Deep Ridgelet Transform: Voice with Koopman Operator Proves Universality of Formal Deep Networks

Sho Sonoda[1] SHO.SONODA@RIKEN.JP
Yuka Hashimoto[2,1] YUKA.HASHIMOTO@NTT.COM
Isao Ishikawa[3,1] ISHIKAWA.ISAO.ZX@EHIME-U.AC.JP
Masahiro Ikeda[1] MASAHIRO.IKEDA@RIKEN.JP
[1] *Center for Advanced Intelligence Project (AIP), RIKEN*
[2] *NTT Network Service Systems Laboratories, NTT Corporation*
[3] *Center for Data Science, Ehime University*

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

Abstract

We identify hidden layers inside a deep neural network (DNN) with group actions on the data domain, and formulate a formal deep network as a dual voice transform with respect to the Koopman operator, a linear representation of the group action. Based on the group theoretic arguments, particularly by using Schur's lemma, we show a simple proof of the universality of DNNs.

Keywords: deep neural network, group representation, Koopman operator, Schur's lemma, voice transform

1. Introduction

An ultimate goal of deep learning theories is to characterize the network parameters obtained by deep learning. We may formulate this characterization problem as a functional equation problem: Let $\mathcal{H}$ denote a class of data generating functions, and let $\mathtt{DNN}[\gamma]$ denote a certain deep neural network with parameter γ. Given a function $f \in \mathcal{H}$, find an unknown parameter γ so that network $\mathtt{DNN}[\gamma]$ represents function f, i.e.

$$\mathtt{DNN}[\gamma] = f. \tag{1}$$

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

When the network has only one infinitely-wide hidden layer, though it is not deep but shallow, the characterization problem has been well investigated. For example, the learning dynamics and the global convergence property (of SGD) are well studied in the mean field

theory (Nitanda and Suzuki, 2017; Rotskoff and Vanden-Eijnden, 2018; Mei et al., 2018; Chizat and Bach, 2018) and the Langevin dynamics theory (Suzuki, 2020), and even closed-form solution operator to a "shallow" NN equation, the original ridgelet transform, has already been presented (Sonoda et al., 2021a,b, 2022a,b).

On the other hand, when the network has more than one hidden layer, the problem is far from solved, and it is common to either consider infinitely-deep mathematical models such as Neural ODEs (Sonoda and Murata, 2017; E, 2017; Li and Hao, 2018; Haber and Ruthotto, 2017; Chen et al., 2018), or handcraft inner feature maps depending on the problem. For example, construction methods such as the Telgarsky sawtooth function (or the Yarotsky scheme) and bit extraction techniques (Cohen et al., 2016; Telgarsky, 2016; Yarotsky, 2017, 2018; Yarotsky and Zhevnerchuk, 2020) have been proposed to demonstrate the depth separation and the minmax optimality of deep learning methods. 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 and Sellke, 2017; Lin and Jegelka, 2018; Kidger and Lyons, 2020; Park et al., 2021; Cai, 2023). Needless to say, there is no guarantee that these handcrafted feature maps are acquired by deep learning, so these analyses are considered to be analyses of possible worlds.

In this study, we introduce a *formal deep network* as an infinite mixture of the *Koopman operators*, and solve the DNN equation for the first time by using the *voice transform*. In other words, we present a first ridgelet transform for DNNs, which us understood as the constructive proof of the $\mathcal{H}$-universality of DNNs without handcrafting network architecture. Besides, the proof is simple by using Schur's lemma.

2. Technical Backgrounds

We briefly overview a few technical backgrounds. *Schur's lemma* and the *Haar measure* play key roles in the proof of the main results. The *(dual) voice transform* and the *Koopman operator* (as a unitary representation) are key aspects of the DNN considered in this study.

Notation. For any measure space X, $L^2(X)$ denotes the Hilbert space of all square-integrable functions f on X. For any topological space X, $C_c(X)$ denotes the Banach space of all compactly supported functions f on X.

