Journal of Statistical Planning and Inference 233 (2024) 106184

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

A unified Fourier slice method to derive ridgelet transform for a variety of depth-2 neural networks

Sho Sonoda [a,*], Isao Ishikawa [b,a], Masahiro Ikeda [a]

[a] RIKEN Center for Advanced Intelligence Project (AIP), 1-4-1 Nihonbashi, Chuo-ku, 103-0027, Tokyo, Japan
[b] Center for Data Science, Ehime University (CDSE), 3 Bunkyocho, Matsuyama, 790-8577, Ehime, Japan

ARTICLE INFO

MSC:
68T07
42C40
43A85

Keywords:
Neural network
d-plane ridgelet transform
Fourier slice theorem
Group convolution
Noncompact symmetric space

ABSTRACT

To investigate neural network parameters, it is easier to study the distribution of parameters than to study the parameters in each neuron. The ridgelet transform is a pseudo-inverse operator that maps a given function f to the parameter distribution γ so that a network $\mathtt{NN}[\gamma]$ reproduces f, i.e. $\mathtt{NN}[\gamma] = f$. For depth-2 fully-connected networks on a Euclidean space, the ridgelet transform has been discovered up to the closed-form expression, thus we could describe how the parameters are distributed. However, for a variety of modern neural network architectures, the closed-form expression has not been known. In this paper, we explain a systematic method using Fourier expressions to derive ridgelet transforms for a variety of modern networks such as networks on finite fields $\mathbb{F}_p$, group convolutional networks on abstract Hilbert space $\mathcal{H}$, fully-connected networks on noncompact symmetric spaces G/K, and pooling layers, or the d-plane ridgelet transform.

1. Introduction

Neural networks are learning machines that support today's AI technology. Mathematically, they are nonlinear functions determined by a network of functions with learnable parameters (called *neurons*) connecting in parallel and series. Since the learning process is automated, we do not fully understand the parameters obtained through learning. An integral representation is a powerful tool for mathematical analysis of these parameters. One of the technical difficulties in analyzing the behavior of neural networks is that their parameters are extremely nonlinear. An integral representation is a method of indirectly analyzing the parameters through their distribution, rather than directly analyzing the parameters of each neuron. The set of all the signed (or probability) parameter distributions forms a linear (or convex) space, making it possible to perform far more insightful analysis than directly analyzing individual parameters.

For instances, characterization of neural network parameters such as the *ridgelet transform* (Murata, 1996; Candès, 1998; Sonoda and Murata, 2017) and the *representer theorems* for ReLU networks (Savarese et al., 2019; Ongie et al., 2020; Parhi and Nowak, 2021; Unser, 2019), and convergence analysis of stochastic gradient descent (SGD) for deep learning such as the *mean field theory* (Nitanda and Suzuki, 2017; Chizat and Bach, 2018; Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018; Sirignano and Spiliopoulos, 2020) and the *infinite-dimensional Langevin dynamics* (Suzuki, 2020; Nitanda et al., 2022), have been developed using integral representations.

* Corresponding author.
E-mail address: sho.sonoda@riken.jp (S. Sonoda).

https://doi.org/10.1016/j.jspi.2024.106184
Received 1 September 2023; Received in revised form 25 February 2024; Accepted 11 April 2024
Available online 15 April 2024

1.1. Integral representation

The integral representation of a *depth-2 fully-connected neural network* is defined as below.

Definition 1.1. Let $\sigma : \mathbb{R} \to \mathbb{C}$ be a measurable function, called *activation function*, and fix it. For any signed measure γ on $\mathbb{R}^m \times \mathbb{R}$, called a *parameter distribution*, we define the *integral representation of depth-2 fully-connected neural network* as

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

Here, for each hidden parameter $(\boldsymbol{a}, b)$, feature map $\boldsymbol{x} \mapsto \sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)$ corresponds to a single hidden neuron with activation function σ, weight $\gamma(\boldsymbol{a}, b)$ corresponds to an output coefficient, and the integration implies that all the possible neurons are assigned in advance. Since the *only* free parameter is *parameter distribution* γ, we can identify network $S[\gamma]$ with point γ in a function space.

We note that this representation covers *both* infinite (or continuous) and finite widths. Indeed, while the integration may be understood as an infinite width layer, the integration with a finite sum of point masses such as $\gamma_p := \sum_{i=1}^p c_i \delta_{(\boldsymbol{a}_i, b_i)}$ can represent a finite width layer:

$$S[\gamma_p](\boldsymbol{x}) = \sum_{i=1}^p c_i \sigma(\boldsymbol{a}_i \cdot \boldsymbol{x} - b_i) = C\sigma(A\boldsymbol{x} - \boldsymbol{b}), \quad \boldsymbol{x} \in \mathbb{R}^m$$

where the third term is the so-called "matrix" representation with matrices $A \in \mathbb{R}^{p \times m}, C \in \mathbb{R}^{1 \times p}$ and vector $\boldsymbol{b} \in \mathbb{R}^p$ followed by component-wise activation σ. Singular measures as above can be mathematically justified without any inconsistency if the class of the parameter distributions are set to a class of Borel measures or Schwartz distributions.

There are at least four advantages to introducing integral representations:

1. Aggregation of parameters, say $\{(\boldsymbol{a}_i, b_i, c_i)\}_{i=1}^p$, into a single function (parameter distribution) $\gamma(\boldsymbol{a}, b)$,
2. Ability to represent finite models and continuous models in the same form,
3. Linearization of networks and convexification of learning problems, and
4. Presence of the ridgelet transform.

Advantages 1 and 2 have already been explained. Advantage 3 is described in the next subsection, and Advantage 4 is emphasized throughout the paper. On the other hand, there are two disadvantages:

1. Extensions to deep networks are hard[1], and
2. Advanced knowledge on functional analysis are required.

1.2. Linearization and convexification effect

The third advantage of the integral representation is the so-called *linearization* (and *convexification*) tricks. That is, while the network is *nonlinear* with respect to the raw parameters $\boldsymbol{a}$ and b, namely,

$$S[\delta_{(\alpha_1 \boldsymbol{a}_1 + \alpha_2 \boldsymbol{a}_2, b)}] \neq \alpha_1 S[\delta_{(\boldsymbol{a}_1, b)}] + \alpha_2 S[\delta_{(\boldsymbol{a}_2, b)}], \quad \alpha_1, \alpha_2 \in \mathbb{C}$$

(and similarly for b), it is *linear* with respect to the parameter distribution γ, namely,

$$S[\alpha_1\gamma_1 + \alpha_2\gamma_2] = \alpha_1 S[\gamma_1] + \alpha_2 S[\gamma_2], \quad \alpha_1, \alpha_2 \in \mathbb{C}.$$

Furthermore, linearizing neural networks leads to convexifying learning problems. Specifically, for a convex function $\ell : \mathbb{R} \to \mathbb{R}$, the loss function defined as $L[\gamma] := \ell(S[\gamma])$ satisfies the following:

$$L[t\gamma_1 + (1-t)\gamma_2] \leq tL[\gamma_1] + (1-t)L[\gamma_2], \quad t \in [0, 1].$$

It may sound paradoxical that a convex loss function on a function space has local minima in raw parameters, but we can understand this through the chain rule for functional derivative: Suppose that a parameter distribution γ is parametrized by a raw parameter, say θ, then

$$\frac{\partial L[\gamma(\theta)]}{\partial \theta} = \left\langle \frac{\partial \gamma(\theta)}{\partial \theta}, \frac{\partial L[\gamma]}{\partial \gamma} \right\rangle.$$

In other words, a local minimum ($\partial_\theta L = 0$) in raw parameter θ can arise not only from the global optimum ($\partial_\gamma L = 0$) but also from the case when two derivatives $\partial_\theta \gamma$ and $\partial_\gamma L$ are orthogonal.

The trick of lifting nonlinear objects in a linear space has been studied since the age of Frobenius, one of the founders of the linear representation theory of groups. In the context of neural network study, as well as the recent studies mentioned above, either the integral representation by Barron (1993) or the convex neural network by Bengio et al. (2006) are often referred. In the context of

[1] Good News: After the initial submission of this manuscript, the authors have successfully developed the ridgelet transform for *deep* networks in Sonoda et al. (2023a,b).

deep learning theory, this linearization/convexification trick has been employed to show the global convergence of the SGD training of shallow ReLU networks (Nitanda and Suzuki, 2017; Chizat and Bach, 2018; Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018; Sirignano and Spiliopoulos, 2020; Suzuki, 2020; Nitanda et al., 2022), and to characterize parameters in ReLU networks (Savarese et al., 2019; Ongie et al., 2020; Parhi and Nowak, 2021; Unser, 2019).

1.3. Ridgelet transform

The fourth advantage of the integral representation is the so-called the *ridgelet transform* R, or a right inverse operator of the integral representation operator S. For example, the ridgelet transform for depth-2 fully-connected network (1) is given as below.

Definition 1.2. For any measurable functions $f : \mathbb{R}^m \to \mathbb{C}$ and $\rho : \mathbb{R} \to \mathbb{C}$,

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

In principle, the *ridgelet function* ρ can be chosen independently of the activation function σ of neural network S. The following theorem holds.

Theorem 1.1 (*Reconstruction Formula*). *Suppose σ and ρ are a tempered distribution ($\mathcal{S}'$) on $\mathbb{R}$ and a rapidly decreasing function ($\mathcal{S}$) on $\mathbb{R}$, respectively. Then, for any square integrable function f, the following reconstruction formula*

$$S[R[f;\rho]] = (\!(\sigma,\rho)\!)f \quad \textit{in} \quad L^2(\mathbb{R}^m)$$

holds with the factor being a scalar product of σ and ρ,

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

where $\sharp$ denotes the Fourier transform.

From the perspective of neural network theory, the reconstruction formula claims a detailed/constructive version of the universal approximation theorem. That is, given any target function f, as long as $(\!(\sigma,\rho)\!) \neq 0$, the network $S[\gamma]$ with coefficient $\gamma = R[f;\rho]$ reproduces the original function, and the coefficient is given explicit.

From the perspective of functional analysis, on the other hand, the reconstruction formula states that R and S are analysis and synthesis operators, and thus play the same roles as, for instance, 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 $F^{-1}[F[f]] = f$.

Despite the common belief that neural network parameters are a blackbox, the closed-form expression (2) of ridgelet transform clearly describes how the network parameters are distributed, which is a clear advantage of the integral representation theory (see e.g. Sonoda et al., 2021b). Moreover, the integral representation theory can deal with a wide range of activation functions without approximation, not only ReLU but all the tempered distribution $\mathcal{S}'(\mathbb{R})$ (see e.g. Sonoda and Murata, 2017).

The ridgelet transform is discovered in the late 1990s independently by Murata (1996) and Candès (1998). The term "ridgelet" is named by Candès, based on the facts that the graph of a function $\boldsymbol{x}\mapsto\rho(\boldsymbol{a}\cdot\boldsymbol{x}-b)$ is ridge-shaped, and that the integral transform R can be regarded as a multidimensional counterpart of the wavelet transform.

In fact, the ridgelet transform can be decomposed into the composite of wavelet transform after the Radon transform, namely,

$$R[f;\rho](a\boldsymbol{u},b) = \int_{\mathbb{R}} P[f](\boldsymbol{u},t)\overline{\rho(at-b)}\mathrm{d}a\mathrm{d}b, \quad (a,\boldsymbol{u},b)\in\mathbb{R}\times\mathbb{S}^{m-1}\times\mathbb{R},$$
$$P[f](\boldsymbol{u},t) := \int_{(\mathbb{R}\boldsymbol{u})^\perp} f(t\boldsymbol{u}+\boldsymbol{y})\mathrm{d}\boldsymbol{y}, \quad (\boldsymbol{u},t)\in\mathbb{S}^{m-1}\times\mathbb{R},$$

where $\mathbb{S}^{m-1}$ denotes the m-dimensional unit sphere, $(\mathbb{R}\boldsymbol{u})^\perp \cong \mathbb{R}^{m-1}$ denotes the orthocomplement of the normal vector $\boldsymbol{u}\in\mathbb{S}^{m-1}$, $\mathrm{d}\boldsymbol{y}$ denotes the Hausdorff measure on $(\mathbb{R}\boldsymbol{u})^\perp$ or the Lebesgue measure on $\mathbb{R}^{m-1}$, and $\boldsymbol{a}\in\mathbb{R}^m$ is represented in polar coordinates $\boldsymbol{a} = a\boldsymbol{u}$ with $(a,\boldsymbol{u})\in\mathbb{R}\times\mathbb{S}^{m-1}$ allowing the double covering: $(a,\boldsymbol{u}) = (-a,-\boldsymbol{u})$ (see Sonoda and Murata, 2017, for the proof). Therefore, several authors have remarked that *ridgelet analysis is wavelet analysis in the Radon domain* (Donoho, 2002; Kostadinova et al., 2014; Starck et al., 2010).

In the context of deep learning theory, Savarese et al. (2019), Ongie et al. (2020), Parhi and Nowak (2021) and Unser (2019) investigate the ridgelet transform for the specific case of fully-connected ReLU layers to establish the representer theorem. Sonoda et al. (2021b) have shown that the parameter distribution of a finite model trained by regularized empirical risk minimization (RERM) converges to the ridgelet spectrum $R[f;\rho]$ in an over-parametrized regime, meaning that we can understand the parameters at local minima to be a finite approximation of the ridgelet transform. In other words, analyzing neural network parameters can be turned to analyzing the ridgelet transform.

1.4. Scope and contribution of this study

On the other hand, one of the major shortcomings of ridgelet analysis is that the closed-form expression is known for relatively small class of networks. Indeed, until Sonoda et al. (2022a,b), it was known only for depth-2 fully-connected layer: $\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)$. In the age of deep learning, a variety of layers have become popular such as the convolution and pooling layers (Fukushima, 1980; LeCun et al., 1998; Ranzato et al., 2007; Krizhevsky et al., 2012). Furthermore, the fully-connected layers on manifolds have also been developed such as the hyperbolic network (Ganea et al., 2018; Shimizu et al., 2021). Since the conventional ridgelet transform was discovered heuristically in the 1990s, and the derivation heavily depends on the specific structure of affine map $\boldsymbol{a}\cdot\boldsymbol{x}-b$, the ridgelet transforms for those modern architectures have been unknown for a long time.

In this study, we explain a systematic method to find the ridgelet transforms via the *Fourier expression* of neural networks, and obtain *new ridgelet transforms* in a unified manner. The Fourier expression of $S[\gamma]$ is essentially a change-of-frame from neurons $\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)$ to plane waves (or harmonic oscillators) $\exp(i\boldsymbol{x}\cdot\boldsymbol{\xi})$. Since the Fourier transform is extensively developed on a variety of domains, once a network $S[\gamma]$ is translated into a Fourier expression, we can systematically find a particular coefficient γ_f satisfying $S[\gamma_f]=f$ via the Fourier inversion formula. In fact, the traditional ridgelet transform is re-discovered. Associated with the change-of-frame in $S[\gamma]$, the ridgelet transform $R[f]$ is also given a Fourier expression, but this form is known as the *Fourier slice theorem* of ridgelet transform $R[f]$ (see e.g. Kostadinova et al., 2014). Hence, we call our proposed method as the *Fourier slice method*.

Besides the classical networks, we deal with *four* types of networks:

1. Networks on finite fields $\mathbb{F}_p$ in Section 3,
2. Group convolution networks on Hilbert spaces $\mathcal{H}$ in Section 4,
3. Fully-connected networks on noncompact symmetric spaces G/K in Section 5, and
4. Pooling layers (also known as the *d-plane ridgelet transform*) in Section 6.

The first three cases are already published thus we only showcase them, while the last case (pooling layer and d-plane ridgelet) involves *new* results.

For all the cases, the reconstruction formula $S[R[f]]=f$ is understood as a constructive proof of the *universal approximation theorem* for corresponding networks. The group convolution layer case widely extends the ordinary convolution layer with periodic boundary, which is also the main subject of the so-called *geometric deep learning* (Bronstein et al., 2021). The case of fully-connected layer on symmetric spaces widely extends the recently emerging concept of *hyperbolic networks* (Ganea et al., 2018; Gulcehre et al., 2019; Shimizu et al., 2021), which can be cast as another geometric deep learning. The pooling layer case includes the original fully-connected layer and the pooling layer; and the corresponding ridgelet transforms include previously developed formulas such as the Radon transform formula by Savarese et al. (2019) and related to the previously developed "d-plane ridgelet transforms" by Rubin (2004) and Donoho (2001).

1.5. General notations

For any integer $d>0$, $\mathcal{S}(\mathbb{R}^d)$ and $\mathcal{S}'(\mathbb{R}^d)$ denote the classes of Schwartz test functions (or rapidly decreasing functions) and tempered distributions on $\mathbb{R}^d$, respectively. Namely, $\mathcal{S}'$ is the topological dual of $\mathcal{S}$. We note that $\mathcal{S}'(\mathbb{R})$ includes truncated power functions $\sigma(b)=b_+^k=\max\{b,0\}^k$ such as step function for $k=0$ and ReLU for $k=1$.

