G)_{\text{sph}}^\perp$ are $\rho(G)$-invariant closed subspaces. Viewing the base eigenfunction $\phi_0$ as an element of $L^2(\Gamma \backslash G)^K$, we define in a similar manner
\[
\mathcal{B}_\delta = \text{the closure of the span of}\{\rho(g)\phi_0 : g \in G\} \subset L^2(\Gamma \backslash G)_{\text{sph}},
\]
and
\[
(\rho, L^2(\Gamma \backslash G)_{\text{sph}}) = (\rho, \mathcal{B}_\delta) \oplus (\rho, \mathcal{W}),
\]
$\mathcal{W}$ being the orthogonal complement of $\mathcal{B}_\delta$ in $L^2(\Gamma \backslash G)_{\text{sph}}$. 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 that we make of in Section 5 is that for all $f \in L^2(\Gamma \backslash G)^K$,
\[
m^{\text{BR}}(f) = m^{\text{BR}*}(f) = \langle f, \phi_0 \rangle. \tag{2.1}
\]
Finally, the direct integral decomposition (cf. Corollary 14.9.5) of $(\rho, L^2(\Gamma \backslash G))$ reads
\[
(\rho, L^2(\Gamma \backslash G)) \cong \int_{Z}^\oplus (\pi_\zeta \mathcal{H}_\zeta) \, d\mu_Z(\zeta), \tag{2.2}
\]
where $(\pi_\zeta, \mathcal{H}_\zeta)$ is an irreducible unitary representation of $G$ for $\mu_Z$-a.e. $\zeta$.

2.5. Sobolev norms on unitary representations of $K$. Given a unitary representation $(\pi, V)$ of $K$ with invariant inner product $(\cdot, \cdot)_V$ and a basis $\{X_j\}$ of $\mathfrak{k}$, we define a Sobolev norm $\| \cdot \|_{S^m(V)} (m \in \mathbb{N})$ on $V$ by
\[
\|v\|_{S^m(V)}^2 := \|v\|_V^2 + \sum_{U} \|d\pi(U)v\|_V^2 \quad \text{for any } v \in V, \tag{2.3}
\]
where $\| \cdot \|_V$ denotes the corresponding norm and the sum runs over all monomials in the chosen basis of order $0 \leq i \leq m$.

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

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

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

**Lemma 2.2.** If $m > \frac{(d+1)d}{2}$, then for any $f \in S^m(\Gamma \backslash G)$,
\[
m^{\text{BR}}(f), m^{\text{BR}*}(f) \ll S^m(f).
\]

**Proof.** Given $f \in C(\Gamma \backslash G) \cap L^2(\Gamma \backslash G)$, define
\[
f_K(x) := \max_{k \in K} |f(xk)| \quad \text{for } x \in \Gamma \backslash G.
\]
By construction, $f_K$ is $\rho(K)$-invariant and $|f(x)| \leq f_K(x)$ for all $x \in \Gamma \setminus G$, hence $|m_{\text{BR}}(f)| \leq m_{\text{BR}}(f_K)$. Using [2.1] and the Sobolev embedding theorem on the compact manifold $K$ gives

$$|m_{\text{BR}}(f)| \leq m_{\text{BR}}(f_K) = \|f_K\| \leq \|f\|S^m(\Gamma \setminus G)$$

for all $m > \frac{\dim(K)}{2}$ (cf. [1, Theorem 2.30]). The claim for $m_{\text{BR}}^*$ follows similarly. \hfill \Box

3. Complementary series representations and Eisenstein integrals

In this section we recall the definition and $K$-type structure of the complementary series representations of $G = \text{SO}^0(d+1,1)$, 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.

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

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

Similarly, the irreducible representations of $\text{SO}(2m+1)$ 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 $\text{SO}(2m+1)$ and $\text{SO}(4m)$ are self-dual, and that the dual $\tau^*$ of an irreducible representation $\tau = (\tau_1, \tau_2, \ldots, \tau_{2m+1})$ of $\text{SO}(4m+2)$ is given by $\tau^* = (\tau_1, \tau_2, \ldots, -\tau_{2m+1})$.

A key result that we will make repeated use of is the branching law for restrictions of representations of $K$ to $M$, cf. [17, Chapter IX.3] or [40, pp. 377-380]. Let $\tau$ be an irreducible representation of $K$ (we henceforth call irreducible representations of $K$ or $M$ a $K$ or $M$-type, respectively). 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

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

and

$$\tau_1 \geq \sigma_1 \geq \tau_2 \geq \sigma_2 \geq \tau_{m} \geq \sigma_{m} \quad \text{for } d = 2m,$$

and $m_{\sigma,\tau} = 0$ otherwise. If $m_{\sigma,\tau} = 1$, we say that $\tau$ contains $\sigma$, and write $\sigma \subset \tau$. The fact that each $M$-type occurs at most once in $\tau$ will allow us to make repeated use of Schur’s lemma in the following manner: if a linear operator on $V_\tau$ ($V_\tau$ being a vector space on which $\tau$ is realized) commutes
with $\tau(M)$, then it acts as a scalar on each $\sigma \subseteq \tau$. 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 $(\pi, H)$ be a unitary representation of $G$. For each $\tau \in \hat{K}$, we define an operator $P_\tau$ by

$$P_\tau = \int_K \overline{\chi_\tau(k)} \pi(k) \, dk,$$

where $\chi_\tau(k) = \dim(V_\tau) \cdot \text{tr}(\tau(k))$ (the trace being defined with respect to any invariant inner product on any realization of $\tau$). Note that $P_\tau$ is the orthogonal projection onto the space $H_\tau$, where

$$H_\tau := \{ v \in H : P_\tau v = v \}.$$

This gives rise to a decomposition of $H$ as the orthogonal direct sum

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

Each $H_\tau$ has the property that $\pi(k)|H_\tau \cong \tau(k)$ ($\forall k \in K$). If $H_\tau \neq 0$, we say that $\pi$ contains $\tau$, and write $\tau \subseteq \pi$. There is a similar decomposition of $H$ with respect to $M$-types, and projection operators $P_\sigma = \int_M \chi_\sigma(m) \pi(m) \, dm$.

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

$$\sum_{\tau \subseteq V} \dim(\tau)^\alpha \|P_\tau v\|_{\mathcal{S}^m(V)} \ll \|v\|_{\mathcal{S}^{m+m_0(\alpha+1)}(V)},$$

where the implied constant depends only on $K$.

**Proof.** Though the argument is standard (cf. [37, Chapter 4.4.2] or [16, Lemmas 10.3 and 10.4]), we briefly recount it: using the Harish-Chandra isomorphism and the theorem of the highest weight, we find an element $\Omega_K$ 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 $c_\tau$ on $\tau$, where $|c_\tau| \geq \dim(\tau)$. Letting $m_0$ be the order of $\Omega_K$, for any $\ell \in \mathbb{N}$, we have

$$\sum_{\tau \subseteq V} \dim(\tau)^\alpha \|P_\tau v\|_{\mathcal{S}^m(V)} = \sum_{\tau \subseteq V} \frac{\dim(\tau)^\alpha}{|c_\tau|^{\ell}} \|P_\tau d\pi(\Omega_K^\ell) v\|_{\mathcal{S}^m(V)}$$

$$\leq \left( \sum_{\tau \subseteq V} \dim(\tau)^{2\alpha - 2\ell} \sum_{\tau \subseteq V} \|P_\tau d\pi(\Omega_K^\ell) v\|_{\mathcal{S}^m(V)}^2 \right)^{1/2}$$

$$\ll \left( \sum_{\tau \subseteq V} \dim(\tau)^{2\alpha - 2\ell} \|v\|_{\mathcal{S}^{m+\ell m_0}(V)} \right)^{1/2}.$$

The sum $\sum_{\tau \subseteq V} \frac{1}{\dim(\tau)}$ is finite by [16, Lemma 13], so choosing $\ell = \lceil \alpha + 1 \rceil$ gives $(\sum_{\tau \subseteq V} \dim(\tau)^{2\alpha - 2\ell})^{1/2} < \infty$. \qed
3.2. Complementary series representations. We will now recall Hirai’s classification \cite{hirai} of the non-tempered unitary dual of $G$ and construct the models of the complementary series that we will work with, see \cite{yoshida} Chapter 5.5 (cf. also \cite{kuwamura} Sections 3.1-3.3).

