UNIFIED UNIVERSALITY THEOREM FOR DEEP AND SHALLOW JOINT-GROUP-EQUIVARIANT MACHINES

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a constructive universal approximation theorem for learning machines equipped with joint-group-equivariant feature maps, based on the group representation theory. "Constructive" here indicates that the distribution of parameters is given in a closed-form expression known as the ridgelet transform. Joint-group-equivariance encompasses a broad class of feature maps that generalize classical group-equivariance. Notably, this class includes fully-connected networks, which are *not* group-equivariant *but* are joint-group-equivariant. Moreover, our main theorem also unifies the universal approximation theorems for both shallow and deep networks. While the universality of shallow networks has been investigated in a unified manner by the ridgelet transform, the universality of deep networks has been investigated in a case-by-case manner.

1 INTRODUCTION

The proof of a universality theorem contains hints for understanding the internal data processing mechanisms inside learning machines such as neural networks. For example, the first universality theorem*s* for depth-2 neural networks were shown in 1989 with *four* different proofs by Cybenko (1989), Hornik et al. (1989), Funahashi (1989), and Carroll & Dickinson (1989). Among them, Cybenko's proof using Hahn-Banach and Hornik et al.'s proof using Stone-Weierstrass are existential proofs, meaning that it is not clear how to assign the parameters. On the other hand, Funahashi's proof reducing to the Fourier transform and Carroll and Dickinson's proof reducing to the Radon transform are constructive proofs, meaning that it is clear how to assign the parameters. The latter constructive methods, which reduce to integral transforms, were refined as the so-called integral representation by Barron (1993) and further culminated as the *ridgelet transform*, the main objective of this study, discovered by Murata (1996) and Candès (1998).

To show the universality in a constructive manner, we formulate the the problem as a functional equation: Let $\mathtt{LM}[\gamma]$ denote a certain learning machine (such as a deep network) with parameter γ, and let $\mathcal{F}$ denote a class of functions to be expressed by the learning machine. Given a function $f \in \mathcal{F}$, find an unknown parameter γ so that the machine $\mathtt{LM}[\gamma]$ represents function f, i.e.

$$\mathtt{LM}[\gamma] = f, \tag{1}$$

which we call a *learning equation*. This equation is understood as a stronger formulation of learning than an ordinary formulation by the empirical risk minimization such as minimizing $\sum_{i=1}^{n} |\mathtt{LM}[\gamma](x_i) - f(x_i)|^2$ with respect to γ, as the latter is understood as a weak form (or a variational form) of this equation. Therefore, characterizing the solution space of this equation leads to understanding the parameters obtained by risk minimization. Following previous studies (Murata, 1996; Candès, 1998; Sonoda et al., 2021a;b; 2022a;b), we call a solution operator $\mathtt{R}$ that satisfies $\mathtt{LM}[\mathtt{R}[f]] = f$ a *ridgelet transform*. Once such a solution operator $\mathtt{R}$ is found, we can conclude a *universality* of the learning machine in consideration because the reconstruction formula $\mathtt{LM}[\mathtt{R}[f]] = f$ implies for any $f \in \mathcal{F}$ there exists a machine that represents f. In particular, when $\mathtt{R}[f]$ is found in a closed-form manner, then it leads to a *constructive* proof of the universality since $\mathtt{R}[f]$ could indicate how to assign parameters.

For depth-2 neural networks (particularly with an infinitely-wide hidden layer), the equation has been solved with several closed-form ridgelet transforms. For example, the closed-form ridgelet transforms have been obtained for depth-2 fully-connected layers (Sonoda et al., 2021b), depth-2 fully-connected layers on manifolds (Sonoda et al., 2022b), depth-2 group convolution layers (Sonoda et al., 2022a), and depth-2 fully-connected layers on finite fields (Yamasaki et al., 2023). The essential technique to obtain these ridgelet transforms are to construct a Fourier expression corresponding to the network in consideration. We refer to Sonoda et al. (2024b) for more tecnnical backgrounds behind these results. Furthermore, Sonoda et al. (2021a) have revealed that the distribution of parameters inside depth-2 fully-connected networks obtained by regularized empirical risk minimization assymptotically converges to the ridgelet transform. In other words, the ridgelet transform can also explain the solutions obtained by risk minimization.

On the other hand, for depth-n neural networks, the equation is far from solved, and it is common to either consider infinitely-deep mathematical models such as Neural ODEs (Sonoda & Murata, 2017b; E, 2017; Li & Hao, 2018; Haber & Ruthotto, 2017; Chen et al., 2018), or handcraft solutions depending on the network specifications. For example, construction methods such as the so-called Telgarsky sawtooth function (or the Yarotsky scheme) and bit extraction techniques (Cohen et al., 2016; Telgarsky, 2016; Yarotsky, 2017; 2018; Yarotsky & Zhevnerchuk, 2020; Daubechies et al., 2022; Cohen et al., 2022; Siegel, 2023; Petrova & Wojtaszczyk, 2023; Grohs et al., 2023) have been developed (not only to investigate the expressivity but also) to demonstrate the depth separation, super-convergence, and minmax optimality of deep ReLU networks. Various feature maps have also been handcrafted in the contexts of geometric deep learning (Bronstein et al., 2021) and deep narrow networks (Lu et al., 2017; Hanin & Sellke, 2017; Lin & Jegelka, 2018; Kidger & Lyons, 2020; Park et al., 2021; Li et al., 2023; Cai, 2023; Kim et al., 2024). However, for the purpose of understanding the parameters obtained by risk minimization (in a manner presented by Sonoda et al. (2021a)), these results are less satisfactory because there is no guarantee that these handcrafted solutions are obtaiend by risk minimization.

Recently, Sonoda et al. (2024a) developed a novel technique to show the universality based on the group representation theory, and discovered a rich class of ridgelet transforms for learning machines with joint-group-*invariant* feature maps. However, their technique was essentially limited to depth-2 networks and could not cover depth-n networks defined by composites of nonlinear activation functions such as $\sigma(A_2\sigma(A_1\boldsymbol{x} - \boldsymbol{b}_1) - \boldsymbol{b}_2)$. By carefully reviewing their group theoretic arguments, we found that the joint-*invariance* is the bottleneck and it can be resolved by relaxing the assumption to the joint-*equivariance*. In this study, we present a wider class of ridgelet transforms for learning machines with joint-*equivariant* feature maps so to cover the depth-n (as well as depth-2) fully-connected networks.

The contributions of this study include

- We derived the ridgelet transform (solution operator for the learning equation) for learning machines with joint-group-equivariant feature maps. Since the solution of the learning equation can be written in closed form for any $f \in \mathcal{F}$, it is a constructive and unified proof of the universal approximation theorem for joint-group-equivariant machines.
- As a corollary, we have shown the constructive universal approximation property of deep fully-connected neural networks. Until this study, the universality of deep networks has been shown in a different manner from the universality of shallow networks, but our results discuss them on common ground. Now we can understand the approximation schemes of various learning machines in a unified manner.
- In addition, as an example of a learning machine whose universality has not been known, we presented a network with quadratic forms and showed its universality.