2.1. Irreducible Unitary Representation and Schur's Lemma

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

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

Theorem 1 (Schur's lemma) *A unitary representation $(\pi, \mathcal{H})$ of G is irreducible iff $C(\pi)$ only contains scalar multiples of the identity, i.e. $C(\pi) = \{c\,\mathrm{Id}_{\mathcal{H}} \mid c \in \mathbb{C}\}$ or $\{0\}$.*

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

2.2. Calculus on Locally Compact Group

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

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

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

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

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

2.3. Voice Transform, or Generalized Wavelet Transform

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

Definition 2 *Given a unitary representation $(\pi, \mathcal{H})$ of group G on a Hilbert space $\mathcal{H}$, the voice transform is defined as*

$$V_\phi[f](g) := \langle f, \pi_g[\phi] \rangle_{\mathcal{H}}, \quad g \in G, \ f, \phi \in \mathcal{H}. \tag{2}$$

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

For example, the wavelet transform

$$W[f; \psi](b, a) = \int_{\mathbb{R}} f(x) \overline{\frac{1}{\sqrt{a}} \psi\left(\frac{x - b}{a}\right)} \mathrm{d}x, \quad (b, a) \in \mathbb{R} \times \mathbb{R}_+$$

is the voice transform with 1-dim. Affine group ("$ax + b$-group") acting on $L^2(\mathbb{R})$.

One of the strengths of this general theory is that a pseudo-inverse is given simply by the dual V_ψ^*.

Theorem 3 (Reconstruction Formula) *Let $\pi : G \to U(\mathcal{H})$ be a square integrable representation and fix an admissible vector $\psi \in \mathcal{H}$. For any $\gamma \in L^2(G)$, put the weak integral*

$$V_\psi^*[\gamma] := \int_G \gamma(g)\pi_g[\psi]\mathrm{d}g. \tag{3}$$

Then, for any $f \in \mathcal{H}$,

$$V_\psi^*[V_\psi[f]] = f. \tag{4}$$

Here, $\psi \in \mathcal{H}$ is called an *admissible* vector when $\|V_\psi[\psi]\|_{L^2(G)} = 1$, and π is said *square integrable* when there exists at least one admissible vector. We refer to Berge (2021, Proposition 2.33 and Corollary 2.34) for more details.

2.4. Koopman Operator

The Koopman operator is first appeared in Koopman (1931) and Neumann (1932) in the dynamical systems theory, and have been applied for data science since around 2000s (e.g., by Mezić, 2005). We refer to Brunton et al. (2022), Mauroy et al. (2020), and Eisner et al. (2015) for more details.

Definition 4 (Koopman Operator) *Let X be a topological space, and $g : X \to X$ be a map. For any continuous function $\psi \in C(X)$, the Koopman operator with respect to g is the following composition operator:*

$$K_g[\psi] := \psi \circ g. \tag{5}$$

The definition of Koopman operators seems a trivial rewriting, but the strength is the so-called *linearization effect* that in the raw form $\psi \circ g$ the dependence on g is nonlinear, whereas in the operator form $K_g[\psi]$ the dependence on K_g is linear, i.e. $K_g[\psi_1 + \psi_2] = K_g[\psi_1] + K_g[\psi_2]$. (More precisely, K_g also preserves the product of functions, i.e. $K_g[\psi_1\psi_2] = K_g[\psi_1]K_g[\psi_2]$, making it an algebra homomorphism.)