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

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

for all $v \in L^2(K)$, $g \in G$, and $k \in K$. We let $\lambda$ and $\rho$ denote the left- and right-regular representations of $K$ on $L^2(K)$. 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 $\tau$. Given $v \in \hat{M}$, we define

$$L^2(K : v) := \{ v \in L^2(K) : \rho(m)v = v(m)v \text{ for all } m \in M \} = \left\{ v \in L^2(K) : \int_M \chi_v(m)\rho(m)v \, dm = v \right\}.$$

Noting that $U^s(g)$ and $\rho(m)$ commute for all $g \in G$, $m \in M$ shows that $(U^s, L^2(K : v))$ is a representation of $G$; in fact $(U^s, L^2(K : v))$ is isomorphic to $\dim(v)$ copies of $\text{ind}_{M}^{G} \chi_{K} \otimes s \otimes 1$ (cf. \cite{kazhdan} Chapter VII). Note that $L^2(K : v)_{\tau} \subset L^2(K)$ for all $\tau \in \hat{K}$. From Hirai’s classification of the unitary dual of $G$ \cite{hirai}, for each $v \in \hat{M}$ there exists an interval $I_v \subset (\frac{d}{2}, d)$ such that if $s \in I_v$, then there exists an irreducible, unitarizable subrepresentation $\mathcal{U}(v, s)$ of $(U^s, L^2(K : v))$. Furthermore, every non-tempered representation of $G$ may be realized in this way, cf. \cite{yoshida} Theorem 5.5.1.5 \cite{kazhdan} Theorem 8.37. Hirai also gave a description of the $K$-types of $\mathcal{U}(v, s)$: each $\tau$ that contains $v$ occurs exactly once in $\mathcal{U}(v, s)$.

3.3. Eisenstein integrals. Here we develop the Eisenstein integrals needed to obtain the desired asymptotic expansion of matrix coefficients (cf. \cite{kuwamura} Section 3.3, \cite{mackey} Chapter 6). Before starting, we make the following elementary but important observation:

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

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

**Proof.** Since $v \in L^2(K : v)$, $v = \int_M \rho(m)v\chi_v(m) \, dm$. Hence

$$v(m) = \int_M v(mm_1)\chi_v(m_1) \, dm_1 = \int_M v(m_1m)\chi_v(m_{1}^{-1}m_1m) \, dm_1 = \int_M v(m_1^{-1}m)\chi_v(m_1^{-1}m) \, dm_1 = [P_v v](m).$$
Now the orthogonality of the $M$-types of $L^2(K)$ gives $P_{v^*}v = 0$, yielding the claim.

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 $\mathcal{U}(v, s)$. The inner product that makes $\mathcal{U}(v, s)$ a unitary representation is denoted $\langle \cdot, \cdot \rangle_{\mathcal{U}(v, s)}$. For each $K$-type $\tau$ of $\mathcal{U}(v, s)$ we define a vector $\chi_\tau \in \mathcal{U}(v, s)_\tau$ by

$$
\chi_\tau = \sum_{i=1}^{\dim(\tau)} v_i(e) v_i,
$$

(3.1)

where $\{v_i\}_{i=1}^{\dim(\tau)}$ is an orthonormal basis of $\mathcal{U}(v, s)_\tau$. Note that $\chi_\tau$ is independent of the choice of basis, and for all $v \in \mathcal{U}(v, s)_\tau$, $k \in K$, we have

$$
v(k) = \langle v, U^s(k)\chi_\tau \rangle_K.
$$

(3.2)

**Lemma 3.3.** For all $v \in \mathcal{U}(v, s)_\tau \cap \mathcal{U}(v, s)_{v^*}$, we have:

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

Proof. Let $\{v_j\}$ be an orthonormal basis of $\mathcal{U}(v, s)_\tau \cap \mathcal{U}(v, s)_{v^*}$ with respect to $\langle \cdot, \cdot \rangle_K$. Using (3.1) and the Schur orthogonality relations, we get

$$
\int_M |v(m)|^2 dm = \int_M |\langle v, U^s(m)\chi_\tau \rangle_K|^2 dm = \sum_{i,j} |v_i(e)v_j(e)|^2 \int_M \langle U^s(m)v_i, v_i \rangle_K \langle U^s(m)v_j, v_j \rangle_K dm = \frac{\|v\|_K^2 \sum_i |v_i(e)|^2}{\dim(v)}.
$$

Letting $\{v_j\} \cup \{w_i\}$ be an orthonormal basis for all of $\mathcal{U}(v, s)_\tau$ and recalling $w_i(e) = 0$ for all $i$ by Lemma 3.2 we have

$$
\sum_j |v_j(e)|^2 = \sum_j |v_j(e)|^2 + \sum_i |w_i(e)|^2 = \dim(\tau).
$$

(3.3)

For $K$-types $\tau_1, \tau_2$, we define an operator

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

by

$$
T_{\tau_2}^{\tau_1}v = \int_M v(m)U^s(m)\chi_{\tau_2} dm \quad \text{for all } v \in \mathcal{U}(v, s)_{\tau_1}.
$$

(3.4)

We have the following interpretation of the Eisenstein integral (a similar (but different) formula appears in [21 Theorem 3.4]):
Theorem 3.4. For any $K$-types $\tau_1$, $\tau_2$ of $\mathcal{U}(v, s)$ and $g \in G$, we have

$$P_{\tau_2} U^s(g) P_{\tau_1} = \int_K e^{(s-d)H(gk)} U^s(\kappa(gk)) T_{\tau_1}^2 U^s(k^{-1}) \, dk,$$

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

Proof. Following [38, Theorem 6.2.4, pp. 42-43], the key fact is that

$$dk = e^{dH(gk)} d\mu_K(\kappa(gk)).$$

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

$$\langle U^s(g)v, w \rangle_K = \int_K e^{-sH(g^{-1}k)} v(\kappa(g^{-1}k)) \overline{w(k)} \, dk$$

$$= \int_K e^{-sH(g^{-1}k)} v(\kappa(g^{-1}k)) \overline{w(k)} e^{(d-s)H(g^{-1}k)} d\mu_K(\kappa(g^{-1}k))$$

$$= \int_K e^{(d-s)H(g^{-1}k)} v(k) \overline{w(\kappa(gk))} \, dk = \int_K e^{(s-d)H(gk)} v(k) \overline{w(\kappa(gk))} \, dk,$$

where we used $g^{-1} \kappa(gk) = k \exp(-H(gk))(n_{kg}^{-1}) \exp(H(gk))$. By (3.2), we get

$$v(k) \overline{w(\kappa(gk))} = \langle v, U^s(k) \chi_{\tau_1} \rangle_K \langle w, U^s(\kappa(gk)) \chi_{\tau_2} \rangle_K$$

$$= \langle U^s(k^{-1})v, \chi_{\tau_1} \rangle_K \langle U^s(\kappa(gk)) \chi_{\tau_2}, w \rangle_K$$

$$= \langle (U^s(k^{-1})v, \chi_{\tau_1}) \cdot U^s(\kappa(gk)) \chi_{\tau_2}, w \rangle_K.$$

Hence

$$\langle U^s(g)v, w \rangle_K = \int_K e^{(s-d)H(gk)} \langle (U^s(k^{-1})v, \chi_{\tau_1}) \cdot U^s(\kappa(gk)) \chi_{\tau_2}, w \rangle_K \, dk$$

$$= \int_K e^{(s-d)H(gk)} \langle U^s(k^{-1})v, \chi_{\tau_1} \rangle_{\tau_2} \cdot U^s(\kappa(gk)) \chi_{\tau_2}, w \rangle_K \, dk.$$

Thus, if we define $T_0 : \mathcal{U}(v, s)_{\tau_1} \rightarrow \mathcal{U}(v, s)_{\tau_2}$ by

$$T_0 v = \langle v, \chi_{\tau_1} \rangle_{\tau_2} = v(e) \chi_{\tau_2},$$

then

$$P_{\tau_2} U^s(g) P_{\tau_1} = \int_K e^{(s-d)H(gk)} U^s(\kappa(gk)) T_0 U^s(k^{-1}) \, dk.$$
We now replace $T_0$ by an operator that commutes with $U^s(M)$ as follows: writing $g = k_1a_1k_2$, we have

\[
P_{\tau_2}U^s(g)P_{\tau_1} = U^s(k_1)P_{\tau_2}U^s(a_1)P_{\tau_1}U^s(k_2)
\]

\[
= \int_M U^s(k_1)U^s(m)P_{\tau_2}U^s(a_1)P_{\tau_1}U^s(m^{-1})U^s(k_2) \, dm
\]

\[
= U^s(k_1)\int_M \int_K e^{(s-d)H(a_1k)}U^s(m)\tau_2U^s(k^{-1}m^{-1}) \, dk \, dm \, U^s(k_2)
\]

\[
= \int_K e^{(s-d)H(a_1k)}U^s(k_1\tau_2)(\int_M U^s(m)T_0U^s(m) \, dm)U^s(k^{-1}k_2) \, dk
\]

\[
= \int_K e^{(s-d)H(gk)}U^s(\tau_2U^s(k)) \left(\int_M U^s(m)T_0U^s(m) \, dm\right)U^s(k^{-1}) \, dk.
\]