2 PRELIMINARIES

We quickly overview the original integral representation and the ridgelet transform, a mathematical model of depth-2 fully-connected network and its right inverse. Then, we list a few facts in the group representation theory. In particular, *Schur's lemma* and the *Haar measure* play key roles in the proof of the main results.

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

2.1 Integral Representation and Ridgelet Transform for Depth-2 Fully-Connected Network

Definition 1. For any measurable functions $\sigma : \mathbb{R} \to \mathbb{C}$ and $\gamma : \mathbb{R}^m \times \mathbb{R} \to \mathbb{C}$, put

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

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

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

$$S_\sigma[\gamma_p](\boldsymbol{x}) = \sum_{i=1}^p c_i \sigma(\boldsymbol{a}_i \cdot \boldsymbol{x} - b_i). \tag{3}$$

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

Next, we introduce the ridgelet transform, which is known to be a right-inverse operator to S_σ.

Definition 2. For any measurable functions $\rho : \mathbb{R} \to \mathbb{C}$ and $f : \mathbb{R}^m \to \mathbb{C}$, put

$$R_\rho[f](\boldsymbol{a}, b) := \int_{\mathbb{R}^m} f(\boldsymbol{x})\overline{\rho(\boldsymbol{a} \cdot \boldsymbol{x} - b)}\mathrm{d}\boldsymbol{x}, \quad (\boldsymbol{a}, b) \in \mathbb{R}^m \times \mathbb{R}. \tag{4}$$

We call R_ρ a ridgelet transform.

To be precise, it satisfies the following reconstruction formula.

Theorem 1 (Reconstruction Formula)**.** *Suppose σ and ρ are a tempered distribution ($\mathcal{S}'$) and a rapid decreasing function ($\mathcal{S}$) respectively. There exists a bilinear form $(\!(\sigma, \rho)\!)$ such that*

$$S_\sigma \circ R_\rho[f] = (\!(\sigma, \rho)\!) f, \tag{5}$$

for any square integrable function $f \in L^2(\mathbb{R}^m)$. Further, the bilinear form is given by $(\!(\sigma, \rho)\!) = \int_{\mathbb{R}} \sigma^\sharp(\omega)\overline{\rho^\sharp(\omega)}|\omega|^{-m}\mathrm{d}\omega$, where $\sharp$ denotes the 1-dimensional Fourier transform.

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

The original ridgelet transform was discovered by Murata (1996) and Candès (1998). It is recently extended to a few modern networks by the Fourier slice method (see e.g. Sonoda et al., 2024b). In this study, we present a systematic scheme to find the ridgelet transform for a variety of given network architecture based on the group theoretic arguments.

2.2 Irreducible Unitary Representation and Schur's Lemma

In the main theorem, we use *Schur's lemma*, a fundamental theorem from unitary group representation theory. Group representation is a method for investigating properties of an abstract group G by mapping G to another (much computable) group of invertible linear operators. We refer to Folland (2015) for more details on group representation and harmonic analysis on groups.

In this study, we assume group G to be *locally compact*. This is a sufficient condition for having invariant measures. It is not a strong assumption. For example, any finite group, discrete group, compact group, and finite-dimensional Lie group are locally compact, while an infinite-dimensional Lie group is *not* locally compact.

Let $\mathcal{H}$ be a nonzero Hilbert space, and $\mathcal{U}(\mathcal{H})$ be the group of unitary operators on $\mathcal{H}$. A *unitary representation* π of G on $\mathcal{H}$ is a group homomorphism that is continuous with respect to the strong operator topology—that is, a map $\pi : G \to \mathcal{U}(\mathcal{H})$ satisfying $\pi_{gh} = \pi_g \pi_h$ and $\pi_{g^{-1}} = \pi_g^{-1}$, and for any $\psi \in \mathcal{H}$, the map $G \ni g \mapsto \pi_g[\psi] \in \mathcal{H}$ is continuous.

Suppose $\mathcal{M}$ is a closed subspace of $\mathcal{H}$. $\mathcal{M}$ is called an *invariant* subspace when $\pi_g[\mathcal{M}] \subset \mathcal{M}$ for all $g \in G$. Particularly, π is called *irreducible* when it does not admit any nontrivial invariant subspace $\mathcal{M} \neq \{0\}$ nor $\mathcal{H}$. The following theorem is a fundamental result of group representation theory that characterizes the irreducibility.

Theorem 2 (Schur's lemma)**.** *A unitary representation $\pi : G \to \mathcal{U}(\mathcal{H})$ is irreducible iff any bounded operator T on $\mathcal{H}$ that commutes with π is always a constant multiple of the identity. In other words, if $\pi_g \circ T = T \circ \pi_g$ for all $g \in G$, then $T = c \,\mathrm{Id}_{\mathcal{H}}$ for some $c \in \mathbb{C}$.*

See Folland (2015, Theorem 3.5(a)) for the proof. We use this as a key step in the proof of our main theorem.

As a concrete example of an irreducible representation, we use the following regular representation of the affine group $\mathrm{Aff}(m)$ on $L^2(\mathbb{R}^m)$.

Theorem 3. *Let $G := \mathrm{Aff}(m) := GL(m) \ltimes \mathbb{R}^m$ be the affine group acting on $X = \mathbb{R}^m$ by $(L, \boldsymbol{t}) \cdot \boldsymbol{x} = L\boldsymbol{x} + \boldsymbol{t}$, and let $\mathcal{H} := L^2(\mathbb{R}^m)$ be the Hilbert space of square-integrable functions. Let $\pi : \mathrm{Aff}(m) \to \mathcal{U}(L^2(\mathbb{R}^m))$ be the regular representation of the affine group $\mathrm{Aff}(m)$ on $L^2(\mathbb{R}^m)$, namely $\pi_g[f](\boldsymbol{x}) := |\det L|^{-1/2} f(L^{-1}(\boldsymbol{x} - \boldsymbol{t}))$ for any $g = (L, \boldsymbol{t}) \in G$. Then π is irreducible.*

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

2.3 Calculus on Locally Compact Group

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

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

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

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

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

3 Main Results

We introduce unitary representations π and $\widehat{\pi}$, a *joint-equivariant feature map* $\phi : X \times \Xi \to Y$, a *joint-equivariant machine* $\mathtt{LM}[\gamma; \phi] : X \to Y$, and present the ridgelet transform $\mathtt{R}[f; \psi] : \Xi \to \mathbb{C}$ for joint-equivariant machines, yielding the universality $\mathtt{LM}[\mathtt{R}[f; \psi]; \phi] = c_{\phi,\psi} f$. We note that π plays a key role in the main theorem, and the joint-equivariance is an essential property of depth-n fully-connected network.

Let G be a locally compact group equipped with a left invariant measure $\mathrm{d}g$. Let X and Ξ be G-spaces equipped with G-invariant measures $\mathrm{d}x$ and $\mathrm{d}\xi$, called the *data domain* and the *parameter domain*. respectively. Let Y be a separable Hilbert space, called the *output domain*. Let $\mathcal{U}(Y)$ be the space of unitary operators on Y, and let $\upsilon : G \to \mathcal{U}(Y)$ be a unitary representation of G on Y. We call a Y-valued map ϕ on the data-parameter domain $X \times \Xi$, i.e. $\phi : X \times \Xi \to Y$, a *feature map*.

