Universality of Group Convolutional Neural Networks Based on Ridgelet Analysis on Groups

Sho Sonoda
RIKEN AIP
sho.sonoda@riken.jp

Isao Ishikawa
Ehime University
ishikawa.isao.zx@ehime-u.ac.jp

Masahiro Ikeda
RIKEN AIP
masahiro.ikeda@riken.jp

Abstract

We show the universality of depth-2 group convolutional neural networks (GCNNs) in a unified manner based on the ridgelet theory. Despite widespread use in applications, the approximation property of (G)CNNs has not been well investigated. The universality of (G)CNNs has been shown since the late 2010s. Yet, our understanding on how (G)CNNs represent functions is incomplete because the past universality theorems have been shown case-by-case manner by manually/carefully designing the network parameters depending on the variety of convolution layers, and either existential or indirect manner by converting/modifying the (G)CNNs into other universal approximators such as invariant polynomials and fully-connected networks. In this study, we formulate a versatile depth-2 continuous GCNN $S[\gamma]$ as a nonlinear mapping between group representations, and directly obtain an analysis operator, called the *ridgelet trasform*, that maps a given function f to the network parameter γ so that $S[\gamma] = f$. The proposed GCNN covers typical GCNNs such as the cyclic convolution on multi-channel images, networks on permutation-invariant inputs (Deep Sets), and $\mathrm{E}(n)$-equivariant networks. The closed-form expression of the ridgelet transform can describe how the network parameters are organized to represent a function. While it has been known only for fully-connected networks, this study is the first to obtain the *ridgelet transform for GCNNs*. By discretizing the closed-form expression, we can systematically generate a constructive proof of the *cc-universality of finite GCNNs*. As a consequence, our universality proofs are more unified, direct and constructive than previous proofs.

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]. Despite the rapid growth of diversity, 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 (G)CNNs are *universal*; that is, they can approximate some class of continuous maps with any precision [9–15]. These studies are still limited because the proofs are shown (1) case-by-case by carefully assigning the parameters for a network to approximate a given function f, which means that once the network architecture is modified, then we need to manually reassign the parameters from scratch, and (2) either existential or indirect by converting/modifying the (G)CNNs into other universal approximators such as invariant polynomials and FNNs.

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

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

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 context, (G)CNN studies are at the stage of case-by-case and indirect proofs, indicating that we know about (G)CNN parameters only indirectly. (See § 6.1 for more details).

In this study, we show the universality of depth-2 GCNNs by devising a general notion of group convolution and developing the *ridgelet transform for GCNNs*—an analysis operator that maps a given function f to the weight parameter γ in a single hidden layer of a neural network. Consequently, our universality proof is more *unified, direct and constructive* because our GCNN covers a wide range of typical GCNNs, and the ridgelet transform clearly indicate how to assign the network parameters.

In the following, we describe the formulation of GCNNs to overview our main contributions.

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 an 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 $(i,j) \in [m_1 - w_1 + 1] \times [m_2 - w_2 + 1]$.

For technical reasons, we suppose that the output channels (indexed by $\ell \in [n_{out}]$) are aggregated soon after the activation function, which may be slightly different from an ordinary formulation of CNNs, but we can understand this as a part of the subsequent layer.

In the standard formulation of GCNNs, a multi-channel image is understood as a vector-valued function on a group G or a homogeneous space G/H, such as a product group $G = \mathbb{Z}_{m_1} \times \mathbb{Z}_{m_2}$ of cyclic groups $\mathbb{Z}_{m_i} := \mathbb{Z}/m_i\mathbb{Z}$ $(i = 1, 2)$. (More geometrically, Cohen and Welling [32] phrased it as 'a section of a fiber bundle'). The convolution in the pixel directions (i,j) is reformulated as a group convolution with respect to the product group, and the inner product in the channel direction k is understood as the convolution with respect to the trivial action of G on a 'fiber' $\mathbb{R}^{n_{in}}$.

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, $\mathcal{X}$ be an *arbitrary* Hilbert space of feature vector x and filter a, and $\gamma : \mathcal{X} \times \mathbb{R} \to \mathbb{C}$ be a function called the *parameter distribution*. We formulate a group convolution layer in an integral form, called the *integral representation*, as

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

This is an infinite-dimensional reparametrization of a depth-2 GCNN; that is, 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), and the integration over (a,b) implies that all the possible neurons are assigned in advance. Thus $S[\gamma]$ can be understood as a *continuous neural network*. We note, however, that if we put γ as 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 represent a finite model

$$S[\gamma_n](x)(g) = \sum_{\ell=1}^{n} c^{\ell} \sigma((a^{\ell} * x)(g) - b^{\ell}), \quad x \in \mathcal{X},\ g \in G. \tag{3}$$

In summary, $S[\gamma]$ is a mathematical model of shallow neural networks with *any* width ranging from finite to continuous. In particular, the sparsity/low-rankness of parameters are reflected as the localization/concentration of parameter distribution γ.

An advantage to use the integral representation is the *linearization trick* (on a single hidden layer). Whereas a finite network $S[\gamma_n]$ is *nonlinear* in the original parameters (a,b), the integral representation $S[\gamma]$ 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. In this study, we have obtained its closed-form expression, that is,

$$R[f;\rho](a,b) := \int_{\mathcal{X}} f(x)(e)\overline{\rho(\langle a,x\rangle_{\mathcal{X}} - b)}\mathrm{d}a\mathrm{d}b, \quad (a,b) \in \mathcal{X} \times \mathbb{R}, \tag{4}$$

where $f : \mathcal{X} \to \mathbb{C}^G$ is a target vector-valued nonlinear function to be approximated, called a *feature map*, $e \in G$ is the identity element, and $\rho : \mathbb{R} \to \mathbb{C}$ is an auxiliary function, called the *ridgelet function*. Provided that f is *group equivariant*, then under mild regularity assumptions, it satisfies the *reconstruction formula*

$$S[R[f;\rho]] = (\!(\sigma,\rho)\!)f, \tag{5}$$

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. Particularly, the reconstruction formula $S[R[f;\rho]] = (\!(\sigma,\rho)\!)f$ corresponds to the Fourier inversion formula $F^{-1}[F[f]] = f$.