Lemma 5 *Let G be a group of invertible maps $g : X \to X$ with product $gh = g \circ h$ and left action $g \cdot x = g(x)$. Then $K : G \to U(L^2(X))$ is a unitary representation of G acting from right on $L^2(X, \mathrm{d}_l x)$. Namely, for any $g, h \in G$ and $\psi, \phi \in L^2(X, \mathrm{d}_l x)$,*

$$\langle K_g[\psi], K_g[\phi] \rangle_{L^2(X,\mathrm{d}_l x)} = \int_X \psi \circ g(x)\overline{\phi \circ g(x)}\mathrm{d}_l x = \int_X \psi(x)\overline{\phi(x)}\mathrm{d}_l x = \langle \psi, \phi \rangle_{L^2(X,\mathrm{d}_l x)},$$
$$K_g[K_h[\psi]] = \psi \circ h \circ g = K_{hg}[\psi].$$

3. Formal Deep Network

We define a *formal deep network* in two steps: First, introduce a subnetwork, then define an entire network. The key concept is to identify each hidden layer, say g, with an element of a group G acting on the input space X, and the composite of hidden layers, say $g \circ h$, with the group operation gh. Since a group is closed under its operation by definition, the proposed network can represent literally *any depth* such as a single hiden layer g, double hidden layers $g \circ h$, triple hidden layers $g \circ h \circ k$, and infinite hidden layers $g \circ h \circ \cdots$.

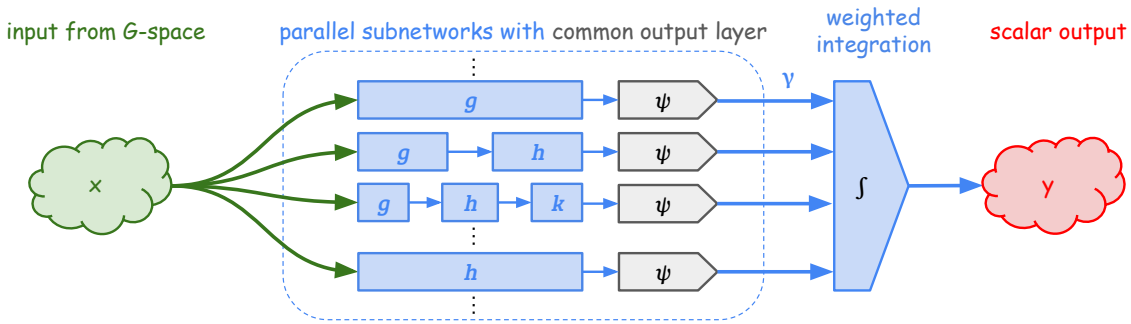

Figure 1: Formal Deep Network is an infinite mixture (or parallel connection) of subnetworks. The hidden layers $g, g \circ h, \cdots$ are formulated as group actions so that the proposed network can deal with any depth and any input domain. Thanks to the integral form, the only parameter is the weight function $\gamma(g)$ in the output layer.

3.1. Formal Deep Subnetwork

Let G be a locally compact group equipped with an invariant measure $\mathrm{d}g$, let X be a G-space equipped with invariant measure $\mathrm{d}x$ (induced from $\mathrm{d}g$), and let $L^2(X)$ be the Hilbert space of square integrable functions on X. To identify the group action with the hidden layer, we write the group action $g \cdot x$ as $g(x)$.

Definition 6 *For any function $\psi \in L^2(X)$ and element $g \in G$, put*

$$K_g[\psi](x) := \psi \circ g(x), \quad x \in X. \tag{6}$$

That is, we identify the Koopman operator $K_g[\psi]$ as a deep network composed of hidden layer g and output layer ψ. We say it is *formal* because neither ψ nor g has specific network implementations. In order to investigate function approximation by composite maps, we extract only the mathematical structure of the composite mapping of deep networks.

Remark 7 *We may consider a DNN equation*

$$K_g[\psi] = f, \tag{7}$$

with regarding both g and ψ with parameters. In fact, under some mild conditions, K_g has a pseudo-inverse operator $K_g^\dagger$ satisfying $K_g[K_g^\dagger[f]] = f$. Thus, given a function f on X,

$$(\psi_f, g_f) := (K_g^\dagger[f], g) \tag{8}$$

is a solution of the DNN equation (7), namely $K_{g_f}[\psi_f] = f$. However, the obtained solution is (1) less informative because hidden layer g_f can be independent of f thus remain as a hyper-parameter, and (2) less feasible because computing pseudo-inverse $K_g^\dagger$ is in general another hard problem.

