Fully-Connected Network on Noncompact Symmetric Space and Ridgelet Transform based on Helgason-Fourier Analysis

Sho Sonoda [1] **Isao Ishikawa** [2 1] **Masahiro Ikeda** [1]

Abstract

Neural network on Riemannian symmetric space such as hyperbolic space and the manifold of symmetric positive definite (SPD) matrices is an emerging subject of research in geometric deep learning. Based on the well-established framework of the Helgason-Fourier transform on the noncompact symmetric space, we present a fully-connected network and its associated ridgelet transform on the noncompact symmetric space, covering the hyperbolic neural network (HNN) and the SPDNet as special cases. The ridgelet transform is an analysis operator of a depth-2 continuous network spanned by neurons, namely, it maps an arbitrary given function to the weights of a network. Thanks to the coordinate-free reformulation, the role of nonlinear activation functions is revealed to be a wavelet function. Moreover, the reconstruction formula is applied to present a constructive proof of the universality of finite networks on symmetric spaces.

1. Introduction

Geometric deep learning is an emerging research direction that aims to devise neural networks on non-Euclidean spaces (Bronstein et al., 2021). In this study, we focus on devising a fully-connected layer on a noncompact symmetric space $X = G/K$ (Helgason, 1984; 2008). In general, it is more challenging to devise a fully-connected layer on a manifold than to devise a convolution layer because neither the scalar product, bias translation, nor pointwise activation can be trivially defined. A noncompact symmetric space is a Riemannian manifold X with nonpositive curvature, as well as a homogeneous space G/K of Lie groups G and K. It covers several important spaces in the recent literature of representation learning, such as the hyperbolic space and the manifold of symmetric positive definite (SPD) matrices, or the SPD manifold. On those spaces, several neural networks have been developed such as *hyperbolic neural networks (HNNs)* and *SPDNets*.

Neural Network on Hyperbolic Space. The hyperbolic space is a symmetric space with a constant negative curvature. Following the success of Poincaré embedding (Krioukov et al., 2010; Nickel & Kiela, 2017; 2018; Sala et al., 2018), the hyperbolic space has been recognized as an effective space for embedding tree-structured data; and hyperbolic neural networks (HNNs) (Ganea et al., 2018; Gulcehre et al., 2019; Shimizu et al., 2021) have been developed to promote effective use of hyperbolic geometry for saving parameters against Euclidean counterparts. The previous studies such as HNN (Ganea et al., 2018) and HNN++ (Shimizu et al., 2021) have replaced each operation with gyrovector calculus, but there are still rooms for arguments such as on the expressive power of the proposed network and on the role of nonlinear activation functions.

Neural Network on SPD Manifold. The SPD manifold equipped with the standard Riemannian metric has nonconstant nor nonpositive curvature. The metric is isomorphic to the Fisher information metric for multivariate centered normal distributions. Since covariance matrices are positive definite, the SPD manifold has been investigated and applied in a longer and wider literature than the hyperbolic space. Besides, the SPD manifold has also attracted attention as a space for graph embedding (Lopez et al., 2021; Cruceru et al., 2021). To reduce the computational cost without harming the Riemannian geometry, several distances have been proposed such as the affine-invariant Riemannian metric (AIRM) (Pennec et al., 2006), the Stein metric (Sra, 2012), the Bures–Wasserstein metric (Bhatia et al., 2019), the Log-Euclidean metric (Arsigny et al., 2006; 2007), and the vector-valued distance (Lopez et al., 2021). Furthermore, neural networks on SPD manifolds have been developed, such as SPDNet (Huang & Gool, 2017; Dong et al., 2017; Gao et al., 2019; Brooks et al., 2019b;a), deep manifold-to-manifold transforming network (DMT-Net) (Zhang et al., 2018), and ManifoldNet (Chakraborty et al., 2018; 2022).

[1] RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan [2] Ehime University, Ehime, Japan. Correspondence to: Sho Sonoda <sho.sonoda@riken.jp>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

Although those networks are aware of underlying geometry, except for a universality result on *horospherical* HNNs by Wang (2021), previous studies lack theoretical investigations, such as on the expressive power and on the effect of nonlinear activation functions. The purpose of this study is to define a fully-connected layer on a noncompact symmetric space in a unified manner from the perspective of harmonic analysis on symmetric space, and derive an associated *ridgelet transform*—an analysis operator that maps a function f on X to the weight parameters, written γ, of a network. In the end, the ridgelet transform is given as a *closed-form expression*, the reconstruction formula further elicits a constructive proof of the *universality* of finite models, and the role/effect of an activation function will be understood as a wavelet function.

Harmonic Analysis on Symmetric Space. The Helgason-Fourier transform has been introduced in (Helgason, 1965) as a Fourier transform on the noncompact symmetric space X. This is an integral transform of functions f on X with respect to the eigenfunctions of the Laplace-Beltrami operator Δ_X on X. We refer to Helgason (1984, Introduction) and Helgason (2008, Ch.III) for more details.

The Integral Representation $S[\gamma](\boldsymbol{x})$ on Euclidean Space is an infinite-dimensional linear representation of a depth-2 fully-connected neural network, given by the following integral operator: For every $\boldsymbol{x} \in \mathbb{R}^m$,

$$S[\gamma](\boldsymbol{x}) = \int_{\mathbb{R}^m \times \mathbb{R}} \gamma(\boldsymbol{a}, b)\sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)\mathrm{d}\boldsymbol{a}\mathrm{d}b. \tag{1}$$

Here, each function $\boldsymbol{x} \mapsto \sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)$ represents a single neuron, or a feature map of input $\boldsymbol{x}$ parametrized by $(\boldsymbol{a}, b)$. The integration over $(\boldsymbol{a}, b)$ implies that all the possible neurons are assigned in advance, and thus $S[\gamma]$ can be understood as a *continuous neural network*. We note, however, that if we take γ to be a finite sum of Dirac's measures such as $\gamma_p := \sum_{i=1}^p c_i \delta_{(\boldsymbol{a}_i, b_i)}$, then the integral representation can also exactly reproduce a finite model: For every $\boldsymbol{x} \in \mathbb{R}^m$,

$$S[\gamma_p](\boldsymbol{x}) = \sum_{i=1}^p c_i \sigma(\boldsymbol{a}_i \cdot \boldsymbol{x} - b_i).$$

In summary, $S[\gamma]$ is a mathematical model of shallow neural networks with *any* width ranging from finite to infinite.

The Ridgelet Transform $R[f;\rho](\boldsymbol{a}, b)$ is a right inverse (or pseudo-inverse) operator of the integral representation operator S. For the Euclidean neural network given in (1), the ridgelet transform is given as a *closed-form expression*: For every $(\boldsymbol{a}, b) \in \mathbb{R}^m \times \mathbb{R}$,

$$R[f;\rho](\boldsymbol{a}, b) := \int_{\mathbb{R}^m} f(\boldsymbol{x})\overline{\rho(\boldsymbol{a} \cdot \boldsymbol{x} - b)}\mathrm{d}\boldsymbol{x}. \tag{2}$$

Here, $f : \mathbb{R}^m \to \mathbb{C}$ is a target function to be approximated, and $\rho : \mathbb{R} \to \mathbb{C}$ is an auxiliary function, called the *ridgelet function*. Under mild conditions, the reconstruction formula

$$S[R[f;\rho]] = (\!(\sigma, \rho)\!) f,$$

holds, where $(\!(\cdot, \cdot)\!)$ denote a scalar product of σ and ρ given by a weighted inner-product in the Fourier domain as

$$(\!(\sigma, \rho)\!) := (2\pi)^{m-1} \int_{\mathbb{R}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-m}\mathrm{d}\omega,$$

where $\cdot^\sharp$ denotes the Fourier transform in $b \in \mathbb{R}$. Therefore, as long as the product $(\!(\sigma, \rho)\!)$ is neither 0 nor ∞, we can normalize ρ to satisfy $(\!(\sigma, \rho)\!) = 1$ so that $S[R[f;\rho]] = f$.

In other words, R and S are analysis and synthesis operators, and thus play the same roles as the Fourier (F) and inverse Fourier (F^{-1}) transforms respectively, in the sense that the reconstruction formula $S[R[f;\rho]] = (\!(\sigma, \rho)\!) f$ corresponds to the Fourier inversion formula $F^{-1}[F[f]] = f$. In the meanwhile, different from the case of the Fourier transform, there are infinitely many different ρ's satisfying $(\!(\sigma, \rho)\!) = 1$. This means that R is not strictly an inverse operator to S, which is unique if it exists, but a right inverse operator, indicating that S has a nontrivial null space. Sonoda et al. (2021b) have revealed that the null space is spanned by the ridgelet transforms $R[\,\cdot\,;\rho_0]$ with degenerated ridgelet functions satisfying $(\!(\sigma, \rho_0)\!) = 0$. This means that any parameter distribution γ satisfying $S[\gamma] = f$ can always be represented as (not always single but) a linear combination of ridgelet transforms.

Despite the common belief that neural network parameters are a blackbox, the closed-form expression of ridgelet transform (2) clearly describes how the network parameters are organized, which is a clear advantage of the integral representation theory. Furthermore, the integral representation theory can deal with a wide range of activation functions without any modification, not only ReLU but all the tempered distribution $\mathcal{S}'(\mathbb{R})$ (see, e.g., Sonoda & Murata, 2017).

Relations between Continuous and Finite Models. (1) The relation between a general continuous model $\int \gamma(v)\sigma(v, x)\mathrm{d}v$ with a general feature map $x \mapsto \sigma(v, x)$ parametrized by v, and the finite model $\sum_{i=1}^p c_i \sigma(v_i, x)$ is well investigated in the *Maurey-Jones-Barron (MJB) theory*, claiming the L^p-density of finite models in the space of continuous models (see, e.g., Kainen et al., 2013). The density is fundamental to show that a certain property of finite models is preserved when the model is extended to a continuous model, so that we can concentrate on investigating the continuous model instead of the finite model. (2) In addition, Sonoda et al. (2021a) have shown that the parameter distribution of a finite model trained by regularized empirical risk minimization (RERM) converges to a certain unique ridgelet spectrum $R[f;\sigma_*]$ with special σ_* in an over-parametrized regime. This means that we can understand the parameters at local minima to be a finite approximation of the ridgelet transform, and thus we can investigate the ridgelet transform to study the minimizer of the learning problem.

Historical Overview. The idea of the integral representation first emerged in the 1990s to investigate the expressive power of infinitely-wide shallow neural networks (Irie & Miyake, 1988; Funahashi, 1989; Carroll & Dickinson, 1989; Ito, 1991; Barron, 1993), and the original ridgelet transform is discovered independently by Murata (1996), Candès (1998) and Rubin (1998). In the context of sparse signal processing, ridgelet analysis has been developed as a multidimensional counterpart of wavelet analysis (Donoho, 2002; Starck et al., 2010; Kutyniok & Labate, 2012; Kostadinova et al., 2014). In the context of deep learning theory, continuous models have been employed in the so-called *mean-field theory* to show the global convergence of the SGD training of shallow ReLU networks (Nitanda & Suzuki, 2017; Mei et al., 2018; Rotskoff & Vanden-Eijnden, 2018; Chizat & Bach, 2018; Sirignano & Spiliopoulos, 2020; Suzuki, 2020), and new *ridgelet transforms for ReLU networks* have been developed to investigate the expressive power of ReLU networks (Sonoda & Murata, 2017), and to establish the *representer theorem* for ReLU networks (Savarese et al., 2019; Ongie et al., 2020; Parhi & Nowak, 2020; Unser, 2019).

Contributions of This Study

One of the major shortcomings of conventional ridgelet analysis has been that the closed-form expression like (2) is known only for the case of Euclidean network: $\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)$. In this study, we explain a natural way to find the ridgelet transform via the Fourier expression, then obtain a series of new ridgelet transforms for noncompact symmetric space $X=G/K$ in a unified manner by replacing the Euclidean-Fourier transform with the Helgason-Fourier transform on noncompact symmetric space. The reconstruction formula $S[R[f]]=f$ can provide a constructive proof of the universal approximation property of finite neural networks on an arbitrary noncompact symmetric space. As far as we have noticed, Wang (2021) is the only author who shows the universality of HNNs. Following the classical arguments by Cybenko (1989), her proof is based on the Hahn-Banach theorem. As a result, it is simple but non-constructive. On the other hand, our results (1) are more informative because of the constructive nature, (2) cover a wider range of spaces, i.e., any noncompact symmetric space $X=G/K$, and (3) cover a wider range of nonlinear activation functions, i.e., any tempered distribution $\sigma\in\mathcal{S}'(\mathbb{R})$, without any modification.

Remarks for Avoiding Potential Confusions. As clarified in the discussion, the HNNs devised in this study is the same as the one investigated by Wang (2021), but different from the ones proposed by Ganea et al. (2018) and Shimizu et al. (2021). Both HNNs can be regarded as extensions of the Euclidean NN (ENN), but as a hyperbolic counterpart of the Euclidean hyperplane, Wang (2021) and we employed the *horosphere*, while Ganea et al. (2018) and Shimizu et al. (2021) employed the *set of geodesics*, called the *Poincaré hyperplane*. As a consequence, our main results *do* cover our *horospherical HNN*, but do not cover their *geodesic HNNs*. Yet, we conjecture that the proof technique can be applied for those geodesic HNNs. On the other hand, the SPDNet devised in this study is essentially the same as the ones proposed in previous studies (Huang & Gool, 2017; Dong et al., 2017; Gao et al., 2019; Brooks et al., 2019b;a).