An advantage of the ridgelet transform is the *closed-form expression*. Despite the common belief that neural network parameters are a blackbox, the closed-form expression can clearly describe 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] 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 they [45] have shown that finite networks trained by regularized empirical risk minimization (RERM) converges to a certain unique ridgelet transform. (We note that NTK and the Gibbs distribution can also describe the parameter distribution, but NTK is limited to the kernel regime, and the Gibbs distribution is given only implicitly.) As an application, Savarese et al. [46] and their followers [47–49] have established the *representer theorems* for ReLU-FNNs by using the ridgelet transform. Although the parallel results for CNNs have not yet been published, we anticipate that the ridgelet transform could facilitate our understanding of deep learning solutions.

Challenges and Contributions. The closed-form expression of the ridgelet transform has been known only for FNNs, which was discovered in the 1990s independently by Murata [26], Candès [27] and Rubin [28]. (We refer to [50–52] for ridgelet analysis in the 2000s, and [53–55] for more recent results.) One of the difficulties to obtain 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" (see e.g. [13]), but it is somewhat misleading because such an equivalence holds only when both CNNs and FNNs are very carefully designed. While FNNs are defined on the Euclidean space $\mathbb{R}^m$, GCNNs are defined on a more abstract space $\mathcal{X}$. For example, since the convolution on the Euclidean space can be written using Töplitz matrices, one could consider a formulation such as $\int_{\mathbb{R}^{k\times m}\times\mathbb{R}^k} \gamma(A,b)\sigma(Ax-b)\mathrm{d}A\mathrm{d}b$ where the parameter A is an $k \times m$-matrix. However, this only leads to another ridgelet transform that covers less symmetries G. In fact, it is a version of the so-called k-plane ridgelet transform developed in the 2000s [50].

To circumvent this difficulty, we formulate GCNNs as general as possible by dealing with the feature space $\mathcal{X}$, group G, and representation T in a coordinate-free manner. Eventually, we have shown the reconstruction formula for a wide range of GCNNs (as displayed in § 5), 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 Notation and Basic Terminologies

Notation. For any integer $n > 0$, $[n]$ denotes the set $\{1,\dots,n\}$. For any sets G and $\mathbb{K}$, $\mathbb{K}^G$ denotes the collection $\{G \to \mathbb{K}\}$ of all mappings from G to $\mathbb{K}$. For any topological space X, $C(X)$ and $C_c(X)$ denote the collections of all continuous functions on X, and continuous functions on X with compact support, respectively. We note that when X is compact, then $C(X) = C_c(X)$. For any measure space X and number $p \in [1,\infty]$, $L^p(X)$ denotes the space of p-integrable functions on X.

2.1 Fourier Analysis on $\mathbb{R}^d$

We refer to [56, 57, 53] for more details on Fourier transform and tempered distributions ($\mathcal{S}'$).

Schwartz Distributions. 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, $\mathcal{S}'(\mathbb{R})$ and $\mathcal{S}(\mathbb{R})$ are assigned as 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 equation. When $f \in L^2(\mathbb{R}^d)$, the inversion formula holds "in L^2", which is not a pointwise equation because the equation "$f = g$ 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 $\tanh$.

2.2 Group Representation

Let G be a group, let $\mathcal{X}$ be a vector space over a field $\mathbb{K}$, and let $GL(\mathcal{X})$ be the general linear group on $\mathcal{X}$. A *group representation* T of the group G on the vector space $\mathcal{X}$ is a group homomorphism from G to $GL(\mathcal{X})$, that is, a map $T : G \to GL(\mathcal{X}); g \mapsto T_g$ 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 $G \times \mathcal{X} \to \mathcal{X}; (g, x) \mapsto T_g[x]$ be continuous. Here, $\mathcal{X}$ is called the *representation space*, or sometimes the *G-module*.

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

$$L_g[x](h) := x(g^{-1}h), \quad g, h \in G,\ x \in \mathcal{X} = \mathbb{K}^G. \tag{6}$$

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 $\mathcal{X} = L^2(G)$ is a unitary representation.

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

$$\langle T_g[x], y\rangle_{\mathcal{X}} = \langle x, T_{g^{-1}}^*[y]\rangle_{\mathcal{X}}, \quad g \in G,\ x, y \in \mathcal{X}. \tag{7}$$

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

$$f_{a,x}(g) := a[T_g[x]], \quad g \in G,\ x \in \mathcal{X},\ a \in \mathcal{X}' \tag{8}$$

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

$$f_{a,x}(g) = \langle T_g[x], a\rangle_{\mathcal{X}}, \quad g \in G,\ a, x \in \mathcal{X}. \tag{9}$$

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. (See, e.g., [58, 59]). In this study, we show the so-called *cc-univesality*, one of the standard universalities in the machine learning 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 $g \in \mathtt{NN}$ such that

$$\left\| f - g|_K \right\|_{C(K)} := \sup_{x \in K} |f(x) - g(x)| < \varepsilon, \tag{10}$$

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

3 Functions on Abstract Hilbert Space $\mathcal{X}$

We introduce an extended group convolution on $\mathcal{X}$, a uniform norm and the group-equivariance for functions on $\mathcal{X}$, an induced measure and an induced Fourier transform on $\mathcal{X}$, and a projection to $\mathcal{X}_m$.

3.1 (G, T)-Convolution $*_T : \mathcal{X} \times \mathcal{X} \to \mathbb{C}^G$

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

$$(a *_T x)(g) := \langle x, T_g^*[a] \rangle_{\mathcal{X}} = \langle T_{g^{-1}}[x], a \rangle_{\mathcal{X}}. \tag{11}$$

We remark (1) that this is simply a paraphrase of the *matrix element of a group representation* (see the previous section), and (2) that this is *not necessarily* a binary operation because $\mathcal{X} \neq \mathbb{C}^G$ in general. Nevertheless, we call it a *convolution* simply because it covers a wide range of 'group convolutions' in today's GCNN literature.

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

$$\langle x, T_g^*[a] \rangle_{L^2(G)} = \int_G x(h) \overline{a(g^{-1}h)} \mathrm{d}\mu(h) = \int_G x(h) \widetilde{a}(h^{-1}g) \mathrm{d}\mu(h) = (x *_G \widetilde{a})(g), \tag{12}$$

where $\widetilde{a}(g) := \overline{a(g^{-1})}$ is an involution.

