SIAM J. MATH. DATA SCI.
Vol. 7, No. 1, pp. 110–135

Function-Space Optimality of Neural Architectures with Multivariate Nonlinearities*

Rahul Parhi† and Michael Unser‡

Abstract. We investigate the function-space optimality (specifically, the Banach-space optimality) of a large class of shallow neural architectures with multivariate nonlinearities/activation functions. To that end, we construct a new family of Banach spaces defined via a regularization operator, the k-plane transform, and a sparsity-promoting norm. We prove a representer theorem that states that the solution sets to learning problems posed over these Banach spaces are completely characterized by neural architectures with multivariate nonlinearities. These optimal architectures have skip connections and are tightly connected to orthogonal weight normalization and multi-index models, both of which have received recent interest in the neural network community. Our framework is compatible with a number of classical nonlinearities including the rectified linear unit activation function, the norm activation function, and the radial basis functions found in the theory of thin-plate/polyharmonic splines. We also show that the underlying spaces are special instances of reproducing kernel Banach spaces and variation spaces. Our results shed light on the regularity of functions learned by neural networks trained on data, particularly with multivariate nonlinearities, and provide new theoretical motivation for several architectural choices found in practice.

Key words. multi-index models, multivariate nonlinearities, neural networks, regularization, representer theorem

MSC codes. 46E27, 47A52, 68T05, 82C32, 94A12

DOI. 10.1137/23M1620971

1. Introduction.

In supervised machine learning, the goal is to predict an output $y \in \mathcal{Y}$ (e.g., a label or response) from an input $\boldsymbol{x} \in \mathcal{X}$ (e.g., a feature or example), where $\mathcal{X}$ and $\mathcal{Y}$ denote the domain of the inputs and outputs, respectively. One solves this task by "training" a model to fit a set of data which consists of a finite number of input-output pairs $\{(\boldsymbol{x}_m, y_m)\}_{m=1}^M \subset \mathcal{X} \times \mathcal{Y}$. The goal is to "learn" a function $f : \mathcal{X} \to \mathcal{Y}$ with $f(\boldsymbol{x}_m) \approx y_m$, $m = 1, \dots, M$, such that f can accurately predict the output $y \in \mathcal{Y}$ of a new input $\boldsymbol{x} \in \mathcal{X}$. This task is usually formulated as an optimization problem of the form

$$\min_{f \in \mathcal{F}} \sum_{m=1}^{M} \mathcal{L}(y_m, f(\boldsymbol{x}_m)) + \lambda \Phi(f), \tag{1.1}$$

*Received by the editors December 1, 2023; accepted for publication (in revised form) October 14, 2024; published electronically January 9, 2025.
https://doi.org/10.1137/23M1620971
Funding: This work was supported in part by the Swiss National Science Foundation under grant 200020_219356 / 1.
†Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093 USA, and part of this work was done while the author was with the Biomedical Imaging Group, École polytechnique fédérale de Lausanne, CH-1015 Lausanne, Switzerland (rahul@ucsd.edu).
‡Biomedical Imaging Group, École polytechnique fédérale de Lausanne, CH-1015 Lausanne, Switzerland (michael.unser@epfl.ch).

where $\mathcal{F}$ is a prescribed *model class* of functions that map $\mathcal{X} \to \mathcal{Y}$, $\mathcal{L}(\cdot,\cdot)$ is a *loss function,* and $\Phi : \mathcal{F} \to \mathbb{R}_{\geq 0}$ is a *regularization functional* that injects prior knowledge/regularity on the function to be learned. The hyperparameter $\lambda > 0$ controls the tradeoff between data fidelity and regularity. Without the inclusion of the regularization functional in (1.1), the problem is typically ill-posed. Indeed, in many practical scenarios the problem is *overparameterized* as the dimension of the model class $\mathcal{F}$ greatly exceeds the number M of data. A classical choice of model class is a reproducing kernel Hilbert space (RKHS). The accompanying regularization functional is the squared Hilbert norm of the RKHS. In this scenario, the RKHS representer theorem establishes that there exists a solution to (1.1) that takes the form of a linear combination of reproducing kernels centered at the data sites [10, 29, 55, 70, 71]. This provides an exact characterization of the function-space optimality of kernel methods.

Recently, there has been a line of work that investigates the function-space optimality of neural networks [4, 5, 41, 42, 43, 44, 46, 54, 62, 66]. Crucially, these works define and study (non-Hilbertian) Banach spaces defined by sparsity or variation. These spaces have an analytic description via the Radon transform [30, 41, 43]. The accompanying neural network representer theorems for these spaces were first established in [43] and then studied and refined by a number of authors [5, 44, 62, 66]. While these results characterize the function-space optimality of neural networks, they only consider univariate nonlinearities. We refer the reader to the recent survey [45] for an up-to-date summary of this research direction. The purpose of this paper is to further extend the existing results on the function-space optimality of neural architectures, with a particular focus on *multivariate nonlinearities,* which have gained recent interest in the neural network community [2, 21, 23, 38].

The form of a neuron with an m-variate nonlinearity, $1 \leq m \leq d$, is

$$\boldsymbol{x} \mapsto \rho(\mathbf{A}\boldsymbol{x} - \boldsymbol{t}), \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{1.2}$$

where $\rho : \mathbb{R}^m \to \mathbb{R}$ is the nonlinearity (or activation function), $\mathbf{A} \in \mathbb{R}^{m \times d}$ is a weight matrix that controls the orientation of the neuron, and $\boldsymbol{t} \in \mathbb{R}^m$ is a bias which controls the offset of neuron. When $m = 1$, these atoms can be written as

$$\boldsymbol{x} \mapsto \rho(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x} - t), \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{1.3}$$

with $\rho : \mathbb{R} \to \mathbb{R}$, $\boldsymbol{\alpha} \in \mathbb{R}^d$, and $t \in \mathbb{R}$, which is the form of a classical neuron with a univariate nonlinearity. Neurons with m-variate nonlinearities as in (1.2) have been studied under many different names, including *m-sparse functions* [1, 18], *generalized ridge functions* [28], $(d-m)$-*plane ridge functions* [47], and *multi-index models* [8, 9, 15, 32, 35]. Notably, multi-index models have gained recent interest from the neural network community [1, 4, 18, 48]. A shallow neural architecture with an m-variate nonlinearity $\rho : \mathbb{R}^m \to \mathbb{R}$ takes the form

$$\boldsymbol{x} \mapsto \sum_{n=1}^{N} v_n\, \rho(\mathbf{A}_n\boldsymbol{x} - \boldsymbol{t}_n), \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{1.4}$$

where, for $n = 1, \ldots, N$, $v_n \in \mathbb{R}$, $\mathbf{A}_n \in \mathbb{R}^{m \times d}$, and $\boldsymbol{t}_n \in \mathbb{R}^m$. Such architectures are sometimes called *generalized translation networks* [38, 39, 40] and are classically known to be universal approximators if and only if[1] $\rho : \mathbb{R}^m \to \mathbb{R}$ is not a polynomial [39, Corollary 3.3]. These architectures have also recently been studied in the context of ridgelet analysis for a variety of shallow neural architectures [61]. In this paper, we characterize, for all integers m with $1 \leq m \leq d$, the function-space optimality of neural architectures with m-variate nonlinearities of the form (1.4), for a large class of nonlinearities. We show that these architectures are optimal solutions to data-fitting problems posed over (non-Hilbertian) Banach spaces defined via a sparsity-promoting norm in the domain of the k-plane transform. When $m = 1$, our framework is compatible with univariate nonlinearities and classical neural architectures, including the ReLU activation function. At the opposite extreme ($m = d$) our framework encompasses sparse kernel expansions and radial basis functions [3, 50, 63]. To the best of our knowledge, the results for $1 < m < d$ are new.

1.1. Main contributions and road map. Our results shed light on the regularity of the functions learned by neural networks trained on data. They provide new theoretical motivation for several architectural choices often found in practice, particularly with multivariate nonlinearities. These results hinge on recent developments regarding the distributional extension and invertibility of the k-plane transform and its dual [47]. The main contributions and organization of this paper are summarized in the remainder of this section.

New neural network Banach spaces. We propose and study the properties of a new family of *native spaces,* defined by[2]

$$\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d) := \left\{ f : \mathbb{R}^d \to \mathbb{R} \text{ is measurable} : \begin{array}{l} \|\mathrm{K}_{d-k} \mathscr{R}_k \mathrm{L} f\|_{\mathcal{M}} < \infty, \\ \operatorname*{ess\,sup}_{\boldsymbol{x} \in \mathbb{R}^d} |f(\boldsymbol{x})|(1 + \|\boldsymbol{x}\|_2)^{-n_{\mathrm{L}}} < \infty \end{array} \right\} \subset \mathcal{S}'(\mathbb{R}^d), \tag{1.5}$$

where k is an integer such that $0 \leq k < d$, $\mathcal{S}'(\mathbb{R}^d)$ denotes the space of tempered distributions, L is a *k-plane-admissible* pseudodifferential operator (in the sense of Definition 3.1), $\mathscr{R}_k$ denotes the k-plane transform, and K_{d-k} is the filtering operator of computed tomography (CT) which is such that $\mathscr{R}_k^* \mathrm{K}_{d-k} \mathscr{R}_k = \mathrm{Id}$. The $\mathcal{M}$-norm denotes the total variation norm (in the sense of measures). It can be viewed as a "generalization" of the L^1-norm that can also be applied to distributions such as the Dirac impulse. Said differently, if $f \in \mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ is such that $\mathrm{K}_{d-k} \mathscr{R}_k \mathrm{L} f$ is a bona fide function (not a distribution), then

$$\|\mathrm{K}_{d-k} \mathscr{R}_k \mathrm{L} f\|_{\mathcal{M}} = \|\mathrm{K}_{d-k} \mathscr{R}_k \mathrm{L} f\|_{L^1}. \tag{1.6}$$

The growth restriction of degree n_{L} plays the role of a proxy to the order of L; more specifically, n_{L} is the highest polynomial degree annihilated by L. The growth restriction in the definition of the native space ensures that the null space of the operator $\mathrm{K}_{d-k} \mathscr{R}_k \mathrm{L}$ is finite-dimensional. In subsection 3.1, we prove that, when equipped with an appropriate direct-sum topology, $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ forms a Banach space that is isometrically isomorphic to the Cartesian product of a space of (Radon) measures with the space of polynomials of degree at most n_{L}. They add to the growing list of "neural Banach spaces" that are currently being actively investigated [57].

[1] This equivalence holds under the global assumption that $\rho : \mathbb{R}^m \to \mathbb{R}$ does not grow faster than a polynomial, i.e., it is a *tempered* function.

[2] We refer to a function as *measurable* when it is measurable with respect to the Lebesgue σ-algebra.

Representer theorems for neural networks with multivariate nonlinearities. We prove a representer theorem (Theorem 3.8) that states that, under mild assumptions on the loss function and L, the solution set to the optimization problem

$$\min_{f\in\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)} \sum_{m=1}^{M} \mathcal{L}(y_m, f(\boldsymbol{x}_m)) + \lambda\|\mathrm{K}_{d-k}\mathscr{R}_k \mathrm{L} f\|_{\mathcal{M}} \tag{1.7}$$

is completely characterized by shallow neural architectures with $(d-k)$-variate activation functions matched to the operator L and widths bounded by the number M of data (independent of the dimension d of the data). This result sheds light on the role of biases, skip connections, and the use of structured weight matrices in neural architectures. Indeed, these architectures take the form

$$\boldsymbol{x} \mapsto c(\boldsymbol{x}) + \sum_{n=1}^{N} v_n\, \rho_{\mathrm{L}}(\mathbf{A}_n\boldsymbol{x} - \boldsymbol{t}_n) \tag{1.8}$$

with $N \leq M$, where, for $n = 1,\dots,N$, $\mathbf{A}_n \in \mathbb{R}^{(d-k)\times d}$ is such that $\mathbf{A}_n\mathbf{A}_n^{\mathsf{T}} = \mathbf{I}_{d-k}$ (identity matrix), $\boldsymbol{t}_n \in \mathbb{R}^{d-k}$, and $v_n \in \mathbb{R}\setminus\{0\}$. The function c is a polynomial of degree at most n_{L} and the function $\rho_{\mathrm{L}} : \mathbb{R}^{d-k} \to \mathbb{R}$ is a $(d-k)$-variate nonlinearity matched to the operator L. Finally, the regularization cost of (1.8) is $\sum_{n=1}^{N}|v_n| = \|\boldsymbol{v}\|_1$. The term $\boldsymbol{x} \mapsto c(\boldsymbol{x})$ that appears in (1.8) can be viewed as a (generalized) skip connection in neural network parlance. Note that (1.8) is exactly a sparse combination of multi-index models with learnable orientations $\mathbf{A}_n$, offsets $\boldsymbol{t}_n$, and fixed profiles specified by the multivariate nonlinearity ρ_{L} as well as a generalized translation network as in (1.4). Thus, if the data lie on a low-dimensional subspace (or union of subspaces), the neural architecture could automatically adapt to this structure and avoid the curse of dimensionality.

Connections to prior work. In section 4, we instantiate our results on the function-space optimality of neural architectures. First, we discuss implications of our representer theorem to the training and regularization of neural networks. These results provide new insight into the role of overparameterization and the use of *orthogonal weight normalization* in network architectures, which corresponds to the property that $\mathbf{A}_n\mathbf{A}_n^{\mathsf{T}} = \mathbf{I}_{d-k}$ in (1.8) [2, 24, 25, 33]. This property has been shown to increase the stability [2] and generalization properties [25] of neural architectures. We then discuss specific examples of neural architectures that are compatible with our framework. These architectures include the popular ReLU [19], the norm activation function/nonlinearity [23], and the radial basis functions found in the theory of thin-plate/polyharmonic splines [12, 71]. In particular, our theory provides a way to interpolate between the completely anisotropic atoms found in neural architectures with univariate nonlinearities $(k = (d-1))$ to the completely isotropic atoms found in the theory of sparse kernel expansions and radial basis functions $(k = 0)$ in a similar vein to how α-molecules interpolate between ridgelets (anisotropic) and wavelets (isotropic) [22].

In section 5, we discuss how the native space $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ can be viewed as an example of a reproducing kernel Banach space (RKBS) [5, 34, 62, 72] as well as an example of a variation space [4, 11, 31, 37, 56, 58, 57]. These are classical approaches used for the understanding of neural networks through the lens of functional analysis and approximation theory. Thus, any abstract result for RKBSs or variation spaces immediately applies to $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$.

