Joint Group Invariant Functions on Data-Parameter Domain Induce Universal Neural Networks

Sho Sonoda[1] SHO.SONODA@RIKEN.JP
Hideyuki Ishi[2] HIDEYUKI-ISHI@OMU.AC.JP
Isao Ishikawa[3,1] ISHIKAWA.ISAO.ZX@EHIME-U.AC.JP
Masahiro Ikeda[1] MASAHIRO.IKEDA@RIKEN.JP
[1] *Center for Advanced Intelligence Project (AIP), RIKEN*
[2] *Osaka Central Advanced Mathematical Institute (OCAMI), Osaka Metropolitan University*
[3] *Center for Data Science, Ehime University*

Editors: Sophia Sanborn, Christian Shewmake, Simone Azeglio, Nina Miolane

Abstract

The symmetry and geometry of input data are considered to be encoded in the internal data representation inside the neural network, but the specific encoding rule has been less investigated. In this study, we present a systematic method to induce a generalized neural network and its right inverse operator, called the *ridgelet transform*, from a *joint group invariant function* on the data-parameter domain. Since the ridgelet transform is an inverse, (1) it can describe the arrangement of parameters for the network to represent a target function, which is understood as the *encoding rule*, and (2) it implies the *universality* of the network. Based on the group representation theory, we present a new simple proof of the universality by using Schur's lemma in a unified manner covering a wide class of networks, for example, the original ridgelet transform, formal *deep* networks, and the dual voice transform. Since traditional universality theorems were demonstrated based on functional analysis, this study sheds light on the group theoretic aspect of the approximation theory, connecting geometric deep learning to abstract harmonic analysis.

Keywords: ridgelet transform, universality, joint group invariant function, Schur's lemma

1. Introduction

The internal data representation of neural networks is expected to reflect the symmetry and geometry of the data domain. In geometric deep learning (Bronstein et al., 2021), several authors have developed novel network architectures that are compatible with the geometric structure of the data (e.g. group equivariant networks). However, these methods typically require handcrafting the network architecture for each specific symmetry and geometry. In this study, we present a systematic method to induce a generalized neural network and its right inverse operator, called the *ridgelet transform*, from a *joint group invariant function* on the data-parameter domain. Since the ridgelet transform is an inverse, (1) it explicitly describes the arrangement of parameters for the network to represent a target function, and (2) it implies the *universality* of the network.

Remark 1 *Our reviewers have kindly let us know that Cohen et al. (2019), Finzi et al. (2021), and Aslan et al. (2023) have proposed versatile group equivariant network architectures that cover a wide class of groups in a unified manner, and Ravanbakhsh et al. (2017) have investigated the symmetry in the parameters. Since our results are applicable to* any *network architectures, it would be interesting to find the ridgelet transform for each network.*

© 2023 S. Sonoda, H. Ishi, I. Ishikawa & M. Ikeda.

The proof of a universality theorem contains hints for understanding the internal data processing mechanisms inside neural networks. The year 1989 was the beginning of the universality theorem and a great year, as four different proofs were presented by Cybenko (1989), Hornik et al. (1989), Funahashi (1989), and Carroll and Dickinson (1989). Among them, Cybenko's proof using Hahn-Banach and Hornik et al.'s proof using Stone-Weierstrass are existential proofs, meaning that it is not clear how to assign the parameters. On the other hand, Funahashi's proof reducing to the Fourier transform and Carroll and Dickinson's proof reducing to the Radon transform are constructive proofs, meaning that it is clear how to assign the parameters. The latter constructive methods, which reduce to integral transforms, were refined as the so-called integral representation by Barron (1993) and further culminated as the ridgelet transform discovered by Murata (1996) and Candès (1998).

The ridgelet transform, the main topic of this study, is a pseudo-inverse operator of the integral representation neural network and is a detailed analysis tool that can describe the relationship between data and parameters due to its analytical representation. In the 2000s, thanks to the efforts of Donoho and others, research on ridgelet transforms evolved into geometric multiscale analysis (GMA, see e.g. Donoho, 2002), leading to the development of various x-lets such as curvelets (Candès and Donoho, 2004), contourlet (Do and Vetterli, 2005), shearlet (Labate et al., 2005), bandelet (Pennec and Mallat, 2005), and grouplet (Mallat, 2009b). These lines of studies mainly focused on developing multidimensional wavelet transforms for image processing (i.e., 2D signals) (Starck et al., 2010; Mallat, 2009a) and gradually moved apart from neural networks.

In the 2020s, the concept of integral representations has re-emerged as tools for analyzing deep learning theories, bringing renewed attention to ridgelet transforms. Precisely, they are often referred to by different names such as overparametrization, continuous/infinite width, mean field theory (Nitanda and Suzuki, 2017; Mei et al., 2018; Rotskoff and Vanden-Eijnden, 2018; Chizat and Bach, 2018; Sirignano and Spiliopoulos, 2020), and Langevin dynamics (Suzuki, 2020). Sonoda et al. (2022b,a) have developed ridgelet transforms for various networks, such as group convolutional networks and networks on manifolds, and have shown constructive universality theorems. In these proofs, reducing the network to Fourier transforms was an essential step to find the ridgelet transforms. In this study, we can find the ridgelet transforms even when there is no clear path to reducing them to Fourier transforms, as long as we can find a group invariant function.

The theory of function expansion based on group representations is well investigated in abstract harmonic analysis (Folland, 2015). There are two main streams: one is the generalization of *Fourier transform*, which expands functions on group G as a sum/integration of multiple irreducible unitary representations (Sugiura, 1990), and the other is the generalization of *wavelet transform* called the *voice transform*, which expands functions in representation space $\mathcal{H}$ as a sum/integration of functions generated by a single square-integrable unitary representation (Holschneider, 1998; Berge, 2021). For example, recent studies by Miyato et al. (2022) and Koyama et al. (2023) belong to the Fourier stream, while this study belongs to the wavelet/voice stream. Yet, it is precisely a new integral transform that differs from the conventional voice transform. The generalized ridgelet transform discovered in this study was motivated by the research objective of geometrically analyzing the parameters of neural networks, and we believe it is a missing link for connecting geometric deep learning to abstract harmonic analysis.

2. Preliminaries

We showcase the original integral representation and the ridgelet transform, a mathematical model of depth-2 fully-connected network and its right inverse, then list a few facts in the group representation theory.

Notation. For any topological space X, $C_c(X)$ denotes the Banach space of all compactly supported functions f on X. $\mathcal{S}(\mathbb{R}^d)$ and $\mathcal{S}'(\mathbb{R}^d)$ denote the classes of rapidly decreasing functions (or Schwartz test functions) and tempered distributions on $\mathbb{R}^d$, respectively.

