Universality of group convolutional neural networks based on ridgelet analysis on groups

Anonymous Author(s)
Affiliation
Address
email

Abstract

We investigate the approximation property of group convolutional neural networks (GCNNs) based on the ridgelet theory. We regard a group convolution as a matrix element of a group representation, and formulate a versatile GCNN as a nonlinear mapping between group representations, which covers typical GCNN literatures such as a cyclic convolution on a multi-channel image, permutation-invariant datasets (Deep Sets), and $\mathrm{E}(n)$-equivariant convolutions. The ridgelet transform is an analysis operator of a depth-2 network, namely, it maps an arbitrary given target function f to the weight γ of a network $S[\gamma]$ so that the network represents the function as $S[\gamma] = f$. It has been known only for fully-connected networks, and this study is the first to present the *ridgelet transform for (G)CNNs*. Since the ridgelet transform is given as a closed-form integral operator, it provides a constructive proof of the *cc-universality of GCNNs*. Unlike previous universality arguments on CNNs, we do not need to convert/modify the networks into other universal approximators such as invariant polynomials and fully-connected networks.

1 Introduction

In the research field of geometric deep learning [1], *group convolutional neural networks (GCNNs)* have been developed to capture the inductive bias behind a variety of datasets such as sets and point clouds [2, 3], graphs [4, 5], manifolds, groups and homogeneous spaces [6–8, 4]. In the theoretical aspect, however, the *approximation property* of CNNs is less investigated than that of fully-connected neural networks (FNNs). To this date, several authors have shown that some CNNs are *universal*, that is, they can approximate some class of continuous maps with any precision [9–15]. These studies are still limited because (1) the network components such as the activation function and convolution are limited, and/or (2) the proofs are indirect or existential since they rely on the universality of other universal approximators such as FNNs and invariant polynomials. (See § 6 for more details).

The approximation property of *FNNs* has been investigated in the 1990s, with gradually increasing the resolution of proofs from abstract to concrete, starting from existential proofs based on the Hahn-Banach theorem [16] and the Stone-Weierstrass theorem [17], indirect proofs based on the Fourier transform [18, 19], the Radon transform [20, 21], B-splines [22, 23], to constructive proofs based on the integral representation [24], ridge functions [25], and the *ridgelet transform* [26–28]. For deep-ReLU-FNNs, further approximation properties have been investigated [29–31] in the 2010s.

In this study, we show the universality of GCNNs by developing the *ridgelet transform for GCNNs*—an integral transform (or analysis operator) that maps a given function f to the weight parameter γ in a single hidden layer of a neural network. By virtue of the ridgelet transform, our universality proof is more *unified and constructive*. Some previous studies have also presented construction algorithms of network parameters, but they are case-by-case, and if the network architecture changes, the construction algorithm has to be manually reprogrammed from scratch. On the contrary, our ridgelet

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

transform is devised for a wide range of GCNNs in a unified manner, and its discretization (numerical integration, or finite sum approximation) process automatically turns out to be a construction algorithm of network parameters.

To overview the main contributions, we describe the formulation of GCNNs.

A Typical Convolution Layer for Images. Given an $m_1 \times m_2$-dimensional n_{in}-channel input image $\boldsymbol{x} \in \mathbb{R}^{m_1 \times m_2 \times n_{in}}$, a typical *convolution layer* with $w_1 \times w_2$-dimensional $n_{in} \times n_{out}$-channel filter $\boldsymbol{a} \in \mathbb{R}^{w_1 \times w_2 \times n_{in} \times n_{out}}$ and n_{out}-channel bias $\boldsymbol{b} \in \mathbb{R}^{n_{out}}$ followed by the elementwise activation function $\sigma : \mathbb{R} \to \mathbb{R}$ and the aggregation with output coefficients $\boldsymbol{c} \in \mathbb{R}^{n_{out}}$ is given by

$$S[\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}](\boldsymbol{x})(i,j) = \sum_{\ell=1}^{n_{out}} c^\ell \sigma \left(\sum_{k=1}^{n_{in}} \sum_{p=1}^{w_1} \sum_{q=1}^{w_2} a_{pq}^{k\ell} x_{i+p,j+q}^k - b^\ell \right) \tag{1}$$

for each pixel at $(i,j) \in (m_1 - w_1 + 1) \times (m_2 - w_2 + 1)$.

Here, in this study, the output channels (indexed by n_{out}) are aggregated soon after the activation function, which may be slightly different from an ordinary formulation of CNNs. We impose this aggregation for technical reasons, but we may understand this as a part of the subsequent layer.

In the standard formulation of GCNNs, the convolution in the pixel directions (i,j) is reformulated as a group convolution with respect to the product group $G = \mathbb{Z}_{m_1} \times \mathbb{Z}_{m_2}$ of cyclic groups $\mathbb{Z}_{m_i}$ $(i = 1, 2)$. So, the input image is understood as a function on a group G (or a homogeneous space G/H on which G acts). In addition, the inner product in the channel direction k is also understood as a part of the convolution with respect to the trivial action of G on the 'fiber' $\mathbb{R}^{n_{in}}$ ([32]). In this sense, the multi-channel image is understood as a vector-valued function (or a section of a fiber bundle).

The Integral Representation $S[\gamma]$ of Group Convolution Layer. Let G be an *arbitrary* group, $\sigma : \mathbb{R} \to \mathbb{R}$ be an *arbitrary* nonlinear function, and $\mathcal{H}$ be an *arbitrary* vector space of feature vector x and filter a. In this study, we formulate a group convolution layer in an integral form, called the *integral representation*, as

$$S[\gamma](x)(g) := \int_{\mathcal{H} \times \mathbb{R}} \gamma(a,b) \sigma\big((a * x)(g) - b\big) \mathrm{d}a\mathrm{d}b, \quad x \in \mathcal{H},\ g \in G. \tag{2}$$

This is an infinite-dimensional linear representation of a depth-2 group convolutional neural network. Here, each function $x \mapsto \sigma((a * x)(g) - b)$ represents a single convolutional neuron, or a feature map of input x parametrized by (a,b). The integration over (a,b) implies that all the possible neurons are assigned in advance, and thus $S[\gamma]$ can be understood as a *continuous neural network*. We note, however, that if we take γ to be a finite sum of Dirac's measures such as $\gamma_n := \sum_{\ell=1}^n c^\ell \delta_{(a^\ell, b^\ell)}$, then the integral representation can also exactly reproduce a finite model: $S[\gamma_n](x)(g) = \sum_{\ell=1}^n c^\ell \sigma((a^\ell * x)(g) - b^\ell)$ for $x \in \mathcal{H}, g \in G$. In summary, $S[\gamma]$ is a mathematical model of shallow neural networks with *any* width ranging from finite to infinite. In particular, if the parameter distribution γ is localized and concentrated, then it means the filters a^ℓ are sparse and low-rank.

An advantage to use the integral representation is the *linearization* effect (of a single hidden layer). Whereas a finite network is *nonlinear* in the original parameters (a,b), the integral representation is *linear* in the parameter distribution γ. It is first emerged in the 1990s to investigate the expressive power of infinitely-wide shallow FNNs [18–21, 24, 26–28]; and it is as well common in today's deep learning theory, for example, to investigate the learning dynamics of SGD such as neural tangent kernel (NTK) [33–35], lazy learning [36], lottery tickets [37], mean field theory [38–42], and Langevin dynamics [43].

The Ridgelet Transform $R[f;\rho](a,b)$ is a right inverse (or pseudo-inverse) operator of the integral representation operator S. It is given as a closed-form integral operator

$$R[f;\rho](a,b) := \int_{\mathcal{H}} f(x)(e) \overline{\rho(\langle a, x \rangle_{\mathcal{H}} - b)} \mathrm{d}a\mathrm{d}b, \quad (a,b) \in \mathcal{H} \times \mathbb{R}, \tag{3}$$

where $f : \mathcal{H} \to \mathbb{C}^G$ is a target vector-valued function to be approximated, $e \in G$ is the identity element, and $\rho : \mathbb{R} \to \mathbb{C}$ is an auxiliary function, called the *ridgelet function*. Under mild conditions, it satisfies the *reconstruction formula*

$$S[R[f;\rho]] = (\!(\sigma, \rho)\!) f, \tag{4}$$

where $(\!(\cdot,\cdot)\!)$ denote a scalar product of σ and ρ. Therefore, as long as the product $(\!(\sigma,\rho)\!)$ is neither 0 nor ∞, we can normalize ρ to satisfy $(\!(\sigma,\rho)\!) = 1$ so that $S[R[f;\rho]] = f$.

In other words, R and S are analysis and synthesis operators, and thus play the same roles as the Fourier (F) and inverse Fourier (F^{-1}) transforms respectively, in the sense that the reconstruction formula $S[R[f;\rho]] = (\!(\sigma,\rho)\!) f$ corresponds to the Fourier inversion formula $F^{-1}[F[f]] = f$.

An advantage to use the ridgelet transform is the *closed-form expression*. Despite the common belief that neural network parameters are a blackbox, the closed-form expression of ridgelet transform clearly describes how the network parameters are organized. Previous studies on the CNN universality have also provided several construction algorithms of parameters, but these are only *particular* solutions for a CNN to represent a target function f, and not necessary related to, for example, deep learning solutions. For *FNNs*, on the other hand, Sonoda et al. [44, 45] have shown that any parameter distribution γ satisfying $S[\gamma] = f$ can always be represented as (not always single but) a linear combination of ridgelet transforms, and that finite networks trained by regularized empirical risk minimization (RERM) converges to a certain unique ridgelet transform. (We note that other theories such as NTK and Gibbs distribution can also provide the final distribution, but they are more limited or indirect.) Furthermore, Savarese et al. [46] and their followers [47–50] have established the *representer theorems* for ReLU-FNNs by using the ridgelet transform. For CNNs, similar results have not yet been presented, but we expect that the ridgelet transform for GCNN will be a key to facilitate our understanding of deep learning solutions.