Fourier transform. To avoid potential confusion, we use two symbols $\widehat{\cdot}$ and $\cdot^\sharp$ for the Fourier transforms in $\boldsymbol{x}\in\mathbb{R}^m$ and $b\in\mathbb{R}$, respectively. For example,

$$\widehat{f}(\boldsymbol{\xi}) := \int_{\mathbb{R}^m} f(\boldsymbol{x})e^{-i\boldsymbol{x}\cdot\boldsymbol{\xi}}\mathrm{d}\boldsymbol{x}, \quad \boldsymbol{\xi}\in\mathbb{R}^m$$
$$\rho^\sharp(\omega) := \int_{\mathbb{R}} \rho(b)e^{-ib\omega}\mathrm{d}b, \quad \omega\in\mathbb{R}$$
$$\gamma^\sharp(\boldsymbol{a},\omega) = \int_{\mathbb{R}} \gamma(\boldsymbol{a},b)e^{-ib\omega}\mathrm{d}b, \quad (\boldsymbol{a},\omega)\in\mathbb{R}^m\times\mathbb{R}.$$

Furthermore, with a slight abuse of notation, when σ is a tempered distribution (i.e., $\sigma\in\mathcal{S}'(\mathbb{R})$), then $\sigma^\sharp$ is understood as the Fourier transform of distributions. Namely, $\sigma^\sharp$ is another tempered distribution satisfying $\int_{\mathbb{R}}\sigma^\sharp(\omega)\phi(\omega)\mathrm{d}\omega=\int_{\mathbb{R}}\sigma(\omega)\phi^\sharp(\omega)\mathrm{d}\omega$ for any test function $\phi\in\mathcal{S}(\mathbb{R})$. We refer to Grafakos (2008) for more details on the Fourier transform for distributions.

2. Method

We explain the basic steps to find the parameter distribution γ satisfying $S[\gamma]=f$. The basic steps is three-fold: (**Step 1**) Turn the network into the *Fourier expression*, (**Step 2**) *change variables* inside the feature map into principal and auxiliary variables, and (**Step 3**) put unknown γ in the *separation-of-variables form* to find a particular solution. In the following, we conduct the basic steps for the classical setting, i.e., the case of the fully-connected layer, for the explanation purpose. However, the "catch" of this procedure is that it is applicable to a wide range of networks as we will see in the subsequent sections.

2.1. Basic steps to solve $S[\gamma] = f$

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

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

Step 2. 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 feature map $\sigma^\sharp(\omega)e^{i\omega\boldsymbol{a}\cdot\boldsymbol{x}}$ splits into a product of a principal factor (in $\boldsymbol{\xi}$ and $\boldsymbol{x}$) and an auxiliary factor (in ω), namely

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

Now, we can see that the integration inside brackets $[\cdots]$ becomes the Fourier inversion with respect to $\boldsymbol{\xi}$ and $\boldsymbol{x}$.

Step 3. Because of the Fourier inversion, it is natural to assume that the unknown function γ has a *separation-of-variables form* as

$$\gamma^\sharp_{f,\rho}(\boldsymbol{\xi}/\omega, \omega) := \widehat{f}(\boldsymbol{\xi})\overline{\rho^\sharp(\omega)}, \tag{3}$$

with using an arbitrary function $\rho \in S(\mathbb{R})$. Namely, it is composed of a principal factor $\widehat{f}$ containing the target function f, and an auxiliary factor $\rho^\sharp$ set only for convergence of the integration in ω. Then, we have

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

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 $c = (\!(\sigma, \rho)\!) \in \mathbb{C}$.

Furthermore, $\gamma_{f,\rho}$ is the ridgelet transform because it is rewritten as

$$\gamma^\sharp_{f,\rho}(\boldsymbol{a}, \omega) = \widehat{f}(\omega\boldsymbol{a})\overline{\rho^\sharp(\omega)},$$

and thus calculated as

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

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

In conclusion, the separation-of-variables expression (3) is the way to naturally find the ridgelet transform. We note that Steps 1 and 2 can be understood as the *change-of-frame* from the frame spanned by neurons $\{\boldsymbol{x} \mapsto \sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b) \mid (\boldsymbol{a}, b) \in \mathbb{R}^m \times \mathbb{R}\}$, with which we are less familiar, to the frame spanned by (the tensor product of an auxiliary function and the) plane wave $\{\boldsymbol{x} \mapsto \sigma^\sharp(\omega)e^{i\boldsymbol{\xi}\cdot\boldsymbol{x}} \mid (\boldsymbol{\xi}, \omega) \in \mathbb{R}^m \times \mathbb{R}\}$, with which we are much familiar. Hence, in particular, the map $\gamma(\boldsymbol{a}, b) \mapsto \gamma^\sharp(\boldsymbol{a}/\omega, \omega)|\omega|^{-m}$ can be understood as the associated coordinate transformation.

3. Case I: NN on finite field $\mathbb{F}_p := \mathbb{Z}/\mathbb{Z}_p$

As one of the simplest applications, we showcase the results by Yamasaki et al. (2023), a neural network on the finite field $\mathbb{F}_p := \mathbb{Z}/p\mathbb{Z} \cong \{0, 1, \ldots, p-1 \mod p\}$ (with prime number p). This study aimed to design a quantum algorithm that efficiently computes the ridgelet transform, and the authors developed this example based on the demand to represent all data and parameters in finite qubits. To be precise, the authors dealt with functions on discrete space $\mathbb{F}_p^m$, as a discretization of functions on a continuous space $\mathbb{R}^m$.

3.1. Fourier transform

For any positive integers n, m, let $\mathbb{Z}_n^m := (\mathbb{Z}/n\mathbb{Z})^m$ denote the product of cyclic groups. We note that the set of all real functions f on $\mathbb{Z}_n^m$ is identified with the $(n \times m)$-dimensional real vector space, i.e. $\{f : \mathbb{Z}_n^m \to \mathbb{R}\} \cong \mathbb{R}^{n \times m}$, because each value $f(i,j)$ of function f at $(i,j) \in \mathbb{Z}_n^m$ can be identified with the (i,j)th component a_{ij} of vector $\boldsymbol{a} = (a_{ij}) \in \mathbb{R}^{n \times m}$. In particular, $L^2(\mathbb{Z}_n^m) = \mathbb{R}^{n \times m}$.

We use the Fourier transform on a product of cyclic groups as below.

Definition 3.1 (*Fourier Transform on* $\mathbb{Z}_n^m$)**.** For any $f \in L^2(\mathbb{Z}_n^m)$, put

$$\widehat{f}(\boldsymbol{\xi}) := \sum_{\boldsymbol{x} \in \mathbb{Z}_n^m} f(\boldsymbol{x}) e^{-2\pi i \boldsymbol{\xi} \cdot \boldsymbol{x}/n}, \quad \boldsymbol{\xi} \in \mathbb{Z}_n^m.$$

Theorem 3.1 (*Inversion Formula*)**.** *For any* $f \in L^2(\mathbb{Z}_n^m)$,

$$f(\boldsymbol{x}) = \frac{1}{|\mathbb{Z}_n^m|} \sum_{\boldsymbol{\xi} \in \mathbb{Z}_n^m} \widehat{f}(\boldsymbol{\xi}) e^{2\pi i \boldsymbol{\xi} \cdot \boldsymbol{x}/n}, \quad \boldsymbol{x} \in \mathbb{Z}_n^m.$$

The proof is immediate from the so-called *orthogonality of characters*, an identity $\sum_{g \in \mathbb{Z}_n} e^{2\pi i g(t-s)/n} = |\mathbb{Z}_n| \delta_{ts}$ $(t, s \in \mathbb{Z}_n)$, where δ_{ts} being the Kronecker's δ.

We note that despite the Fourier transform itself can be defined on any cyclic group $\mathbb{Z}_n$, namely n needs not be prime, a finite field $\mathbb{F}_p (= \mathbb{Z}_p)$ is assumed to perform the change-of-variables step.

3.2. Network design

Remarkably, the $\mathbb{F}_p$-version of arguments is obtained by formally replacing every integration in the $\mathbb{R}$-version of arguments with summation.

Definition 3.2 (*NN on* $\mathbb{F}_p^m$)**.** For any functions $\gamma \in L^2(\mathbb{F}_p^m \times \mathbb{F}_p)$ and $\sigma \in L^\infty(\mathbb{F}_p)$, put

$$S[\gamma](\boldsymbol{x}) := \sum_{(\boldsymbol{a},b) \in \mathbb{F}_p^m \times \mathbb{F}_p} \gamma(\boldsymbol{a}, b) \sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b), \quad \boldsymbol{x} \in \mathbb{F}_p^m$$

Again, in Yamasaki et al. (2023), it is introduced as a discretized version of a function on a continuous space $\mathbb{R}^m$.

3.3. Ridgelet transform

Theorem 3.2 (*Reconstruction Formula*)**.** *For any function* $\rho \in L^\infty(\mathbb{F}_p)$*, put*

$$R[f;\rho](\boldsymbol{a}, b) := \sum_{\boldsymbol{x} \in \mathbb{F}_p^m} f(\boldsymbol{x}) \overline{\rho(\boldsymbol{a} \cdot \boldsymbol{x} - b)}, \quad (\boldsymbol{a}, b) \in \mathbb{F}_p^m \times \mathbb{F}_p$$

$$(\!(\sigma, \rho)\!) := \frac{1}{|\mathbb{F}_p|^{m-1}} \sum_{\omega \in \mathbb{F}_p} \sigma^\sharp(\omega) \overline{\rho^\sharp(\omega)}.$$

Then, for any $f \in L^2(\mathbb{F}_p^m)$,

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

In other words, the fully-connected network on finite field $\mathbb{F}_p^m$ can strictly represent any square integrable function on $\mathbb{F}_p^m$.

Finally, the following proof shows that a new example of neural networks can be obtained by systematically following the same three steps as in the original arguments.

Sketch Proof. ***Step 1***. Turn to the Fourier expression:

$$\begin{aligned} S[\gamma](\boldsymbol{x}) &:= \sum_{(\boldsymbol{a},b) \in \mathbb{F}_p^m \times \mathbb{F}_p} \gamma(\boldsymbol{a}, b) \sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b) \\ &= \frac{1}{|\mathbb{F}_p|} \sum_{(\boldsymbol{a},\omega) \in \mathbb{F}_p^m \times \mathbb{F}_p} \gamma^\sharp(\boldsymbol{a}, \omega) \sigma(\omega) e^{2\pi i \omega \boldsymbol{a} \cdot \boldsymbol{x}/p} \end{aligned}$$

Step 2. Change variables $\boldsymbol{\xi} = \omega \boldsymbol{a}$

$$= \frac{1}{|\mathbb{F}_p|} \sum_{(\boldsymbol{\xi},\omega) \in \mathbb{F}_p^m \times \mathbb{F}_p} \gamma^\sharp(\boldsymbol{\xi}/\omega, \omega) \sigma(\omega) e^{2\pi i \boldsymbol{\xi} \cdot \boldsymbol{x}/p}$$

Step 3. Put separation-of-variables form $\gamma^\sharp(\boldsymbol{\xi}/\omega, \omega) = \widehat{f}(\boldsymbol{\xi}) \overline{\rho^\sharp(\omega)}$

$$= \left(|\mathbb{F}_p|^{m-1} \sum_{\omega \in \mathbb{F}_p} \sigma^\sharp(\omega) \overline{\rho^\sharp(\omega)} \right) \left(\frac{1}{|\mathbb{F}_p|^m} \sum_{\xi \in \mathbb{F}_p^m} \widehat{f}(\xi) e^{2\pi i \xi \cdot x/p} \right)$$
$$= (\!(\sigma, \rho)\!) f(\boldsymbol{x}),$$

and we can verify $\gamma = R[f;\rho]$. □

4. Case II: Group convolutional NN on Hilbert space $\mathcal{H}$

Next, we showcase the results by Sonoda et al. (2022b). Since there are various versions of convolutional neural networks (CNNs), their approximation properties (such as the universality) have been investigated individually depending on the network architecture. The method presented here defines the generalized group convolutional neural network (GCNN) that encompasses a wide range of CNNs, and provides a powerful result by unifying the expressivity analysis in a constructive and simple manner by using ridgelet transforms.

4.1. Fourier transform

Since the input to CNNs is a signal (or a function), the Fourier transform corresponding to a *naive* integral representation is the Fourier transform on the space of signals, which is typically an infinite-dimensional space $\mathcal{H}$ of functions. Although the Fourier transform on the infinite-dimensional Hilbert space has been well developed in the probability theory, the mathematics tends to become excessively advanced for the expected results. One of the important ideas of this study is to induce the Fourier transform of $\mathbb{R}^m$ in a *finite*-dimensional subspace $\mathcal{H}_m$ of $\mathcal{H}$ instead of using the Fourier transform on the entire space $\mathcal{H}$. To be precise, we simply take an m-dimensional orthonormal frame $F_m = \{h_i\}_{i=1}^m$ of $\mathcal{H}$, put the linear span $\mathcal{H}_m := \operatorname{span} F_m = \{\sum_{i=1}^m c_i h_i \mid c_i \in \mathbb{R}\}$, and identify each element $f = \sum_{i=1} c_i h_i \in \mathcal{H}_m \subset \mathcal{H}$ with point $\boldsymbol{c} = (c_1, \dots, c_m) \in \mathbb{R}^m$. Obviously, this embedding depends on the choice of m-frame F_m, yet drastically simplifies the main theory itself.

Definition 4.1 (*Fourier Transform on a Hilbert Space $\mathcal{H}_m \subset \mathcal{H}$*)**.** Let $\mathcal{H}$ be a Hilbert space, $\mathcal{H}_m \subset \mathcal{H}$ be an m-dimensional subspace, and λ be the Lebesgue measure induced from $\mathbb{R}^m$. Put

$$\widehat{f}(\xi) := \int_{\mathcal{H}_m} f(x) e^{-i\langle x, \xi \rangle} \mathrm{d}\lambda(x), \quad \xi \in \mathcal{H}_m$$

Then, obviously from the construction, we have the following.

Theorem 4.1. *For any $f \in L^2(\mathcal{H}_m)$,*

$$\frac{1}{(2\pi)^m} \int_{\mathcal{H}_m} \widehat{f}(\xi) e^{i\langle x, \xi \rangle} \mathrm{d}\lambda(\xi) = f(x), \quad x \in \mathcal{H}_m$$

4.2. Network design

Another important idea is to deal with various group convolutions in a uniform manner by using the linear representation of groups.

Definition 4.2 (*Generalized Group Convolution*)**.** Let G be a group, $\mathcal{H}$ be a Hilbert space, and $T : G \to GL(\mathcal{H})$ be a group representation of G. The (G,T)-convolution is given by

$$(a * x)(g) := \langle T_{g^{-1}}[x], a \rangle_{\mathcal{H}}, \quad a, x \in \mathcal{H}.$$

As clarified in Sonoda et al. (2022b), the *generalized convolution* covers a variety of typical examples such as (1) classical group convolution $\int_G x(h) a(h^{-1}g) \mathrm{d}h$, (2) discrete cyclic convolution for multi-channel digital images, (3) DeepSets, or permutation equivariant maps, (4) continuous cyclic convolution for signals, and (5) E(n)-equivariant maps.

Definition 4.3 (*Group CNN*)**.** Let $\mathcal{H}_m \subset \mathcal{H}$ be an m-dimensional subspace equipped with the Lebesgue measure λ. Put

$$S[\gamma](x)(g) := \int_{\mathcal{H}_m \times \mathbb{R}} \gamma(a,b) \sigma((a * x)(g) - b) \mathrm{d}\lambda(a) \mathrm{d}b, \quad x \in \mathcal{H}, g \in G$$

Here, the integration runs all the possible convolution filters a. For the sake of simplicity, however, the domain $\mathcal{H}_m$ of filters is restricted to an m-dimensional subspace of entire space $\mathcal{H}$.

4.3. Ridgelet transform

In the following, $e \in G$ denotes the identity element.

Definition 4.4 (*(G,T)-Equivariance*)**.** A (nonlinear) map $f : \mathcal{H} \to \mathbb{C}^G$ is (G,T)-equivariant when

$$f(T_g[x])(h) = f(x)(g^{-1}h), \quad \forall x \in \mathcal{H}_m, g, h \in G$$

We note that the proposed network is inherently (G,T)-equivariant

$$S[\gamma](T_g[x])(h) = S[\gamma](x)(g^{-1}h), \quad \forall x \in \mathcal{H},\ g, h \in G.$$

Definition 4.5 (*Ridgelet Transform*)**.** For any measurable functions $f : \mathcal{H}_m \to \mathbb{C}^G$ and $\rho : \mathbb{R} \to \mathbb{C}$, put

$$R[f;\rho](a,b) := \int_{\mathcal{H}_m} f(x)(e)\overline{\rho(\langle a,x\rangle_{\mathcal{H}} - b)}\mathrm{d}\lambda(x).$$

It is remarkable that the product of a and x inside the ρ is not convolution $a * x$ but scalar product $\langle a,x\rangle$. This is essentially because (1) f will be assumed to be group equivariant, and (2) the network is group equivariant by definition.

Theorem 4.2 (*Reconstruction Formula*)**.** *Suppose that f is (G,T)-equivariant and $f(\bullet)(e) \in L^2(\mathcal{H}_m)$, then $S[R[f;\rho]] = (\!(\sigma,\rho)\!)f$.*

In other words, a continuous GCNN can represent any square-integrable *group-equivariant* function. Again, the proof is performed by systematically following the three steps as below.