2.1. Quick Introduction to Integral Representation and Ridgelet Transform

Definition 2 *For any measurable function $\sigma : \mathbb{R} \to \mathbb{C}$ and Borel measure γ on $\mathbb{R}^m \times \mathbb{R}$, put*

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

We call $S_\sigma[\gamma]$ an (integral representation of) neural network, and γ a parameter distribution.

The integration over all the hidden parameters $(\boldsymbol{a}, b) \in \mathbb{R}^m \times \mathbb{R}$ means all the neurons $\{\boldsymbol{x} \mapsto \sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b) \mid (\boldsymbol{a}, b) \in \mathbb{R}^m \times \mathbb{R}\}$ are summed (or integrated, to be precise) with weight γ, hence formally $S_\sigma[\gamma]$ is understood as a continuous neural network with a single hidden layer. We note, however, when γ is a finite sum of point measures such as $\gamma_p = \sum_{i=1}^p c_i \delta_{(\boldsymbol{a}_i, b_i)}$, then it can also reproduce a finite width network

$$S_\sigma[\gamma_p](\boldsymbol{x}) = \sum_{i=1}^p c_i \sigma(\boldsymbol{a}_i \cdot \boldsymbol{x} - b_i). \tag{2}$$

In other words, the integral representation is a mathmatical model of depth-2 network with *any* width (ranging from finite to continuous).

Definition 3 *For any measurable functions $\rho : \mathbb{R} \to \mathbb{C}$ and $f : \mathbb{R}^m \to \mathbb{C}$, put*

$$R_\rho[f](\boldsymbol{a}, b) := \int_{\mathbb{R}^m} f(\boldsymbol{x})\overline{\rho(\boldsymbol{a} \cdot \boldsymbol{x} - b)}\mathrm{d}\boldsymbol{x}, \quad (\boldsymbol{a}, b) \in \mathbb{R}^m \times \mathbb{R}. \tag{3}$$

We call R_ρ a ridgelet transform.

The ridgelet transform is known to be a right-inverse operator to S_σ. To be precise, the following reconstruction formula holds.

Theorem 4 (Reconstruction Formula) *Suppose σ and ρ are a tempered distribution ($\mathcal{S}'$) and a rapid decreasing function ($\mathcal{S}$) respectively. There exists a bilinear form $(\!(\sigma, \rho)\!)$ such that*

$$S_\sigma \circ R_\rho[f] = (\!(\sigma, \rho)\!) f, \tag{4}$$

for any square integrable function $f \in L^2(\mathbb{R}^m)$. Further, the bilinear form is given by

$$(\!(\sigma, \rho)\!) = \int_{\mathbb{R}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-m}\mathrm{d}\omega \tag{5}$$

where $\sharp$ denotes the 1-dimensional Fourier transform.

See Sonoda et al. (2021, Theorem 6) for the proof. In particular, according to Sonoda et al. (2021, Lemma 9), for any activation function σ, there always exists ρ satisfying $(\!(\sigma, \rho)\!) = 1$. Here, σ being a tempered distribution means that typical activation functions are covered such as ReLU, step function, tanh, gaussian, etc... We can interpret the reconstruction formula as a universality theorem of continuous neural networks, since for any given data generating function f, a network with output weight $\gamma_f = R_\rho[f]$ reproduces f (up to factor $(\!(\sigma, \rho)\!)$), i.e. $S[\gamma_f] = f$. In other words, the ridgelet transform indicates how the network parameters should be organized so that the network represents an individual function f.

In this study, we showcase a new proof of the reconstruction formula based on the group theoretic arguments, and present a systematic scheme to find the ridgelet transform for a variety of given network architecture based on the symmetry in the data-parameter domain.

2.2. Irreducible Unitary Representation and Schur's Lemma

Let G be a locally compact group, $\mathcal{H}$ be a nonzero Hilbert space, and $U(\mathcal{H})$ be the group of unitary operators on $\mathcal{H}$. For example, any finite group, discrete group, compact group, and finite-dimensional Lie group are locally compact, while an infinite-dimensional Lie group is not locally compact. A *unitary representation* π of G on $\mathcal{H}$ is a group homomorphism that is continuous with respect to the strong operator topology—that is, a map $\pi : G \to U(\mathcal{H})$ satisfying $\pi(gh) = \pi(g)\pi(h)$ and $\pi(g^{-1}) = \pi(g)^{-1} = \pi(g)^*$, and for any $\psi \in \mathcal{H}$ map $G \ni g \mapsto \pi(g)[\psi] \in \mathcal{H}$ is continuous. Suppose $\mathcal{M}$ is a closed subspace of $\mathcal{H}$. $\mathcal{M}$ is called an *invariant* subspace when $\pi(g)\mathcal{M} \subset \mathcal{M}$ for all $g \in G$. Particularly, π is called *irreducible* when it does not admit any nontrivial invariant subspace $\mathcal{M} \neq \{0\}$ nor $\mathcal{H}$.

Let $C(\pi)$ be the set of all bounded linear operators T on Hilbert space $\mathcal{H}$ that commutes with π, namely $C(\pi) := \{T \in B(\mathcal{H}) \mid T\pi(g) = \pi(g)T \text{ for all } g \in G\}$.

Lemma 5 (Schur's lemma) *A unitary representation π of G is irreducible iff $C(\pi)$ only contains scalar multiples of the identity, i.e., $C(\pi) = \{c\,\mathrm{Id} \mid c \in \mathbb{C}\}$ or $\{0\}$.*

See Folland (2015, Theorem 3.5(a)) for the proof.

2.3. Calculus on Locally Compact Group

By Haar's theorem, if G is a locally compact group, then there uniquely exist left and right invariant measures $\mathrm{d}_l g$ and $\mathrm{d}_r g$, satisfying for any $s \in G$ and $f \in C_c(G)$,

$$\int_G f(sg)\mathrm{d}_l g = \int_G f(g)\mathrm{d}_l g, \quad \text{and} \quad \int_G f(gs)\mathrm{d}_r g = \int_G f(g)\mathrm{d}_r g.$$

Let X be a G-space with transitive left (resp. right) G-action $g \cdot x$ (resp. $x \cdot g$) for any $(g, x) \in G \times X$. Then, we can further induce the left (resp. right) invariant measure $\mathrm{d}_l x$ (resp. $\mathrm{d}_r x$) so that for any $f \in C_c(G)$,

$$\int_X f(x)\mathrm{d}_l x := \int_G f(g \cdot o)\mathrm{d}_l g, \quad \text{resp.} \quad \int_X f(x)\mathrm{d}_r x := \int_G f(o \cdot g)\mathrm{d}_r g,$$

where $o \in G$ is a fixed point called the origin.