Therefore

\[
P_{\tau_2}U^s(g)P_{\tau_1} = \int_K e^{(s-d)H(gk)}U^s(\tau_2U^s(k))T'U^s(k^{-1}) \, dk,
\]

where

\[
T = \int_M U^s(m)T_0U^s(m) \, dm.
\]

Letting $T_{\tau_1}^{\tau_2} = T$ gives the first identity claimed in the theorem. The second identity follows from the fact that $\langle U^s(m^{-1})v, \chi_{\tau_1} \rangle_K = v(m)$ for all $v \in U(v, s)_{\tau_1}$ and $m \in M$. \hfill \Box

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

**Corollary 3.5.** If $v \in U(v, s)_{\tau_1}$ is orthogonal to $U(v, s)_{\tau_2}$, then

\[
T_{\tau_1}^{\tau_2}v = 0.
\]

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

**Proof.** We have $T_{\tau_1}^{\tau_2}v = \int_M v(m)U^s(m)\chi_{\tau_2} \, dm$. By Lemma 3.2 if $v$ is orthogonal to $U(v, s)_{\tau_2}$, then $v(m) = 0$ and thus $T_{\tau_1}^{\tau_2}v = 0$. Since $T_{\tau_1}^{\tau_2}$ commutes with $U^s(M)$, Schur’s lemma then gives that if $T_{\tau_1}^{\tau_2}v$ is non-zero, it must be in $U(v, s)_{\tau_2} \cap U(v, s)_{\tau_2^*}$. \hfill \Box

**Lemma 3.6.** For any $\tau_1, \tau_2 \in \hat{K}$, 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. Let \( u \in \mathcal{U}(v, s)_{\tau_1}, \ v \in \mathcal{U}(v, s)_{\tau_2} \). Then

\[
\langle T_{\tau_2}^\tau u, v \rangle_K = \int_M u(m) \langle U^s(m) \chi_{\tau_2}, v \rangle_K dm \\
= \int_M \langle u, U^s(m) \chi_{\tau_1} \rangle_K \langle U^s(m) \chi_{\tau_2}, v \rangle_K dm \\
= \left\langle u, \int_M (U^s(m)^{-1}) v, \chi_{\tau_2} \rangle_K \cdot U^s(m) \chi_{\tau_1} dm \right\rangle \\
= \langle u, T_{\tau_2}^\tau v \rangle_K.
\]

\( \square \)

**Corollary 3.7.** For any \( \tau_1, \tau_2 \in \hat{K} \),

\[
\| T_{\tau_2}^\tau \|_K \leq \frac{\sqrt{\dim(\tau_1) \dim(\tau_2)}}{\dim(v)}.
\]

**Proof.** Given a unit vector \( v \in \mathcal{U}(v, s)_{\tau_1} \cap \mathcal{U}(v, s)_{v^*} \), we have

\[
\| T_{\tau_2}^\tau \|_K^2 = \langle T_{\tau_2}^\tau v, T_{\tau_2}^\tau v \rangle_K = \int_M \overline{v(m)} \langle T_{\tau_2}^\tau v, U^s(m) \chi_{\tau_2} \rangle_K dm \\
= \int_M \int_M \overline{v(m_1)} v(m_2) \langle U^s(m_2) \chi_{\tau_2}, U^s(m_1) \chi_{\tau_2} \rangle_K dm_2 dm_1 \\
= \int_M \int_M \overline{v} \langle U^s(m_1) \chi_{\tau_1} \rangle_K \langle v, U^s(m_2) \chi_{\tau_2} \rangle_K \chi_{\tau_2}(m_2^{-1} m_1) dm_2 dm_1 \\
= \int_M \left( \int_M \overline{v} \langle U^s(m_2 m_3) \chi_{\tau_1} \rangle_K \langle v, U^s(m_2) \chi_{\tau_2} \rangle_K dm_2 \right) \chi_{\tau_2}(m_3) dm_3 \\
= \| v \|_K^2 \int_M \overline{\chi_{\tau_1}(m)} \chi_{\tau_2}(m) dm \\
\leq \frac{\dim(v)}{\dim(\tau_1)} \sqrt{\int_M |\chi_{\tau_1}(m)|^2 dm} \sqrt{\int_M |\chi_{\tau_2}(m)|^2 dm}.
\]

Lemma 3.3 together with the identity \( \| \chi_{\tau_i} \|_K^2 = \dim(\tau_i) \) completes the proof. \( \square \)

**Corollary 3.8.** For \( \tau \in \hat{K} \) and \( v \in \mathcal{U}(v, s)_{\tau} \), we have

\[
T_{\tau}^\tau v = \frac{\dim(\tau)}{\dim(v)} P_{v^*} v.
\]

**Proof.** Following the proof of Corollary 3.7, we obtain

\[
\| T_{\tau}^\tau \|_K^2 = \frac{1}{\dim(v)} \int_M |\chi_{\tau}(m)|^2 dm = \frac{\dim(\tau)^2}{\dim(v)^2}.
\]

So by Corollary 3.5, we have

\[
T_{\tau}^\tau = c \cdot P_{v^*},
\]
where $c$ is one of $\pm \frac{\dim(\tau)}{\dim(\upsilon)}$. Since
\[ c = \langle T_\tau^* \upsilon, \upsilon \rangle_K = \int_M |\upsilon(m)|^2 \, dm \geq 0 \]
for any unit vector $\upsilon \in \mathcal{U}(\upsilon, s) \cap \mathcal{U}(\upsilon, s)^*$, we get $c = \frac{\dim(\tau)}{\dim(\upsilon)}$. \hfill \Box

4. Asymptotic expansions of matrix coefficients

We fix a complementary series representation $\mathcal{U}(\upsilon, s)$ of $G$. The main goal of this section is to obtain an effective expansions of matrix coefficients for $\mathcal{U}(\upsilon, s)$. More precisely, we will write a matrix coefficient $\langle U_s(a)t, v \rangle_{\mathcal{U}(\upsilon, s)}$ 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 the Harish-Chandra $c$-function 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 $\text{dist}_K$ denote the bi-invariant Riemannian metric on $K$ induced from the (negative of the) Killing form on $\mathfrak{k}$, with $|\cdot|_K$ denoting the corresponding norm on $\mathfrak{k}$.

**Lemma 4.1.** Let $(\pi, V)$ be a finite-dimensional unitary representation of $K$ with invariant inner product $\langle \cdot, \cdot \rangle_V$. Then for all $\upsilon \in V$ and $k \in K$, we have
\[ \| \pi(k)\upsilon - \upsilon \|_V \ll \text{dist}_K(k, e) \| \upsilon \|_{S^1(V)}. \]
The implied constant depends solely on the choice of basis defining $\| \cdot \|_{S^1(V)}$.