We further remark that the so-called "equivalence between convolutional networks and fully-connected networks" (e.g. Petersen & Voigtlaender, 2020) can hold *only when* networks are carefully designed (e.g., when X is a finite set). While a convolution on X is a binary operation $f*g$ of functions $f,g:X\to\mathbb{R}$, a scalar-product on X is a binary operation $\langle x,y\rangle$ of points $x,y\in X$, and there are *no* canonical rules to identify functions and points in general. Moreover, there are *no* canonical scalar-products for points in general manifolds. Therefore, even if we are given a convolutional network on X, we cannot directly translate it as a fully-connected network.

Notations

$|\cdot|_E$ (or simply $|\cdot|$) denotes the Euclidean norm of $\mathbb{R}^m$.

$GL(m,\mathbb{R})$ denotes the set of $m\times m$ real regular matrices, or the general linear group. $O(m)$ denotes the set of $m\times m$ orthogonal matrices, or the orthogonal group. $D(m), D_+(m)$ and $D_{\pm 1}(m)$ denote the sets of $m\times m$ diagonal matrices with real entries, positive entries, and ± 1, respectively. $T_+(m), T_1(m)$ and $T_0(m)$ denote the sets of $m\times m$ upper triangular matrices with positive entries, ones, and zeros on the diagonal, respectively. $\mathbb{P}_m$ denotes the set of $m\times m$ symmetric positive definite matrices.

On a (possibly noncompact) manifold X, $C_c(X)$ denotes the compactly supported continuous functions, $C_c^\infty(X)$ denotes the compactly supported infinitely differentiable functions, and $L^2(X)$ denotes the square-integrable functions. When X is a symmetric space, $\mathrm{d}x$ is supposed to be the left-invariant measure.

For any integer $d>0$, $\mathcal{S}(\mathbb{R}^d)$ denotes the Schwartz test functions (or rapidly decreasing functions) on $\mathbb{R}^d$, and $\mathcal{S}'(\mathbb{R}^d)$ the tempered distributions on $\mathbb{R}^d$ (i.e., the topological dual of $\mathcal{S}(\mathbb{R}^d)$). We eventually set the class of activation functions to be tempered distributions $\mathcal{S}'(\mathbb{R})$, which covers

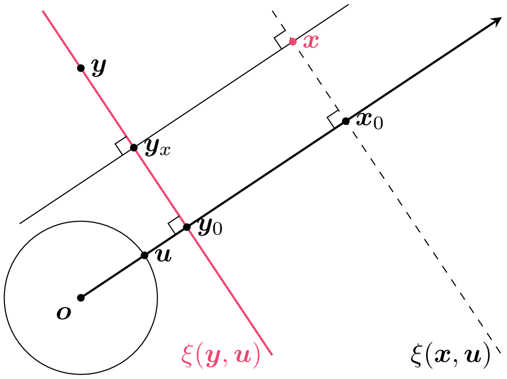

Figure 1: The Euclidean fully-connected layer $\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)$ is recast as the signed distance $d(\boldsymbol{x},\xi)$ from a point $\boldsymbol{x}$ to a hyperplane $\xi(\boldsymbol{y},\boldsymbol{u})$ followed by a wavelet function $\sigma(r\cdot)$, where $\boldsymbol{y}$ satisfies $r\boldsymbol{y}\cdot\boldsymbol{u}=b$ and $\xi(\boldsymbol{y},\boldsymbol{u})$ passes through the point $\boldsymbol{y}$ with a normal $\boldsymbol{u}$.

truncated power functions $\sigma(b)=b_+^k=\max\{b,0\}^k$ covering the step function for $k=0$ and the rectified linear unit (ReLU) for $k=1$. We refer to Grafakos (2008); Gel'fand & Shilov (1964); Sonoda & Murata (2017) for more details on Schwartz distributions and Fourier analysis on them.

To avoid potential confusion, we use two symbols $\widehat{\cdot}$ and $\cdot^\sharp$ for the Fourier transforms in the input variable $x\in X=G/K$ (or $\boldsymbol{x}\in\mathbb{R}^m$) and the bias variable $b\in\mathbb{R}$, respectively. For example, $\widehat{f}(\boldsymbol{\xi}):=\int_{\mathbb{R}^m}f(\boldsymbol{x})e^{-i\boldsymbol{x}\cdot\boldsymbol{\xi}}\mathrm{d}\boldsymbol{x}$ for $\boldsymbol{\xi}\in\mathbb{R}^m$, $\rho^\sharp(\omega):=\int_{\mathbb{R}}\rho(b)e^{-ib\omega}\mathrm{d}b$ for $\omega\in\mathbb{R}$, and $\gamma^\sharp(\boldsymbol{a},\omega)=\int_{\mathbb{R}}\gamma(\boldsymbol{a},b)e^{-ib\omega}\mathrm{d}b$ for $(\boldsymbol{a},\omega)\in\mathbb{R}^m\times\mathbb{R}$.

2. Fully-Connected Layer on Euclidean Space

We briefly review the Euclidean fully-connected layer $\boldsymbol{x}\mapsto\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)$ on $\mathbb{R}^m$, where $\boldsymbol{x}\in\mathbb{R}^m$ is an input vector, $(\boldsymbol{a},b)\in\mathbb{R}^m\times\mathbb{R}$ is hidden parameters, $\boldsymbol{a}\cdot\boldsymbol{x}$ is the Euclidean scalar product, and $\sigma:\mathbb{R}\to\mathbb{R}$ is an arbitrary given nonlinear function. In particular, expressions (3) and (4) are keys to devise fully connected layer on a symmetric space with a variety of activation function σ and to derive the associated ridgelet transform.

2.1. Coordinate-Free Reformulation of Euclidean Fully-Connected Layer

For any $(\boldsymbol{x},\boldsymbol{u})\in\mathbb{R}^m\times\mathbb{S}^{m-1}$, put

$$\xi(\boldsymbol{x},\boldsymbol{u}):=\{\boldsymbol{y}\in\mathbb{R}^m\mid\boldsymbol{u}\cdot(\boldsymbol{x}-\boldsymbol{y})=0\},$$

the hyperplane passing through a point $\boldsymbol{x}\in\mathbb{R}^m$ and orthogonal to a unit vector $\boldsymbol{u}\in\mathbb{S}^{m-1}$.

First, we change the parameters in polar coordinates as

$$(\boldsymbol{a},b)=(r\boldsymbol{u},r\boldsymbol{u}\cdot\boldsymbol{y}),\quad(r,\boldsymbol{u},\boldsymbol{y})\in\mathbb{R}_{\geqslant 0}\times\mathbb{S}^{m-1}\times\mathbb{R}^m.$$

We note that the mapping from $\boldsymbol{y}$ to b is not injective, but it is rather understood as the mapping from (any representative point of) hyperplane $\xi((b/r)\boldsymbol{u},\boldsymbol{u})=\{\boldsymbol{y}\mid r\boldsymbol{u}\cdot\boldsymbol{y}=b\}$ to b.

Then, the fully-connected layer $\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)$ is rewritten as

$$\begin{aligned}\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)&=\sigma(r\boldsymbol{u}\cdot(\boldsymbol{x}-\boldsymbol{y}))=\sigma(rd(\boldsymbol{x},\boldsymbol{y}_x))\\&=\sigma(rd(\boldsymbol{x},\xi(\boldsymbol{y},\boldsymbol{u}))),\end{aligned}\tag{3}$$

where $d(\boldsymbol{x},\boldsymbol{y}):=\operatorname{sign}(\boldsymbol{x}-\boldsymbol{y})|\boldsymbol{x}-\boldsymbol{y}|_E$ denotes the signed Euclidean distance, and $\boldsymbol{y}_x$ denotes the closest point to $\boldsymbol{x}$ in the hyperplane $\{\boldsymbol{y}\mid r\boldsymbol{u}\cdot\boldsymbol{y}=b\}$. Figure 1 illustrates the relations of symbols.

The last two expressions are coordinate-free, but the final expression (3) is much appreciated because $\boldsymbol{y}$ can be an arbitrary representative point of hyperplane $\{r\boldsymbol{u}\cdot\boldsymbol{y}=b\}$. Meanwhile, the scaled nonlinear function $\sigma(r\cdot)$ is understood as a wavelet function, which plays a role of multiscale analysis (Mallat, 2009) such as singularity detection with scale r running from 0 to ∞.

In summary, a fully-connected layer $\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)$ is recast as $\sigma(rd(\boldsymbol{x},\xi))$ "wavelet analysis with respect to a wavelet function σ on the signed distance $d(\boldsymbol{x},\xi)$ between point $\boldsymbol{x}$ and hyperplane ξ." Since the wavelet transform can detect a point singularity, wavelet analysis on the distance between a point and a hyperplane can detect a singularity in the normal direction to the hyperplane.

2.2. How to Solve $S[\gamma]=f$ and Find $\gamma=R[f]$

We explain the basic steps to find the parameter distribution γ satisfying $S[\gamma]=f$. The basic steps is threefold: (Step 1) Turn the network into the *Fourier expression*, (Step 2) *change variables* to split the feature map into useful and junk factors, and (Step 3) put the unknown γ the *separation-of-variables form* to find a particular solution.

The following procedure is valid, for example, when $\sigma\in\mathcal{S}'(\mathbb{R}),\rho\in\mathcal{S}(\mathbb{R}),f\in L^2(\mathbb{R}^m)$ and $\gamma\in L^2(\mathbb{R}^m\times\mathbb{R})$. See Kostadinova et al. (2014) and Sonoda & Murata (2017) for more details on the valid combinations of function classes.

Step 1. To begin with, we turn the network into the *Fourier expression* as below.

$$\begin{aligned}S[\gamma](\boldsymbol{x})&:=\int_{\mathbb{R}^m\times\mathbb{R}}\gamma(\boldsymbol{a},b)\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)\mathrm{d}\boldsymbol{a}\mathrm{d}b\\&=\int_{\mathbb{R}^m}[\gamma(\boldsymbol{a},\cdot)*_b\sigma](\boldsymbol{a}\cdot\boldsymbol{x})\mathrm{d}\boldsymbol{a}\\&=\frac{1}{2\pi}\int_{\mathbb{R}^m\times\mathbb{R}}\gamma^\sharp(\boldsymbol{a},\omega)\sigma^\sharp(\omega)e^{i\omega\boldsymbol{a}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{a}\mathrm{d}\omega.\end{aligned}$$

Here, $*_b$ denotes the convolution in b; and the last equation follows from the identity (i.e., the Fourier inversion formula) $\phi(b)=\frac{1}{2\pi}\int_{\mathbb{R}}\phi^\sharp(\omega)e^{i\omega b}\mathrm{d}\omega$ with $\phi(b)=[\gamma(\boldsymbol{a},\cdot)*_b\sigma](b)$ and $b=\boldsymbol{a}\cdot\boldsymbol{x}$.

Step 2. Next, we change variables $(\boldsymbol{a},\omega) = (\boldsymbol{\xi}/\omega, \omega)$ with $\mathrm{d}\boldsymbol{a}\mathrm{d}\omega = |\omega|^{-m}\mathrm{d}\boldsymbol{\xi}\mathrm{d}\omega$ so that the modified feature map $\sigma^\sharp(\omega)e^{i\omega\boldsymbol{a}\cdot\boldsymbol{x}}$ split into useful and junk factors as

$$= \frac{1}{2\pi}\int_{\mathbb{R}}\left[\int_{\mathbb{R}^m}\gamma^\sharp(\boldsymbol{\xi}/\omega,\omega)e^{i\boldsymbol{\xi}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{\xi}\right]\sigma^\sharp(\omega)|\omega|^{-m}\mathrm{d}\omega. \quad (4)$$

Step 3. Finally, since inside the bracket $[\cdots]$ is the Fourier inversion with respect to $\boldsymbol{\xi}$, it is natural to put γ to be a *separation-of-variables* expression

$$\gamma^\sharp_{f,\rho}(\boldsymbol{\xi}/\omega,\omega) := \widehat{f}(\boldsymbol{\xi})\overline{\rho^\sharp(\omega)}, \quad (5)$$

with the given function $f \in L^2(\mathbb{R}^m)$ and an arbitrary function $\rho \in \mathcal{S}(\mathbb{R})$. Then, we have

$$\begin{aligned} S[\gamma_{f,\rho}](\boldsymbol{x}) &= (\!(\sigma,\rho)\!)\frac{1}{(2\pi)^m}\int_{\mathbb{R}^m}\widehat{f}(\boldsymbol{\xi})e^{i\boldsymbol{\xi}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{\xi}\mathrm{d}\omega \\ &= (\!(\sigma,\rho)\!)f(\boldsymbol{x}), \end{aligned}$$

where we put

$$(\!(\sigma,\rho)\!) := (2\pi)^{m-1}\int_{\mathbb{R}}\sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-m}\mathrm{d}\omega.$$

In other words, the separation-of-variables expression $\gamma_{f,\rho}$ is a particular solution to the integral equation $S[\gamma] = cf$ with factor $c = (\!(\sigma,\rho)\!) \in \mathbb{C}$.