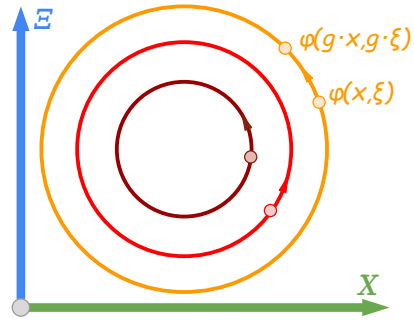

Figure 1: A joint G-invariant function ϕ is a function on the data-parameter domain $X \times \Xi$ that is constant along each G-orbit $G(x, \xi) := \{(g \cdot x, g \cdot \xi) \mid g \in G\}$.

3. Main Results

We introduce generalized neural networks and generalized ridgelet transforms induced from joint group invariant functions on data-parameter domain, and present a simple group theoretic proof of the reconstruction formula.

Let G be a locally compact group equipped with a left invariant measure $\mathrm{d}g$. Let X and Ξ be G-spaces equipped with G-invariant measures $\mathrm{d}x$ and $\mathrm{d}\xi$, called the data domain and the parameter domain, respectively. Particularly, we call the product space $X \times \Xi$ the *data-parameter* domain (like time-frequency domain). By abusing notation, we use the same symbol $\cdot$ for the G-actions on X and Ξ (e.g., $g \cdot x$ and $g \cdot \xi$).

Let π and $\widehat{\pi}$ be left-regular actions of G on $L^2(X)$ and $L^2(\Xi)$, respectively. Namely, for any $g \in G, f \in L^2(X)$ and $\gamma \in L^2(\Xi)$,

$$\pi_g[f](x) := f(g^{-1} \cdot x), \quad \text{and} \quad \widehat{\pi}_g[\gamma](\xi) := \gamma(g^{-1} \cdot \xi). \tag{6}$$

Definition 6 (Joint G-Invariant Function) *We say a function ϕ on $X \times \Xi$ is joint G-invariant when it satisfies for all $g \in G$ and $(x, \xi) \in X \times \Xi$,*

$$\phi(g \cdot x, g \cdot \xi) = \phi(x, \xi). \tag{7}$$

By $\mathcal{A}$, we symbolize the algebra of all joint G-invariant functions.

Here, $\mathcal{A}$ is indeed an *algebra* because if ϕ and ψ are joint G-invariant, then so are $\phi + \psi$ and $\phi\psi$. Namely, $\phi, \psi \in \mathcal{A} \implies \phi + \psi, \phi\psi \in \mathcal{A}$. As visualized in Figure 1, a joint G-invariant function is constant along each G-orbit $\{(g \cdot x, g \cdot \xi) \mid g \in G\}$. Hence finding a joint G-invariant function is not difficult.

Definition 7 (Generalized Neural Network Induced from Invariant ϕ) *For any joint invariant function $\phi \in \mathcal{A}$ and Borel measure γ on Ξ, put*

$$\mathtt{NN}[\gamma; \phi](x) := \int_{\Xi} \gamma(\xi)\phi(x, \xi)\mathrm{d}\xi, \quad x \in X. \tag{8}$$

We call the integral transform $\mathtt{NN}[\bullet; \phi]$ a ϕ-transform, and each individual image $\mathtt{NN}[\gamma; \phi]$ a ϕ-network for short.

The ϕ-network is an extension of the original neural network because when $X = \mathbb{R}^m, \Xi = \mathbb{R}^m \times \mathbb{R}$ and $\phi(\boldsymbol{x}, (\boldsymbol{a}, b)) := \sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)$ with some activation function $\sigma : \mathbb{R} \to \mathbb{R}$, it reduces to a fully-connected network $\int_{\mathbb{R}^m \times \mathbb{R}} \gamma(\boldsymbol{a}, b)\sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)\mathrm{d}\boldsymbol{a}\mathrm{d}b$.

Definition 8 (Generalized Ridgelet Transform Induced from Invariant ϕ) *For any joint invariant map $\phi \in \mathcal{A}$ and measurable function f on X, put*

$$\mathtt{R}[f; \phi](\xi) := \int_X f(x)\overline{\phi(x, \xi)}\mathrm{d}x, \quad \xi \in \Xi. \tag{9}$$

We call the integral transform $\mathtt{R}[\bullet; \phi]$ a ϕ-ridgelet transform for short.

As long as the integrals are convergent, it is the dual operator of ϕ-transform, since

$$\langle \gamma, \mathtt{R}[f; \phi] \rangle_{L^2(\Xi)} = \int_{X \times \Xi} \gamma(\xi)\phi(x, \xi)\overline{f(x)}\mathrm{d}x\mathrm{d}\xi = \langle \mathtt{NN}[\gamma; \phi], f \rangle_{L^2(X)}. \tag{10}$$

Theorem 9 *Let G be a locally compact group. For any joint invariant functions $\phi, \psi \in \mathcal{A}$, suppose that composite $\mathtt{NN}_\phi \circ \mathtt{R}_\psi : L^2(X) \to L^2(X)$ is bounded, and that regular representation $(\pi, L^2(X))$ is irreducible. Then, there exists a bilinear form $(\!(\phi, \psi)\!) \in \mathbb{C}$ such that for any function $f \in L^2(X)$,*

$$\mathtt{NN}_\phi \circ \mathtt{R}_\psi[f] = (\!(\phi, \psi)\!) f. \tag{11}$$

In other words, the ψ-ridgelet transform $\mathtt{R}_\psi$ is understood as a group theoretic generalization of the original ridgelet transform, as it is a right inverse operator of ϕ-transform $\mathtt{NN}_\phi$.

Proof We write $\mathtt{NN}[\bullet; \phi]$ as $\mathtt{NN}_\phi$ and $\mathtt{R}[\bullet; \phi]$ as $\mathtt{R}_\phi$ for short. By the left-invariances of $\mathrm{d}x$ and ψ, for all $g \in G$, we have

$$\begin{aligned}
\mathtt{R}_\psi[\pi_g[f]](\xi) &= \int_X f(g^{-1} \cdot x)\overline{\psi(x, \xi)}\mathrm{d}x && = \langle \pi_g[f], \psi(\bullet, \xi) \rangle_{L^2(X)} \\
&= \int_X f(x)\overline{\psi(g \cdot x, \xi)}\mathrm{d}x && = \langle f, \pi_g^*[\psi](\bullet, \xi) \rangle_{L^2(X)} \\
&= \int_X f(x)\overline{\psi(x, g^{-1} \cdot \xi)}\mathrm{d}x && = \langle f, \widehat{\pi}_g[\psi](\bullet, \xi) \rangle_{L^2(X)} \\
&= \widehat{\pi}_g[\mathtt{R}_\psi[f]](\xi).
\end{aligned} \tag{12}$$

Here, π^* denotes the dual representation of π with respect to $L^2(X)$-product. Similarly,

$$\begin{aligned}
\mathtt{NN}_\phi[\widehat{\pi}_g[\gamma]](x) &= \int_\Xi \gamma(g^{-1} \cdot \xi)\phi(x, \xi)\mathrm{d}\xi && = \langle \widehat{\pi}_g[\gamma], \phi(x, \bullet) \rangle_{L^2(\Xi)} \\
&= \int_\Xi \gamma(\xi)\phi(x, g \cdot \xi)\mathrm{d}\xi && = \langle \gamma, \widehat{\pi}_g^*[\phi](x, \bullet) \rangle_{L^2(\Xi)} \\
&= \int_\Xi \gamma(\xi)\phi(g^{-1} \cdot x, \xi)\mathrm{d}\xi && = \langle \gamma, \pi_g[\phi](x, \bullet) \rangle_{L^2(\Xi)} \\
&= \pi_g[\mathtt{NN}_\phi[\gamma]](x).
\end{aligned} \tag{13}$$

Here, $\widehat{\pi}^*$ denotes the dual representation of $\widehat{\pi}$ with respect to $L^2(\Xi)$-product.