**Proof.** Since $\text{dist}_K$ is bi-invariant, there exists $J \in \mathfrak{k}$ with $|J|_K = 1$ such that the $k = \exp(\text{dist}_K(k, e)J)$. This gives
\[
\pi(k)\upsilon - \upsilon = \int_0^{\text{dist}_K(k, e)} \frac{d}{dt} \pi(\exp(tJ))\upsilon \, dt \\
= \int_0^{\text{dist}_K(k, e)} \pi(\exp(tJ))d\pi(\text{Ad}_{\exp(-tJ)}J)\upsilon \, dt.
\]
Since $\pi$ is unitary, $|J|_K = 1$, and $K$ is compact,
\[
\| \pi(\exp(tJ))d\pi(\text{Ad}_{\exp(-tJ)}J)\upsilon \|_V \leq \max_{X \in \mathfrak{k}} \| d\pi(X)\upsilon \|_V \ll \| \upsilon \|_{S^1(V)}.
\]
\hfill \Box

We now arrive at the main technical result of this section (cf. [21 Theorem 3.23]): The following is the so-called Harish-Chandra $c$-function:
\[ C_+(s) = \int_\mathbb{N} U^s(\kappa(\pi)^{-1})e^{-sH(\pi)} \, d\pi. \] (4.1)
Recall that $\| \cdot \|_{S(K)}$ denotes the $L^2(K)$ Sobolev norm and set

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

**Theorem 4.2.** For all $u \in \mathcal{U}(v, s)_{\tau_1}, v \in \mathcal{U}(v, s)_{\tau_2},$ and $t \geq 0,$ we have

$$\langle U^s(a_t)u, v \rangle_K = e^{(s-d)t} \langle T_{\tau_2}^t \mathcal{C}_+(s)u, v \rangle_K$$

$$+ O_s \left( e^{(s-d-\eta_s)t} \| T_{\tau_1}^t \|_K \| u \|_K \| v \|_{S^1(K)} \right),$$

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

**Proof.** Applying Theorem 3.4, we have

$$\langle U^s(a_t)u, v \rangle_K = \int_K e^{(s-d)H(\alpha_k)} \langle U^s(\kappa(\alpha_k))T_{\tau_1}^t U^s(k^{-1})u, v \rangle_K \, dk.$$ 

Since the function $k \mapsto e^{(s-d)H(\alpha_k)} \langle U^s(\kappa(\alpha_k))T_{\tau_1}^t U^s(k^{-1})u, v \rangle_{U(v, s)}$ is right $M$-invariant, we may use the integration formula \cite[Consequence 3, p. 147]{16} to obtain

$$= \int_{\mathcal{N}} e^{(s-d)H(\alpha_k(\overline{\pi}))} \langle U^s(\kappa(\alpha_k(\overline{\pi})))T_{\tau_1}^t U^s(\kappa(\overline{\pi})^{-1})u, v \rangle_K e^{-dH(\overline{\pi})} d\overline{\pi}.$$ 

The identities

$$H(\alpha_k(\overline{\pi})) = H(\alpha_k(\overline{\theta})) + H(\overline{\pi}) - H(\overline{\theta})$$

$$\kappa(\alpha_k(\overline{\pi})) = \kappa(\alpha_k(\overline{\theta}))$$

then give that the previous integral is equal to

$$e^{(s-d)t} \int_{\mathcal{N}} \langle T_{\tau_1}^t U^s(\kappa(\overline{\pi})^{-1})u, U^s(\kappa(\alpha_k(\overline{\theta}))^{-1})v \rangle_K e^{(s-d)H(\alpha_k(\overline{\theta})^{-1}) - sH(\overline{\pi})} d\overline{\pi}.$$ 

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

$$e^{(s-d)t} \int_{\mathbb{R}^d} \langle T_{\tau_1}^t U^s(\kappa(\overline{\pi})^{-1})u, U^s(\kappa(\alpha_k(\overline{\pi})^{-1})v \rangle_K e^{(s-d)H(\alpha_k(\overline{\pi})^{-1}) - sH(\overline{\pi})} d\overline{\pi}.$$ 

(4.3)

The integral being absolutely convergent due to $s > \frac{d}{2}$ and then change to spherical coordinates: $x = r \theta, r \geq 0, \theta \in S^{d-1}.$ Since the isomorphism $\mathcal{N} \cong \mathbb{R}^d$ can be chosen so that $e^{H(\overline{\pi}, \overline{\theta})} \asymp 1 + r^2$ (cf. \cite[Lemma 7.20]{16}) we have

$$\langle U^s(a_t)u, v \rangle_K = e^{(s-d)t} \langle T_{\tau_2}^t \mathcal{C}_+(s)u, v \rangle_K +$$

$$O \left( e^{(s-d)t} \int_{\mathbb{S}^{d-1}} \langle T_{\tau_1}^t U^s(\kappa(\overline{\pi})^{-1})u, w(e^{-t}r, \theta) \rangle_K dm(\theta) (1 + r^2)^{-s} r^{d-1} dr \right),$$

where $dm(\theta)$ denotes the spherical measure, and

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

Since $\overline{\pi} \mapsto \kappa(\overline{\pi})$ is smooth,

$$d_K(\kappa(\overline{\pi}), e) \ll r \quad \text{for all } r > 0, \theta \in S^{d-1}. $$
Using \((1 + r^2)^{s-d} = 1 + O_s(r)\) (with the implied constant depending continuously on \(s\)) and Lemma 4.1, we have

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

This gives

\[
\langle U^s(at)u, v \rangle_K = e^{(s-d)t} \langle T_{\tau_2}^s C_+ (s) u, v \rangle_K
\]

\[+ O_s \left( \|T_{\tau_1}^s \|_K \|u\|_K \|v\|_{S^1(K)} e^{(s-d)t} \int_0^\infty \min\{1, e^{-r} \} (1 + r^2)^{-s} r^{d-1} dr \right). \]

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

\[
\int_0^\infty \min\{1, e^{-r} \} (1 + r^2)^{-s} r^{d-1} dr \leq e^{-t} + e^{-t} \int_1^{e^t} r \cdot r^{-2s} \cdot r^{d-1} dr + \int_{e^t}^\infty r^{-2s} \cdot r^{d-1} dr \ll_s e^{-t} + e^{(d-2s)t}.
\]

\[\square\]

### 4.2. The invariant inner product on \(\mathcal{U}(v, s)\)

The intertwining operator \(A(v, s)\) on \(\mathcal{U}(v, s)\) is defined so that

\[
\langle u, v \rangle_{\mathcal{U}(v, s)} = \langle u, A(v, s) v \rangle_K
\]

for all \(K\)-finite vectors \(u, v \in \mathcal{U}(v, s)\). The key intertwining property of \(A(v, s)\) reads

\[
A(v, s) U^s(g) = U^{d-s}(g) A(v, s) \quad \text{for all } g \in G.
\]

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

\[
A(v, s) = \sum_{\tau \supset v} a(v, s, \tau) P_\tau.
\]

The positive definiteness of the inner product \(\langle \cdot, \cdot \rangle_{\mathcal{U}(v, s)}\) implies that for all \(\tau\) in \(\mathcal{U}(v, s)\),

\[
a(v, s, \tau) > 0.
\]

**Proposition 4.3.** Let \(d \geq 2\), and assume that \(\mathcal{U}(v, s)\) has a non-trivial \(M\)-invariant vector, and let \(\tau_1, \tau_2\) be two \(K\)-types of \(\mathcal{U}(v, s)\). If \(T_{\tau_1}^s \neq 0\), then

\[
\frac{a(v, s, \tau_2)}{a(v, s, \tau_1)} \ll_s (\dim(\tau_1) \dim(\tau_2))^{d/(d-1)},
\]

where the implied constant does not depend on \(v\), and is bounded over compact subsets of \((\frac{d}{2}, d)\).
Proof. Let $\omega \in N_K(A)$ be such that $\omega \omega^{-1} = a^{-1}$ for all $a \in A$. Since $\langle \cdot, \cdot \rangle_{U(v,s)}$ is $U^s$-invariant, given $u \in U(v,s)_{\tau_1}$, $v \in U(v,s)_{\tau_2}$, and $t \geq 0$, we have

$$a(v,s,\tau_2)\langle U^s(a_t)u,v \rangle_K = \langle U^s(a_t)u,v \rangle_{U(v,s)}$$

$$= \langle u,U^s(a_{-t})v \rangle_{U(v,s)}$$

$$= \langle u,U^s(\omega a_{t}\omega^{-1})v \rangle_{U(v,s)}$$

$$= \langle U^s(a_t)U^s(\omega^{-1})v,U^s(\omega^{-1})u \rangle_{U(v,s)}$$

$$= \langle a(v,s,\tau_1)(U^s(a_t)U^s(\omega^{-1})v),U^s(\omega^{-1})u \rangle_K.$$

So assuming $\langle U^s(a_t)u,v \rangle_K \neq 0$, we get that for all $t \geq 0$,

$$\frac{a(v,s,\tau_2)}{a(v,s,\tau_1)} = \frac{\langle U^s(a_t)U^s(\omega^{-1})v,U^s(\omega^{-1})u \rangle_K}{\langle U^s(a_t)u,v \rangle_K}.$$