2. Mathematical preliminaries and notation.

The Schwartz space of smooth and rapidly decreasing functions on $\mathbb{R}^d$ is denoted by $\mathcal{S}(\mathbb{R}^d)$. Its continuous dual is the space $\mathcal{S}'(\mathbb{R}^d)$ of tempered distributions. We let $L^p(\mathbb{R}^d)$ denote the Lebesgue space for $1 \leq p \leq \infty$ and define the weighted L^∞-space

$$L^\infty_{-\alpha}(\mathbb{R}^d) := \left\{ f : \mathbb{R}^d \to \mathbb{R} \text{ is measurable} : \|f\|_{L^\infty_{-\alpha}} := \operatorname*{ess\,sup}_{\boldsymbol{x}\in\mathbb{R}^d} |f(\boldsymbol{x})|(1+\|\boldsymbol{x}\|_2)^{-\alpha} < \infty \right\}. \tag{2.1}$$

This is the space of growth-restricted functions with rate $\alpha \in \mathbb{R}$. It is a Banach space that can be identified as the continuous dual of the weighted L^1-space

$$L^1_\alpha(\mathbb{R}^d) := \left\{ f : \mathbb{R}^d \to \mathbb{R} \text{ is measurable} : \|f\|_{L^1_\alpha} := \int_{\mathbb{R}^d} |f(\boldsymbol{x})|(1+\|\boldsymbol{x}\|_2)^{\alpha} \, \mathrm{d}\boldsymbol{x} < \infty \right\}. \tag{2.2}$$

The Banach space of continuous functions vanishing at $\pm\infty$ on $\mathbb{R}^d$ equipped with the L^∞-norm is denoted by $C_0(\mathbb{R}^d)$. By the Riesz–Markov–Kakutani representation theorem [14, Chapter 7], its continuous dual can be identified with the Banach space of finite Radon measures, denoted $\mathcal{M}(\mathbb{R}^d)$. Since $\mathcal{S}(\mathbb{R}^d)$ is dense in $C_0(\mathbb{R}^d)$, we have, by duality, that $\mathcal{M}(\mathbb{R}^d)$ is continuously embedded in $\mathcal{S}'(\mathbb{R}^d)$. Given a space $\mathcal{X}$ and a norm $\|\cdot\|$, the completion of $\mathcal{X}$ in $\|\cdot\|$ is a Banach space, denoted by $\overline{(\mathcal{X}, \|\cdot\|)}$. For example, we have, for $1 \leq p < \infty$, that $L^p(\mathbb{R}^d) = \overline{(\mathcal{S}(\mathbb{R}^d), \|\cdot\|_{L^p})}$, and $C_0(\mathbb{R}^d) = \overline{(\mathcal{S}(\mathbb{R}^d), \|\cdot\|_{L^\infty})}$.

The Fourier transform of $\varphi \in \mathcal{S}(\mathbb{R}^d)$ is defined as

$$\widehat{\varphi}(\boldsymbol{\xi}) := \mathscr{F}\{\varphi\}(\boldsymbol{\xi}) = \int_{\mathbb{R}^d} \varphi(\boldsymbol{x}) e^{-\mathrm{i}\boldsymbol{\xi}^\mathsf{T}\boldsymbol{x}} \, \mathrm{d}\boldsymbol{x}, \quad \boldsymbol{\xi} \in \mathbb{R}^d, \tag{2.3}$$

where $\mathrm{i}^2 = -1$. Consequently, the inverse Fourier transform of $\widehat{\varphi} \in \mathcal{S}(\mathbb{R}^d)$ is given by

$$\mathscr{F}^{-1}\{\widehat{\varphi}\}(\boldsymbol{x}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \widehat{\varphi}(\boldsymbol{\xi}) e^{\mathrm{i}\boldsymbol{\xi}^\mathsf{T}\boldsymbol{x}} \, \mathrm{d}\boldsymbol{\xi}, \quad \boldsymbol{x} \in \mathbb{R}^d. \tag{2.4}$$

These operators are extended to act on $\mathcal{S}'(\mathbb{R}^d)$ by duality.

Any continuous linear shift-invariant (LSI) operator $\mathrm{L} : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ is a convolution operator specified by a unique kernel $h \in \mathcal{S}'(\mathbb{R}^d)$ such that $\mathrm{L}\,\varphi = h * \varphi$. Such operators can also be specified in the Fourier domain by

$$\mathrm{L}\,\varphi = \mathscr{F}^{-1}\{\widehat{L}\widehat{\varphi}\}, \tag{2.5}$$

where $\widehat{L} \in \mathcal{S}'(\mathbb{R}^d)$ is the Fourier transform of the kernel $h \in \mathcal{S}'(\mathbb{R}^d)$. The tempered distribution h is the *impulse response* of L and the tempered distribution $\widehat{L}$ is the Fourier symbol or *frequency response* of L. We shall generally use upright, roman letters for LSI operators and use the italic variant with a hat to denote its frequency response.

2.1. The k-plane transform.

We are going to adopt the parameterization of the k-plane transform from [47, section 4]. There, the space of k-planes is parameterized by the Cartesian product of the Stiefel manifold with $\mathbb{R}^{d-k}$, where k is an integer such that $1 \leq k < d$. Let

$$V_{d-k}(\mathbb{R}^d) := \{\mathbf{A} \in \mathbb{R}^{(d-k)\times d} : \mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{I}_{d-k}\} \tag{2.6}$$

denote the Stiefel manifold. Then, the k-plane transform of $\varphi \in \mathcal{S}(\mathbb{R}^d)$ is defined as

$$\mathscr{R}_k\{\varphi\}(\mathbf{A}, \boldsymbol{t}) = \int_{\mathbb{R}^d} \varphi(\boldsymbol{x})\delta(\mathbf{A}\boldsymbol{x} - \boldsymbol{t})\,\mathrm{d}\boldsymbol{x}, \quad (\mathbf{A}, \boldsymbol{t}) \in (V_{d-k}(\mathbb{R}^d), \mathbb{R}^{d-k}), \tag{2.7}$$

where $\delta \in \mathcal{S}'(\mathbb{R}^{d-k})$ is the $(d-k)$-variate Dirac impulse[3] and the integral is understood as the action of $\delta(\mathbf{A}(\cdot) - \boldsymbol{t}) \in \mathcal{S}'(\mathbb{R}^d)$ on $\varphi \in \mathcal{S}(\mathbb{R}^d)$. Further, the dual transform (often called the "backprojection") of $g \in L^\infty(V_{d-k}(\mathbb{R}^d) \times \mathbb{R}^{d-k})$ is given by

$$\mathscr{R}_k^*\{g\}(\boldsymbol{x}) = \int_{V_{d-k}(\mathbb{R}^d)} g(\mathbf{A}, \mathbf{A}\boldsymbol{x})\,\mathrm{d}\mathbf{A}, \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{2.8}$$

where $\mathrm{d}\mathbf{A}$ denotes integration against the Haar measure of $V_{d-k}(\mathbb{R}^d)$. Since we impose that the rows of $\mathbf{A}$ are orthonormal, we have that $(\mathbf{A}, \boldsymbol{t})$ and $(\mathbf{U}\mathbf{A}, \mathbf{U}\boldsymbol{t})$ define the same k-plane, for any orthogonal transformation $\mathbf{U} \in \mathrm{O}_{d-k}(\mathbb{R})$ (the orthogonal group in dimension $(d-k)$). The main advantage of the proposed parameterization is that it will allow us to identify the symmetries of k-plane domain as "isotropic" symmetries.

Letting $\Xi_k := V_{d-k}(\mathbb{R}^d) \times \mathbb{R}^{d-k}$ denote the k-plane domain, we define the space of Schwartz functions on Ξ_k, denoted by $\mathcal{S}(\Xi_k)$, as the space of smooth functions that are rapidly decreasing in the $\boldsymbol{t} \in \mathbb{R}^{d-k}$ variable [20]. More specifically, we have that $\mathcal{S}(\Xi_k) = C^\infty(V_{d-k}(\mathbb{R}^d)) \widehat{\otimes} \mathcal{S}(\mathbb{R}^{d-k})$, where $\widehat{\otimes}$ denotes the topological tensor product, which is the completion of the algebraic tensor product with respect to the projective topology [64, Chapter 43]. We state in Proposition 2.1 a classical result regarding the continuity and invertibility of the k-plane transform.

Proposition 2.1 (see [16, 20, 27, 47, 51, 59, 60]). *The operator $\mathscr{R}_k$ continuously maps $\mathcal{S}(\mathbb{R}^d)$ into $\mathcal{S}(\Xi_k)$. Moreover,*

$$\mathscr{R}_k^* \mathrm{K}_{d-k} \mathscr{R}_k = c_{d,k}(-\Delta)^{\frac{k}{2}} \mathscr{R}_k^* \mathscr{R}_k = c_{d,k} \mathscr{R}_k^* \mathscr{R}_k (-\Delta)^{\frac{k}{2}} = \mathrm{Id} \tag{2.9}$$

on $\mathcal{S}(\mathbb{R}^d)$ with

$$c_{d,k} = \frac{1}{(2\pi)^k} \frac{|\mathbb{S}^{k-1}|}{|\mathbb{S}^{d-k-1}|} \frac{1}{\prod_{n=k}^{d-1} |\mathbb{S}^{n-1}|}, \tag{2.10}$$

where $|\cdot|$ denotes the surface area. The underlying operators are the d-variate Laplacian operator Δ and the filtering operator[4] $\mathrm{K}_{d-k} = c_{d,k}(-\Delta_{d-k})^{k/2}$, where Δ_{d-k} denotes the $(d-k)$-variate Laplacian applied to the $\boldsymbol{t} \in \mathbb{R}^{d-k}$ variable. The filtering operator is equivalently specified by the frequency response $\widehat{K}_{d-k}(\boldsymbol{\omega}) = c_{d,k}\|\boldsymbol{\omega}\|_2^k$, $\boldsymbol{\omega} \in \mathbb{R}^{d-k}$.

The k-plane transform has tight connections with the Fourier transform. This is summarized in the so-called *Fourier slice theorem.*

[3]The distribution $\delta \in \mathcal{S}'(\mathbb{R}^{d-k})$ is such that $\langle \delta, \phi \rangle = \phi(\mathbf{0})$ for all $\phi \in \mathcal{S}(\mathbb{R}^{d-k})$.

[4]In CT, this filter is referred to as the backprojection filter found in the filtered backprojection algorithm for CT image reconstruction.

Proposition 2.2 ([47, Corollary 7.5]). *Given* $\varphi \in \mathcal{S}(\mathbb{R}^d)$, *we have that*

$$\widehat{\mathscr{R}_k\{\varphi\}(\mathbf{A},\cdot)}(\boldsymbol{\omega}) = \widehat{\varphi}(\mathbf{A}^\mathsf{T}\boldsymbol{\omega}), \quad \boldsymbol{\omega} \in \mathbb{R}^{d-k}, \mathbf{A} \in V_{d-k}(\mathbb{R}^d), \tag{2.11}$$

where the Fourier transform on the left-hand side is the $(d-k)$*-variate transform and the Fourier transform on the right-hand side is the* d*-variate Fourier transform.*

Remark 2.3. The Fourier slice theorem can be extended to apply to members of $\mathcal{S}'(\mathbb{R}^d)$ as long as some additional care is taken regarding in what sense the equality in (2.11) holds [47, Theorem 7.7].

Let $\mathcal{S}_k := \mathscr{R}_k\left(\mathcal{S}(\mathbb{R}^d)\right)$ denote the range of the k-plane transform. The range $\mathcal{S}_k$ is a strict subspace of $\mathcal{S}(\Xi_k)$ that satisfies certain consistency conditions (see [36, Chapter 4] for a detailed discussion and references on this matter). We have the following additional result regarding the continuity and invertibility of the k-plane transform.

Proposition 2.4 ([47, Corollary 5.3]). *The operator* $\mathscr{R}_k : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}_k$ *is a homeomorphism with inverse* $\mathscr{R}_k^{-1} = \mathscr{R}_k^* \mathrm{K}_k : \mathcal{S}_k \to \mathcal{S}(\mathbb{R}^d)$.

Proposition 2.4 motivates the following *distributional extension* of the k-plane transform and related operators.

Definition 2.5 ([47, Definition 6.1]).

1. *The* distributional k-plane transform

$$\mathscr{R}_k : \mathcal{S}'(\mathbb{R}^d) \to (\mathrm{K}_{d-k}\mathscr{R}_k(\mathcal{S}(\mathbb{R}^d)))' \tag{2.12}$$

is defined to be the dual map of the homeomorphism $\mathscr{R}_k^* : \mathrm{K}_{d-k}\mathscr{R}_k\left(\mathcal{S}(\mathbb{R}^d)\right) \to \mathcal{S}(\mathbb{R}^d)$.

2. *The* distributional filtered k-plane transform

$$\mathrm{K}_{d-k}\mathscr{R}_k : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}_k' \tag{2.13}$$

is defined to be the dual map of the homeomorphism $\mathscr{R}_k^* \mathrm{K}_{d-k} : \mathcal{S}_k \to \mathcal{S}(\mathbb{R}^d)$.

3. *The* distributional backprojection

$$\mathscr{R}_k^* : \mathcal{S}_k' \to \mathcal{S}'(\mathbb{R}^d) \tag{2.14}$$

is defined to be the dual map of the homeomorphism $\mathscr{R}_k : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}_k$.

4. *The* extended distributional backprojection

$$\mathscr{R}_k^* : \mathcal{S}'(\Xi_k) \to \mathcal{S}'(\mathbb{R}^d) \tag{2.15}$$

is defined to be the dual map of the continuous operator $\mathscr{R}_k : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\Xi_k)$, *which is well-defined since* $\mathcal{S}_k$ *is continuously embedded in* $\mathcal{S}(\Xi_k)$.

Based on these definitions, we state in Theorem 2.6 a result on the invertibility of the filtered k-plane transform on $\mathcal{S}'(\mathbb{R}^d)$, which is the dual of Propositions 2.1 and 2.4.

Theorem 2.6. *It holds that* $\mathscr{R}_k^* \mathrm{K}_{d-k} \mathscr{R}_k = \mathrm{Id}$ *on* $\mathcal{S}'(\mathbb{R}^d)$. *Moreover, the filtered* k*-plane transform* $\mathrm{K}_{d-k}\mathscr{R}_k : \mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'_k$ *is a homeomorphism with the inverse given by the backprojection* $(\mathrm{K}_{d-k}\mathscr{R}_k)^{-1} = \mathscr{R}_k^* : \mathcal{S}'_k \to \mathcal{S}'(\mathbb{R}^d)$.

While this setup provides an attractive formulation to handle the distributional extension of the k-plane transform and its dual, it turns out that the distribution spaces $\mathcal{S}'_k$ and $\left(\mathrm{K}_{d-k}\mathscr{R}_k\left(\mathcal{S}(\mathbb{R}^d)\right)\right)'$ are actually *equivalence classes* of distributions [47]. Luckily, by working with certain Banach subspaces that continuously embed into these distribution spaces, one can identify a concrete member of the equivalence classes via continuous projection operators. To this end, the results of this paper hinge on a nontrivial result of [47] regarding the invertibility of the distributional dual k-plane transform on the space of isotropic Radon measures.

Consider the operator

$$\mathrm{P}_{\mathrm{iso}}\{g\}(\mathbf{A}, \boldsymbol{t}) = \fint_{\mathrm{O}_{d-k}(\mathbb{R})} g(\mathbf{U}\mathbf{A}, \mathbf{U}\boldsymbol{t}) \, \mathrm{d}\sigma(\mathbf{U}), \tag{2.16}$$

where $\fint$ denotes the average integral and σ is the Haar measure on $\mathrm{O}_{d-k}(\mathbb{R})$. This is a well-defined operator that maps $C_0(\Xi_k) \to C_0(\Xi_k)$ and, in particular, is the self-adjoint continuous projector which extracts the isotropic part of a function [47, equation (8.10)]. By the Riesz–Markov–Kakutani representation theorem, we can extend $\mathrm{P}_{\mathrm{iso}} = \mathrm{P}_{\mathrm{iso}}^*$ by duality to act on $\mathcal{M}(\Xi_k) = (C_0(\Xi_k))'$, the Banach space of finite Radon measures on Ξ_k. To this end, define

$$\mathcal{M}_{\mathrm{iso}}(\Xi_k) := \mathrm{P}_{\mathrm{iso}}\left(\mathcal{M}(\Xi_k)\right), \tag{2.17}$$

the Banach subspace of isotropic finite Radon measures on Ξ_k. This Banach subspace is complemented in $\mathcal{M}(\Xi_k)$ [47, Theorem 8.2] and so

$$\mathcal{M}(\Xi_k) = \mathcal{M}_{\mathrm{iso}}(\Xi_k) \oplus (\mathcal{M}_{\mathrm{iso}}(\Xi_k))^{\mathsf{c}} \tag{2.18}$$

with $(\mathcal{M}_{\mathrm{iso}}(\Xi_k))^{\mathsf{c}} := (\mathrm{Id} - \mathrm{P}_{\mathrm{iso}})\left(\mathcal{M}(\Xi_k)\right)$. We also note that

$$\mathcal{M}_{\mathrm{iso}}(\Xi_k) = (C_{0,\mathrm{iso}}(\Xi_k))', \tag{2.19}$$

where

$$C_{0,\mathrm{iso}}(\Xi_k) := \mathrm{P}_{\mathrm{iso}}\left(C_0(\Xi_k)\right) = \overline{(\mathcal{S}_k, \|\cdot\|_{L^\infty})}, \tag{2.20}$$

where the last equality is from [47, equation (8.12)].