As a consequence, $\mathtt{NN}_\phi \circ \mathtt{R}_\psi : L^2(X) \to L^2(X)$ commutes with π as below

$$\mathtt{NN}_\phi \circ \mathtt{R}_\psi \circ \pi_g = \mathtt{NN}_\phi \circ \widehat{\pi}_g \circ \mathtt{R}_\psi = \pi_g \circ \mathtt{NN}_\phi \circ \mathtt{R}_\psi \tag{14}$$

for all $g \in G$. Hence by Schur's lemma (Lemma 5), there exist a constant $C_{\phi,\psi} \in \mathbb{C}$ such that $\mathtt{NN}_\phi \circ \mathtt{R}_\psi = C_{\phi,\psi} \operatorname{Id}_{L^2(X)}$. By the construction of left-hand side, $C_{\phi,\psi}$ is bilinear in ϕ and ψ. ■

4. Examples

4.1. Original Ridgelet Transform

This study started from a group theoretic proof of the original reconstruction formula (Theorem 4). The proof is in fact new, thought-provoking and valuable, so we leave it in Appendix A. Below is a sketch of the full proof.

Example 1 *Let G be the affine group $\mathrm{Aff}(m) = GL(m) \ltimes \mathbb{R}^m$, $X = \mathbb{R}^m$ be the data domain with G-action*

$$g \cdot \boldsymbol{x} := L\boldsymbol{x} + \boldsymbol{t}, \quad g = (L, \boldsymbol{t}) \in G,\ \boldsymbol{x} \in \mathbb{R}^m = X$$

and $\Xi = \mathbb{R}^m \times \mathbb{R}$ be the parameter domain with dual G-action

$$g \cdot (\boldsymbol{a}, b) = (L^{-\top}\boldsymbol{a}, b + \boldsymbol{t}^\top L^{-\top}\boldsymbol{a}), \quad g = (L, \boldsymbol{t}) \in G,\ (\boldsymbol{a}, b) \in \mathbb{R}^m \times \mathbb{R} = \Xi. \tag{15}$$

We can see $\phi(\boldsymbol{x}, (\boldsymbol{a}, b)) := \sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)$ is joint G-invariant. In fact,

$$\phi(g \cdot \boldsymbol{x}, g \cdot (\boldsymbol{a}, b)) = \sigma\left(L^{-\top}\boldsymbol{a} \cdot (L\boldsymbol{x} + \boldsymbol{t}) - (b + \boldsymbol{t}^\top L^{-\top}\boldsymbol{a})\right) = \sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b) = \phi(\boldsymbol{x}, (\boldsymbol{a}, b)).$$

Further, by Lemma 12, the regular representation π_g of $G = \mathrm{Aff}(m)$ is known to be irreducible. Hence we can retain the original neural network and ridgelet transform:

$$\mathtt{NN}[\gamma](\boldsymbol{x}) = \int_{\mathbb{R}^m \times \mathbb{R}} \gamma(\boldsymbol{a}, b)\sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)\mathrm{d}\boldsymbol{a}\mathrm{d}b, \quad \textit{and} \quad \mathtt{R}[f](\boldsymbol{a}, b) = \int_{\mathbb{R}^m} f(\boldsymbol{x})\overline{\rho(\boldsymbol{a} \cdot \boldsymbol{x} - b)}\mathrm{d}\boldsymbol{x},$$

satisfying $\mathtt{NN} \circ \mathtt{R} = (\!(\sigma, \rho)\!) \operatorname{Id}_{L^2(\mathbb{R}^m)}$.

Additionally, a geometric interpretation of dual G-action (15) is discussed in Appendix B.

4.2. Deep Ridgelet Transform

Sonoda et al. (2023) presented the ridgelet transform for *deep* neural networks. We noticed their network can also be induced from an invariant function. In other words, from the group representation theory perspective, function approximation with *any depth* is unified.

Example 2 *Let G be any locally compact group, data domain X be any G-space, rewriting its G-action $g \cdot x$ as $g(x)$ so as to formally identify g with a hidden layer map, and parameter domain Ξ be the group G itself with dual G-action*

$$g \cdot \xi = \xi g^{-1}. \tag{16}$$

We can see $\phi(x,\xi) := \psi \circ \xi(x)$ *is joint* G*-invariant. In fact,*

$$\phi(g\cdot x, g\cdot \xi) = \psi \circ (g\cdot\xi)(g\cdot x) = \psi\circ(\xi\circ g^{-1})(g(x)) = \psi\circ\xi(x) = \phi(x,\xi)$$

Therefore, assuming that the regular representation $\pi_g = \psi\circ g$ *is irreducible on an invariant subspace* $\mathcal{H}$ *of* $L^2(X)$*, we can retain the formal deep network and deep ridgelet transform:*

$$\mathtt{NN}[\gamma](x) := \int_\Xi \gamma(\xi)\psi\circ\xi(x)\mathrm{d}\xi, \quad \textit{and} \quad \mathtt{R}[f](\xi) = \int_X f(x)\overline{\psi\circ\xi(x)}\mathrm{d}x,$$

satisfying $\mathtt{NN}\circ\mathtt{R} = (\!(\sigma,\rho)\!)\,\mathrm{Id}_{\mathcal{H}}$.

4.3. Voice Transform, or Generalized Wavelet Transform

The voice transform is also known as the *Gilmore–Perelomov coherent states* and the *generalized wavelet transform* (Perelomov, 1986; Ali et al., 2014). It is well investigated in the research field of *coorbit theory* (Feichtinger and Gröchenig, 1988, 1989a,b). We refer to Berge (2021) for a quick review of voice transform and coorbit theory.

Definition 10 *Given a unitary representation* $(\pi,\mathcal{H})$ *of group* G *on a Hilbert space* $\mathcal{H}$*, the voice transform is defined as*

$$V_\phi[f](g) := \langle f, \pi_g[\phi]\rangle_{\mathcal{H}}, \quad g\in G,\ f,\phi\in\mathcal{H}. \tag{17}$$

This unifies several integral transforms from the perspective of group theory such as short-time Fourier transform (STFT), wavelet transform (Grossmann et al., 1985, 1986; Holschneider, 1998; Laugesen et al., 2002; Gressman et al., 2003), and continuous shearlet transform (Labate et al., 2005; Guo and Labate, 2007; Kutyniok and Labate, 2012).