Letting $t \to \infty$, Theorem 4.2 gives

$$\frac{a(v,s,\tau_2)}{a(v,s,\tau_1)} = \frac{\langle T^s_{\tau_2}C_+(s)U^s(\omega^{-1})v,U^s(\omega^{-1})u \rangle_K}{\langle T^s_{\tau_2}C_+(s)u,v \rangle_K}$$

$$= \frac{\langle T^s_{\tau_1}U^s(\omega^{-1})u,C_+(s)U^s(\omega^{-1})v \rangle_K}{\langle T^s_{\tau_1}C_+(s)u,v \rangle_K}.$$

Since $C_+(s)|_{U(v,s)_\tau} \in \text{End}(U(v,s)_\tau)$, and $C_+(s)$ commutes with $U^s(M)$, $C_+(s)$ acts as a scalar on each $M$-type of $U(v,s)_\tau$. We now choose $u$ to be a unit vector in $U(v,s)_{\tau_1} \cap U(v,s)_{\nu^*}$ such that

$$\|T^s_{\tau_1}u\|_K = \|T^s_{\tau_1}\|_K.$$

Let $v = T^s_{\tau_2}u$, and denote the scalar which $C_+(s)$ acts on $U(v,s)_{\tau} \cap U(v,s)_{\nu^*}$ as by $c(v,s,\tau)$. This gives

$$\frac{a(v,s,\tau_2)}{a(v,s,\tau_1)} = \frac{\langle T^s_{\tau_2}U^s(\omega^{-1})u,C_+(s)U^s(\omega^{-1})T^s_{\tau_2}u \rangle_K}{c(v,s,\tau_1)\|T^s_{\tau_2}\|_K^2}.$$

By Corollary 3.5, we have $T^s_{\tau_1}U^s(\omega^{-1})u \in U(v,s)_{\tau_2} \cap U(v,s)_{\nu^*}$. Hence

$$\langle T^s_{\tau_1}U^s(\omega^{-1})u,C_+(s)U^s(\omega^{-1})T^s_{\tau_2}u \rangle_K$$

$$= \langle T^s_{\tau_1}U^s(\omega^{-1})u,P_vC_+(s)U^s(\omega^{-1})T^s_{\tau_2}u \rangle_K$$

$$= c(v,s,\tau_2)\langle T^s_{\tau_1}U^s(\omega^{-1})u,P_vU^s(\omega^{-1})T^s_{\tau_2}u \rangle_K.$$

Therefore

$$\frac{a(v,s,\tau_2)}{a(v,s,\tau_1)} \leq \left| \frac{c(v,s,\tau_2)}{c(v,s,\tau_1)} \right|.$$

Using the parameterization of $K$ and $M$ types in Section 3.1, since $U(v,s)$ has an $M$-invariant vector, we may identify $v$ with the element $(v,0,\ldots,0) \in \mathbb{Z}^{|\frac{d}{2}|}$, where $v \geq 0$. Under this parameterization, all $K$-types $\tau$ of $U(v,s)$ are
then given as vectors \((\tau, 0, \ldots, 0) \in \mathbb{Z}[\frac{d+1}{2}]\) with \(\tau \geq v\). The scalars \(c(v, s, \tau)\) are given explicitly by [8, Theorem 8.2]: if \(d\) is even, then
\[
c(v, s, \tau) = \frac{\Gamma(s+1-d-v)\Gamma(s+1-v)}{\Gamma(s-d+1-\tau)\Gamma(s+\tau)} \cdot \frac{(d-1)!}{(\frac{d}{2}-1)!} \cdot \frac{1}{\prod_{j=1}^{d/2}(s-j)},
\]
and if \(d\) is odd,
\[
c(v, s, \tau) = \frac{\Gamma(s+1-d-v)\Gamma(s+1-v)}{\Gamma(s-d+1-\tau)\Gamma(s+\tau)} \cdot \frac{2^{d-2s}(d-1)!\Gamma(2s-d)}{(s-\frac{d}{2})^2} \cdot \frac{1}{\prod_{j=2}^{d-1/2}(s-j)}.
\]
In both cases we obtain
\[
\left|\frac{c(v, s, \tau_2)}{c(v, s, \tau_1)}\right| = \frac{\Gamma(s-d+1-\tau_1)\Gamma(s+\tau_1)}{\Gamma(s-d+1-\tau_2)\Gamma(s+\tau_2)} \leq \frac{\Gamma(s+1)}{\Gamma(d-s+1)} \frac{\Gamma(d-s+\tau_2)}{\Gamma(d-s+\tau_1)}.
\]
By [13, Theorem 1],
\[
\frac{\Gamma(s+\tau)}{\Gamma(d-s+\tau)} \approx_s 1 + s^{2s-d},
\]
so
\[
\left|\frac{c(v, s, \tau_2)}{c(v, s, \tau_1)}\right| \leq (1 + \tau_1^{2s-d})(1 + \tau_2^{2s-d}) \leq (1 + s^d)(1 + s^d).
\]
Now using the Weyl dimension formula (cf. [16, pp. 81-83, 107-109]),
\[
\dim(\tau) = \frac{(\tau + d - 2)!\tau!(d+1)}{\tau!^d(\tau+1)!} \gg \tau^{d-1}.
\]
This finishes the proof.

Proposition 4.3 allows us to restate Theorem 4.2 in terms of \(\langle \cdot, \cdot \rangle_{U(v, s)}\). 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 4.2, we have

**Theorem 4.4.** There exists \(m \in \mathbb{N}\) such that for any \(U(v, s)\) with an \(M\)-invariant vector, for all \(u \in U(v, s)_{\tau_1}, v \in U(v, s)_{\tau_2}\), and \(t \geq 0\),
\[
\langle U^s(a_t)u, v \rangle_{U(v, s)} = e^{(s-d)t}\langle T_{\tau_1}^sC_+(s)u, v \rangle_{U(v, s)} + O_s\left(e^{(s-d-\eta)t}\|u\|_{S^m(v, s)}\|v\|_{S^m(v, s)}\right),
\]
where the implied constant is uniformly bounded over \(s\) in compact subsets of \((\frac{d}{2}, d)\).

**Proof.** Since \(v \in U(v, s)_{\tau_2}\), using the expression for \(\langle \cdot, \cdot \rangle_{U(v, s)}\), Theorem 4.2 gives
\[
\langle U^s(a_t)u, v \rangle_{U(v, s)} = a(v, s, \tau_2)\langle U^s(a_t)u, v \rangle_K
\]
\[
= e^{(s-d)t}a(v, s, \tau_2)(T_{\tau_1}^s C_+(s)u, v)_K + O_s\left(e^{(s-d-\eta)t}a(v, s, \tau_2)\|T_{\tau_1}^s u\|_K\|v\|_{S^1(K)}\right)
\]
\[
= e^{(s-d)t}(T_{\tau_1}^s C_+(s)u, v)_{U(v, s)} + O_s\left(e^{(s-d-\eta)t}\sqrt{\frac{a(v, s, \tau_2)}{a(v, s, \tau_1)}}\|T_{\tau_1}^s u\|_K\|v\|_{S^1(v, s)}\right).
\]
When \( d \geq 2 \), Corollary 3.7 and Proposition 4.3 together give that for all \( \tau_1, \tau_2 \subset \mathcal{U}(v, s) \),
\[
\sqrt{\frac{a(v, s, \tau_2)}{a(v, s, \tau_1)}} \| T_{\tau_2}^\tau \|_K \ll_s \left( \dim(\tau_1) \dim(\tau_2) \right)^\frac{3}{2}.
\] (4.4)

Lemma 3.3 provides the bound claimed in the proposition. For \( d = 1 \), the left-hand side of (4.4) is bounded by the product of the weights of \( \tau_1 \) and \( \tau_2 \); this and the relation between the Sobolev norm and the weights imply the claim.

**Theorem 4.5.** There exists \( m \in \mathbb{N} \) such that for any non-trivial \( v \in \hat{M} \), \( s \in \mathcal{L} \), and for all \( M \)-invariant vectors \( u, v \in S^m(v, s) \),
\[
\| (U^s(a_t)u, v)_{\mathcal{U}(v, s)} \| \ll_s e^{(s-d-n_1)t} \| u \|_{S^m(v, s)} \| v \|_{S^m(v, s)},
\]
where the implied constant is uniformly bounded over \( s \) in compact subsets of \( \left( \frac{d}{2}, d \right) \).