Challenges and Contributions. Since the original ridgelet transform was discovered in the 1990s independently by Murata [26], Candès [27] and Rubin [28], the closed-form expression of the ridgelet transform has been known only for FNNs. (We refer to [51–53] for ridgelet analysis in the 2000s, and [54–56] for more recent results.) One of the difficulties to devise the ridgelet transform for CNNs is that there is no unique way to formulate an "integral representation of CNNs". (We note that some authors claim the "equivalence of CNNs and FNNs," which is simply misleading because such an equivalence holds only for very carefully designed CNNs and FNNs.) While FNNs are defined on Euclidean spaces $\mathbb{R}^m$, (G)CNNs are defined on a more abstract space of features, denoted $\mathcal{H}$. For example, since the convolution on the Euclidean space can be written using Töplitz matrices, one could consider a formulation such as $\int \gamma(A,b)\sigma(Ax-b)\mathrm{d}A\mathrm{d}b$ where the parameter A is a matrix, but this leads to a completely different ridgelet transform covering less symmetries (related to the ridgelet transform in the 2000s [51]).

To circumvent this difficulty, we formulate GCNNs as *geometric* as possible by dealing with the feature space $\mathcal{H}$, group G, and representation T in a coordinate-free manner. Eventually, we have shown the reconstruction formula for a wide range of GCNNs (as instantiated in the Examples section), with a relatively *simple* proof. This study is the first to derive the ridgelet transform for a general class of GCNNs. As an application, we show the cc-universality of GCNNs for a general class of group equivalent continuous vector-valued functions in a unified and constructive manner.

2 Preliminaries

2.1 Notation

For any topological space X, $C(X), C_0(X)$ and $C_c(X)$ denote the collections of all continuous functions on X, continuous functions on X that vanish at infinity, and continuous functions on X with compact support, respectively. We note that when X is compact, then these three spaces are equivalent, i.e., $C(X) = C_0(X) = C_c(X)$.

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 $\mathcal{S}$. In this study, we employ $\mathcal{S}'(\mathbb{R})$ and $\mathcal{S}(\mathbb{R})$ to be classes of activation and ridgelet functions, respectively. 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 The Fourier transform on the Euclidean space $\mathbb{R}^d$ and its inversion formula has been defined on (at least) three different function classes: $L^1(\mathbb{R}^d), L^2(\mathbb{R}^d)$ and $\mathcal{S}'(\mathbb{R}^d)$. When $f \in L^1(\mathbb{R}^d)$ and $\widehat{f} \in L^1(\mathbb{R}^d)$, the inversion formula holds "at every continuous point $\boldsymbol{x}$ of f", which

is a pointwise equality. When $f \in L^2(\mathbb{R}^d)$, the inversion formula holds "in L^2", which is not a pointwise equality because the equality "$f = g$ in L^2" is defined as "$f(\boldsymbol{x}) = g(\boldsymbol{x})$ a.e.". Similarly, when $f \in \mathcal{S}'(\mathbb{R}^m)$, the inversion formula holds "in $\mathcal{S}'$". We use the third definition for computing the Fourier transform of activation functions $\sigma \in \mathcal{S}'(\mathbb{R})$ such as ReLU and $\tanh$. We refer to [57, 58, 54] for more details on the Fourier transform for tempered distributions ($\mathcal{S}'$).

2.2 Group Representation

Let G be a group, let $\mathcal{H}$ be a vector space over a field $\mathbb{K}$, and let $GL(\mathcal{H})$ be the general linear group on $\mathcal{H}$. A *group representation* T of the group G on the vector space $\mathcal{H}$ is a group homomorphism from G to $GL(\mathcal{H})$, that is, a map $T : G \to GL(\mathcal{H}); g \mapsto T_g$ satisfying $T_{gh} = T_g T_h$ for all $g, h \in G$. When G is a topological group, we further assume that the action $G \times \mathcal{H} \to \mathcal{H}; (g, x) \mapsto T_g[x]$ be continuous. Here, $\mathcal{H}$ is called the *representation space*, or sometimes the G*-module*.

Regular Representation. Let $\mathcal{H}$ be the vector space of all functions on G, i.e., $\mathcal{H} = \mathbb{K}^G$. The *(left) regular representation* L is a group representation defined on $\mathcal{H}$ as

$$L_g[x](h) := x(g^{-1}h), \quad g, h \in G,\ x \in \mathcal{H} = \mathbb{K}^G. \tag{5}$$

In particular, when G is a locally compact Haussdorf (LCH) group, then it has a (left) invariant measure μ, and we can define the collection $L^2(G)$ of all square integrable functions on G with respect to the canonical inner product $\langle x, y \rangle_{L^2(G)} := \int_G x(g)\overline{y(g)}\mathrm{d}\mu(g)$ for any measurable functions $x, y : G \to \mathbb{C}$. It is known that the regular representation on $\mathcal{H} = L^2(G)$ is a unitary representation.

Dual Representation. For any group representation $T : G \to GL(\mathcal{H})$, the *dual representation* $T^* : G \to GL(\mathcal{H}')$ is a group representation defined on the dual vector space $\mathcal{H}'$ as the transpose of $T_{g^{-1}}$, that is, $T^*_g = T^\top_{g^{-1}}$. When $\mathcal{H}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, then it satisfies the following relation:

$$\langle T_g[x], y \rangle_{\mathcal{H}} = \langle x, T^*_{g^{-1}}[y] \rangle_{\mathcal{H}}, \quad g \in G,\ x, y \in \mathcal{H}. \tag{6}$$

Matrix Element. For any group representation $T : G \to GL(\mathcal{H})$, the *matrix element* (or the *matrix coefficient*) of T is a bilinear functional $f_{a,x}$ on G defined by

$$f_{a,x}(g) := a[T_g[x]], \quad g \in G,\ x \in \mathcal{H},\ a \in \mathcal{H}' \tag{7}$$

where x is a vector in $\mathcal{H}$ and $a \in \mathcal{H}'$ is a continuous linear functional on $\mathcal{H}$. When $\mathcal{H}$ is a Hilbert space, then (abusing the notations $a \in \mathcal{H}'$ and $a \in \mathcal{H}$) it can be written as

$$f_{a,x}(g) = \langle T_g[x], a \rangle_{\mathcal{H}}, \quad g \in G,\ a, x \in \mathcal{H}. \tag{8}$$

In the next section, we use this quantity as the generalized form of the *group convolution*.

2.3 Universality

The notion of *universality* in machine learning can be rephrased as the *density* in mathematics, and thus it has several definitions [59, 60]. In this study, we show the so-called *cc-univesality*, one of the standard universalities in the neural network theory.

cc-**Universality.** Let X be a topological space, and let NN be a collection of functions (e.g., neural networks) on X. The *cc*-univesality of NN is defined as the density of NN in $C(X)$ endowed with the topology of *compact convergence*, that is, for any compact subset $K \subset X$, continuous function $f \in C(K)$, and all $\varepsilon > 0$, there exists a function $g \in \mathrm{NN}$ such that

$$\left\| f - g|_K \right\|_{C(K)} := \sup_{x \in K} |f(x) - g(x)| < \varepsilon, \tag{9}$$

where $g|_K$ denotes the restriction of g to K.

3 Formulations

We introduce the notions of (G, T)-convolution and (G, T)-equivariance, which are natural extensions of group convolution and group equivariance, respectively. Then, we introduce analysis tools for abstract feature vectors $f : \mathcal{H} \to \mathbb{C}^G$ such as a norm and a measure.

3.1 (G,T)-Convolution

Definition 1. Let G be a group, let $\mathcal{H}$ be a Hilbert space with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ over a field $\mathbb{K}$, and let $T : G \to GL(\mathcal{H})$ be a representation of G on $\mathcal{H}$. For any $a, x \in \mathcal{H}$ and $g \in G$, we define the (G,T)-*convolution* as

$$(a *_T x)(g) := \langle x, T_g^*[a]\rangle_{\mathcal{H}} = \langle T_{g^{-1}}[x], a\rangle_{\mathcal{H}}. \tag{10}$$

First of all, this is simply a paraphrase of the *matrix element of a group representation* (see the previous section). Nevertheless, we call it a *convolution* simply because it covers a wide range of previously proposed 'group convolutions' in the GCNN literature.

For example, an orthodox group convolution is reproduced when T is the regular representation (of a LCH group G) on $\mathcal{H} = L^2(G)$, i.e., $T_g^*[a](h) = a(g^{-1}h)$. In fact,

$$\langle x, T_g^*[a]\rangle_{L^2(G)} = \int_G x(h)\overline{a(g^{-1}h)}\mathrm{d}\mu(h) = \int_G x(h)\tilde{a}(h^{-1}g)\mathrm{d}\mu(h) = (x *_G \tilde{a})(g), \tag{11}$$

where $\tilde{a}(g) := \overline{a(g^{-1})}$ is an involution. Specifically, the cyclic convolution for an n-channel image $\boldsymbol{x} = (x_{ij}^k) \in \mathbb{R}^{m_1\times m_2\times n}$ is understood as the case when $G = \mathbb{Z}_{m_1} \times \mathbb{Z}_{m_2}$, $\mathcal{H} = \mathbb{R}^{m_1\times m_2\times n}$ and

$$T_{(p,q)}^*[\boldsymbol{a}](i,j,k) = a_{i-p,j-q}^k \quad \text{thus} \quad \langle \boldsymbol{x}, T_{(p,q)}^*[\boldsymbol{a}]\rangle_{\mathbb{R}^{m_1\times m_2\times n}} = \sum_{i,j,k} x_{ij}^k a_{i-p,j-q}^k. \tag{12}$$

On the other hand, the 'input feature space' $\mathcal{H}$ can be an arbitrary abstract vector space, which is more abstract than the conventional group convolution where the feature space is instantiated as a function space on G, such as $L^2(G)$ and $L^2(G/H)$. We consider this as an advantage for geometric understanding of CNNs, since the theory becomes free from the specific instantiation of the feature x.

3.2 (G,T)-Equivariance

Associated with the extended convolution, we introduce the notion of equivariance as below.