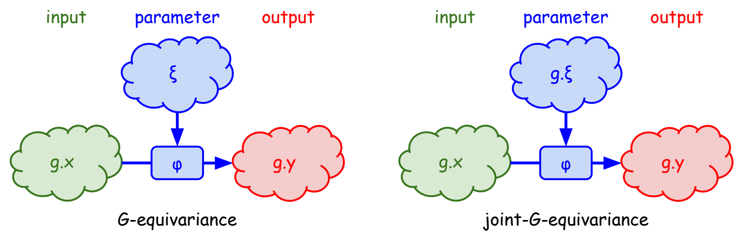

Figure 1: An ordinary G-equivariant feature map $\phi : X \times \Xi \to Y$ is a subclass of joint-G-equivariant map where the G-action on parameter domain Ξ is *trivial*, i.e. $g \cdot \xi = \xi$

Let $L^2(X;Y)$ denote the space of Y-valued square-integrable functions on X equipped with the inner product $\langle \phi, \psi \rangle_{L^2(X;Y)} := \int_X \langle \phi(x), \psi(x) \rangle_Y \mathrm{d}x$; and let $L^2(\Xi)$ denote the space of $\mathbb{C}$-valued square-integrable functions on Ξ.

If there is no risk of confusion, we use the same symbol $\cdot$ for the G-actions on X, Y, and Ξ (e.g., $g \cdot x$, $g \cdot y$, and $g \cdot \xi$). On the other hand, to avoid the confusion between G-actions on output domain Y and Y-valued function $f : X \to Y$, both "$g \cdot f(x)$" and "$\upsilon_g[f(x)]$" (if needed) always imply G-action on Y, and "$\pi_g[f](x)$" (introduced soon below) for G-actions on $f : X \to Y$.

Additionally, we introduce two unitary representations π and $\widehat{\pi}$ of G on function spaces $L^2(X;Y)$ and $L^2(\Xi)$ as follows: For each $g \in G$, $f \in L^2(X;Y)$ and $\gamma \in L^2(\Xi)$,

$$\pi_g[f](x) := \upsilon_g[f(g^{-1} \cdot x)] = g \cdot f(g^{-1} \cdot x), \quad x \in X \tag{6}$$

$$\widehat{\pi}_g[\gamma](\xi) := \gamma(g^{-1} \cdot \xi), \quad \xi \in \Xi. \tag{7}$$

In the main theorem, the irreducibility of π will be a sufficient condition for the universality. On the other hand, the irreducibility of $\widehat{\pi}$ is not necessary. For those who are less familiar with group representations, we have shown that π and $\widehat{\pi}$ are unitary representations in Lemmas 6 and 7.

3.1 Joint-Equivariant Feature Map

We introduce the joint-group-equivariant feature map, extending the classical notion of group-equivariant feature maps. The major motivation to introduce this is that depth-n fully-connected networks, the main subject of this study, are not equivariant but joint-equivariant.

Definition 3 (Joint-G-Equivariant Feature Map)**.** We say a feature map $\phi : X \times \Xi \to Y$ is *joint-G-equivariant* when

$$\phi(g \cdot x, g \cdot \xi) = g \cdot \phi(x, \xi), \quad (x, \xi) \in X \times \Xi, \tag{8}$$

holds for all $g \in G$. Especially, when G-action on Y is trivial, i.e. $\phi(g \cdot x, g \cdot \xi) = \phi(x, \xi)$, we say it is *joint-$G$-invariant*.

Remark 1 (Relation to classical G-equivariance). The joint-G-equivariance is not a restriction but an extension of the classical notion of *G-equivariance*, i.e. $\phi(g \cdot x, \xi) = g \cdot \phi(x, \xi)$. In fact, G-equivariance is a special case of joint-G-equivariance where G acts trivially on parameter domain, i.e. $g \cdot \xi = \xi$ (see also Figure 1). Thus, all G-equivariant maps are automatically joint-G-equivariant.

Remark 2 (Interpretations of joint-G-equivariance). We have two interpretations from algebraic and geometric perspectives. First, from an algebraic perspective, ϕ is a homomorphism (or a G-map) between G-sets from $X \times \Xi$ to Y. So we may denote the collection of all joint-G-equivariant maps as $\hom_G(X \times \Xi, Y)$. Second, from a more geometric perspective, ϕ is a vector-field $\Xi \to Y^X$ with structure group G acting on fiber Y^X by π. Here we identify $\phi : X \times \Xi \to Y$ with $\phi_c : \Xi \to Y^X$ by the so-called *currying* $\phi_c(\xi) := \phi(\bullet, \xi)$. In other words, ϕ_c is a global section of a trivial G-bundle $\Xi \times Y^X \to \Xi$. Consequently, we can understand the G-action on parameter domain (or base space) Ξ is induced from the change-of-basis π so that the section ϕ_c to be G-equivariant

$$\pi_g[\phi_c(\xi)](x) = g \cdot \phi(g^{-1} \cdot x, \xi) =: \phi_c(g \cdot \xi)(x). \tag{9}$$

Finally, two aspects are unified as an adjunction: $\hom_G(X \times \Xi, Y) \cong \hom_G(\Xi, Y^X)$.

In the following, we list several construction methods of joint-equivariant maps in Lemmas 1, 2 and 3 (in the next subsection), indicating the richness of the proposed concept. Whereas to construct a (non-joint) G-equivariant network, we must carefully and precisely design the network architecture (see, e.g., a textbook of geometric deep learning Bronstein et al., 2021), to construct a joint-G-equivariant network, we can easily and systematically obtain the one.

First, we can synthesize a joint-equivariant map from (not equivariant but) *any* map $\phi_0 : X \to Y$.

Lemma 1. *Let X and Y be G-sets. Fix an arbitrary map $\phi_0 : X \to Y$, and put $\phi(x, g) := \pi_g[\phi_0](x) = g \cdot \phi_0(g^{-1} \cdot x)$ for every $x \in X$ and $g \in G$. Then, $\phi : X \times G \to Y$ is joint-G-equivariant.*

Proof. For any $g, h \in G$, we have $\phi(g \cdot x, g \cdot h) = (gh) \cdot \phi_0((gh)^{-1} \cdot (g \cdot x)) = g \cdot \phi(x, h)$. □

In general, a G-set is understood as a representation of G. So, the case of $X = Y = \Xi = G$ with $\phi : G \times G \to G$ is understood as a primitive type of joint-G-equivariant maps $\phi : X \times \Xi \to Y$.

The next lemma suggests the compatibility with function compositions, or deep structures.

