SIAM Review
Vol. 68, No. 1, pp. 127–149

Compositional Function Spaces for Deep Learning*

Rahul Parhi†
Robert D. Nowak‡

Abstract. We present a variational framework for studying functions learned by deep neural networks with rectified linear unit nonlinearities. We introduce a function space built from compositions of functions of second-order Radon-domain bounded variation. The compositional form of these functions captures the structure of deep neural networks. We prove a representer theorem that shows that deep neural networks with finite width solve regularized data-fitting problems over this space. The critical width is controlled by the square of the number of training data. This perspective explains the effect of weight-decay regularization in neural network training, the importance of skip connections, and the role of sparsity in neural networks. By considering the function-space perspective, we provide sharp links between deep learning and variational methods.

Key words. deep learning, neural networks, regularization, representer theorem, sparsity

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

DOI. 10.1137/25M1802948

Contents

*Published electronically February 9, 2026. This paper originally appeared in *SIAM Journal on Mathematics of Data Science,* Volume 4, Number 2, 2022, pages 464–489, under the title "What Kinds of Functions Do Deep Neural Networks Learn? Insights from Variational Spline Theory."
https://doi.org/10.1137/25M1802948
Funding: The original work [52] was partially supported by NSF grants DMS-2134140 and CCF-2427440, ONR MURI grant N00014-20-1-2787, AFOSR/AFRL grant FA9550-18-1-0166, and the NSF Graduate Research Fellowship Program under grant DGE-1747503. This SIGEST article is partially supported by NSF grant CCF-2427440.
†Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093 USA (rahul@ucsd.edu).
‡Department of Electrical and Computer Engineering, University of Wisconsin–Madison, Madison, WI 53705 USA (rdnowak@wisc.edu).

1. Introduction. The goal of statistics and machine learning is to learn an approximation of data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^D$ via a function $f : \mathbb{R}^d \to \mathbb{R}^D$ such that $f(\boldsymbol{x}_i) \approx \boldsymbol{y}_i$; for prediction problems in particular, given a new sample $(\boldsymbol{x}_{\text{new}}, \boldsymbol{y}_{\text{new}}) \in \mathbb{R}^d \times \mathbb{R}^D$, the goal is also that $f(\boldsymbol{x}_{\text{new}}) \approx \boldsymbol{y}_{\text{new}}$. Without any assumptions on the problem, there are infinitely many functions that can agree with any given set of data arbitrarily well. Therefore, this problem is inherently *ill-posed.* To circumvent this issue, some form of *regularization* must be imposed on the learning problem. In this paper, we take the perspective of *explicit regularization* as opposed to that of *implicit regularization.* Thus, our perspective is agnostic to any optimization algorithm.

A classical solution to this problem is to use *kernel methods.* These are functions that are solutions to function-space optimization problems over reproducing kernel Hilbert spaces (RKHSs) [1, 61, 72]. While these optimization problems are defined on infinite-dimensional spaces, the RKHS representer theorem guarantees the existence of a unique, parametric solution to the problem. The parametric form lies in the span of reproducing kernels centered at the data sites, i.e., it takes the form of a *kernel machine.* This allows the problem to be recast as a finite-dimensional problem on kernel machine coefficients. However, the dawn of the deep learning era has shown that deep neural networks often outperform kernel methods in a wide variety of tasks ranging from speech recognition [29] to image classification [32]. Moreover, deep neural networks are the foundation of large language models and related systems at the forefront of modern artificial intelligence. Thus, there is great interest in developing a theory of deep learning that parallels our current understanding of kernel methods.

Toward this goal, our prior works [51, 52] developed *Banach-space representer theorems* for neural networks with rectified linear unit (ReLU) activation functions by studying optimization problems on the Banach space of functions of second-order Radon-domain bounded variation (which we denote by $\mathscr{R}\mathrm{BV}^2$). The ReLU activation function is a common choice in contemporary neural networks. In the univariate case, this space coincides with the classical second-order bounded variation space. In that setting, the neural network solutions are exactly the well-known locally adaptive linear splines [23, 40, 71]. In the multivariate case, this space is fundamentally distinct from those studied in analysis [22, 54, 64, 65].