Example 2. 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 $G = \mathbb{Z}_{m_1} \times \mathbb{Z}_{m_2}$, $\mathcal{X} = \mathbb{R}^{m_1 \times m_2 \times n}$, and $T_{(p,q)}^*[\boldsymbol{a}](i, j, k) = a_{i-p, j-q}^k$, then

$$\langle \boldsymbol{x}, T_{(p,q)}^*[\boldsymbol{a}] \rangle_{\mathbb{R}^{m_1 \times m_2 \times n}} = \sum_{i,j,k} x_{ij}^k a_{i-p, j-q}^k. \tag{13}$$

While the post-activation feature $\sigma((a * x)(g) - b)$ is a function on G, the input feature x can be an arbitrary abstract vector, which is more general than typical GCNN formulations where feature x is supposed to be a vector-valued function on G or G/H. This is an advantage for better understanding of CNNs, since the theory becomes free from the specification of x.

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

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

$$f(T_g[x])(h) = L_g[f(x)](h) = f(x)(g^{-1}h), \quad x \in \mathcal{X},\ g, h \in G. \tag{14}$$

Here, we restrict the definition for a special case of the regular representation L. This is simply due to the fact that our GCNN satisfies this case.

Definition 3. Let G be a topological group. For any vector-valued function $f : \mathcal{X} \to \mathbb{C}^G$, put

$$\|f\|_{C(\mathcal{X}; C(G))} := \|f\|_{C(\mathcal{X}) \to C(G)} := \sup_{x \in \mathcal{X}} \Big| \sup_{g \in G} |f(x)(g)| \Big|. \tag{15}$$

By $C_{equi}(\mathcal{X}; C(G))$, we denote the normed vector space of all *continuous* (G, T)-*equivariant* $C(G)$-*valued* functions on $\mathcal{X}$ equipped with the uniform norm $\| \cdot \|_{C(\mathcal{X}; C(G))}$.

We note that the topology of uniform norm $\| \cdot \|_{C(\mathcal{X}; C(G))}$ is stronger than the topology of compact convergence, which is employed in the *cc*-universality argument. In fact, $C_{equi}(\mathcal{X}; C(G))$ need not be complete (or Banach) to show the *cc*-universality.

3.3 Induced Lebesgue Measure λ and Induced Fourier Transform $\widehat{\cdot}$ on Subspace $\mathcal{X}_m$

Let $\mathcal{X}_m$ denote an m-dimensional subspace of $\mathcal{X}$, and let $\{\mathsf{e}_i\}_{i\in[m]}$ be an orthonormal basis of $\mathcal{X}_m$.

Induced Lebesgue Measure on $\mathcal{X}_m$. We induce the Lebesgue measure λ on $\mathcal{X}_m$ by pushing forward the Lebesgue measure $\mathrm{d}\boldsymbol{x}$ on $\mathbb{R}^m$ via an isometric linear embedding $\phi : \mathbb{R}^m \to \mathcal{X}_m$. For example, take 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 $\mathcal{X}_m$ as the push forward measure $\lambda = \phi_\sharp \mathrm{d}\boldsymbol{x}$ so that the volume of a hypercube $Q = \{\sum_{i\in[m]} c_i \mathsf{e}_i \mid c_i \in [a_i, b_i]\}$ in $\mathcal{X}_m$ is calculated as $\lambda(Q) = \prod_{i\in[m]} |b_i - a_i|$, and the integration of a measurable function $f : \mathcal{X}_m \to \mathbb{C}$ over a measurable set $E \subset \mathcal{X}_m$ is calculated as

$$\int_E f(x)\mathrm{d}\lambda(x) = \int_{\mathbb{R}^m} 1_E \left(\sum_{i=1}^m x_i \mathsf{e}_i\right) f\left(\sum_{i=1}^m x_i \mathsf{e}_i\right) \prod_{i=1}^m \mathrm{d}x_i = \int_{\phi^{-1}(E)} f \circ \phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}. \tag{16}$$

As far as there is no risk of confusion, we denote $\mathrm{d}x$ instead of $\mathrm{d}\lambda(x)$.

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

$$\widehat{f}(y) := \int_{\mathcal{X}_m} f(x) e^{-i\langle x,y\rangle_{\mathcal{X}_m}} \mathrm{d}\lambda(x), \quad f(x) \overset{\star}{=} \frac{1}{(2\pi)^m} \int_{\mathcal{X}_m} \widehat{f}(y) e^{i\langle x,y\rangle_{\mathcal{X}_m}} \mathrm{d}\lambda(y). \tag{17}$$

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

We remark (1) that once the subspace $\mathcal{X}_m$ is fixed, the induced Fourier transform is unique up to the orthogonal transformation of the basis $\{\mathsf{e}_i\}_{i\in[m]}$, and (2) that the induced Fourier transform "on $\mathcal{X}$" should not be confused with the Fourier transform "on group G". Especially, this *cannot* map a convolution $x *_T a$, an element in $\mathbb{C}^G$, to a point product such as "$\widehat{x} \cdot \widehat{a}$".

3.4 Projection $P : \mathcal{X} \to \mathcal{X}_m$ and Extension Operator P^*

In order to induce the Lebesgue measure λ, we assume that the dimension of $\mathcal{X}_m$ to be finite. As a side effect of this assumption, the image $T_G[\mathcal{X}_m] := \{T_g[x] \mid g \in G, x \in \mathcal{X}_m\}$ can extend toward the outside of $\mathcal{X}_m$. (That is, $\mathcal{X}_m$ is not always G-invariant: $T_G[\mathcal{X}_m] \not\subset \mathcal{X}_m$.) To avoid an "undefined error" such as to input x outside of $\mathcal{X}_m$ for a function f defined only on $\mathcal{X}_m$, we introduce projection P and extension P^* as below. When $\dim \mathcal{X} < \infty$, we can omit P by putting $\mathcal{X}_m = \mathcal{X}$ (so $P = \mathrm{Id}$), because by the definition of the group representation, always $T_G[\mathcal{X}] = \mathcal{X}$.