Definition 2. We say a function $f : \mathcal{H} \to \mathbb{C}^G$ is (G,T)-*equivariant* when

$$f(T_g[x])(h) = L_g[f(x)](h) = f(x)(g^{-1}h), \quad x \in \mathcal{H},\ g,h \in G. \tag{13}$$

Here, we restrict the term 'equivariance' for a special case of the regular representation, which is simply because our GCNN satisfies this case.

3.3 (G,T)-Equivariant Continuous Vector-Valued Function $f : \mathcal{H} \to C(G)$

Let G be a topological group. For any vector-valued function $f : \mathcal{H} \to \mathbb{C}^G$, put

$$\|f\|_{C(\mathcal{H};C(G))} := \|f\|_{C(\mathcal{H})\to C(G)} := \sup_{x\in\mathcal{H}} \Big| \sup_{g\in G} |f(x)(g)| \Big|. \tag{14}$$

By $C_{equi}(\mathcal{H}; C(G))$, we denote the normed vector space of all (G,T)-*equivariant* $C(G)$-*valued continuous* functions on $\mathcal{H}$. We note that in the *cc*-universality argument, we deal with the topology of compact convergence, which is weaker than the topology of uniform norm $\|\cdot\|_{C(\mathcal{H};C(G))}$. Specifically, $C_{equi}(\mathcal{H}; C(G))$ need not be a Banach space.

3.4 Fourier Transform on The Abstract Feature Space $\mathcal{H}$

For the sake of simplicity, we assume that the real dimension $m := \dim_{\mathbb{R}} \mathcal{H}$ of $\mathcal{H}$ is *finite*.

Lebesgue Measure on $\mathcal{H}$. We define the Lebesgue measure λ on $\mathcal{H}$ by pushing forward the Lebesgue measure $\mathrm{d}\boldsymbol{x}$ on $\mathbb{R}^m$ via an isometric linear embedding $\phi : \mathbb{R}^m \to \mathcal{H}$. For example, let $\{\mathsf{e}_i\}_{i\in[m]}$ be an orthonormal basis of $\mathcal{H}$, and put a linear embedding $\phi(\boldsymbol{x}) := \sum_{i\in[m]} x_i \mathsf{e}_i$, then (it preserves the length and) we can induce the Lebesgue measure λ on $\mathcal{H}$ as the push forward measure

$\lambda = \phi_\sharp \mathrm{d}\boldsymbol{x}$ so that the volume of a hypercube $E = \{\sum_{i\in[m]} c_i \mathsf{e}_i \mid c_i \in [a_i, b_i]\}$ in $\mathcal{H}$ is calculated as in $\mathbb{R}^m$: $\lambda(E) = \prod_{i\in[m]} |b_i - a_i|$, and the integration is calculated as

$$\int_E f(x)\mathrm{d}\lambda(x) = \int_{\mathbb{R}^m} 1_X \left(\sum_{i=1}^m x_i \mathsf{e}_i\right) f\left(\sum_{i=1}^m x_i \mathsf{e}_i\right) \prod_{i=1}^m \mathrm{d}x_i = \int_{\phi^{-1}(X)} f \circ \phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{15}$$

for any measurable subset $E \subset \mathcal{H}$ and measurable function $f : \mathcal{H} \to \mathbb{C}$. As far as there is no risk of confusion, we denote $\mathrm{d}x$ instead of $\mathrm{d}\lambda(x)$.

Fourier Transform on $\mathcal{H}$. Based on λ, we induce the Fourier transform on $\mathbb{R}^m$ as below: For any function $f : \mathcal{H} \to \mathbb{C}$,

$$\widehat{f}(y) := \int_{\mathcal{H}} f(x) e^{-i\langle x, y\rangle_{\mathcal{H}}} \mathrm{d}\lambda(x), \quad f(x) \overset{\star}{=} \frac{1}{(2\pi)^m} \int_{\mathcal{H}} \widehat{f}(y) e^{i\langle x, y\rangle_{\mathcal{H}}} \mathrm{d}\lambda(y). \tag{16}$$

Here, the equality $\overset{\star}{=}$ holds in at least three different senses (see the comments in § 2.1).

Again, $\mathcal{H}$ remains abstract. Depending on the embedding ϕ, each instantiation of $\mathcal{H}$ looks different. Nevertheless, these share the equivalent structures induced from $\mathbb{R}^m$. We consider this as also an advantage for geometric understanding of CNNs.

4 Main Results

Based on the formulations in the previous sections, we define the (G, T)-convolutional neural networks and the corresponding ridgelet transform.

Throughout this section, $T : G \to GL(\mathcal{H})$ is a representation of a group G on a (potentially infinite-dimensional) Hilbert space $\mathcal{H}$ over a field $\mathbb{K}$ endowed with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, and $\mathcal{H}_m \subset \mathcal{H}$ is a finite-dimensional closed subspace of $\mathcal{H}$ equipped with the Lebesgue measure λ. Let $P : \mathcal{H} \to \mathcal{H}_m$ denote the orthogonal projection onto $\mathcal{H}_m$, and let $\mathcal{H}_m^\perp$ denote the orthogonal complement of $\mathcal{H}_m$ in $\mathcal{H}$. Let $m := \dim_{\mathbb{R}} \mathcal{H}_m$ be the real dimension of $\mathcal{H}_m$, and $k := \dim_{\mathbb{R}} \mathbb{K}$ be the real dimension of $\mathbb{K}$, that is, $k = 1$ for $\mathbb{K} = \mathbb{R}$ and $k = 2$ for $\mathbb{K} = \mathbb{C}$.

4.1 Integral Representation of (G, T)-Convolutional Neural Network

Definition 3. For any functions $\gamma : \mathcal{H}_m \times \mathbb{K} \to \mathbb{C}$ and $\sigma : \mathbb{K} \to \mathbb{C}$, we define the *integral representation of* (G, T)*-convolutional neural network* as a vector-valued function $\mathcal{H} \to \mathbb{C}^G$,

$$S[\gamma](x)(g) := \int_{\mathcal{H}_m \times \mathbb{K}} \gamma(a, b)\sigma((a *_T x)(g) - b)\mathrm{d}\lambda(a)\mathrm{d}b, \quad x \in \mathcal{H},\ g \in G. \tag{17}$$

Here, we call γ a coefficient function, and σ an activation function. If there is no risk of confusion, we abbreviate $\mathrm{d}\lambda(a)$ as $\mathrm{d}a$.

It is easy to see that a (G, T)-CNN is (G, T)-equivariant. In fact, for every $g, h \in G$,

$$S[\gamma](T_g[x])(h) = \int_{\mathcal{H}_m \times \mathbb{K}} \gamma(a, b)\sigma(\langle T_{(g^{-1}h)^{-1}}[x], a\rangle_{\mathcal{H}} - b)\mathrm{d}a\mathrm{d}b = S[\gamma](x)(g^{-1}h). \tag{18}$$

In addition, at the identity element $g = e$, it is reduced to a fully-connected network: $S[\gamma](x)(e) = \int_{\mathcal{H}_m \times \mathbb{K}} \gamma(a, b)\sigma(\langle x, a\rangle_{\mathcal{H}} - b)\mathrm{d}a\mathrm{d}b$ satisfying a *projection property*:

$$S[\gamma](P[x])(e) = S[\gamma](x)(e), \quad x \in \mathcal{H}. \tag{19}$$

In order for the (G, T)-convolution $a *_T x$ to be well-defined, $\mathcal{H}$ needs to be a G-invariant space: $T_G[\mathcal{H}] := \{T_g[x] \mid g \in G, x \in \mathcal{H}\} \subset \mathcal{H}$, which is automatically satisfied when T is a group representation, whereas $T_G[\mathcal{H}_m] \not\subset \mathcal{H}_m$ in general. On the other hand, in order for the Lebesgue measure λ to be well-defined, $\mathcal{H}_m$ needs to be a finite-dimensional space.

4.2 Ridgelet Transform and Scalar Product of Activation Function

Definition 4. For any functions $f : \mathcal{H}_m \to \mathbb{C}^G$ and $\rho : \mathbb{K} \to \mathbb{C}$, we define the *ridgelet transform* as

$$R[f;\rho](a,b) := \int_{\mathcal{H}_m} f(x)(e)\overline{\rho(\langle x, a\rangle_{\mathcal{H}} - b)}\mathrm{d}x, \quad (a,b) \in \mathcal{H}_m \times \mathbb{K}. \tag{20}$$

Here e denotes the identity element of G.

Definition 5. For any tempered distribution $\sigma \in \mathcal{S}'(\mathbb{K})$ and function $\rho \in \mathcal{S}(\mathbb{K})$, put a *scalar product* as

$$(\!(\sigma, \rho)\!) := (2\pi)^{m-k} \int_{\mathbb{K}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-m}\mathrm{d}\omega. \tag{21}$$

Here, $\cdot^\sharp$ denotes the Fourier transform on $\mathbb{K}$, which is identified with the Fourier transform on $\mathbb{R}^k$ with $k = \dim_{\mathbb{R}} \mathbb{K}$. We note that $\sigma^\sharp$ is defined in the sense of tempered distributions.

The derivations of the ridgelet transform and the scalar product are clarified in the proof of the reconstruction formula. Some readers who are familiar with the classical ridgelet transform may notice that the ridgelet transform for GCNN is formally the same as the one for fully-connected NNs. This may look asymmetric, since $\rho((a *_T x)(g) - b)$ looks intuitively more symmetric than $\rho(\langle x, a\rangle_{\mathcal{H}} - b)$ for the ridgelet transform for GCNNs. In fact, this symmetry is not broken, because at the identity element $g = e$, the convolution is reduced to the inner product: $\langle x, a\rangle_{\mathcal{H}} = (a *_T x)(e)$, thus we can understand the 'ridgelet transform with $\rho((a *_T x)(g) - b)$' is applied at the single point $g = e$, which works because (in the main theorem) the target function is assumed to be (G, T)-equivariant.