In our prior works [51, 52], we established that optimization problems over this space can be recast as finite-dimensional neural network training problems. In particular, the solutions to minimizing the sum of losses/errors of a neural network model plus a regularization term proportional to the sum of squared neural network weights

are solutions to these problems. This form of regularization corresponds to the commonly used technique of *weight decay* [33] in gradient-descent methods for neural network training. Therefore, the Banach space of functions of second-order Radon-domain bounded variation exactly captures the *regularity* of finite-width neural networks trained with weight decay.

In this paper, we revisit the characterization and extensions of this space to deep (multilayer) neural networks with ReLU activation functions from the original version of the present paper [52]. A special property of deep ReLU networks is that their input-output relation is continuous piecewise linear [43]. The reverse is also true in that any continuous piecewise-linear function can be represented with a sufficiently wide and deep ReLU network [2]. Thus, one can interpret a deep ReLU network as a multivariate spline of degree one. This connection between deep neural networks and splines has been observed by a number of authors [3, 6, 7, 15, 51, 52, 54, 67]. In particular, one can view a deep neural network as a *hierarchical spline* [58] to emphasize the compositional nature of deep neural networks. Due to this special property, we will work exclusively with ReLU activation functions.

1.1. Contributions. This paper presents a variational framework to understand the properties of functions learned by deep neural networks fit to data and it revisits the compositional function spaces introduced in [52]. We prove a *representer theorem* for fully connected feedforward deep ReLU networks; i.e., we show that solutions to certain function-space optimization problems are exactly realizable by such networks. These solutions have skip connections and rank-controlled weight matrices. The contributions of this paper are as follows:

1. We extend the scalar-valued second-order Radon-domain bounded variation space of [51] to functions that are vector-valued while maintaining a tight connection to weight-decay regularization. We then consider the compositional version of this space whose members are compositions of functions from the vector-valued spaces.
2. We prove a representer theorem that shows that deep ReLU networks with skip connections and rank-controlled weight matrices are solutions to regularized data-fitting problems over functions from this compositional space. Furthermore, we show that the critical width is controlled by the square of the number of training data.
3. We show that the function-space problem can be recast as a deep ReLU network training problem. The regularizer for the function-space problem can be expressed in terms of neural network parameters and coincides with notions of weight-decay and path-norm regularization. This provides insight into the kind of regularity these common regularization schemes impose on the learned functions.

1.2. Connections to Empirical Studies in Deep Learning. Our results provide theoretical support and insight for a number of empirical findings in deep learning. We show that the common regularization method of weight decay corresponds to Radon-domain total variation (TV) regularization. This characterizes the function-space properties of neural networks trained with weight decay—the functions they represent are regular in a precise sense. The optimal solutions to the function-space problem require "skip connections" between layers, which provides a new theoretical explanation for their benefits in practice [27]. The sparse nature of our solutions sheds new light on the roles of sparsity and redundancy in deep learning, ranging from "drop-out" [30] to the "lottery ticket hypothesis" [25]. Finally, rank-controlled weight

matrices are a natural by-product of our variational framework that has precedent in practical studies of deep neural networks. Indeed, it has been empirically observed that low-rank weight matrices can speed up learning [4] and improve the accuracy [26], robustness [59], and computational efficiency [73] of deep neural networks.

1.3. Related Prior Work. The function-space perspective of neural networks has received a lot of renewed attention due to the seminal works of Savarese et al. [60] and Ongie et al. [48]. In approximation theory, the function-space perspective played a key role dating back to the 1990s with the breakthrough work of Barron [8, 9] and subsequent work on variation spaces [5, 35, 36, 42]. Building on the work in [48, 60], the papers [51, 52] proved representer theorems for shallow and deep ReLU networks for optimization problems posed over certain function spaces. The techniques used there are rooted in spline theory and the study of continuous-domain inverse problems [23, 40, 78]. A common theme in all of these works is to leverage the sparsifying nature of TV regularization on spaces of measures to learn *sparse solutions*. Other related work includes "deep kernel learning" [14] where the authors consider compositional function spaces based on RKHSs. Another line of research considers compositional functions from variation spaces [21] to study the approximation theory of deep neural networks.