Let $\mathcal{X}_m^\perp$ denote the orthogonal complement of $\mathcal{X}_m$ in $\mathcal{X}$. Let $P : \mathcal{X} \to \mathcal{X}_m$ denote the orthogonal projection onto $\mathcal{X}_m$. For any function $f : \mathcal{X}_m \to \mathbb{C}^G$, put

$$P^*f(z)(g) := f(P(z))(g), \quad z \in \mathcal{X}, g \in G. \tag{18}$$

This extends f (on a subspace $\mathcal{X}_m$) to the entire space $\mathcal{X}$ as a constant function on $\mathcal{X}_m^\perp$; that is, $P^*f(x \oplus y) = f(x \oplus 0)$ for each $x \oplus y \in \mathcal{X}_m \oplus \mathcal{X}_m^\perp$.

4 Main Results

We introduce the (G, T)-convolutional neural networks and the corresponding ridgelet transform, and present the reconstruction formula for continuous GCNNs and the cc-universality for finite GCNNs.

Throughout this section, we fix a representation $T : G \to GL(\mathcal{X})$ of a group G on a (potentially infinite-dimensional) Hilbert space $\mathcal{X}$ over a field $\mathbb{K}$ endowed with an inner product $\langle\cdot,\cdot\rangle_{\mathcal{X}}$, and fix an m-dimensional closed subspace $\mathcal{X}_m$ of $\mathcal{X}$ equipped with an induced Lebesgue measure λ. Let $k := \dim_{\mathbb{R}} \mathbb{K}$ denote the real dimension of $\mathbb{K}$, that is, $k = 1$ for $\mathbb{K} = \mathbb{R}$ and $k = 2$ for $\mathbb{K} = \mathbb{C}$. Let e denote the identity element of G.

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

Definition 4. For any functions $\gamma : \mathcal{X}_m \times \mathbb{K} \to \mathbb{C}$ and $\sigma : \mathbb{K} \to \mathbb{C}$, we define the *integral representation of* (G, T)*-convolutional neural network* as a vector-valued function $\mathcal{X} \to \mathbb{C}^G$,

$$S[\gamma](x)(g) := \int_{\mathcal{X}_m \times \mathbb{K}} \gamma(a, b)\sigma((a *_T x)(g) - b)\mathrm{d}\lambda(a)\mathrm{d}b, \quad x \in \mathcal{X},\ g \in G. \tag{19}$$

Here, we call γ a parameter distribution, 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 $g,h \in G$,

$$S[\gamma](T_g[x])(h) = \int_{\mathcal{X}_m \times \mathbb{K}} \gamma(a,b)\sigma(\langle T_{(g^{-1}h)^{-1}}[x], a\rangle_{\mathcal{X}} - b)\mathrm{d}a\mathrm{d}b = S[\gamma](x)(g^{-1}h). \tag{20}$$

In addition, at the identity element $g = e$, it is reduced to a FNN:

$$S[\gamma](x)(e) = \int_{\mathcal{X}_m \times \mathbb{K}} \gamma(a,b)\sigma(\langle x, a\rangle_{\mathcal{X}} - b)\mathrm{d}a\mathrm{d}b, \quad x \in \mathcal{X} \tag{21}$$

and it satisfies a *projection property*:

$$S[\gamma](P[x])(e) = S[\gamma](x)(e), \quad x \in \mathcal{X}. \tag{22}$$

4.2 Ridgelet Transform and Scalar Product of Activation Function

Definition 5. For any functions $f : \mathcal{X}_m \to \mathbb{C}^G$ and $\rho : \mathbb{K} \to \mathbb{C}$, we define the *ridgelet transform* as

$$R[f;\rho](a,b) := \int_{\mathcal{X}_m} f(x)(e)\overline{\rho(\langle x, a\rangle_{\mathcal{X}} - b)}\mathrm{d}x, \quad (a,b) \in \mathcal{X}_m \times \mathbb{K}. \tag{23}$$

Here e denotes the identity element of G.

Definition 6. For any tempered distribution $\sigma \in \mathcal{S}'(\mathbb{K})$ and function $\rho \in \mathcal{S}(\mathbb{K})$, put a *scalar product* as

$$(\!(\sigma,\rho)\!) := (2\pi)^{m-k} \int_{\mathbb{K}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-m}\mathrm{d}\omega. \tag{24}$$

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 may notice that the ridgelet transform for GCNN is formally the same as the one for FNNs, and may wonder why inner product $\langle x, a\rangle_{\mathcal{X}}$ instead of group convolution $(a *_T x)(g)$. Indeed, this is a consequence of two facts (1) that a group convolution at the identity e is reduced to an inner product: $(a *_T x)(e) = \langle x, a\rangle_{\mathcal{X}}$, and (2) that when f is (G,T)-equivariant, then the value $f(x)(g)$ at each $g \in G$ is determined by translating the value $f(x)(e)$ at the identity.

4.3 Reconstruction Formula, or the Universality of Continuous GCNNs

We state the first half of our main results. For $f : \mathcal{X}_m \to \mathbb{C}^G$, we write $f_e(x) := f(x)(e)$ for short.

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

$$P^*f(T_g[z])(h) = f(z)(g^{-1}h), \quad \textit{for every } z \in \mathcal{X} \textit{ and } g,h \in G; \tag{25}$$

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., $f_e, \widehat{f}_e \in L^1(\mathcal{X}_m)$, (A2b) f_e is square-integrable, i.e., $f_e \in L^2(\mathcal{X}_m)$, or (A2c) f_e is a tempered distribution, i.e., $f_e \in \mathcal{S}'(\mathcal{X}_m)$. Then, the following reconstruction formula holds:

$$S[R[f;\rho]](x)(g) = \int_{\mathcal{X}_m \times \mathbb{K}} R[f;\rho](a,b)\sigma((a *_T x)(g) - b)\mathrm{d}a\mathrm{d}b \overset{\star}{=} (\!(\sigma,\rho)\!) f(x)(g), \tag{26}$$

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 of Finite GCNNs

Finally, we state the second half of our main results. Let $\mathtt{NN}$ be the collection of finite GCNNs, that is,

$$\mathtt{NN} := \bigcup_{n\in\mathbb{N}} \left\{ f_n(x)(g) = \sum_{i=1}^{n} c_i \sigma((a_i *_T x)(g) - b_i) \middle| (a_i, b_i, c_i) \in \mathcal{X}_m \times \mathbb{K} \times \mathbb{C}, i \in [n] \right\}. \quad (27)$$