Proposition 2.7 (see [47, Theorems 8.1 and 8.2 and Corollary 8.3]). *The Banach space* $\mathcal{M}_{\mathrm{iso}}(\Xi_k)$ *continuously embeds into* $\mathcal{S}'_k$. *Further, the distributional backprojection operator* $\mathscr{R}_k^*$ *is invertible on* $\mathcal{M}_{\mathrm{iso}}(\Xi_k)$, *so that*

$$\mathrm{K}_{d-k}\mathscr{R}_k\mathscr{R}_k^* = \mathrm{Id} \ \ \text{on } \mathcal{M}_{\mathrm{iso}}(\Xi_k). \tag{2.21}$$

Furthermore, the null space of $\mathscr{R}_k^* : \mathcal{M}(\Xi_k) \to \mathcal{S}'(\mathbb{R}^d)$ *is* $(\mathcal{M}_{\mathrm{iso}}(\Xi_k))^{\mathsf{c}}$ *and so* $\mathscr{R}_k^*\left(\mathcal{M}(\Xi_k)\right) = \mathscr{R}_k^*\left(\mathcal{M}_{\mathrm{iso}}(\Xi_k)\right)$.

2.1.1. The case $k=0$. When $k=0$, the k-plane transform of $\varphi \in \mathcal{S}(\mathbb{R}^d)$ becomes

$$\mathscr{R}_0\{\varphi\}(\mathbf{A}, \boldsymbol{t}) = \int_{\mathbb{R}^d} \varphi(\boldsymbol{x})\delta(\mathbf{A}\boldsymbol{x} - \boldsymbol{t})\,\mathrm{d}\boldsymbol{x} = \varphi(\mathbf{A}^\mathsf{T}\boldsymbol{t}). \tag{2.22}$$

Indeed, when $k=0$, the matrix $\mathbf{A} \in \mathbb{R}^{d\times d}$ is now an orthogonal matrix (i.e., $\mathbf{A}^\mathsf{T}\mathbf{A} = \mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{I}_d$) and $\boldsymbol{t} \in \mathbb{R}^d$. The last equality of the above integral follows from the change of variables $\boldsymbol{y} = (\mathbf{A}\boldsymbol{x} - \boldsymbol{t})$. In this case, it is clear that the range $\mathcal{S}_0$ is exactly the closed subspace of isotropic functions in $\mathcal{S}(\Xi_0)$, denoted by $\mathcal{S}_{\mathrm{iso}}(\Xi_0)$. The fundamental results discussed previously trivially hold in this limit setting. In particular, it is easy to check that

$$c_{d,0}\,\mathscr{R}_0^*\mathscr{R}_0 = \mathrm{Id} \text{ on } \mathcal{S}(\mathbb{R}^d), \tag{2.23}$$

which is the same statement as $\mathscr{R}_0^*\,\mathrm{K}_d\,\mathscr{R}_0 = \mathrm{Id}$ since $\mathrm{K}_d = c_{d,0}\,\mathrm{Id}$ (Proposition 2.1). Therefore, in the remainder of the paper, when working with the k-plane transform for general k such that $0 \le k < d$, we shall not treat separately the case $k=0$. We do warn the reader, however, to be aware that the underlying mathematics of the $(k=0)$-plane transform is much simpler than that of $1 \le k < d$.

2.2. Polynomial spaces and related projectors. The null space of the regularizer in the learning problem (1.7) is the space of polynomials of degree n_L (see Lemma 3.5), which depends on the operator L. This space is denoted $\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)$. To this end, we will be interested in working with a biorthogonal system for this null space. We will use the biorthogonal system from [67, section 2.2].

Remark 2.8. When the null space of L is trivial, we make the identifications $n_\mathrm{L} = (-1)$ and $\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d) = \{0\}$, noting that any sum taken from $n=0$ to (-1) is understood as 0.

The space $\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)$ is spanned by the monomial/Taylor basis

$$m_{\boldsymbol{n}}(\boldsymbol{x}) = \frac{\boldsymbol{x}^{\boldsymbol{n}}}{\boldsymbol{n}!}, \tag{2.24}$$

where $\boldsymbol{n} = (n_1, \dots, n_d) \in \mathbb{N}_0^d$ is a multi-index. Accordingly, we have that

$$\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d) = \left\{ \sum_{|\boldsymbol{n}|\le n_\mathrm{L}} b_{\boldsymbol{n}} m_{\boldsymbol{n}} : b_{\boldsymbol{n}} \in \mathbb{R} \right\} \subset \mathcal{S}'(\mathbb{R}^d). \tag{2.25}$$

Importantly, $\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)$ forms a finite-dimensional Banach subspace of $\mathcal{S}'(\mathbb{R}^d)$. Since all norms are equivalent in finite dimensions, the exact choice does not matter, but, for concreteness, we equip $\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)$ with the norm

$$\|p\|_{\mathcal{P}_{n_\mathrm{L}}} := \|(b_{\boldsymbol{n}})_{|\boldsymbol{n}|\le n_\mathrm{L}}\|_2, \tag{2.26}$$

where the $b_{\boldsymbol{n}}$ are the coefficients of p in the monomial/Taylor basis. While the dual space $\mathcal{P}'_{n_\mathrm{L}}(\mathbb{R}^d)$ is also finite-dimensional, its "abstract" elements are actually equivalence classes in $\mathcal{S}'(\mathbb{R}^d)$. Following the approach from [67, section 2.2], we identify every dual element $p^* \in \mathcal{P}'_{n_\mathrm{L}}(\mathbb{R}^d)$ as a function in $\mathcal{S}(\mathbb{R}^d)$ by working with a concrete dual basis $\{m^*_{\boldsymbol{n}}\}_{|\boldsymbol{n}|\le n_\mathrm{L}}$ that satisfies the biorthogonality property $\langle m^*_{\boldsymbol{n}}, m_{\boldsymbol{n}'}\rangle = \delta[\boldsymbol{n}-\boldsymbol{n}']$, where $\delta[\cdot]$ is the Kronecker impulse which takes the value 1 when its input is 0 and 0 otherwise. Our specific choice is

$$m^*_{\boldsymbol{n}} := (-1)^{|\boldsymbol{k}|}\partial^{\boldsymbol{n}}\kappa_{\text{iso}} \in \mathcal{S}(\mathbb{R}^d), \tag{2.27}$$

where $\kappa_{\text{iso}} \in \mathcal{S}(\mathbb{R}^d)$ is an isotropic function constructed in [67, Lemma 1]. Its frequency response is such that $\widehat{\kappa}_{\text{iso}}(\boldsymbol{\xi}) = \widehat{\kappa}_{\text{rad}}(\|\boldsymbol{\xi}\|_2)$, where the radial frequency profile $\widehat{\kappa}_{\text{rad}} \in \mathcal{S}(\mathbb{R})$ satisfies $0 \le \widehat{\kappa}_{\text{rad}} \le 1$ and, for $|\omega| \ge 1$, $\widehat{\kappa}_{\text{rad}}(\omega) = 0$.

The biorthogonal system $\{(m^*_{\boldsymbol{n}}, m_{\boldsymbol{n}})\}_{|\boldsymbol{n}|\le n_{\mathrm{L}}}$ allows us to define the projection onto $\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$ by the operator

$$\mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)}\{f\} = \sum_{\boldsymbol{n}\le n_{\mathrm{L}}} \langle m^*_{\boldsymbol{n}}, f\rangle m_{\boldsymbol{n}}. \tag{2.28}$$

This projector continuously maps $\mathcal{S}'(\mathbb{R}^d) \to \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$ (because $m^*_{\boldsymbol{n}} \in \mathcal{S}(\mathbb{R}^d)$). Moreover, since $L^\infty_{-n_{\mathrm{L}}}(\mathbb{R}^d)$ continuously embeds into $\mathcal{S}'(\mathbb{R}^d)$, the restricted operator

$$\mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)} : L^\infty_{-n_{\mathrm{L}}}(\mathbb{R}^d) \to \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d) \tag{2.29}$$

is continuous as well. Last, the finite dimensionality of $\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$ ensures that $\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$ is complemented in $L^\infty_{-n_{\mathrm{L}}}(\mathbb{R}^d)$ [52, Lemma 4.21]. Thus, the complementary projector

$$(\mathrm{Id} - \mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)}) : L^\infty_{-n_{\mathrm{L}}}(\mathbb{R}^d) \to L^\infty_{-n_{\mathrm{L}}}(\mathbb{R}^d) \tag{2.30}$$

is guaranteed to exist and be continuous.

3. Main results.

We first define the class of operators that are admissible for the learning problem in (1.7). These operators form a subclass of the so-called *spline-admissible* operators [69, Definition 1].

Definition 3.1. *A continuous LSI operator* $\mathrm{L} : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ *is said to be* k-plane-admissible *with a polynomial null space of degree* n_{L} *if*

1. *its adjoint* L^* *is a continuous injection that maps* $\mathcal{S}(\mathbb{R}^d)$ *into* $L^1_{n_{\mathrm{L}}}(\mathbb{R}^d)$*;*
2. *it is isotropic in the sense that its frequency response is continuous and satisfies* $\widehat{L}(\boldsymbol{\xi}) = \widehat{L}_{\text{rad}}(\|\boldsymbol{\xi}\|_2)$ *for some continuous univariate radial frequency profile* $\widehat{L}_{\text{rad}} : \mathbb{R} \to \mathbb{R}$*;*
3. *the radial frequency profile* $\widehat{L}_{\text{rad}}$ *does not vanish over* $\mathbb{R}$*, except for a zero of order* $\gamma_{\mathrm{L}} \in (n_{\mathrm{L}}, n_{\mathrm{L}}+1]$ *at the origin, so that there exists a constant* $C > 0$ *satisfying*

$$\lim_{\omega\to 0} \frac{\widehat{L}_{\text{rad}}(\omega)}{|\omega|^{\gamma_{\mathrm{L}}}} = C; \tag{3.1}$$

4. *there exist* $\gamma'_{\mathrm{L}} > (d-k)$, $C' > 0$, *and* $R > 0$ *such that*

$$|\widehat{L}_{\text{rad}}(\omega)| \ge C'|\omega|^{\gamma'_{\mathrm{L}}} \tag{3.2}$$

for all $|\omega| > R$.

Remark 3.2. Item 3 implies that the extension by duality $\mathrm{L} : L^\infty_{-n_\mathrm{L}}(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ of a k-plane-admissible operator annihilates polynomials of degree at most n_L.

Remark 3.3. Item 4 guarantees that the order of the operator is sufficiently high to ensure that the point evaluation functional is well-defined on the native space $\mathcal{M}^k_\mathrm{L}(\mathbb{R}^d)$. In particular, this condition is reminiscent of the condition $s > (d-k)$ on the smoothness index of an L^1-Sobolev space defined on (subsets of) $\mathbb{R}^{d-k}$ to ensure the continuity of its members. Indeed, when $s > (d-k)$, the Sobolev embedding theorem guarantees that the L^1-Sobolev space of order s embeds into the space of continuous functions. In our setting, the condition $\gamma'_\mathrm{L} > (d-k)$ plays the role of the smoothness index. This property is key to establishing the existence of solutions to the learning problem (1.7).

3.1. Native spaces. The primary technical contribution of this paper is the careful treatment of the distributional extension and (pseudo)invertibility of the operator $\mathrm{L}_{\mathscr{R}_k} := \mathrm{K}_{d-k} \mathscr{R}_k \mathrm{L}$, where L is a k-plane-admissible operator (Definition 3.1). This in turn allows us to define our native (Banach) spaces. Indeed, for the definition of the native space in (1.5) to be coherent, the action of $\mathrm{L}_{\mathscr{R}_k}$ on $L^\infty_{-n_\mathrm{L}}(\mathbb{R}^d)$ must be well-defined. Furthermore, as we shall see in Lemma 3.5, the null space of $\mathrm{L}_{\mathscr{R}_k}$ is exactly the space of polynomials of degree at most n_L. We then use a technique from spline theory to "factor out" the null space of $\mathrm{L}_{\mathscr{R}_k}$ and identify the subspace of $\mathcal{M}^k_\mathrm{L}(\mathbb{R}^d)$ on which $\mathrm{L}_{\mathscr{R}_k}$ is invertible [10, 12]. This allows us to identify $\mathcal{M}^k_\mathrm{L}(\mathbb{R}^d)$ as the direct sum of two Banach spaces. Thus, it forms a Banach space when equipped with the composite norm.

Lemma 3.4. *Let* L *be a k-plane-admissible operator in the sense of Definition* 3.1. *Then, the operator* $\mathrm{L}_{\mathscr{R}_k} = \mathrm{K}_{d-k} \mathscr{R}_k \mathrm{L}$ *continuously maps* $L^\infty_{-n_\mathrm{L}}(\mathbb{R}^d) \to \mathcal{S}'_k$.

Proof. It suffices to prove that the adjoint operator $\mathrm{L}^*_{\mathscr{R}_k} = \mathrm{L}^* \mathscr{R}^*_k \mathrm{K}_{d-k}$ continuously maps $\mathcal{S}_k \to L^1_{n_\mathrm{L}}(\mathbb{R}^d)$. The result then follows by duality. Since $\mathscr{R}^*_k \mathrm{K}_{d-k} : \mathcal{S}_k \to \mathcal{S}(\mathbb{R}^d)$ is a homeomorphism (Proposition 2.4), the lemma follows from item 1 in Definition 3.1. ∎

Lemma 3.5. *Let* L *be a k-plane-admissible operator in the sense of Definition* 3.1. *Then, the null space of the operator* $\mathrm{L}_{\mathscr{R}_k} = \mathrm{K}_{d-k} \mathscr{R}_k \mathrm{L} : L^\infty_{-n_\mathrm{L}}(\mathbb{R}^d) \to \mathcal{S}'_k$ *is the space of polynomials of degree at most* n_L *on* $\mathbb{R}^d$, *denoted by* $\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)$.