Lemma 2 (Depth-n Joint-Equivariant Feature Map $\phi_{1:n}$)**.** *Given a sequence of joint-G-equivariant feature maps $\phi_i : X_{i-1} \times \Xi_i \to X_i$ $(i = 1, \ldots, n)$, let $\Xi_{1:n} := \Xi_1 \times \cdots \times \Xi_n$ be the n-fold parameter space with the component-wise G-action $g \cdot \xi_{1:n} := (g \cdot \xi_1, \ldots, g \cdot \xi_n)$ for each n-fold parameters $\xi_{1:n} \in \Xi_{1:n}$, and let $\phi_{1:n} : X_0 \times \Xi_{1:n} \to X_n$ be the depth-n feature map given by*

$$\phi_{1:n}(x, \xi_{1:n}) := \phi_n(\bullet, \xi_n) \circ \cdots \circ \phi_1(x, \xi_1). \tag{10}$$

Then, $\phi_{1:n}$ is joint-G-equivariant.

In other words, the composition of joint-equivariant maps defines a cascade product of morphisms: $\hom_G(\Xi_2, X_2^{X_1}) \times \hom_G(\Xi_1, X_1^{X_0}) \to \hom_G(\Xi_1 \times \Xi_2, X_2^{X_0})$. See Appendix A.2 for the proof.

3.2 Joint-Equivariant Machine and Ridgelet Transform

We further introduce the joint-equivariant machine, extending the integral representation.

Definition 4 (Joint-Equivariant Machine)**.** Fix an arbitrary joint-equivariant feature map $\phi : X \times \Xi \to Y$. For any scalar-valued measurable function $\gamma : \Xi \to \mathbb{C}$, define a Y-valued map on X by

$$\mathtt{LM}[\gamma; \phi](x) := \int_\Xi \gamma(\xi)\phi(x, \xi)\mathrm{d}\xi, \quad x \in X, \tag{11}$$

where the integral is understood as the Bochner integral. We also write $\mathtt{LM}_\phi := \mathtt{LM}[\bullet; \phi]$ for short. If needed, we call the image $\mathtt{LM}[\gamma; \phi] : X \to Y$ a joint-equivariant *machine*, and the integral transform $\mathtt{LM}[\bullet; \phi]$ of γ a joint-equivariant *transform*.

The joint-equivariant machine extends the original integral representation. It inherits the concept of integrating all the possible parameters ξ and indirectly select which parameters to use by weighting on them, which *linearize* parametrization by lifting nonlinear parameters ξ to linear parameter γ.

Recall that the G-action on parameter domain Ξ is also linearized by lifting it to $\widehat{\pi}$ on $L^2(\Xi)$. The joint-equivariance of $\phi : \Xi \to Y^X$ is inherited under the linearization to $\mathtt{LM}_\phi : L^2(\Xi) \to L^2(X; Y)$.

Lemma 3. *A joint-G-equivariant machine $\mathtt{LM}_\phi : L^2(\Xi) \to L^2(X; Y)$ is joint-G-equivariant, i.e. $\mathtt{LM}_\phi \in \hom_G(L^2(\Xi), L^2(X; Y))$.*

Proof. $\mathtt{LM}_\phi[\widehat{\pi}_g[\gamma]](g \cdot x) = \int_\Xi \gamma(g^{-1} \cdot \xi)\phi(g \cdot x, \xi)\mathrm{d}\xi = \int_\Xi \gamma(\xi)\phi(g \cdot x, g \cdot \xi)\mathrm{d}\xi = g \cdot \mathtt{LM}_\phi[\gamma](x)$. □

Definition 5 (Ridgelet Transform for Joint-Equivariant Machine)**.** For any joint-equivariant feature map $\psi : X \times \Xi \to Y$ and Y-valued Borel measurable function f on X, put a scalar-valued map by

$$\mathtt{R}[f; \psi](\xi) := \int_X \langle f(x), \psi(x, \xi)\rangle_Y \mathrm{d}x, \quad \xi \in \Xi. \tag{12}$$

We also write $\mathtt{R}_\psi := \mathtt{R}[\bullet; \psi]$ for short. If there is no risk of confusion, we call both the image $\mathtt{R}[f; \psi] : X \to Y$ and the integral transform $\mathtt{R}[\bullet; \psi]$ of f a ridgelet transform.

Intuitively, it measures the similarity between target function f and feature $\psi(\bullet,\xi)$ at ξ. As long as the integrals are convergent, the ridgelet transform is the dual operator of the joint-equivariant transform (with common ϕ):

$$\langle \gamma, \mathtt{R}[f;\phi]\rangle_{L^2(\Xi)} = \int_{X\times\Xi} \gamma(\xi)\langle \phi(x,\xi), f(x)\rangle_Y \mathrm{d}x\mathrm{d}\xi = \langle \mathtt{LM}[\gamma;\phi], f\rangle_{L^2(X;Y)}. \tag{13}$$

Similarly to the joint-equivariant machine, the ridgelet transform is again joint-G-*invariant*. In fact,

$$\mathtt{R}_\psi[\pi_g[f]](g\cdot\xi) = \int_X \langle \upsilon_g[f(g^{-1}\cdot x)], \psi(x, g\cdot\xi)\rangle_Y \mathrm{d}\xi = \int_X \langle f(x), \upsilon_g^*[\psi(g\cdot x, g\cdot\xi)]\rangle_Y \mathrm{d}\xi = \mathtt{R}_\psi[f](\xi).$$

Hence, geometrically, if we regard $\mathtt{LM}_\phi : L^2(\Xi) \to L^2(X;Y)$ a vector field of trivial G-bundle $L^2(\Xi)\times L^2(X;Y) \to L^2(\Xi)$, then $\mathtt{R}_\phi : L^2(X;Y) \to L^2(\Xi)$ corresponds to a G-connection.

3.3 Main Theorem

At last, we state the main theorem, that is, the reconstruction formula for joint-equivariant machines.

Theorem 4 (Reconstruction Formula)**.** *Assume (1) feature maps $\phi, \psi : X\times\Xi \to Y$ are joint-G-equivariant, (2) composite operator $\mathtt{LM}_\phi \circ \mathtt{R}_\psi : L^2(X;Y) \to L^2(X;Y)$ is bounded (i.e., Lipschitz continuous), and (3) the unitary representation $\pi : G \to \mathcal{U}(L^2(X;Y))$ defined in (6) is irreducible. Then, there exists a bilinear form $(\!(\phi,\psi)\!) \in \mathbb{C}$ (independent of f) such that for any Y-valued square-integrable function $f \in L^2(X;Y)$,*

$$\mathtt{LM}_\phi \circ \mathtt{R}_\psi[f] = \int_\Xi \left[\int_X \langle f(x), \psi(x,\xi)\rangle_Y \mathrm{d}x\right] \phi(\bullet,\xi)\mathrm{d}\xi = (\!(\phi,\psi)\!) f. \tag{14}$$

In practice, once the irreducibility of G-action π on $L^2(X;Y)$ is verified, the ridgelet transform $\mathtt{R}_\psi$ becomes a right inverse operator of joint-equivariant transform $\mathtt{LM}_\phi$ as long as $(\!(\phi,\psi)\!) \neq 0, \infty$. The proof is given in Appendix A.3. Despite the wide coverage of examples, the proof is brief.