Since the reconstruction formula $S[\gamma_f] = f$ with $\gamma_f = R[f;\rho]$ holds for an arbitrary function f, we can construct a sequence $\{f_n\}_{n\in\mathbb{N}}$ of finite (G,T)-CNNs that converges to an arbitrary target function f, namely

$$f_n \to f \quad \text{as} \quad n \to \infty, \quad (28)$$

by discretizing the continuous network $S[\gamma_f]$ and distribution γ_f into finite sums

$$f_n(x)(g) := S[\gamma_n](x)(g) = \sum_{i=1}^{n} c_i \sigma((a_i *_T x)(g) - b_i) \quad \text{with} \quad \gamma_n := \sum_{i=1}^{n} c_i \delta_{(a_i,b_i)} \quad (29)$$

in a 'nice' manner so that $\gamma_n \to \gamma_f = R[f;\rho]$ as $n \to \infty$. 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 a regularity assumption on the activation function σ, we introduce the forward difference operator Δ_θ^n with difference $\theta > 0$, defined as

$$\Delta_\theta^1[\sigma](t) := \sigma(t+\theta) - \sigma(t), \quad \Delta_\theta^{n+1}[\sigma](t) := \Delta_\theta^1 \circ \Delta_\theta^n[\sigma](t). \quad (30)$$

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. Then, $\mathtt{NN}$ is cc-universal, that is, for any continuous (G,T)-equivariant $C(G)$-valued function $f \in C_{equi}(\mathcal{X}_m; C(G))$, and for any compact sets $K \subset \mathcal{X}_m$ and $L \subset G$, there exists a sequence $\{f_n\}_{n\in\mathbb{N}} \subset \mathtt{NN}$ of finite GCNNs satisfying*

$$\|f - f_n\|_{C(K;C(L))} = \sup_{x\in K} \sup_{g\in L} |f(x)(g) - f_n(x)(g)| \to 0, \quad n \to \infty. \quad (31)$$

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

5 Examples

We display the ridgelet transforms and reconstruction formulas for a few typical GCNNs. Besides, we calculated in Examples 5 and 8 the ridgelet transforms of a differential filter, which is often reported to be acquired as a feature map in the first layer of deep CNNs for image recognition.

5.1 Finite Periodic Convolution Layer

Example 3 (For 1-dimensional periodic signals)**.** The periodic convolution corresponds to the case when $\mathbb{K} = \mathbb{R}, G = \mathbb{Z}_m \cong [m] = \{0, 1, \ldots, m-1\}$, $\mathcal{X} = L^2(G) \cong \mathbb{R}^m$ 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 4 (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 $\mathcal{X} \cong \mathbb{R}^{m\times n\times c}$. Let x_{ij}^k denote the (i,j)-th component in the k-th channel of $x \in \mathcal{X}$. 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 5 (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 $f(x)(0) = x_1 - x_0$,

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

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 6. Let $\mathcal{X} = \mathbb{R}^m, G \leqslant \mathrm{S}_m$ and $T_g[x] = (x_{g^{-1}(1)}, x_{g^{-1}(2)}, \ldots, x_{g^{-1}(m)})$. Thus $\langle a, x\rangle_{\mathcal{X}} = \frac{1}{m}\sum_{i\in[m]} a_ix_i$, and $(a *_T x)(g) = \frac{1}{m}\sum_{p\in[m]} a_px_{g(p)}$. So,

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

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

5.3 Continuous Periodic Convolution Layer

Example 7. 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 $\mathcal{X}$ to be an m-dimensional subspace $\mathcal{X} := \{\sum_{|n|<m} x_ne^{in\theta} \mid x_{-n} = x_n \in \mathbb{R}\}$ equipped with an inner product $\langle x, y\rangle_{\mathcal{X}} := \int_{\mathbb{T}} x(\theta)\overline{y(\theta)}\mathrm{d}\theta$. We note that the constraint $x_n = x_{-n}$ implies $\sum_{|n|<m} x_ne^{in\theta} = \sum_{|n|<m} x_n\cos(n\theta)$ and thus any signal $x \in \mathcal{X}$ 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 $(a * x)(\alpha) = \int_{\mathbb{T}} a(\theta)x(\alpha - \theta)\mathrm{d}\theta = \sum_{|n|<m} a_nx_ne^{in\alpha} = \sum_{|n|<m} a_nx_n\cos(n\alpha)$ (by the convolution theorem and the constraint). Therefore,

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

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

Example 8 (Differential operator (with convergence factor))**.** A differential operator $\frac{\mathrm{d}}{\mathrm{d}\theta}$ is calculated as $x = \sum_{|n|<m} x_ne^{in\theta} \mapsto f(x) = \sum_{|n|<m} nx_ne^{in\theta}$. Since $\boldsymbol{x} \mapsto f(\boldsymbol{x})(0) = \sum_{|n|<m} nx_n$ is not integrable on $\mathbb{R}^m$, we impose a convergence factor ϕ_t as follows. $f_t(x)(\theta) := f(x)\phi_t(x) = \frac{\mathrm{d}}{\mathrm{d}\theta}x(\theta)\phi_t(x) = \sum_{|n|<m} nx_n\phi_t(x)e^{in\theta}$. Here, $(\phi_t)_{t>0} \subset \mathcal{S}(\mathcal{X})$ is a family of convergence factors that satisfies (1) the first moment $\int_{\mathcal{X}} |x|_{\mathcal{X}}|\phi_t(x)|\mathrm{d}x$ 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 $\phi_t(x) = \exp(-|x|_{\mathcal{X}}^2/4t)$. Hence,

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

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

Example 9. 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 $\mathcal{X} \subset L^2(\mathbb{R}^n)$ and $T_{(U,s)}[x](t) := x(U^{-1}(t - s))$. Then,

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

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

We note that a more memory 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, 60].

6 Discussion

6.1 Related Works on (G)CNN Universality

Non-Group CNN. Zhou [9, 10] is the earliest to show the cc-universality of deep ReLU (non-group) CNNs. 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 depth also increases width, which is not covered in our GCNN.