Proof. Let $\mathcal{N}(\mathrm{L}_{\mathscr{R}_k})$ denote the null space of $\mathrm{L}_{\mathscr{R}_k}$. More precisely,

$$\mathcal{N}(\mathrm{L}_{\mathscr{R}_k}) = \{f \in L^\infty_{-n_\mathrm{L}}(\mathbb{R}^d) : \mathrm{L}_{\mathscr{R}_k}\{f\} = 0\}. \tag{3.3}$$

From Remark 3.2, L annihilates polynomials of degree at most n_L. This implies that $\mathcal{N}(\mathrm{L}_{\mathscr{R}_k}) \supset \mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)$. However, given $f \in L^\infty_{-n_\mathrm{L}}(\mathbb{R}^d) \subset \mathcal{S}'(\mathbb{R}^d)$, from the Fourier slice theorem,

$$\widehat{\mathrm{L}_{\mathscr{R}_k}\{f\}(\mathbf{A}, \cdot)}(\boldsymbol{\omega}) = c_{d,k} \|\boldsymbol{\omega}\|_2^k \widehat{L}(\mathbf{A}^\mathsf{T}\boldsymbol{\omega}) \widehat{f}(\mathbf{A}^\mathsf{T}\boldsymbol{\omega}) = c_{d,k} \|\boldsymbol{\omega}\|_2^k \widehat{L}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2) \widehat{f}(\mathbf{A}^\mathsf{T}\boldsymbol{\omega}). \tag{3.4}$$

Note that the product in the last equality must be a well-defined tempered distribution (in the $\boldsymbol{\omega}$ variable) since $\mathrm{L}_{\mathscr{R}_k} : L^\infty_{-n_\mathrm{L}}(\mathbb{R}^d) \to \mathcal{S}'_k$ is well-defined by Lemma 3.4. Due to the vanishing property in item 3 of Definition 3.1, this quantity is 0 if and only if $\widehat{f}$ is supported only at $\mathbf{0}$, in which case f is a polynomial. Combined with the growth restriction $f \in L^\infty_{-n_\mathrm{L}}(\mathbb{R}^d)$, we have that f must be a polynomial of degree at most n_L. Therefore, $\mathcal{N}(\mathrm{L}_{\mathscr{R}_k}) \subset \mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)$. ∎

From Lemmas 3.4 and 3.5, we deduce that the native space in (1.5) is well-defined and that the null space of $\mathrm{L}_{\mathscr{R}_k}$ is the finite-dimensional Banach space $\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$. The next two technical theorems (Theorems 3.6 and 3.7) establish the Banach structure of the native space.

Theorem 3.6. *Let* L *be a* k*-plane-admissible operator in the sense of Definition* 3.1*. Then, the operator* $\mathrm{L}_{\mathscr{R}_k} = \mathrm{K}_{d-k}\mathscr{R}_k\mathrm{L}$ *maps* $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d) \to \mathcal{M}_{\mathrm{iso}}(\Xi_k)$*. Furthermore, there exists an operator* $\mathrm{L}_{\mathscr{R}_k}^{\dagger}$ *that continuously maps* $\mathcal{M}_{\mathrm{iso}}(\Xi_k) \to L^{\infty}_{-n_{\mathrm{L}}}(\mathbb{R}^d) \subset \mathcal{S}'(\mathbb{R}^d)$ *and is such that*

$$\mathrm{L}_{\mathscr{R}_k}\mathrm{L}_{\mathscr{R}_k}^{\dagger}\{\mu\} = \mu \text{ for all } \mu \in \mathcal{M}_{\mathrm{iso}}(\Xi_k), \tag{3.5}$$

$$\mathrm{L}_{\mathscr{R}_k}^{\dagger}\mathrm{L}_{\mathscr{R}_k}\{f\} = (\mathrm{Id} - \mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)})\{f\} \text{ for all } f \in \mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d). \tag{3.6}$$

This operator is realized by

$$\mathrm{L}_{\mathscr{R}_k}^{\dagger} = (\mathrm{Id} - \mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)})\,\mathrm{L}^{-1}\mathscr{R}_k^*, \tag{3.7}$$

where L^{-1} *is the operator specified by the frequency response* $\boldsymbol{\xi} \mapsto 1/\widehat{L}(\boldsymbol{\xi})$*. Furthermore,* $\mathrm{L}_{\mathscr{R}_k}^{\dagger}$ *is an integral operator specified by the kernel* $(\boldsymbol{x}, (\mathbf{A}, \boldsymbol{t})) \mapsto g_{\mathbf{A},\boldsymbol{t}}(\boldsymbol{x})$ *that takes the form*

$$g_{\mathbf{A},\boldsymbol{t}}(\boldsymbol{x}) = \rho_{\mathrm{L}}(\mathbf{A}\boldsymbol{x} - \boldsymbol{t}) - \sum_{|\boldsymbol{k}| \le n_{\mathrm{L}}} \langle m_{\boldsymbol{k}}^*, \rho_{\mathrm{L}}(\mathbf{A}(\cdot) - \boldsymbol{t})\rangle m_{\boldsymbol{k}}(\boldsymbol{x}), \tag{3.8}$$

where $\rho_{\mathrm{L}} = \mathscr{F}_{d-k}^{-1}\{1/\widehat{L}_{\mathrm{rad}}(\|\cdot\|_2)\}$ *and* $\{(m_{\boldsymbol{n}}^*, m_{\boldsymbol{n}})\}_{|\boldsymbol{n}| \le n_{\mathrm{L}}}$ *are specified in subsection* 2.2*, and where* $\mathscr{F}_{d-k}^{-1}$ *denotes the* $(d-k)$*-variate inverse Fourier transform. This kernel satisfies the stability/continuity bound*

$$\sup_{\substack{\boldsymbol{x} \in \mathbb{R}^d \\ (\mathbf{A},\boldsymbol{t}) \in \Xi_k}} |g_{\mathbf{A},\boldsymbol{t}}(\boldsymbol{x})|(1 + \|\boldsymbol{x}\|_2)^{-n_{\mathrm{L}}} < \infty \tag{3.9}$$

with

$$(\mathbf{A}, \boldsymbol{t}) \mapsto g_{\mathbf{A},\boldsymbol{t}}(\boldsymbol{x}_0) \in C_{0,\mathrm{iso}}(\Xi_k) \tag{3.10}$$

for any fixed $\boldsymbol{x}_0 \in \mathbb{R}^d$*. Thus, for* $\mu \in \mathcal{M}_{\mathrm{iso}}(\Xi_k)$*,*

$$\mathrm{L}_{\mathscr{R}_k}^{\dagger}\{\mu\}(\boldsymbol{x}) = \int_{\Xi_k} g_{\mathbf{A},\boldsymbol{t}}(\boldsymbol{x})\,\mathrm{d}\mu(\mathbf{A}, \boldsymbol{t}), \quad \boldsymbol{x} \in \mathbb{R}^d. \tag{3.11}$$

Theorem 3.7. *Consider the setting of Theorem* 3.6*. Then, the following hold.*

1. *The range space* $\mathcal{V} := \mathrm{L}_{\mathscr{R}_k}^{\dagger}(\mathcal{M}_{\mathrm{iso}}(\Xi_k))$ *is a Banach space when equipped with the norm*

$$\|f\|_{\mathcal{V}} := \|\mathrm{L}_{\mathscr{R}_k} f\|_{\mathcal{M}}. \tag{3.12}$$

   This Banach space is isometrically isomorphic to $\mathcal{M}_{\mathrm{iso}}(\Xi_k)$*.*

2. *The native space* $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ *is decomposable as the direct sum of Banach spaces*

$$\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d) = \mathcal{V} \oplus \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d) = \mathrm{L}_{\mathscr{R}_k}^{\dagger}(\mathcal{M}_{\mathrm{iso}}(\Xi_k)) \oplus \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d). \tag{3.13}$$

3. *The native space* $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ *forms a bona fide Banach space when equipped with the norm*

$$\|f\|_{\mathcal{M}_{\mathrm{L}}^k} := \|\mathrm{L}_{\mathscr{R}_k} f\|_{\mathcal{M}} + \|\mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)} f\|_{\mathcal{P}_{n_{\mathrm{L}}}}. \tag{3.14}$$

Furthermore, $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ *is isometrically isomorphic to* $\mathcal{M}_{\mathrm{iso}}(\Xi_k) \times \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$ *via the map*

$$f = \mathrm{L}_{\mathscr{R}_k}^{\dagger}\{u\} + p \mapsto (u, p), \tag{3.15}$$

where $u = \mathrm{L}_{\mathscr{R}_k}\{f\}$ *and* $p = \mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)} f$.

4. *For any* $\boldsymbol{x}_0 \in \mathbb{R}^d$, *the shifted Dirac impulse (point evaluation functional)* $\delta(\cdot - \boldsymbol{x}_0) : f \mapsto f(\boldsymbol{x}_0)$ *is weak*-continuous*[5] *on* $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$.

The proofs of Theorems 3.6 and 3.7 appear in Appendices A and B, respectively.

3.2. Optimality of neural architectures with multivariate nonlinearities. Having established the properties of the native space $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ in subsection 3.1, we can now prove our main theorem regarding the function-space optimality of neural architectures with multivariate nonlinearities.

Theorem 3.8. *Let* $\mathcal{L}(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ *be convex, coercive, and lower-semicontinuous in its second argument and let* L *be a k-plane-admissible operator in the sense of Definition* 3.1. *Then, for any finite data set* $\{(\boldsymbol{x}_m, y_m)\}_{m=1}^M \subset \mathbb{R}^d \times \mathbb{R}$ *for which the data-fitting problem is well-posed*[6] *over* $\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$, *the solution set to the data-fitting variational problem*

$$S := \operatorname*{arg\,min}_{f \in \mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)} \sum_{m=1}^M \mathcal{L}(y_m, f(\boldsymbol{x}_m)) + \lambda \|\mathrm{K}_{d-k} \mathscr{R}_k \mathrm{L} f\|_{\mathcal{M}} \tag{3.16}$$

is nonempty, convex, and weak-compact. If* $\mathcal{L}(\cdot,\cdot)$ *is strictly convex (or if it imposes the equality* $y_m = f(\boldsymbol{x}_m)$ *for* $m = 1, \ldots, M$*), then* S *is the weak*-closure of the convex hull of its extreme points, which can all be expressed as*

$$f_{\mathrm{extreme}}(\boldsymbol{x}) = c(\boldsymbol{x}) + \sum_{n=1}^N v_n \, \rho_{\mathrm{L}}(\mathbf{A}_n \boldsymbol{x} - \boldsymbol{t}_n), \tag{3.17}$$

where the number N *of atoms satisfies* $N \le (M - \dim \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d))$, *and, for* $n = 1, \ldots, N$, $v_n \in \mathbb{R} \setminus \{0\}$, $\mathbf{A}_n \in V_{d-k}(\mathbb{R}^d)$, *and* $\boldsymbol{t}_n \in \mathbb{R}^{d-k}$. *The function* $c \in \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$ *is a polynomial of degree at most* n_{L} *and* $\rho_{\mathrm{L}} : \mathbb{R}^{d-k} \to \mathbb{R}^d$ *is a* $(d-k)$*-variate nonlinearity given by* $\rho_{\mathrm{L}} = \mathscr{F}_{d-k}^{-1}\{1/\widehat{L}_{\mathrm{rad}}(\|\cdot\|_2)\}$, *where* $\mathscr{F}_{d-k}^{-1}$ *is the* $(d-k)$*-variate inverse Fourier transform. Finally, the regularization cost, which is common to all solutions, is*

$$\|\mathrm{K}_{d-k} \mathscr{R}_k \mathrm{L} f_{\mathrm{extreme}}\|_{\mathcal{M}} = \sum_{n=1}^N |v_n| = \|\boldsymbol{v}\|_1. \tag{3.18}$$

The proof of the theorem requires the following proposition.

[5]In particular, we prove that $(\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d), \|\cdot\|_{\mathcal{M}_{\mathrm{L}}^k})$ can be identified as the dual of some primary Banach space, which allows us to equip $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ with a weak* topology.

[6]The data-fitting problem is well-posed over the null space when the classical least-squares polynomial-fitting problem admits a unique solution with respect to the data $\{(\boldsymbol{x}_m, y_m)\}_{m=1}^M \subset \mathbb{R}^d \times \mathbb{R}$.

Proposition 3.9 ([47, Lemma 10.2]). *The isotropic shifted Dirac impulse* $\delta_{\mathrm{iso}}(\cdot - (\mathbf{A}_0, \boldsymbol{t}_0)) \in \mathcal{M}_{\mathrm{iso}}(\Xi_k)$ *defined by*

$$\delta_{\mathrm{iso}}(\cdot - (\mathbf{A}_0, \boldsymbol{t}_0)) := \mathrm{P}_{\mathrm{iso}}\{\delta(\cdot - (\mathbf{A}_0, \boldsymbol{t}_0))\}, \tag{3.19}$$

where $\delta(\cdot - (\mathbf{A}_0, \boldsymbol{t}_0)) = \delta(\cdot - \mathbf{A}_0)\delta(\cdot - \boldsymbol{t}_0) \in \mathcal{M}(\Xi_k)$ *is the "classical" Dirac impulse on* Ξ_k*, satisfies the following properties:*

1. *Sampling: For any* $\psi \in C_{0,\mathrm{iso}}(\Xi_k)$*,*
$$\langle \delta_{\mathrm{iso}}(\cdot - (\mathbf{A}_0, \boldsymbol{t}_0)), \psi \rangle_k = \psi(\mathbf{A}_0, \boldsymbol{t}_0). \tag{3.20}$$
2. *Rotation invariance: For any* $\mathbf{U} \in \mathrm{O}_{d-k}(\mathbb{R})$*,*
$$\delta_{\mathrm{iso}}(\cdot - (\mathbf{A}_0, \boldsymbol{t}_0)) = \delta_{\mathrm{iso}}(\cdot - (\mathbf{U}\mathbf{A}_0, \mathbf{U}\boldsymbol{t}_0)). \tag{3.21}$$
3. *Unit norm:* $\|\delta_{\mathrm{iso}}(\cdot - (\mathbf{A}_0, \boldsymbol{t}_0))\|_{\mathcal{M}} = 1$.
4. *Linear combination: For any set* $\{(\mathbf{A}_n, \boldsymbol{t}_n)\}_{n=1}^N \subset \Xi_k$ *of distinct points,*
$$\left\| \sum_{n=1}^N a_n \delta_{\mathrm{iso}}(\cdot - (\mathbf{A}_n, \boldsymbol{t}_n)) \right\|_{\mathcal{M}} = \sum_{n=1}^N |a_n| = \|\boldsymbol{a}\|_1. \tag{3.22}$$
5. *Extreme points of* $B_{\mathcal{M}_{\mathrm{iso}}} := \{e \in \mathcal{M}_{\mathrm{iso}}(\Xi_k) : \|e\|_{\mathcal{M}} \leq 1\}$*: If* $e \in \operatorname{Ext} B_{\mathcal{M}_{\mathrm{iso}}}$*, then* $e = \pm\delta_{\mathrm{iso}}(\cdot - (\mathbf{A}_n, \boldsymbol{t}_n))$ *for some* $(\mathbf{A}_n, \boldsymbol{t}_n) \in \Xi_k$*.*

Proof of Theorem 3.8. The proof relies on the abstract representer theorem in [68] (see also [6, 7, 65]). From the assumptions on the loss function combined with the weak*-continuity of the point evaluation functional on $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ (item 4 in Theorem 3.7), our setting coincides with the hypotheses of [68, Theorem 3]. First, this abstract result ensures that the solution set S is nonempty, convex, and weak*-compact. Second, it ensures that, when the loss function is strictly convex (or if it imposes the equality $y_m = f(\boldsymbol{x}_m)$ for $m = 1, \dots, M$), S is the weak*-closure of the convex hull of its extreme points, which can all be expressed as

$$f_{\mathrm{extreme}}(\boldsymbol{x}) = c(\boldsymbol{x}) + \sum_{n=1}^N v_n e_n(\boldsymbol{x}), \tag{3.23}$$

where the number N of atoms satisfies $N \leq (M - \dim \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d))$, $c(\cdot)$ is in the null space of the regularizer (i.e., $c \in \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$), and, for $n = 1, \dots, N$, $v_n \in \mathbb{R} \setminus \{0\}$ and e_n is an extreme point of the unit regularization ball

$$B := \left\{ f \in \mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d) : \|\mathrm{K}_{d-k} \mathscr{R}_k \mathrm{L} f\|_{\mathcal{M}} \leq 1 \right\}. \tag{3.24}$$

From item 2 in Theorem 3.7, we have the direct-sum decomposition of the native space as

$$\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d) = \mathcal{V} \oplus \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d) = \mathrm{L}_{\mathscr{R}_k}^\dagger(\mathcal{M}_{\mathrm{iso}}(\Xi_k)) \oplus \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d). \tag{3.25}$$

Since $\mathrm{L}_{\mathscr{R}_k}^\dagger : \mathcal{M}_{\mathrm{iso}}(\Xi_k) \to \mathcal{V}$ is an isometric isomorphism (item 1 in Theorem 3.7), the extreme points of B take the form $\mathrm{L}_{\mathscr{R}_k}^\dagger(\operatorname{Ext} B_{\mathcal{M}_{\mathrm{iso}}})$ plus a polynomial term in $\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$. From item 5 in Proposition 3.9, it then follows that

$$e_n = \mathrm{L}_{\mathscr{R}_k}^\dagger\{\pm\delta_{\mathrm{iso}}(\cdot - (\mathbf{A}_n, \boldsymbol{t}_n))\} + p = \pm\rho_{\mathrm{L}}(\mathbf{A}_n(\cdot) - \boldsymbol{t}_n) + \widetilde{p} \tag{3.26}$$

with $(\mathbf{A}_n, \boldsymbol{t}_n) \in \Xi_k$ and $p, \widetilde{p} \in \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$.