1.4. Subsequent Work. Since the original version of the present paper [52] there has been a growing body of research that considers the function-space perspective and $\mathscr{R}\,\mathrm{BV}^2$-type spaces. A number of these directions is briefly discussed below.

Reproducing Kernel Banach Spaces and Representer Theorems. Recent papers have extended the representer theorem framework to reproducing kernel Banach spaces (RKBSs), providing a kernel-based/feature-map perspective parallel to the $\mathscr{R}\,\mathrm{BV}^2$ framework. Notable works include [12, 11, 66, 75] that specifically consider nonreflexive RKBSs and sparsity-promoting norms. As noted in [12], when specializing to ReLU atoms, these RKBSs are equivalent (as Banach spaces) to $\mathscr{R}\,\mathrm{BV}^2$. Other directions have considered generalizations of $\mathscr{R}\,\mathrm{BV}^2$ by considering general families of activation functions [69] and neurons with multivariate nonlinearities [56]. Closer to the setup of the present paper (and its original version), there have also been recent efforts toward studying deep and compositional versions of these spaces [13, 18, 28, 74].

Approximation Theory. The approximation properties of $\mathscr{R}\,\mathrm{BV}^2$-type spaces have received considerable recent attention. In particular, [65] established sharp bounds on approximation rates, metric entropy, and n-widths of $\mathscr{R}\,\mathrm{BV}^2$ and related neural Banach spaces such as Barron spaces and variation spaces. Building on that work, the authors of [19] showed that shallow neural networks provide optimal approximation rates for a broader class of functions than the standard variation spaces by introducing the notion of *weighted* variation spaces. It still remains an open problem to precisely characterize the approximation spaces of shallow neural networks, though these works are making progress in that direction. Other works have also focused on relating $\mathscr{R}\,\mathrm{BV}^2$-type spaces to other spaces studied in approximation theory [54, 34, 41, 64, 22].

Statistical Learning Theory. The function-space perspective has also been fruitful for studying statistical properties of neural networks and related methods. The authors of [54] show that shallow ReLU networks trained with weight decay (to a global minimizer) are minimax optimal for learning functions in $\mathscr{R}\,\mathrm{BV}^2$. These ideas were extended in [76] for nonparametric function estimation in $\mathscr{R}\,\mathrm{BV}^2$-type spaces with various neural architectures, in [37] for classification problems with $\mathscr{R}\,\mathrm{BV}^2$ decision boundaries, and in [38] to study the generalization properties of flat minima of neural network training problems. More recently, the authors of [57] showed that ReLU

neural networks are especially well suited to learning single- and multi-index models. The authors of [49] build on the $\mathscr{R}\mathrm{BV}^2$ framework to develop nonparametric goodness-of-fit tests, while the authors of [50] build on the framework to study the non-Gaussianity of randomly initialized neural networks. Other works apply insights from [52] to study the inductive bias of neural networks used for denoising problems [77] and implicit neural representations [44, 63].

1.5. Roadmap. In section 2 we introduce the notation and mathematical preliminaries used in the remainder of the paper and extend the results of [51] to vector-valued functions. In section 3 we prove our main result, the representer theorem for deep ReLU networks. In section 4 we discuss applications of our representer theorem to the training and regularization of deep ReLU networks.

2. Mathematical Preliminaries. Let X be a locally compact Hausdorff space. The Riesz–Markov–Kakutani representation theorem says that $\mathcal{M}(X)$, the Banach space of finite Radon measures on X, is the continuous dual of $C_0(X)$, the Banach space of continuous functions vanishing at infinity equipped with the $L^\infty(X)$-norm [24, Chapter 7]. In particular, it holds that

$$\|u\|_{\mathcal{M}} = \sup_{\substack{\varphi \in C_0(X) \\ \|\varphi\|_{L^\infty}=1}} \langle u, \varphi \rangle, \tag{2.1}$$

where $\langle \cdot, \cdot \rangle$ denotes the canonical pairing between $\mathcal{M}(X)$ and $C_0(X)$.