Finite Group CNN. Maron et al. [12], Sannai et al. [61], Keriven and Peyré [62], Ravanbakhsh [63] 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 $\mathcal{X} = \mathbb{R}^{n^k \times a}$ (a-channel k-th order n-dimensional tensors), the output space is $\mathcal{X}' = \mathbb{R}^{n^l \times b}$ (b-channel l-th order n-dimensional tensors), the group G is a subgroup of a symmetric group S_n, and the group action (or representation) T is the left-translation (or left-regular representation). In this setup, they presented the cc-universality of deep-ReLU-GCNNs in the space of continuous G-equivariant functions $C_{equi}(\mathcal{X};\mathcal{X}')$. The proofs are indirect because they are based on *invariant polynomials* or *MLPs*. The finite group cases are essentially covered as Example 6 (Deep Sets).

Lie Group CNN. Yarotsky [11] carefully designed deep GCNNs with Lie groups acting on *infinite-dimensional* input/output spaces, and show a version of universality in the space of continuous G-equivariant functions $C_{equi}(L^2(G;\mathbb{R}^d);L^2(G;\mathbb{R}^{d'}))$. To be precise, G is either a compact group, translation group $\mathbb{R}^d$, or 2-dimensional roto-translation group $\mathrm{SE}(2)$, and the input/output spaces $\mathcal{X}$ and $\mathcal{X}'$ are square-integrable functions on G. The proposed networks are not covered in our GCNNs, but several infinite group cases are covered in Examples 7 and 9.

Remarkably, Kumagai et al. [14] introduced an integral representation that covers LCH groups, and showed the universality. The proposed integral representation is based on the Haar measure, thus slightly different from ours. The proofs are indirect because the network is converted to an MLP.

6.2 Review of Assumptions

Group G. We only assume G to be a topological group, to deal with continuous functions on G. Thus, a quite large class of groups are covered, for example, all the finite groups such as $\mathbb{Z}_n$ and S_n, compact groups such as $\mathrm{SO}(n)$ and $\mathrm{U}(n)$, and non-compact groups such as $\mathbb{R}^n$ and $\mathrm{E}(n)$ as well.

Representation Space $\mathcal{X}$. Unlike previous studies, it does *not* need to be a function space such as $C(G)$ and $L^2(G/H)$, but it only needs to be an abstract Hilbert space, which is one of the major advantages for geometric understanding of GCNNs. On the other hand, we introduce an auxiliary finite-dimensional subspace $\mathcal{X}_m$ (and projection P), to use the Fourier inversion formula on the finite-dimensional Euclidean space $\mathbb{R}^m$ in the proof. We conjecture that the extension to an infinite-dimensional setting would be a routine for some specialists in functional analysis.

Group Representation T. It does *not* need to be unitary, irreducible, nor square-integrable, since the proof is based only on a few basic properties of the linear group representation.

Network Architecture. The ridgelet theory supports a wide class of *activation functions*, namely, the tempered distributions ($\mathcal{S}'$). The extension to *deep GCNNs* remains an important open question.

Acknowledgments and Disclosure of Funding

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

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 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, 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 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 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, 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, Akiyoshi Sannai, and Makoto Kawano. Universal approximation with neural networks on function spaces. *Journal of Experimental & Theoretical Artificial Intelligence*, pages 1–12, 2022.

[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, 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 *Advances in Neural Information Processing Systems 31*, pages 8571–8580, 2018.

[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, 2019.

[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, 2019.

[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*, 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 *Advances in Neural Information Processing Systems 31*, pages 7146–7155, 2018.

[41] Lénaïc Chizat and Francis Bach. On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport. In *Advances in Neural Information Processing Systems 32*, pages 3036–3046, 2018.

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

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

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

[51] 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.

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

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

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

[55] Sho Sonoda, Isao Ishikawa, and Masahiro 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, 2022.

[56] Loukas Grafakos. *Classical Fourier Analysis*. Graduate Texts in Mathematics. Springer New York, 2 edition, 2008.

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

[58] 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.

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

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

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

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

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

[64] 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

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 $x \in \mathcal{X}$ and $b \in \mathbb{K}$, respectively. For example,

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

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 $\phi \in \mathcal{S}'(\mathbb{R}^d)$,

$$\triangle^{t/2}[\phi](\boldsymbol{v}) := \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} |\boldsymbol{u}|^t \widehat{\phi}(\boldsymbol{u}) e^{i\boldsymbol{u}\cdot\boldsymbol{v}} \mathrm{d}\boldsymbol{u}, \tag{32}$$
$$\langle \triangle \rangle^{t/2}[\phi](\boldsymbol{v}) := \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} (1 + |\boldsymbol{u}|^2)^{t/2} \widehat{\phi}(\boldsymbol{u}) e^{i\boldsymbol{u}\cdot\boldsymbol{v}} \mathrm{d}\boldsymbol{u}. \tag{33}$$

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 $T : G \to GL(\mathcal{X})$ of a group G on a (potentially infinite-dimensional) Hilbert space $\mathcal{X}$ over a field $\mathbb{K}$ equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{X}}$, which is a G-invariant vector space: $T_G[\mathcal{X}] = \mathcal{X}$, and a finite-dimensional closed subspace $\mathcal{X}_m \subset \mathcal{X}$ equipped with the Lebesgue measure λ. Let $\mathcal{X}_m^\perp$ be the orthogonal complement of $\mathcal{X}_m$ in $\mathcal{X}$, i.e., $x \oplus y \in \mathcal{X}_m \oplus \mathcal{X}_m^\perp = \mathcal{X}$, and let $P : \mathcal{X} \to \mathcal{X}_m$ denote the orthogonal projection onto $\mathcal{X}_m$. Let $m := \dim_{\mathbb{R}} \mathcal{X}_m$ denotes the real dimension of $\mathcal{X}_m$, 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. $a \in \mathcal{X}_m$, and (A") that $\gamma^\sharp \sigma^\sharp \in L^1(\mathcal{X}_m \times \mathbb{K})$, which will be eventually justified because later in (38), 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*:

$$S[\gamma](x)(g) = \frac{1}{(2\pi)^k} \int_{\mathcal{X}_m} \int_{\mathbb{K}} \gamma(a, b) \sigma(\langle T_{g^{-1}}[x], a \rangle_{\mathcal{X}} - b) \mathrm{d}b \mathrm{d}a \tag{34}$$
$$= \frac{1}{(2\pi)^k} \int_{\mathcal{X}_m} \int_{\mathbb{K}} \gamma^\sharp(a, \omega) \sigma^\sharp(\omega) \exp\left(i\omega \langle T_{g^{-1}}[x], a \rangle_{\mathcal{X}}\right) \mathrm{d}\omega \mathrm{d}a. \tag{35}$$