Remark 3. (1) When π is not irreducible (thus reducible) and admits an irreducible decomposition $L^2(X;Y) = \bigoplus_{i=1}^\infty \mathcal{H}_i$, reconstruction formula (14) holds for $f \in \mathcal{H}_k$ for some k. (2) The irreducibility is assumed only for π, and not for $\widehat{\pi}$. This asymmetry originates from the fact that our main theorem focuses only on (the universality of) $\mathtt{LM}_\phi : X \to Y$, not on its dual $\mathtt{R}_\psi : \Gamma \to \mathbb{R}$. However, when $\widehat{\pi}$ is irreducible, then we can state $\mathtt{R}_\psi \circ \mathtt{LM}_\phi[\gamma] = \gamma$ for any $\gamma \in L^2(\Xi)$ (the order of composition is reverted from $\mathtt{LM}_\phi \circ \mathtt{R}_\psi$). (3) The regularity of feature maps ϕ,ψ needs to be studied in a case-by-case manner. In fact, in the examples of fully-connected networks (§ 5) and quadratic-form networks (§ 6), the joint-equivariance holds for any activation function. However, the constant $(\!(\phi,\psi)\!)$ could degenerate to zero or diverge when feature maps are inappropriate. For the case of depth-2 fully-connected networks, it is known that the constant is zero if and only if the activation function is a polynomial function (see e.g., Sonoda & Murata, 2017a). To this date, however, such a condition is shown in a case-by-case manner, and the general principle is not known.

4 Example: Depth-n Joint-Equivariant Machine

As pointed out in Lemma 2, the depth-n feature map $\phi_{1:n}$ is joint-G-equivariant. Therefore, the following Y-valued depth-n joint-equivariant machine $\mathtt{DLM}[\gamma;\phi_{1:n}]$ is $L^2(X;Y)$-universal.

Corollary 1 (Deep Ridgelet Transform)**.** *For any maps $\gamma : X \to \mathbb{C}$ and $f \in L^2(X;Y)$, put*

$$\mathtt{DLM}[\gamma;\phi_{1:n}](x) := \int_{\Xi_1\times\cdots\times\Xi_n} \gamma(\xi_1,\ldots,\xi_n)\phi_n(\bullet,\xi_n)\circ\cdots\circ\phi_1(x,\xi_1)\mathrm{d}\boldsymbol{\xi}, \quad x\in X, \tag{15}$$

$$\mathtt{R}[f;\psi_{1:n}](\boldsymbol{\xi}) := \int_\Xi \langle f(x), \psi_n(\bullet,\xi_n)\circ\cdots\circ\psi_1(x,\xi_n)\rangle_Y \mathrm{d}x, \quad \boldsymbol{\xi}\in\Xi_1\times\cdots\times\Xi_n. \tag{16}$$

Under the assumptions that $\mathtt{DLM}_{\phi_{1:n}} \circ \mathtt{R}_{\psi_{1:n}}$ is bounded, and that π is irreducible, there exists a bilinear form $(\!(\phi_{1:n},\psi_{1:n})\!)$ satisfying $\mathtt{DLM}_{\phi_{1:n}} \circ \mathtt{R}_{\psi_{1:n}} = (\!(\phi_{1:n},\psi_{1:n})\!)\, \mathrm{Id}_{L^2(X;Y)}$.

Again, it extends the original integral representation, and inherits the *linearization* trick of nonlinear parameters $\boldsymbol{\xi}$ by integrating all the possible parameters (beyond the difference of layers) and indirectly select which parameters to use by weighting on them.

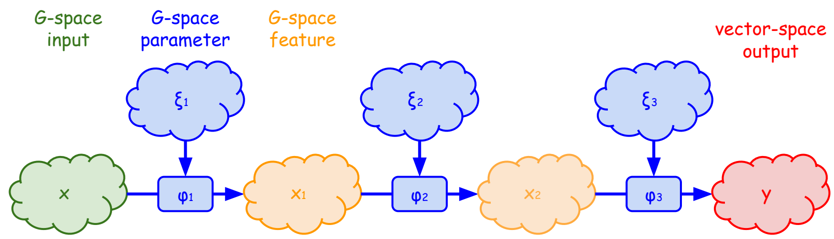

Figure 2: Deep Y-valued joint-G-equivariant machine on G-space X is $L^2(X;Y)$-universal when unitary representation π of G on $L^2(X;Y)$ is irreducible, and the distribution of parameters for the machine to represent a given map $f : X \to Y$ is exactly given by the ridgelet transform $\mathtt{R}[f]$

5 Example: Depth-n Fully-Connected Network

We explain the case of depth-n (precisely, depth-$n+1$) fully-connected network. We use the following fact.

Lemma 4 (Folland (2015, Theorem 7.12))**.** *Let π_1 and π_2 be representations of locally compact groups G_1 and G_2, and let $\pi_1 \otimes \pi_2$ be their outer tensor product, which is a representation of the product group $G_1 \times G_2$. Then, π_1 and π_2 are irreducible if and only if $\pi_1 \otimes \pi_2$ is irreducible.*

Set $X = Y = \mathbb{R}^m$ (input and output domains), and for each $i \in \{1, \ldots, n\}$, set $\Xi_i := \mathbb{R}^{p_i \times d_i} \times \mathbb{R}^{p_i} \times \mathbb{R}^{d_{i+1} \times q_i}$ (parameter domain), $\sigma_i : \mathbb{R}^{p_i} \to \mathbb{R}^{q_i}$ (activation functions), and define the feature map (vector-valued fully-connected neurons) as

$$\phi_i(\boldsymbol{x}_i, \boldsymbol{\xi}_i) := C_i \sigma_i(A_i \boldsymbol{x}_i - \boldsymbol{b}_i), \quad \boldsymbol{x}_i \in \mathbb{R}^{d_i}, \boldsymbol{\xi}_i = (A_i, \boldsymbol{b}_i, C_i) \in \Xi_i \tag{17}$$

Specifically, $d_1 = d_{n+1} = m$. If there is no risk of confusion, we omit writing i for simplicity.

Let $O(m)$ denote the orthogonal group in dimension m. Let $G := O(m) \times \mathrm{Aff}(m)$ be the product group of $O(m)$ and $\mathrm{Aff}(m) = GL(m) \ltimes \mathbb{R}^m$. We suppose G acts on the input and output domains as below: For any $g = (Q, L, \boldsymbol{t}) \in G = O(m) \times (GL(m) \ltimes \mathbb{R}^m)$,

$$g \cdot \boldsymbol{x} := L\boldsymbol{x} + \boldsymbol{t}, \ \boldsymbol{x} \in X, \quad \text{and} \quad g \cdot \boldsymbol{y} := \upsilon_g[\boldsymbol{y}] := Q\boldsymbol{y}, \ \boldsymbol{y} \in Y. \tag{18}$$

Namely, the group actions of both $O(m)$ on X and $\mathrm{Aff}(m)$ on Y are trivial.