4.3 Reconstruction Formula

Here, we state the first half of our main results. For a function $f : \mathcal{H}_m \to \mathbb{C}^G$, put $f_e(x) := f(x)(e)$ for any $x \in \mathcal{H}_m$, and let $P_\sharp f : \mathcal{H} \to \mathbb{C}^G$ denote a pullback defined by $P_\sharp f(z)(g) := f(P[z])(g)$ for every $z \in \mathcal{H}$ and $g \in G$. We note that when $\mathcal{H}_m = \mathcal{H}$ and thus $P = \mathrm{Id}$, then $P_\sharp f = f$. Since $P_\sharp f(x \oplus y) = f(x \oplus 0)$ for any $x \oplus y \in \mathcal{H}_m \oplus \mathcal{H}_m^\perp$, $P_\sharp f$ is an extension of f to $\mathcal{H}$ as a constant function over $\mathcal{H}_m^\perp$.

Theorem 1 (Main Theorem 1/2)**.** *Given a function $f : \mathcal{H}_m \to \mathbb{C}^G$, assume (A1) that $P_\sharp f : \mathcal{H} \to \mathbb{C}^G$ is (G, T)-equivariant, i.e.,*

$$P_\sharp f(T_g[z])(h) = f(z)(g^{-1}h), \quad \textit{for every } z \in \mathcal{H} \textit{ and } g, h \in G; \tag{22}$$

and (A2) that f satisfies at least one of the following conditions: (A2a) both f_e and $\widehat{f}_e$ are absolute-integrable, i.e., $f_e, \widehat{f}_e \in L^1(\mathcal{H}_m)$, (A2b) f_e is square-integrable, i.e., $f_e \in L^2(\mathcal{H}_m)$, or (A2c) f_e is a tempered distribution, i.e., $f_e \in \mathcal{S}'(\mathcal{H}_m)$. Then, the following reconstruction formula holds:

$$S[R[f;\rho]](x)(g) = \int_{\mathcal{H}_m \times \mathbb{K}} R[f;\rho](a,b)\sigma((a *_T x)(g) - b)\mathrm{d}a\mathrm{d}b \overset{\star}{=} (\!(\sigma, \rho)\!) f(x)(g), \tag{23}$$

where the equality $\overset{\star}{=}$ holds at every continuous point x_c of f for (A2a), in L^2 for (A2b), and in $\mathcal{S}'$ for (A2c), respectively.

The proof is given in Appendix A.1.

4.4 cc-Universality

Finally, we state the second half of our main results. Since the reconstruction formula holds for an arbitrary function f, by discretizing the reconstruction formula $S[\gamma] = f$ with $\gamma = R[f;\rho]$ into a finite sum

$$f_n(x)(g) := \sum_{i=1}^{n} c_i \sigma((a_i *_T x)(g) - b_i), \tag{24}$$

in a 'nice' manner so that the discretized coefficient function $\gamma_n := \sum_{i=1}^n c_i \delta_{(a_i,b_i)}$ reproduces $\gamma = R[f;\rho]$ in a limit of $n \to \infty$, we can construct a sequence $\{f_n\}_{n\in\mathbb{N}}$ of finite (G,T)-CNNs that converges to an arbitrary target function f,

$$f_n \to f, \quad n \to \infty. \tag{25}$$

This is the primitive idea behind the constructive proof of the following *cc*-universality of finite (G,T)-CNNs based on ridgelet analysis.

To state an assumption on the activation function σ, we introduce the forward difference operator Δ_θ^n with difference $\theta > 0$, defined as

$$\Delta_\theta^1[\sigma](t) := \sigma(t+\theta) - \sigma(t), \quad \Delta_\theta^{n+1}[\sigma](t) := \Delta_\theta^1 \circ \Delta_\theta^n[\sigma](t). \tag{26}$$

Theorem 2 (Main Theorem 2/2)**.** *For an activation function $\sigma \in \mathcal{S}'(\mathbb{K})$, assume (A3) that there exist $n \in \mathbb{N}$ and $\theta > 0$ such that $\Delta_\theta^n[\sigma]$ is bounded and Lipschitz continuous. Let NN be the collection of finite (G,T)-CNNs, that is, $\mathrm{NN} := \bigcup_{n\in\mathbb{N}} \{f_n(x)(g) = \sum_{i=1}^n c_i \sigma((a_i *_T x)(g) - b_i) \mid (a_i, b_i, c_i) \in \mathcal{H}_m \times \mathbb{K} \times \mathbb{C}, i \in [n]\}$. Then, NN is cc-universal, that is, for any (G,T)-equivariant $C(G)$-valued continuous function $f \in C_{equi}(\mathcal{H}_m; C(G))$, and for any compact sets $K \subset \mathcal{H}_m$ and $L \subset G$, there exists a sequence $\{f_n\}_{n\in\mathbb{N}} \subset \mathrm{NN}$ of finite GCNNs such that*

$$\|f - f_n\|_{C(K;C(L))} = \sup_{x\in K} \sup_{g\in L} |f(x)(g) - f_n(x)(g)| \to 0, \quad n \to \infty. \tag{27}$$

The proof is given in Appendix A.2. Here, $f \in C_{equi}(\mathcal{H}_m; C(G))$ means that $P_\sharp f$ is (G,T)-equivariant.

4.5 Remarks

The topological group G covers a quite large class of groups. For example, it contains all the finite groups like $\mathbb{Z}_n$ and S_n, compact groups like $\mathrm{SO}(n)$ and $\mathrm{U}(n)$, and non-compact groups like $\mathbb{R}^n$ and $\mathrm{E}(n)$ as well. We assumed this simply for considering the class $C(G)$ of continuous functions. We note that when G is a finite group, then $C(G)$ is simply identified with a $|G|$-dimensional real/complex-vector space $\mathbb{K}^{|G|}$, because $C(G) := \{x : G \to \mathbb{K}, \text{ continuous}\} = \mathbb{K}^G$.

The G-module $\mathcal{H}$ is an abstract vector space and thus needs *not* be a space of functions on G such as $C(G)$ and $L^2(G/H)$, which is one of the major advantages; On the other hand, $\mathcal{H}_m$ needs to be finite-dimensional, since the proof is based on the Fourier inversion formula on the finite-dimensional Euclidean space $\mathbb{R}^m$. We conjecture that the extension to an infinite-dimensional setting would be a routine for some specialists in functional analysis. As we will see in the examples, however, ordinary input vectors are always essentially finite-dimensional due to the limitations in computational resources, and thus we consider finite-dimensional setting is reasonable.

The representation T needs *not* to be unitary, irreducible, nor square-integrable, since the proof is based only a few basic properties of the group homomorphism.

5 Examples

5.1 Finite Periodic Convolution Layer

Example 1 (For 1-dimensional periodic signals)**.** The periodic convolution corresponds to the case when $\mathbb{K} = \mathbb{R}, G = \mathbb{Z}_m \cong [m] = \{0, 1, \ldots, m-1\}$, $\mathcal{H} = L^2(G) \cong \mathbb{R}^m$ equipped with the inner product $\langle x, y\rangle := \frac{1}{m}\sum_{i\in[m]} x_i y_i$, and $T_i[x](j) := x_{j-i}$ thus $(a *_T x)(i) = \frac{1}{m}\sum_{j\in[m]} a_j x_{i+j}$. Therefore, the ridgelet transform and the reconstruction formula are given by

$$R[f;\rho](a,b) = \int_{\mathbb{R}^m} f(\boldsymbol{x})(0)\overline{\rho\left(\frac{1}{m}\sum_{i\in[m]} a_i x_i - b\right)}\mathrm{d}\boldsymbol{x},$$

$$S[R[f;\rho]](x)(i) = \int_{\mathbb{R}^m\times\mathbb{R}} R[f;\rho](\boldsymbol{a},b)\sigma\left(\frac{1}{m}\sum_{j\in[m]} a_j x_{i+j} - b\right)\mathrm{d}\boldsymbol{a}\mathrm{d}b = (\!(\sigma,\rho)\!) f(x)(i).$$

Example 2 (For 2-dimensional multi-channel periodic images)**.** A 2-dimensional c-channel image is identified with a vector-valued function $x : \mathbb{Z}_m \times \mathbb{Z}_n \to \mathbb{R}^c$, thus $\mathcal{H} \cong \mathbb{R}^{m\times n\times c}$. Let x_{ij}^k denote the (i,j)-th component in the k-th channel of $x \in \mathcal{H}$. Let $G = \mathbb{Z}_m \times \mathbb{Z}_n$, and put $T_{(p,q)}[x]_{ij}^k := x_{i-p,j-q}^k$. Therefore, the ridgelet transform and the reconstruction formula are given by

$$R[f;\rho](a,b) = \int_{\mathbb{R}^{mnc}} f(x)_{(0,0)} \overline{\rho\left(\frac{1}{mnc}\sum_{k\in[c]}\sum_{i,j\in[m]} a_{ij}^k x_{ij}^k - b\right)} \mathrm{d}\boldsymbol{x}.$$

$$S[R[f;\rho]](x)_{ij} = \int_{\mathbb{R}^{mnc}\times\mathbb{R}} R[f;\rho](a,b)\sigma\left(\frac{1}{mnc}\sum_{k\in[c]}\sum_{p,q\in[m]} a_{pq}^k x_{p+i,q+j}^k - b\right)\mathrm{d}\boldsymbol{a}\mathrm{d}b = (\!(\sigma,\rho)\!) f(x)_{ij}.$$

Example 3 (Difference operator (with cutoff function))**.** A difference operator on $x : [m] \to \mathbb{R}$ is given by $x = \sum_{i\in[m]} x_i\delta_i \mapsto f(x) = \sum_{i\in[m]}(x_{i+1} - x_i)\delta_i$, which is (G,T)-equivariant: $f(T_k^*[x])(i) = T_k^*[x]_{i+1} - T_k^*[x]_i = x_{i+1-k} - x_{i-k} = f(x)(i-k)$. Since $f(x)(0) = x_1 - x_0$,

$$R\left[f|_K;\rho\right](a,b) = \int_{\mathbb{R}^m} (x_1 - x_0) 1_K(\boldsymbol{x}) \overline{\rho\left(\frac{1}{m}\sum_{i\in[m]} a_i x_i - b\right)} \mathrm{d}\boldsymbol{x}.$$

We note since $\boldsymbol{x} \mapsto x_1 - x_0$ is not integrable in $\mathbb{R}^m$, we restrict f to a compact set $K \subset \mathbb{R}^m$, and impose the indicator function 1_K as an auxiliary cutoff function.