In the end, $\gamma_{f,\rho}$ turns out to be the ridgelet transform: The Fourier inversion of $\gamma^\sharp_{f,\rho}(\boldsymbol{a},\omega) = \widehat{f}(\omega\boldsymbol{a})\overline{\rho^\sharp(\omega)}$ is calculated as

$$\begin{aligned} \gamma_{f,\rho}(\boldsymbol{a},b) &= \frac{1}{2\pi}\int_{\mathbb{R}}\widehat{f}(\omega\boldsymbol{a})\overline{\rho^\sharp(\omega)e^{-i\omega b}}\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{\mathbb{R}^m\times\mathbb{R}} f(\boldsymbol{x})\overline{\rho^\sharp(\omega)e^{i\omega(\boldsymbol{a}\cdot\boldsymbol{x}-b)}}\mathrm{d}\omega \\ &= \int_{\mathbb{R}^m\times\mathbb{R}} f(\boldsymbol{x})\overline{\rho(\boldsymbol{a}\cdot\boldsymbol{x}-b)}\mathrm{d}\boldsymbol{x}, \end{aligned}$$

which is exactly the definition of the ridgelet transform $R[f;\rho]$.

In conclusion, the separation-of-variables expression (5) is the way to naturally find the ridgelet transform. We note that Steps 1 and 2 to obtain (4) can be understood as the *change-of-frame* from the neurons $\sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b)\mathrm{d}\boldsymbol{a}\mathrm{d}b$, which we are less familiar with, to the tensor product of a plane wave and a junk: $e^{i\boldsymbol{\xi}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{\xi} \otimes \sigma^\sharp(\omega)|\omega|^{-m}\mathrm{d}\omega$, which we are much familiar with. Hence, the map $\gamma(\boldsymbol{a},b) \mapsto \gamma^\sharp(\boldsymbol{\xi}/\omega,\omega)$ is understood to be the associated transformation of the coefficient γ.

3. Harmonic Analysis on Noncompact Symmetric Space

Readers may skip the first subsection, § 3.1, by understanding general notations in symmetric space X as specific ones in the hyperbolic space or the SPD manifold as listed in Table 1.

3.1. Noncompact Riemannian Symmetric Space

We follow the notation by Helgason (2008, Ch. II) except for the conflict cases. For example, we assign "$u \in \partial X$" instead of "$b \in B$" for the boundary of X, since b is assigned for the bias in a fully-connected layer in this study.

Let G be a connected semisimple Lie group with finite center, and let $G = KAN$ be its Iwasawa decomposition. Namely, it is a unique diffeomorphic decomposition of G into subgroups K, A, and N, where K is maximal compact, A is maximal abelian, and N is maximal nilpotent. For example, when $G = GL(m,\mathbb{R})$ (general linear group), then $K = O(m)$ (orthogonal group), $A = D_+(m)$ (all positive diagonal matrices), and $N = T_1(m)$ (all upper triangular matrices with ones on the diagonal).

Let $\mathrm{d}g, \mathrm{d}k, \mathrm{d}a$, and $\mathrm{d}n$ be left G-invariant measures on G, K, A, and N respectively. Following Helgason (1984, Proposition 5.1, Ch. I), we normalize the measures so that $\int_K \mathrm{d}k = 1$, and

$$\begin{aligned} \int_G f(g)\mathrm{d}g &= \int_{KAN} f(kan)e^{2\varrho\log a}\mathrm{d}k\mathrm{d}a\mathrm{d}n \\ &= \int_{NAK} f(nak)e^{-2\varrho\log a}\mathrm{d}n\mathrm{d}a\mathrm{d}k \\ &= \int_{ANK} f(ank)\mathrm{d}a\mathrm{d}n\mathrm{d}k, \end{aligned}$$

for any $f \in C_c(G)$, with a constant $\varrho \in \mathfrak{a}^*$ defined below.

Let $\mathfrak{g}, \mathfrak{k}, \mathfrak{a}$, and $\mathfrak{n}$ be the Lie algebras of G, K, A, and N respectively. By a fundamental property of abelian Lie algebra, both $\mathfrak{a}$ and its dual $\mathfrak{a}^*$ are the same dimensional vector spaces, and thus they can be identified with $\mathbb{R}^r$ for some r, namely $\mathfrak{a} = \mathfrak{a}^* = \mathbb{R}^r$. We call $r := \dim\mathfrak{a}$ the rank of X. For example, when $G = GL(m,\mathbb{R})$, then $\mathfrak{g} = \mathfrak{gl}_m = \mathbb{R}^{m\times m}$ (all $m \times m$ real matrices), $\mathfrak{k} = \mathfrak{o}_m$ (all skew-symmetric matrices), $\mathfrak{a} = D(m)$ (all diagonal matrices), and $\mathfrak{n} = T_0(m)$ (all strictly upper triangular matrices).

Let $X := G/K$ be a noncompact symmetric space, namely, a Riemannian manifold composed of all the left cosets

$$X := G/K := \{x = gK \mid g \in G\}.$$

Using the identity element e of G, let $o = eK$ be the origin of X. By the construction of X, group G acts transitively on X, and let $g[x] := ghK$ (for $x = hK$) denote the G-action of $g \in G$ on X. Specifically, any point $x \in X$ can always be written as $x = g[o]$ for some $g \in G$. Let $\mathrm{d}x$ denote the left G-invariant measure on X. Following the normalization above, we normalize $\mathrm{d}x$ so that $\int_X f(x)\mathrm{d}x := \int_G f(g[o])\mathrm{d}g = \int_{AN} f(an[o])\mathrm{d}a\mathrm{d}n$ for any $f \in C_c(X)$.

in symmetric space	in hyperbolic space	in SPD manifold
$X = G/K$	hyperbolic space $\mathbb{H}^m$	SPD manifold $\mathbb{P}_m$
$\partial X := K/M$	boundary (or ideal sphere) $\partial\mathbb{H}^m$	boundary $\partial\mathbb{P}_m$
$\mathfrak{a}^*$	frequency domain $\mathbb{R}^1$	frequency domain $\mathbb{R}^m$
$\xi(x,u)$	horosphere $\xi(x,u)$	horosphere $\xi(x,u)$
$\langle x,u\rangle := -H(g^{-1}k)$	signed distance $\langle x,u\rangle$	vector-valued distance $\langle x,u\rangle$

Table 1: Correspondence of notations in $X = G/K, \mathbb{H}^m$ and $\mathbb{P}_m$

Let $M := C_K(A) := \{k \in K \mid ka = ak \text{ for all } a \in A\}$ be the centralizer of A in K, and let

$$\partial X := K/M := \{u = kM \mid k \in K\}$$

be the boundary (or ideal sphere) of X, which is known to be a compact manifold. Let $\mathrm{d}u$ denote the uniform probability measure on ∂X. For example, when $K = O(m)$ and $A = D_+(m)$, then $M = D_{\pm 1}$ (the subgroup of K consisting of diagonal matrices with entries ± 1).

Let

$$\Xi := G/MN := \{\xi = gMN \mid g \in G\}$$

be the space of horospheres. Here, basic horospheres are: An N-orbit $\xi_o := N[o] = \{n[o] \mid n \in N\}$, which is a horosphere passing through the origin $x = o$ with normal $u = eM$; and $ka[\xi_o] = kaN[o]$, which is a horosphere through point $x = ka[o]$ with normal $u = kM$. In fact, any horosphere can be represented as $\xi(kan[o], kM)$ since $kaN = kanN$ for any $n \in N$. We refer to Helgason (2008, Ch.I, § 1) and Bartolucci et al. (2021, § 3.5) for more details on the horospheres and boudaries.

As a consequence of the Iwasawa decomposition, for any $g \in G$ there uniquely exists an r-dimensional vector $H(g) \in \mathfrak{a}$ satisfying $g \in Ke^{H(g)}N$. For any $(x,u) = (g[o], kM) \in X \times \partial X$, put

$$\langle x,u\rangle := -H(g^{-1}k) \in \mathfrak{a} \cong \mathbb{R}^r,$$

which is understood as the r-dimensional vector-valued distance, called the *composite distance*, from the origin $o \in X$ to the horosphere $\xi(x,u)$ through point x with normal u. Here, the vector-valued distance means that the ℓ^2-norm coincides with the Riemannian length, that is, $|\langle x,u\rangle| = |d(o,\xi(x,u))|$. We refer to Helgason (2008, Ch.II, § 1, 4) and Kapovich et al. (2017, § 2) for more details on the vector-valued composite distance.

Let $\Sigma \subset \mathfrak{a}^*$ be the set of (restricted) roots of $\mathfrak{g}$ with respect to $\mathfrak{a}$. For $\alpha \in \Sigma$, let $\mathfrak{g}_\alpha$ denote the corresponding root space, and call $m_\alpha := \dim(\mathfrak{g}_\alpha)$ the multiplicity of α. Let $\mathfrak{a}^+$ be the Weyl chamber corresponding to $\mathfrak{n}$, i.e. $\mathfrak{n} = \sum_{\alpha\in\Sigma^+}\mathfrak{g}_\alpha$, where Σ^+ is the set of $\alpha \in \Sigma$ that are positive on $\mathfrak{a}^+$. Put $\varrho := \sum_{\alpha\in\Sigma^+}\frac{m_\alpha}{2}\alpha \in \mathfrak{a}^*$. Let W be the Weyl group of G/K, and let $|W|$ denote its order. Let $\boldsymbol{c}(\lambda)$ be the Harish-Chandra $\boldsymbol{c}$-function for G. We refer to Helgason (1984, Theorem 6.14, Ch. IV) for the closed-form expression of the $\boldsymbol{c}$-function.

3.2. Helgason-Fourier Transform on Symmetric Space

For any function f on X, the Helgason-Fourier transform is defined as

$$\widehat{f}(\lambda,u) := \int_X f(x)e^{(-i\lambda+\varrho)\langle x,u\rangle}\mathrm{d}x, \ (\lambda,u) \in \mathfrak{a}^* \times \partial X$$

where the exponent $(-i\lambda+\varrho)\langle x,u\rangle$ is understood as the action of functional $-i\lambda+\varrho \in \mathfrak{a}^*$ on a vector $\langle x,u\rangle \in \mathfrak{a}$. The inversion formula (Helgason, 2008, Theorems 1.3 and 1.5, Ch. III) is given by

$$f(x) = \int_{\mathfrak{a}^*\times\partial X}\widehat{f}(\lambda,u)e^{(i\lambda+\varrho)\langle x,u\rangle}\frac{\mathrm{d}\lambda\mathrm{d}u}{|W||\boldsymbol{c}(\lambda)|^2}.$$

Here, the equality holds at every point $x \in X$ when $f \in C_c^\infty(X)$, and in L^2 when $f \in L^2(X)$. In particular, the following Plancherel theorem holds: For any $f_1, f_2 \in L^2(X)$, $\int_X f_1(x)\overline{f_2(x)}\mathrm{d}x = \int_{\mathfrak{a}^*\times\partial X}\widehat{f}_1(\lambda,u)\overline{\widehat{f}_2(\lambda,u)}\frac{\mathrm{d}\lambda\mathrm{d}u}{|W||\boldsymbol{c}(\lambda)|^2}$.

The integral kernel $e^{(-i\lambda+\varrho)\langle x,u\rangle}$ is an X-counterpart of a plane wave $e^{-i\lambda\boldsymbol{u}\cdot\boldsymbol{x}}$ in the Euclidean-Fourier transform $\int_{\mathbb{R}^m} f(\boldsymbol{x})e^{-i\lambda\boldsymbol{u}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{x}$ (expressed in polar coordinate). While the plane wave $e^{-i\lambda\boldsymbol{u}\cdot\boldsymbol{x}}$ is a joint eigenfunction of all the invariant differential operators (that is, all the polynomials of the Laplacian $\varDelta$) on $\mathbb{R}^m$, the X-plane wave $e^{(-i\lambda+\varrho)\langle x,u\rangle}$ is a joint eigenfunction of all the invariant differential operators (e.g., polynomials of the Laplace-Beltrami operator $\varDelta_X$) on X. In particular, the Plancherel measure $|\boldsymbol{c}(\lambda)|^{-2}\mathrm{d}\lambda\mathrm{d}u$ plays a parallel role to $\lambda^{-m}\mathrm{d}\lambda\mathrm{d}\boldsymbol{u}$ in polar coordinates.

3.3. Poincaré Ball Model of Hyperbolic Space

Here, we briefly introduce the Poincaré ball $\mathbb{B}^m$ as a Riemannian manifold. In Appendix A, we further explain the homogeneous space aspect of the Poincaré disk $\mathbb{B}^2$. In the following, the boldface such as $\boldsymbol{x}$ and $\boldsymbol{u}$ emphasizes that the symbols should be understood as the Cartesian coordinates, rather than a point itself on a manifold.