The $\mathcal{M}$-norm is exactly the TV norm (in the sense of measures). The pairing $\langle \cdot, \cdot \rangle$ in (2.1) is concretely specified as the integral

$$\langle u, \varphi \rangle = \int_X \varphi(\boldsymbol{x}) \, \mathrm{d}u(\boldsymbol{x}). \tag{2.2}$$

Furthermore, $\mathcal{M}(X)$ can be viewed as a "generalization" of $L^1(X)$ in the sense that for any $f \in L^1(X)$, $\|f\|_{L^1(X)} = \|f\|_{\mathcal{M}(X)}$, but $\mathcal{M}(X)$ is a strictly larger space that also includes the shifted Dirac impulses $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in X$, with the property that $\|\delta(\cdot - \boldsymbol{x}_0)\|_{\mathcal{M}(X)} = 1$. We also remark that the $\mathcal{M}$-norm is the continuous-domain analogue of the ℓ^1-norm and is thus sparsity-promoting. In this paper, we mostly work with $X = \mathbb{S}^{d-1} \times \mathbb{R}$, where $\mathbb{S}^{d-1} = \{\boldsymbol{u} \in \mathbb{R}^d : \|\boldsymbol{u}\|_2 = 1\}$, the $(d-1)$-sphere.

The Radon transform of a function $f \in L^1(\mathbb{R}^d)$ is given by

$$\mathscr{R}\{f\}(\boldsymbol{w}, b) := \int_{\{\boldsymbol{x} : \boldsymbol{w}^\mathsf{T}\boldsymbol{x} = b\}} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \quad (\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}, \tag{2.3}$$

where $\mathrm{d}\boldsymbol{x}$ is the integration against the $(d-1)$-dimensional Lebesgue measure on the hyperplane $\{\boldsymbol{x} : \boldsymbol{w}^\mathsf{T}\boldsymbol{x} = b\}$. Observe that the Radon domain is parameterized by a *direction* $\boldsymbol{w} \in \mathbb{S}^{d-1}$ and an *offset* $b \in \mathbb{R}$. When working with the Radon transform of functions defined on $\mathbb{R}^d$, the following *ramp filter* arises in the Radon inversion formula:

$$\mathrm{K} = \left(-\partial_t^2\right)^{\frac{d-1}{2}}, \tag{2.4}$$

where ∂_t denotes the partial derivative with respect to the offset variable of the Radon domain and fractional powers are understood via the Fourier transform. In this paper, we require the *distributional extension* of the Radon transform. It turns out that the Radon transform is well defined for any tempered distribution $f \in \mathscr{S}'(\mathbb{R}^d)$ [39, 55].

2.1. Function Spaces of Scalar-Valued Shallow ReLU Networks. The main contribution of [51] is a representer theorem for shallow ReLU networks with scalar outputs. That work showed that finite-width ReLU networks with width bounded by the number of training data are solutions to regularized data-fitting problems over the second-order Radon-domain bounded variation space. In this section we review that result, collect some relevant properties of this function space, and provide a short proof of [51, Theorem 1].

The space of functions of second-order bounded variation in the Radon domain is then given by

$$\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d) = \left\{ f : \mathbb{R}^d \to \mathbb{R} \text{ measurable} : \begin{array}{c} \mathscr{R}\,\mathrm{TV}^2(f) < \infty \\ \operatorname*{ess\,sup}_{\boldsymbol{x}\in\mathbb{R}^d} |f(\boldsymbol{x})|(1+\|\boldsymbol{x}\|_2)^{-1} < \infty \end{array} \right\}, \tag{2.5}$$

where

$$\mathscr{R}\,\mathrm{TV}^2(f) = c_d \|\partial_t^2\,\mathrm{K}\mathscr{R} f\|_{\mathcal{M}} \tag{2.6}$$

denotes the second-order TV of a function with respect to the offset variable of the Radon domain, where $c_d^{-1} = 2(2\pi)^{d-1}$ is a constant that arises when working with the Radon transform. This quantity is a *seminorm* whose null space (when restricted to $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$) is the space of affine functions. All the operators that appear in (2.6) must be understood in the distributional sense. We refer the reader to [51, section 3] and [39, 55] for details about the distributional extension of the Radon transform and related operators.