A calculation in the Fourier domain reveals that $\mathrm{L}\{\rho_{\mathrm{L}}(\mathbf{A}(\cdot) - \boldsymbol{t})\} = \delta(\mathbf{A}(\cdot) - \boldsymbol{t})$ for $(\mathbf{A}, \boldsymbol{t}) \in \Xi_k$. Next, we invoke the property that $\mathrm{K}_{d-k} \mathscr{R}_k \{\delta(\mathbf{A}(\cdot) - \boldsymbol{t})\} = \delta_{\mathrm{iso}}(\cdot - (\mathbf{A}, \boldsymbol{t}))$ [47, equation (9.16)], which yields that

$$\mathrm{K}_{d-k} \mathscr{R}_k \mathrm{L}\{f_{\mathrm{extreme}}\} = \sum_{n=1}^{N} v_n \delta_{\mathrm{iso}}(\cdot - (\mathbf{A}_n, \boldsymbol{t}_n)). \tag{3.27}$$

Finally, by item 4 in Proposition 3.9, we have that $\|\mathrm{K}_{d-k} \mathscr{R}_k \mathrm{L} f_{\mathrm{extreme}}\|_{\mathcal{M}} = \sum_{n=1}^{N} |v_n| = \|\boldsymbol{v}\|_1$, which proves the theorem. ■

3.2.1. Discussion. The main takeaway from Theorem 3.8 is that *sparse* neural architectures (sparse in the sense that there are fewer neurons than data) are solutions to variational problems over $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$. In particular, the regularity of functions imposed by the Banach structure of $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ explains the variational optimality of the architectures in (3.17). Furthermore, by the isomorphism in item 3 of Theorem 3.7, we see that $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ is a *nonreflexive Banach space* (since $\mathcal{M}_{\mathrm{iso}}(\Xi_k)$ is nonreflexive), which shows that $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ differs in a fundamental way from a Hilbert space. Said differently, Theorem 3.8 provides a function-space framework for neural networks that differs in a fundamental way from the (Hilbertian) framework of the neural tangent kernel [26].

The two extremes of the theorem ($k = 0$ and $k = (d-1)$) capture well-studied problems. Indeed, when $k = 0$, we can take advantage of the fact that the effect of the k-plane transform essentially disappears. Indeed, when $k = 0$ we have (see subsection 2.1.1) that

$$\|\mathrm{K}_d \mathscr{R}_0 \mathrm{L} f\|_{\mathcal{M}(\Xi_0)} = c_{d,0} \|(\mathbf{A}, \boldsymbol{t}) \mapsto \mathrm{L}\{f\}(\mathbf{A}^{\mathsf{T}} \boldsymbol{t})\|_{\mathcal{M}(\Xi_0)} = \|\mathrm{L} f\|_{\mathcal{M}(\mathbb{R}^d)}. \tag{3.28}$$

Therefore, the variational problem in (3.16) reduces to the well-studied variational problem for L-splines [13, 69]. Since L is isotropic from the admissibility assumptions (Definition 3.1), the atoms ρ_{L} are radial basis functions.

For these problems, the classical theory [13, 69] suggests that the extreme point solutions are built from atoms of the form $\rho_{\mathrm{L}}(\cdot - \boldsymbol{\tau}_n)$, $\boldsymbol{\tau}_n \in \mathbb{R}^d$, where $\rho_{\mathrm{L}} : \mathbb{R}^d \to \mathbb{R}$ is the (canonical) Green's function of L defined in the Fourier domain by $\widehat{\rho}_{\mathrm{L}} = 1/\widehat{L}$. We can quickly see that Theorem 3.8 recovers this result since the atoms take the form

$$\boldsymbol{x} \mapsto \rho_{\mathrm{L}}(\mathbf{A}_n \boldsymbol{x} - \boldsymbol{t}_n) = \rho_{\mathrm{L}}(\mathbf{A}_n^{\mathsf{T}} \mathbf{A} \boldsymbol{x} - \mathbf{A}_n^{\mathsf{T}} \boldsymbol{t}_n) = \rho_{\mathrm{L}}(\boldsymbol{x} - \mathbf{A}_n^{\mathsf{T}} \boldsymbol{t}_n) = \rho_{\mathrm{L}}(\boldsymbol{x} - \boldsymbol{\tau}_n), \tag{3.29}$$

where we made the substitution $\boldsymbol{\tau}_n = \mathbf{A}_n^{\mathsf{T}} \boldsymbol{t}_n \in \mathbb{R}^d$ in the last equality. Here, we used the fact that $\mathbf{A}\mathbf{A}^{\mathsf{T}} = \mathbf{A}^{\mathsf{T}}\mathbf{A} = \mathbf{I}_d$ when $\mathbf{A} \in V_d(\mathbb{R}^d)$ (i.e., the $k = 0$ Stiefel manifold is the space of $(d \times d)$ orthogonal matrices).

At the opposite extreme, when $k = (d-1)$, the atoms take the form

$$\boldsymbol{x} \mapsto \rho_{\mathrm{L}}(\boldsymbol{\alpha}^{\mathsf{T}} \boldsymbol{x} - t), \tag{3.30}$$

where $\rho_{\mathrm{L}} : \mathbb{R} \to \mathbb{R}$ is the Green's function of the univariate operator $\mathrm{L}_{\mathrm{rad}}$, specified by the frequency response $\widehat{L}_{\mathrm{rad}}$ of the univariate radial profile of L. These atoms are classical neurons with univariate nonlinearities. This problem was first studied in [43] with $\mathrm{L}_{\mathrm{rad}} = \partial_t^m$, which corresponds to nonlinearities proportional to the truncated power functions $t \mapsto t_+^{m-1}$ (which is the ReLU when $m = 2$), and then generalized to other regularization operators L in [66].

4. Observations and examples of compatible neural architectures.

Since Theorem 3.8 guarantees the existence of a solution to (3.16) that takes the form in (3.17), we can always find an admissible solution by solving the neural network training problem

$$\begin{aligned} &\min_{\boldsymbol{\theta}} \quad \left(\sum_{m=1}^{M} \mathcal{L}(y_m, f_{\boldsymbol{\theta}}(\boldsymbol{x}_m)) + \lambda \sum_{n=1}^{N} |v_n| \right) \\ &\text{s.t.} \quad \mathbf{A}_n \mathbf{A}_n^{\mathsf{T}} = \mathbf{I}_{d-k},\ n = 1, \dots, N, \end{aligned} \tag{4.1}$$

for some fixed width $N \geq M$ with

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = c(\boldsymbol{x}) + \sum_{n=1}^{N} v_n \, \rho_{\mathrm{L}}(\mathbf{A}_n \boldsymbol{x} - \boldsymbol{t}_n), \quad \boldsymbol{x} \in \mathbb{R}^d,\ \boldsymbol{\theta} = \{v_n, \mathbf{A}_n, \boldsymbol{t}_n\}_{n=1}^{N} \cup \{c(\cdot)\}. \tag{4.2}$$

The assumption that $N \geq M$ ensures that a solution to the variational problem in (3.16) exists in the neural network parameter space (indexed by $\boldsymbol{\theta}$) thanks to the bound in Theorem 3.8. This assumption implies that, as long as the neural network problem is critically parameterized or overparameterized, its solutions will also be solutions to the variational problem in Theorem 3.8. Thus, this result provides insight on the role of overparameterization. We also remark that the constraint on the weight matrices in (4.1) corresponds to orthogonal weight normalization. The latter has become a popular architectural choice as it has been shown to increase the stability and improve the generalization performance of neural networks [2, 24, 25, 33].

The nonlinearity $\rho_{\mathrm{L}} : \mathbb{R}^{d-k} \to \mathbb{R}^{d-k}$ that appears in (3.17) can be viewed as the Green's function of the operator $\mathrm{L}_{d-k} : \mathcal{S}(\mathbb{R}^{d-k}) \to \mathcal{S}'(\mathbb{R}^{d-k})$ which shares the radial profile $\widehat{L}_{\mathrm{rad}}$ of the k-plane-admissible operator L. That is to say,

$$\widehat{L}_{d-k}(\boldsymbol{\omega}) = \widehat{L}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2). \tag{4.3}$$

Furthermore, due to the intertwining properties of the k-plane transform [47, Corollary 7.8], it turns out that the regularization operator in (3.16) has the alternative specification

$$\mathrm{K}_{d-k} \mathscr{R}_k \mathrm{L} = \mathrm{L}_{d-k} \mathrm{K}_{d-k} \mathscr{R}_k\,. \tag{4.4}$$

The framework of Theorem 3.8 encapsulates many neural architectures. The prototypical example of such an operator is the fractional Laplacian $\mathrm{L} = (-\Delta)^{\frac{\alpha}{2}}$ and, so, $\mathrm{L}_{d-k} = (-\Delta_{d-k})^{\frac{\alpha}{2}}$. The radial profile for this family of operators is

$$\widehat{L}_{\mathrm{rad}}(\omega) = |\omega|^{\alpha}. \tag{4.5}$$

From Definition 3.1, the reader can readily verify that this operator is k-plane-admissible for $\alpha > (d-k)$. This simple operator encapsulates several known results. At one extreme ($k = (d-1)$), we recover[7] the classical ReLU neurons by the choice $\mathrm{L} = (-\Delta)$ [41, 43]. At the opposite extreme ($k = 0$), from subsection 3.2.1, we see that we recover a sparse variant of the

[7] Technically, $\rho_{(-\Delta)}(t) = |t|/2$ in this case, but since $t \mapsto |t|/2$ differs from the ReLU $t \mapsto t_+$ by a null space component (affine function), the ReLU and absolute value nonlinearity are treated the same in this framework.

classical thin-plate/polyharmonic spline radial basis functions of Duchon [12] by the choice $\mathrm{L} = (-\Delta)^{\frac{\alpha}{2}}$, $\alpha > d$.

In the intermediate regime $1 \leq k \leq (d-2)$, we can choose $\mathrm{L}_{d-k} = (-\Delta_{d-k})^{\frac{1+(d-k)}{2}}$ so that $\rho_{\mathrm{L}}(\boldsymbol{t}) \propto \|\boldsymbol{t}\|_2$, $\boldsymbol{t} \in \mathbb{R}^{d-k}$, is the norm activation function that has been used for neural architectures in [23]. These observations follow from the fact that the Green's function of the fractional Laplacian $(-\Delta_n)^{\frac{\alpha}{2}}$ (which acts on n-variables) for $\alpha > n$ takes the form

$$k_{\alpha,n}(\boldsymbol{t}) = \mathscr{F}^{-1}\left\{\frac{1}{\|\cdot\|_2}\right\}(\boldsymbol{t}) = \begin{cases} A_{\alpha,n}\|\boldsymbol{t}\|_2^{\alpha-n}, & \alpha - n \notin 2\mathbb{N}, \\ B_{m,n}\|\boldsymbol{t}\|_2^{2m}\log\|\boldsymbol{t}\|_2, & \alpha - n = 2m, m \in \mathbb{N}, \\ \Delta_n^{-m}\{\delta\}, & -\alpha/2 = m, m \in \mathbb{N}, \end{cases} \tag{4.6}$$

with $\boldsymbol{t} \in \mathbb{R}^n$, $A_{\alpha,n} = \frac{\Gamma((n-\alpha)/2)}{2^\alpha \pi^{n/2}\Gamma(\alpha/2)}$, and $B_{m,n} = \frac{(-1)^{1+m}}{2^{2m+n-1}\pi^{n/2}\Gamma(m+n/2)m!}$ [17, 53]. In general, there exist many nonlinearities that are compatible with the presented framework. All that needs to be verified is the admissibility conditions (Definition 3.1) of the underlying regularization operator.

5. Connections to RKBS methods and variation spaces.

After [43], a recent line of research has been trying to understand neural networks through the lens of RKBSs [5, 62]. These works consider Banach spaces defined on, say, $\mathbb{R}^d$ whose members are defined as integral combinations of atoms from some continuously indexed dictionary $\mathcal{D}$. The elements of the dictionary are assumed to be continuously indexed by $\xi \in \Xi$, where Ξ is assumed to be some locally compact Hausdorff space. That is, $\mathcal{D} = \{\varphi_\xi\}_{\xi\in\Xi}$, with the additional hypothesis that $\xi \mapsto \varphi_\xi(\boldsymbol{x}) \in C_0(\Xi)$ for any $\boldsymbol{x} \in \mathbb{R}^d$.

It turns out that the space

$$\mathcal{B}(\mathbb{R}^d) := \left\{f : \mathbb{R}^d \to \mathbb{R} \text{ is measurable} : \text{there exists } \mu \in \mathcal{M}(\Xi) \text{ s.t. } f = \int_\Xi \varphi_\xi \,\mathrm{d}\mu(\xi)\right\} \tag{5.1}$$

forms a Banach space when equipped with the norm

$$\|f\|_{\mathcal{B}} := \inf_{\mu\in\mathcal{M}(\Xi)} \|\mu\|_{\mathcal{M}} \quad \text{s.t.} \quad f = \int_\Xi \varphi_\xi \,\mathrm{d}\mu(\xi). \tag{5.2}$$

The assumptions on $\mathcal{D} = \{\varphi_\xi\}_{\xi\in\Xi}$ ensure that the point evaluation is continuous on $\mathcal{B}(\mathbb{R}^d)$ (i.e., $\delta(\cdot - \boldsymbol{x}_0) \in \mathcal{B}'(\mathbb{R}^d)$). Such Banach spaces are referred to as *RKBSs* [34, 72]. An RKBS formed from integral combinations of atoms from some continuously indexed dictionary is an *integral RKBS* (I-RKBS) [62]. With this formalism, [5, 62] study many properties of $\mathcal{B}(\mathbb{R}^d)$ as well as data-fitting problems over these spaces with associated representer theorems. We remark that, thanks to the assumption $\xi \mapsto \varphi_\xi(\boldsymbol{x}) \in C_0(\Xi)$ for any $\boldsymbol{x} \in \mathbb{R}^d$, these authors implicitly ensure that the point evaluation is actually weak*-continuous on $\mathcal{B}(\mathbb{R}^d)$, which is stronger than standard continuity. This property is critical in proving the existence of solutions to data-fitting problems over these spaces.