5.2 (Deep Sets) Permutation Equivariant Maps on A Finite Set

Example 4. Let $\mathcal{H} = \mathbb{R}^m, G \leqslant S_m$ and $T_g[x] = (x_{g^{-1}(1)}, x_{g^{-1}(2)}, \ldots, x_{g^{-1}(m)})$. Thus $\langle a, x\rangle_{\mathcal{H}} = \frac{1}{m}\sum_{i\in[m]} a_i x_i$, and $(a *_T x)(g) = \frac{1}{m}\sum_{p\in[m]} a_p x_{g(p)}$. So,

$$R[f;\rho](a,b) = \int_{\mathbb{R}^m} f(\boldsymbol{x})(e) \overline{\rho\left(\frac{1}{m}\sum_{i\in[m]} a_i x_i - b\right)} \mathrm{d}\boldsymbol{x},$$

$$S[R[f;\rho]](x)(g) = \int_{\mathbb{R}^m\times\mathbb{R}} R[f;\rho](\boldsymbol{a},b)\sigma\left(\frac{1}{m}\sum_{p\in[m]} a_p x_{g(p)} - b\right)\mathrm{d}\boldsymbol{a}\mathrm{d}b = (\!(\sigma,\rho)\!) f(x)(g).$$

5.3 Continuous Periodic Convolution Layer

Example 5. Let $G = \mathbb{T} := \mathbb{R}/2\pi\mathbb{Z} \cong \{e^{i\theta} \mid \theta \in [-\pi,\pi]\}$ be the 1-dimensional torus group, which is one of the most basic continuous group. As a consequence of the Fourier series expansion, $L^2(\mathbb{T})$ is spanned by $\{e^{in\theta} \mid n \in \mathbb{N}\}$. Hence, we can take $\mathcal{H}$ to be an m-dimensional subspace $\mathcal{H} := \{\sum_{|n|<m} x_n e^{in\theta} \mid x_{-n} = x_n \in \mathbb{R}\}$ equipped with an inner product $\langle x, y\rangle_{\mathcal{H}} := \int_{\mathbb{T}} x(\theta)\overline{y(\theta)}\mathrm{d}\theta$. We note that the constraint $x_n = x_{-n}$ implies $\sum_{|n|<m} x_n e^{in\theta} = \sum_{|n|<m} x_n \cos(n\theta)$ and thus any signal $x \in \mathcal{H}$ is a bandlimited real-valued continuous signal with each coefficient x_n being the n-th frequency spectrum. Put $T_\alpha[x](\theta) := x(\theta - \alpha)$, then $(a * x)(\alpha) = \int_{\mathbb{T}} a(\theta)x(\alpha-\theta)\mathrm{d}\theta = \sum_{|n|<m} a_n x_n e^{in\alpha} = \sum_{|n|<m} a_n x_n \cos(n\alpha)$ (by the convolution theorem and the constraint). Therefore,

$$R[f;\rho](a,b) := \int_{\mathbb{R}^m} f(\boldsymbol{x})(0) \overline{\rho\left(\sum_{|n|<m} a_n x_n - b\right)} \mathrm{d}\boldsymbol{x},$$

$$S[R[f;\rho]](x)(\theta) = \int_{\mathbb{R}^m\times\mathbb{R}} R[f;\rho](\boldsymbol{a},b)\sigma\left(\sum_{|n|<m} a_n x_n \cos(n\theta) - b\right)\mathrm{d}\boldsymbol{a}\mathrm{d}b = (\!(\sigma,\rho)\!) f(x)(\theta).$$

Example 6 (Differential operator (with convergence factor))**.** A differential operator $\frac{\mathrm{d}}{\mathrm{d}\theta}$ is calculated as $x = \sum_{|n|<m} x_n e^{in\theta} \mapsto f(x) = \sum_{|n|<m} n x_n e^{in\theta}$. Since $\boldsymbol{x} \mapsto f(\boldsymbol{x})(0) = \sum_{|n|<m} n x_n$ is not integrable on $\mathbb{R}^m$, we impose a convergence factor ϕ_t as follows. $f_t(x)(\theta) := f(x)\phi_t(x) = \frac{\mathrm{d}}{\mathrm{d}\theta}x(\theta)\phi_t(x) = \sum_{|n|<m} n x_n \phi_t(x) e^{in\theta}$. Here, $(\phi_t)_{t>0} \subset \mathcal{S}(\mathcal{H})$ is a family of convergence factors that satisfies (1) the first moment $\int_{\mathcal{H}} |x|_{\mathcal{H}}|\phi_t(x)|\mathrm{d}x$ exists at every t, (2) $\phi_t \to 1$ in the weak sense as $t \to \infty$, and (3) (continuous and) (G,T)-equivariant. For example, we can take a Gaussian $\phi_t(x) = \exp(-|x|_{\mathcal{H}}^2/4t)$. Hence,

$$R[f_t;\rho](a,b) = \int_{\mathbb{R}^m} \sum_{|n|<m} n x_n \phi_t(\boldsymbol{x}) \overline{\rho\left(\sum_{|n|<m} a_n x_n - b\right)} \mathrm{d}\boldsymbol{x}.$$

5.4 Euclidean group $\mathrm{E}(n)$ equivariant map

Example 7. The Euclidean group $\mathrm{E}(n)$ is a semidirect product $\mathbb{R}^n \rtimes \mathrm{O}(n)$ of the translational group $\mathbb{R}^n$ and the orthogonal group $\mathrm{O}(n)$, which acts on $\mathbb{R}^n$ as $(U, s) \cdot t := Ut + s$ for any $t \in \mathbb{R}^n$ and $(U, s) \in O(n) \times \mathbb{R}^n$. So, put $\mathcal{H} \subset L^2(\mathbb{R}^n)$ and $T_{(U,s)}[x](t) := x(U^{-1}(t - s))$. Then,

$$R[f;\rho](a,b) := \int_{\mathcal{H}} f(x)(I,0)\overline{\rho\left(\int_{\mathbb{R}^n} a(t)\overline{x(t)}\mathrm{d}t - b\right)}\mathrm{d}x$$

$$S[R[f;\rho]](x)(U,t) = \int_{\mathcal{H}\times\mathbb{R}} R[f;\rho](a,b)\sigma\left(\int_{\mathbb{R}^n} a(U^{-1}(t-s))\overline{x(s)}\mathrm{d}s - b\right)\mathrm{d}a\mathrm{d}b = f(x)(U,t)$$

We note that a more efficient representation for $L^2(\mathbb{R}^n)$ and/or a more general representation such as $L^2(\mathrm{SO}(2))$ and $L^2(\mathrm{E}(3))$, have been developed in the context of *steerable* CNNs [32, 61].

6 Related Work (on CNN Universality)

Non-Group CNN. Zhou [9, 10] have shown the cc-universality of specific deep ReLU (non-group) CNNs. This seems to be the earliest result. In [10], he presented (Theorem 1) the cc-universality in $C(\mathbb{R}^d;\mathbb{R})$ in the limit of depth $J \to \infty$, and (Theorem 2) an approximation error rate with respect to J. The CNN is carefully designed so that increasing the depth results in increasing the width. The activation function is limited to ReLU, which is essential for the proof. Since the proposed network is different from standard CNNs (as well as group CNNs), the "universality of CNN" shown in this paper is at a concept level. Specifically, our result does not directly cover his results.

Finite Group CNN. Maron et al. [12], Sannai et al. [62], Keriven and Peyré [63], Ravanbakhsh [64] and Petersen and Voigtlaender [13] presented the cc-(or L^p-)universality results of *finite*-group CNNs. Maron et al. [12] is often cited as one of the earliest publications, where the input space is $\mathcal{H} = \mathbb{R}^{n^k \times a}$ (a-channel k-th order n-dimensional tensors), the output space is $\mathcal{H}' = \mathbb{R}^{n^l \times b}$ (b-channel l-th order n-dimensional tensors), the group G is a subgroup of a symmetric group S_n, and the group action is the left-translation (i.e. the group representation T is the left-regular representation). In this setup, they presented the cc-universality of deep-ReLU-GCNNs in the space of continuous G-equivariant functions $C_{equi}(\mathcal{H};\mathcal{H}')$. The proofs are constructive but still existential because they are based on existential theorems of either *invariant polynomials* (Stone-Weierstrass) or *MLPs* (Hahn-Banach). We note that a Stone-Weierstrass theorem states the existence of *polynomials* without construction, and Taylor/Fourier series expansion is not the one. These finite-group cases are essentially covered as Example 4 (Deep Sets). Again, the proof based on the ridgelet transform can further provide the construction of parameters, and thus be more constructive.

Infinite Group CNN. Yarotsky [11] deals with *infinite* groups G and *infinite-dimensional* input/output spaces, and show a version of universalities of carefully designed deep GCNNs in the space of G-equivariant continuous functions $C_{equi}(L^2(G;\mathbb{R}^d); L^2(G;\mathbb{R}^{d'}))$ where G is either translation group $\mathbb{R}^d$ and 2-dimensional roto-translation group $\mathrm{SE}(2)$. This study aims to investigate a new convolution-like structure that explicitly encodes the group-equivariance structure as explicit inductive bias, particularly when the input/output feature spaces are infinite-dimensional spaces (i.e. $L^2(G;\mathbb{R}^d)$) and the group G is an infinite group. (We note that eventually the input/output spaces are finite-dimensional since the spaces are discretized as finite-dimensional Euclidean spaces). Thus, the proposed networks are not directly covered in our study. However, the cases of infinite-dimensional groups acting on a function space are covered in Examples 5 and 7.

References

[1] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Veličković. Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges. *arXiv preprint: 2104.13478*, 2021.

[2] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[3] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep Sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[4] Risi Kondor and Shubhendu Trivedi. 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*, volume 80 of *Proceedings of Machine Learning Research*, pages 2747–2755, 2018.

[5] Haggai Maron, Heli Ben-Hamu, Nadav Shamir, and Yaron Lipman. Invariant and Equivariant Graph Networks. In *International Conference on Learning Representations*, 2019.