Let π be the induced representation of G on the vector-valued square-integrable functions $L^2(X;Y)$, defined by

$$\pi_g[\boldsymbol{f}](\boldsymbol{x}) := |\det L|^{-1/2} Q \boldsymbol{f}(L^{-1}(\boldsymbol{x} - \boldsymbol{t})), \quad \boldsymbol{x} \in X, \ \boldsymbol{f} \in L^2(X;Y) \tag{19}$$

for each $g = (Q, L, \boldsymbol{t}) \in O(m) \times (GL(m) \ltimes \mathbb{R}^m)$.

Lemma 5. *The above $\pi : G \to \mathcal{U}(L^2(\mathbb{R}^m; \mathbb{R}^m))$ is irreducible.*

Proof. Recall the representations of $O(m)$ on $\mathbb{R}^m$ and of $\mathrm{Aff}(m)$ on $L^2(\mathbb{R}^m)$ are respectively irreducible (see Theorem 3), and $L^2(\mathbb{R}^m; \mathbb{R}^m)$ is equivalent to the tensor product $\mathbb{R}^m \otimes L^2(\mathbb{R}^m)$. Hence by Lemma 4, the representation π of the product group $O(m) \times \mathrm{Aff}(m)$ on the tensor product $\mathbb{R}^m \otimes L^2(\mathbb{R}^m) = L^2(\mathbb{R}^m; \mathbb{R}^m)$ is irreducible. $\square$

Additionally, we put the dual action of G on parameter domain Ξ_i as below:

$$g \cdot (A_i, \boldsymbol{b}_i, C_i) := \begin{cases} (A_i L^{-1}, \boldsymbol{b}_i + A_i L^{-1} \boldsymbol{t}, C_i), & i = 1 \\ (A_i, \boldsymbol{b}_i, C_i), & i \neq 1, n \\ (A_i, \boldsymbol{b}_i, QC_i), & i = n \end{cases} \tag{20}$$

for all $g = (Q, L, \boldsymbol{t}) \in O(m) \times (GL(m) \ltimes \mathbb{R}^m)$, $(A_i, \boldsymbol{b}_i, C_i) \in \Xi_i$.

Then, the composition of feature maps $\phi_{1:n}(\boldsymbol{x}, \boldsymbol{\xi}_{1:n}) := \phi_n(\bullet, \boldsymbol{\xi}_n) \circ \cdots \circ \phi_1(\boldsymbol{x}, \boldsymbol{\xi}_1)$ is joint-G-equivariant. In fact,

$$\begin{aligned}
\phi_1(g \cdot \boldsymbol{x}, g \cdot \boldsymbol{\xi}_1) &= C_1 \sigma\left(A_1 L^{-1}(L\boldsymbol{x} + \boldsymbol{t}) - (\boldsymbol{b}_1 + A_1 L^{-1}\boldsymbol{t})\right) = C_1\sigma(A_1\boldsymbol{x} - \boldsymbol{b}_1) = \phi_1(\boldsymbol{x}, \boldsymbol{\xi}_1),\\
\phi_i(\boldsymbol{x}, g \cdot \boldsymbol{\xi}_i) &= C_i\sigma(A_i\boldsymbol{x} - \boldsymbol{b}_i) = \phi_i(\boldsymbol{x}, \boldsymbol{\xi}_i), \quad i \neq 1, n\\
\phi_n(\boldsymbol{x}, g \cdot \boldsymbol{\xi}_n) &= QC_n\sigma(A_n\boldsymbol{x} - \boldsymbol{b}_n) = g \cdot \phi_n(\boldsymbol{x}, \boldsymbol{\xi}_n),
\end{aligned}$$

Therefore $\phi_{1:n}(g \cdot \boldsymbol{x}, g \cdot \boldsymbol{\xi}_{1:n}) = g \cdot \phi_{1:n}(\boldsymbol{x}, \boldsymbol{\xi}_{1:n})$.

So by putting depth-n neural network and the corresponding ridgelet transform as below

$$\mathtt{DNN}[\gamma; \phi_{1:n}](\boldsymbol{x}) = \int_{\Xi_{1:n}} \gamma(\boldsymbol{\xi}_{1:n})\phi_{1:n}(\boldsymbol{x}, \boldsymbol{\xi}_{1:n})\mathrm{d}\boldsymbol{\xi}_{1:n}, \tag{21}$$

$$\mathtt{R}[\boldsymbol{f}; \psi_{1:n}](\boldsymbol{a}, b) = \int_{\mathbb{R}^m} \boldsymbol{f}(\boldsymbol{x}) \cdot \overline{\psi_{1:n}(\boldsymbol{x}, \boldsymbol{\xi}_{1:n})}\mathrm{d}\boldsymbol{x}, \tag{22}$$

Theorem 4 yields the reconstruction formula $\mathtt{DNN}_{\phi_{1:n}} \circ \mathtt{R}_{\psi_{1:n}} = (\!(\phi_{1:n}, \psi_{1:n})\!) \, \mathrm{Id}_{L^2(\mathbb{R}^m;\mathbb{R}^m)}$.

6 Example: Quadratic-form with Nonlinearity

Here, we present a new network for which the universality was not known.

Let M denote the class of all $m \times m$-symmetric matrices equipped with the Lebesgue measure $\mathrm{d}A = \bigwedge_{i \geq j} \mathrm{d}a_{ij}$. Set $X = \mathbb{R}^m$, $\Xi = M \times \mathbb{R}^m \times \mathbb{R}$, and

$$\phi(\boldsymbol{x}, \xi) := \sigma(\boldsymbol{x}^\top A\boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{b} + c) \tag{23}$$

for any fixed function $\sigma : \mathbb{R} \to \mathbb{R}$. Namely, it is a quadratic-form in x followed by nonlinear activation function σ.

Then, it is joint-invariant with $G = \mathrm{Aff}(m)$. In fact, we can put the group actions of $g = (\boldsymbol{t}, L) \in \mathbb{R}^m \rtimes GL(m)$ on X and Ξ by

$$(\boldsymbol{t}, L) \cdot \boldsymbol{x} := \boldsymbol{t} + L\boldsymbol{x}, \tag{24}$$

$$(\boldsymbol{t}, L) \cdot (A, \boldsymbol{b}, c) := (L^{-\top}AL^{-1}, L^{-\top}\boldsymbol{b} - 2L^{-\top}AL^{-1}\boldsymbol{t}, c + \boldsymbol{t}^\top L^{-\top}AL^{-1}\boldsymbol{t} - \boldsymbol{t}^\top L^{-\top}\boldsymbol{b}). \tag{25}$$

Then, the joint-G-action on $X \times \Xi$ remains the feature map joint-invariant as below.