Sketch Proof. ***Step 1***. Turn to a Fourier expression:

$$\begin{aligned} S[\gamma](x)(g) &= \int_{\mathcal{H}_m\times\mathbb{R}} \gamma(a,b)\sigma(\langle T_{g^{-1}}[x],a\rangle_{\mathcal{H}} - b)\mathrm{d}a\mathrm{d}b \\ &= \frac{1}{2\pi}\int_{\mathcal{H}_m\times\mathbb{R}} \gamma^\sharp(a,\omega)\sigma^\sharp(\omega)e^{i\omega\langle T_{g^{-1}}[x],a\rangle_{\mathcal{H}}}\mathrm{d}a\mathrm{d}\omega \end{aligned}$$

Step 2. Change variables $(a,\omega) = (\xi/\omega,\omega)$ with $\mathrm{d}a\mathrm{d}\omega = |\omega|^{-m}\mathrm{d}\xi\mathrm{d}\omega$:

$$= \frac{1}{2\pi}\int_{\mathcal{H}_m\times\mathbb{R}} \gamma^\sharp(\xi/\omega,\omega)\sigma^\sharp(\omega)e^{i\langle T_{g^{-1}}[x],\xi\rangle_{\mathcal{H}}}|\omega|^{-m}\mathrm{d}\xi\mathrm{d}\omega.$$

Step 3. Put separation-of-variables form $\gamma^\sharp_{f,\rho}(\xi/\omega,\omega) := \widehat{f}(\xi)(e)\overline{\rho^\sharp(\omega)}$

$$\begin{aligned} &= \frac{1}{2\pi}\int_{\mathcal{H}_m} \widehat{f}(\xi)(e)e^{i\langle T_{g^{-1}}[x],\xi\rangle_{\mathcal{H}}}\mathrm{d}\lambda(\xi)\int_{\mathbb{R}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-m}\mathrm{d}\omega \\ &= (\!(\sigma,\rho)\!)f(x)(g), \end{aligned}$$

and we can verify $\gamma_{f,\rho} = R[f;\rho]$. □

4.4. Literature in geometric deep learning

General convolution networks for geometric/algebraic domains have been developed for capturing the invariance/equivariance to the symmetry in a data-efficient manner (Bruna and Mallat, 2013; Cohen and Welling, 2016; Zaheer et al., 2017; Kondor and Trivedi, 2018; Cohen et al., 2019; Kumagai and Sannai, 2020). To this date, quite a variety of convolution networks have been proposed for grids, finite sets, graphs, groups, homogeneous spaces and manifolds. We refer to Bronstein et al. (2021) for a detailed survey on the so-called *geometric deep learning*.

Since a universal approximation theorem (UAT) is a corollary of a reconstruction formula, $S[R[f;\rho]] = (\!(\sigma,\rho)\!)f$, the 3-steps Fourier expression method have provided a variety of UATs for $\sigma(ax-b)$-type networks in a *unified manner*. Here, we remark that the UATs of individual convolution networks have already shown (Maron et al., 2019; Zhou, 2020; Yarotsky, 2022). However, in addition to above mentioned advantages, the *wide coverage* of activation functions σ is also another strong advantage. In particular, we do not need to rely on the specific features of ReLU, nor need to rely on Taylor expansions/density arguments/randomized assumptions to deal with nonlinear activation functions.

5. Case III: NN on noncompact symmetric space $X = G/K$

Then, we showcase the results by Sonoda et al. (2022a). When the data is known to be on a certain manifold, it is natural to consider developing NNs on the manifold, in order to explicitly incorporate the inductive bias. Since there are no such thing as the standard inner products or affine mappings on manifolds, various NNs have been proposed based on geometric considerations and implementation constraints. The main idea of this study is to start from the Fourier transform on a manifold and induce a NN on the manifold and its reconstruction formula. On compact groups such as spheres $\mathbb{S}^{m-1}$, the Fourier analysis is well known as the Peter–Weyl theorem, but in this study, the authors focused on noncompact symmetric spaces G/K such as hyperbolic space $\mathbb{H}^m$ and space $\mathbb{P}_m$ of positive definite matrices and developed NNs based on the Helgason–Fourier transform on noncompact symmetric space.

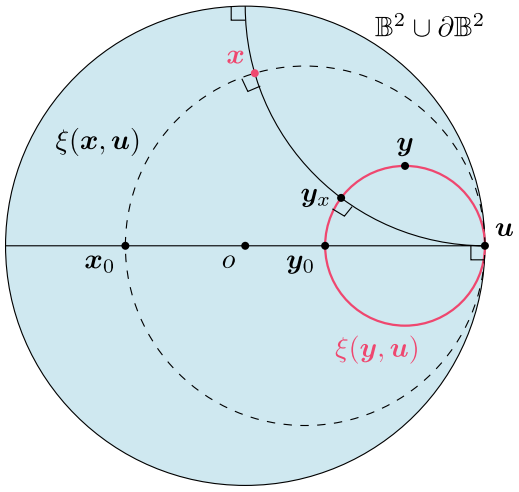

Fig. 1. Poincare Disk $\mathbb{B}^2$ is a noncompact symmetric space $SU(1,1)/SO(2)$. Poincaré disk $\mathbb{B}^2$, boundary $\partial\mathbb{B}^2$, point $\boldsymbol{x}$ (magenta), horocycle $\xi(\boldsymbol{y},\boldsymbol{u})$ (magenta) through point $\boldsymbol{y}$ tangent to the boundary at $\boldsymbol{u}$, 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 $\boldsymbol{y}_0$. See Appendix A for more details. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.1. Noncompact symmetric space G/K

We refer to Helgason (1984, Introduction) and Helgason (2008, Chapter III). A noncompact symmetric space is a homogeneous space G/K with nonpositive sectional curvature on which G acts transitively. Two important examples are hyperbolic space $\mathbb{H}^m$ (Fig. 1) and SPD manifold $\mathbb{P}_m$. See also Appendices A and B for more details on these spaces.

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.

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).

Definition 5.1. 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.

Example 5.1 (*Hyperbolic Space* $\mathbb{H}^m = SO^+(1,m)/O(m)$)**.** It is used for embedding words and tree-structured dataset.

Example 5.2 (*SPD Manifold* $\mathbb{P}_m = GL(m)/O(m)$)**.** It is a manifold of positive definite matrices, such as covariance matrices.

5.2. Boundary ∂X, horosphere ξ, and vector-valued composite distance $\langle x,u\rangle$

We further introduce three geometric objects in noncompact symmetric space G/X. In comparison to Euclidean space $\mathbb{R}^m$, the boundary ∂X corresponds to "the set of all infinite points" $\lim_{r\to+\infty}\{r\boldsymbol{u} \mid |\boldsymbol{u}| = 1, \boldsymbol{u} \in \mathbb{R}^m\}$, a horosphere ξ through point $x \in X$ with normal $u \in \partial X$ corresponds to a straight line $\boldsymbol{\xi}$ through point $\boldsymbol{x} \in \mathbb{R}^m$ with normal $\boldsymbol{u} \in \mathbb{S}^{m-1}$, and the vector distance $\langle x,u\rangle$ between origin $o \in X$ and horosphere $\xi(x,u)$ corresponds to the Riemannian distance between origin $\boldsymbol{0} \in \mathbb{R}^m$ and straight line $\boldsymbol{\xi}(\boldsymbol{x},\boldsymbol{u})$.

Definition 5.2. 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).

Definition 5.3. 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 boundaries.

Definition 5.4. 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.

5.3. Fourier transform

Based on the preparations so far, we introduce the Fourier transform on G/K, known as the Helgason–Fourier transform. 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.

Definition 5.5 (*Helgason–Fourier Transform*)**.** For any measurable function f on X, put

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

with a certain constant vector $\rho \in \mathfrak{a}^*$. Here, the exponent $(-i\lambda + \rho)\langle x, u\rangle$ is understood as the action of functional $-i\lambda + \rho \in \mathfrak{a}^*$ on a vector $\langle x, u\rangle \in \mathfrak{a}$.

This is understood as a "Fourier transform" because $e^{(-i\lambda+\rho)\langle x,u\rangle}$ is the eigenfunction of Laplace–Beltrami operator Δ_X on X.

Theorem 5.1 (*Inversion Formula*)**.** *For any square integrable function $f \in L^2(X)$,*

$$f(x) = \frac{1}{|W|} \int_{\mathfrak{a}^* \times \partial X} \widehat{f}(\lambda, u) e^{(i\lambda+\rho)\langle x,u\rangle} \frac{\mathrm{d}\lambda \mathrm{d}u}{|\boldsymbol{c}(\lambda)|^2}, \quad x \in X.$$

We refer to Helgason (2008, Theorems 1.3 and 1.5, Ch. III) for more details on the inversion formula.

5.4. Network design

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

Definition 5.6 (*NN on Noncompact Symmetric Space G/K*)**.** For any measurable functions $\sigma : \mathbb{R} \to \mathbb{C}$ and $\gamma : \mathfrak{a}^* \times \partial X \times \mathbb{R} \to \mathbb{C}$, put

$$S[\gamma](x) := \int_{\mathfrak{a}^* \times \partial X \times \mathbb{R}} \gamma(a, u, b) \sigma(a\langle x, u\rangle - b) e^{\rho\langle x,u\rangle} \mathrm{d}a\mathrm{d}u\mathrm{d}b, \quad x \in G/K.$$

Remarkably, the scalar product $\boldsymbol{a} \cdot \boldsymbol{x}$ (or $a\boldsymbol{u} \cdot \boldsymbol{x}$ in polar coordinate) in the Euclidean setting is replaced with a distance function $a\langle x, u\rangle$ in the manifold setting. In Sonoda et al. (2022a), the authors instantiate two important examples as below.

Example 5.3 (*Continuous Horospherical Hyperbolic NN*)**.** On the *Poincare ball model* $\mathbb{B}^m := \{\boldsymbol{x} \in \mathbb{R}^m \mid |\boldsymbol{x}| < 1\}$ equipped with the Riemannian metric $\mathfrak{g} = 4(1 - |\boldsymbol{x}|)^{-2} \sum_{i=1}^m \mathrm{d}x_i \otimes \mathrm{d}x_i$,

$$S[\gamma](\boldsymbol{x}) := \int_{\mathbb{R} \times \partial \mathbb{B}^m \times \mathbb{R}} \gamma(a, \boldsymbol{u}, b) \sigma(a\langle \boldsymbol{x}, \boldsymbol{u}\rangle - b) e^{\rho\langle \boldsymbol{x},\boldsymbol{u}\rangle} \mathrm{d}a\mathrm{d}\boldsymbol{u}\mathrm{d}b, \quad \boldsymbol{x} \in \mathbb{B}^m$$

$$\rho = (m-1)/2, \ \langle \boldsymbol{x}, \boldsymbol{u}\rangle = \log\left(\frac{1-|\boldsymbol{x}|_E^2}{|\boldsymbol{x}-\boldsymbol{u}|_E^2}\right), \quad (\boldsymbol{x}, \boldsymbol{u}) \in \mathbb{B}^m \times \partial\mathbb{B}^m$$

Example 5.4 (*Continuous Horospherical SPD Net*)**.** On the SPD manifold $\mathbb{P}_m$,

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

$\boldsymbol{\rho} = (-\frac{1}{2},\dots,-\frac{1}{2},\frac{m-1}{4})$, $\langle \boldsymbol{x},\boldsymbol{u}\rangle = \frac{1}{2}\log\lambda\left(uxu^{\top}\right)$, $\quad (x,u)\in\mathbb{P}_m\times\partial\mathbb{P}_m$ where $\lambda(x)$ denotes the diagonal in the *Cholesky decomposition* of x.

5.5. Ridgelet transform

Definition 5.7 (*Ridgelet Transform*)**.** For any measurable functions $f : X\to\mathbb{C}$ and $\rho:\mathbb{R}\to\mathbb{C}$, put

$$\begin{aligned}
R[f;\rho](a,u,b) &:= \int_X \boldsymbol{c}[f](x)\overline{\rho(a\langle x,u\rangle - b)}e^{\rho\langle x,u\rangle}\mathrm{d}x,\\
\boldsymbol{c}[f](x) &:= \int_{\mathfrak{a}^*\times\partial X}\widehat{f}(\lambda,u)e^{(i\lambda+\rho)\langle x,u\rangle}\frac{\mathrm{d}\lambda\mathrm{d}u}{|W|\,|\boldsymbol{c}(\lambda)|^4},\\
(\!(\sigma,\rho)\!) &:= \frac{|W|}{2\pi}\int_{\mathbb{R}}\sigma^{\sharp}(\omega)\overline{\rho^{\sharp}(\omega)}|\omega|^{-r}\mathrm{d}\omega.
\end{aligned}$$

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

Theorem 5.2 (*Reconstruction Formula*)**.** *Let $\sigma\in\mathcal{S}'(\mathbb{R})$, $\rho\in\mathcal{S}(\mathbb{R})$. Then, for any square integrable function f on X, we have*

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

In other words, the fully-connected network on noncompact symmetric space G/K can represent any square-integrable function. Again, the proof is performed by systematically following the three steps as below.

Sketch Proof. We identify the scale parameter $a\in\mathfrak{a}^*$ with vector $\boldsymbol{a}\in\mathbb{R}^r$. ***Step 1***. Turn to a Fourier expression:

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

Step 2. Change variables $(\boldsymbol{a},\omega) = (\lambda/\omega,\omega)$ with $\mathrm{d}\boldsymbol{a}\mathrm{d}\omega = |\omega|^{-r}\mathrm{d}\lambda\mathrm{d}\omega$:

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

Step 3. Put separation-of-variables form $\gamma^{\sharp}_{f,\rho}(\lambda/\omega,u,\omega) = \widehat{f}(\lambda,u)\overline{\rho^{\sharp}(\omega)}|\boldsymbol{c}(\lambda)|^{-2}$

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

and we can verify $\gamma_{f,\rho} = R[f;\rho]$. □

5.6. Literature in hyperbolic neural networks

The hyperbolic neural network (HNN) (Ganea et al., 2018; Gulcehre et al., 2019; Shimizu et al., 2021) is another emerging direction of geometric deep learning, inspired by the empirical observations that some datasets having tree or hierarchical structure can be efficiently embedded into hyperbolic spaces (Krioukov et al., 2010; Nickel and Kiela, 2017, 2018; Sala et al., 2018). We note that designing a FC layer $\sigma(\langle a,x\rangle - b)$ on manifold X is less trivial, because neither *scalar product* $\langle a,x\rangle$, *bias subtraction* $-b$, nor *elementwise activation* of σ is trivially defined on X in general, and thus we have to face those primitive issues.

The design concept of the original HNN (Ganea et al., 2018) is to reconstruct basic operations in the ordinary neural networks such as linear maps, bias translations, pointwise nonlinearities and softmax layers by using the Gyrovector operations in a tractable and geometric manner. For example, in HNN++ by Shimizu et al. (2021), the Poincaré multinomial logistic regression (MLR) layer $v(\boldsymbol{x})$ and fully-connected (FC) layer $\mathcal{F}(\boldsymbol{x})$, corresponding to the 1-affine layer $\boldsymbol{a}\cdot\boldsymbol{x} - b$ and k-affine layer $A^{\top}\boldsymbol{x} - \boldsymbol{b}$ without activation respectively, are designed as nonlinear maps $v:\mathbb{H}^m\to\mathbb{R}$ and $\mathcal{F}:\mathbb{H}^m\to\mathbb{H}^n$ so that $v(\boldsymbol{x};\boldsymbol{a},b)$ coincides with the distance between output $\boldsymbol{y} = \mathcal{F}(\boldsymbol{x};\{\boldsymbol{a}_i,b_i\}_{i\in[n]})$ and Poincaré hyperplane $H(o,\boldsymbol{e}_{\boldsymbol{a},b})$. Here, a Poincaré hyperplane $H(\boldsymbol{x},\boldsymbol{z})$ is defined as the collection of all geodesics through point $\boldsymbol{x}$ and normal to $\boldsymbol{z}$. Furthermore, they are designed so that the discriminative hyperplane coincides a Poincaré hyperplane. The nonlinear activation function $\sigma:\mathbb{R}^m\to\mathbb{R}^n$ is cast as a map $\sigma:\mathbb{H}^m\to\mathbb{H}^k$ via lifting $\sigma(\boldsymbol{x}) := \exp_0\circ\sigma\circ\log_0(\boldsymbol{x})$ for any $\boldsymbol{x}\in\mathbb{H}^m$. However, in practice, activation can be omitted since the FC layer is inherently nonlinear.