[6] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations*, 2018.

[7] Taco S Cohen, Mario Geiger, and Maurice Weiler. A General Theory of Equivariant CNNs on Homogeneous Spaces. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[8] Risi Kondor, Zhen Lin, and Shubhendu Trivedi. Clebsch-Gordan Nets: a Fully Fourier Space Spherical Convolutional Neural Network. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[9] Ding-Xuan Zhou. Deep distributed convolutional neural networks: Universality. *Analysis and Applications*, 16(06):895–919, 2018.

[10] Ding-Xuan Zhou. Universality of deep convolutional neural networks. *Applied and Computational Harmonic Analysis*, 48(2):787–794, 2020.

[11] Dmitry Yarotsky. Universal Approximations of Invariant Maps by Neural Networks. *Constructive Approximation*, 2021.

[12] Haggai Maron, Ethan Fetaya, Nimrod Segol, and Yaron Lipman. On the Universality of Invariant Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4363–4371, 2019.

[13] Philipp Petersen and Felix Voigtlaender. Equivalence of approximation by convolutional neural networks and fully-connected networks. *Proceedings of the American Mathematical Society*, 148(4):1567–1581, 2020.

[14] Wataru Kumagai and Akiyoshi Sannai. Universal Approximation Theorem for Equivariant Maps by Group CNNs. *arXiv preprint: 2012.13882*, 2020.

[15] Sho Okumoto and Taiji Suzuki. Learnability of convolutional neural networks for infinite dimensional input via mixed and anisotropic smoothness. In *International Conference on Learning Representations*, 2022.

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

[17] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[18] Bunpei Irie and Sei Miyake. Capabilities of three-layered perceptrons. In *IEEE International Conference on Neural Networks*, pages 641–648. IEEE, 1988.

[19] Ken-Ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural Networks*, 2(3):183–192, 1989.

[20] 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.

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

[22] H.N Mhaskar and Charles A Micchelli. Approximation by superposition of sigmoidal and radial basis functions. *Advances in Applied Mathematics*, 13(3):350–373, 1992.

[23] H. N. Mhaskar. Neural Networks for Optimal Approximation of Smooth and Analytic Functions. *Neural Computation*, 8:164–177, 1996.

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

[25] Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

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

[27] Emmanuel Jean Candès. *Ridgelets: theory and applications*. PhD thesis, Standford University, 1998.

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

[29] Matus Telgarsky. Benefits of depth in neural networks. In *29th Annual Conference on Learning Theory*, pages 1–23, 2016.

[30] Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94:103–114, 2017.

[31] Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018.

[32] Taco S Cohen and Max Welling. Steerable CNNs. In *International Conference on Learning Representations 2017*, pages 1–14, 2017.

[33] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 8571–8580, Montreal, BC, 2018. Curran Associates, Inc.

[34] Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. In *Advances in Neural Information Processing Systems 32*, pages 8572–8583, Vancouver, BC, 2019. Curran Associates, Inc.

[35] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 322–332, Long Beach, 2019. PMLR.

[36] Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On Lazy Training in Differentiable Programming. In *Advances in Neural Information Processing Systems 32*, pages 2937–2947, 2019.

[37] Jonathan Frankle and Michael Carbin. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *International Conference on Learning Representations 2019*, pages 1–42, 2019.

[38] Atsushi Nitanda and Taiji Suzuki. Stochastic Particle Gradient Descent for Infinite Ensembles. *arXiv preprint: 1712.05438*, 2017.

[39] Song Mei, Andrea Montanari, and Phan-Minh 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.

[40] Grant Rotskoff and Eric Vanden-Eijnden. Parameters as interacting particles: long time convergence and asymptotic error scaling of neural networks. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 7146–7155, Montreal, BC, 2018. Curran Associates, Inc.

[41] Lenaic Chizat and Francis 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. Curran Associates, Inc.

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

[43] Taiji 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.

[44] Sho Sonoda, Isao Ishikawa, and Masahiro Ikeda. Ghosts in Neural Networks: Existence, Structure and Role of Infinite-Dimensional Null Space. *arXiv preprint: 2106.04770*, 2021.

[45] Sho Sonoda, Isao Ishikawa, and Masahiro Ikeda. Ridge Regression with Over-Parametrized Two-Layer Networks Converge to Ridgelet Spectrum. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021*, volume 130, pages 2674–2682. PMLR, 2021.

[46] Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Proceedings of the 32nd Conference on Learning Theory*, volume 99, pages 2667–2690, 2019.

[47] Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A Function Space View of Bounded Norm Infinite Width ReLU Nets: The Multivariate Case. In *International Conference on Learning Representations*, 2020.

[48] Rahul Parhi and Robert D. Nowak. Neural Networks, Ridge Splines, and TV Regularization in the Radon Domain. *arXiv preprint: 2006.05626*, 2020.

[49] Rahul Parhi and Robert D Nowak. Banach Space Representer Theorems for Neural Networks and Ridge Splines. *Journal of Machine Learning Research*, 22(43):1–40, 2021.

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

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

[52] Jean-Luc Starck, Fionn Murtagh, and Jalal M. Fadili. The ridgelet and curvelet transforms. In *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity*, pages 89–118. Cambridge University Press, 2010.

[53] Gitta Kutyniok and Demetrio Labate. *Shearlets: Multiscale Analysis for Multivariate Data*. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, 1 edition, 2012.

[54] Sho Sonoda and Noboru Murata. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 43(2):233–268, 2017.

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

[56] Sho Sonoda, Isao Ishikawa, and Masahiro Ikeda. Fully-Connected Network on Noncompact Symmetric Space and Ridgelet Transform based on Helgason-Fourier Analysis. *arXiv preprint: 2203.01631*, 2022.

[57] Loukas Grafakos. *Classical Fourier Analysis*. Graduate Texts in Mathematics. Springer New York, second edition, 2008.

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

[59] Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, Characteristic Kernels and RKHS Embedding of Measures. *Journal of Machine Learning Research*, 12 (Jul):2389–2410, 2010.

[60] Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8: 143–195, 1999.

[61] Maurice Weiler and Gabriele Cesa. General E(2)-Equivariant Steerable CNNs. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[62] Akiyoshi Sannai, Yuuki Takai, and Matthieu Cordonnier. Universal approximations of permutation invariant/equivariant functions by deep neural networks. *arXiv preprint: 1903.01939*, mar 2019.

[63] Nicolas Keriven and Gabriel Peyré. Universal Invariant and Equivariant Graph Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[64] Siamak Ravanbakhsh. Universal Equivariant Multilayer Perceptrons. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 7996–8006. PMLR, 2020.

[65] P C Kainen, Věra Kůrková, and Marcello Sanguineti. Approximating multivariable functions by feedforward neural nets. In *Handbook on Neural Information Processing*, volume 49 of *Intelligent Systems Reference Library*, pages 143–181. Springer Berlin Heidelberg, 2013.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A]

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs

Additional Notation In the proofs, we use two symbols $\widehat{\cdot}$ and $\cdot^\sharp$ for the Fourier transforms in $x \in \mathcal{H}$ and $b \in \mathbb{K}$, respectively. For example,

$$\widehat{f}(\xi) := \int_{\mathcal{H}} f(x) e^{-i\langle \xi, x\rangle_{\mathcal{H}}} \mathrm{d}x, \quad \xi \in \mathcal{H}$$
$$\rho^\sharp(\omega) := \int_{\mathbb{K}} \rho(b) e^{-i\omega b} \mathrm{d}b, \quad \omega \in \mathbb{K}$$
$$\gamma^\sharp(a, \omega) = \int_{\mathbb{K}} \gamma(a, b) e^{-i\omega b} \mathrm{d}b, \quad (a, \omega) \in \mathcal{H} \times \mathbb{K}.$$

With a slight abuse of notation, when σ is a tempered distribution (i.e., $\sigma \in \mathcal{S}'(\mathbb{K})$), then $\sigma^\sharp$ is understood as the Fourier transform of distributions. Namely, $\sigma^\sharp$ is another tempered distribution satisfying $\int_{\mathbb{K}} \sigma^\sharp(\omega)\phi(\omega)\mathrm{d}\omega = \int_{\mathbb{K}} \sigma(\omega)\phi^\sharp(\omega)\mathrm{d}\omega$ for any test function $\phi \in \mathcal{S}(\mathbb{K})$.

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

$$\triangle^{t/2}[\phi](\boldsymbol{v}) := \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\boldsymbol{u}|^t \widehat{\phi}(\boldsymbol{u}) e^{i\boldsymbol{u}\cdot\boldsymbol{v}} \mathrm{d}\boldsymbol{u}, \tag{28}$$

$$\langle \triangle \rangle^{t/2}[\phi](\boldsymbol{v}) := \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} (1 + |\boldsymbol{u}|^2)^{t/2} \widehat{\phi}(\boldsymbol{u}) e^{i\boldsymbol{u}\cdot\boldsymbol{v}} \mathrm{d}\boldsymbol{u}. \tag{29}$$

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

A.1 Theorem 1

Proof. In the following, we fix a representation $T : G \to GL(\mathcal{H})$ of a group G on a (potentially infinite-dimensional) Hilbert space $\mathcal{H}$ over a field $\mathbb{K}$ equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, which is a G-invariant vector space: $T_G[\mathcal{H}] = \mathcal{H}$, and a finite-dimensional closed subspace $\mathcal{H}_m \subset \mathcal{H}$

equipped with the Lebesgue measure λ. Let $\mathcal{H}_m^\perp$ be the orthogonal complement of $\mathcal{H}_m$ in $\mathcal{H}$, i.e., $x \oplus y \in \mathcal{H}_m \oplus \mathcal{H}_m^\perp = \mathcal{H}$, and let $P : \mathcal{H} \to \mathcal{H}_m$ denote the orthogonal projection onto $\mathcal{H}_m$. Let $m := \dim_{\mathbb{R}} \mathcal{H}_m$ denotes the real dimension of $\mathcal{H}_m$, and let $k := \dim_{\mathbb{R}} \mathbb{K}$ denotes the real dimension of $\mathbb{K}$, which is either 1 or 2.