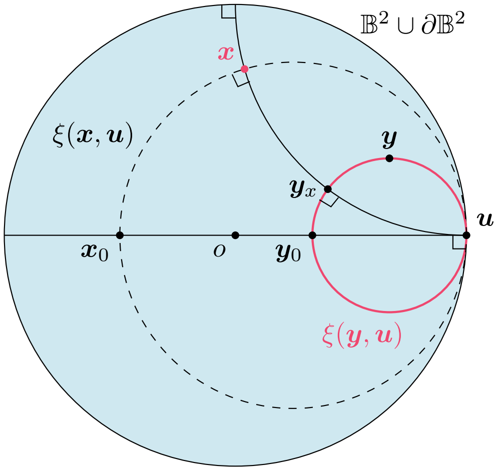
$\mathbb{B}^2 \cup \partial\mathbb{B}^2$
$\boldsymbol{x}$
$\xi(\boldsymbol{x}, \boldsymbol{u})$
$\boldsymbol{y}$
$\boldsymbol{y}_x$
$\boldsymbol{u}$
$\boldsymbol{x}_0$
o
$\boldsymbol{y}_0$
$\xi(\boldsymbol{y}, \boldsymbol{u})$

Figure 2: Poincaré disk $\mathbb{B}^2$, boundary $\partial\mathbb{B}^2$, point $\boldsymbol{x}$ (magenta), horocycle $\xi(\boldsymbol{y}, \boldsymbol{u})$ (magenta) through point $\boldsymbol{y}$ tangent to the boundary at $\boldsymbol{u}$, and two geodesics (solid black) orthogonal to the boundary at $\boldsymbol{u}$ through $\boldsymbol{o}$ and $\boldsymbol{x}$ respectively. The signed composite distance $\langle \boldsymbol{y}, \boldsymbol{u} \rangle$ from the origin $\boldsymbol{o}$ to the horocycle $\xi(\boldsymbol{y}, \boldsymbol{u})$ can be visualized as the Riemannian distance from $\boldsymbol{o}$ to point $\boldsymbol{y}_0$. Similarly, the distance between point $\boldsymbol{x}$ and horocycle $\xi(\boldsymbol{y}, \boldsymbol{u})$ is understood as the Riemannian distance between $\boldsymbol{x}$ and $\boldsymbol{y}_x$ along the geodesic, or equivalently, $\boldsymbol{x}_0$ and $\boldsymbol{y}_0$.

Let $\mathbb{B}^m := \{\boldsymbol{x} \in \mathbb{R}^m \mid |\boldsymbol{x}|_E < 1\}$ be a Riemannian manifold equipped with metric

$$\mathfrak{g}_{\boldsymbol{x}} := \left(\frac{2}{1 - |\boldsymbol{x}|_E^2} \right)^2 \sum_{i=1}^m \mathrm{d}x_i \wedge \mathrm{d}x_i, \quad \boldsymbol{x} \in \mathbb{B}^m.$$

This is the Poincaré ball model of m-dimensional hyperbolic space $\mathbb{H}^m$. The Riemannian distance between $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{B}^m$ is given by

$$d_P(\boldsymbol{x}, \boldsymbol{y}) = \cosh^{-1} \left(1 + \frac{2|\boldsymbol{x} - \boldsymbol{y}|_E^2}{(1 - |\boldsymbol{x}|_E^2)(1 - |\boldsymbol{y}|_E^2)} \right),$$

and the Riemannian volume measure at $\boldsymbol{x} \in \mathbb{B}^m$ is given by

$$\mathrm{d}\,\mathrm{vol}_{\mathfrak{g}}(\boldsymbol{x}) = \left(\frac{2}{1 - |\boldsymbol{x}|_E^2} \right)^m \mathrm{d}\boldsymbol{x},$$

with respect to the Lebesgue measure $\mathrm{d}\boldsymbol{x}$. Let $\partial\mathbb{B}^m := \{\boldsymbol{u} \in \mathbb{R}^m \mid |\boldsymbol{u}|_E = 1\} = \mathbb{S}^{m-1}$ be the boundary (or ideal sphere) equipped with the uniform spherical measure $\mathrm{d}\boldsymbol{u}$. Since $\mathbb{H}^m$ is rank-one, we identify $\mathfrak{a}^* \cong \mathbb{R}^1$ equipped with the Lebesgue measure.

In the Poincaré ball model $\mathbb{B}^m$, any boundary point $\boldsymbol{u}$ on the boundary $\partial\mathbb{B}^m$ is infinitely far from any inner point $\boldsymbol{x}$ in $\mathbb{B}^m$; any geodesic is a Euclidean arc that is orthogonal to the boundary $\partial\mathbb{B}^m$; any hyperbolic ball/sphere is a Euclidean ball/sphere in $\mathbb{B}^m$; and any horosphere is a Euclidean ball that is tangent to the boundary $\partial\mathbb{B}^m$. Hence a horosphere is understood as a "hyperbolic sphere of infinite radius", and it is identified by two parameters $(\boldsymbol{x}, \boldsymbol{u}) \in \mathbb{B}^m \times \partial\mathbb{B}^m$ as "a horosphere $\xi(\boldsymbol{x}, \boldsymbol{u})$ passing through $\boldsymbol{x}$ tangent to the boundary at $\boldsymbol{u}$." We note that since a hyperplane in the Euclidean space can also be understood as a "Euclidean sphere of infinite radius", we can understand horospheres as a hyperbolic counterpart of hyperplanes in the Euclidean space.

The signed composite distance $\langle \boldsymbol{x}, \boldsymbol{u} \rangle$ from the origin $\boldsymbol{o}$ to the horosphere $\xi(\boldsymbol{x}, \boldsymbol{u})$ is calculated as

$$\langle \boldsymbol{x}, \boldsymbol{u} \rangle := d_P(\boldsymbol{o}, \xi(\boldsymbol{x}, \boldsymbol{u})) = d_P(\boldsymbol{o}, \boldsymbol{x}_0) = \log \left(\frac{1 - |\boldsymbol{x}|_E^2}{|\boldsymbol{x} - \boldsymbol{u}|_E^2} \right).$$

Here, we put $\boldsymbol{x}_0 := t\boldsymbol{u}$ for some $|t| < 1$ so that $(\boldsymbol{x}_0 - \boldsymbol{x}, \boldsymbol{u} - \boldsymbol{x})_E = 0$, i.e., Thales' theorem.

The Helgason-Fourier transform and the inversion formula are instantiated as

$$\widehat{f}(\lambda, \boldsymbol{u}) = \int_{\mathbb{B}^m} f(\boldsymbol{x}) e^{(-i\lambda + \varrho)\langle \boldsymbol{x}, \boldsymbol{u} \rangle} \left(\frac{2}{1 - |\boldsymbol{x}|_E^2} \right)^m \mathrm{d}\boldsymbol{x},$$
$$f(\boldsymbol{x}) = \frac{c_m^2}{2} \int_{\mathbb{R} \times \mathbb{S}^{m-1}} \widehat{f}(\lambda, \boldsymbol{u}) e^{(i\lambda + \varrho)\langle \boldsymbol{x}, \boldsymbol{u} \rangle} \frac{\mathrm{d}\lambda \mathrm{d}\boldsymbol{u}}{|\boldsymbol{c}(\lambda)|^2},$$

for any $(\lambda, \boldsymbol{u}) \in \mathbb{R} \times \mathbb{S}^{m-1}$ and $\boldsymbol{x} \in \mathbb{B}^m$ respectively, where $c_m^2 = 2^{2\varrho}/(2\pi \,\mathrm{vol}(\mathbb{S}^{m-1}))$, $\varrho = (m-1)/2$, and the Plancherel measure $|\boldsymbol{c}(\lambda)|^{-2}$ is given by

$$(2^{k-1}(2k-1)!!)^{-2} \prod_{j=0}^{k-1} (\lambda^2 + j^2),$$

when $m = 2k + 1$, and

$$(2^{k-1}(2k-2)!!)^{-2} \frac{\pi\lambda \tanh(\pi\lambda)}{\lambda^2 + (1/2)^2} \prod_{j=0}^{k-1} \left(\lambda^2 + \left(\tfrac{2j-1}{2}\right)^2 \right),$$

when $m = 2k$.

4. Fully-Connected Layer on Symmetric Space

We define the fully-connected layer on the noncompact symmetric space, present the associated ridgelet transform and reconstruction formula, and finally state the cc-universality of finite networks.

4.1. Network Definition

In accordance with the geometric perspective, it is natural to define the network as below.

Definition 4.1. Let $\sigma : \mathbb{R} \to \mathbb{C}$ be a measurable function. For any function $\gamma : \mathfrak{a}^* \times \partial X \times \mathbb{R} \to \mathbb{C}$, the continuous neural network on the symmetric space X is given by

$$S[\gamma](x)$$

$$:= \int_{\mathfrak{a}^* \times \partial X \times \mathbb{R}} \gamma(a,u,b)\sigma(a\langle x,u\rangle - b)e^{\varrho\langle x,u\rangle}\mathrm{d}a\mathrm{d}u\mathrm{d}b.$$

Here, we call $x \in X$ the input, $a \in \mathfrak{a}^*$ the scale, $u \in \partial X$ the normal (of horosphere), and $b \in \mathbb{R}$ the bias. $\varrho \in \mathfrak{a}^*$ is a constant vector depending on G/K.

If we take $y \in X$ satisfying $a\langle y,u\rangle = b$, then we can rewrite $a\langle x,u\rangle - b$ as $ad(x, \xi(y,u))$, which can be understood as an X-counterpart of the coordinate-free expression (3). For technical reasons (i.e., for connecting the Helgason-Fourier transform), we impose an auxiliary weight $e^{\varrho\langle x,u\rangle}$.

4.2. Ridgelet Transform

Definition 4.2. Let $\rho : \mathbb{R} \to \mathbb{C}$ and $f : X \to \mathbb{C}$ be measurable functions. Put

$$R[f;\rho](a,u,b) := \int_X \boldsymbol{c}[f](x)\overline{\rho(a\langle x,u\rangle - b)}e^{\varrho\langle x,u\rangle}\mathrm{d}x,$$

$$\boldsymbol{c}[f](x) := \int_{\mathfrak{a}^* \times \partial X} \widehat{f}(\lambda,u)e^{(i\lambda+\varrho)\langle x,u\rangle}\frac{\mathrm{d}\lambda\mathrm{d}u}{|W||\boldsymbol{c}(\lambda)|^4},$$

$$(\!(\sigma,\rho)\!) := \frac{|W|}{2\pi}\int_{\mathbb{R}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-r}\mathrm{d}\omega.$$

Here $\boldsymbol{c}[f]$ is defined as a multiplier satisfying $\widehat{\boldsymbol{c}[f]}(\lambda,u) = \widehat{f}(\lambda,u)|\boldsymbol{c}(\lambda)|^{-2}$.

4.3. Reconstruction Formula

Theorem 4.3 (Reconstruction Formula on Symmetric Space)**.** *Let $X = G/K$ be a noncompact symmetric space defined as above. Let $\sigma \in \mathcal{S}'(\mathbb{R})$, $\rho \in \mathcal{S}(\mathbb{R})$. Then,*

$$\begin{aligned} &S[R[f;\rho]](x) \\ &= \int_{\mathfrak{a}^* \times \partial X \times \mathbb{R}} R[f;\rho](a,u,b)\sigma(a\langle x,u\rangle - b)e^{\varrho\langle x,u\rangle}\mathrm{d}a\mathrm{d}u\mathrm{d}b \\ &= (\!(\sigma,\rho)\!)f(x), \end{aligned}$$

where the equality holds at every point $x \in X$ when $f \in C_c^\infty(X)$, and in L^2 when $f \in L^2(X)$.

The proof is given in Appendix B.1, which is parallel to § 2.2.

As a result, while the Euclidean ridgelet transform is a scalar product of function $f(\boldsymbol{x})$ and co-feature map $\rho(\boldsymbol{a}\cdot\boldsymbol{x} - b)$, we revealed that the ridgelet transform on a symmetric space X is a scalar product of function f and co-feature map $\rho(a\langle x,u\rangle - b)$ *with* auxiliary weights $e^{\varrho\langle x,u\rangle}$ in the input data domain X and $|\boldsymbol{c}(\lambda)|^{-2}$ in the Fourier domain $\mathfrak{a}^* \times \partial X$. In geometric deep learning, it has been an open question how to naturally formulate the *affine map* $\boldsymbol{a}\cdot\boldsymbol{x} - b$ and *element-wise activation* σ for each point x on a manifold *without* depending on the specific choice of coordinates. From the perspective of harmonic analysis on the symmetric space, our answer is to embed the data $x \in X$ into the flat space $\mathfrak{a} = \mathbb{R}^r$ via the vector-valued composite distance $\langle x,u\rangle$.

4.4. cc-Universality

By discretizing the reconstruction formula $S[R[f;\rho]] = f$, we can construct a finite network f_n that approximates an arbitrary given function f. This is the primitive idea behind the constructive proof of the following cc-universality.

Let Δ_θ^n be a forward difference operator with difference $\theta > 0$, defined by $\Delta_\theta^1[\sigma](t) := \sigma(t+\theta) - \sigma(t)$ and $\Delta_\theta^{n+1}[\sigma](t) := \Delta_\theta^1 \circ \Delta_\theta^n[\sigma](t)$.