Example 3 *Let* G *be any group, data domain* X *be any* G*-space, and parameter domain* Ξ *be the group* G *itself with dual* G*-action* $g\cdot\xi = g\xi$*. We can see* $\theta(x,\xi) := \psi(\xi^{-1}\cdot x)$ *is joint* G*-invariant. In fact,*

$$\theta(g\cdot x, g\cdot\xi) = \psi((g\cdot\xi)^{-1}\cdot(g\cdot x)) = \psi(\xi^{-1}\cdot x) = \theta(x,\xi).$$

Therefore, assuming that the regular representation π_g *is irreducible, we can retain a dual voice transform and voice transform:*

$$\mathtt{NN}[\gamma](x) := \int_\Xi \gamma(\xi)\phi(\xi^{-1}\cdot x)\mathrm{d}\xi, \quad \textit{and} \quad \mathtt{R}[f](\xi) = \int_X f(x)\psi(\xi^{-1}\cdot x)\mathrm{d}x,$$

satisfying $\mathtt{NN}\circ\mathtt{R} = (\!(\sigma,\rho)\!)\,\mathrm{Id}_{L^2(X)}$*. This is a special case of the voice transform when* $\mathcal{H} = L^2(X)$*, and* $\pi_g[\psi] = \psi(g^{-1}\cdot\bullet)$*.*

We note that the voice transform $V_\phi[f](g) := \langle f,\pi_g[\phi]\rangle_{\mathcal{H}}$ and the ϕ-ridgelet transform $\mathtt{R}_\phi[f](\xi) := \langle f,\phi(\bullet,\xi)\rangle_{L^2(X)}$ have common parts, but are different in general. While the example above and the original wavelet transform $W_\psi[f](b,a) := \int_{\mathbb{R}} f(x)\psi((x-b)/a)\mathrm{d}x/\sqrt{a}$ are simultaneously the voice and ridgelet transforms, a ridgelet transform can be a voice transform only when the representation $(\pi,\mathcal{H})$ is the regular representation on $L^2(X)$, and a voice transform can be a ridgelet transform only when the parameter domain Ξ is the group G itself and the feature map ϕ is generated by G-action on a single function ψ. Hence pursuing parallel results for the coorbit theory would be an interesting future work.

5. Discussion

We presented a systematic method to induce a generalized neural network and its ridgelet transform, from a joint group invariant function on the data-parameter domain. Namely, given a joint group invariant function, the marginalization of parameter ξ (resp. data x) induces the network (resp. the ridgelet transform). Based on the group theoretic arguments, we demonstrated a simple proof of the reconstruction formula by using Schur's lemma, which implies the universality of the network. Since conventional universality theorems were shown using functional analytic tools, the group theoretic proof is a new contribution to the approximation theory, connecting geometric deep learning to abstract harmonic analysis. Further, since the proposed network covers both shallow and deep networks, the group representation theory can offer a unified perspective on function approximation with *any depth*.

In the past, Sonoda et al. (2022a,b) have developed the ridgelet transforms for neural networks on manifolds and function spaces using the Fourier transforms on manifolds and function spaces, and proposed a systematic scheme to derive a ridgelet transform for neural networks on a given domain based on the Fourier transform on there. Compared to our group theoretic method, the Fourier transform method is indirect and requires additional knowledge (not only on the symmetry on the data domain but also) on the Fourier transform on there. We conjecture that those Fourier-based ridgelet transforms can also be derived in our group-theoretic method.

Acknowledgments

The authors are extremely grateful to the three anonymous reviewers for their valuable comments and suggestions, which have helped improve the quality of our manuscript. This work was supported by JSPS KAKENHI 20K03657, JST PRESTO JPMJPR2125, JST CREST JPMJCR2015 and JPMJCR1913, and JST ACTX JPMJAX2004.

References

S. T. Ali, J.-P. Antoine, and J.-P. Gazeau. *Coherent States, Wavelets, and Their Generalizations*. Theoretical and Mathematical Physics. Springer New York, 2 edition, 2014.

B. Aslan, D. Platt, and D. Sheard. Group invariant machine learning by fundamental domain projections. In *Proceedings of the 1st NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, volume 197 of *Proceedings of Machine Learning Research*, pages 181–218. PMLR, 2023.

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

E. Berge. A Primer on Coorbit Theory. *Journal of Fourier Analysis and Applications*, 28 (2):1–61, 2021.

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

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

E. J. Candès and D. L. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C2 singularities. *Communications on Pure and Applied Mathematics*, 57(2):219–266, feb 2004.

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

L. Chizat and F. Bach. On the Global Convergence of Gradient Descent for Overparameterized Models using Optimal Transport. In *Advances in Neural Information Processing Systems 32*, pages 3036–3046, Montreal, BC, 2018.

T. S. Cohen, M. Geiger, and M. Weiler. A General Theory of Equivariant CNNs on Homogeneous Spaces. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

M. N. Do and M. Vetterli. The contourlet transform: an efficient directional multiresolution image representation. *IEEE Transactions on Image Processing*, 14(12):2091–2106, 2005.

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

H. G. Feichtinger and K. Gröchenig. A unified approach to atomic decompositions via integrable group representations. In *Function Spaces and Applications*, pages 52–73, Berlin, Heidelberg, 1988. Springer Berlin Heidelberg.

H. G. Feichtinger and K. H. Gröchenig. Banach spaces related to integrable group representations and their atomic decompositions, I. *Journal of Functional Analysis*, 86(2): 307–340, 1989a.

H. G. Feichtinger and K. H. Gröchenig. Banach spaces related to integrable group representations and their atomic decompositions. Part II. *Monatshefte für Mathematik*, 108 (2):129–148, 1989b.

M. Finzi, M. Welling, and A. G. G. Wilson. A Practical Method for Constructing Equivariant Multilayer Perceptrons for Arbitrary Matrix Groups. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3318–3328. PMLR, 2021.

G. B. Folland. *A Course in Abstract Harmonic Analysis*. Chapman and Hall/CRC, New York, second edition, 2015.

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

P. Gressman, D. Labate, G. Weiss, and N. W. Edward. 8 - Affine, Quasi-Affine and Co-Affine Wavelets. In *Beyond Wavelets*, volume 10, pages 215–223. Elsevier, 2003.

A. Grossmann, J. Morlet, and T. Paul. Transforms associated to square integrable group representations. I. General results. *Journal of Mathematical Physics*, 26(10):2473–2479, oct 1985.