$$\begin{aligned}
\phi(g \cdot \boldsymbol{x}, g \cdot \boldsymbol{\xi}) &= \sigma((L\boldsymbol{x} + \boldsymbol{t})^\top L^{-\top}AL^{-1}(L\boldsymbol{x} + \boldsymbol{t}) + (L\boldsymbol{x} + \boldsymbol{t})^\top(L^{-\top}\boldsymbol{b} - 2L^{-\top}AL^{-1}\boldsymbol{t}) + ...)\\
&= \sigma(\boldsymbol{x}^\top A\boldsymbol{x} + 2\boldsymbol{x}^\top AL^{-1}\boldsymbol{t} + \boldsymbol{t}^\top L^{-\top}AL^{-1}\boldsymbol{t} +\\
&\qquad + \boldsymbol{x}^\top\boldsymbol{b} - 2\boldsymbol{x}^\top AL^{-1}\boldsymbol{t} + \boldsymbol{t}^\top L^{-\top}\boldsymbol{b} - 2\boldsymbol{t}^\top L^{-\top}AL^{-1}\boldsymbol{t}\\
&\qquad + c + \boldsymbol{t}^\top L^{-\top}AL^{-1}\boldsymbol{t} - \boldsymbol{t}^\top L^{-\top}\boldsymbol{b})\\
&= \sigma(\boldsymbol{x}^\top A\boldsymbol{x} + \boldsymbol{x}^\top\boldsymbol{b} + c) = \phi(g \cdot \boldsymbol{x}, g \cdot \boldsymbol{\xi}).
\end{aligned}$$

By Theorem 3, the regular representation π of $\mathrm{Aff}(m)$ on $L^2(\mathbb{R}^m)$ is irreducible. Hence as a consequence of the general result, the following network is $L^2(\mathbb{R}^m)$-universal.

$$\mathtt{NN}[\gamma](\boldsymbol{x}) := \int_{M \times \mathbb{R}^m \times \mathbb{R}} \gamma(A, \boldsymbol{b}, c)\sigma(\boldsymbol{x}^\top A\boldsymbol{x} + \boldsymbol{x}^\top\boldsymbol{b} + c)\mathrm{d}A\mathrm{d}\boldsymbol{b}\mathrm{d}c. \tag{26}$$

7 Discussion

We have developed a systematic method for deriving a ridgelet transform for a wide range of learning machines defined by joint-group-equivariant feature maps, yielding the universal approximation theorems as corollaries. Traditionally, the techniques used in the expressive power analysis of deep networks were different from those used in the analysis of shallow networks, as overviewed in the introduction. Our main theorem unifies the approximation schemes of both deep and shallow networks from the perspective of joint-group-action on the data-parameter domain. Technically, this unification is due to Schur's lemma, a basic and useful result in the representation theory. Thanks to this lemma, the proof of the main theorem is simple, yet the scope of application is wide. The significance of this study lies in revealing the close relationship between machine learning theory and modern algebra. With this study as a catalyst, we expect a major upgrade to machine learning theory from the perspective of modern algebra.

REFERENCES

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

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

Yongqiang Cai. Achieve the Minimum Width of Neural Networks for Universal Approximation. In *The Eleventh International Conference on Learning Representations*, 2023.

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

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, pp. 607–611. IEEE, 1989.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems*, volume 31, pp. 6572–6583, Palais des Congrès de Montréal, Montréal CANADA, 2018. Curran Associates, Inc.

Albert Cohen, Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk. Optimal Stable Nonlinear Approximation. *Foundations of Computational Mathematics*, 22(3):607–648, 2022.

Nadav Cohen, Or Sharir, and Amnon Shashua. On the Expressive Power of Deep Learning: A Tensor Analysis. In *29th Annual Conference on Learning Theory*, volume 49, pp. 1–31, 2016.

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

I Daubechies, R DeVore, S Foucart, B Hanin, and G Petrova. Nonlinear Approximation and (Deep) ReLU Networks. *Constructive Approximation*, 55(1):127–172, 2022.

Weinan E. A Proposal on Machine Learning via Dynamical Systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

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

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

Philipp Grohs, Andreas Klotz, and Felix Voigtlaender. Phase Transitions in Rate Distortion Theory and Deep Learning. *Foundations of Computational Mathematics*, 23(1):329–392, 2023.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):1–22, 2017.

Boris Hanin and Mark Sellke. Approximating Continuous Functions by ReLU Nets of Minimal Width. *arXiv preprint: 1710.11278*, 2017.

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

Patrick Kidger and Terry Lyons. Universal Approximation with Deep Narrow Networks. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2306–2327. PMLR, 2020.

Namjun Kim, Chanho Min, and Sejun Park. Minimum width for universal approximation using ReLU networks on compact domain. In *The Twelfth International Conference on Learning Representations*, 2024.

Li'Ang Li, Yifei Duan, Guanghua Ji, and Yongqiang Cai. Minimum Width of Leaky-ReLU Neural Networks for Uniform Universal Approximation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19460–19470. PMLR, 2023.

Qianxiao Li and Shuji Hao. An Optimal Control Approach to Deep Learning and Applications to Discrete-Weight Neural Networks. In *Proceedings of The 35th International Conference on Machine Learning*, volume 80, pp. 2985–2994, Stockholm, 2018. PMLR.

Hongzhou Lin and Stefanie Jegelka. ResNet with one-neuron hidden layers is a Universal Approximator. In *Advances in Neural Information Processing Systems*, volume 31, Montreal, BC, 2018. Curran Associates, Inc.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The Expressive Power of Neural Networks: A View from the Width. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

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

Sejun Park, Chulhee Yun, Jaeho Lee, and Jinwoo Shin. Minimum Width for Universal Approximation. In *International Conference on Learning Representations*, 2021.

Guergana Petrova and Przemyslaw Wojtaszczyk. Limitations on approximation by deep and shallow neural networks. *Journal of Machine Learning Research*, 24(353):1–38, 2023.

Jonathan W Siegel. Optimal Approximation Rates for Deep ReLU Neural Networks on Sobolev and Besov Spaces. *Journal of Machine Learning Research*, 24(357):1–52, 2023.

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

Sho Sonoda and Noboru Murata. Transportation analysis of denoising autoencoders: a novel method for analyzing deep neural networks. In *NIPS 2017 Workshop on Optimal Transport & Machine Learning (OTML)*, pp. 1–10, Long Beach, 2017b.

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, pp. 2674–2682. PMLR, 2021a.

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

Sho Sonoda, Isao Ishikawa, and Masahiro Ikeda. Universality of Group Convolutional Neural Networks Based on Ridgelet Analysis on Groups. In *Advances in Neural Information Processing Systems 35*, pp. 38680–38694, New Orleans, Louisiana, USA, 2022a. Curran Associates, Inc.

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, pp. 20405–20422, Baltimore, Maryland, USA, 2022b. PMLR.

Sho Sonoda, Hideyuki Ishi, Isao Ishikawa, and Masahiro Ikeda. Joint Group Invariant Functions on Data-Parameter Domain Induce Universal Neural Networks. In *Proceedings of the 2nd NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, Proceedings of Machine Learning Research, pp. 129–144. PMLR, 2024a.

Sho Sonoda, Isao Ishikawa, and Masahiro Ikeda. A unified Fourier slice method to derive ridgelet transform for a variety of depth-2 neural networks. *Journal of Statistical Planning and Inference*, 233:106184, 2024b.

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

Hayata Yamasaki, Sathyawageeswar Subramanian, Satoshi Hayakawa, and Sho Sonoda. Quantum Ridgelet Transform: Winning Lottery Ticket of Neural Networks with Quantum Computation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 39008–39034, Honolulu, Hawaii, USA, 2023. PMLR.