Theorem 4.4 (cc-universality of finite networks on symmetric space)**.** *Suppose that there exists $k \geqslant 0$ and $\theta > 0$ such that $\Delta_\theta^k[\sigma] \in L^\infty(\mathbb{R})$ and Lipschitz continuous. Then, the finite neural networks of the form*

$$f_n(x) = \sum_{i=1}^n c_i\sigma(a_i\langle x,u_i\rangle - b_i)e^{\varrho\langle x,u_i\rangle}, \quad x \in X$$

are cc-universal, that is, for any compact set $Z \subset X$, and continuous function $f \in C(Z)$, there exists a sequence of finite networks such that $\|f_n - f\|_{C(Z)} \to 0$ as $n \to \infty$.

The proof is given in Appendix B.2.

5. Examples: HNNs

We instantiate a continuous (horospherical) hyperbolic neural network (HNN) on the Poincaré ball model $\mathbb{B}^m$. In Appendix C, we further instantiate a continuous neural network on the SPD manifold $\mathbb{P}_m$ (SPDNet).

5.1. Continuous HNN

Definition 5.1. For any $\boldsymbol{x} \in \mathbb{B}^m$, put

$$\begin{aligned} &S[\gamma](\boldsymbol{x}) \\ &:= \int_{\mathbb{R}\times\mathbb{S}^{m-1}\times\mathbb{R}} \gamma(a,\boldsymbol{u},b)\sigma(a\langle\boldsymbol{x},\boldsymbol{u}\rangle - b)e^{\varrho\langle\boldsymbol{x},\boldsymbol{u}\rangle}\mathrm{d}a\mathrm{d}\boldsymbol{u}\mathrm{d}b, \end{aligned}$$

where $\langle\boldsymbol{x},\boldsymbol{u}\rangle = \log\frac{1-|\boldsymbol{x}|^2}{|\boldsymbol{x}-\boldsymbol{u}|^2}$ for any $(\boldsymbol{x},\boldsymbol{u}) \in \mathbb{B}^m \times \partial\mathbb{B}^m$.

We note that the weight function $\exp(\langle\boldsymbol{x},\boldsymbol{u}\rangle) = \frac{1-|\boldsymbol{x}|^2}{|\boldsymbol{x}-\boldsymbol{u}|^2}$ is known as the Poisson kernel.

Definition 5.2. For any $(a,\boldsymbol{u},b) \in \mathbb{R}\times\mathbb{S}^{m-1}\times\mathbb{R}$,

$$\begin{aligned} &R[f;\rho](a,\boldsymbol{u},b) \\ &= \int_{\mathbb{B}^m} \boldsymbol{c}[f](\boldsymbol{x})\overline{\rho(a\langle\boldsymbol{x},\boldsymbol{u}\rangle - b)}e^{\varrho\langle\boldsymbol{x},\boldsymbol{u}\rangle}\frac{2^m\mathrm{d}\boldsymbol{x}}{(1-|\boldsymbol{x}|^2)^m}, \end{aligned}$$

where for any $\boldsymbol{x} \in \mathbb{B}^m$,

$$\boldsymbol{c}[f](\boldsymbol{x}) = \int_{\mathbb{R}\times\mathbb{S}^{m-1}} \widehat{f}(\lambda,\boldsymbol{u})e^{(i\lambda+\varrho)\langle\boldsymbol{x},\boldsymbol{u}\rangle}\frac{\mathrm{d}\lambda\mathrm{d}\boldsymbol{u}}{|W||\boldsymbol{c}(\lambda)|^4}.$$

As a consequence of the general results, the following reconstruction formula holds.

Corollary 5.3. *For any $\sigma \in \mathcal{S}'(\mathbb{R}), \rho \in \mathcal{S}(\mathbb{R})$,*

$$S[R[f;\rho]](\boldsymbol{x}) = (\!(\sigma, \rho)\!) f(\boldsymbol{x}),$$

where

$$(\!(\sigma, \rho)\!) := \frac{1}{2\pi} \int_{\mathbb{R}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-1}\mathrm{d}\omega,$$

where the equality holds at every point $x \in \mathbb{B}^m$ when $f \in C_c^\infty(\mathbb{B}^m)$, and in L^2 when $f \in L^2(\mathbb{B}^m)$.

6. Discussion

We have devised the fully-connected layer on noncompact symmetric space $X = G/K$, and presented the closed-form expression of the ridgelet transform. The reconstruction formula $S[R[f]] = f$ is further applied to present a constructive proof of the cc-universality of finite fully-connected networks on X. This is the first universality result that covers a wide range of space X and activation functions σ, associated with a constructive proof in a unified manner. In fact, we do not need to restrict X to be the hyperbolic space or the SPD manifold, nor need to restrict σ to be ReLU.

Parallel to the Euclidean case explained in § 2.1, the fully-connected layer $\sigma(a\langle x, u\rangle - b)$ on X can also be understood as a wavelet function on a composite distance $d(x, \xi)$ from the point x to a horosphere ξ. To see this, we use the fact that a set $\xi(x, u) := \{y \in X \mid \langle x, u\rangle = \langle y, u\rangle\}$ is a horosphere through point x with normal u. Given a, b and u, put $\xi' := \{y \in X \mid a\langle y, u\rangle = b\}$. Then, for an arbitrary base point $y \in \xi'$, the horosphere $\xi(y, u)$ is a subset of ξ'. Following the notations in Figure 2, suppose $u = kM$, and let $x_0 \in \xi(x, u)$ and $y_0 \in \xi'$ be points satisfying $x = ka_x[o]$ and $y_0 = ka_\xi[o]$ for some $a_x, a_\xi \in A$ respectively. Then, $\langle x_0, u\rangle - \langle y_0, u\rangle = d(x_0, y_0)$, and thus we have

$$\sigma(a\langle x, u\rangle - b) = \sigma(ad(x_0, y_0)) = \sigma(ad(x, \xi(y_0, u))).$$

The ordinary wavelet transform can detect/localize a singularity at a point in a signal (see, e.g., Mallat, 2009), such as the singularity of signal $f(t) = 1/|t|$ at the origin $t = 0$. Hence, a wavelet on a distance $d(x, \xi)$ turns out to be a detector of a sigularity along a horosphere ξ.

Based on this coordinate-free reformulation, given a family Ξ of geometric objects $\xi \subset X$, we can devise a fully-connected layer on *an arbitrary metric space X* as

$$S[\gamma](x) := \int_{\mathbb{R}\times\Xi} \gamma(a, \xi)\sigma(ad(x, \xi))\mathrm{d}a\mathrm{d}\xi.$$

If we have a nice coordinates such as $(s, t) \in \mathbb{R}^m \times \mathbb{R}^m$ satisfying $d(x(t), \xi(s)) = t - s$, then we can turn it to the Fourier expression and hopefully obtain the ridgelet transform.

Comparison to HNNs. While Wang (2021) and we employed a horosphere as the geometric object ξ, the original HNNs (Ganea et al., 2018; Shimizu et al., 2021) employed not a horosphere but a set $\xi_{geo.}(\boldsymbol{x}, \boldsymbol{u})$ of geodesics perpendicular to a normal vector $\boldsymbol{u} \in \mathbb{S}^{m-1}(\subset T_{\boldsymbol{x}}\mathbb{B}^m)$ at a point $\boldsymbol{x} \in \mathbb{B}^m$, called the *Poincaré hyperplane*. Since a Euclidean hyperplane can be understood as a set of Euclidean geodesics as well as a Euclidean sphere with infinite radius, both the Poincaré hyperplane and horosphere can be regarded as a hyperbolic counterpart of the Euclidean hyperplane. In fact, there are two types of Radon transform on the hyperbolic space: geodesic and horospherical Radon transforms. We conjecture that both networks can be understood as wavelet analysis on the Radon domain, but the original HNNs are based on the geodesic Radon transform, while ours are based on the horospherical Radon transform.

One of our reviewers kindly notified us that Yu & De Sa (2022) introduced the same weight function $\exp(\langle \boldsymbol{x}, \boldsymbol{u}\rangle)$, or the Poisson kernel, in graph learning by investigating the hyperbolic Laplacian. This is not a coincidence since the Helgason-Fourier transform decomposes function by the eigenfunctions of the Laplace-Beltrami operator on X.

Comparison to SPDNets. In a higher rank symmetric space, the Riemannian distance is not a complete two-point invariance, but the vector-valued distance is (see, e.g., Kapovich et al., 2017). Lopez et al. (2021) have recently utilized it. The original SPDNets (Huang & Gool, 2017; Dong et al., 2017; Gao et al., 2019; Brooks et al., 2019b;a) are composed of the BiMap layer $x \mapsto w^\top xw$ for $x \in \mathbb{P}_m$ with an orthonormal projection matrix $w \in \mathbb{R}^{m\times k}$ satisfying $w^\top w = I$, which extends the scalar product, and the ReEig layer $x \mapsto u^\top \max(0, \lambda - b)u$ via the spectral decomposition $x = u^\top \lambda u$, which extends the pointwise activation with ReLU. While we applied nonlinear activation σ on $\log \lambda(x) \in \mathfrak{a}^* \cong \mathbb{R}^r$, the original ReEig layer applied σ on $\lambda(x) \in A \cong \mathbb{R}^r_+$. It would be a routine to modify our main results to the original formulations.

Acknowledgements

The authors are grateful to anonymous reviewers for their valuable comments. This work was supported by JSPS KAKENHI 18K18113, JST CREST JPMJCR2015 and JPMJCR1913, JST PRESTO JPMJPR2125, and JST ACT-X JPMJAX2004.

References

Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. Log-Euclidean metrics for fast and simple calculus on diffusion tensors. *Magnetic Resonance in Medicine*, 56(2): 411–421, 2006. doi: https://doi.org/10.1002/mrm.20965.

1

Arsigny, V., Fillard, P., Pennec, X., and Ayache, N. Geometric Means in a Novel Vector Space Structure on Symmetric Positive-Definite Matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007. doi: 10.1137/050637996. 1

Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993. 3

Bartolucci, F., De Mari, F., and Monti, M. Unitarization of the Horocyclic Radon Transform on Symmetric Spaces. In De Mari, F. and De Vito, E. (eds.), *Harmonic and Applied Analysis: From Radon Transforms to Machine Learning*, pp. 1–54. Springer International Publishing, 2021. 6

Bhatia, R., Jain, T., and Lim, Y. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019. ISSN 0723-0869. doi: https://doi.org/10.1016/j.exmath.2018.01.002. 1

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv preprint: 2104.13478*, 2021. 1

Brooks, D., Schwander, O., Barbaresco, F., Schneider, J.-Y., and Cord, M. Riemannian batch normalization for SPD neural networks. In *Advances in Neural Information Processing Systems 32*, 2019a. 1, 3, 9

Brooks, D. A., Schwander, O., Barbaresco, F., Schneider, J.-Y., and Cord, M. Exploring Complex Time-series Representations for Riemannian Machine Learning of Radar Data. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3672–3676, 2019b. 1, 3, 9

Candès, E. J. *Ridgelets: theory and applications*. PhD thesis, Standford University, 1998. 3

Carroll, S. M. and Dickinson, B. W. Construction of neural nets using the Radon transform. In *International Joint Conference on Neural Networks 1989*, volume 1, pp. 607–611, 1989. 3

Chakraborty, R., Yang, C.-H., Zhen, X., Banerjee, M., Archer, D., Vaillancourt, D., Singh, V., and Vemuri, B. A Statistical Recurrent Model on the Manifold of Symmetric Positive Definite Matrices. In *Advances in Neural Information Processing Systems 31*, 2018. 1

Chakraborty, R., Bouza, J., Manton, J. H., and Vemuri, B. C. ManifoldNet: A Deep Neural Network for Manifold-Valued Data With Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):799–810, 2022. 1

Chizat, L. and Bach, F. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. In *Advances in Neural Information Processing Systems 32*, pp. 3036–3046, 2018. 3

Cruceru, C., Becigneul, G., and Ganea, O.-E. Computationally Tractable Riemannian Manifolds for Graph Embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):7133–7141, 2021. 1

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems (MCSS)*, 2(4):303–314, 1989. 3

Dong, Z., Jia, S., Zhang, C., Pei, M., and Wu, Y. Deep Manifold Learning of Symmetric Positive Definite Matrices with Application to Face Recognition. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pp. 4009–4015, 2017. 1, 3, 9

Donoho, D. L. Emerging applications of geometric multiscale analysis. *Proceedings of the ICM, Beijing 2002*, I: 209–233, 2002. 3

Funahashi, K.-I. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2 (3):183–192, 1989. 3

Ganea, O., Becigneul, G., and Hofmann, T. Hyperbolic Neural Networks. In *Advances in Neural Information Processing Systems 31*, 2018. 1, 3, 9

Gao, Z., Wu, Y., Bu, X., Yu, T., Yuan, J., and Jia, Y. Learning a robust representation via a deep network on symmetric positive definite manifolds. *Pattern Recognition*, 92:1–12, 2019. 1, 3, 9

Gel'fand, I. M. and Shilov, G. E. *Generalized Functions, Vol. 1: Properties and Operations*. Academic Press, New York, 1964. 4

Grafakos, L. *Classical Fourier Analysis*. Graduate Texts in Mathematics. Springer New York, second edition, 2008. 4