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

\[
u = \sum_{\tau_1 \subset \mathcal{U}(v, s)} u_{\tau_1}, \quad v = \sum_{\tau_2 \subset \mathcal{U}(v, s)} v_{\tau_2},
\]
where \( u_{\tau_1} = P_{\tau_1} u \) and \( u_{\tau_2} = P_{\tau_2} v \). By [37, Theorem 4.4.2.1],
\[
\langle U^s(a_t)u, v \rangle_{\mathcal{U}(v, s)} = \sum_{\tau_1, \tau_2} \langle U^s(a_t)u_{\tau_1}, v_{\tau_2} \rangle_{\mathcal{U}(v, s)},
\]
with the sum converging absolutely. Applying Proposition 4.4 gives
\[
\langle U^s(a_t)u_{\tau_1}, v_{\tau_2} \rangle_{\mathcal{U}(v, s)} = e^{(s-d)t} t^{T_{\tau_2}^\tau} C_+(s) u_{\tau_1}, v_{\tau_2} \rangle_{\mathcal{U}(v, s)} + O_s(e^{(s-d-n_1)t} \| u_{\tau_1} \|_{S^m'(v, s)} \| v_{\tau_2} \|_{S^m'(v, s)})
\]
for some \( m' \in \mathbb{N} \). Since \( M \subset K \), each \( u_{\tau_1} \) is also \( M \)-invariant. Using the fact that \( C_+(s) \) preserves the \( M \)-types of each \( K \)-type, \( C_+(s) u_{\tau_1} \) is orthogonal to \( \mathcal{U}(v, s)_{v^r} \), and so by Corollary 3.5
\[
T_{\tau_2}^\tau C_+(s) u_{\tau_1} = 0.
\]
Hence
\[
\| (U^s(a_t)u_{\tau_1}, v_{\tau_2})_{\mathcal{U}(v, s)} \| \ll_s e^{(s-d-n_1)t} \| u_{\tau_1} \|_{S^m'(v, s)} \| v_{\tau_2} \|_{S^m'(v, s)}.
\]
Therefore, applying Lemma 3.1 we have for \( m \in \mathbb{N} \) large enough:
\[
\| (U^s(a_t)u, v)_{\mathcal{U}(v, s)} \| \ll_s e^{(s-d-n_1)t} \left( \sum_{\tau_1} \| u_{\tau_1} \|_{S^m'(v, s)} \right) \left( \sum_{\tau_2} \| v_{\tau_2} \|_{S^m'(v, s)} \right)
\ll_s e^{(s-d-n_1)t} \| u \|_{S^m(v, s)} \| v \|_{S^m(v, s)}.
\]
\[\square\]
Proposition 4.6. Let \( v \in \hat{M} \) be the trivial representation. Then there exists \( m \in \mathbb{N} \) such that for all \( u, v \in S^m(v, s) \) and \( t \geq 0 \),

\[
\langle U^s(a_t)u, v \rangle_{U(v,s)} = e^{(s-d)t} \left( \sum_{\tau_1, \tau_2 \in \check{K}} \langle T^{T_2}_{\tau_1}C_+(s) P_{\tau_1} u, P_{\tau_2} v \rangle_{U(v,s)} \right) + O_s(e^{(s-d-\eta)t}) \left( \sum_{\tau_1} \|u_{\tau_1}\|_{S^m(v,s)} \right) \left( \sum_{\tau_2} \|v_{\tau_2}\|_{S^m(v,s)} \right)
\]

and the sum

\[
\sum_{\tau_1, \tau_2 \in \check{K}} \langle T^{T_2}_{\tau_1}C_+(s) P_{\tau_1} u, P_{\tau_2} v \rangle_{U(v,s)}
\]

converges absolutely. 

Proof. Following the proof of Theorem 4.5, we write \( u = \sum_{\tau_1} u_{\tau_1}, \; v = \sum_{\tau_2} v_{\tau_2} \), where \( u_{\tau_1} = P_{\tau_1} \) and \( v_{\tau_2} = P_{\tau_2} \). Applying Proposition 4.4 to each matrix coefficient \( \langle U^s(a_t)u_{\tau_1}, v_{\tau_2} \rangle_{U(v,s)} \) gives

\[
\langle U^s(a_t)u, v \rangle_{U(v,s)} = e^{(s-d)t} \left( \sum_{\tau_1, \tau_2} \langle T^{T_2}_{\tau_1}C_+(s) P_{\tau_1} u_{\tau_1}, P_{\tau_2} v_{\tau_2} \rangle_{U(v,s)} \right) + O_s(e^{(s-d-\eta)t}) \left( \sum_{\tau_1} \|u_{\tau_1}\|_{S^m(v,s)} \right) \left( \sum_{\tau_2} \|v_{\tau_2}\|_{S^m(v,s)} \right)
\]

Using Lemma 3.1 as in the proof of Theorem 4.5, we get

\[
\left( \sum_{\tau_1} \|u_{\tau_1}\|_{S^m(v,s)} \right) \left( \sum_{\tau_2} \|v_{\tau_2}\|_{S^m(v,s)} \right) \ll \|u\|_{S^m(v,s)} \|v\|_{S^m(v,s)}
\]

for some \( m \in \mathbb{N} \). It therefore just remains to prove that the sum

\[
\sum_{\tau_1, \tau_2} \langle T^{T_2}_{\tau_1}C_+(s) u_{\tau_1}, v_{\tau_2} \rangle_{U(v,s)}
\]

converges absolutely. Looking at an individual summand, we have

\[
|\langle T^{T_2}_{\tau_1}C_+(s) u_{\tau_1}, v_{\tau_2} \rangle_{U(v,s)}| = a(v, s, \tau_2) |\langle T^{T_2}_{\tau_1}C_+(s) u_{\tau_1}, v_{\tau_2} \rangle_{K}| \leq a(v, s, \tau_2) \|T^{T_2}_{\tau_1} \|_K \|C_+(s) u_{\tau_1} \|_K \|v_{\tau_2} \|_K
\]

Since \( U^s K \) is unitary on \( L^2(K) \), from the definition of \( C_+(s) \) (see Theorem 4.2), \( \|C_+(s) u_{\tau_1} \|_K \ll_s \|u_{\tau_1} \|_K \), giving

\[
|\langle T^{T_2}_{\tau_1}C_+(s) u_{\tau_1}, v_{\tau_2} \rangle_{U(v,s)}| \ll_s a(v, s, \tau_2) \|T^{T_2}_{\tau_1} \|_K \|u_{\tau_1} \|_K \|v_{\tau_2} \|_K \leq \sqrt{\frac{a(v, s, \tau_2)}{a(v, s, \tau_1)}} \|T^{T_2}_{\tau_1} \|_K \|u_{\tau_1} \|_{U(v,s)} \|v_{\tau_2} \|_{U(v,s)}
\]

Using the bound 4.4 and Lemma 3.1 then gives the desired convergence of the sum.
Lemma 4.7. Let \( \mathbf{v} \) be a \( K \)-invariant vector in \( \mathcal{U}(v, s) \). Then for all \( t \geq 0 \),
\[
|\langle U^s(a_t)\mathbf{v}, \mathbf{v} \rangle_{\mathcal{U}(v, s)}| \ll_s e^{(s-d)t} \|\mathbf{v}\|_{\mathcal{U}(v, s)}^2,
\]
where the implied constant is uniformly bounded over \( s \) in compact subsets of \((\frac{d}{2}, d)\).