A. Grossmann, J. Morlet, and T. Paul. Transforms associated to square integrable group representations. II : examples. *Annales de l'I.H.P. Physique théorique*, 45(3):293–309, 1986.

K. Guo and D. Labate. Optimally Sparse Multidimensional Representation Using Shearlets. *SIAM Journal on Mathematical Analysis*, 39(1):298–318, 2007.

M. Holschneider. *Wavelets: An Analysis Tool*. Oxford mathematical monographs. Oxford University Press, 1998.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

M. Koyama, K. Fukumizu, K. Hayashi, and T. Miyato. Neural Fourier Transform: A General Approach to Equivariant Representation Learning. *arXiv preprint: 2305.18484*, 2023.

G. Kutyniok and D. Labate. *Shearlets: Multiscale Analysis for Multivariate Data*. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, 1 edition, 2012.

D. Labate, W.-Q. Lim, G. Kutyniok, and G. Weiss. Sparse multidimensional representation using shearlets. In *Proceedings of Society of Photo-Optical Instrumentation Engineers (SPIE), Wavelets XI*, volume 5914, page 59140U, sep 2005.

R. S. Laugesen, N. Weaver, G. L. Weiss, and E. N. Wilson. A characterization of the higher dimensional groups associated with continuous wavelets. *The Journal of Geometric Analysis*, 12(1):89–102, 2002.

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

S. Mallat. Geometrical grouplets. *Applied and Computational Harmonic Analysis*, 26(2): 161–180, 2009b.

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

T. Miyato, M. Koyama, and K. Fukumizu. Unsupervised Learning of Equivariant Structure from Sequences. In *Advances in Neural Information Processing Systems*, 2022.

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

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

E. L. Pennec and S. Mallat. Sparse geometric image representations with bandelets. *IEEE Transactions on Image Processing*, 14(4):423–438, 2005.

A. Perelomov. *Generalized Coherent States and Their Applications*. Theoretical and Mathematical Physics. Springer-Verlag Berlin Heidelberg, 1986.

S. Ravanbakhsh, J. Schneider, and B. Póczos. Equivariance Through Parameter-Sharing. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2892–2901. PMLR, 2017.

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

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

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

S. Sonoda, I. Ishikawa, and M. Ikeda. Universality of Group Convolutional Neural Networks Based on Ridgelet Analysis on Groups. In *Advances in Neural Information Processing Systems 35*, New Orleans, Louisiana, USA, 2022a.

S. Sonoda, I. Ishikawa, and M. Ikeda. 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*, volume 162, pages 20405–20422, Baltimore, Maryland, USA, 2022b.

S. Sonoda, Y. Hashimoto, I. Ishikawa, and M. Ikeda. Deep Ridgelet Transform: Voice with Koopman Operator Proves Universality of Formal Deep Networks. In *to appear in Proceedings of the 2nd NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, Proceedings of Machine Learning Research. PMLR, 2023.

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

M. Sugiura. *Unitary Representations and Harmonic Analysis: An Introduction*, volume 44 of *North-Holland Mathematical Library*. North-Holland, second edition, 1990.

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

Appendix A. Group Theoretic Proof for Original Ridgelet Transform

This study started from a group theoretic proof of the original reconstruction formula (Theorem 4). The proof is in fact new, thought-provoking and valuable, so we present it here in a self-consistent manner. A non group theoretic proof by reducing to a Fourier expression is given in Sonoda et al. (2021, Theorem 6).

A.1. Preliminaries

We will use the following facts without proofs.

Lemma 11 *Suppose σ and ρ are a tempered distribution ($\mathcal{S}'$) and a Schwartz test function, respectively. Then, $S_\sigma \circ R_\rho : L^2(\mathbb{R}^m) \to L^2(\mathbb{R}^m)$ is bounded.*

See Sonoda et al. (2021, Lemmas 7 and 8) for the proof.

Lemma 12 *The regular representation π of the affine group $\mathrm{Aff}(\mathbb{R}^m) := GL(m) \ltimes \mathbb{R}^m$ on $L^2(\mathbb{R}^m)$, namely $\pi(g)[f](\boldsymbol{x}) := |\det L|^{-1/2} f(L^{-1}(\boldsymbol{x} - \boldsymbol{t}))$ for any $f \in L^2(\mathbb{R}^m)$ and $g = (L, \boldsymbol{t}) \in \mathrm{Aff}(\mathbb{R}^m)$, is irreducible.*

See Folland (2015, Theorem 6.42) for the proof.

A.2. Proof

In the following, we identify the group G acting on data domain $\mathbb{R}^m$ with the affine group $\mathrm{Aff}(\mathbb{R}^m)$, and introduce the so-called twisted dual group action that leaves a function θ invariant. Then, we see the regular action π of G on functions space $L^2(\mathbb{R}^m)$ commutes with composite $S_\sigma \circ R_\rho$. Hence, by Schur's lemma, $S_\sigma \circ R_\rho$ is a constant multiple of identity, which concludes the assertion of Theorem 4.

Proof Let G be the affine group $\mathrm{Aff}(\mathbb{R}^m) = GL(\mathbb{R}^m) \ltimes \mathbb{R}^m$. For any $g = (L, \boldsymbol{t}) \in G$, let

$$g \cdot \boldsymbol{x} := L\boldsymbol{x} + \boldsymbol{t}, \quad \boldsymbol{x} \in \mathbb{R}^m \tag{18}$$

be its action on $\mathbb{R}^m$, and let

$$\begin{aligned} \pi(g)[f](\boldsymbol{x}) &:= |\det L|^{-1/2} f(g^{-1} \cdot \boldsymbol{x}) \\ &= |\det L|^{-1/2} f(L^{-1}(\boldsymbol{x} - \boldsymbol{t})), \quad f \in L^2(\mathbb{R}^m) \end{aligned} \tag{19}$$

be its left-regular action on $L^2(\mathbb{R}^m)$.

Besides, putting

$$\theta((\boldsymbol{a}, b), \boldsymbol{x}) := \boldsymbol{a} \cdot \boldsymbol{x} - b, \quad (\boldsymbol{a}, b) \in \mathbb{R}^m \times \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^m \tag{20}$$

we define the *twisted dual action* of G on $\mathbb{R}^m \times \mathbb{R}$ as

$$g \cdot (\boldsymbol{a}, b) := (L^{-\top}\boldsymbol{a}, b + \boldsymbol{a} \cdot (L^{-1}\boldsymbol{t})), \quad (\boldsymbol{a}, b) \in \mathbb{R}^m \times \mathbb{R} \tag{21}$$

so that the following invariance hold:

$$\theta(g \cdot (\boldsymbol{a}, b), g \cdot \boldsymbol{x}) = \theta((\boldsymbol{a}, b), \boldsymbol{x}) = \boldsymbol{a} \cdot \boldsymbol{x} - b. \tag{22}$$