Gulcehre, C., Denil, M., Malinowski, M., Razavi, A., Pascanu, R., Hermann, K. M., Battaglia, P., Bapst, V., Raposo, D., Santoro, A., and de Freitas, N. Hyperbolic Attention Networks. In *International Conference on Learning Representations*, 2019. 1

Helgason, S. Radon-Fourier transforms on symmetric spaces and related group representations. *Bulletin of the American Mathematical Society*, 71(5):757–763, 1965. doi: bams/1183527303. 2

Helgason, S. *Groups and Geometric Analysis: Integral Geometry, Invariant Differential Operators, and Spherical Functions*, volume 83 of *Mathematical Surveys and Monographs*. American Mathematical Society, 1984. 1, 2, 5, 6, 13

Helgason, S. *Geometric Analysis on Symmetric Spaces: Second Edition*, volume 39 of *Mathematical Surveys and Monographs*. American Mathematical Society, second edition, 2008. 1, 2, 5, 6

Huang, Z. and Gool, L. V. A Riemannian Network for SPD Matrix Learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pp. 2036–2042, 2017. 1, 3, 9

Irie, B. and Miyake, S. Capabilities of three-layered perceptrons. In *IEEE International Conference on Neural Networks*, pp. 641–648, 1988. 3

Ito, Y. Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory. *Neural Networks*, 4(3):385–394, 1991. 3

Kainen, P. C., Kůrková, V., and Sanguineti, M. Approximating multivariable functions by feedforward neural nets. In Bianchini, M., Maggini, M., and Jain, L. C. (eds.), *Handbook on Neural Information Processing*, volume 49 of *Intelligent Systems Reference Library*, pp. 143–181. Springer Berlin Heidelberg, 2013. 2

Kapovich, M., Leeb, B., and Porti, J. Anosov subgroups: dynamical and geometric characterizations. *European Journal of Mathematics*, 3(4):808–898, 2017. ISSN 2199-6768. doi: 10.1007/s40879-017-0192-y. 6, 9

Kostadinova, S., Pilipović, S., Saneva, K., and Vindas, J. The ridgelet transform of distributions. *Integral Transforms and Special Functions*, 25(5):344–358, 2014. 3, 4

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguñá, M. Hyperbolic geometry of complex networks. *Phys. Rev. E*, 82(3):36106, 2010. 1

Kutyniok, G. and Labate, D. *Shearlets: Multiscale Analysis for Multivariate Data*. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, 1 edition, 2012. doi: 10.1007/978-0-8176-8316-0. 3

Lopez, F., Pozzetti, B., Trettel, S., Strube, M., and Wienhard, A. Vector-valued Distance and Gyrocalculus on the Space of Symmetric Positive Definite Matrices. In *Advances in Neural Information Processing Systems 34*, 2021. 1, 9

Mallat, S. *A Wavelet Tour of Signal Processing, Third Edition: The Sparse Way*. Academic Press, 2009. 4, 9

Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33): E7665–E7671, 2018. 3

Murata, N. An integral representation of functions using three-layered betworks and their approximation bounds. *Neural Networks*, 9(6):947–956, 1996. 3

Nickel, M. and Kiela, D. Poincaré Embeddings for Learning Hierarchical Representations. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 1

Nickel, M. and Kiela, D. Learning Continuous Hierarchies in the Lorentz Model of Hyperbolic Geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 3779–3788, 2018. 1

Nitanda, A. and Suzuki, T. Stochastic Particle Gradient Descent for Infinite Ensembles. *arXiv preprint: 1712.05438*, 2017. 3

Ongie, G., Willett, R., Soudry, D., and Srebro, N. A Function Space View of Bounded Norm Infinite Width ReLU Nets: The Multivariate Case. In *International Conference on Learning Representations*, 2020. 3

Parhi, R. and Nowak, R. D. Neural Networks, Ridge Splines, and TV Regularization in the Radon Domain. *arXiv preprint: 2006.05626*, 2020. 3

Pennec, X., Fillard, P., and Ayache, N. A Riemannian Framework for Tensor Computing. *International Journal of Computer Vision*, 66(1):41–66, 2006. ISSN 1573-1405. doi: 10.1007/s11263-005-3222-z. 1

Petersen, P. and Voigtlaender, F. Equivalence of approximation by convolutional neural networks and fully-connected networks. *Proceedings of the American Mathematical Society*, 148(4):1567–1581, 2020. doi: 10.1090/proc/14789. 3

Rotskoff, G. and Vanden-Eijnden, E. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In *Advances in Neural Information Processing Systems 31*, pp. 7146–7155, 2018. 3

Rubin, B. The Calderón reproducing formula, windowed X-ray transforms, and radon transforms in L^p-spaces. *Journal of Fourier Analysis and Applications*, 4(2):175–197, 1998. 3

Sala, F., De Sa, C., Gu, A., and Re, C. Representation Tradeoffs for Hyperbolic Embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 4460–4469, 2018. 1

Savarese, P., Evron, I., Soudry, D., and Srebro, N. How do infinite width bounded norm networks look in function space? In *Proceedings of the 32nd Conference on Learning Theory*, volume 99, pp. 2667–2690, 2019. 3

Shimizu, R., Mukuta, Y., and Harada, T. Hyperbolic Neural Networks++. In *International Conference on Learning Representations*, 2021. 1, 3, 9

Sirignano, J. and Spiliopoulos, K. Mean Field Analysis of Neural Networks: A Law of Large Numbers. *SIAM Journal on Applied Mathematics*, 80(2):725–752, 2020. 3

Sonoda, S. and Murata, N. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 43(2):233–268, 2017. 2, 3, 4, 15

Sonoda, S., Ishikawa, I., and Ikeda, M. Ridge Regression with Over-Parametrized Two-Layer Networks Converge to Ridgelet Spectrum. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021*, volume 130, pp. 2674–2682, 2021a. 2

Sonoda, S., Ishikawa, I., and Ikeda, M. Ghosts in Neural Networks: Existence, Structure and Role of Infinite-Dimensional Null Space. *arXiv preprint: 2106.04770*, 2021b. 2, 15

Sra, S. A new metric on the manifold of kernel matrices with application to matrix geometric means. In *Advances in Neural Information Processing Systems 25*, 2012. 1

Starck, J.-L., Murtagh, F., and Fadili, J. M. The ridgelet and curvelet transforms. In *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*, pp. 89–118. Cambridge University Press, 2010. 3

Suzuki, T. Generalization bound of globally optimal nonconvex neural network training: Transportation map estimation by infinite dimensional Langevin dynamics. In *Advances in Neural Information Processing Systems 33*, pp. 19224–19237, 2020. 3

Terras, A. *Harmonic Analysis on Symmetric Spaces—Higher Rank Spaces, Positive Definite Matrix Space and Generalizations*. Springer New York, 2016. 17

Unser, M. A Representer Theorem for Deep Neural Networks. *Journal of Machine Learning Research*, 20(110): 1–30, 2019. 3

Wang, M.-X. Laplacian Eigenspaces, Horocycles and Neuron Models on Hyperbolic Spaces. 2021. 2, 3, 9

Yu, T. and De Sa, C. HyLa: Hyperbolic Laplacian Features For Graph Learning. *arXiv preprint: 2202.06854*, feb 2022. 9

Zhang, T., Zheng, W., Cui, Z., and Li, C. Deep Manifold-to-Manifold Transforming Network. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pp. 4098–4102, 2018. 1

A. Poincaré Disk D as Noncompact Riemannian Symmetric Space

Following Helgason (1984, Intro. § 4), we review the homogeneous space aspect of a hyperbolic space. Note that the Riemannian metric here drops the factor $\times 2^2$. Let $D := \{z \in \mathbb{C} \mid |z| < 1\}$ be the unit open disk in $\mathbb{C}$ equipped with the Riemannian metric $g_z(u,v) = (u,v)/(1-|z|^2)^2$ for any tangent vectors $u, v \in T_z D$ at $z \in D$, where $(\cdot,\cdot)$ denotes the Euclidean inner product in $\mathbb{R}^2$. Let $\partial D := \{u \in \mathbb{C} \mid |u| = 1\}$ be the boundary of D equipped with the uniform probability measure $\mathrm{d}u$. Namely, D is the *Poincaré disk model of hyperbolic plane* $\mathbb{H}^2$. On this model, the Poincaré metric between two points $z, w \in D$ is given by $d(z,w) = \tanh^{-1} |(z-w)/(1-zw^*)|$, and the volume element is given by $\mathrm{d}z = (1-(x^2+y^2))^{-2}\mathrm{d}x\mathrm{d}y$.

Consider now the group

$$G = SU(1,1) := \left\{ \begin{pmatrix} \alpha & \beta \\ \beta^* & \alpha^* \end{pmatrix} \middle| (\alpha,\beta) \in \mathbb{C}^2, |\alpha|^2 - |\beta|^2 = 1 \right\},$$

which acts on D (and ∂D) by

$$g \cdot z := \frac{\alpha z + \beta}{\beta^* z + \alpha^*}, \quad z \in D \cup \partial D.$$

The G-action is transitive, conformal, and maps circles, lines, and the boundary into circles, lines, and the boundary. In addition, consider the subgroups

$$K := SO(2) = \left\{ k_\phi := \begin{pmatrix} e^{i\phi} & 0 \\ 0 & e^{-i\phi} \end{pmatrix} \middle| \phi \in [0, 2\pi) \right\},$$

$$A := \left\{ a_t := \begin{pmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{pmatrix} \middle| t \in \mathbb{R} \right\},$$

$$N := \left\{ n_s := \begin{pmatrix} 1+is & -is \\ is & 1-is \end{pmatrix} \middle| s \in \mathbb{R} \right\},$$

$$M := C_K(A) = \left\{ k_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, k_\pi = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \right\}$$

The subgroup $K := SO(2)$ fixes the origin $o \in D$. So we have the identifications

$$D = G/K = SU(1,1)/SO(2), \quad \text{and} \quad \partial D = K/M = \mathbb{S}^1.$$

On this model, the following are known (1) that $m = \dim \mathfrak{a} = 1$, $|W| = 1$, $\varrho = 1$, and $|\boldsymbol{c}(\lambda)|^{-2} = \frac{\pi\lambda}{2}\tanh(\frac{\pi\lambda}{2})$ for $\lambda \in \mathfrak{a}^* = \mathbb{R}$, (2) that the geodesics are the circular arcs perpendicular to the boundary ∂D, and (3) that the horocycles are the circles tangent to the boundary ∂D. Hence, let $\xi(x,u)$ denote the horocycle ξ through $x \in D$ and tangent to the boundary at $u \in \partial D$; and let $\langle x, u\rangle$ denote the signed distance from the origin $o \in D$ to the horocycle $\xi(x,u)$.

In order to compute the distance $\langle z, u\rangle$, we use the following fact: The distance from the origin o to a point $z = re^{iu}$ is $d(o,z) = \tanh^{-1}|(0-z)/(1-0z^*)| = \frac{1}{2}\log\frac{1+r}{1-r}$. Hence, let $c \in D$ be the center of the horocycle $\xi(z,u)$, and let $w \in D$ be the closest point on the horocycle $\xi(z,u)$ to the origin. By definition, $\langle z, u\rangle = d(o,w)$. But we can find the w via the cosine rule:

$$\cos zou = \frac{|u|^2 + |z|^2 - |z-u|^2}{2|u||z|} = \cos zoc = \frac{|z|^2 + |\frac{1}{2}(1+|w|)|^2 - |\frac{1}{2}(1-|w|)|^2}{2|z||\frac{1}{2}(1+|w|)|},$$

which yields the tractable formula:

$$\langle z, u\rangle = \frac{1}{2}\log\frac{1+|w|}{1-|w|} = \frac{1}{2}\log\frac{1-|z|^2}{|z-u|^2}, \quad (z,u) \in D \times \partial D.$$

B. Proofs

B.1. Theorem 4.3 (Reconstruction Formula)

Proof. We identify the scale parameter $a \in \mathfrak{a}^*$ with vector $\boldsymbol{a} \in \mathbb{R}^r$.