Proof. \( \mathcal{U}(v, s) \) has a non-zero \( K \)-invariant unit vector (that is moreover unique) if and only if \( \mathbf{v} \) is the trivial representation of \( M \). In the parameterization from Section 3.1, the relevant \( \nu \) is the trivial representation of \( M \). As a first step, we prove the following bound on matrix coefficients of \( \mathcal{L} \) functions in a Sobolev space of sufficiently high order. As a first step, we
\[
|\langle \mathcal{L}_\mathbf{m}, \mathbf{v} \rangle_{\mathcal{U}(v, s)}| \ll \|\mathbf{v}\|_{\mathcal{U}(v, s)}^2.
\]

5. Leading term for Sobolev functions

In this section, we will extend Roblin’s result [28, Theorem 3.4] on mixing of the geodesic flow for continuous functions with compact support to general functions in a Sobolev space of sufficiently high order. As a first step, we prove the following bound on matrix coefficients of \( L^2(\Gamma \backslash G) \):

Lemma 5.1. There exists \( m \in \mathbb{N} \) such that for all \( f_1, f_2 \in S^m(\Gamma \backslash G) \) and \( t \geq 0 \),
\[
|\langle \mathbf{v}, e_t \mathbf{v} \rangle| \ll e^{(\delta-d)t} S^m(f_1) S^m(f_2).
\]

Proof. We start by assuming that \( f_1 \) and \( f_2 \) are \( \rho(K) \)-invariant. The decomposition of the functions according to \( L^2(\Gamma \backslash G)_{\text{ap}} = B_\delta \oplus \mathcal{W} \) reads
\[
f_i = \langle f_i, \phi_0 \rangle \phi_0 + f'_i, \quad i = 1, 2,
\]
where \( f'_i \in \mathcal{W} \). As a consequence of Theorem 2.1, \( \mathcal{W} \) does not weakly contain any complementary series representation \( \mathcal{U}(v, s) \) with \( s > s_1 \). This enables us to use e.g. [30, Theorem 2.1, 2], to get that for any \( \varepsilon > 0 \), for any \( f_1, f_2 \in L^2(\Gamma \backslash G)^K \) and \( t > 0 \),
\[
|\langle \rho(a_t) f'_1, f'_2 \rangle| \ll e^{(s_1-d+\varepsilon)t} \|f'_1\| \|f'_2\|.
\]

By Lemma 4.7,
\[
|\langle \rho(a_t) \phi_0, \phi_0 \rangle| \ll e^{(\delta-d)t},
\]
so choosing \( 0 < \varepsilon < \delta - s_1 \) gives
\[
|\langle \rho(a_t) \phi_1, \phi_2 \rangle| \ll e^{(\delta-d)t} (\|f_1\| \|f_2\| + \|f'_1\| \|f'_2\|) \leq e^{(\delta-d)t} \|f_1\| \|f_2\|.
\]
We then use [30, Proposition 2.5] to extend this bound to all $K$-finite functions $f_1, f_2$ in $L^2(\Gamma \backslash G)$:
\[
|\langle \rho(a) f_1, f_2 \rangle| \leq e^{(d-d)t} \|f_1\| \|f_2\| \sqrt{\dim(\rho(K)f_1)} \dim(\rho(K)f_2).
\]
To pass to Sobolev functions, we first observe that, by e.g. [16, p. 206 (1)], \(\dim(\rho(K)\mathcal{P}_f) \leq \dim(\tau)^3\) for all \(f \in L^2(\Gamma \backslash G)\). Then for all \(f_1, f_2 \in \mathcal{S}^m(\Gamma \backslash G)\),
\[
|\langle \rho(a) f_1, f_2 \rangle| \leq e^{(d-d)t} \sum_{\tau_1, \tau_2 \in \mathcal{K}} \dim(\tau_1) \dim(\tau_2) \|\mathcal{P}_{\tau_1} f_1\| \|\mathcal{P}_{\tau_2} f_2\|
\]
\[
\leq e^{(d-d)t} \|f_1\|_{\mathcal{S}^m(\Gamma \backslash G)} \|f_2\|_{\mathcal{S}^m(\Gamma \backslash G)}
\]
for \(m \in \mathbb{N}\) large enough, by Lemma 3.1.

Proposition 5.2. For simplicity, we write \(\|f\|_{\mathcal{S}^m} = \|f\|_{\mathcal{S}^m(\Gamma \backslash G)}\). There exists \(m \in \mathbb{N}\) such that for all \(M\)-invariant \(f_1, f_2 \in \mathcal{S}^m(\Gamma \backslash G)\),
\[
\lim_{t \to \infty} e^{(d-d)t} \langle \rho(a) f_1, f_2 \rangle = m^{BR}(f_1)m^{BR}(f_2).
\]

Proof. Using the density of \(C^\infty_\mathcal{E}(\Gamma \backslash G)^M\) in \(\mathcal{S}^m(\Gamma \backslash G)^M\), given \(\varepsilon > 0\), there exist \(f_1^\varepsilon, f_2^\varepsilon \in C^\infty_\mathcal{E}(\Gamma \backslash G)\) such that
\[
\|f_i - f_i^\varepsilon\|_{\mathcal{S}^m} \leq \varepsilon \quad i = 1, 2.
\]
We then write, using Lemma 5.1
\[
e^{(d-d)t} \langle \rho(a) f_1, f_2 \rangle
\]=
\[
e^{(d-d)t} \left( \langle \rho(a) f_1^\varepsilon, f_2^\varepsilon \rangle + \langle \rho(a)(f_1 - f_1^\varepsilon), f_2 \rangle + \langle \rho(a)f_1^\varepsilon, f_2 - f_2^\varepsilon \rangle \right)
\]=
\[
e^{(d-d)t} \langle \rho(a) f_1^\varepsilon, f_2^\varepsilon \rangle + O(\varepsilon(\|f_1\|_{\mathcal{S}^m} + \|f_2\|_{\mathcal{S}^m})).
\]
From [28, Theorem 3.4], we have
\[
\lim_{t \to \infty} e^{(d-d)t} \langle \rho(a) f_1^\varepsilon, f_2^\varepsilon \rangle = m^{BR}(f_1)m^{BR}(f_2).
\]
So
\[
\limsup_{t \to \infty} e^{(d-d)t} \langle \rho(a) f_1, f_2 \rangle = m^{BR}(f_1)m^{BR}(f_2) + O(\varepsilon(\|f_1\|_{\mathcal{S}^m} + \|f_2\|_{\mathcal{S}^m})).
\]
By Lemma 2.2
\[
m^{BR}(f_1)m^{BR}(f_2) = m^{BR}(f_1)m^{BR}(f_2) + O(\varepsilon(\|f_1\|_{\mathcal{S}^m} + \|f_2\|_{\mathcal{S}^m})),
\]
giving
\[
\limsup_{t \to \infty} e^{(d-d)t} \langle \rho(a) f_1, f_2 \rangle = m^{BR}(f_1)m^{BR}(f_2) + O(\varepsilon(\|f_1\|_{\mathcal{S}^m} + \|f_2\|_{\mathcal{S}^m})).
\]
Since \(\varepsilon > 0\) was arbitrary, we in fact have
\[
\limsup_{t \to \infty} e^{(d-d)t} \langle \rho(a) f_1, f_2 \rangle = m^{BR}(f_1)m^{BR}(f_2).
\]
An analogous calculation with \(\liminf\) in place of \(\limsup\) proves the proposition.\qed
6. Proof of Theorem 1.1

Proof of Theorem 1.1. Recalling the isomorphism
\[
(\rho, L^2(\Gamma \setminus G)) \cong \int_Z (\pi_\zeta, \mathcal{H}_\zeta) \, d\mu_Z(\zeta),
\]
given \(0 < r < \delta - s_1\), we partition \(Z\) as
\[
Z = Z^- \cup Z^+,
\]
where
\[
Z^- = \{ \zeta \in Z : (\pi_\zeta, \mathcal{H}_\zeta) \cong \mathcal{U}(v, s), \text{ where } v \in \mathcal{M} \text{ and } s \in [s_1 + r, \delta] \}\]
and \(Z^+ = Z \setminus Z^-\) (cf. (2.2)). This partition is then used to decompose \((\rho, L^2(\Gamma \setminus G))\) as
\[
(\rho, L^2(\Gamma \setminus G)) = (\rho, L^2(\Gamma \setminus G)^-) \oplus (\rho, L^2(\Gamma \setminus G)^+),
\]
where (with the dependency on \(r\) being slightly suppressed)
\[
(\rho, L^2(\Gamma \setminus G)^\pm) \cong \int_{Z^\pm} (\pi_\zeta, \mathcal{H}_\zeta) \, d\mu_Z(\zeta).
\]
Observe that \(\mathcal{B}_\delta\) occurs as a subrepresentation of \((\rho, L^2(\Gamma \setminus G)^-)\). We may further decompose \((\rho, L^2(\Gamma \setminus G)^-)\) accordingly: let \(L^2(\Gamma \setminus G)_0\) be the orthogonal complement of \(\mathcal{B}_\delta\) in \(L^2(\Gamma \setminus G)^-\). We thus have
\[
(\rho, L^2(\Gamma \setminus G)) = (\rho, \mathcal{B}_\delta) \oplus (\rho, L^2(\Gamma \setminus G)_0) \oplus (\rho, L^2(\Gamma \setminus G)^+).
\]
Note that Theorem 2.1 and the duality between eigenfunctions of the Laplacian on \(\mathcal{M} = \Gamma \setminus G/K\) and spherical representations in the decomposition of \(L^2(\Gamma \setminus G)\) imply that no spherical representation is weakly contained in \((\rho, L^2(\Gamma \setminus G)_0)\). At the level of functions, we write
\[
f_i = f_i^0 + f_i^- + f_i^+, \quad i = 1, 2
\]
\[(f_i^0 \in \mathcal{B}_\delta, f_i^- \in L^2(\Gamma \setminus G)_0, f_i^+ \in L^2(\Gamma \setminus G)^+).\]