In this study, on the other hand, we take X to be a *noncompact symmetric space* G/K, which is a generalized version of the hyperbolic space $\mathbb{H}^m$. Following the philosophy of the Helgason–Fourier transform, we regard the scalar product $\boldsymbol{u}\cdot\boldsymbol{x}$ of unit vector $\boldsymbol{u}\in\mathbb{S}^{m-1}$ and point $\boldsymbol{x}\in\mathbb{R}^m$ as the signed distance between the origin $\boldsymbol{o}$ and plane $\xi(\boldsymbol{x},\boldsymbol{u})$ through point $\boldsymbol{x}$ with normal $\boldsymbol{u}$. Then,

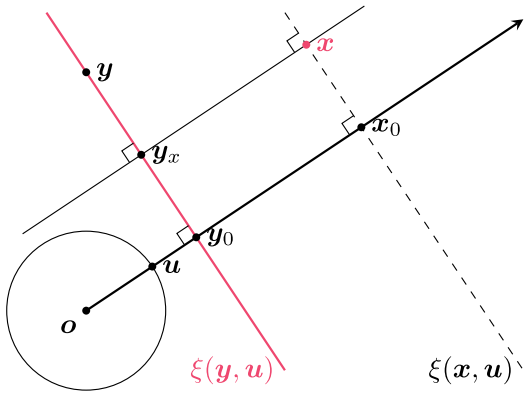

Fig. 2. The Euclidean fully-connected layer $\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)$ is recast as the signed distance $d(\boldsymbol{x},\xi)$ from a point $\boldsymbol{x}$ to a hyperplane $\xi(\boldsymbol{y},\boldsymbol{u})$ followed by nonlinearity $\sigma(r\bullet)$, where $\boldsymbol{y}$ satisfies $r\boldsymbol{y}\cdot\boldsymbol{u}=b$ and $\xi(\boldsymbol{y},\boldsymbol{u})$ passes through the point $\boldsymbol{y}$ with normal $\boldsymbol{u}$.

we recast it to the vector-valued distance, denoted $\langle u,x\rangle$, between the origin o and horocycle $\xi(x,u)$ through point x with normal u. As a result, we can naturally define bias subtraction $-b$ and elementwise activation of $\sigma:\mathbb{R}\to\mathbb{R}$ because the signed distance is identified with a vector.

More geometrically, $\boldsymbol{u}\cdot\boldsymbol{x}-b$ in $\mathbb{R}^m$ is understood as the distance between point $\boldsymbol{x}$ and plane ξ satisfying $\boldsymbol{u}\cdot\boldsymbol{x}-b=0$ (see Fig. 2). Similarly, $\langle u,x\rangle-b$ is understood as the distance between point x and horocycle ξ satisfying $\langle u,x\rangle-b=0$. Hence, as a general principle, we may formulate a versatile template of affine layers on X as

$$S[\gamma](x) := \int_{\mathbb{R}\times\Xi} \gamma(a,\xi)\sigma(ad(x,\xi))\mathrm{d}a\mathrm{d}\xi. \tag{4}$$

For example, in the original HNN, the Poincaré hyperplane H is employed as the geometric object. 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.

The strengths of our results are summarized as that we obtained the ridgelet transform in a unified manner for a wide class of input domain X in a geometric manner, i.e., independent of the coordinates; in particular, that it is the first result to define the neural network and obtained the ridgelet transform on noncompact space.

6. Case IV: Pooling layer and d-plane ridgelet transform

Finally, we present several new results. Technically, we consider networks with *multi*variate activation functions $\sigma:\mathbb{R}^k\to\mathbb{C}$. In all the sections up to this point, we have considered *uni*variate activation function $\sigma:\mathbb{R}\to\mathbb{C}$ (i.e. $k=1$). In the context of neural networks, it is understood as a mathematical model of pooling layers such as

$$\sigma(\boldsymbol{b}) = \frac{1}{k}\sum_{i=1}^{k} b_i \quad \text{(average pooling)},$$
$$\sigma(\boldsymbol{b}) = \max_{i\in[k]}\{b_i\} \quad \text{(max pooling), and}$$
$$\sigma(\boldsymbol{b}) = |\boldsymbol{b}|_p \quad (\ell^p\text{-norm}).$$

Meanwhile, in the context of sparse signal processing in the 2000s such as Donoho (2001) and Rubin (2004) (see also Section 7.2), it can also be understood as the ridgelet transform corresponding to the so-called d-plane transform (see also Section 6.1).

As mentioned in Section 1.3, the ridgelet transforms have profound relations to the *Radon* and wavelet transforms. In the language of probability theory, a Radon transform is understood as a *marginalization*, and the traditional problem of Johann Radon is the inverse problem of reconstructing the original joint distribution from several marginal distributions. Hence, depending on the choice of variables to be marginalized, there are countless different Radon transforms. In other words, the Radon transform can also be a rich source for finding a variety of novel networks and ridgelet transforms. In this section, we derive the ridgelet transform corresponding to the d-plane transform. (Nonetheless, the proofs are shown by the 3-steps Fourier expression method.)

Additional notations

Let m,d,k be positive integers satisfying $m=d+k$; let $M_{m,k} := \{A\in\mathbb{R}^{m\times k} \mid \operatorname{rank}A=k\}$ be a set of all full-column-rank (i.e., injective) matrices equipped with the Lebesgue measure $\mathrm{d}A=\bigwedge_{ij}\mathrm{d}a_{ij}$; let $V_{m,k} := \{U=[\boldsymbol{u}_1,\dots,\boldsymbol{u}_k]\in\mathbb{R}^{m\times k} \mid U^\top U=I_k\}$ be the Stiefel manifold of orthonormal k-frames in $\mathbb{R}^m$ equipped with invariant measure $\mathrm{d}U$; let $O(k) := \{V\in\mathbb{R}^k \mid V^\top V=I_k\}$ be the orthogonal group in $\mathbb{R}^k$ equipped with invariant measure $\mathrm{d}V$. In addition, let $GV_{m,k} := \{A=aU\in\mathbb{R}^{m\times k} \mid a\in\mathbb{R}_+, U\in V_{m,k}\}$ be a

similitude group equipped with the product measure $\mathrm{d}a\mathrm{d}U$. For a rectangular matrix $A \in M_{m,k}$, we write $|\det A| := |\det A^\top A|^{1/2}$ for short. In the following, we use $\widehat{\cdot}$ and $\cdot^\sharp$ for the Fourier transforms in $\boldsymbol{x} \in \mathbb{R}^m$ and $\boldsymbol{b} \in \mathbb{R}^k$, respectively. For any $s \in \mathbb{R}$, let $\Delta^{s/2}$ denote the fractional Laplacian defined as a Fourier multiplier: $\Delta^{s/2}[f](\boldsymbol{x}) := \frac{1}{(2\pi)^m}\int_{\mathbb{R}^m} |\boldsymbol{\xi}|^s \widehat{f}(\boldsymbol{\xi}) e^{i\boldsymbol{\xi}\cdot\boldsymbol{x}} \mathrm{d}\boldsymbol{\xi}$.

6.1. d-plane transform

The d-plane transform is a Radon transform that marginalizes a d-dimensional affine subspace (d-plane) in an m-dimensional space. In the special cases when $d = m-1$ (hyperplane) and $d = 1$ (straight line), they are respectively called the (strict) Radon transform and the X-ray transform. The ridgelet transforms to be introduced in this section correspond to d-plane Radon transform, and the classical ridgelet transform corresponds to the strict Radon transform ($d = m-1$). We refer to Chapter 1 of Helgason (2010).

Definition 6.1 (*d-plane*)**.** A d-plane $\boldsymbol{\xi} \subset \mathbb{R}^m$ is a d-dimensional affine subspace in $\mathbb{R}^m$. Here, *affine* emphasizes that it does *not* always pass through the origin $\boldsymbol{o} \in \mathbb{R}^m$. Let $G_{m,d}$ denote the collection of all d-planes in $\mathbb{R}^m$, called the *affine Grassmannian manifold*.

A d-plane is parametrized by its orthonormal directions $U = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_k] \in V_{m,k}$ and coordinate vector $\boldsymbol{b} \in \mathbb{R}^k$ from the origin $\boldsymbol{o}$ as below

$$\boldsymbol{\xi}(U, \boldsymbol{b}) := U\boldsymbol{b} + \ker U = \sum_{i=1}^k b_i \boldsymbol{u}_i + \left\{ \sum_{j=1}^d c_j \boldsymbol{v}_j \middle| c_j \in \mathbb{R} \right\},$$

where $[\boldsymbol{v}_1, \dots, \boldsymbol{v}_d] \in V_{m,d}$ is a d-frame satisfying $\boldsymbol{v}_j \perp \boldsymbol{u}_i$ for any $\forall i, j$. The first term $U\boldsymbol{b}$ is the displacement vector from the origin $\boldsymbol{o}$, its norm $|U\boldsymbol{b}|$ is the distance from the origin $\boldsymbol{o}$ and d-plane $\boldsymbol{\xi}$, and the second term $\ker U = (\operatorname{span} U)^\perp$ is the d-dimensional linear subspace that is parallel to $\boldsymbol{\xi}$.

Recall that for each direction $U \in V_{m,d}$, the whole space $\mathbb{R}^m$ can be decomposed into a disjoint union of d-planes as $\mathbb{R}^m = \cup_{\boldsymbol{b}\in\mathbb{R}^k} \boldsymbol{\xi}(U, \boldsymbol{b})$. In this perspective, the d-plane transform of f at $\boldsymbol{\xi}$ is defined as a *marginalization* of f in $\boldsymbol{\xi}$.

Definition 6.2 (*d-plane Transform*)**.** For any integrable function $f \in L^1(\mathbb{R}^m)$ and d-plane $\boldsymbol{\xi} = (U, \boldsymbol{b}) \in V_{m,k} \times \mathbb{R}^k$, put

$$P_d[f](\boldsymbol{\xi}) := \int_{\boldsymbol{\xi}} f(\boldsymbol{x}) \mathrm{dL}^d(\boldsymbol{x}) = \int_{\ker U} f(U\boldsymbol{b} + \boldsymbol{y}) \mathrm{d}\boldsymbol{y},$$

where L^d is the d-dimensional Hausdorff measure on $\boldsymbol{\xi}$.

Particularly, the strict Radon transform corresponds to $d = m-1$, and the X-ray transform corresponds to $d = 1$.

The d-plane transform has the following Fourier expression.

Lemma 6.1 (*Fourier Slice Theorem for d-plane Transform*)**.** *For any $f \in L^1(\mathbb{R}^m)$,*

$$P_d[f]^\sharp(U, \boldsymbol{\omega}) = \widehat{f}(U\boldsymbol{\omega}), \quad (U, \boldsymbol{\omega}) \in V_{m,k} \times \mathbb{R}^k,$$

where $\widehat{\cdot}$ and $\cdot^\sharp$ denote the Fourier transforms in $\boldsymbol{x}$ and $\boldsymbol{b}$ respectively. In other words,

$$\int_{\mathbb{R}^k} P_d[f](U, \boldsymbol{b}) e^{-i\boldsymbol{b}\cdot\boldsymbol{\omega}} \mathrm{d}\boldsymbol{b} = \int_{\mathbb{R}^m} f(\boldsymbol{x}) e^{-iU\boldsymbol{\omega}\cdot\boldsymbol{x}} \mathrm{d}\boldsymbol{x}.$$

Using the Fourier slice theorem, we can invert the d-plane transform.

Lemma 6.2 (*Inversion Formula for d-plane Radon Transform*)**.** *For any $f \in L^1(\mathbb{R}^m)$,*

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^m} \int_{V_{m,k}\times\mathbb{R}^k} \widehat{f}(U\boldsymbol{\omega}) |U\boldsymbol{\omega}|^{m-k} e^{iU\boldsymbol{\omega}\cdot\boldsymbol{x}} \mathrm{d}\boldsymbol{\omega}\mathrm{d}U, \quad \boldsymbol{x} \in \mathbb{R}^m.$$

Proof. By the Fourier slice theorem,

$$\begin{aligned}
&\frac{1}{(2\pi)^m} \int_{V_{m,k}\times\mathbb{R}^k} (P_d[f])^\sharp(U, \boldsymbol{\omega}) e^{iU\boldsymbol{\omega}\cdot\boldsymbol{x}} |U\boldsymbol{\omega}|^{m-k} \mathrm{d}U\mathrm{d}\boldsymbol{\omega} \\
&= \frac{1}{(2\pi)^m} \int_{V_{m,k}\times\mathbb{R}^k} \widehat{f}(U\boldsymbol{\omega}) e^{iU\boldsymbol{\omega}\cdot\boldsymbol{x}} |U\boldsymbol{\omega}|^{m-k} \mathrm{d}U\mathrm{d}\boldsymbol{\omega} \\
&= \frac{1}{(2\pi)^m} \int_{\mathbb{R}^m} \widehat{f}(\boldsymbol{\xi}) e^{i\boldsymbol{\xi}\cdot\boldsymbol{x}} \mathrm{d}\boldsymbol{\xi} = f(\boldsymbol{x}).
\end{aligned}$$

Here, we change variable $\boldsymbol{\xi} = U\boldsymbol{\omega}$ and use the matrix polar integration formula Lemma C.1. □

Remark 1 (*Relations to Marginalization of Probability Distributions*)**.** In short, a d-plane $\boldsymbol{\xi}$ is a subset in $\mathbb{R}^m$, it is identified with a single variable as well, and d-plane $\boldsymbol{\xi}$ (as a variable) is marginalized.

Let us consider a two-variables (or bivariate) case. The marginalization of a probability distribution $f(x_1, x_2)$ in x_1 (resp. x_2) refers to an integral transform of f into its first (resp. second) variable defined by $f_1(x_2) = \int_{\mathbb{R}} f(x_1, x_2)\mathrm{d}x_1$ (rep. $f_2(x_1) = \int_{\mathbb{R}} f(x_1, x_2)\mathrm{d}x_2$).

On the other hand, the d-plane transform of an integrable function f on $\mathbb{R}^2$ (i.e. $f \in L^1(\mathbb{R}^2)$) with $d = 1$ (which is reduced to the classical Radon transform) is given by

$$P_d[f](s, \boldsymbol{u}) = \int_{\mathbb{R}} f(s\boldsymbol{u} + t\boldsymbol{u}^\perp)\mathrm{d}t, \quad (t, \boldsymbol{u}) \in \mathbb{R} \times \mathbb{S}^1$$

where $\mathbb{S}^1$ denotes the set of unit vectors in $\mathbb{R}^2$, i.e. $\mathbb{S}^1 = \{\boldsymbol{u} \in \mathbb{R}^2 \mid |\boldsymbol{u}| = 1\}$, and $\boldsymbol{u}^\perp$ denotes an orthonormal vector, or a unit vector satisfying $\boldsymbol{u} \cdot \boldsymbol{u}^\perp = 0$ (there always exist two $\boldsymbol{u}^\perp$'s for each $\boldsymbol{u}$). Each $(s, \boldsymbol{u}) \in \mathbb{R} \times \mathbb{S}^1$ indicates a d-plane $\boldsymbol{\xi}(s, \boldsymbol{u}) = \{s\boldsymbol{u} + t\boldsymbol{u}^\perp \mid t \in \mathbb{R}\}$.

In particular, by fixing an orthonormal basis $\{\boldsymbol{u}_1, \boldsymbol{u}_2\} \in \mathbb{R}^2$, and identifying bivariate function $f(x_1, x_2)$ with univariate function $f(x_1\boldsymbol{u}_1 + x_2\boldsymbol{u}_2)$, the marginalization of probability distributions is identified with the following specific cases:

$$f_1(x_2) = P_d[f](x_2, \boldsymbol{u}_1), \quad f_2(x_1) = P_d[f](x_1, \boldsymbol{u}_2).$$

6.2. Network design

We define the d-plane (or k-affine) layer. Here, k is the co-dimension of d, satisfying $d + k = m$. In addition to the full-column-matrices cases $(A, \boldsymbol{b}) \in M_{m,k} \times \mathbb{R}^k$, we consider the degenerated cases $(A = aU, \boldsymbol{b}) \in GV_{m,k} \times \mathbb{R}^k$ and $(A = U, \boldsymbol{b}) \in V_{m,k} \times \mathbb{R}^k$, which correspond to several previous studies.

Definition 6.3. Let $\sigma : \mathbb{R}^k \to \mathbb{C}$ be a measurable function. Let M denote either $M_{m,k}, GV_{m,k}$ or $V_{m,k}$. For any function $\gamma : M \times \mathbb{R}^k \to \mathbb{C}$, the continuous neural network with *d-plane (or k-affine) layer* is given by

$$S[\gamma](\boldsymbol{x}) := \int_{M \times \mathbb{R}^k} \gamma(A, \boldsymbol{b})\sigma(A^\top \boldsymbol{x} - \boldsymbol{b})\mathrm{d}A\mathrm{d}\boldsymbol{b}, \quad \boldsymbol{x} \in \mathbb{R}^m.$$

Since the null space $\ker A^\top := \{\boldsymbol{x} \in \mathbb{R}^m \mid A^\top \boldsymbol{x} = 0\}$ is d-dimensional, each d-plane neuron $\sigma(A^\top \boldsymbol{x} - \boldsymbol{b})$ has d-dimensions of constant directions. Therefore, d-plane networks are able to capture d-dimensional singularities in a target function f.

6.3. Ridgelet transforms and reconstruction formulas

We present three variants of solutions for d-plane networks. We note that typical pooling layers σ such as average pooling, max pooling, and ℓ^p-norm are contained in the class of *tempered distributions* ($\mathcal{S}'$) on $\mathbb{R}^k$. The first and second theorems present dense ($A \in M_{m,k}$) and sparse ($A \in GV_{m,k}$) solutions of parameters for the same class of activation functions. Since $GV_{m,k}$ is a measure-zero subset of $M_{m,k}$, the second solution is much sparser than the first solution. The third theorem present the sparsest ($A \in V_{m,k}$) solution, by restricting the class of activation functions. It is supposed to capture characteristic solutions modern activation functions such as ReLU.