Proposition 2.1 (special case of [51, Theorem 1]). *Consider the problem of interpolating the data* $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}$. *Then, under the hypothesis that* $y_i = y_j$ *whenever* $\boldsymbol{x}_i = \boldsymbol{x}_j$, *the solution set to*

$$\min_{f\in\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} \mathscr{R}\,\mathrm{TV}^2(f) \quad s.t. \quad f(\boldsymbol{x}_i) = y_i, \quad i = 1,\dots,N, \tag{2.7}$$

is the weak-closure*[1] *of the convex hull of its extreme points which all take the form*

$$f_{\mathrm{ReLU}}(\boldsymbol{x}) = \sum_{k=1}^K v_k(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k)_+ + \boldsymbol{c}^\mathsf{T}\boldsymbol{x} + c_0, \tag{2.8}$$

where $K \le N$, $(\cdot)_+ := \max\{0,\cdot\}$ *denotes the ReLU,* $v_k \in \mathbb{R}\setminus\{0\}$, $\boldsymbol{w}_k \in \mathbb{S}^{d-1}$, $b_k \in \mathbb{R}$, $\boldsymbol{c} \in \mathbb{R}^d$, *and* $c_0 \in \mathbb{R}$.

Remark 2.2. Proposition 2.1 says that there always exists a solution to (2.7) that is realizable by a shallow ReLU network with a *skip connection* [27], which is the affine term in (2.8). In other words, Proposition 2.1 is a *representer theorem* for shallow ReLU networks.

Remark 2.3. The fact that $\boldsymbol{w}_k \in \mathbb{S}^{d-1}$ in (2.8) does not constrain the network is due to the positive homogeneity of the ReLU. Indeed, given any shallow ReLU network with $\boldsymbol{w}_k \in \mathbb{R}^d\setminus\{\boldsymbol{0}\}$, we can use the fact that ReLU is positively homogeneous of degree 1 to rewrite the implementation as

[1] In particular, $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ is a dual Banach space and the weak*-topology coincides with the weak topology induced by its predual, which is explicated in Lemma 2.4.

$$\boldsymbol{x} \mapsto \sum_{k=1}^{K} v_k \|\boldsymbol{w}_k\|_2 (\widetilde{\boldsymbol{w}}_k^{\mathsf{T}} \boldsymbol{x} - \widetilde{b}_k)_+ + \boldsymbol{c}^{\mathsf{T}} \boldsymbol{x} + c_0, \tag{2.9}$$

where $\widetilde{\boldsymbol{w}}_k := \boldsymbol{w}_k / \|\boldsymbol{w}_k\|_2 \in \mathbb{S}^{d-1}$ and $\widetilde{b}_k := b_k / \|\boldsymbol{w}_k\|_2 \in \mathbb{R}$.

The $\mathscr{R}\,\mathrm{TV}^2$-seminorm has an explicit description in terms of network parameters. Indeed, by [51, Lemma 25] if

$$f_{\mathrm{ReLU}}(\boldsymbol{x}) = \sum_{k=1}^{K} v_k (\boldsymbol{w}_k^{\mathsf{T}} \boldsymbol{x} - b_k)_+ + \boldsymbol{c}^{\mathsf{T}} \boldsymbol{x} + c_0, \tag{2.10}$$

and the network is written in *reduced form,* i.e., the input weights and biases $\{(\boldsymbol{w}_k, b_k)\}_{k=1}^K$ are unique up to certain symmetries [51], then

$$\mathscr{R}\,\mathrm{TV}^2(f_{\mathrm{ReLU}}) = \sum_{k=1}^{K} |v_k| \|\boldsymbol{w}_k\|_2. \tag{2.11}$$

If the network is not in reduced form, then an infimum above the right-hand side must be taken over all representations (cf. [62, Equation (6)]). Equation (2.11) is sometimes referred to as the *path-norm* of the network [45]. Moreover, we see that (2.11) is a kind of ℓ^1-norm on the network parameters, giving insight into the sparsity-promoting aspect of the $\mathscr{R}\,\mathrm{TV}^2$-seminorm on the network, where the sparsity is with respect to the number of active neurons. Equation (2.11) also reveals that path-norm regularization is equivalent to $\mathscr{R}\,\mathrm{TV}^2$-regularization, which is subsequently equivalent to weight-decay regularization [53]. Thus, weight-decay regularization favors functions that are regular in the sense of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$.