Without loss of generality, we can assume (A') that $\gamma(a, \bullet) * \sigma \in \mathcal{S}(\mathbb{K})$ for a.e. $a \in \mathcal{H}_m$, and (A") that $\gamma^\sharp \sigma^\sharp \in L^1(\mathcal{H}_m \times \mathbb{K})$, which will be eventually justified because later in (34), we set $\gamma^\sharp(a, \omega) = \widehat{f}(\omega a)\overline{\rho^\sharp(\omega)}$.

Step 1 (Fourier expression). Using an identity: For any function $\phi \in \mathcal{S}(\mathbb{K})$ and $b \in \mathbb{K}$, $\phi(b) = \frac{1}{(2\pi)^k} \int_{\mathbb{K}} \phi^\sharp(\omega) e^{ib\omega} \mathrm{d}\omega$, namely the Fourier inversion formula, we can turn S into a *Fourier expression*:

$$S[\gamma](x)(g) = \frac{1}{(2\pi)^k} \int_{\mathcal{H}_m} \int_{\mathbb{K}} \gamma(a, b) \sigma(\langle T_{g^{-1}}[x], a \rangle_{\mathcal{H}} - b) \mathrm{d}b \mathrm{d}a \tag{30}$$

$$= \frac{1}{(2\pi)^k} \int_{\mathcal{H}_m} \int_{\mathbb{K}} \gamma^\sharp(a, \omega) \sigma^\sharp(\omega) \exp\left(i\omega \langle T_{g^{-1}}[x], a \rangle_{\mathcal{H}}\right) \mathrm{d}\omega \mathrm{d}a. \tag{31}$$

By the assumption (A'), the first equation holds at every point $b = (a *_T x)(g)$, and by the assumption (A"), the Fourier expression is *uniformly* absolutely convergent:

$$\int_{\mathcal{H}_m \times \mathbb{K}} |\gamma^\sharp(a, \omega) \sigma^\sharp(\omega) \exp(i\omega (a * x)(g))| \mathrm{d}a \mathrm{d}\omega = \|\gamma^\sharp \sigma^\sharp\|_{L^1(\mathcal{H}_m \times \mathbb{K})} < \infty, \tag{32}$$

for *all* $(x, g) \in \mathcal{H} \times G$. Hence, we can change the order of integration freely.

Step 2 (Reconstruction). By changing the variables as $(a, \omega) = (\xi/\omega, \omega)$ with $\mathrm{d}a\mathrm{d}\omega = |\omega|^{-m} \mathrm{d}\xi \mathrm{d}\omega$, we have

$$S[\gamma](x)(g) = \frac{1}{(2\pi)^k} \int_{\mathcal{H}_m \times \mathbb{K}} \gamma^\sharp(\xi/\omega, \omega) \sigma^\sharp(\omega) \exp\left(i\langle T_{g^{-1}}[x], \xi \rangle_{\mathcal{H}}\right) |\omega|^{-m} \mathrm{d}\omega \mathrm{d}\xi. \tag{33}$$

Hence, using a given f satisfying the assumptions (A1) and (A2), and some function $\rho \in \mathcal{S}(\mathbb{K})$, suppose that $\gamma_{f,\rho}$ satisfies the following *separation-of-variables* form:

$$\gamma_{f,\rho}^\sharp(\xi/\omega, \omega) = \widehat{f}(\xi)(e)\overline{\rho^\sharp(\omega)}. \tag{34}$$

Then,

$$S[\gamma_{f,\rho}](x)(g) = \frac{1}{(2\pi)^k} \int_{\mathcal{H}_m \times \mathbb{K}} \widehat{f}(\xi)(e)\overline{\rho^\sharp(\omega)} \sigma^\sharp(\omega) \exp\left(i\langle T_{g^{-1}}[x], \xi \rangle_{\mathcal{H}}\right) |\omega|^{-m} \mathrm{d}\omega \mathrm{d}\xi \tag{35}$$

$$= \left((2\pi)^{m-k} \int_{\mathbb{K}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)} |\omega|^{-m} \mathrm{d}\omega \right)$$

$$\times \left(\frac{1}{(2\pi)^m} \int_{\mathcal{H}_m} \widehat{f}(\xi)(e) \exp\left(i\langle T_{g^{-1}}[x], \xi \rangle_{\mathcal{H}}\right) \mathrm{d}\xi \right) \tag{36}$$

$$\overset{\star}{=} (\!(\sigma, \rho)\!) f(PT_{g^{-1}}[x])(e) \tag{37}$$

$$= (\!(\sigma, \rho)\!) f(x)(g). \tag{38}$$

where we put

$$(\!(\sigma, \rho)\!) := (2\pi)^{m-k} \int_{\mathbb{K}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)} |\omega|^{-m} \mathrm{d}\omega. \tag{39}$$

Here, the equality $\overset{\star}{=}$ holds at every continuous point x_c of f for (A2a), in L^2 for (A2b), and in $\mathcal{S}'$ for (A2c), respectively.

Step 3 (Ridgelet transform). Since we put

$$\gamma_{f,\rho}^\sharp(a, \omega) = \widehat{f}(\omega a)(e)\overline{\rho^\sharp(\omega)}, \tag{40}$$

it is calculated as

$$\gamma_{f,\rho}(a,b) = \frac{1}{(2\pi)^k}\int_{\mathbb{K}} \widehat{f}(\omega a)(e)\overline{\rho^\sharp(\omega)}e^{i\omega b}\mathrm{d}\omega \tag{41}$$

$$= \frac{1}{(2\pi)^k}\int_{\mathbb{K}\times\mathcal{H}_m} f(x)(e)\overline{\rho^\sharp(\omega)}e^{i\omega(b-\langle a,x\rangle_{\mathcal{H}})}\mathrm{d}\omega\mathrm{d}x \tag{42}$$

$$= \int_{\mathcal{H}_m} f(x)(e)\overline{\rho(\langle a,x\rangle_{\mathcal{H}}-b)}\mathrm{d}x, \tag{43}$$

which is the definition of the ridgelet transform for GCNN. □

A.2 Theorem 2

Proof. Fix arbitrary compact sets $K \subset \mathcal{H}_m$ and $L \subset G$, positive number $\varepsilon > 0$, and function $f \in C_{equi}(K; C(G))$. An n-term finite (G,T)-CNN is given by

$$f_n(x)(g) := \sum_{i=1}^n c_i\sigma\left((a_i *_T x)(g) - b_i\right), \quad x \in \mathcal{H}_m,\ g \in G \tag{44}$$

with parameters $(a_i, b_i, c_i) \in \mathcal{H}_m \times \mathbb{K} \times \mathbb{C}$. Observe that any finite (G,T)-CNN is (G,T)-equivariant, that is,

$$f_n(T_g[x])(h) = \sum_{i=1}^n c_i\sigma\left(\langle T_{(g^{-1}h)^{-1}}x, a_i\rangle_{\mathcal{H}} - b_i\right) = f_n(x)(g^{-1}h). \tag{45}$$

Put $\overline{K} := \{T_{g^{-1}}[x] \mid x \in K, g \in L\}$, which is compact because T is continuous, and put $f_e(x) := f(x)(e)$, which is compactly supported, i.e., $f_e \in C(K) \subset C(\overline{K})$. By Theorem 3, there exist a finite number $N \in \mathbb{N}$ and an N-term $\mathbb{C}$-valued fully-connected network $F_N(x) = \sum_{i=1}^N c_i\sigma(\langle a_i, x\rangle_{\mathcal{H}} - b_i)$ satisfying $\|F_N - f_e\|_{C(\overline{K})} < \varepsilon$. Put $f_N(x)(g) := F_N(T_{g^{-1}}[x])$. Then, it is a (G,T)-CNN because

$$f_N(x)(g) = \sum_{i=1}^N c_i\sigma\left(\langle T_{g^{-1}}, a_i\rangle_{\mathcal{H}} - b_i\right) = \sum_{i=1}^N c_i\sigma\left((a_i *_T x)(g) - b_i\right), \tag{46}$$

and it is an ε-neighbour of f because

$$\|f_N - f\|_{C(K;C(L))} = \sup_{x\in K}\sup_{g\in L}|f_N(x)(g) - f(x)(g)| \tag{47}$$

$$= \sup_{x\in K}\sup_{g\in L}|F_N(T_{g^{-1}}[x]) - f_e(T_{g^{-1}}[x])| \tag{48}$$

$$= \sup_{g\in L}\sup_{x'\in\overline{K}}|F_N(x') - f_e(x')|, \quad x' = T_{g^{-1}}[x] \tag{49}$$

$$< \varepsilon, \tag{50}$$

which concludes the assertion. □

Theorem 3 (cc-universality of scalar-valued finite fully-connected NNs on $\mathbb{R}^m$)**.** *Suppose that*

1. $\mathcal{H} = \mathcal{H}_m = \mathbb{R}^m$,
2. $f \in C(\mathcal{H}; \mathbb{C})$ *(not vector-valued* $C(\mathcal{H}; \mathbb{C}^G)$ *but scalar-valued), and*
3. *there exists* $k \geqslant 0$ *and* $\theta > 0$ *such that* $\Delta_\theta^k[\sigma] \in L^\infty(\mathbb{R})$ *and Lipschitz continuous.*

Then, the finite neural networks of the form $f_n(\boldsymbol{x}) = \sum_{i=1}^n c_i\sigma(\boldsymbol{a}_i \cdot \boldsymbol{x} - b_i)$ *are cc-universal, that is, for any compact set* $K \subset \mathbb{R}^m$*, positive number* $\varepsilon > 0$*, and continuous function* $f \in C(K)$*, there exists a finite network* f_n *such that* $\|f - f_n\|_{C(K)} < \varepsilon$*.*

Proof. Since $\sum_{i=1}^n c_i\Delta_\theta^k[\sigma](\boldsymbol{a}_i \cdot \boldsymbol{x} - b_i)$ is rewritten as another finite model $\sum_{i=1}^{n'} c_i'\sigma(\boldsymbol{a}_i' \cdot \boldsymbol{x} - b_i')$, it suffice to consider the case $k = 0$. In the following, we assume that $\sigma(= \Delta_\theta^0[\sigma])$ is bounded and Lipschitz continuous.