Our native spaces $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ are compatible with the I-RKBS framework. We first note that item 4 in Theorem 3.7 ensures that the point evaluation is weak*-continuous and, hence, continuous on $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$. Thus, $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ is an RKBS. Next, we have the direct-sum decomposition

$$\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d) = \mathrm{L}_{\mathscr{R}_k}^\dagger(\mathcal{M}_{\mathrm{iso}}(\Xi_k)) \oplus \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d) = \mathrm{L}_{\mathscr{R}_k}^\dagger(\mathcal{M}(\Xi_k)) \oplus \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d), \tag{5.3}$$

where the first equality is from item 2 in Theorem 3.7 and the second inequality follows since the null space of $\mathscr{R}_k^*$ is $(\mathcal{M}_{\mathrm{iso}}(\Xi_k))^{\mathsf{c}}$ (recall that $\mathrm{L}_{\mathscr{R}_k}^{\dagger} = (\mathrm{Id} - \mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)})\,\mathrm{L}^{-1}\mathscr{R}_k^*$ and see Proposition 2.7). This immediately implies that $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ is the direct sum of an I-RKBS with $\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$, where the dictionary is composed of the kernels $g_{\mathbf{A},\boldsymbol{t}}$ from (3.8), continuously indexed by $(\mathbf{A},\boldsymbol{t}) \in \Xi_k$. This correspondence allows us to directly apply any I-RKBS developments to $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$.

The study of variation spaces to understand neural networks is a classical endeavor [31, 37]. These spaces have received renewed interest [4, 11, 56, 58, 57] as a means toward the understanding of the reason why neural networks seem to "break" the curse of dimensionality through the lens of nonlinear approximation theory. It turns out that the variation space for a dictionary $\mathcal{D}$ exactly coincides with the I-RKBS as long as the members of $\mathcal{D}$ are sufficiently regular (see [57, Lemma 3]). Indeed, in that case, the variation space for $\mathcal{D}$ is the Banach space $(\mathcal{B}(\mathbb{R}^d), \|\cdot\|_{\mathcal{B}})$ defined in (5.1). Thus, $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d)$ can also be viewed as a variation space. The investigation of the implications of these tight connections to I-RKBS and variation spaces toward the understanding of neural architectures with multivariate nonlinearities is a direction for future work.

Appendix A. Proof of Theorem 3.6.

Proof. We first note that $\mathrm{L}_{\mathscr{R}_k}$ maps $\mathcal{M}_{\mathrm{L}}^k(\mathbb{R}^d) \to \mathcal{M}_{\mathrm{iso}}(\Xi_k)$ by design due to the isotropic symmetry of the k-plane domain. The remainder of the proof is divided into three parts.

Part (I): Existence/continuity of $\mathrm{L}_{\mathscr{R}_k}^{\dagger}$ and the right-inverse property (3.5). We first note that $\mathrm{L}^* : \mathcal{S}(\mathbb{R}^d) \to L_{n_{\mathrm{L}}}^1(\mathbb{R}^d)$ is a continuous injection (item 1 in Definition 3.1). Therefore, there exists an inverse operator $\mathrm{L}^{*-1} : \mathrm{L}^*\left(\mathcal{S}(\mathbb{R}^d)\right) \to \mathcal{S}(\mathbb{R}^d)$ such that $\mathrm{L}^{*-1}\mathrm{L}^* = \mathrm{Id}$ on $\mathcal{S}(\mathbb{R}^d)$. In particular, L^{*-1} is the LSI operator specified by the frequency response $\boldsymbol{\xi} \mapsto 1/\widehat{L}(\boldsymbol{\xi})$ (since L^* is necessarily self-adjoint from item 2 in Definition 3.1).

Next, we define the operator

$$\mathrm{L}_{\mathscr{R}_k}^{\dagger *} := \mathscr{R}_k\,\mathrm{L}^{-1*}(\mathrm{Id} - \mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)}^*), \tag{A.1}$$

where

$$\mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)}^*\{f\} = \sum_{|\boldsymbol{n}| \le n_{\mathrm{L}}} \langle m_{\boldsymbol{n}}, f\rangle m_{\boldsymbol{n}}^*, \tag{A.2}$$

which is the projection of f onto the dual space $\left(\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)\right)' \subset \mathcal{S}(\mathbb{R}^d)$. By recalling from (2.24) that $m_{\boldsymbol{n}}(\boldsymbol{x}) = \boldsymbol{x}^{\boldsymbol{n}}/\boldsymbol{n}!$, we see that (A.2) is a well-defined operator as long as f has sufficient decay (e.g., $f \in \mathrm{L}^*\left(\mathcal{S}(\mathbb{R}^d)\right) \subset L_{n_{\mathrm{L}}}^1(\mathbb{R}^d)$). This reveals that

$$\mathrm{L}_{\mathscr{R}_k}^{\dagger *} : \mathrm{L}^*\left(\mathcal{S}(\mathbb{R}^d)\right) \to \mathcal{S}_k. \tag{A.3}$$

To check the continuity of this operator, we characterize the boundedness of the (Schwartz) kernel of $\mathrm{L}_{\mathscr{R}_k}^{\dagger *}$. This kernel can be formally identified with $((\mathbf{A},\boldsymbol{t}),\boldsymbol{x}) \mapsto h_{\boldsymbol{x}}(\mathbf{A},\boldsymbol{t}) := \mathrm{L}_{\mathscr{R}_k}^{\dagger *}\{\delta(\cdot - \boldsymbol{x})\}(\mathbf{A},\boldsymbol{t})$. By the Fourier slice theorem,

$$\begin{aligned}
\widehat{h_{\boldsymbol{x}}(\mathbf{A},\cdot)}(\boldsymbol{\omega}) &= \frac{\mathscr{F}\left\{\delta(\cdot-\boldsymbol{x}) - \sum_{|\boldsymbol{n}|\le n_{\mathrm{L}}} \langle m_{\boldsymbol{n}}, \delta(\cdot-\boldsymbol{x})\rangle m_{\boldsymbol{n}}^*\right\}(\mathbf{A}^\mathsf{T}\boldsymbol{\omega})}{\widehat{L}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2)} \\
&= \frac{e^{-\mathrm{i}\boldsymbol{\omega}^\mathsf{T}\mathbf{A}\boldsymbol{x}} - \sum_{|\boldsymbol{n}|\le n_{\mathrm{L}}} \frac{\boldsymbol{x}^{\boldsymbol{n}}}{\boldsymbol{n}!}\, \widehat{m}_{\boldsymbol{n}}^*(\mathbf{A}^\mathsf{T}\boldsymbol{\omega})}{\widehat{L}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2)} \\
&= \frac{e^{-\mathrm{i}\boldsymbol{\omega}^\mathsf{T}\mathbf{A}\boldsymbol{x}} - \sum_{|\boldsymbol{n}|\le n_{\mathrm{L}}} \frac{\boldsymbol{x}^{\boldsymbol{n}}}{\boldsymbol{n}!}\, (-\mathrm{i}\mathbf{A}^\mathsf{T}\boldsymbol{\omega})^{\boldsymbol{n}}\, \widehat{\kappa}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2)}{\widehat{L}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2)} \\
&= \frac{e^{-\mathrm{i}\boldsymbol{\omega}^\mathsf{T}\mathbf{A}\boldsymbol{x}} - \sum_{n=0}^{n_{\mathrm{L}}} \frac{(-\mathrm{i}\boldsymbol{\omega}^\mathsf{T}\mathbf{A}\boldsymbol{x})^n}{n!}\, \widehat{\kappa}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2)}{\widehat{L}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2)}, \\
&= \frac{e^{-\mathrm{i}\boldsymbol{\omega}^\mathsf{T}\mathbf{A}\boldsymbol{x}} - \widehat{\kappa}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2) \sum_{n=0}^{n_{\mathrm{L}}} \frac{(-\mathrm{i}\boldsymbol{\omega}^\mathsf{T}\mathbf{A}\boldsymbol{x})^n}{n!}}{\widehat{L}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2)},
\end{aligned} \tag{A.4}$$

where the penultimate line holds by the multinomial expansion. Note that the quantity in (A.4) is well-defined despite the pole of multiplicity γ_{L} at $\boldsymbol{\omega}=\mathbf{0}$. Since $\gamma_{\mathrm{L}} \in (n_{\mathrm{L}}, n_{\mathrm{L}}+1]$, the form of the numerator ensures a proper pole-zero cancellation with the denominator. Indeed, by Taylor's theorem, when $t\in\mathbb{R}$ is in a neighborhood of 0, we have that

$$e^t - \sum_{n=1}^{n_{\mathrm{L}}} \frac{t^n}{n!} = O(t^{n_{\mathrm{L}}+1}). \tag{A.5}$$

By the identification of the numerator[8] in (A.4) with the above display combined with the property of item 3 in Definition 3.1, we have that (A.4) is well-defined. Next, since $\widehat{\kappa}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2) \le 1$ for $\|\boldsymbol{\omega}\|_2 < 1$ and $\widehat{\kappa}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2) = 0$ for $\|\boldsymbol{\omega}\|_2 \ge 1$ (subsection 2.2), on one hand we have for $\|\boldsymbol{\omega}\|_2 < 1$ that $|\widehat{h_{\boldsymbol{x}}(\mathbf{A},\cdot)}(\boldsymbol{\omega})|$ is bounded by a constant which depends on $\mathbf{A}$ and $\boldsymbol{x}$. To see the dependence on $\mathbf{A}$ and $\boldsymbol{x}$, we note that

$$\begin{aligned}
&\sum_{n=0}^{n_{\mathrm{L}}} \frac{1}{n!} |\mathrm{i}\boldsymbol{\omega}^\mathsf{T}\mathbf{A}\boldsymbol{x}|^n \\
&\le \sum_{n=0}^{n_{\mathrm{L}}} \frac{1}{n!} \|\boldsymbol{\omega}\|_2^n \|\mathbf{A}\boldsymbol{x}\|_2^n \\
&\le \sum_{n=0}^{n_{\mathrm{L}}} \frac{(n_{\mathrm{L}}-n)!}{n_{\mathrm{L}}!} \frac{n_{\mathrm{L}}!}{n!(n_{\mathrm{L}}-n)!} \|\mathbf{A}\boldsymbol{x}\|_2^n \\
&\le \sum_{n=0}^{n_{\mathrm{L}}} \frac{n_{\mathrm{L}}!}{n!(n_{\mathrm{L}}-n)!} \|\mathbf{A}\boldsymbol{x}\|_2^n \\
&= (1+\|\mathbf{A}\boldsymbol{x}\|_2)^{n_{\mathrm{L}}}.
\end{aligned} \tag{A.6}$$

Thus, there exists a universal constant $C_0 > 0$ such that

$$|\widehat{h_{\boldsymbol{x}}(\mathbf{A},\cdot)}(\boldsymbol{\omega})| \le C_0(1+\|\mathbf{A}\boldsymbol{x}\|_2)^{n_{\mathrm{L}}}, \quad \|\boldsymbol{\omega}\|_2 < 1. \tag{A.7}$$

[8]This identification is valid by the substitution $t = -\mathrm{i}\boldsymbol{\omega}^\mathsf{T}\mathbf{A}\boldsymbol{x}$ and noting that $\widehat{\kappa}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2) = 1$ for $\|\boldsymbol{\omega}\|_2 < R_0$, for some $R_0 \le 1/2$ [67, p. 6].

On the other hand, when $\|\boldsymbol{\omega}\|_2 \geq 1$, we have that

$$|\widehat{h_{\boldsymbol{x}}(\mathbf{A},\cdot)}(\boldsymbol{\omega})| \leq \frac{1}{\widehat{L}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2)} \leq \frac{C_0(1+\|\mathbf{A}\boldsymbol{x}\|_2)^{n_{\mathrm{L}}}}{\widehat{L}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2)}. \tag{A.8}$$

From item 4 in Definition 3.1, we have that

$$\frac{1}{\widehat{L}_{\mathrm{rad}}(\|\boldsymbol{\omega}\|_2)} \leq \frac{\|\boldsymbol{\omega}\|_2^{-\gamma'_{\mathrm{L}}}}{C'}, \tag{A.9}$$

where $\gamma'_{\mathrm{L}} > (d-k)$. This property ensures that $\widehat{h_{\boldsymbol{x}}(\mathbf{A},\cdot)} \in L^1(\mathbb{R}^{d-k})$. Combining (A.7)–(A.9), implies that

$$(1+\|\mathbf{A}\boldsymbol{x}\|_2)^{-n_{\mathrm{L}}}\|\widehat{h_{\boldsymbol{x}}(\mathbf{A},\cdot)}\|_{L^1} < \infty. \tag{A.10}$$

Since this bound is uniform in $\boldsymbol{x} \in \mathbb{R}^d$ and $\mathbf{A} \in V_{d-k}(\mathbb{R}^d)$ and the inverse Fourier transform $\mathscr{F}_{d-k}^{-1} : L^1(\mathbb{R}^{d-k}) \to C_0(\mathbb{R}^{d-k})$ is a bounded operator (Riemann–Lebesgue lemma), we have that

$$\sup_{\substack{\boldsymbol{x}\in\mathbb{R}^d \\ (\mathbf{A},\boldsymbol{t})\in\Xi_k}} |h_{\boldsymbol{x}}(\mathbf{A},\boldsymbol{t})|(1+\|\mathbf{A}\boldsymbol{x}\|_2)^{-n_{\mathrm{L}}} < \infty \tag{A.11}$$

with the property that $h_{\boldsymbol{x}} \in C_{0,\mathrm{iso}}(\Xi_k)$, due to the isotropic symmetry of the k-plane transform. Finally, if we write

$$\mathbf{A}\boldsymbol{x} = \begin{bmatrix} \boldsymbol{\alpha}_1^{\mathsf{T}}\boldsymbol{x} \\ \vdots \\ \boldsymbol{\alpha}_{d-k}^{\mathsf{T}}\boldsymbol{x} \end{bmatrix}, \tag{A.12}$$

where $\boldsymbol{\alpha}_n$ is the nth row of $\mathbf{A}$, then we have that

$$\|\mathbf{A}\boldsymbol{x}\|_2 = \sqrt{(\boldsymbol{\alpha}_1^{\mathsf{T}}\boldsymbol{x})^2+\cdots+(\boldsymbol{\alpha}_{d-k}^{\mathsf{T}}\boldsymbol{x})^2} \leq |\boldsymbol{\alpha}_1^{\mathsf{T}}\boldsymbol{x}|+\cdots+|\boldsymbol{\alpha}_{d-k}^{\mathsf{T}}\boldsymbol{x}| \leq (d-k)\|\boldsymbol{x}\|_2. \tag{A.13}$$

This results in

$$\sup_{\substack{\boldsymbol{x}\in\mathbb{R}^d \\ (\mathbf{A},\boldsymbol{t})\in\Xi_k}} |h_{\boldsymbol{x}}(\mathbf{A},\boldsymbol{t})|(1+\|\boldsymbol{x}\|_2)^{-n_{\mathrm{L}}} < \infty. \tag{A.14}$$

This bound implies that the operator is actually well-defined on the larger space $L^1_{n_{\mathrm{L}}}(\mathbb{R}^d) \supset \mathrm{L}^*\left(\mathcal{S}(\mathbb{R}^d)\right)$. In fact, since $h_{\boldsymbol{x}} \in C_{0,\mathrm{iso}}(\Xi_k)$ we have for any $f \in L^1_{n_{\mathrm{L}}}(\mathbb{R}^d)$ that

$$\mathrm{L}^{\dagger *}_{\mathscr{R}_k}\{f\} = \int_{\mathbb{R}^d} f(\boldsymbol{x})h_{\boldsymbol{x}}(\cdot)\,\mathrm{d}\boldsymbol{x} \in C_{0,\mathrm{iso}}(\Xi_k). \tag{A.15}$$

This, combined with (A.14), ensures that the operator

$$\mathrm{L}^{\dagger *}_{\mathscr{R}_k} : (L^1_{n_{\mathrm{L}}}(\mathbb{R}^d), \|\cdot\|_{L^1_{n_{\mathrm{L}}}}) \to (C_{0,\mathrm{iso}}(\Xi_k), \|\cdot\|_{L^\infty}) \tag{A.16}$$

is continuous and, subsequently, that its adjoint

$$(\mathrm{L}^{\dagger *}_{\mathscr{R}_k})^* = \mathrm{L}^{\dagger}_{\mathscr{R}_k} = (\mathrm{Id} - \mathrm{P}_{\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)})\,\mathrm{L}^{-1}\,\mathscr{R}_k^* : (\mathcal{M}_{\mathrm{iso}}(\Xi_k), \|\cdot\|_{\mathcal{M}}) \to (L^\infty_{-n_\mathrm{L}}(\mathbb{R}^d), \|\cdot\|_{L^\infty_{-n_\mathrm{L}}}) \tag{A.17}$$

is also continuous.