To see this, use matrix expressions with extended variables

$$\theta((\boldsymbol{a},b),\boldsymbol{x}) = \begin{pmatrix}\boldsymbol{a}^\top & b\end{pmatrix}\begin{pmatrix}I_m & 0\\ 0 & -1\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ 1\end{pmatrix} =: \tilde{\boldsymbol{a}}^\top \tilde{I}\tilde{\boldsymbol{x}}, \tag{23}$$

$$\widetilde{g\cdot\boldsymbol{x}} := \begin{pmatrix}g\cdot\boldsymbol{x}\\ 1\end{pmatrix} = \begin{pmatrix}L & \boldsymbol{t}\\ 0 & 1\end{pmatrix}\begin{pmatrix}\boldsymbol{x}\\ 1\end{pmatrix} =: \tilde{L}\tilde{\boldsymbol{x}} \tag{24}$$

and calculate

$$\tilde{\boldsymbol{a}}^\top \tilde{I}\tilde{\boldsymbol{x}} = (\tilde{\boldsymbol{a}}^\top \tilde{I}\tilde{L}^{-1}\tilde{I}^{-1})\tilde{I}(\tilde{L}\tilde{\boldsymbol{x}}) = (\tilde{I}\tilde{L}^{-\top}\tilde{I}\tilde{\boldsymbol{a}})^\top \tilde{I}(\tilde{L}\tilde{\boldsymbol{x}}), \tag{25}$$

which suggests $\widetilde{g\cdot(\boldsymbol{a},b)} := \tilde{I}\tilde{L}^{-\top}\tilde{I}\tilde{\boldsymbol{a}}$, and we have

$$\begin{aligned}\tilde{I}\tilde{L}^{-\top}\tilde{I} &= \begin{pmatrix}I_m & 0\\ 0 & -1\end{pmatrix}\begin{pmatrix}L & \boldsymbol{t}\\ 0 & 1\end{pmatrix}^{-\top}\begin{pmatrix}I_m & 0\\ 0 & -1\end{pmatrix}\\ &= \begin{pmatrix}I_m & 0\\ 0 & -1\end{pmatrix}\begin{pmatrix}L^{-\top} & 0\\ -\boldsymbol{t}^\top L^{-\top} & 1\end{pmatrix}\begin{pmatrix}I_m & 0\\ 0 & -1\end{pmatrix} = \begin{pmatrix}L^{-\top} & 0\\ \boldsymbol{t}^\top L^{-\top} & 1\end{pmatrix}.\end{aligned}$$

Further, we define its regular-action by

$$\begin{aligned}\widehat{\pi}(g)[\gamma](\boldsymbol{a},b) &:= |\det L|^{1/2}\gamma(g^{-1}\cdot(\boldsymbol{a},b))\\ &= |\det L|^{1/2}\gamma(L^\top\boldsymbol{a}, b-\boldsymbol{a}\cdot\boldsymbol{t}), \quad (\boldsymbol{a},b)\in\mathbb{R}^m\times\mathbb{R}.\end{aligned} \tag{26}$$

Then we can see that, for all $g=(L,\boldsymbol{t})\in G$,

$$R_\rho\circ\pi(g) = \widehat{\pi}(g)\circ R_\rho, \quad \text{and} \quad S_\sigma\circ\widehat{\pi}(g) = \pi(g)\circ S_\sigma. \tag{27}$$

In fact, at every $g=(L,\boldsymbol{t})\in G$ and $(\boldsymbol{a},b)\in\mathbb{R}^m\times\mathbb{R}$,

$$R_\rho[\pi(g)[f]](\boldsymbol{a},b) = |\det L|^{-1/2}\int_{\mathbb{R}^m} f(g^{-1}\cdot\boldsymbol{x})\overline{\rho(\theta((\boldsymbol{a},b),\boldsymbol{x}))}\mathrm{d}\boldsymbol{x}$$

by putting $\boldsymbol{x} = g\cdot\boldsymbol{y} = L\boldsymbol{y}+\boldsymbol{t}$ with $\mathrm{d}\boldsymbol{x} = |\det L|\mathrm{d}\boldsymbol{y}$,

$$\begin{aligned}&= |\det L|^{1/2}\int_{\mathbb{R}^m} f(\boldsymbol{y})\overline{\rho(\theta((\boldsymbol{a},b),g\cdot\boldsymbol{y}))})\mathrm{d}\boldsymbol{y}\\ &= |\det L|^{1/2}\int_{\mathbb{R}^m} f(\boldsymbol{y})\overline{\rho(\theta(g^{-1}\cdot(\boldsymbol{a},b),\boldsymbol{y}))})\mathrm{d}\boldsymbol{y}\\ &= \widehat{\pi}(g)[R_\rho[f]](\boldsymbol{a},b).\end{aligned} \tag{28}$$

Similarly, at every $g=(L,\boldsymbol{t})\in G$ and $\boldsymbol{x}\in\mathbb{R}^m$,

$$S_\sigma[\widehat{\pi}(g)[\gamma]](\boldsymbol{x}) = |\det L|^{1/2}\int_{\mathbb{R}^m\times\mathbb{R}}\gamma(g^{-1}\cdot(\boldsymbol{a},b))\sigma(\theta((\boldsymbol{a},b),\boldsymbol{x}))\mathrm{d}\boldsymbol{a}\mathrm{d}b$$

by putting $(\boldsymbol{a}, b) := g \cdot (\boldsymbol{\xi}, \eta) = (L^{-\top}\boldsymbol{\xi}, \eta + \boldsymbol{\xi} \cdot (L^{-1}\boldsymbol{t}))$ with $\mathrm{d}\boldsymbol{a}\mathrm{d}b = |\det L|\mathrm{d}\boldsymbol{\xi}\mathrm{d}\eta$,

$$\begin{aligned}
&= |\det L|^{-1/2} \int_{\mathbb{R}^m \times \mathbb{R}} \gamma(\boldsymbol{\xi}, \eta)\sigma(\theta(g \cdot (\boldsymbol{\xi}, \eta), \boldsymbol{x}))\mathrm{d}\boldsymbol{\xi}\mathrm{d}\eta \\
&= |\det L|^{-1/2} \int_{\mathbb{R}^m \times \mathbb{R}} \gamma(\boldsymbol{\xi}, \eta)\sigma(\theta((\boldsymbol{\xi}, \eta), g^{-1} \cdot \boldsymbol{x}))\mathrm{d}\boldsymbol{\xi}\mathrm{d}\eta \\
&= \pi(g)[S_\sigma[\gamma]](\boldsymbol{x}). \qquad (29)
\end{aligned}$$

Hence $S_\sigma \circ R_\rho$ commutes with $\pi(g)$ because

$$S_\sigma \circ R_\rho \circ \pi(g) = S_\sigma \circ \widehat{\pi}(g) \circ R_\rho = \pi(g) \circ S_\sigma \circ R_\rho.$$

Since $S_\sigma \circ R_\rho : L^2(\mathbb{R}^m) \to L^2(\mathbb{R}^m)$ is bounded (Lemma 11), and $(\pi, L^2(\mathbb{R}^m))$ is an irreducible unitary representation of G (Lemma 12), Schur's lemma (Lemma 5) yields that there exist a constant $C_{\sigma,\rho} \in \mathbb{C}$ such that

$$S_\sigma \circ R_\rho[f] = C_{\sigma,\rho} f \qquad (30)$$

for all $f \in L^2(\mathbb{R}^m)$.