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

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In *Proceedings of the 31st Conference On Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 639–649. PMLR, 2018.

Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13005–13015. Curran Associates, Inc., 2020.

A Proofs

A.1 Unitarity of Representations

Lemma 6. *π is a unitary representation of G on $L^2(X;Y)$.*

Proof. Recall that the representation υ of G on Y is unitary. So, for any $g, h \in G$ and $f \in L^2(X;Y)$,

$$\pi_g[\pi_h[f]](x) = g \cdot (h \cdot f(h^{-1} \cdot (g^{-1} \cdot x))) = (gh) \cdot f((gh)^{-1} \cdot x) = \pi_{gh}[f](x),$$

and for any $g \in G$ and $f_1, f_2 \in L^2(X;Y)$,

$$\begin{aligned}\langle \pi_g[f_1], \pi_g[f_2] \rangle_{L^2(X;Y)} &= \int_X \langle \upsilon_g[f_1(g^{-1} \cdot x)], \upsilon_g[f_2(g^{-1} \cdot x)] \rangle_Y \mathrm{d}x \\ &= \int_X \langle f_1(x), \upsilon_g^*[\upsilon_g[f_2(x)]] \rangle_Y \mathrm{d}x = \langle f_1, f_2 \rangle_{L^2(X;Y)}. \qquad \square\end{aligned}$$

Lemma 7. *$\widehat{\pi}$ is a unitary representation of G on $L^2(\Xi)$.*

Proof. For any $g, h \in G$ and $\gamma \in L^2(\Xi)$,

$$\widehat{\pi}_g[\widehat{\pi}_h[\gamma]](\xi) = \gamma(h^{-1} \cdot (g^{-1} \cdot \xi) = \gamma((gh)^{-1} \cdot \xi) = \widehat{\pi}_{gh}[f](x),$$

and for any $g \in G$ and $\gamma_1, \gamma_2 \in L^2(\Xi)$,

$$\begin{aligned}\langle \widehat{\pi}_g[\gamma_1], \widehat{\pi}_g[\gamma_2] \rangle_{L^2(\Xi)} &= \int_\Xi \gamma_1(g^{-1} \cdot \xi) \overline{\gamma_2(g^{-1} \cdot \xi)} \mathrm{d}\xi \\ &= \int_\Xi \gamma_1(\xi) \overline{\gamma_2(\xi)} \mathrm{d}\xi = \langle \gamma_1, \gamma_2 \rangle_{L^2(\Xi)}. \qquad \square\end{aligned}$$

A.2 Proof of Lemma 2

Proof. For any $g \in G, x \in X$, and $\xi_{1:n} \in \Xi_{1:n}$, we have

$$\begin{aligned}\phi_{1:n}(g \cdot x, g \cdot \xi_{1:n}) &= \phi_n(\bullet, g \cdot \xi_n) \circ \cdots \circ \phi_2(\bullet, g \cdot \xi_2) \circ \phi_1(g \cdot x, g \cdot \xi_1) \\ &= \phi_n(\bullet, g \cdot \xi_n) \circ \cdots \circ \phi_2(g \cdot \bullet, g \cdot \xi_2) \circ \phi_1(x, \xi_1) \\ &\vdots \\ &= \phi_n(g \cdot \bullet, g \cdot \xi_n) \circ \cdots \circ \phi_2(\bullet, \xi_2) \circ \phi_1(x, \xi_1) \\ &= g \cdot \phi_n(\bullet, \xi_n) \circ \cdots \circ \phi_2(\bullet, \xi_2) \circ \phi_1(x, \xi_1) \\ &= g \cdot \phi_{1:n}(x, \xi_{1:n}). \qquad \square\end{aligned}$$

A.3 PROOF OF THEOREM 4

Proof. By using the unitarity of representation $\upsilon : G \to \mathcal{U}(Y)$, left-invariance of measure $\mathrm{d}x$, and G-equivariance of feature map ψ, for all $g \in G$, we have

$$\begin{aligned} \mathtt{R}_\psi[\pi_g[f]](\xi) &= \int_X \langle g \cdot f(g^{-1} \cdot x), \psi(x,\xi) \rangle_Y \mathrm{d}x = \int_X \langle f(x), g^{-1} \cdot \psi(g \cdot x, \xi) \rangle_Y \mathrm{d}x \\ &= \int_X \langle f(x), \psi(x, g^{-1} \cdot \xi) \rangle_Y \mathrm{d}x = \widehat{\pi}_g[\mathtt{R}_\psi[f]](\xi). \end{aligned} \tag{27}$$

Similarly,

$$\begin{aligned} \mathtt{LM}_\phi[\widehat{\pi}_g[\gamma]](x) &= \int_\Xi \gamma(g^{-1} \cdot \xi) \phi(x,\xi) \mathrm{d}\xi = \int_\Xi \gamma(\xi) \phi(x, g \cdot \xi) \mathrm{d}\xi \\ &= \int_\Xi \gamma(\xi) \left(g \cdot \phi(g^{-1} \cdot x, \xi) \right) \mathrm{d}\xi = \pi_g[\mathtt{LM}_\phi[\gamma]](x). \end{aligned} \tag{28}$$

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

$$\mathtt{LM}_\phi \circ \mathtt{R}_\psi \circ \pi_g = \mathtt{LM}_\phi \circ \widehat{\pi}_g \circ \mathtt{R}_\psi = \pi_g \circ \mathtt{LM}_\phi \circ \mathtt{R}_\psi \tag{29}$$

for all $g \in G$. Hence by Schur's lemma (Theorem 2), there exist a constant $C_{\phi,\psi} \in \mathbb{C}$ such that $\mathtt{LM}_\phi \circ \mathtt{R}_\psi = C_{\phi,\psi} \, \mathrm{Id}_{L^2(X)}$. Since $\mathtt{LM}_\phi \circ \mathtt{R}_\psi$ is bilinear in ϕ and ψ, $C_{\phi,\psi}$ is bilinear in ϕ and ψ. □

B EXAMPLE: CLASSICAL EQUIVARIANT NETWORK

As mentioned in Remark 1, all the classical equivariant feature maps are automatically joint-equivariant. Hence by assuming the irreducibility of an appropriate unitary representation π of group G on $L^2(X;Y)$, we can conclude the universality from our main theorem.

In fact, however, the universality of classical group-equivariant networks have already been shown in a unified manner by using the ridgelet transform in Sonoda et al. (2022a). The feature map is defined in Eq.19 as

$$\phi(x,(a,b))(h) = \sigma((a *_T x)(h) - b), \quad x, a \in \mathcal{H}, b \in \mathbb{R}, h \in G.$$

It is indeed joint-equivariant because

$$\phi(g \cdot x, g \cdot (a,b))(h) = \sigma(\langle (h^{-1}g) \cdot x, a \rangle - b) = \phi(x,(a,b))(g^{-1}h)$$

with G-action on parameters (a,b) being trivial. As showcased in Section 5, this feature map covers a wide range of typical group-equivariant networks such as Deep Sets and $\mathrm{E}(n)$-equivariant maps.