We now prove the right-inverse property (3.5). Recall that $\mathrm{L}^{*-1}\,\mathrm{L}^* = \mathrm{Id}$ on $\mathcal{S}(\mathbb{R}^d)$. Thus, by duality, we have the identity

$$\mathrm{L}\,\mathrm{L}^{-1} = \mathrm{Id} \text{ on } \mathcal{S}'(\mathbb{R}^d). \tag{A.18}$$

Since $\mathcal{M}_{\mathrm{iso}}(\Xi_k)$ continuously embeds into $\mathcal{S}'_k$ (Proposition 2.7), given $u \in \mathcal{M}_{\mathrm{iso}}(\Xi_k)$, we have that

$$\begin{aligned}
\mathrm{L}_{\mathscr{R}_k}\,\mathrm{L}^{\dagger}_{\mathscr{R}_k}\{u\} &= \mathrm{K}_{d-k}\,\mathscr{R}_k\,\mathrm{L}(\mathrm{Id} - \mathrm{P}_{\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)})\,\mathrm{L}^{-1}\,\mathscr{R}_k^*\{u\} \\
&= \mathrm{K}_{d-k}\,\mathscr{R}_k\,\mathrm{L}\,\mathrm{L}^{-1}\,\mathscr{R}^*\{u\} - \mathrm{K}_{d-k}\,\mathscr{R}_k\,\underbrace{\mathrm{L}\left\{\mathrm{P}_{\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)}\left\{\mathrm{L}^{-1}\,\mathscr{R}_k^*\{u\}\right\}\right\}}_{=0} \\
&= \mathrm{K}_{d-k}\,\mathscr{R}_k\,\mathscr{R}_k^*\{u\} \\
&= u,
\end{aligned} \tag{A.19}$$

where the last equality follows from Proposition 2.7.

Part (II)*: The pseudo-left-inverse property* (3.6). Observe that, for any $f \in L^\infty_{-n_\mathrm{L}}(\mathbb{R}^d)$,

$$\mathrm{L}^{-1}\,\mathrm{L}\{f\} = f + p \tag{A.20}$$

for some $p \in \mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)$. Therefore, given $f \in \mathcal{M}^k_\mathrm{L}(\mathbb{R}^d) \subset L^\infty_{-n_\mathrm{L}}(\mathbb{R}^d)$, we have that

$$\begin{aligned}
\mathrm{L}^{\dagger}_{\mathscr{R}_k}\,\mathrm{L}_{\mathscr{R}_k}\{f\} &= (\mathrm{Id} - \mathrm{P}_{\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)})\,\mathrm{L}^{-1}\,\mathscr{R}^*\,\mathrm{K}_{d-k}\,\mathscr{R}_k\,\mathrm{L}\{f\} \\
&= (\mathrm{Id} - \mathrm{P}_{\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)})\,\mathrm{L}^{-1}\,\mathrm{L}\{f\} \\
&= (\mathrm{Id} - \mathrm{P}_{\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)})\{f + p\} \\
&= f + p - \mathrm{P}_{\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)})\{f\} - \underbrace{\mathrm{P}_{\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)})\{p\}}_{=p} \\
&= (\mathrm{Id} - \mathrm{P}_{\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)})\{f\}.
\end{aligned} \tag{A.21}$$

Part (III)*: The form* (3.8)*, stability* (3.9)*, and continuity* (3.10) *of the kernel.* From (A.17) we immediately see that the kernel takes the form

$$g_{\mathbf{A},\boldsymbol{t}}(\boldsymbol{x}) = \rho_\mathrm{L}(\mathbf{A}\boldsymbol{x} - \boldsymbol{t}) - \sum_{|\boldsymbol{n}| \leq n_\mathrm{L}} \langle m^*_{\boldsymbol{n}}, \rho_\mathrm{L}(\mathbf{A}(\cdot) - \boldsymbol{t})\rangle m_{\boldsymbol{n}}(\boldsymbol{x}), \tag{A.22}$$

where $\langle m^*_{\boldsymbol{n}}, \rho_\mathrm{L}(\mathbf{A}(\cdot) - \boldsymbol{t})\rangle$ is well-defined since $m^*_{\boldsymbol{n}} \in \mathcal{S}(\mathbb{R}^d)$. Next, we note that the kernel of $\mathrm{L}^{\dagger}_{\mathscr{R}_k}$ is the "transpose" of the kernel of $\mathrm{L}^{\dagger *}_{\mathscr{R}_k}$. Consequently, we have the equality $g_{\mathbf{A},\boldsymbol{t}}(\boldsymbol{x}) = h_{\boldsymbol{x}}(\mathbf{A}, \boldsymbol{t})$. Therefore, (A.14) is equivalent to the stability bound

$$\sup_{\substack{\boldsymbol{x} \in \mathbb{R}^d \\ (\mathbf{A},\boldsymbol{t}) \in \Xi_k}} |g_{\mathbf{A},\boldsymbol{t}}(\boldsymbol{x})|(1 + \|\boldsymbol{x}\|_2)^{-n_\mathrm{L}} < \infty. \tag{A.23}$$

Finally, in Part (I) of the proof we showed that $h_{\boldsymbol{x}} \in C_{0,\mathrm{iso}}(\Xi_k)$, which proves (3.10). ■

Appendix B. Proof of Theorem 3.7.

Proof.

1. From (3.5) in Theorem 3.6, we readily deduce that

$$\begin{aligned}&\mathrm{L}^{\dagger}_{\mathscr{R}_k} : \mathcal{M}_{\mathrm{iso}}(\Xi_k) \to \mathcal{V},\\ &\mathrm{L}_{\mathscr{R}_k} : \mathcal{V} \to \mathcal{M}_{\mathrm{iso}}(\Xi_k)\end{aligned} \tag{B.1}$$

are continuous bijections. Therefore, if we equip $\mathcal{V}$ with the norm in (3.12), $\mathcal{V}$ is isometrically isomorphic to $\mathcal{M}_{\mathrm{iso}}(\Xi_k)$.

2. To check that the sum is direct, we must verify that $\mathcal{V} \cap \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d) = \{0\}$. By construction, $\mathcal{V} \subset \mathcal{M}^k_{\mathrm{L}}(\mathbb{R}^d)$. Therefore, by (3.6) in Theorem 3.6,

$$\mathrm{L}^{\dagger}_{\mathscr{R}_k} \mathrm{L}_{\mathscr{R}_k} = \mathrm{Id} - \mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)} \text{ on } \mathcal{V}. \tag{B.2}$$

Yet, from item 1,

$$\mathrm{L}^{\dagger}_{\mathscr{R}_k} \mathrm{L}_{\mathscr{R}_k} = \mathrm{Id} \text{ on } \mathcal{V}. \tag{B.3}$$

Thus, $(\mathrm{Id} - \mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)})\left(\mathcal{M}^k_{\mathrm{L}}(\mathbb{R}^d)\right) = \mathcal{V}$. Consequently, $\mathcal{V}$ and $\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$ are complementary Banach subspaces of $\mathcal{M}^k_{\mathrm{L}}(\mathbb{R}^d)$ and so $\mathcal{V} \cap \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d) = \{0\}$.

3. Since $\mathcal{V}$ and $\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$ are complementary Banach subspaces of $\mathcal{M}^k_{\mathrm{L}}(\mathbb{R}^d)$, we can decompose any $f \in \mathcal{M}^k_{\mathrm{L}}(\mathbb{R}^d)$ as

$$\begin{aligned} f &= (\mathrm{Id} - \mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)})\{f\} + \mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)}\{f\}\\ &= \mathrm{L}^{\dagger}_{\mathscr{R}_k} \mathrm{L}_{\mathscr{R}_k}\{f\} + \mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)}\{f\}\\ &= \mathrm{L}^{\dagger}_{\mathscr{R}_k}\{u\} + p, \end{aligned} \tag{B.4}$$

where the second line follows from (3.6) in Theorem 3.6. Therefore, we can equip $\mathcal{M}^k_{\mathrm{L}}(\mathbb{R}^d)$ with the composite norm

$$\begin{aligned} \|f\|_{\mathcal{M}^k_{\mathrm{L}}} &:= \|\mathrm{L}^{\dagger}_{\mathscr{R}_k}\{u\}\|_{\mathcal{V}} + \|p\|_{\mathcal{P}_{n_{\mathrm{L}}}}\\ &= \|\mathrm{L}^{\dagger}_{\mathscr{R}_k} \mathrm{L}_{\mathscr{R}_k} f\|_{\mathcal{V}} + \|\mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)} f\|_{\mathcal{P}_{n_{\mathrm{L}}}}\\ &= \|\underbrace{\mathrm{L}_{\mathscr{R}_k} \mathrm{L}^{\dagger}_{\mathscr{R}_k}}_{=\mathrm{Id}} \mathrm{L}_{\mathscr{R}_k} f\|_{\mathcal{M}} + \|\mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)} f\|_{\mathcal{P}_{n_{\mathrm{L}}}}\\ &= \|\mathrm{L}_{\mathscr{R}_k} f\|_{\mathcal{M}} + \|\mathrm{P}_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)} f\|_{\mathcal{P}_{n_{\mathrm{L}}}}. \end{aligned} \tag{B.5}$$

This norm is an isometric isomorphism with $\mathcal{V} \times \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$ by design and thus an isometric isomorphism with $\mathcal{M}_{\mathrm{iso}}(\Xi_k) \times \mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)$ by item 1.

4. We first note that, in order to equip a Banach space with a weak* topology, it must be identifiable as the dual of some primary Banach space.
Next, notice that

$$\langle m_{\boldsymbol{n}}, f\rangle = \langle m_{\boldsymbol{n}}, \mathrm{L}^* \varphi\rangle = \langle \mathrm{L}\, m_{\boldsymbol{n}}, \varphi\rangle = \langle 0, \varphi\rangle = 0 \tag{B.6}$$

for all $f = \mathrm{L}^* \varphi \in \mathrm{L}^*\left(\mathcal{S}(\mathbb{R}^d)\right)$ with $\varphi \in \mathcal{S}(\mathbb{R}^d)$ and $|\boldsymbol{n}| \le n_{\mathrm{L}}$, where we took advantage of the null space property of L (Remark 3.2). Therefore, $\mathrm{P}^*_{\mathcal{P}_{n_{\mathrm{L}}}(\mathbb{R}^d)}\{f\} = 0$ for all $f \in \mathrm{L}^*\left(\mathcal{S}(\mathbb{R}^d)\right)$.

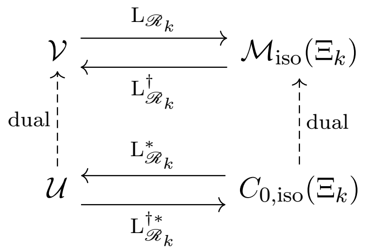

Figure 1. *Relationships between the function spaces.*

This shows that the operator $\mathrm{L}_{\mathscr{R}_k}^*$ (which, the reader can check, maps $\mathcal{S}_k \to \mathrm{L}^* \left(\mathcal{S}(\mathbb{R}^d)\right)$) is such that

$$
\begin{aligned}
\mathrm{L}_{\mathscr{R}_k}^{\dagger *} \mathrm{L}_{\mathscr{R}_k}^* \{\psi\} &= \mathscr{R}_k \mathrm{L}^{-1*}(\mathrm{Id} - \mathrm{P}^*_{\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)}) \mathrm{L}^* \underbrace{\mathscr{R}_k^* \mathrm{K}_{d-k}\{\psi\}}_{\in \mathcal{S}(\mathbb{R}^d)} \\
&= \mathscr{R}_k \mathrm{L}^{-1*} \mathrm{L}^* \mathscr{R}_k^* \mathrm{K}_{d-k}\{\psi\} \\
&= \mathscr{R}_k \mathscr{R}_k^* \mathrm{K}_{d-k}\{\psi\} \\
&= \psi
\end{aligned} \tag{B.7}
$$

for all $\psi \in \mathcal{S}_k$. Thus, $\mathrm{L}_{\mathscr{R}_k}^{\dagger *}$ is a left-inverse of $\mathrm{L}_{\mathscr{R}_k}^*$. This implies that the normed space $(\mathcal{S}_k, \|\cdot\|_{L^\infty})$ is (isometrically) isomorphic to the normed space $(\mathrm{L}^* \left(\mathcal{S}(\mathbb{R}^d)\right), \|\cdot\|_{\mathcal{U}})$ with $\|u\|_{\mathcal{U}} := \|\mathrm{L}_{\mathscr{R}_k}^{\dagger *}\{u\}\|_{L^\infty}$. Recall that $C_{0,\mathrm{iso}}(\Xi_k) = \overline{(\mathcal{S}_k, \|\cdot\|_{L^\infty})}$ from (2.20) and define the Banach space $\mathcal{U} := \overline{(\mathrm{L}^* \left(\mathcal{S}(\mathbb{R}^d)\right), \|\cdot\|_{\mathcal{U}})}$. We can now invoke the bounded linear transformation theorem [49, Theorem I.7, p. 9] on both $\mathrm{L}_{\mathscr{R}_k}^{\dagger *}$ and $\mathrm{L}_{\mathscr{R}_k}^*$ to find that these operators have the continuous extensions

$$
\begin{aligned}
&\mathrm{L}_{\mathscr{R}_k}^{\dagger *} : \mathcal{U} \to C_{0,\mathrm{iso}}(\Xi_k), \\
&\mathrm{L}_{\mathscr{R}_k}^{*} : C_{0,\mathrm{iso}}(\Xi_k) \to \mathcal{U},
\end{aligned} \tag{B.8}
$$

which establishes that $\mathcal{U}$ and $C_{0,\mathrm{iso}}(\Xi_k)$ are (isometrically) isomorphic Banach spaces. From item 3, we know that $\mathcal{M}_\mathrm{L}^k(\mathbb{R}^d)$ is isometrically isomorphic to $\mathcal{M}_{\mathrm{iso}}(\Xi_k) \times \mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)$. Since $\mathcal{M}_{\mathrm{iso}}(\Xi_k) = (C_{0,\mathrm{iso}}(\Xi_k))'$ (see (2.19)) and $\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d) = \left(\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)\right)''$ (since $\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)$ is finite-dimensional and hence reflexive), we see that there is a predual of $\mathcal{M}_\mathrm{L}^k(\mathbb{R}^d)$ that is isometrically isomorphic to $C_{0,\mathrm{iso}}(\Xi_k) \times \left(\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)\right)'$.