3.2. Formal Deep Network

To circumvent the difficulties of the single Koopman operator formulation, we come to impose additional integration layer as below.

Definition 8 *For any function $\psi \in L^2(X)$ and measure γ on X, put*

$$\mathtt{DNN}[\gamma;\psi](x) := \int_G \gamma(g)\, \underbrace{\psi \circ g}_{=K_g[\psi]}(x)\mathrm{d}g, \quad x \in X \tag{9}$$

We call $\mathtt{DNN}[\gamma;\psi]$ a formal deep network with weight γ and sub-output layer ψ.

The integration over G means that the entire network $\mathtt{DNN}[\gamma;\psi]$ is a γ-weighted parallel connection of (at most infinite) subnetworks $\{\psi \circ g \mid g \in G\}$. Thanks to the integral form, we do not need to directly specify which hidden map $g \in G$ to use. Instead, we specify indirectly via the weight function γ. For example, if γ has a high intensity at $g_0 \in G$, then the subnetwork g_0 is considered to be essential for the entire network to express given f.

We remark that the integral form is another *linearization trick*, since in the single operator form $K_g[\psi]$ the dependence on raw g is still nonlinear, whereas in the integral form $\langle \gamma, K_\bullet[\psi] \rangle$ the dependence on γ is linear, i.e. $\mathtt{DNN}[\gamma_1+\gamma_2] = \mathtt{DNN}[\gamma_1] + \mathtt{DNN}[\gamma_2]$.

4. Main Results

In the following, we use right invariant measure $\mathrm{d}_r g$ for $L^2(G)$ and left invariant measure $\mathrm{d}_l x$ for $L^2(X)$ so that the Koopman operator K becomes a unitary representation of G acting from right on $L^2(X, \mathrm{d}_l x)$ (as discussed in Lemma 5). Then, the formal deep network $\mathtt{DNN}[\gamma;\psi]$ can be identified with the dual voice transform generated from the Koopman operator. Therefore, it is natural to define the *ridgelet transform*, or a solution operator to the DNN equation, as the voice transform with respect to the Koopman operator as below.

Definition 9 (Deep Ridgelet Transform) *For any functions $f, \psi \in L^2(X)$, put*

$$R_\psi[f](g) := \langle f, K_g[\psi] \rangle_{L^2(X)} = \int_X f(x)\overline{K_g[\psi](x)}\mathrm{d}_l x, \quad g \in G. \tag{10}$$

Since K_g is a unitary representation of G, this is a voice transform. It is straightforward to see that $\mathtt{DNN}$ is the adjoint of R as below:

$$\langle \gamma, R_\psi[f] \rangle_{L^2(G)} = \int_{X\times G} \gamma(g) K_g[\psi](x)\overline{f(x)}\mathrm{d}_l x \mathrm{d}_r g = \langle \mathtt{DNN}_\psi[\gamma], f \rangle_{L^2(X)}. \tag{11}$$

Namely, $\mathtt{DNN}$ is the dual voice transform. Hence according to the general result of the voice transform theory, the reconstruction formula $\mathtt{DNN}[R[f]] = f$ should hold under the assumption that the unitary representation K is *irreducible*. In general, however, K is not irreducible on the entire space $L^2(X)$. So we state the theorem for an invariant subspace $\mathcal{H}$ of $L^2(X)$ on which the restriction of K is irreducible.