Recall that $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ is defined by a seminorm and the null space of $\mathscr{R}\,\mathrm{TV}^2(\cdot)$ is nontrivial; it is the space of affine functions on $\mathbb{R}^d$. It turns out that $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ can be turned into a *bona fide* Banach space when equipped with an appropriate norm. We collect some of the relevant properties of this Banach space in the next lemma.

LEMMA 2.4. *The space $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ equipped with the norm*

$$\|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} := \mathscr{R}\,\mathrm{TV}^2(f) + |f(\mathbf{0})| + \sum_{k=1}^{d} |f(\boldsymbol{e}_k) - f(\mathbf{0})|, \tag{2.12}$$

where $\{\boldsymbol{e}_k\}_{k=1}^d$ denotes the canonical basis of $\mathbb{R}^d$, has the following properties:

1. *It is a Banach space.*
2. *It is a dual Banach space and, in particular, for any $\boldsymbol{x}_0 \in \mathbb{R}^d$, the point evaluation functional $f \mapsto f(\boldsymbol{x}_0)$ is weak* continuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$.*
3. *Modulo affine functions, the extreme points of the unit ball*

$$\{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d) : \mathscr{R}\,\mathrm{TV}^2(f) \le 1\} \tag{2.13}$$

take the form $\{\boldsymbol{x} \mapsto \pm(\boldsymbol{w}^{\mathsf{T}} \boldsymbol{x} - b)_+\}_{(\boldsymbol{w},b) \in \mathbb{S}^{d-1} \times \mathbb{R}}$.

The proof of Lemma 2.4 appears in Appendix A. With this lemma, we immediately have a short proof of Proposition 2.1.

Proof of Proposition 2.1. From item 2 in Lemma 2.4, we can endow $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ with a weak* topology. The remainder of the proof relies on recent characterizations of abstract representer theorems (see [16, 17, 68, 70]). From the weak*-continuity

of the point evaluation functional on $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ (item 2 in Lemma 2.4), our setting coincides with the hypotheses of [70, Theorem 3]. First, this abstract result ensures that the solution set is nonempty, convex, and weak*-compact. Second, it ensures that it is the weak*-closure of the convex hull of its extreme points, which can all be expressed as

$$f_{\text{extreme}} = \sum_{k=1}^{K} v_k e_k + q, \tag{2.14}$$

where the number K of atoms satisfies $K \leq N$, q is in the null space of the regularizer (i.e., it is an affine function), and, for $k = 1, \ldots, K$, $v_k \in \mathbb{R} \setminus \{0\}$ and e_k is an extreme point of the unit regularization ball. The characterization of extreme points in item 3 in Lemma 2.4 completes the proof. ∎

2.2. Function Spaces of Vector-Valued Shallow ReLU Networks. Since a deep neural network is built from compositions of vector-valued shallow neural networks, we require the extension of the function spaces of scalar-valued networks to the vector-valued case and corresponding representer theorems. The extension of the result in subsection 2.1 to the vector-valued case follows from standard techniques with a slight nuance in the choice of vector norm to maintain tight connections to weight-decay and path-norm regularization.

Clearly, the vector-valued space is the D-fold Cartesian product

$$\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D) := \underbrace{\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d) \times \cdots \times \mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)}_{D \text{ times}}. \tag{2.15}$$

The primary challenge is endowing $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$ with a (semi-)norm that maintains tight connections with weight decay. By definition, every $f = (f_1, \ldots, f_D) \in \mathscr{R}\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$ satisfies

$$\partial_t^2 \mathrm{K}\mathscr{R} f_j \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R}) \tag{2.16}$$

for all $j \in \{1, \ldots, D\} =: [D]$. Write

$$\underline{\partial_t^2 \mathrm{K}\mathscr{R}} f := \begin{bmatrix} \partial_t^2 \mathrm{K}\mathscr{R} f_1 \\ \vdots \\ \partial_t^2 \mathrm{K}\mathscr{R} f_D \end{bmatrix} \tag{2.17}$$