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:

$$\int_{\mathcal{X}_m \times \mathbb{K}} |\gamma^\sharp(a, \omega) \sigma^\sharp(\omega) \exp(i\omega(a * x)(g))| \mathrm{d}a \mathrm{d}\omega = \|\gamma^\sharp \sigma^\sharp\|_{L^1(\mathcal{X}_m \times \mathbb{K})} < \infty, \tag{36}$$

for *all* $(x, g) \in \mathcal{X} \times G$. 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

$$S[\gamma](x)(g) = \frac{1}{(2\pi)^k}\int_{\mathcal{X}_m\times\mathbb{K}} \gamma^\sharp(\xi/\omega,\omega)\sigma^\sharp(\omega)\exp\left(i\langle T_{g^{-1}}[x],\xi\rangle_{\mathcal{X}}\right)|\omega|^{-m}\mathrm{d}\omega\mathrm{d}\xi. \tag{37}$$

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:

$$\gamma^\sharp_{f,\rho}(\xi/\omega,\omega) = \widehat{f}(\xi)(e)\overline{\rho^\sharp(\omega)}. \tag{38}$$

Then,

$$\begin{aligned}
S[\gamma_{f,\rho}](x)(g) &= \frac{1}{(2\pi)^k}\int_{\mathcal{X}_m\times\mathbb{K}} \widehat{f}(\xi)(e)\overline{\rho^\sharp(\omega)}\sigma^\sharp(\omega)\exp\left(i\langle T_{g^{-1}}[x],\xi\rangle_{\mathcal{X}}\right)|\omega|^{-m}\mathrm{d}\omega\mathrm{d}\xi && (39)\\
&= \left((2\pi)^{m-k}\int_{\mathbb{K}}\sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-m}\mathrm{d}\omega\right)\\
&\quad\times\left(\frac{1}{(2\pi)^m}\int_{\mathcal{X}_m}\widehat{f}(\xi)(e)\exp\left(i\langle T_{g^{-1}}[x],\xi\rangle_{\mathcal{X}}\right)\mathrm{d}\xi\right) && (40)\\
&\overset{\star}{=} (\!(\sigma,\rho)\!)f(PT_{g^{-1}}[x])(e) && (41)\\
&= (\!(\sigma,\rho)\!)f(x)(g). && (42)
\end{aligned}$$

where we put

$$(\!(\sigma,\rho)\!) := (2\pi)^{m-k}\int_{\mathbb{K}}\sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-m}\mathrm{d}\omega. \tag{43}$$

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

$$\gamma^\sharp_{f,\rho}(a,\omega) = \widehat{f}(\omega a)(e)\overline{\rho^\sharp(\omega)}, \tag{44}$$

it is calculated as

$$\begin{aligned}
\gamma_{f,\rho}(a,b) &= \frac{1}{(2\pi)^k}\int_{\mathbb{K}}\widehat{f}(\omega a)(e)\overline{\rho^\sharp(\omega)}e^{i\omega b}\mathrm{d}\omega && (45)\\
&= \frac{1}{(2\pi)^k}\int_{\mathbb{K}\times\mathcal{X}_m} f(x)(e)\overline{\rho^\sharp(\omega)}e^{i\omega(b-\langle a,x\rangle_{\mathcal{X}})}\mathrm{d}\omega\mathrm{d}x && (46)\\
&= \int_{\mathcal{X}_m} f(x)(e)\overline{\rho(\langle a,x\rangle_{\mathcal{X}}-b)}\mathrm{d}x, && (47)
\end{aligned}$$

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

A.2 Theorem 2

Proof. Fix arbitrary compact sets $K \subset \mathcal{X}_m$ 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

$$f_n(x)(g) := \sum_{i=1}^n c_i\sigma\left((a_i *_T x)(g) - b_i\right), \quad x \in \mathcal{X}_m,\ g \in G \tag{48}$$

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

$$f_n(T_g[x])(h) = \sum_{i=1}^n c_i\sigma\left(\langle T_{(g^{-1}h)^{-1}}x, a_i\rangle_{\mathcal{X}} - b_i\right) = f_n(x)(g^{-1}h). \tag{49}$$

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 $F_N(x) = \sum_{i=1}^N c_i \sigma(\langle a_i, x\rangle_{\mathcal{X}} - b_i)$ 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

$$f_N(x)(g) = \sum_{i=1}^N c_i \sigma\left(\langle T_{g^{-1}}, a_i\rangle_{\mathcal{X}} - b_i\right) = \sum_{i=1}^N c_i \sigma\left((a_i *_T x)(g) - b_i\right), \tag{50}$$

and it is an ε-neighbour of f because

$$\begin{aligned}
\|f_N - f\|_{C(K;C(L))} &= \sup_{x\in K}\sup_{g\in L} |f_N(x)(g) - f(x)(g)| && (51)\\
&= \sup_{x\in K}\sup_{g\in L} |F_N(T_{g^{-1}}[x]) - f_e(T_{g^{-1}}[x])| && (52)\\
&= \sup_{g\in L}\sup_{x'\in\overline{K}} |F_N(x') - f_e(x')|, \quad x' = T_{g^{-1}}[x] && (53)\\
&< \varepsilon, && (54)
\end{aligned}$$

which concludes the assertion. □

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

1. $\mathcal{X} = \mathcal{X}_m = \mathbb{R}^m$,
2. $f \in C(\mathcal{X};\mathbb{C})$ *(not vector-valued* $C(\mathcal{X};\mathbb{C}^G)$ *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

$$S[R[f_c;\rho]](\boldsymbol{x}) = f_c(\boldsymbol{x}) \text{ at every point } \boldsymbol{x} \in \mathbb{R}^m. \tag{55}$$