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

$$S[R[f_c;\rho]](\boldsymbol{x}) = f_c(\boldsymbol{x}) \text{ at every point } \boldsymbol{x} \in \mathbb{R}^m. \tag{51}$$

For example, take a compactly-supported smooth function $\rho_0 \in C_c^\infty(\mathbb{K})$, write $k = \dim_{\mathbb{R}} \mathbb{K}(= 1 \text{ or } 2)$, and put $\rho(b) := \triangle_b^{m/2}[\rho_0](b) = (2\pi)^{-k}\int_{\mathbb{K}} |\omega|^m \rho_0^\sharp(\omega) e^{ib\cdot\omega}\mathrm{d}\omega$. Then, $(\!(\sigma,\rho)\!) = (2\pi)^{m-k}\int_{\mathbb{K}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-m}\mathrm{d}\omega = (2\pi)^{m-k}\int_{\mathbb{K}} \sigma^\sharp(\omega)\overline{\rho_0^\sharp(\omega)}\mathrm{d}\omega = (2\pi)^m\int_{\mathbb{K}} \sigma(b)\overline{\rho_0(b)}\mathrm{d}b = \langle \sigma, \rho_0\rangle_{L^2(\mathbb{K})}$, which is an ordinary functional inner product, and it is easy to find a ρ_0 satisfying $\langle \sigma, \rho_0\rangle_{L^2(\mathbb{K})} \neq 0$. By normalizing $\rho' := \rho/(\!(\sigma,\rho)\!)$, we can find the ρ'. We refer to Sonoda and Murata [54] and Sonoda et al. [44] for more details on the scalar product $(\!(\sigma,\rho)\!)$.

Step 2 ($R[f_c;\rho]$). To show a discretization f_n of the reconstruction formula converges to f_c in $C(K)$, it is convenient to regard the integration $\int_{\mathbb{R}^m\times\mathbb{R}}[\cdots]\mathrm{d}\boldsymbol{a}\mathrm{d}b$ in S as the Bochner integral, and the integrand $\gamma(\boldsymbol{a},b)\sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b)$ as a vector-valued function from $\mathbb{R}^m \times \mathbb{R}$ to $C(K)$.

Since f_c is C^∞-smooth, $R[f_c;\rho](\boldsymbol{a},b)$ is bounded and decays rapidly in $\boldsymbol{a}$, and thus $R[f_c;\rho]\sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b)$ is Bochner integrable, that is,

$$\int_{\mathbb{R}^m\times\mathbb{R}} \sup_{\boldsymbol{x}\in K} \left|R[f_c;\rho](\boldsymbol{a},b)\sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b)\right|\mathrm{d}\boldsymbol{a}\mathrm{d}b < \infty. \tag{52}$$

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

$$\begin{aligned}
|R[f_c;\rho](\boldsymbol{a},b)| &= \frac{1}{2\pi}\left|\int_{\mathbb{R}} \widehat{f}_c(\omega\boldsymbol{a})\overline{\rho^\sharp(\omega)}e^{i\omega b}\mathrm{d}\omega\right| \\
&= \frac{1}{2\pi}\left|\int_{\mathbb{R}} \langle\omega\boldsymbol{a}\rangle^s\langle\omega\boldsymbol{a}\rangle^{-s}\langle b\rangle^t\langle b\rangle^{-t}\widehat{f}_c(\omega\boldsymbol{a})\overline{\rho^\sharp(\omega)}e^{i\omega b}\mathrm{d}\omega\right| \\
&\leqslant \frac{1}{2\pi}\left|\int_{\mathbb{R}} \langle\omega\boldsymbol{a}\rangle^s\widehat{f}_c(\omega\boldsymbol{a})\langle\omega\rangle^{-s}\overline{\rho^\sharp(\omega)}\langle\triangle_\omega\rangle^t e^{i\omega b}\mathrm{d}\omega\right|\langle\boldsymbol{a}\rangle^{-s}\langle b\rangle^{-t},
\end{aligned} \tag{53}$$

which asserts the integrability as below

$$\int_{\mathbb{R}^m\times\mathbb{R}} \sup_{\boldsymbol{x}\in K} |R[f_c;\rho](\boldsymbol{a},b)\sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b)|\mathrm{d}\boldsymbol{a}\mathrm{d}b \lesssim \int_{\mathbb{R}^m\times\mathbb{R}} \langle\boldsymbol{a}\rangle^{-s}\langle b\rangle^{-t}\mathrm{d}\boldsymbol{a}\mathrm{d}b < \infty. \tag{54}$$

Step 3 ($f_c \sim f_V \sim f_n$). Next, take a compact domain ($m+1$-dimensional hypercube) $V := \{(\boldsymbol{a},b) \in \mathbb{R}^m \times \mathbb{R} \mid |a_i| \leqslant \delta/2, |b| \leqslant \delta/2\}$, and put a band-limited function

$$f_V(\boldsymbol{x}) := \int_V R[f_c;\rho](\boldsymbol{a},b)\sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b)\mathrm{d}\boldsymbol{a}\mathrm{d}b, \tag{55}$$

so that $\|f_c - f_V\|_{C(K)} < \varepsilon/3$ (by letting δ sufficiently large). Then, let $V = \bigsqcup_{i\in I_n} V_{ni}$ be a decomposition of the domain V into the union of disjoint family of $|I_n| = n^{m+1}$ cubes with volume $\mathrm{vol}(V_n) = (\delta/n)^{m+1}$ and the longest diagonal $d_n = \sqrt{m+1}\delta/n$. From each cube, take a point $(\boldsymbol{a}_{ni}, b_{ni}) \in V_{ni}$ as a center of gravity, that is, so that $c_{ni} = \int_{V_{ni}} R[f_c;\rho](\boldsymbol{a},b)\mathrm{d}\boldsymbol{a}\mathrm{d}b = R[f_c;\rho](\boldsymbol{a}_{ni}, b_{ni})\,\mathrm{vol}(V_n)$, and put $w_{ni} := R[f_c;\rho](\boldsymbol{a}_{ni}, b_{ni})$, then put a finite network as

$$f_n(\boldsymbol{x}) := \sum_{i\in I_n} c_{ni}\sigma(\boldsymbol{a}_{ni}\cdot\boldsymbol{x} - b_{ni}). \tag{56}$$

Step 4 ($f_V \sim f_n$). We show $f_n \to f_V$ in $C(K)$. First, the integrands converge to the limit at almost every $(\boldsymbol{a},b) \in V_{ni}$ as

$$\begin{aligned}
&\sup_{\boldsymbol{x}\in K}\left|R[f_c;\rho](\boldsymbol{a},b)\sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b) - w_{ni}\sigma(\boldsymbol{a}_{ni}\cdot\boldsymbol{x} - b_{ni})\right| \\
&\leqslant \sup_{\boldsymbol{x}\in K}\left|R[f_c;\rho](\boldsymbol{a},b)\right|\left|\sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b) - \sigma(\boldsymbol{a}_{ni}\cdot\boldsymbol{x} - b_{ni})\right|
\end{aligned} \tag{57}$$

$$+ \left|R[f_c;\rho](\boldsymbol{a}_{ni}, b_{ni}) - R[f_c;\rho](\boldsymbol{a}, b)\right| \left|\sigma(\boldsymbol{a}_{ni}\cdot\boldsymbol{x} - b_{ni})\right| \tag{58}$$

$$\leqslant \|R[f_c;\rho]\|_\infty \operatorname{Lip}(\sigma) \sup_{\boldsymbol{x}\in K} \left|(\boldsymbol{a}-\boldsymbol{a}_{ni})\cdot\boldsymbol{x} - (b - b_{ni})\right|$$

$$+ \operatorname{Lip}(R[f_c;\rho]) \left|(\boldsymbol{a}-\boldsymbol{a}_{ni}, b - b_{ni})\right| \|\sigma\|_{L^\infty(\mathbb{R})} = O(\delta/n) \to 0 \quad n\to\infty. \tag{59}$$

Besides, the integrands are uniformly bounded as

$$\sup_{\boldsymbol{x}\in K} \left|w_{ni}\sigma(\boldsymbol{a}_{ni}\cdot\boldsymbol{x} - b_{ni})\right| \leqslant \left|R[f_c;\rho](\boldsymbol{a}, b)\right| \|\sigma\|_{L^\infty(\mathbb{R})}, \text{ for a.e. } (\boldsymbol{a}, b)\in V_{ni}. \tag{60}$$

Therefore, by the dominated convergence theorem for the Bochner integral, we have

$$\|f_V - f_n\|_{C(K)} = \sup_{\boldsymbol{x}\in K} \left| \sum_{i\in I_n} \int_{V_{ni}} R[f_c;\rho](\boldsymbol{a}, b)\sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b)\mathrm{d}\boldsymbol{a}\mathrm{d}b - \sum_{i\in I_n} c_{ni}\sigma(\boldsymbol{a}_{ni}\cdot\boldsymbol{x} - b_{ni}) \right| \tag{61}$$

$$\leqslant \sum_{i\in I_n} \int_{V_{ni}} \sup_{\boldsymbol{x}\in K} \left|R[f_c;\rho](\boldsymbol{a}, b)\sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b) - w_{ni}\sigma(\boldsymbol{a}_{ni}\cdot\boldsymbol{x} - b_{ni})\right| \mathrm{d}\boldsymbol{a}\mathrm{d}b \tag{62}$$

$$\to 0, \quad n\to\infty. \tag{63}$$

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

To sum up, we have shonw the *cc*-universality:

$$\|f - f_n\|_{C(K)} \leqslant \|f - f_c\|_{C(K)} + \|f_c - f_V\|_{C(K)} + \|f_V - f_n\|_{C(K)} < \varepsilon. \tag{64}$$

□

Notes. In the proof, we employed a naive discretization based on the regular grids in V. However, since we know the closed-form expression of the ridgelet transform, we can discretize it more effectively. For example, a better discretization scheme is investigated in the so-called *Maurey-Jones-Barron (MJB) theory* and the dimension independent *Barron's bound* ([see, e.g., 65]).