The matrix coefficients we are interested in now decompose as
\[
\langle \rho(a_i) f_1, f_2 \rangle = \langle \rho(a_i) f_1^0, f_2^0 \rangle + \langle \rho(a_i) f_1^-, f_2^- \rangle + \langle \rho(a_i) f_1^+, f_2^+ \rangle. \tag{6.1}
\]
We deal with the three summands in turn: by construction, \((\rho, L^2(\Gamma \setminus G)^+)\) does not weakly contain any \(\mathcal{U}(v, s)\) with \(s > s_1 + r\) (this also uses the fact \((\rho, L^2(\Gamma \setminus G))\) does not weakly contain any \(\mathcal{U}(v, s)\) with \(s > \delta\); cf. [21, Proposition 3.23]).

We assume \(m\) is large enough so that so that all the results of the previous sections hold for \(f_1, f_2 \in S^m(\Gamma \setminus G)\). By [21, Proposition 3.29] (or [30, Theorem 2.1] combined with the argument from the proof of Lemma 5.1),
\[
|\langle \rho(a_i) f_1^+, f_2^+ \rangle| \leq e^{(s_1 + r + \xi - \delta)t} \|f_1^+\|_{S^m(\Gamma \setminus G)} \|f_2^+\|_{S^m(\Gamma \setminus G)}
\]
\[
\leq e^{(s_1 + r + \xi - \delta)t} \|f_1\|_{S^m(\Gamma \setminus G)} \|f_2\|_{S^m(\Gamma \setminus G)} \tag{6.2}
\]
for all \(\xi > 0\).
We will now use Theorem 4.5 to bound \( \langle \rho(a_t) f_1^-, f_2^- \rangle \). Let \( \mathbb{Z}_r^- \subset \mathbb{Z}_r^+ \) be such that \( (\rho, L^2(\Gamma \setminus G))_s^\sim = \int_{\mathbb{Z}_r^-} (\pi_\zeta, \mathcal{H}_\zeta) d\mu(z) \). The corresponding decomposition of the functions \( f_1^-, f_2^- \) reads \( f_i^- = \int_{\mathbb{Z}_r^-} (f_i^-)_z d\mu_\zeta(z) \) \( (i = 1, 2) \). Since \( f_i \) is \( \mathcal{M} \)-invariant, so is \( f_i^- \) and \( \mu_\zeta \)-a. e. \( (f_i^-)_\zeta \). The matrix coefficient \( \langle \rho(a_t) f_1^-, f_2^- \rangle \) is now written as

\[
\langle \rho(a_t) f_1^-, f_2^- \rangle = \int_{\mathbb{Z}_r^-} \langle \pi_\zeta(a_t)(f_1^-)_\zeta, (f_2^-)_\zeta \rangle \mathcal{H}_{\zeta} d\mu_\zeta(z).
\]

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

\[
\lambda(\delta, r) = \max_{s \in [s_1 + r, \delta]} s - d - \eta_s = \max_{s \in [s_1 + r, \delta]} s - d - \min\{1, 2s - d\},
\]

we apply Theorem 4.5 to each \( \langle \pi_\zeta(a_t)(f_1^-)_\zeta, (f_2^-)_\zeta \rangle \mathcal{H}_{\zeta} \) and obtain

\[
|\langle \rho(a_t) f_1^-, f_2^- \rangle| \leq e^{\lambda(\delta, r)t} \int_{\mathbb{Z}_r^-} \| (f_1^-)_z \|_{\mathcal{H}_{\zeta}} \| (f_2^-)_z \|_{\mathcal{H}_{\zeta}} d\mu_\zeta(z)
\]

\[
\leq e^{\lambda(\delta, r)t} \sqrt{\int_{\mathbb{Z}_r^-} \| (f_1^-)_z \|^2_{\mathcal{H}_{\zeta}} d\mu_\zeta(z)} \sqrt{\int_{\mathbb{Z}_r^-} \| (f_2^-)_z \|^2_{\mathcal{H}_{\zeta}} d\mu_\zeta(z)}
\]

\[
= e^{\lambda(\delta, r)t} \| f_1^- \|_{S^m(\Gamma \setminus G)} \| f_2^- \|_{S^m(\Gamma \setminus G)}
\]

\[
\leq e^{\lambda(\delta, r)t} \| f_1 \|_{S^m(\Gamma \setminus G)} \| f_2 \|_{S^m(\Gamma \setminus G)} \leq e^{\lambda(\delta, r)t} \| f_1 \|_{S^m(\Gamma \setminus G)} \| f_2 \|_{S^m(\Gamma \setminus G)},
\]

where the implied constant remains bounded as \( r \to 0 \) if \( s_0 > \frac{d}{2} \) but is unbounded otherwise. 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 4.6 gives

\[
\langle \rho(a_t) f_1^0, f_2^0 \rangle = \Phi(f_1^0, f_2^0) \ e^{(\delta - d)t} + O_\delta \left( e^{(\delta - d- \eta_s)t} \| f_1^0 \|_{S^m(\Gamma \setminus G)} \| f_2^0 \|_{S^m(\Gamma \setminus G)} \right)
\]

\[
= \Phi(f_1^0, f_2^0) \ e^{(\delta - d)t} + O_\delta \left( e^{(\delta - d- \eta_s)t} \| f_1^0 \|_{S^m(\Gamma \setminus G)} \| f_2^0 \|_{S^m(\Gamma \setminus G)} \right),
\]

where \( \Phi(f_1^0, f_2^0) \) is given by the sum (4.5) under the aforementioned isomorphism.
Recall $\eta_s$ from (4.2), and note
\[
\beta := \min \left\{ 1, \eta, \delta - s_1 - r - \xi, \delta - d - \lambda(\delta, r) \right\}
\]
\[
= \min \left\{ 1, 2\delta - d, \delta - s_1 - r - \xi, \min_{s \in [s_1 + r, \delta]} \left( \delta - s + \min\{1, 2s - d\} \right) \right\}
\]
\[
\geq \min \left\{ 1, 2\delta - d, \delta - s_1 - r, \min_{s \in [s_1 + r, \delta]} \left( \delta - s + \min\{1, 2s - d\} \right) - \xi \right\}
\]
\[
= \min \{1, 2\delta - d, \delta - s_1 - r, \delta + s_1 + r - d\} - \xi - \xi
\]
\[
= \min \{1, \delta - s_1 - r\} - \xi
\]
\[
\geq \min \{1, \delta - s_1 - r\} - (\xi + r).
\]

Entering (6.2), (6.3), and (6.4) into (6.1) gives
\[
e^{(d-\delta)t}(\rho(a_t)f_1, f_2) = \Phi(f_1^0, f_2^0) + O_{\varepsilon, \xi}(e^{-\beta t}\|f_1\|_{S^m(\Gamma \setminus G)}\|f_2\|_{S^m(\Gamma \setminus G)}).
\]

Writing $\varepsilon = \xi + r$, we thus have
\[
e^{(d-\delta)t}(\rho(a_t)f_1, f_2) = \Phi(f_1^0, f_2^0) + O_{\varepsilon}(e^{-(\eta-\varepsilon)t}\|f_1\|_{S^m(\Gamma \setminus G)}\|f_2\|_{S^m(\Gamma \setminus G)}).
\]

Choosing $0 < \varepsilon < \eta$ gives
\[
\lim_{t \to \infty} e^{(d-\delta)t}(\rho(a_t)f_1, f_2) = \Phi(f_1^0, f_2^0),
\]
so $\Phi(f_1^0, f_2^0) = m^{BR}(f_1)m^{BR}(f_2)$ by Proposition 5.2, completing the proof.

\[\square\]

References


Department of Mathematics, Yale University, New Haven, CT 06520

Email address: samuel.edwards@yale.edu

Department of Mathematics, Yale University, New Haven, CT 06520 and Korea Institute for advanced study, Seoul, Korea

Email address: hee.oh@yale.edu