In the following, $c_{m,k} := \int_{\mathbb{S}^{k-1} \times V_{m,k-1}} \mathrm{d}U\mathrm{d}\boldsymbol{\omega}$.

Theorem 6.1. *Let $\sigma \in \mathcal{S}'(\mathbb{R}^k), \rho \in \mathcal{S}(\mathbb{R}^k), f \in H^d(\mathbb{R}^m)$. Put*

$$R[f; \rho](A, \boldsymbol{b}) := \frac{1}{\delta(A)} \int_{\mathbb{R}^m} \triangle^{d/2}\, [f](\boldsymbol{x})\overline{\rho(A^\top \boldsymbol{x} - \boldsymbol{b})}\mathrm{d}\boldsymbol{x}, \quad (A, \boldsymbol{b}) \in M_{m,k} \times \mathbb{R}^k$$

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

where $\delta(A)$ is defined as $2^{-k} \prod_{i=1}^d d_i^d \prod_{i<j}(d_i^2 - d_j^2)$ with $d_1 > \cdots > d_k > 0$ being the singular values of A. Then, for almost every $\boldsymbol{x} \in \mathbb{R}^m$, we have

$$S[R[f; \rho]](\boldsymbol{x}) = \int_{M_{m,k} \times \mathbb{R}^k} R[f; \rho](A, \boldsymbol{b})\sigma(A^\top \boldsymbol{x} - \boldsymbol{b})\mathrm{d}A\mathrm{d}\boldsymbol{b} = (\!(\sigma, \rho)\!) f(\boldsymbol{x}).$$

Theorem 6.2. *Let s be a real number; let $\sigma \in \mathcal{S}'(\mathbb{R}^k), \rho \in \mathcal{S}(\mathbb{R}^k), f \in H^s(\mathbb{R}^m)$. Put*

$$R_s[f; \rho](aU, \boldsymbol{b}) := a^{m-s-1} \int_{\mathbb{R}^m} \triangle^{s/2}\, [f](\boldsymbol{x})\overline{\rho(A^\top \boldsymbol{x} - \boldsymbol{b})}\mathrm{d}\boldsymbol{x}, \quad (aU, \boldsymbol{b}) \in GV_{m,k} \times \mathbb{R}^k$$

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

Then, for almost every $\boldsymbol{x} \in \mathbb{R}^m$, we have

$$S[R_s[f; \rho]](\boldsymbol{x}) = \int_{GV_{m,k} \times \mathbb{R}^k} R_s[f; \rho](aU, \boldsymbol{b})\sigma(aU^\top \boldsymbol{x} - \boldsymbol{b})\mathrm{d}a\mathrm{d}U\mathrm{d}\boldsymbol{b} = (\!(\sigma, \rho)\!)_s f(\boldsymbol{x}).$$

Theorem 6.3. *For any real number t, suppose that $\sigma \in \mathcal{S}'(\mathbb{R}^k)$ satisfy $\sigma^\sharp(\boldsymbol{\omega}) = |\boldsymbol{\omega}|^t$ (i.e., σ is the Green function of $\triangle_{\boldsymbol{b}}^{-t/2}$). Let $f \in H^d(\mathbb{R}^m)$. Put*

$$R[f](U, \boldsymbol{b}) := \triangle_{\boldsymbol{b}}^{(d-t)/2}\, P_d[f](U, \boldsymbol{b}) = P_d[\triangle^{(d-t)/2}\, f](U, \boldsymbol{b}), \quad (U, \boldsymbol{b}) \in V_{m,k} \times \mathbb{R}^k$$

where $\triangle_{\boldsymbol{b}}$ denotes the fractional Laplacian in $\boldsymbol{b} \in \mathbb{R}^k$, and P_d is the d-plane transform. Then, for almost every $\boldsymbol{x} \in \mathbb{R}^m$, we have

$$S[R[f]](\boldsymbol{x}) = \int_{V_{m,k}\times\mathbb{R}^k} R[f](U,\boldsymbol{b})\sigma(U^\top \boldsymbol{x} - \boldsymbol{b})\mathrm{d}U\,\mathrm{d}\boldsymbol{b} = \frac{1}{(2\pi)^d c_{m,k}} f(\boldsymbol{x}).$$

As consequences, these reconstruction formulas are understood as constructive universality theorems for d-plane networks. We note (1) that, as far as we have noticed, the first result was not known, (2) that the second result extends the "d-plane ridgelet transform" by Donoho (2001) and Rubin (2004) (see Section 6.4), and (3) that the third result extends the Radon formulas (Theorem 7.2) by Carroll and Dickinson (1989) and Ito (1991) as the special case $k = 1$ and $t = -1$, and recent results on ReLU-nets such as in Savarese et al. (2019), Ongie et al. (2020) and Parhi and Nowak (2021) as the special case $k = 1$ and $t = -2$.

The proof is performed by systematically following the three steps as below.

Sketch Proof. We present the first case. See Appendix D for full proofs.

Step 1 Turn to the Fourier expression:

$$S[\gamma](\boldsymbol{x}) = \frac{1}{(2\pi)^k}\int \gamma^\sharp(A,\boldsymbol{\omega})\sigma^\sharp(\boldsymbol{\omega})e^{i(A\boldsymbol{\omega})\cdot\boldsymbol{x}}\frac{\mathrm{d}A\mathrm{d}\boldsymbol{\omega}}{\delta(A)}$$

Step 2 Use singular value decomposition (SVD)

$$A = UDV^\top, \quad (U,D,V) \in V_{m,k}\times\mathbb{R}^k_+\times O(k),$$

with $\mathrm{d}A/\delta(A) = \mathrm{d}U\mathrm{d}D\mathrm{d}V$ to have

$$= \frac{1}{(2\pi)^k}\int \gamma^\sharp(A,\boldsymbol{\omega})\sigma^\sharp(\boldsymbol{\omega})e^{i(UDV^\top\boldsymbol{\omega})\cdot\boldsymbol{x}}\mathrm{d}U\,\mathrm{d}D\mathrm{d}V\,\mathrm{d}\boldsymbol{\omega}.$$

Change variables $\boldsymbol{\omega}' = V^\top\boldsymbol{\omega}$ (V fixed) and $\boldsymbol{y} = D\boldsymbol{\omega}'$ ($\boldsymbol{\omega}'$ fixed)

$$= \frac{1}{(2\pi)^k}\int \gamma^\sharp(A,V\boldsymbol{\omega}')\sigma^\sharp(V\boldsymbol{\omega}')e^{i(U\boldsymbol{y})\cdot\boldsymbol{x}}\prod_{i=1}^k|\omega_i'|^{-1}\mathrm{d}U\,\mathrm{d}\boldsymbol{y}\mathrm{d}V\,\mathrm{d}\boldsymbol{\omega}'.$$

Step 3 Put a separation-of-variables form (note: $A\boldsymbol{\omega} = AV\boldsymbol{\omega}' = U\boldsymbol{y}$)

$$\gamma^\sharp_{f,\rho}(A,V\boldsymbol{\omega}') = \widehat{f}(U\boldsymbol{y})|U\boldsymbol{y}|^{m-k}\overline{\rho(V\boldsymbol{\omega}')}$$

Then, $\gamma_{f,\rho}$ turns out to be a particular solution because

$$\begin{aligned} S[\gamma_{f,\rho}](\boldsymbol{x}) &= \frac{c_{m,k}}{(2\pi)^k}\left(\int_{O(k)\times\mathbb{R}^k}\sigma^\sharp(V\boldsymbol{\omega}')\overline{\rho^\sharp(V\boldsymbol{\omega}')}\prod_{i=1}^k|\omega_i'|^{-1}\mathrm{d}V\,\mathrm{d}\boldsymbol{\omega}'\right)\left(\int_{V_{m,k}\times\mathbb{R}^k}\widehat{f}(U\boldsymbol{y})|U\boldsymbol{y}|^{m-k}e^{i(U\boldsymbol{y})\cdot\boldsymbol{x}}\mathrm{d}U\,\mathrm{d}\boldsymbol{y}\right)\\ &= (\!(\sigma,\rho)\!)f(\boldsymbol{x}). \end{aligned}$$

Finally, the matrix ridgelet transform can be calculated as below

$$\begin{aligned} \gamma_{f,\rho}(A,\boldsymbol{b}) &= \frac{1}{(2\pi)^k}\int_{\mathbb{R}^k}|A\boldsymbol{\omega}|^{m-k}\widehat{f}(A\boldsymbol{\omega})\overline{\rho(\boldsymbol{\omega})}e^{i\boldsymbol{\omega}\cdot\boldsymbol{b}}\mathrm{d}\boldsymbol{\omega}\\ &= \frac{1}{(2\pi)^k}\int_{\mathbb{R}^k}\left[\int_{\mathbb{R}^m}\triangle^{(m-k)/2}[f](\boldsymbol{x})e^{-iA\boldsymbol{\omega}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{x}\right]\overline{\rho^\sharp(\boldsymbol{\omega})}e^{i\boldsymbol{\omega}\cdot\boldsymbol{b}}\mathrm{d}\boldsymbol{\omega}\\ &= \int_{\mathbb{R}^m}\triangle^{(m-k)/2}[f](\boldsymbol{x})\left[\frac{1}{(2\pi)^k}\int_{\mathbb{R}^k}\rho^\sharp(\boldsymbol{\omega})e^{i\boldsymbol{\omega}\cdot(A^\top\boldsymbol{x}-\boldsymbol{b})}\mathrm{d}\boldsymbol{\omega}\right]^*\mathrm{d}\boldsymbol{x}\\ &= \int_{\mathbb{R}^m}\triangle^{\frac{m-k}{2}}[f](\boldsymbol{x})\overline{\rho(A^\top\boldsymbol{x}-\boldsymbol{b})}\mathrm{d}\boldsymbol{x}\\ &=: R[f;\rho](A,\boldsymbol{b}) \quad \square \end{aligned}$$

6.4. Literature in d-plane ridgelet transform

In the past, two versions of the d-plane ridgelet transform have been proposed. One is a tight frame (i.e., a discrete transform) by Donoho (2001), and the other is a continuous transform by Rubin (2004). The d-plane ridgelet by Donoho can be regarded as the discrete version of the d-plane ridgelet transform by Rubin.

Theorem 6.4 (*Continuous d-plane Ridgelet Transform by Rubin (2004)*)**.**

$$U_a[f](\boldsymbol{\xi}) := \int_{\mathbb{R}^m} f(\boldsymbol{x})u_a(|\boldsymbol{x}-\boldsymbol{\xi}|)\,\mathrm{d}\boldsymbol{x}, \quad \boldsymbol{\xi}\in G_{m,d}$$

$$V_a^*[\phi](\boldsymbol{x}) := \int_{G_{m,d}}\phi(\boldsymbol{\xi})v_a(|\boldsymbol{x}-\boldsymbol{\xi}|)\,\mathrm{d}\boldsymbol{\xi}, \quad \boldsymbol{x}\in\mathbb{R}^m$$

$$\int_0^\infty V_a^*[U_a[f]]\frac{\mathrm{d}a}{a^{1+d}} = cf,$$

where $|\boldsymbol{x}-\boldsymbol{\xi}|$ *denotes the Euclidean distance between point* $\boldsymbol{x}$ *and* d*-plane* $\boldsymbol{\xi}$, $u_a(\cdot) = u(\cdot/a)/a^d$ *and* $v_a(\cdot) = v(\cdot/a)/a^d$.

Recall that an affine d-plane $\boldsymbol{\xi} \in G_{m,k}$ is parametrized by an orthonormal k-frame $U \in V_{m,k}$ and a coordinate vector $\boldsymbol{b} \in \mathbb{R}^k$ as $\boldsymbol{\xi}(U,\boldsymbol{b}) := \{\boldsymbol{x} \in \mathbb{R}^m \mid U^\top \boldsymbol{x} = \boldsymbol{b}\} = U\boldsymbol{b} + \ker U$. Because $|\boldsymbol{x} - \boldsymbol{\xi}(U,\boldsymbol{b})| = |U^\top \boldsymbol{x} - \boldsymbol{b}|$ for any point $\boldsymbol{x} \in \mathbb{R}^m$, the quantity $U^\top \boldsymbol{x} - \boldsymbol{b}$ is understood as the Euclidean *vector-distance* between point $\boldsymbol{x}$ and d-plane $\boldsymbol{\xi}$. Therefore, the d-plane ridgelet transform by Rubin is understood as a special case of $\sigma(aU^\top \boldsymbol{x} - \boldsymbol{b})$ as in Theorem 6.2 where both σ and ρ are radial functions. We remark that a more redundant parametrization $\sigma(A^\top \boldsymbol{x} - \boldsymbol{b})$ as in Theorem 6.1 is natural for the purpose of neural network study, simply because neural network parameters are not strictly restricted to $\sigma(aU^\top \boldsymbol{x} - \boldsymbol{b})$ during the training.

7. Literature overview

7.1. Ridgelet transform in the 1990s

One of the major problems in neural network study in the 1990s was to investigate the expressive power of (fully-connected) shallow neural networks, and the original ridgelet transform was discovered in this context independently by Murata (1996), and Candès (1998). Later in the 2010s, the classes of f and σ have been extended to the distributions by Kostadinova et al. (2014) and Sonoda and Murata (2017) to include the modern activation functions such as ReLU.

The idea of using integral transforms for function approximation is fundamental and has a long history in approximation theory (see, e.g. DeVore and Lorentz, 1993). In the literature of neural network study, the integral representation by Barron (1993) is one of the representative works, where the so-called Barron class and Maurey–Jones–Barron (MJB) approximation error upperbound have been established, which play an important role both in the approximation and estimation theories of neural networks. We refer to Kainen et al. (2013) for more details on the MJB theory.

One obvious strength of the ridgelet transform is the closed-form expression. Before the ridgelet transform, two pioneering results were proposed. One is the Fourier formula by Irie and Miyake (1988) and Funahashi (1989):

Theorem 7.1. *For any* $\sigma \in L^1(\mathbb{R})$ *and* $f \in L^2(\mathbb{R}^m)$,

$$f(\boldsymbol{x}) = \frac{1}{(2\pi)^m \sigma^\sharp(1)} \int_{\mathbb{R}^m \times \mathbb{R}} \widehat{f}(\boldsymbol{a})\sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b)e^{ib}\mathrm{d}\boldsymbol{a}\mathrm{d}b.$$

The other is the Radon formula by Carroll and Dickinson (1989) and Ito (1991):

Theorem 7.2. *For* $\sigma(b) := b_+^0$ *(step function) and any* $f \in \mathcal{S}(\mathbb{R}^m)$,

$$f(\boldsymbol{x}) = \frac{1}{2(2\pi)^{m-1}} \int_{\mathbb{S}^{m-1}\times\mathbb{R}} \partial_t(-\triangle_t)^{(m-1)/2} \boldsymbol{P}[f](\boldsymbol{u},t)\sigma(\boldsymbol{u}\cdot\boldsymbol{x} - t)\mathrm{d}\boldsymbol{u}\mathrm{d}t,$$

where P *denotes the Radon transform.*

Both results clearly show the strong relationship between neural networks and the Fourier and Radon transforms. We note that our result Theorem 6.3 includes the Radon formula (Theorem 7.2) as the special case $k = 1$ and $t = -1$.

7.2. Ridgelet transform in the 2000s

In the context of sparse signal processing, the emergence of the ridgelet transform has motivated another direction of research: Exploring a high-dimensional counterpart of the 1-dimensional wavelet transform. Indeed, the wavelet transform for m-dimensional signals such as images and videos *does* exist, and it is given as below

$$W[f;\psi](a,\boldsymbol{b}) := \int_{\mathbb{R}^m\times\mathbb{R}^m} f(\boldsymbol{x})\overline{\psi_a(\boldsymbol{x}-\boldsymbol{b})}\mathrm{d}\boldsymbol{x}, \quad (a,\boldsymbol{b}) \in \mathbb{R}_+ \times \mathbb{R}^m$$

$$f(\boldsymbol{x}) = \int_{\mathbb{R}^m\times\mathbb{R}_+} W[f;\psi]\psi_a(\boldsymbol{x}-\boldsymbol{b})\frac{\mathrm{d}\boldsymbol{b}\mathrm{d}a}{a}, \quad \boldsymbol{x} \in \mathbb{R}^m$$

for $f \in L^2(\mathbb{R}^m)$, where $\psi_a(\boldsymbol{b}) := \psi(\boldsymbol{b}/a)/a^m$ and $\psi \in \mathcal{S}(\mathbb{R}^m)$ is a wavelet function. However, it is considered to be *unsatisfied* in its *localization* ability, because it is essentially a tensor product of 1-dimensional wavelet transforms.