Finally, by directly computing the left-hand-side, namely $S_\sigma \circ R_\rho[f]$, we can verify that the constant $C_{\sigma,\rho}$ is given by

$$C_{\sigma,\rho} = (\!(\sigma, \rho)\!) := (2\pi)^{m-1} \int_{\mathbb{R}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-m}\mathrm{d}\omega. \qquad (31)$$

■

Appendix B. Geometric Interpretation of Dual Action for Original Ridgelet Transform

We explain a geometric interpretation of the dual action (21) in the previous section. We note that in general θ does not require any geometric interpretation as long as it is joint group invariant on data-parameter domain.

For each $(\boldsymbol{a}, b) \in \mathbb{R}^m \times \mathbb{R}$, put $\xi(\boldsymbol{a}, b) := \{\boldsymbol{x} \in \mathbb{R}^m \mid \boldsymbol{a} \cdot \boldsymbol{x} - b = 0\}$. Then it is a hyperplane in $\mathbb{R}^m$ through point $\boldsymbol{x}_0 = b\boldsymbol{a}/|\boldsymbol{a}|^2$ with normal vector $\boldsymbol{u} := \boldsymbol{a}/|\boldsymbol{a}|$.

For any point $\boldsymbol{y}$ in the hyperplane $\xi(\boldsymbol{a}, b)$, by definition $\boldsymbol{a} \cdot \boldsymbol{y} = b$, thus

$$\boldsymbol{a} \cdot \boldsymbol{x} - b = \boldsymbol{a} \cdot (\boldsymbol{x} - \boldsymbol{y}). \qquad (32)$$

But this means $\boldsymbol{a} \cdot \boldsymbol{x} - b$ is a scaled distance between point $\boldsymbol{x}$ and hyperplane $\xi(\boldsymbol{a}, b)$,

$$= |\boldsymbol{a}| d_E(\boldsymbol{x}, \xi(\boldsymbol{a}, b)), \qquad (33)$$

and further a scaled distance between hyperplanes $\xi(\boldsymbol{a}, \boldsymbol{a} \cdot \boldsymbol{x})$ through $\boldsymbol{x}$ with normal $\boldsymbol{a}/|\boldsymbol{a}|$ and $\xi(\boldsymbol{a}, b)$,

$$= |\boldsymbol{a}| d_E(\xi(\boldsymbol{a}, \boldsymbol{a} \cdot \boldsymbol{x}), \xi(\boldsymbol{a}, b)). \qquad (34)$$

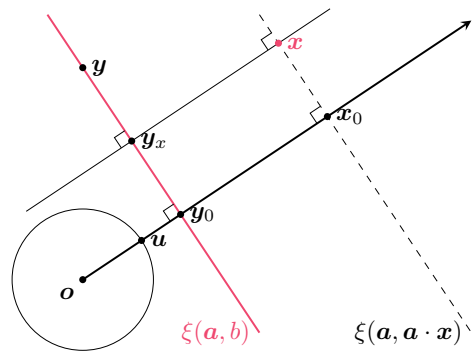

Figure 2: The invariant $\phi((\boldsymbol{a}, b), \boldsymbol{x}) = \sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)$ is the euclidean distance between point $\boldsymbol{x}$ and hyperplane $\xi(\boldsymbol{a}, b)$ followed by scaling and nonlinearity σ

Now, we can interpret the invariant $\theta((\boldsymbol{a}, b), \boldsymbol{x}) := \boldsymbol{a} \cdot \boldsymbol{x} - b$ in a geometric manner, that is, it is the distance between point and hyperplane, or between hyperplanes. We note that we can regard entire $\sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)$—the distance modulated by both scaling and nonlinearity—as the invariant, say ϕ.

Furthermore, the dual action $g \cdot (\boldsymbol{a}, b)$ is understood as a parallel translation of hyperplane $\xi(\boldsymbol{a}, b)$ to $\xi(g \cdot (\boldsymbol{a}, b))$ so as to leave the scaled distance θ invariant, namely

$$d_E(g \cdot \boldsymbol{x}, g \cdot \xi(\boldsymbol{a}, b)) = d_E(\boldsymbol{x}, \xi(\boldsymbol{a}, b)). \tag{35}$$

Indeed, for any $g = (L, \boldsymbol{t}) \in G$,

$$\begin{aligned} g \cdot \xi(\boldsymbol{a}, b) &= \{g \cdot \boldsymbol{x} \mid \boldsymbol{a} \cdot \boldsymbol{x} - b = 0\} \\ &= \{\boldsymbol{y} \mid \boldsymbol{a} \cdot (g^{-1} \cdot \boldsymbol{y}) - b = 0\} && \text{(by letting } \boldsymbol{y} = g \cdot \boldsymbol{x}) \\ &= \{\boldsymbol{y} \mid (L^{-\top}) \cdot \boldsymbol{y} - (b + \boldsymbol{a} \cdot (L^{-1}\boldsymbol{t})) = 0\} \\ &= \xi(g \cdot (\boldsymbol{a}, b)), \end{aligned}$$

meaning that the hyperplane with parameter $(\boldsymbol{a}, b)$ translated by g is identical to the hyperplane with parameter $g \cdot (\boldsymbol{a}, b)$.

To summarize, in the case of fully-connected neural network (and its corresponding ridgelet transform), the invariant is a modulated distance $\sigma(\boldsymbol{a} \cdot \boldsymbol{x} - b)$, and the dual action is the parallel translation of hyperplane so as to keep the distance invariant. Further, from this geometric perspective, we can rewrite the fully-connected neural network in a geometric manner as

$$S[\gamma](\boldsymbol{x}) := \int_{\mathbb{R} \times \Xi} \gamma(\xi) \sigma(a d_E(\boldsymbol{x}, \xi)) \mathrm{d}a \mathrm{d}\xi, \tag{36}$$

where $a \in \mathbb{R}$ denotes signed scale and Ξ denotes the space of all hyperplanes (not always through the origin). Since each hyperplane is parametrized by normal vectors $\boldsymbol{u} \in \mathbb{S}^{m-1}$ and distance $p \geq 0$ from the origin, we can induce the product of spherical measure $\mathrm{d}\boldsymbol{u}$ and Lebesgue measure $\mathrm{d}p$ as a measure $\mathrm{d}\xi$ on the space Ξ of hyperplanes.