Step 1. Since $b \in \mathbb{R}$, the Fourier expression is given by

$$\begin{aligned} S[\gamma](x) &:= \int_{\mathbb{R}^r \times \partial X \times \mathbb{R}} \gamma(\boldsymbol{a}, u, b)\sigma(\boldsymbol{a} \cdot \langle x, u\rangle - b)e^{\varrho\langle x,u\rangle}\mathrm{d}\boldsymbol{a}\mathrm{d}u\mathrm{d}b \\ &= \frac{1}{2\pi}\int_{\mathbb{R}^r \times \partial X \times \mathbb{R}} \gamma^\sharp(\boldsymbol{a}, u, \omega)\sigma^\sharp(\omega)e^{(i\omega\boldsymbol{a}+\varrho)\langle x,u\rangle}\mathrm{d}\boldsymbol{a}\mathrm{d}u\mathrm{d}\omega. \end{aligned}$$

Step 2. By changing the variables as $(\boldsymbol{a}, \omega) = (\boldsymbol{\lambda}/\omega, \omega)$ with $\mathrm{d}\boldsymbol{a}\mathrm{d}\omega = |\omega|^{-r}\mathrm{d}\boldsymbol{\lambda}\mathrm{d}\omega$, and identifying the vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r) \in \mathbb{R}^r$ with $\lambda \in \mathfrak{a}^*$, we have

$$S[\gamma](x) = \frac{1}{2\pi}\int_{\mathbb{R}}\left[\int_{\mathfrak{a}^* \times \partial X} \gamma^\sharp(\lambda/\omega, u, \omega)e^{(i\lambda+\varrho)\langle x,u\rangle}\mathrm{d}\lambda\mathrm{d}u\right]\sigma^\sharp(\omega)|\omega|^{-r}\mathrm{d}\omega.$$

Step 3. Since inside the bracket $[\cdots]$ is the inverse Helgason-Fourier transform (excluding the Plancherel measure $|\boldsymbol{c}(\lambda)|^{-2}$), put a separation-of-variables form as

$$\gamma^\sharp_{f,\rho}(\lambda/\omega, u, \omega) = \widehat{f}(\lambda, u)\overline{\rho^\sharp(\omega)}|\boldsymbol{c}(\lambda)|^{-2},$$

we have a particular solution:

$$S[\gamma_{f,\rho}](x) = \left(\frac{|W|}{2\pi}\int_{\mathbb{R}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-r}\mathrm{d}\omega\right)\left(\int_{\mathfrak{a}^* \times \partial X} \widehat{f}(\lambda, u)e^{(i\lambda+\varrho)\langle x,u\rangle}\frac{\mathrm{d}\lambda\mathrm{d}u}{|W||\boldsymbol{c}(\lambda)|^2}\right) = (\!(\sigma, \rho)\!)f(x),$$

where we put

$$(\!(\sigma, \rho)\!) := \frac{|W|}{2\pi}\int_{\mathbb{R}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-r}\mathrm{d}\omega.$$

Here, the equality holds for every point $x \in X$ when $f \in C_c^\infty(X)$, and in L^2 when $f \in L^2(X)$.

In particular, the ridgelet transform is calculated as

$$\begin{aligned} R[f;\rho](\boldsymbol{a}, u, b) &:= \frac{1}{2\pi}\int_{\mathbb{R}} \gamma^\sharp_{f,\rho}(\boldsymbol{a}, u, \omega)e^{i\omega b}\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{f}(\omega\boldsymbol{a}, u)|\boldsymbol{c}(\omega\boldsymbol{a})|^{-2}\overline{\rho^\sharp(\omega)}e^{i\omega b}\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{\mathbb{R}\times X} \boldsymbol{c}[f](x)\overline{\rho^\sharp(\omega)}e^{(-i\omega\boldsymbol{a}+\varrho)\langle x,u\rangle + i\omega b}\mathrm{d}x\mathrm{d}\omega \\ &= \int_X \boldsymbol{c}[f](x)\overline{\rho(\boldsymbol{a}\cdot\langle x,u\rangle - b)}e^{\varrho\langle x,u\rangle}\mathrm{d}x, \end{aligned}$$

where we put $\boldsymbol{c}[f]$ as a Helgason-Fourier multiplier satisfying $\widehat{\boldsymbol{c}[f]}(\lambda, u) = \widehat{f}(\lambda, u)|\boldsymbol{c}(\lambda)|^{-2}$. □

B.2. Theorem 4.4 (cc-Universality)

Additional Notation. For a function f on a set X, $\|f\|_{C(X)} := \sup_{x\in X}|f(x)|$ denotes the uniform norm on X.

For any integer $d > 0$ and vector $\boldsymbol{v} \in \mathbb{R}^d$, $|\boldsymbol{v}|$ denotes the Euclidean norm, and $\langle \boldsymbol{v}\rangle := \sqrt{1 + |\boldsymbol{v}|^2}$. For any positive number $t > 0$, $\triangle^{t/2}$ and $\langle\triangle\rangle^t$ denote fractional differential operators defined as Fourier multipliers: for any $\phi \in \mathcal{S}'(\mathbb{R}^d)$,

$$\triangle^{t/2}[\phi](\boldsymbol{v}) := \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} |\boldsymbol{u}|^t\widehat{\phi}(\boldsymbol{u})e^{i\boldsymbol{u}\cdot\boldsymbol{v}}\mathrm{d}\boldsymbol{u}, \quad \langle\triangle\rangle^{t/2}[\phi](\boldsymbol{v}) := \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} (1 + |\boldsymbol{u}|^2)^{t/2}\widehat{\phi}(\boldsymbol{u})e^{i\boldsymbol{u}\cdot\boldsymbol{v}}\mathrm{d}\boldsymbol{u}.$$

In particular when $t = 2$, $\triangle^{t/2}$ coincides with the ordinary Laplacian on $\mathbb{R}^d$.

Proof. We will show that for any compact set $Z \subset X$, positive number $\varepsilon > 0$, compactly-supported continuous function $f \in C(Z)$, there exists a finite network f_n such that $\|f - f_n\|_{C(Z)} < \varepsilon$.

Since $\sum_{i=1}^{n} c_i \Delta_\theta^k[\sigma](a_i\langle x, u_i\rangle - b_i)e^{\varrho\langle x,u_i\rangle}$ is rewritten as another finite model $\sum_{i=1}^{n'} c_i' \sigma(a_i'\langle x, u_i'\rangle - b_i')e^{\varrho\langle x,u_i'\rangle}$, it suffice to consider the case $k = 0$. In the following, we assume that $\sigma(= \Delta_\theta^0[\sigma])$ is bounded and Lipschitz continuous. So, put $M_\sigma := \|\sigma\|_{L^\infty(\mathbb{R})}$ and $L_\sigma := \mathrm{Lip}(\sigma)$. As a consequence of the Iwasawa decomposition, the composite distance $\langle x, u\rangle$ is C^∞-smooth and thus Lipschitz continuous. Hence, put $L_c := \sup_{x \in Z} \sup_{u,u' \in \partial X} |\langle x, u\rangle - \langle x, u'\rangle|/d(u, u')$, $L_e := \sup_{x \in Z} \sup_{u,u' \in \partial X} |\exp(\varrho\langle x, u\rangle) - \exp(\varrho\langle x, u'\rangle)|/d(u, u')$, and $M_e := \sup_{x \in Z, u \in \partial X} |\exp(\varrho\langle x, u\rangle)|$.

Step 1 ($f \sim f_c$). By the density of $C_c^\infty(X)$ in $C(Z)$ with respect to the uniform norm, we can take a compactly-supported smooth function $f_c \in C_c^\infty(Z)$ satisfying $\|f - f_c\|_{C(Z)} < \varepsilon/3$. Since f_c is sufficiently smooth and integrable, there exists a compactly-supported smooth function $\rho \in C_c^\infty(\mathbb{R})$ such that

$$S[R[f_c; \rho]](x) = f_c(x) \text{ at every point } x \in X.$$

For example, take a compactly-supported smooth function $\rho_0 \in C_c^\infty(\mathbb{R})$, and put $\rho(b) := \triangle_b^{r/2}[\rho_0](b) = \frac{1}{2\pi}\int_{\mathbb{R}} |\omega|^r \rho_0^\sharp(\omega)e^{ib\omega}\mathrm{d}\omega$. Then, $(\!(\sigma, \rho)\!) = \frac{|W|}{2\pi}\int_{\mathbb{R}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-r}\mathrm{d}\omega = \frac{|W|}{2\pi}\int_{\mathbb{R}} \sigma^\sharp(\omega)\overline{\rho_0^\sharp(\omega)}\mathrm{d}\omega = |W|\int_{\mathbb{R}} \sigma(b)\overline{\rho_0(b)}\mathrm{d}b = |W|\langle\sigma, \rho_0\rangle_{L^2(\mathbb{R})}$, which is an ordinary functional inner product, and it is easy to find a ρ_0 satisfying $\langle\sigma, \rho_0\rangle_{L^2(\mathbb{R})} \neq 0$. By normalizing $\rho' := \rho/(\!(\sigma, \rho)\!)$, we can find the ρ'. We refer to Sonoda & Murata (2017) and Sonoda et al. (2021b) for more details on the scalar product $(\!(\sigma, \rho)\!)$.

Step 2 ($R[f_c; \rho]$). To show a discretization f_n of the reconstruction formula converges to f_c in $C(Z)$, it is convenient to regard the integrand

$$\phi(a, u, b)(x) := R[f_c; \rho](a, u, b)\sigma(a\langle x, u\rangle - b)e^{\varrho\langle x,u\rangle}$$

as a vector-valued function $\phi : \mathfrak{a}^* \times \partial X \times \mathbb{R} \to C(Z)$, and the integration $S[\gamma](x) = \int_{\mathfrak{a}^* \times \partial X \times \mathbb{R}} \phi(a, u, b)(x)\mathrm{d}a\mathrm{d}u\mathrm{d}b$ as a Bochner integral. Since f_c is C^∞-smooth, $R[f_c; \rho](a, u, b)$ is bounded and decays rapidly in a, and thus ϕ is Bochner integrable, that is,

$$\int_{\mathfrak{a}^* \times \partial X \times \mathbb{R}} \|\phi(a, u, b)\|_{C(Z)}\mathrm{d}a\mathrm{d}u\mathrm{d}b < \infty.$$

To see this, the decay property is estimated as follows. For any positive numbers $s, t > 1$,

$$\begin{aligned}
|R[f_c; \rho](a, u, b)| &= \frac{1}{2\pi}\left|\int_{\mathbb{R}} \widehat{f}_c(\omega a, u)|\boldsymbol{c}(\omega a)|^{-2}\overline{\rho^\sharp(\omega)}e^{i\omega b}\mathrm{d}\omega\right| \\
&= \frac{1}{2\pi}\left|\int_{\mathbb{R}} \langle\omega a\rangle^s\langle\omega a\rangle^{-s}\langle b\rangle^t\langle b\rangle^{-t}\widehat{f}_c(\omega a, u)|\boldsymbol{c}(\omega a)|^{-2}\overline{\rho^\sharp(\omega)}e^{i\omega b}\mathrm{d}\omega\right| \\
&\leqslant \frac{1}{2\pi}\left|\int_{\mathbb{R}} \langle\omega a\rangle^s\widehat{f}_c(\omega a, u)|\boldsymbol{c}(\omega a)|^{-2}\langle\omega\rangle^{-s}\overline{\rho^\sharp(\omega)}\langle\triangle_\omega\rangle^t e^{i\omega b}\mathrm{d}\omega\right|\langle a\rangle^{-s}\langle b\rangle^{-t},
\end{aligned}$$

which asserts the integrability as below

$$\int_{\mathfrak{a}^* \times \partial X \times \mathbb{R}} \|\phi(a, u, b)\|_{C(Z)}\mathrm{d}a\mathrm{d}u\mathrm{d}b \leqslant M_\sigma M_e\|R[f_c; \rho]\|_{L^1(X)} \lesssim \int_{\mathfrak{a}^* \times \partial X \times \mathbb{R}} \langle a\rangle^{-s}\langle b\rangle^{-t}\mathrm{d}a\mathrm{d}u\mathrm{d}b < \infty.$$

Step 3 ($f_c \sim f_V \sim f_n$). Next, take a compact domain $V := \{(a, u, b) \in \mathfrak{a}^* \times \partial X \times \mathbb{R} \mid |a_i| \leqslant \delta/2, |b| \leqslant \delta/2\}$, namely the product of an $(r + 1)$-dimensional hypercube and the compact manifold ∂X, and put a band-limited function

$$f_V(x) := \int_V \phi(a, u, b)(x)\mathrm{d}a\mathrm{d}u\mathrm{d}b,$$

so that $\|f_c - f_V\|_{C(Z)} < \varepsilon/3$ (by letting δ sufficiently large). For each $n \in \mathbb{N}$, let $V = \bigsqcup_{i \in I_n} V_{ni}$ be a disjoint decomposition of V into a disjoint family of $|I_n|$ subsets V_{ni} with diameter at most $d_n = O(1/n)$. Since V is a compact manifold, each volume $\mathrm{vol}(V_{ni})$ decays at $O(n^{-\dim V})$ as $n \to \infty$, and the cardinality $|I_n|$ ($\approx d_n$-covering number) grows at the reciprocal $O(n^{\dim V})$. From each subset V_{ni}, take a point (a_{ni}, u_{ni}, b_{ni}) satisfying

$$c_{ni} := \int_{V_{ni}} R[f_c; \rho](a, u, b) \mathrm{d}a\mathrm{d}u\mathrm{d}b = R[f_c; \rho](a_{ni}, u_{ni}, b_{ni}) \,\mathrm{vol}(V_{ni}),$$

and put a finite network as

$$f_n(x) := \sum_{i \in I_n} c_{ni} \sigma(a_{ni}\langle x, u_{ni}\rangle - b_{ni}) e^{\varrho\langle x, u_{ni}\rangle}.$$

In addition, we use

$$\phi_{ni}(x) := \phi(a_{ni}, u_{ni}, b_{ni})(x), \quad \text{and} \quad \phi_n(a, u, b)(x) := \sum_{i \in I_n} \mathbf{1}_{V_{ni}}(a, u, b) \phi_{ni}(x),$$

so that

$$f_n(x) = \sum_{i \in I_n} \phi_{ni}(x) \,\mathrm{vol}(V_{ni}) = \int_V \phi_n(a, u, b)(x) \mathrm{d}a\mathrm{d}u\mathrm{d}b.$$