More precisely, while the 1-dimensional wavelet transform is good at localizing the point singularities such as jumps and kinks in the 1-dimensional signals such as audio recordings, the 2-dimensional wavelet transform is *not* good at localizing the line singularities in 2-dimensional signals such as pictures except when the singularity is straight and parallel to either x- or y-axes. Here, the *singularity of dimension* d is the term by Donoho (2001). For example,

$$f(\boldsymbol{x}) := |x_1^2 + \cdots + x_k^2|^{-\alpha/2}\exp(-|\boldsymbol{x}|^2), \quad 0 < \alpha < k/2$$

is a singular square-integrable function f on $\mathbb{R}^m$ that attains ∞ along the hyperplane $x_1 = \cdots = x_k = 0$.

On the other hand, the ridgelet transform is good at localizing the $(m-1)$-dimensional singularities in any direction because the feature map $\boldsymbol{x} \mapsto \sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)$ is a constant function along the $(m-1)$-dimensional hyperplane normal to $\boldsymbol{a}$. Similarly, the d-plane (or k-affine) ridgelet transform presented in this study is good at localizing the d-dimensional singularities because the feature map $\boldsymbol{x} \mapsto \sigma(A^\top \boldsymbol{x} - \boldsymbol{b})$ is a constant function along the d-dimensional subspace $\ker A^\top = \{\boldsymbol{x} \in \mathbb{R}^m \mid A^\top \boldsymbol{x} = 0\}$.

In search for better localization properties, a variety of "X-lets" have been developed such as curvelet, beamlet, contourlet, and sheerlet under the slogan of geometric multiscale analysis (GMA) (see e.g. Donoho, 2002; Starck et al., 2010). Since ridgelet analysis had already been recognized as *wavelet analysis in the Radon domain*, a variety of generalizations of wavelet transforms and Radon transforms were investigated. In a modern sense, the philosophy of general Radon transforms is to map a function f on a space $X = G/K$ of points x to a function $P[f]$ on another space $\Xi = G/H$ of shapes ξ (see e.g. Helgason, 2010). In the context of singularity localization, the shape ξ such as d-plane determines the shape of singularities, namely, a collection of constant directions in X, and thus the general Radon domain Ξ is understood as the parameter space of the singularities. In this perspective, we can restate the functionality of the ridgelet transform as *wavelet localization in the space Ξ of singularities in X*.

7.3. Ridgelet transform in the 2020s

In the context of deep learning study, the idea of ridgelet transforms have regained the spotlight for the *representer theorem* that characterizes (either deep or shallow) infinitely-wide ReLU networks that minimizes a "representational cost" (Savarese et al., 2019; Ongie et al., 2020; Parhi and Nowak, 2021; Unser, 2019). Here, the representational cost for function f is defined as the infimum of the total variation (TV) norm of the parameter distribution:

$$C[f] := \inf_{\gamma \in \mathcal{G}} \|\gamma\|_{\mathrm{TV}}, \quad \text{s.t.} \quad S[\gamma] = f,$$

where $\mathcal{G}$ is the collection of all signed measures. The TV-norm is a fundamental quantity for the MJB bounds (see e.g. Kainen et al., 2013).

According to Sonoda et al. (2021a), when the class $\mathcal{G}$ of parameter distributions is restricted to $L^2(\mathbb{R}^m \times \mathbb{R})$, then any γ satisfying $S[\gamma] = f$ is uniquely written as a series of ridgelet transforms: $\gamma = R[f;\sigma_*] + \sum_{i=1}^{\infty} R[f_i;\rho_i]$ where σ_* is a certain unique function satisfying $(\!(\sigma, \rho_0)\!) = 1$, yielding $S[R[f;\sigma_*]] = f$; $\{\rho_i\}_{i\in\mathbb{N}}$ is an orthonormal system satisfying $(\!(\sigma, \rho_i)\!) = 0$, yielding $S[R[f_i;\rho_i]] = 0$; and $\{f_i\}_{i\in\mathbb{N}}$ is L^2-functions that is uniquely determined for each γ. We note that σ_* and ρ_i are independent of γ. Hence, the cost is rewritten as a constraint-free expression:

$$C[f] = \inf_{\{f_i\}} \left\| R[f;\sigma_*] + \sum_{i=1}^{\infty} R[f_i;\rho_i] \right\|_{L^1}.$$

As a result, we can conjecture that the minimizer of $C[f]$ is given by ridgelet transform(s). In fact, Ongie et al. (2020) have shown that under some assumptions, the minimizer is given by a derivative of the Radon transform: $\Delta^{(m+1)/2} P[f]$, which is exactly the special case of the ridgelet transform in Theorem 6.3 when $k = 1$ and $t = -1$.

Update: At the same time as the initial submission, Parhi and Unser (2023a) have obtained a representer theorem for *multivariate* activation functions under more careful considerations on the regularization and function spaces based on an extended theory of the d-plane transforms for *distributions* (Parhi and Unser, 2023b). Their result suggests our conjecture was essentially true (modulo finite-order polynomials).

8. Discussion

In the main text, we have seen a variety of examples, but what is essential behind the Fourier expression, changing variables and assuming the separation-of-variables form? In a nutshell, it is *coefficient comparison* for solving equations. Namely, after appropriately changing variables, the network $S[\gamma]$ is rewritten in the Fourier basis, which is thus the coordinate transform from the basis $\{\sigma(ax-b)\}_{a,b}$ to the Fourier basis $\{\exp(i\xi x)\}_{\xi}$. Since we (are supposed to) know the Fourier coefficient $\widehat{f}(\xi)$, we can obtain the unknown function γ by comparing the coefficients in the Fourier domain. From this perspective, we can now understand that the Fourier basis is just a one choice of frames, and the solution steps are summarized as below:

Let $\phi : V \times X \to \mathbb{R}$ and $\psi : \Xi \times X \to \mathbb{R}$ be two feature maps on X parametrized by V and Ξ respectively, and consider their associated integral representations:

$$S[\gamma](x) := \int_V \gamma(v)\phi(v,x)\mathrm{d}v, \quad T[g](x) := \int_\Xi g(\xi)\psi(\xi,x)\mathrm{d}\xi.$$

Here, S and T correspond to the continuous neural network and the inverse Fourier transform respectively. Given a function $f : X \to \mathbb{R}$, suppose that g of $T[g] = f$ is known as, say $g = \widehat{f}$, but γ of $S[\gamma] = f$ is unknown. Then, find a coordinate transform H satisfying $H[\phi](\xi,x) = \psi(\xi,x)$ so that

$$S[\gamma](x) = \int_\Xi H'[\gamma](\xi)\psi(\xi,x)\mathrm{d}\xi,$$

where H' is a dual transform of the coefficients γ associated with the coordinate transform. Then, we can find γ by comparing the coefficients:

$$H'[\gamma] = \widehat{f}.$$

In other words, the mapping H that matches the neuron $\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)$ and Fourier basis $\exp(i\boldsymbol{\xi}\cdot\boldsymbol{x})$ corresponds to the Fourier expression and change of variables in the main text, yielding $H'[\gamma](\boldsymbol{\xi}) = \gamma^\sharp(\boldsymbol{\xi}/\omega, \omega)$.

Table 1
List of layer types $\sigma(ax-b)$ covered in this study. See corresponding sections for the definitions of symbols such as $\mathbb{F}_p, \mathcal{H}_m, G/K, \partial X$ and $\mathfrak{a}^*$.

Layer type	Input x	Parameter (a,b)	Single neuron
Sections 1–2. fully-connected (FC) layer	$\mathbb{R}^m$	$\mathbb{R}^m \times \mathbb{R}$	$\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)$
Section 3. FC layer on finite fields	$\mathbb{F}_p^m$	$\mathbb{F}_p^m \times \mathbb{F}_p$	$\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)$
Section 4. group convolution layer	$\mathcal{H}_m$	$\mathcal{H}_m \times \mathbb{R}$	$\sigma((a \star x)(g)-b)$
Section 5. FC layer on manifolds	G/K	$\mathfrak{a}^* \times \partial X \times \mathbb{R}$	$\sigma(a\langle x,u\rangle - b)$
Section 6. pooling (d-plane ridgelet)	$\mathbb{R}^m$	$\mathbb{R}^{m\times k} \times \mathbb{R}^k$	$\sigma(A^\top\boldsymbol{x}-\boldsymbol{b})$

9. Conclusion

The ultimate goal of this study is to understand neural network parameters. While the ridgelet transform is a strong analysis tool, one of the major short-comings is that the closed-form expression has been known only for small class of neural networks. In this paper, we propose the Fourier slice method, and have shown that various neural networks and their corresponding ridgelet transforms, listed in Table 1, can be systematically obtained by following the three steps of the Fourier slice method.

Needless to say, it is more efficient to analyze networks uniformly in terms of ridgelet analysis than to analyze individual networks manually one by one. As demonstrated in this paper, the coverage of ridgelet analysis is gradually expanding. With the strength of a closed-form expression of the pseudo-inverse operator, the ridgelet transform has several applications. For example, we can/may

1. present a constructive proof of the universal approximation theorem,
2. estimate approximation error bounds by discretizing the reconstruction formula using numerical integration schemes (e.g. MJB theory),
3. describe the distribution of parameters obtained by gradient descent learning (Sonoda et al., 2021b),
4. obtain the general solution to the learning equation $S[\gamma] = f$ (Sonoda et al., 2021a), and
5. construct a representer theorem (Unser, 2019).

The Fourier expression further allows us to view neural networks from the perspective of harmonic analysis and integral geometry. By recasting neural networks in these contexts, we will be able to discover further varieties of novel networks. On the other hand, after the initial submission of this manuscript, the authors have also developed an alternative method of discovery that uses group invariant functions instead of the Fourier expression (Sonoda et al., 2023a,b). The characterization of the networks obtained by the Fourier slice method would be an interesting direction of this study.

Acknowledgments

This work was supported by JSPS KAKENHI 18K18113, JST CREST JPMJCR2015 and JPMJCR1913, JST PRESTO JPMJPR2125, and JST ACT-X JPMJAX2004.

Appendix A. Poincaré disk

Following Helgason (1984)[Introduction, § 4] and Helgason (2008)[Ch.II, § 1], we describe the group theoretic aspect of the Poincaré disk. 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 $|c(\lambda)|^{-2} = \frac{\pi\lambda}{2} \tanh(\frac{\pi\lambda}{2})$ 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:

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

Appendix B. 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

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

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, $\mathrm{d}n = \bigwedge_{1<i<j\le m} \mathrm{d}n_{ij}$ on N,

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

where the second expression is for the polar coordinates $x = k[a]$ 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 precise, 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

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

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, $\omega_m := \prod_{j=1}^m \frac{\Gamma(j/2)}{j(2\pi i)^{j/2}}$, $\boldsymbol{\rho} = (-\frac{1}{2},\dots,-\frac{1}{2},\frac{m-1}{4}) \in \mathbb{C}^m$, and

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

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

Appendix C. Matrix calculus

We refer to Rubin (2018) and Díaz-García and González-Farías (2005) for matrix calculus.

Let m,k be positive integers ($m \ge k$). Let $M_{m,k} \subset \mathbb{R}^{m\times k}$ be the set of all full-column-rank matrices equipped with the volume measure $\mathrm{d}W = \prod_{i,j}\mathrm{d}w_{ij}$. Let $V_{m,k}$ be the Stiefel manifold of orthonormal k-frames in $\mathbb{R}^m$ equipped with the invariant measure $\mathrm{d}U$ normalized to $\int_{V_{m,k}}\mathrm{d}U = 2^k\pi^{mk/2}/\Gamma_k(m/2) =: \sigma_{m,k}$ where Γ_k is the Siegel gamma function. Let $\mathfrak{S}_k \subset \mathbb{R}^{k\times k}$ be the space of real symmetric matrices equipped with the volume measure $\mathrm{d}S = \prod_{i\le j}\mathrm{d}s_{ij}$, which is isometric to the euclidean space $\mathbb{R}^{k(k+1)/2}$. Let $\mathfrak{P}_k \subset \mathbb{R}^{k\times k}$ be the cone of positive definite matrices in $\mathfrak{S}_k$ equipped with the volume measure $\mathrm{d}P = \prod_{i\le j}\mathrm{d}p_{ij}$.

Lemma C.1 (*Matrix Polar Decomposition*). *For any $W \in M_{m,k}$, there uniquely exist $U \in V_{m,k}$ and $P \in \mathfrak{P}_k$ such that*

$$W = UP^{1/2}, \quad P = W^\top W;$$

and for any $f \in L^1(M_{m,k})$,

$$\int_{M_{m,k}} f(W)\mathrm{d}W = \frac{1}{2^k}\int_{V_{m,k}\times\mathfrak{P}_k} f(UP^{1/2})|\det P|^{\frac{m-k-1}{2}}\mathrm{d}P\mathrm{d}U, \quad P = W^\top W.$$

See Rubin (2018, Lemma 2.1) for more details. We remark that while Lemma C.1 is an integration formula on Stiefel manifold, the Grassmannian manifold version is the Blaschke–Petkantschin formula.

Lemma C.2 (*Matrix Polar Integration*). *For any $f \in L^1(\mathbb{R}^k)$,*

$$c_{m,k}\int_{\mathbb{R}^m} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_{V_{m,k}\times\mathbb{R}^k} f(U\boldsymbol{b})|U\boldsymbol{b}|^{m-k}\mathrm{d}U\mathrm{d}\boldsymbol{b},$$

where $c_{m,k} := \int_{\mathbb{S}^{k-1}\times V_{m,k-1}}\mathrm{d}U\mathrm{d}\boldsymbol{\omega}$.

Proof. Recall that $U\boldsymbol{b} = \sum_{i=1}^k b_i\boldsymbol{u}_i \in \mathbb{R}^m$ and thus $|U\boldsymbol{b}|^2 = \boldsymbol{b}^\top U^\top U\boldsymbol{b} = |\boldsymbol{b}|^2$. Hence, using the polar decomposition $\boldsymbol{b} = r\boldsymbol{\omega}$ with $\mathrm{d}\boldsymbol{b} = r^{k-1}\mathrm{d}r\mathrm{d}\boldsymbol{\omega}$,

$$\int_{V_{m,k}\times\mathbb{R}^k} f(U\boldsymbol{b})|U\boldsymbol{b}|^{m-k}\mathrm{d}U\mathrm{d}\boldsymbol{b}$$
$$= \int_{V_{m,k}\times\mathbb{S}^{k-1}\times\mathbb{R}_+} f(U\boldsymbol{\omega}r)r^{m-k}r^{k-1}\mathrm{d}r\mathrm{d}\boldsymbol{\omega}\mathrm{d}U$$

then letting $\boldsymbol{u}_{\boldsymbol{\omega}} := \sum_{i=1}^k \omega_i\boldsymbol{u}_i$, which is a unit vector in $\operatorname{span} U = [\boldsymbol{u}_1,\dots,\boldsymbol{u}_k]$, and letting $U_{-\boldsymbol{\omega}}$ be a rearranged $k-1$-frame in $\operatorname{span} U$ that excludes $\boldsymbol{u}_{\boldsymbol{\omega}}$,

$$= \int_{\mathbb{S}^{k-1}}\left[\int_{V_{m,k-1}\times\mathbb{S}^{k-1}\times\mathbb{R}_+} f(r\boldsymbol{u}_{\boldsymbol{\omega}})r^{m-1}\mathrm{d}r\mathrm{d}\boldsymbol{u}_{\boldsymbol{\omega}}\mathrm{d}U_{-\boldsymbol{\omega}}\right]\mathrm{d}\boldsymbol{\omega}$$
$$= \int_{\mathbb{S}^{k-1}\times V_{m,k-1}}\mathrm{d}U_{k-1}\mathrm{d}\boldsymbol{\omega}\int_{\mathbb{R}^m} f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \quad \boldsymbol{x} = r\boldsymbol{u}_{\boldsymbol{\omega}}. \quad \square$$

Lemma C.3 (*SVD*). *For any column-full-rank matrix $W \in M_{m,k}$, there exist $(U,D,V) \in V_{m,k}\times\mathbb{R}_+^k\times O(k)$ satisfying $W = UDV^\top$ with $d_1 > \cdots > d_k > 0$; and for any $f \in L^1(M_{m,k})$,*

$$\int_{M_{m,k}} f(W)\mathrm{d}W = 2^{-k}\int_{V_{m,k}\times\mathbb{R}_+^k\times O(k)} f(UDV^\top)|\det D|^{m-k}\prod_{i<j}(d_i^2-d_j^2)\mathrm{d}D\mathrm{d}V\mathrm{d}U,$$

where $\mathrm{d}D = \bigwedge_{i=1}^k \mathrm{d}d_i$ and $\mathrm{d}U,\mathrm{d}V$ denote the invariant measures.

See Lemma 1 of Díaz-García and González-Farías (2005) for the proof.

Appendix D. Proofs

D.1. Solution via singular value decomposition

Step 1. We begin with the following Fourier expression:

$$\begin{aligned} S[\gamma](\boldsymbol{x}) &= \int_{M_{m,k}\times\mathbb{R}^k} \gamma(A,\boldsymbol{b})\sigma(A^\top \boldsymbol{x}-\boldsymbol{b})\mathrm{d}A\mathrm{d}\boldsymbol{b} \\ &= \frac{1}{(2\pi)^k}\int_{M_{m,k}\times\mathbb{R}^k} \gamma^\sharp(A,\boldsymbol{\omega})\sigma^\sharp(\boldsymbol{\omega})e^{i(A\boldsymbol{\omega})\cdot\boldsymbol{x}}\mathrm{d}A\mathrm{d}\boldsymbol{\omega}. \end{aligned} \tag{D.1}$$

Here, we assume (D.1) to be absolutely convergent for a.e. $\boldsymbol{x}\in\mathbb{R}^m$, so that we can change the order of integrations freely. But this assumption will be automatically satisfied because we eventually set γ to be the ridgelet transform.