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 [53] 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,

$$\int_{\mathbb{R}^m\times\mathbb{R}} \sup_{\boldsymbol{x}\in K} \left|R[f_c;\rho](\boldsymbol{a},b)\sigma(\boldsymbol{a}\cdot\boldsymbol{x} - b)\right|\mathrm{d}\boldsymbol{a}\mathrm{d}b < \infty. \tag{56}$$

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

$$
\begin{aligned}
|R[f_c;\rho](\boldsymbol{a},b)| &= \frac{1}{2\pi}\left|\int_{\mathbb{R}} \widehat{f}_c(\omega\boldsymbol{a})\overline{\rho^\sharp(\omega)}e^{i\omega b}\mathrm{d}\omega\right| \\
&= \frac{1}{2\pi}\left|\int_{\mathbb{R}} \langle\omega\boldsymbol{a}\rangle^s\langle\omega\boldsymbol{a}\rangle^{-s}\langle b\rangle^t\langle b\rangle^{-t}\widehat{f}_c(\omega\boldsymbol{a})\overline{\rho^\sharp(\omega)}e^{i\omega b}\mathrm{d}\omega\right| \\
&\leqslant \frac{1}{2\pi}\left|\int_{\mathbb{R}} \langle\omega\boldsymbol{a}\rangle^s\widehat{f}_c(\omega\boldsymbol{a})\langle\omega\rangle^{-s}\overline{\rho^\sharp(\omega)}\langle\triangle_\omega\rangle^t e^{i\omega b}\mathrm{d}\omega\right|\langle\boldsymbol{a}\rangle^{-s}\langle b\rangle^{-t}, \qquad (57)
\end{aligned}
$$

which asserts the integrability as below

$$
\int_{\mathbb{R}^m\times\mathbb{R}} \sup_{\boldsymbol{x}\in K}|R[f_c;\rho](\boldsymbol{a},b)\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)|\mathrm{d}\boldsymbol{a}\mathrm{d}b \lesssim \int_{\mathbb{R}^m\times\mathbb{R}}\langle\boldsymbol{a}\rangle^{-s}\langle b\rangle^{-t}\mathrm{d}\boldsymbol{a}\mathrm{d}b < \infty. \qquad (58)
$$

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

$$
f_Q(\boldsymbol{x}) := \int_Q R[f_c;\rho](\boldsymbol{a},b)\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)\mathrm{d}\boldsymbol{a}\mathrm{d}b, \qquad (59)
$$

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

$$
f_n(\boldsymbol{x}) := \sum_{i\in I_n} c_{ni}\sigma(\boldsymbol{a}_{ni}\cdot\boldsymbol{x}-b_{ni}). \qquad (60)
$$

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

$$
\begin{aligned}
&\sup_{\boldsymbol{x}\in K}\Big|R[f_c;\rho](\boldsymbol{a},b)\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b) - w_{ni}\sigma(\boldsymbol{a}_{ni}\cdot\boldsymbol{x}-b_{ni})\Big| \qquad (61)\\
&\leqslant \sup_{\boldsymbol{x}\in K}\Big|R[f_c;\rho](\boldsymbol{a},b)\Big|\Big|\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)-\sigma(\boldsymbol{a}_{ni}\cdot\boldsymbol{x}-b_{ni})\Big| \\
&\qquad + \Big|R[f_c;\rho](\boldsymbol{a}_{ni},b_{ni}) - R[f_c;\rho](\boldsymbol{a},b)\Big|\Big|\sigma(\boldsymbol{a}_{ni}\cdot\boldsymbol{x}-b_{ni})\Big| \qquad (62)\\
&\leqslant \|R[f_c;\rho]\|_\infty\,\mathrm{Lip}(\sigma)\sup_{\boldsymbol{x}\in K}\Big|(\boldsymbol{a}-\boldsymbol{a}_{ni})\cdot\boldsymbol{x}-(b-b_{ni})\Big| \\
&\qquad + \mathrm{Lip}(R[f_c;\rho])\Big|(\boldsymbol{a}-\boldsymbol{a}_{ni},b-b_{ni})\Big|\|\sigma\|_{L^\infty(\mathbb{R})} = O(\delta/n)\to 0 \quad n\to\infty. \qquad (63)
\end{aligned}
$$

Besides, the integrands are uniformly bounded as

$$
\sup_{\boldsymbol{x}\in K}\Big|w_{ni}\sigma(\boldsymbol{a}_{ni}\cdot\boldsymbol{x}-b_{ni})\Big| \leqslant \Big|R[f_c;\rho](\boldsymbol{a},b)\Big|\|\sigma\|_{L^\infty(\mathbb{R})}, \text{ for a.e. } (\boldsymbol{a},b)\in Q_{ni}. \qquad (64)
$$

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

$$
\begin{aligned}
\|f_Q - f_n\|_{C(K)} &= \sup_{\boldsymbol{x}\in K}\left|\sum_{i\in I_n}\int_{Q_{ni}} R[f_c;\rho](\boldsymbol{a},b)\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b)\mathrm{d}\boldsymbol{a}\mathrm{d}b - \sum_{i\in I_n} c_{ni}\sigma(\boldsymbol{a}_{ni}\cdot\boldsymbol{x}-b_{ni})\right| \qquad (65)\\
&\leqslant \sum_{i\in I_n}\int_{Q_{ni}}\sup_{\boldsymbol{x}\in K}\Big|R[f_c;\rho](\boldsymbol{a},b)\sigma(\boldsymbol{a}\cdot\boldsymbol{x}-b) - w_{ni}\sigma(\boldsymbol{a}_{ni}\cdot\boldsymbol{x}-b_{ni})\Big|\mathrm{d}\boldsymbol{a}\mathrm{d}b \qquad (66)\\
&\to 0, \quad n\to\infty. \qquad (67)
\end{aligned}
$$

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

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

$$
\|f - f_n\|_{C(K)} \leqslant \|f - f_c\|_{C(K)} + \|f_c - f_Q\|_{C(K)} + \|f_Q - f_n\|_{C(K)} < \varepsilon. \qquad (68)
$$

□

Notes. In the proof, we employed a naive discretization based on the regular grids in Q. 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., 64]).