Theorem 10 (Reconstruction Formula) *Suppose (1) $\mathcal{H}$ is an invariant subspace of $L^2(X)$ on which K is irreducible, and (2) $\psi \in \mathcal{H}$ satisfies the admissibility condition $c_\psi := \|R_\psi[\psi]\|^2_{L^2(G)}/\|\psi\|^2_{L^2(X)} < \infty$. Then, for any $f \in \mathcal{H}$,*

$$\mathtt{DNN}_\psi[R_\psi[f]] = \int_G R_\psi[f](g)\, \psi \circ g(\bullet) \mathrm{d}_r g = c_\psi f. \tag{12}$$

In other words, the deep ridgelet transform R_ψ solves the DNN equation: Given a function $f \in \mathcal{H}$, find a parameter γ satisfying

$$\mathtt{DNN}_\psi[\gamma] = f. \tag{13}$$

As mentioned in the Introduction, it concludes the $\mathcal{H}$-universality of DNN because for any $f \in \mathcal{H}$, there exists a γ_f (namely $\gamma_f = R[f]$) satisfying $\mathtt{DNN}_\psi[\gamma_f] = f$. In particular, it leads to a constructive proof without handcrafting feature maps because the closed-form expression (10) of the ridgelet transform explicitly indicates which feature map $\psi \circ g$ to use (from the pool of candidate subnetworks $\{\psi \circ g \mid g \in G\}$) by weighting on them.

Proof Put a dual action $\widehat{K}$ of G on $C(G)$ by

$$\widehat{K}_g[\gamma](h) := \gamma(hg^{-1}), \quad g, h \in G, \gamma \in C(G). \tag{14}$$

We can see

$$R_\psi \circ K_g = \widehat{K}_g \circ R_\psi, \quad \text{and} \quad \mathtt{DNN}_\psi \circ \widehat{K}_g = K_g \circ \mathtt{DNN}_\psi. \tag{15}$$

In fact, by the left and right invariances of $\mathrm{d}_l x$ and $\mathrm{d}_r g$ respectively,

$$\begin{aligned}
R_\psi[K_g[f]](h) &= \int_X f \circ g(x) \overline{\psi \circ h(x)} \mathrm{d}_l x \\
&= \int_X f(x) \overline{\psi \circ h \circ g^{-1}(x)} \mathrm{d}_l x \\
&= \int_X f(x) \overline{\psi((hg^{-1}) \circ x)} \mathrm{d}_l x = \widehat{K}_g[R_\psi[f]](h), \\
\mathtt{DNN}_\psi[\widehat{K}_g[\gamma]](x) &= \int_G \gamma(hg^{-1}) K_h[\psi](x) \mathrm{d}_r h \\
&= \int_G \gamma(h) K_{hg}[\psi](x) \mathrm{d}_r h \\
&= \int_G \gamma(h) K_g[K_h[\psi]](x) \mathrm{d}_r h = K_g[\mathtt{DNN}_\psi[\gamma]](x).
\end{aligned}$$

Therefore, K_g commutes with $\mathtt{DNN}_\psi \circ R_\psi$ for all $g \in G$ as below

$$\mathtt{DNN}_\psi \circ R_\psi \circ K_g = \mathtt{DNN}_\psi \circ \widehat{K}_g \circ R_\psi = K_g \circ \mathtt{DNN}_\psi \circ R_\psi. \tag{16}$$

By the assumption that K_g is irreducible, Schur's lemma yields that there exists a constant $c_\psi \in \mathbb{C}$ such that $\mathtt{DNN}_\psi \circ R_\psi = c_\psi \mathrm{Id}_{\mathcal{H}}$. But the admissible condition implies $c_\psi = \|R_\psi[\psi]\|^2/\|\psi\|^2$ because $\|R_\psi[\psi]\|^2 = \langle R_\psi[\psi], R_\psi[\psi]\rangle = \langle \psi, \mathtt{DNN}_\psi[R_\psi[\psi]]\rangle = c_\psi \|\psi\|^2$. ■

5. Discussion

We introduced the formal deep network, and derived the deep ridgelet transform R_ψ to solve the corresponding DNN equation, yielding a constructive proof of $\mathcal{H}$-universality without handcrafting feature maps. By formulating a hidden layer as a group action, our result covers a variety of DNNs with any depth on any G-space X. Further, by introducing an integral form, the network parameter is linearized and the network comes to be identified with a dual voice transform with the unitary group representation being the Koopman operator, resulting in a simple proof based on Schur's lemma.