Step 2. In the following, we aim to turn the integration $\int\cdots e^{i(A\boldsymbol{\omega})\cdot\boldsymbol{x}}\mathrm{d}A\mathrm{d}\boldsymbol{\omega}$ into the Fourier inversion $\int\cdots e^{iU\boldsymbol{\xi}\cdot\boldsymbol{x}}|U\boldsymbol{\xi}|^d\mathrm{d}U\mathrm{d}\boldsymbol{\xi}$ in the matrix polar coordinates. To achieve this, we use the singular value decomposition.

Lemma D.1 (*Singular Value Decomposition, Lemma C.3*)**.** *The matrix space $M_{m,k}$ can be decomposed into $M_{m,k}=V_{m,k}\times\mathbb{R}^k_+\times O(k)$ via singular value decomposition*

$$A = UDV^\top,\quad (U,D,V)\in V_{m,k}\times\mathbb{R}^k_+\times O(k)$$

satisfying $D=\mathrm{diag}[d_1,\dots,d_k]$ $(d_1>\cdots>d_k>0)$ (distinct singular values); and the Lebesgue measure $\mathrm{d}A$ is calculated as

$$\mathrm{d}A = \delta(D)\mathrm{d}D\mathrm{d}U\mathrm{d}V,\quad \delta(D) := 2^{-k}|\det D|^d\Delta(D^2),$$

where $\mathrm{d}D=\bigwedge_{i=1}^k \mathrm{d}d_i$; $\mathrm{d}U$ and $\mathrm{d}V$ are invariant measures on $V_{m,k}$ and $O(k)$ respectively; and $\Delta(D^2):=\prod_{i<j}(d_i^2-d_j^2)$ denotes the Vandermonde polynomial (or the products of differences) of a given (diagonalized) vector $D=[d_1,\dots,d_k]$.

If there is no risk of confusion, we write $UDV^\top$ as A for the sake of readability.

Using SVD, the Fourier expression is rewritten as follows:

$$\text{(D.1)} = \frac{1}{(2\pi)^k}\int_{M_{m,k}\times\mathbb{R}^k}\gamma^\sharp(A,\boldsymbol{\omega})\sigma^\sharp(\boldsymbol{\omega})e^{i(UDV^\top\boldsymbol{\omega})\cdot\boldsymbol{x}}\delta(D)\mathrm{d}U\mathrm{d}D\mathrm{d}V\mathrm{d}\boldsymbol{\omega} \tag{D.2}$$

Changing the variables as $(\boldsymbol{\omega},V)=(V\boldsymbol{\omega}',V)$ with $\mathrm{d}\boldsymbol{\omega}\mathrm{d}V=\mathrm{d}\boldsymbol{\omega}'\mathrm{d}V$,

$$= \frac{1}{(2\pi)^k}\int_{M_{m,k}\times\mathbb{R}^k}\gamma^\sharp(A,V\boldsymbol{\omega}')\sigma^\sharp(V\boldsymbol{\omega}')e^{i(UD\boldsymbol{\omega}')\cdot\boldsymbol{x}}\delta(D)\mathrm{d}U\mathrm{d}D\mathrm{d}V\mathrm{d}\boldsymbol{\omega}' \tag{D.3}$$

Then, extending the domain of D from $\mathbb{R}^k_+$ to $\mathbb{R}^k$, changing the variables as $(d_i,\omega_i')=(y_i/\omega_i',\omega_i')$ with $\mathrm{d}d_i\mathrm{d}\omega_i'=|\omega_i'|^{-1}\mathrm{d}y_i\mathrm{d}\omega_i'$, and writing $\boldsymbol{y}:=[y_1,\dots,y_k]$,

$$= \frac{1}{(4\pi)^k}\int_{V_{m,k}\times\mathbb{R}^k\times O(k)\times\mathbb{R}^k}\gamma^\sharp(A,V\boldsymbol{\omega}')\sigma^\sharp(V\boldsymbol{\omega}')e^{i(U\boldsymbol{y})\cdot\boldsymbol{x}}\delta(D)\prod_{i=1}^k|\omega_i'|^{-1}\mathrm{d}U\mathrm{d}\boldsymbol{y}\mathrm{d}V\mathrm{d}\boldsymbol{\omega}'. \tag{D.4}$$

Step 3. Therefore, it is natural to suppose γ to satisfy a *separation-of-variables* form

$$\gamma^\sharp(A,V\boldsymbol{\omega}')\sigma^\sharp(V\boldsymbol{\omega}')\delta(D)\prod_{i=1}^k|\omega_i'|^{-1} = \widehat{f}(U\boldsymbol{y})|U\boldsymbol{y}|^d\phi^\sharp(V,\boldsymbol{\omega}') \tag{D.5}$$

with an auxiliary convergence factor ϕ. Then, we have

$$\begin{aligned} \text{(D.4)} &= \frac{1}{(4\pi)^k}\left(\int_{O(k)\times\mathbb{R}^k}\phi^\sharp(V,\boldsymbol{\omega}')\mathrm{d}V\mathrm{d}\boldsymbol{\omega}'\right)\left(\int_{V_{m,k}\times\mathbb{R}^k}\widehat{f}(U\boldsymbol{y})|U\boldsymbol{y}|^d e^{iU\boldsymbol{y}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{y}\mathrm{d}U\right) \\ &= \frac{c_\phi}{(2\pi)^m}\int_{\mathbb{R}^m}\widehat{f}(\boldsymbol{y})e^{i\boldsymbol{y}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{y} = c_\phi f(\boldsymbol{x}). \end{aligned}$$

Here, we put $c_\phi := 2^{-k}(2\pi)^d c_{m,k}^{-1}\int_{O(k)\times\mathbb{R}^k}\phi^\sharp(V,\boldsymbol{\omega}')\mathrm{d}V\mathrm{d}\boldsymbol{\omega}'$, and used a matrix polar integration formula given in Lemma C.1.

Finally, the form (D.5) can be satisfied as below. Since $V\boldsymbol{\omega}'=\boldsymbol{\omega}$ and $U\boldsymbol{y}=UD\boldsymbol{\omega}'=A\boldsymbol{\omega}$, it is reduced to

$$\frac{\gamma^\sharp(A,\boldsymbol{\omega})}{\phi^\sharp(V,\boldsymbol{\omega}')} = \frac{\widehat{f}(A\boldsymbol{\omega})|A\boldsymbol{\omega}|^d}{\sigma^\sharp(V\boldsymbol{\omega}')\delta(D)\prod_{i=1}^k|\omega_i'|^{-1}}.$$

Hence we can put

$$\gamma^\sharp(A,\boldsymbol{\omega}) = \frac{\widehat{f}(A\boldsymbol{\omega})|A\boldsymbol{\omega}|^d\overline{\rho^\sharp(\boldsymbol{\omega})}}{\delta(D)}, \tag{D.6}$$

$$\phi^\sharp(V,\boldsymbol{\omega}') = \sigma^\sharp(V\boldsymbol{\omega}')\overline{\rho^\sharp(V\boldsymbol{\omega}')}\prod_{i=1}^k |\omega_i'|^{-1},$$

with additional convergence factor ρ. In this setting, the constant c_ϕ is calculated as

$$\begin{aligned} c_\phi &:= \frac{(2\pi)^d}{2^k c_{m,k}} \int_{O(k)\times\mathbb{R}^k} \sigma^\sharp(V\boldsymbol{\omega}')\overline{\rho^\sharp(V\boldsymbol{\omega}')}\prod_{i=1}^k |\omega_i'|^{-1}\mathrm{d}V\mathrm{d}\boldsymbol{\omega}' \\ &= \frac{(2\pi)^d}{2^k c_{m,k}} \int_{\mathbb{R}^k} \sigma^\sharp(\boldsymbol{\omega})\overline{\rho^\sharp(\boldsymbol{\omega})}\prod_{i=1}^k |\omega_i|^{-1}\mathrm{d}\boldsymbol{\omega} =: (\!(\sigma,\rho)\!); \end{aligned}$$

and γ is obtained by taking the Fourier inversion of (D.6) with respect to $\boldsymbol{\omega}$ as follows:

$$\begin{aligned} \gamma(A,\boldsymbol{b}) &= \frac{1}{(2\pi)^k\delta(D)} \int_{\mathbb{R}^k} |A\boldsymbol{\omega}|^d \widehat{f}(A\boldsymbol{\omega})\overline{\rho^\sharp(\boldsymbol{\omega})}e^{i\boldsymbol{\omega}\cdot\boldsymbol{b}}\mathrm{d}\boldsymbol{\omega}, \\ &= \frac{1}{(2\pi)^k\delta(D)} \int_{\mathbb{R}^k} \left[\int_{\mathbb{R}^m} \Delta^{d/2}\,[f](\boldsymbol{x})e^{-iA\boldsymbol{\omega}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{x}\right] \overline{\rho^\sharp(\boldsymbol{\omega})}e^{i\boldsymbol{\omega}\cdot\boldsymbol{b}}\mathrm{d}\boldsymbol{\omega} \\ &= \frac{1}{\delta(D)} \int_{\mathbb{R}^m} \Delta^{d/2}\,[f](\boldsymbol{x}) \left[\frac{1}{(2\pi)^k}\int_{\mathbb{R}^k} \rho^\sharp(\boldsymbol{\omega})e^{i\boldsymbol{\omega}\cdot(A^\top\boldsymbol{x}-\boldsymbol{b})}\mathrm{d}\boldsymbol{\omega}\right]^* \mathrm{d}\boldsymbol{x} \\ &= \frac{1}{\delta(D)} \int_{\mathbb{R}^m} \Delta^{d/2}\,[f](\boldsymbol{x})\overline{\rho(A^\top\boldsymbol{x}-\boldsymbol{b})}\mathrm{d}\boldsymbol{x} \\ &=: R[f](A,\boldsymbol{b}). \end{aligned}$$

To sum up, we have shown that

$$S[R[f]](\boldsymbol{x}) = \int_{M_{m,k}\times\mathbb{R}^k} R[f](A,\boldsymbol{b})\sigma(A^\top\boldsymbol{x}-\boldsymbol{b})\mathrm{d}A\mathrm{d}\boldsymbol{b} = (\!(\sigma,\rho)\!)f(\boldsymbol{x}).$$

D.2. Restriction to the similitude group

Let us consider the restricted case of the similitude group $GV_{m,k}$. Since it is a measure-zero subspace of $M_{m,k}$, we can obtain the different solution.

Step 1. The continuous network and its Fourier expression are given as

$$\begin{aligned} S[\gamma](\boldsymbol{x}) &:= \int_{GV_{m,k}\times\mathbb{R}^k} \gamma(A,\boldsymbol{b})\sigma(A^\top\boldsymbol{x}-\boldsymbol{b})\mathrm{d}\mu(A)\mathrm{d}\boldsymbol{b} \\ &= \frac{1}{(2\pi)^k}\int_{GV_{m,k}\times\mathbb{R}^k} \gamma^\sharp(A,\boldsymbol{\omega})\sigma^\sharp(\boldsymbol{\omega})e^{i(aU\boldsymbol{\omega})\cdot\boldsymbol{x}}\alpha(a)\mathrm{d}a\mathrm{d}U\mathrm{d}\boldsymbol{\omega} \end{aligned} \tag{D.7}$$

Step 2. (Skipping the matrix decomposition of A and) turning $\boldsymbol{\omega}$ into polar coordinates $\boldsymbol{\omega} = r\boldsymbol{v}$ with $(r,\boldsymbol{v}) \in \mathbb{R}_+\times\mathbb{S}^{k-1}$ and $\mathrm{d}\boldsymbol{\omega} = r^{k-1}\mathrm{d}r\mathrm{d}\boldsymbol{v}$, yielding

$$(\text{D.7}) = \frac{1}{(2\pi)^k}\int_{GV_{m,k}\times\mathbb{R}_+\times\mathbb{S}^{k-1}} \gamma^\sharp(A,r\boldsymbol{v})\sigma^\sharp(r\boldsymbol{v})e^{i(arU\boldsymbol{v})\cdot\boldsymbol{x}}\alpha(a)r^{k-1}\mathrm{d}a\mathrm{d}U\mathrm{d}r\mathrm{d}\boldsymbol{v}$$

Changing the variable $(a,r) = (y/r,r)$ with $\mathrm{d}a\mathrm{d}r = r^{-1}\mathrm{d}y\mathrm{d}r$,

$$= \frac{1}{(2\pi)^k}\int_{GV_{m,k}\times\mathbb{R}_+\times\mathbb{S}^{k-1}} \gamma^\sharp(A,r\boldsymbol{v})\sigma^\sharp(r\boldsymbol{v})e^{i(yU\boldsymbol{v})\cdot\boldsymbol{x}}\alpha(y/r)r^{k-2}\mathrm{d}y\mathrm{d}U\mathrm{d}r\mathrm{d}\boldsymbol{v}$$

and returning $y\boldsymbol{v}$ into the Euclidean coordinate $\boldsymbol{y}$ with $y^{k-1}\mathrm{d}y\mathrm{d}\boldsymbol{v} = \mathrm{d}\boldsymbol{y}$,

$$= \frac{1}{(2\pi)^k}\int_{GV_{m,k}\times\mathbb{R}^k} \gamma^\sharp(A,r\boldsymbol{y}/|\boldsymbol{y}|)\sigma^\sharp(r\boldsymbol{y}/|\boldsymbol{y}|)e^{i(U\boldsymbol{y})\cdot\boldsymbol{x}}\alpha(|\boldsymbol{y}|/r)r^{k-2}|\boldsymbol{y}|^{-(k-1)}\mathrm{d}\boldsymbol{y}\mathrm{d}U\mathrm{d}r. \tag{D.8}$$

Step 3. Since $r\boldsymbol{y}/|\boldsymbol{y}| = r\boldsymbol{v} = \boldsymbol{\omega}$ and $|\boldsymbol{y}|/r = y/r = a$, supposing the separation-of-variables form as

$$\gamma^\sharp(A,\boldsymbol{\omega})\sigma^\sharp(\boldsymbol{\omega})\alpha(a)a^{-\ell}|\boldsymbol{\omega}|^{k-2-\ell}|\boldsymbol{y}|^{-(k-1-\ell)} = \widehat{f}(U\boldsymbol{y})|U\boldsymbol{y}|^d\phi^\sharp(\boldsymbol{\omega}), \tag{D.9}$$

for any number $\ell\in\mathbb{R}$, we have

$$(\text{D.8}) = \frac{1}{(2\pi)^k}\left(\int\int_{\mathbb{R}_+}\phi^\sharp(r)\mathrm{d}r\right)\left(\int_{V_{m,k}\times\mathbb{R}^k}\widehat{f}(U\boldsymbol{y})|U\boldsymbol{y}|^d e^{-(U\boldsymbol{y})\cdot\boldsymbol{x}}\mathrm{d}U\mathrm{d}\boldsymbol{y}\right) = c_\phi f(\boldsymbol{x}).$$

Note that in addition to $U\boldsymbol{y} = A\boldsymbol{\omega}$, we have $|\boldsymbol{y}| = |U\boldsymbol{y}| = a|\boldsymbol{\omega}|$ and thus $|U\boldsymbol{y}|^d = |A\boldsymbol{\omega}|^n a^{d-n}|\boldsymbol{\omega}|^{d-n}$ for any number $n\in\mathbb{R}$. Hence the condition is reduced to

$$\frac{\gamma^\sharp(A,\boldsymbol{\omega})}{\phi^\sharp(\boldsymbol{\omega})} = \frac{\widehat{f}(A\boldsymbol{\omega})|A\boldsymbol{\omega}|^{n+(k-1-\ell)}a^{d-n+\ell}|\boldsymbol{\omega}|^{m-2k-n+2+\ell}}{\sigma^\sharp(\boldsymbol{\omega})\alpha(a)} = \frac{\widehat{f}(A\boldsymbol{\omega})|A\boldsymbol{\omega}|^{s}a^{m-s-1}|\boldsymbol{\omega}|^{d-s+1}}{\sigma^\sharp(\boldsymbol{\omega})\alpha(a)},$$

where we put $s := n + (k - 1 - \ell)$, which can be an arbitrary real number. As a result, to satisfy (D.9), we may put

$$\gamma^\sharp(A,\boldsymbol{\omega}) = \widehat{f}(A\boldsymbol{\omega})|A\boldsymbol{\omega}|^{s}\overline{\rho^\sharp(\boldsymbol{\omega})},$$
$$\alpha(a) = a^{m-s-1},$$
$$\phi^\sharp(\boldsymbol{\omega}) = \sigma^\sharp(\boldsymbol{\omega})\overline{\rho^\sharp(\boldsymbol{\omega})}|\boldsymbol{\omega}|^{-(d-s+1)};$$

which lead to

$$R_s[f] = \int_{\mathbb{R}^m} [\Delta^{s/2}\, f](\boldsymbol{x})\overline{\rho(A^\top\boldsymbol{x}-\boldsymbol{b})}\mathrm{d}\boldsymbol{x},$$
$$(\!(\sigma,\rho)\!)_s \propto \int_{\mathbb{R}^k} \sigma^\sharp(\boldsymbol{\omega})\overline{\rho^\sharp(\boldsymbol{\omega})}|\boldsymbol{\omega}|^{-(d-s+1)}\mathrm{d}\boldsymbol{\omega},$$
$$S[R_s[f]](\boldsymbol{x}) = \int_{GV_{m,k}\times\mathbb{R}^k} R_s[f](A,\boldsymbol{b})\sigma(A^\top\boldsymbol{x}-\boldsymbol{b})a^{m-s-1}\mathrm{d}a\mathrm{d}U\mathrm{d}\boldsymbol{b} = (\!(\sigma,\rho)\!)f(\boldsymbol{x}).$$

By matching the order of the fractional derivative Δ^s, $s = d$ corresponds to the SVD solution. On the other hand, by matching the weight $|\boldsymbol{\omega}|^{-(d-s+1)}$, $k = 1$ and $s = 0$ exactly reproduces the classical result.