From (B.8) and item 1, we have the diagram in Figure 1. Therein, we see that $\mathcal{V} = \mathcal{U}'$ and so $\mathcal{X} = \mathcal{U} \oplus \left(\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)\right)'$ is such that $\mathcal{X}' = \mathcal{M}_\mathrm{L}^k(\mathbb{R}^d)$. To complete the proof, we need to establish that $\delta(\cdot - \boldsymbol{x}_0) \in \mathcal{U} \oplus \left(\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)\right)'$ [49, Theorem IV.20, p. 114]. Clearly, $\delta(\cdot - \boldsymbol{x}_0) \in \left(\mathcal{P}_{n_\mathrm{L}}(\mathbb{R}^d)\right)'$. Therefore, we only need to check that $\delta(\cdot - \boldsymbol{x}_0) \in \mathcal{U}$. This is equivalent to $\mathrm{L}_{\mathscr{R}_k}^{\dagger *}\{\delta(\cdot - \boldsymbol{x}_0)\} \in C_{0,\mathrm{iso}}(\Xi_k)$. Since $\mathrm{L}_{\mathscr{R}_k}^{\dagger *}\{\delta(\cdot - \boldsymbol{x}_0)\}(\mathbf{A}, \boldsymbol{t}) = g_{\mathbf{A},\boldsymbol{t}}(\boldsymbol{x}_0)$, the result follows from (3.10) in Theorem 3.6. ■

REFERENCES

[1] E. Abbe, E. B. Adsera, and T. Misiakiewicz, *The merged-staircase property: A necessary and nearly sufficient condition for SGD learning of sparse functions on two-layer neural networks*, in Proceedings of the Conference on Learning Theory, 2022, pp. 4782–4887.

[2] C. Anil, J. Lucas, and R. Grosse, *Sorting out Lipschitz function approximation*, in Proceedings of the International Conference on Machine Learning, 2019, pp. 291–301.

[3] S. Aziznejad and M. Unser, *Multikernel regression with sparsity constraint*, SIAM J. Math. Data Sci., 3 (2021), pp. 201–224, https://doi.org/10.1137/20M1318882.

[4] F. Bach, *Breaking the curse of dimensionality with convex neural networks*, J. Mach. Learn. Res., 18 (2017), pp. 629–681.

[5] F. Bartolucci, E. De Vito, L. Rosasco, and S. Vigogna, *Understanding neural networks with reproducing kernel Banach spaces*, Appl. Comput. Harmon. Anal., 62 (2023), pp. 194–236, https://doi.org/10.1016/j.acha.2022.08.006.

[6] C. Boyer, A. Chambolle, Y. De Castro, V. Duval, F. de Gournay, and P. Weiss, *On representer theorems and convex regularization*, SIAM J. Optim., 29 (2019), pp. 1260–1281, https://doi.org/10.1137/18M1200750.

[7] K. Bredies and M. Carioni, *Sparsity of solutions for variational inverse problems with finite-dimensional data*, Calc. Var. Partial Differential Equations, 59 (2020), https://doi.org/10.1007/s00526-019-1658-1.

[8] A. Cohen, I. Daubechies, R. DeVore, G. Kerkyacharian, and D. Picard, *Capturing ridge functions in high dimensions from point queries*, Constr. Approx., 35 (2012), pp. 225–243, https://doi.org/10.1007/s00365-011-9147-6.

[9] A. S. Dalalyan, A. Juditsky, and V. Spokoiny, *A new algorithm for estimating the effective dimension-reduction subspace*, J. Mach. Learn. Res., 9 (2008), pp. 1647–1678.

[10] C. de Boor and R. E. Lynch, *On splines and their minimum properties*, J. Math. Mech., 15 (1966), pp. 953–969.

[11] R. DeVore, R. D. Nowak, R. Parhi, and J. W. Siegel, *Weighted variation spaces and approximation by shallow ReLU networks*, Appl. Comput. Harmon. Anal., 74 (2025), 101713, https://doi.org/10.1016/j.acha.2024.101713.

[12] J. Duchon, *Splines minimizing rotation-invariant semi-norms in Sobolev spaces*, in Constructive Theory of Functions of Several Variables, Springer, Berlin, 1977, pp. 85–100.

[13] S. D. Fisher and J. W. Jerome, *Spline solutions to L^1 extremal problems in one and several variables*, J. Approx. Theory, 13 (1975), pp. 73–83, https://doi.org/10.1016/0021-9045(75)90016-7.

[14] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed., Pure Appl. Math. (Hoboken), Hoboken, New Jersey, 1999.

[15] K. Fukumizu, F. R. Bach, and M. I. Jordan, *Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces*, J. Mach. Learn. Res., 5 (2004), pp. 73–99.

[16] I. M. Gel'fand, M. I. Graev, and N. Y. Vilenkin, *Generalized Functions*, *Vol.* 5: *Integral Geometry and Representation Theory*, Academic Press, New York, 1966.

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

[18] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari, *When do neural networks outperform kernel methods?*, Adv. Neural Inf. Process. Syst., 33 (2020), pp. 14820–14830.

[19] X. Glorot, A. Bordes, and Y. Bengio, *Deep sparse rectifier neural networks*, in Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, 2011, pp. 315–323.

[20] F. B. Gonzalez, *On the range of the Radon d-plane transform and its dual*, Trans. Amer. Math. Soc., 327 (1991), pp. 601–619, https://doi.org/10.2307/2001816.

[21] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, *Maxout networks*, in Proceedings of the International Conference on Machine Learning, 2013, pp. 1319–1327.

[22] P. Grohs, S. Keiper, G. Kutyniok, and M. Schäfer, *α-molecules*, Appl. Comput. Harmon. Anal., 41 (2016), pp. 297–336, https://doi.org/10.1016/j.acha.2015.10.009.

[23] C. Gulcehre, K. Cho, R. Pascanu, and Y. Bengio, *Learned-norm pooling for deep feedforward and recurrent neural networks*, in Machine Learning and Knowledge Discovery in Databases: European Conference, Proceedings, Part I, Springer, New York, 2014, pp. 530–546.

[24] L. Huang, X. Liu, B. Lang, A. Yu, Y. Wang, and B. Li, *Orthogonal weight normalization: Solution to optimization over multiple dependent Stiefel manifolds in deep neural networks*, in Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018.

[25] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu, and L. Shao, *Normalization techniques in training DNNs: Methodology, analysis and application*, IEEE Trans. Pattern Anal. Mach. Intell., 45 (2023), pp. 10173–10196, https://doi.org/10.1109/TPAMI.2023.3250241.

[26] A. Jacot, F. Gabriel, and C. Hongler, *Neural tangent kernel: Convergence and generalization in neural networks*, Adv. Neural Inf. Process. Syst., 31 (2018).

[27] F. Keinert, *Inversion of k-plane transforms and applications in computer tomography*, SIAM Rev., 31 (1989), pp. 273–298, https://doi.org/10.1137/1031051.

[28] S. Keiper, *Approximation of generalized ridge functions in high dimensions*, J. Approx. Theory, 245 (2019), pp. 101–129, https://doi.org/10.1016/j.jat.2019.04.006.

[29] G. S. Kimeldorf and G. Wahba, *Some results on Tchebycheffian spline functions*, J. Math. Anal. Appl., 33 (1971), pp. 82–95, https://doi.org/10.1016/0022-247X(71)90184-3.

[30] V. Kůrková, P. C. Kainen, and V. Kreinovich, *Estimates of the number of hidden units and variation with respect to half-spaces*, Neural Networks, 10 (1997), pp. 1061–1068, https://doi.org/10.1016/S0893-6080(97)00028-2.

[31] V. Kůrková and M. Sanguineti, *Bounds on rates of variable-basis and neural-network approximation*, IEEE Trans. Inform. Theory, 47 (2001), pp. 2659–2665, https://doi.org/10.1109/18.945285.

[32] K.-C. Li, *Sliced inverse regression for dimension reduction*, J. Amer. Statist. Assoc., 86 (1991), pp. 316–327, https://doi.org/10.1080/01621459.1991.10475035.

[33] Q. Li, S. Haque, C. Anil, J. Lucas, R. B. Grosse, and J.-H. Jacobsen, *Preventing gradient attenuation in Lipschitz constrained convolutional networks*, Adv. Neural Inf. Process. Syst., 32 (2019).

[34] R. R. Lin, H. Z. Zhang, and J. Zhang, *On reproducing kernel Banach spaces: Generic definitions and unified framework of constructions*, Acta Math. Sin. (Engl. Ser.), 38 (2022), pp. 1459–1483, https://doi.org/10.1007/s10114-022-1397-7.

[35] H. Liu and W. Liao, *Learning functions varying along a central subspace*, SIAM J. Math. Data Sci., 6 (2024), pp. 343–371, https://doi.org/10.1137/23M1557751.

[36] A. Markoe, *Analytic Tomography*, Encyclopedia Math. Appl. 106, Cambridge University Press, Cambridge, 2006, https://doi.org/10.1017/CBO9780511530012 .

[37] H. N. Mhaskar, *On the tractability of multivariate integration and approximation by neural networks*, J. Complexity, 20 (2004), pp. 561–590, https://doi.org/10.1016/j.jco.2003.11.004.

[38] H. N. Mhaskar, *Approximation by Non-Symmetric Networks for Cross-Domain Learning*, preprint, arXiv:2305.03890, 2023.

[39] H. N. Mhaskar and C. A. Micchelli, *Approximation by superposition of sigmoidal and radial basis functions*, Adv. Appl. Math., 13 (1992), pp. 350–373, https://doi.org/10.1016/0196-8858(92)90016-P.

[40] H. N. Mhaskar and C. A. Micchelli, *Degree of approximation by neural and translation networks with a single hidden layer*, Adv. Appl. Math., 16 (1995), pp. 151–183, https://doi.org/10.1006/aama.1995.1008.

[41] G. Ongie, R. Willett, D. Soudry, and N. Srebro, *A function space view of bounded norm infinite width ReLU nets: The multivariate case*, in Proceedings of the International Conference on Learning Representations, 2020, pp. 1–24.

[42] R. Parhi and R. D. Nowak, *The role of neural network activation functions*, IEEE Signal Process. Lett., 27 (2020), pp. 1779–1783, https://doi.org/10.1109/LSP.2020.3027517.

[43] R. Parhi and R. D. Nowak, *Banach space representer theorems for neural networks and ridge splines*, J. Mach. Learn. Res., 22 (2021).

[44] R. Parhi and R. D. Nowak, *What kinds of functions do deep neural networks learn? Insights from variational spline theory*, SIAM J. Math. Data Sci., 4 (2022), pp. 464–489, https://doi.org/10.1137/21M1418642.

[45] R. Parhi and R. D. Nowak, *Deep learning meets sparse regularization: A signal processing perspective*, IEEE Signal Process. Mag., 40 (2023), pp. 63–74, https://doi.org/10.1109/MSP.2023.3286988.

[46] R. Parhi and R. D. Nowak, *Near-minimax optimal estimation with shallow ReLU neural networks*, IEEE Trans. Inform. Theory, 69 (2023), pp. 1125–1140, https://doi.org/10.1109/TIT.2022.3208653.

[47] R. Parhi and M. Unser, *Distributional extension and invertibility of the k-plane transform and its dual*, SIAM J. Math. Anal., 56 (2024), pp. 4662–4686, https://doi.org/10.1137/23M1556721.

[48] S. Parkinson, G. Ongie, and R. Willett, *ReLU Neural Networks with Linear Layers Are Biased Towards Single- and Multi-Index Models*, preprint, arXiv:2305.15598, 2023.
[49] M. Reed and B. Simon, *Methods of Modern Mathematical Physics* I: *Functional Analysis*, Academic Press, New York, 1972.
[50] S. Rosset, G. Swirszcz, N. Srebro, and J. Zhu, ℓ_1 *regularization in infinite dimensional feature spaces*, in Proceedings of the 20th Annual Conference on Learning Theory, Springer, 2007, pp. 544–558.
[51] B. Rubin, *Inversion of k-plane transforms via continuous wavelet transforms*, J. Math. Anal. Appl., 220 (1998), pp. 187–203, https://doi.org/10.1006/jmaa.1997.5852.
[52] W. Rudin, *Functional Analysis*, 2nd ed., Internat. Ser. Pure Appl. Math., McGraw-Hill, New York, 1991.
[53] S. G. Samko, A. A. Kilbas, and O. I. Marichev, *Fractional integrals and derivatives*, in Theory and Applications, S. M. Nikol'skiĭ, ed., Gordon and Breach, New York, 1993, pp. 143–157.
[54] P. Savarese, I. Evron, D. Soudry, and N. Srebro, *How do infinite width bounded norm networks look in function space?* in Proceedings of the 32nd Annual Conference on Learning Theory, 2019, pp. 2667–2690.
[55] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, Adapt. Comput. Mach. Learn., MIT Press, Cambridge, MA, 2002.
[56] J. Shenouda, R. Parhi, K. Lee, and R. D. Nowak, *Variation spaces for multi-output neural networks: Insights on multi-task learning and network compression*, J. Mach. Learn. Res., 25 (2024), pp. 1–40.
[57] J. W. Siegel and J. Xu, *Characterization of the variation spaces corresponding to shallow neural networks*, Constr. Approx., 57 (2023), pp. 1109–1132, https://doi.org/10.1007/s00365-023-09626-4.
[58] J. W. Siegel and J. Xu, *Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks*, Found. Comput. Math., 24 (2024), pp. 481–537, https://doi.org/10.1007/s10208-022-09595-3.
[59] K. T. Smith, D. C. Solmon, and S. L. Wagner, *Practical and mathematical aspects of the problem of reconstructing objects from radiographs*, Bull. Amer. Math. Soc., 83 (1977), pp. 1227–1270, https://doi.org/10.1090/S0002-9904-1977-14406-6.
[60] D. C. Solmon, *The X-ray transform*, J. Math. Anal. Appl., 56 (1976), pp. 61–83, https://doi.org/10.1016/0022-247X(76)90008-1.
[61] S. Sonoda, I. Ishikawa, and M. Ikeda, *A unified Fourier slice method to derive ridgelet transform for a variety of depth-2 neural networks*, J. Stat. Plan. Inferen., 233 (2024), 106184, https://doi.org/10.1016/j.jspi.2024.106184.
[62] L. Spek, T. J. Heeringa, and C. Brune, *Duality for Neural Networks Through Reproducing Kernel Banach Spaces*, preprint, arXiv:2211.05020, 2022.
[63] I. Steinwart, *Sparseness of support vector machines*, J. Mach. Learn. Res., 4 (2003), pp. 1071–1105.
[64] F. Trèves, *Topological Vector Spaces, Distributions and Kernels*, Academic Press, New York-London, 1967.
[65] M. Unser, *A unifying representer theorem for inverse problems and machine learning*, Found. Comput. Math., 21 (2021), pp. 941–960, https://doi.org/10.1007/s10208-020-09472-x.
[66] M. Unser, *From kernel methods to neural networks: A unifying variational formulation*, Found. Comput. Math., 24 (2024), pp. 1779–1818, https://doi.org/10.1007/s10208-023-09624-9.
[67] M. Unser, *Ridges, neural networks, and the Radon transform*, J. Mach. Learn. Res., 24 (2023).
[68] M. Unser and S. Aziznejad, *Convex optimization in sums of Banach spaces*, Appl. Comput. Harmon. Anal., 56 (2022), pp. 1–25, https://doi.org/10.1016/j.acha.2021.07.002.
[69] M. Unser, J. Fageot, and J. P. Ward, *Splines are universal solutions of linear inverse problems with generalized TV regularization*, SIAM Rev., 59 (2017), pp. 769–793, https://doi.org/10.1137/16M1061199.
[70] G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, 1990.
[71] H. Wendland, *Scattered Data Approximation*, Cambridge Monogr. Appl. Comput. Math. 17, Cambridge University Press, Cambridge, 2005.
[72] H. Zhang, Y. Xu, and J. Zhang, *Reproducing kernel Banach spaces for machine learning*, J. Mach. Learn. Res., 10 (2009), pp. 3520–3527.