The assumption that hidden layers form a group may sound too ideal, and it may be more realistic to calculate ridgelet transforms for semigroups. However, we consider it is unlikely that something deviating significantly from the basic idea of the standard voice transform. Rather, more important contributions of this study lie in demonstrating that the theory of function approximation by *composite maps* is also a member of the voice transform kingdom, and/or in indicating a method to avoid another hard problem of calculating the pseudo-inverse of the Koopman operator.

Acknowledgments

The authors are extremely grateful to the three anonymous reviewers for their valuable comments and suggestions, which have helped improve the quality of our manuscript. The authors are grateful to Professor Kenji Fukumizu, Professor Yoshinobu Kawahara, Professor Noboru Murata, Professor Atsushi Nitanda, and Professor Taiji Suzuki for productive comments on the early version of this study. This work was supported by JSPS KAKENHI 20K03657, JST PRESTO JPMJPR2125, JST CREST JPMJCR2015 and JPMJCR1913, and JST ACTX JPMJAX2004.

References

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

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

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

S. L. Brunton, M. Budišić, E. Kaiser, and J. N. Kutz. Modern Koopman Theory for Dynamical Systems. *SIAM Review*, 64(2):229–340, 2022.

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

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

R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud. Neural Ordinary Differential Equations. In *Advances in Neural Information Processing Systems 31*, pages 6572–6583, Palais des Congrès de Montréal, Montréal CANADA, 2018.

L. Chizat and F. 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, Montreal, BC, 2018.

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

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

T. Eisner, B. Farkas, M. Haase, and R. Nagel. *Operator Theoretic Aspects of Ergodic Theory*. Graduate Texts in Mathematics. Springer Cham, 2015.

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

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

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

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

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

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

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

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

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

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

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

P. Kidger and T. Lyons. Universal Approximation with Deep Narrow Networks. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 2306–2327, 2020.

B. O. Koopman. Hamiltonian Systems and Transformation in Hilbert Space. *Proceedings of the National Academy of Sciences*, 17(5):315–318, 1931.

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

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

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

Q. Li and S. 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, pages 2985–2994, Stockholm, 2018.

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

Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang. The Expressive Power of Neural Networks: A View from the Width. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

A. Mauroy, I. Mezić, and Y. Susuki. *The Koopman Operator in Systems and Control: Concepts, Methodologies, and Applications*. Lecture Notes in Control and Information Sciences. Springer Cham, 2020.

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

I. Mezić. Spectral Properties of Dynamical Systems, Model Reduction and Decompositions. *Nonlinear Dynamics*, 41(1):309–325, 2005.

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

J. v. Neumann. Zur Operatorenmethode In Der Klassischen Mechanik. *Annals of Mathematics*, 33(3):587–642, 1932.

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

S. Park, C. Yun, J. Lee, and J. Shin. Minimum Width for Universal Approximation. In *International Conference on Learning Representations*, 2021.

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

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

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

S. Sonoda, I. Ishikawa, and M. 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, 2021a.

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

S. Sonoda, I. Ishikawa, and M. Ikeda. Universality of Group Convolutional Neural Networks Based on Ridgelet Analysis on Groups. In *Advances in Neural Information Processing Systems 35*, pages 38680–38694, New Orleans, Louisiana, USA, 2022a.

S. Sonoda, I. Ishikawa, and M. Ikeda. Fully-Connected Network on Noncompact Symmetric Space and Ridgelet Transform based on Helgason-Fourier Analysis. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 20405–20422, Baltimore, Maryland, USA, 2022b.

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

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

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

D. 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*, pages 639–649, 2018.

D. Yarotsky and A. Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 13005–13015, 2020.