Step 4 ($f_V \sim f_n$). We show $f_n \to f_V$ in $C(Z)$. Put $M_R := \|R[f_c; \rho]\|_{C(V)}$ and $L_R := \mathrm{Lip}(R[f_c; \rho])$. For every $n \in \mathbb{N}$, since

$$\begin{aligned}\|f_V - f_n\|_{C(Z)} &= \sup_{x \in Z} \left| \int_V \phi(a, u, b)(x) \mathrm{d}a\mathrm{d}u\mathrm{d}b - \int_V \phi_n(a, u, b)(x) \mathrm{d}a\mathrm{d}u\mathrm{d}b \right| \\ &\leqslant \int_V \|\phi(a, u, b) - \phi_n(a, u, b)\|_{C(Z)} \mathrm{d}a\mathrm{d}u\mathrm{d}b,\end{aligned}$$

it suffice to show that (1) ϕ_n is a.e. dominated by an integrable function, and (2) converges a.e. to ϕ. In the following, we fix an arbitrary $(a, u, b) \in V_{ni}$. First, ϕ_n is uniformly dominated by a constant function, which is in $L^1(V)$, that is,

$$\|\phi_n(a, u, x)\|_{C(Z)} = \|\phi_{ni}\|_{C(Z)} \leqslant \sup_{(a', u', b') \in V_{ni}} \|\phi(a', u', b')\|_{C(Z)} \leqslant M_R M_\sigma M_e.$$

Second, ϕ_n coverges to ϕ a.e.:

$$\begin{aligned}&\|\phi(a, u, b) - \phi_n(a, u, b)\|_{C(Z)} \\ &= \|\phi(a, u, b) - \phi_{ni}\|_{C(Z)} \\ &= \sup_{x \in Z} \left| R[f_c; \rho](a, u, b) \sigma(a\langle x, u\rangle - b) e^{\varrho\langle x, u\rangle} - R[f_c; \rho](a_{ni}, u_{ni}, b_{ni}) \sigma(a_{ni}\langle x, u_{ni}\rangle - b_{ni}) e^{\varrho\langle x, u_{ni}\rangle} \right| \\ &\leqslant \sup_{x \in Z} \left| R[f_c; \rho](a, u, b) \right| \left| \sigma(a\langle x, u\rangle - b) e^{\varrho\langle x, u\rangle} - \sigma(a_{ni}\langle x, u_{ni}\rangle - b_{ni}) e^{\varrho\langle x, u_{ni}\rangle} \right| \\ &\qquad + \sup_{x \in Z} \left| R[f_c; \rho](a_{ni}, u_{ni}, b_{ni}) - R[f_c; \rho](a, u, b) \right| \left| \sigma(a_{ni}\langle x, u_{ni}\rangle - b_{ni}) e^{\varrho\langle x, u_{ni}\rangle} \right| \\ &\leqslant M_R \left(L_\sigma M_e \sup_{x \in Z} \left| a\langle x, u\rangle - a_{ni}\langle x, u_{ni}\rangle + (b - b_{ni}) \right| + M_\sigma L_e d(u, u_{ni}) \right) \\ &\qquad + M_\sigma M_e L_R \left| d((a, u, b), (a_{ni}, u_{ni}, b_{ni})) \right| \\ &\lesssim d_n = O(1/n) \to 0, \quad n \to \infty.\end{aligned}$$

Therefore, the dominated convergence theorem for the Bochner integral yields

$$\|f_V - f_n\|_{C(Z)} \leqslant \int_V \|\phi(a, u, b) - \phi_n(a, u, b)\|_{C(Z)} \mathrm{d}a\mathrm{d}u\mathrm{d}b \to 0, \quad n \to \infty.$$

Hence by letting n sufficiently large, we have $\|f_n - f_V\|_{C(Z)} < \varepsilon/3$.

To sum up, we have shown the cc-universality:

$$\|f - f_n\|_{C(Z)} \leqslant \|f - f_c\|_{C(Z)} + \|f_c - f_V\|_{C(Z)} + \|f_V - f_n\|_{C(Z)} < \varepsilon.$$

□

C. Further Examples: SPDNets

C.1. SPD Manifold

Following Terras (2016, Chapter 1), we introduce the SPD manifold. On the space $\mathbb{P}_m$ of $m \times m$ symmetric positive definite (SPD) matrices, the Riemannian metric is given by

$$\mathfrak{g}_x := \operatorname{tr}\left((x^{-1}\mathrm{d}x)^2\right), \quad x \in \mathbb{P}_m$$

where x and $\mathrm{d}x$ denote the matrices of entries x_{ij} and $\mathrm{d}x_{ij}$.

Put $G = GL(m, \mathbb{R})$, then the Iwasawa decomposition $G = KAN$ is given by $K = O(m), A = D_+(m), N = T_1(m)$; and the centralizer $M = C_K(A)$ is given by $M = D_{\pm 1}$ (diagonal matrices with entries ± 1). The quotient space G/K is identified with the SPD manifold $\mathbb{P}_m$ via a diffeomorphism onto, $gK \mapsto gg^\top$ for any $g \in G$; and K/M is identified with the boundary $\partial\mathbb{P}_m$, another manifold of all *singular positive semidefinite* matrices. The action of G on $\mathbb{P}_m$ is given by $g[x] := gxg^\top$ for any $g \in G$ and $x \in \mathbb{P}_m$. In particular, the metric $\mathfrak{g}$ is G-invariant. According to the *spectral decomposition*, for any $x \in \mathbb{P}_m$, there uniquely exist $k \in K$ and $a \in A$ such that $x = k[a]$; and according to the *Cholesky (or Iwasawa) decomposition*, there exist $n \in N$ and $a \in A$ such that $x = n[a]$.

When $x = k[\exp(H)] = \exp(k[H])$ for some $H \in \mathfrak{a} = D(m)$ and $k \in K$, then the geodesic segment y from the origin $o = I$ (the identity matrix) to x is given by

$$y(t) = \exp(tk[H]), \quad t \in [0, 1]$$

satisfying $y(0) = o$ and $y(1) = x$; and the Riemannian length of y (i.e., the Riemannian distance from o to x) is given by $d(o, x) = |H|_E$. So, $H \in \mathfrak{a}$ is the *vector-valued distance* from o to $x = k[\exp(H)]$.

The G-invariant measures are given by $\mathrm{d}g = |\det g|^{-m} \bigwedge_{i,j} \mathrm{d}g_{ij}$ on G, $\mathrm{d}k$ to be the uniform probability measure on K, $\mathrm{d}a = \bigwedge_i \mathrm{d}a_i/a_i$ on A, $\mathrm{d}n = \bigwedge_{1<i<j\leqslant m} \mathrm{d}n_{ij}$ on N,

$$\begin{aligned}\mathrm{d}\mu(x) &= |\det x|^{-\frac{m+1}{2}} \bigwedge_{1\leqslant i\leqslant j\leqslant m} \mathrm{d}x_{ij} \quad \text{on} \quad \mathbb{P}_m, \\ &= c_m \prod_{j=1}^m a_j^{-\frac{m-1}{2}} \prod_{1\leqslant i<j\leqslant m} |a_i - a_j| \mathrm{d}a\mathrm{d}k,\end{aligned}$$

where the second expression is for the polar coordinates $x \leftarrow k[a]$ with $(k, a) \in K \times A$ and $c_m := \pi^{(m^2+m)/4} \prod_{j=1}^m j^{-1}\Gamma^{-1}(j/2)$, and $\mathrm{d}u$ to be the uniform probability measure on $\partial\mathbb{P}_m := K/M$.

The vector-valued composite distance from the origin o to a horosphere $\xi(x, u)$ is calculated as

$$\langle x = g[o], u = kM\rangle = \frac{1}{2} \log \lambda(k^\top[x]),$$

where $\lambda(y)$ denotes the diagonal vector λ in the *Cholesky decomposition* $y = \nu[\lambda] = \nu\lambda\nu^\top$ of y for some $(\nu, \lambda) \in NA$.

Proof. Since $\langle x, kM\rangle := -H(g^{-1}k) = \langle k^\top[x], eM\rangle$, it suffices to consider the case $(x, u) = (g[o], eM)$. Namely, we solve $g^{-1} = kan$ for unknowns $(k, a, n) \in KAN$. (To be preceise, we only need a because $\langle x, eM\rangle = -\log a$.) Put the Cholesky decomposition $x = \nu[\lambda] = \nu\lambda\nu^\top$ for some $(\nu, \lambda) \in NA$. Then, $a = \lambda^{-1/2}$ because $x^{-1} = (\nu^{-1})^\top\lambda^{-1}\nu^{-1}$, while $x^{-1} = (gg^\top)^{-1} = n^\top a^2 n$. □

The Helgason-Fourier transform and its inversion formula are given by

$$\begin{aligned}\widehat{f}(\boldsymbol{s}, u) &= \int_{\mathbb{P}_m} f(x)\overline{e^{\boldsymbol{s}\cdot\langle x,u\rangle}}\mathrm{d}\mu(x), \\ f(x) &= \omega_m \int_{\Re \boldsymbol{s}=\boldsymbol{\varrho}} \int_{\partial\mathbb{P}_m} \widehat{f}(\boldsymbol{s}, u) e^{\boldsymbol{s}\cdot\langle x,u\rangle} \mathrm{d}u \frac{\mathrm{d}\boldsymbol{s}}{|\boldsymbol{c}(\boldsymbol{s})|^2},\end{aligned}$$

for any $(\boldsymbol{s}, u) \in \mathfrak{a}_{\mathbb{C}}^{*} \times O(m)$ (where $\mathfrak{a}_{\mathbb{C}}^{*} = \mathbb{C}^m$) and $x \in \mathbb{P}_m$. Here, $\omega_m := \prod_{j=1}^{m} \frac{\Gamma(j/2)}{j(2\pi i)\pi^{j/2}}$, $\boldsymbol{\varrho} = (-\frac{1}{2}, \dots, -\frac{1}{2}, \frac{m-1}{4}) \in \mathbb{C}^m$, and

$$\boldsymbol{c}(\boldsymbol{s}) = \prod_{1 \leqslant i \leqslant j < m} \frac{B(\frac{1}{2}, s_i + \cdots + s_j + \frac{j-i+1}{2})}{B(\frac{1}{2}, \frac{j-i+1}{2})},$$

where $B(x, y) := \Gamma(x)\Gamma(y)/\Gamma(x+y)$ is the beta function.

C.2. Continuous SPDNet

Definition C.1. For any $x \in \mathbb{P}_m$, put

$$S[\gamma](x) = \int_{\mathbb{R}^m \times \partial\mathbb{P}_m \times \mathbb{R}} \gamma(\boldsymbol{a}, u, b)\sigma(\boldsymbol{a} \cdot \langle x, u \rangle - b) e^{\boldsymbol{\varrho} \cdot \langle x, u \rangle} \mathrm{d}\boldsymbol{a} \mathrm{d}u \mathrm{d}b,$$

where for any $(x, u) \in \mathbb{P}_m \times \partial\mathbb{P}_m$ with $u = kM$ for some $k \in K$,

$$\langle x, u \rangle = \frac{1}{2} \log \lambda(k^\top[x]).$$

Definition C.2. For any $(\boldsymbol{a}, b) \in \mathbb{R}^m \times \mathbb{R}$,

$$R[f; \rho](\boldsymbol{a}, b) = \int_{\mathbb{P}_m} \boldsymbol{c}[f](x) \overline{\sigma(\boldsymbol{a} \cdot \langle x, u \rangle - b)} e^{\boldsymbol{\varrho} \cdot \langle x, u \rangle} \mathrm{d}\mu(x),$$

where for any $x \in \mathbb{P}_m$,

$$\boldsymbol{c}[f](x) = \int_{\mathbb{R}^m \times \partial\mathbb{P}_m} \widehat{f}(i\boldsymbol{\lambda} + \boldsymbol{\varrho}, u) e^{(i\boldsymbol{\lambda} + \boldsymbol{\varrho}) \cdot \langle x, u \rangle} \frac{\omega_m \mathrm{d}\boldsymbol{\lambda} \mathrm{d}u}{|\boldsymbol{c}(i\boldsymbol{\lambda} + \boldsymbol{\varrho})|^4}.$$

As a consequence of the general results, the following reconstruction formula holds.

Corollary C.3. *For any $\sigma \in \mathcal{S}'(\mathbb{R})$, $\rho \in \mathcal{S}(\mathbb{R})$,*

$$S[R[f; \rho]](x) = (\!(\sigma, \rho)\!) f(x),$$

where

$$(\!(\sigma, \rho)\!) := \frac{1}{2\pi} \int_{\mathbb{R}} \sigma^\sharp(\omega) \overline{\rho^\sharp(\omega)} |\omega|^{-m} \mathrm{d}\omega,$$

where the equality holds at every point $x \in \mathbb{P}_m$ when $f \in C_c^\infty(\mathbb{P}_m)$, and in L^2 when $f \in L^2(\mathbb{P}_m)$.