D.3. Restriction to the stiefel manifold

Let us consider a further restricted case of the Stiefel manifold.

Step 1.

$$S[\gamma](\boldsymbol{x}) := \int_{V_{m,k}\times\mathbb{R}^k} \gamma(U,\boldsymbol{b})\sigma(U^\top\boldsymbol{x}-\boldsymbol{b})\mathrm{d}U\mathrm{d}\boldsymbol{b} = \frac{1}{(2\pi)^k}\int_{V_{m,k}\times\mathbb{R}^k} \gamma^\sharp(U,\boldsymbol{\omega})\sigma^\sharp(\boldsymbol{\omega})e^{i(U\boldsymbol{\omega})\cdot\boldsymbol{x}}\mathrm{d}U\mathrm{d}\boldsymbol{\omega}.$$

Namely, the weight matrix parameter $U \in M_{m,k}$ now simply lies in the Stiefel manifold $V_{m,k}$, and thus it does not contain any scaling factor. We show that this formulation still admit solutions, provided that σ is also appropriately restricted.

Step 3. Skipping the rescaling step (Step 2), let us consider the separation-of-variables form:

$$\gamma^\sharp(U,\boldsymbol{\omega})\sigma^\sharp(\boldsymbol{\omega}) = \widehat{f}(U\boldsymbol{\omega})|U\boldsymbol{\omega}|^d\phi^\sharp(\boldsymbol{\omega}); \tag{D.10}$$

satisfied by

$$\gamma^\sharp(U,\boldsymbol{\omega}) = \widehat{f}(U\boldsymbol{\omega})|U\boldsymbol{\omega}|^{s}\overline{\rho^\sharp(\boldsymbol{\omega})},$$
$$\phi^\sharp(\boldsymbol{\omega}) = \sigma^\sharp(\boldsymbol{\omega})\overline{\rho^\sharp(\boldsymbol{\omega})}|\boldsymbol{\omega}|^{s-d},$$

for any real number $s \in \mathbb{R}$. Here, we used $|U\boldsymbol{\omega}| = |\boldsymbol{\omega}|$.

In order (D.10) to turn to a solution, it is sufficient when

$$\phi^\sharp(\boldsymbol{\omega}) = c_{m,k}^{-1}(2\pi)^{-d}, \tag{D.11}$$

because then

$$\text{(D.10)} = \frac{1}{(2\pi)^k}\int_{V_{m,k}\times\mathbb{R}^k} \widehat{f}(U\boldsymbol{\omega})|U\boldsymbol{\omega}|^d\phi^\sharp(\boldsymbol{\omega})e^{i(U\boldsymbol{\omega})\cdot\boldsymbol{x}}\mathrm{d}U\mathrm{d}\boldsymbol{\omega} = \frac{1}{(2\pi)^m}\int_{\mathbb{R}^m}\widehat{f}(\boldsymbol{y})e^{i\boldsymbol{y}\cdot\boldsymbol{x}}\mathrm{d}\boldsymbol{y} = f(\boldsymbol{x}).$$

Compared to the previous results, (D.11) demands much more strict. Nonetheless, a few examples are such as

$$\sigma^\sharp(\boldsymbol{\omega}) = |\boldsymbol{\omega}|^t, \quad \rho^\sharp(\boldsymbol{\omega}) = c_{m,k}^{-1}(2\pi)^{-d}|\boldsymbol{\omega}|^{d-(s+t)};$$

or equivalently in the real domain,

$$\Delta_{\boldsymbol{b}}^{-\frac{t}{2}}[\sigma](\boldsymbol{b}) = \delta(\boldsymbol{b}), \quad \Delta_{\boldsymbol{b}}^{-\frac{d-(s+t)}{2}}[\rho](\boldsymbol{b}) = c_{m,k}^{-1}(2\pi)^{-d}\delta(\boldsymbol{b}).$$

In particular when $k = 1$, then σ coincides with the Dirac delta ($t = 0$), step function ($t = -1$), and ReLU function ($t = -2$).

Interestingly, the ridgelet transform is reduced to the d-plane transform ($d := m - k$ is the codimension). Since $\gamma^\sharp(U,\boldsymbol{\omega}) = \widehat{f}(U\boldsymbol{\omega})|U\boldsymbol{\omega}|^{d-t}$, we have

$$\begin{aligned}\gamma(U,\boldsymbol{b}) &= \frac{1}{(2\pi)^k}\int_{\mathbb{R}^k} \widehat{f}(U\boldsymbol{\omega})|U\boldsymbol{\omega}|^{d-t} e^{i\boldsymbol{\omega}\cdot\boldsymbol{b}}\mathrm{d}\boldsymbol{\omega} \\ &= \frac{1}{(2\pi)^k}\int_{\mathbb{R}^k} \widehat{\Delta^{(d-t)/2}[f]}(U\boldsymbol{\omega}) e^{i\boldsymbol{\omega}\cdot\boldsymbol{b}}\mathrm{d}\boldsymbol{\omega},\end{aligned}$$

but this is the Fourier expression (a.k.a. Fourier slice theorem) for the d-plane transform, say P_d, of the derivative $\Delta^{(d-t)/2}[f]$. In other words, when the scaling parameter is removed, the reconstruction formula is reduced to the Radon transform:

$$S[R[f]](\boldsymbol{x}) = \int_{V_{m,k}\times\mathbb{R}^k} P_d[\Delta^{(d-t)/2}[f]](U,\boldsymbol{b})\sigma(U^\top\boldsymbol{x}-\boldsymbol{b})\mathrm{d}U\mathrm{d}\boldsymbol{b}.$$

References

Barron, A.R., 1993. Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. Inform. Theory 39, 930–945.

Bartolucci, F., De Mari, F., Monti, M., 2021. Unitarization of the horocyclic Radon transform on symmetric spaces. In: Harmonic and Applied Analysis: From Radon Transforms To Machine Learning. Springer International Publishing, Cham, pp. 1–54.

Bengio, Y., Le Roux, N., Vincent, P., Delalleau, O., Marcotte, P., 2006. Convex neural networks. Adv. Neural Inf. Process. Syst. 18, 123–130.

Bronstein, M.M., Bruna, J., Cohen, T., Veličković, P., 2021. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. arXiv preprint: 2104.13478.

Bruna, J., Mallat, S., 2013. Invariant scattering convolution networks. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1872–1886.

Candès, E.J., 1998. Ridgelets: Theory and Applications (Ph.D. thesis). Standford University.

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

Chizat, L., Bach, F., 2018. On the global convergence of gradient descent for over-parameterized models using optimal transport. Adv. Neural Inf. Process. Syst. 32, 3036–3046.

Cohen, T.S., Geiger, M., Weiler, M., 2019. A general theory of equivariant CNNs on homogeneous spaces. Adv. Neural Inf. Process. Syst. 32.

Cohen, T., Welling, M., 2016. Group equivariant convolutional networks. In: Proceedings of the 33rd International Conference on Machine Learning. Vol. 48, pp. 2990–2999.

DeVore, R.A., Lorentz, G.G., 1993. Constructive Approximation. Springer-Verlag Berlin Heidelberg.

Díaz-García, J.A., González-Farías, G., 2005. Singular random matrix decompositions: Jacobians. J. Multivariate Anal. 93, 296–312.

Donoho, D.L., 2001. Ridge functions and orthonormal ridgelets. J. Approx. Theory 111, 143–179.

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

Fukushima, K., 1980. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol. Cybernet. 36, 193–202.

Funahashi, K.-I., 1989. On the approximate realization of continuous mappings by neural networks. Neural Netw. 2, 183–192.

Ganea, O., Becigneul, G., Hofmann, T., 2018. Hyperbolic neural networks. Adv. Neural Inf. Process. Syst. 31.

Grafakos, L., 2008. Classical Fourier analysis. In: Graduate Texts in Mathematics, second ed. Springer New York.

Gulcehre, C., Denil, M., Malinowski, M., Razavi, A., Pascanu, R., Hermann, K.M., Battaglia, P., Bapst, V., Raposo, D., Santoro, A., de Freitas, N., 2019. Hyperbolic attention networks. In: International Conference on Learning Representations.

Helgason, S., 1984. Groups and geometric analysis: Integral geometry. In: Invariant Differential Operators, and Spherical Functions. In: volume 83 of Mathematical Surveys and Monographs, American Mathematical Society.

Helgason, S., 2008. Geometric Analysis on Symmetric Spaces: Second Edition, second ed. In: volume 39 of Mathematical Surveys and Monographs, American Mathematical Society.

Helgason, S., 2010. Integral Geometry and Radon Transforms. Springer-Verlag New York.

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

Ito, Y., 1991. Representation of functions by superpositions of a step or sigmoid function and their applications to neural network theory. Neural Netw. 4, 385–394.

Kainen, P.C., Kůrková, V., Sanguineti, M., 2013. Approximating multivariable functions by feedforward neural nets. In: Handbook on Neural Information Processing. In: volume 49 of Intelligent Systems Reference Library, Springer Berlin Heidelberg, pp. 143–181.

Kapovich, M., Leeb, B., Porti, J., 2017. Anosov subgroups: dynamical and geometric characterizations. Eur. J. Math. 3, 808–898.

Kondor, R., Trivedi, S., 2018. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In: Proceedings of the 35th International Conference on Machine Learning. Vol. 80, pp. 2747–2755.

Kostadinova, S., Pilipović, S., Saneva, K., Vindas, J., 2014. The ridgelet transform of distributions. Integral Transforms Spec. Funct. 25, 344–358.

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., Boguñá, M., 2010. Hyperbolic geometry of complex networks. Phys. Rev. E 82, 36106.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. ImageNet classification with deep convolutional neural networks. Adv. Neural Inf. Process. Syst. 25, 1097–1105.

Kumagai, W., Sannai, A., 2020. Universal approximation theorem for equivariant maps by group CNNs. arXiv preprint: 2012.13882.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. In: Proceedings of the IEEE. Vol. 86, pp. 2278–2324.

Maron, H., Fetaya, E., Segol, N., Lipman, Y., 2019. On the universality of invariant networks. In: Proceedings of the 36th International Conference on Machine Learning. Vol. 97, pp. 4363–4371.

Mei, S., Montanari, A., Nguyen, P.-M., 2018. A mean field view of the landscape of two-layer neural networks. In: Proceedings of the National Academy of Sciences. Vol. 115, pp. E7665–E7671.

Murata, N., 1996. An integral representation of functions using three-layered networks and their approximation bounds. Neural Netw. 9, 947–956.

Nickel, M., Kiela, D., 2017. Poincaré embeddings for learning hierarchical representations. Adv. Neural Inf. Process. Syst. 30.

Nickel, M., Kiela, D., 2018. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In: Proceedings of the 35th International Conference on Machine Learning. Vol. 80, pp. 3779–3788.

Nitanda, A., Suzuki, T., 2017. Stochastic particle gradient descent for infinite ensembles. arXiv preprint: 1712.05438.

Nitanda, A., Wu, D., Suzuki, T., 2022. Convex analysis of the mean field langevin dynamics. In: Proceedings of the 25th International Conference on Artificial Intelligence and Statistics. Vol. 151, pp. 9741–9757.

Ongie, G., Willett, R., Soudry, D., Srebro, N., 2020. A function space view of bounded norm infinite width ReLU nets: The multivariate case. In: International Conference on Learning Representations.

Parhi, R., Nowak, R.D., 2021. Banach space representer theorems for neural networks and ridge splines. J. Mach. Learn. Res. 22, 1–40.
Parhi, R., Unser, M., 2023a. Function-space optimality of neural architectures with multivariate nonlinearities. arXiv preprint: 2310.03696.
Parhi, R., Unser, M., 2023b. Distributional extension and invertibility of the k-plane transform and its dual. arXiv preprint: 2310.01233.
Ranzato, M., Poultney, C., Chopra, S., LeCun, Y., 2007. Efficient learning of sparse representations with an energy-based model. Adv. Neural Inf. Process. Syst. 19, 1137–1144.
Rotskoff, G., Vanden-Eijnden, E., 2018. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. Adv. Neural Inf. Process. Syst. 31, 7146–7155.
Rubin, B., 2004. Convolution–backprojection method for the k-plane transform, and Calderón's identity for ridgelet transforms. Appl. Comput. Harmon. Anal. 16, 231–242.
Rubin, B., 2018. A note on the Blaschke-Petkantschin formula, Riesz distributions, and Drury's identity. Fract. Calc. Appl. Anal. 21, 1641–1650.
Sala, F., De Sa, C., Gu, A., Re, C., 2018. Representation tradeoffs for hyperbolic embeddings. In: Proceedings of the 35th International Conference on Machine Learning. Vol. 80, pp. 4460–4469.
Savarese, P., Evron, I., Soudry, D., Srebro, N., 2019. How do infinite width bounded norm networks look in function space? In: Proceedings of the 32nd Conference on Learning Theory. Vol. 99, pp. 2667–2690.
Shimizu, R., Mukuta, Y., Harada, T., 2021. Hyperbolic neural networks++. In: International Conference on Learning Representations.
Sirignano, J., Spiliopoulos, K., 2020. Mean field analysis of neural networks: A law of large numbers. SIAM J. Appl. Math. 80, 725–752.
Sonoda, S., Hashimoto, Y., Ishikawa, I., Ikeda, M., 2023a. Deep ridgelet transform: Voice with koopman operator proves universality of formal deep networks. In: Proceedings of the 2nd NeurIPS Workshop on Symmetry and Geometry in Neural Representations.
Sonoda, S., Ishi, H., Ishikawa, I., Ikeda, M., 2023b. Joint group invariant functions on data-parameter domain induce universal neural networks. In: Proceedings of the 2nd NeurIPS Workshop on Symmetry and Geometry in Neural Representations.
Sonoda, S., Ishikawa, I., Ikeda, M., 2021a. Ghosts in neural networks: Existence, structure and role of infinite-dimensional null space. arXiv preprint: 2106.04770.
Sonoda, S., Ishikawa, I., Ikeda, M., 2021b. Ridge regression with over-parametrized two-layer networks converge to ridgelet spectrum. In: Proceedings of the 24th International Conference on Artificial Intelligence and Statistics 2021. Vol. 130, pp. 2674–2682.
Sonoda, S., Ishikawa, I., Ikeda, M., 2022a. Fully-connected network on noncompact symmetric space and ridgelet transform based on Helgason-Fourier analysis. In: Proceedings of the 39th International Conference on Machine Learning. Vol. 162, pp. 20405–20422.
Sonoda, S., Ishikawa, I., Ikeda, M., 2022b. Universality of group convolutional neural networks based on ridgelet analysis on groups. Adv. Neural Inf. Process. Syst. 35, 38680–38694.
Sonoda, S., Murata, N., 2017. Neural network with unbounded activation functions is universal approximator. Appl. Comput. Harmon. Anal. 43, 233–268.
Starck, J.-L., Murtagh, F., Fadili, J.M., 2010. The ridgelet and curvelet transforms. In: Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity. Cambridge University Press, pp. 89–118.
Suzuki, T., 2020. Generalization bound of globally optimal non-convex neural network training: Transportation map estimation by infinite dimensional langevin dynamics. Adv. Neural Inf. Process. Syst. 33, 19224–19237.
Terras, A., 2016. Harmonic analysis on symmetric spaces—Higher rank spaces. In: Positive Definite Matrix Space and Generalizations. Springer New York.
Unser, M., 2019. A representer theorem for deep neural networks. J. Mach. Learn. Res. 20, 1–30.
Yamasaki, H., Subramanian, S., Hayakawa, S., Sonoda, S., 2023. Quantum ridgelet transform: Winning lottery ticket of neural networks with quantum computation. In: Proceedings of the 40th International Conference on Machine Learning. Vol. 202, pp. 39008–39034.
Yarotsky, D., 2022. Universal approximations of invariant maps by neural networks. Constr. Approx. 55, 407–474.
Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J., 2017. Deep sets. Adv. Neural Inf. Process. Syst. 30.
Zhou, D.-X., 2020. Universality of deep convolutional neural networks. Appl. Comput. Harmon. Anal. 48, 787–794.