for the vectorized version of $\partial_t^2 \mathrm{K}\mathscr{R}$. The problem now reduces to choosing a norm on the *vector-valued measure* $\underline{\partial_t^2 \mathrm{K}\mathscr{R}} f \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R};\mathbb{R}^D)$. Given $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_D) \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R};\mathbb{R}^D)$, consider the norm

$$\|\boldsymbol{\nu}\|_{2,\mathcal{M}} := \sup_{\mathbb{S}^d = \bigcup_{i=1}^n A_i} \sum_{i=1}^{n} \|\boldsymbol{\nu}(A_i)\|_2 = \sup_{\substack{\mathbb{S}^d = \bigcup_{i=1}^n A_i \\ n \in \mathbb{N}}} \sum_{i=1}^{n} \left(\sum_{j=1}^{D} |\nu_j(A_i)|^2 \right)^{1/2}. \tag{2.18}$$

With this choice of norm, $(\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R};\mathbb{R}^D), \|\cdot\|_{2,\mathcal{M}})$ is a Banach space. We refer the reader to the monograph [20] for a full treatment of vector-valued measures and the accompanying results. This leads to a seminorm for $f = (f_1, \ldots, f_D) \in \mathscr{R}\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$ defined by

$$\mathscr{R}\mathrm{TV}^2(f) := \|\underline{\partial_t^2 \mathrm{K}\mathscr{R}} f\|_{2,\mathcal{M}} = \left\| \begin{bmatrix} \partial_t^2 \mathrm{K}\mathscr{R} f_1 \\ \vdots \\ \partial_t^2 \mathrm{K}\mathscr{R} f_D \end{bmatrix} \right\|_{2,\mathcal{M}}. \tag{2.19}$$

The choice of norm in (2.18) is one of many possible choices of (equivalent) norms on $\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R}; \mathbb{R}^D)$. It turns out that this is the only choice of norm that guarantees tight connections between the $\mathscr{R}\,\mathrm{TV}^2$ of a function and weight-decay regularization (see the discussion in [62, Appendix A] for more details). We explicate this in (2.26).

Analogous to Lemma 2.4, we can turn $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d; \mathbb{R}^D)$ into a *bona fide* Banach space when equipped with an appropriate norm. We collect some of the relevant properties of this Banach space in the next lemma.

LEMMA 2.5. *The space $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d; \mathbb{R}^D)$ equipped with the norm*

$$\|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)} := \mathscr{R}\,\mathrm{TV}^2(f) + \sum_{j=1}^{D} \left(|f_j(\mathbf{0})| + \sum_{k=1}^{d} |f_j(\boldsymbol{e}_k) - f_j(\mathbf{0})| \right) \tag{2.20}$$

has the following properties:

1. *It is a Banach space.*
2. *It is a dual Banach space and, in particular, for any $\boldsymbol{x}_0 \in \mathbb{R}^d$ and $j \in [D]$, the componentwise point evaluation functional $f \mapsto [f(\boldsymbol{x}_0)]_j$ is weak* continuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d; \mathbb{R}^D)$.*
3. *Modulo affine functions, the extreme points of the unit ball*

$$\{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d; \mathbb{R}^D) : \mathscr{R}\,\mathrm{TV}^2(f) \leq 1\} \tag{2.21}$$

take the form $\{\boldsymbol{x} \mapsto \boldsymbol{u}(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)_+\}_{(\boldsymbol{u},\boldsymbol{w},b) \in \mathbb{S}^{D-1} \times \mathbb{S}^{d-1} \times \mathbb{R}}$.

The proof of Lemma 2.5 appears in Appendix B.

LEMMA 2.6. *Let $f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d; \mathbb{R}^D)$; then f is Lipschitz continuous and satisfies the Lipschitz bound*

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\|_2 \leq \|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)} \|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{2.22}$$

The proof of Lemma 2.6 appears in Appendix C.

THEOREM 2.7. *Consider the problem of interpolating the data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N \subset \mathbb{R}^d \times \mathbb{R}^D$. Then, under the hypothesis that $\boldsymbol{y}_i = \boldsymbol{y}_j$ whenever $\boldsymbol{x}_i = \boldsymbol{x}_j$, the solution set to*

$$\min_{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)} \mathscr{R}\,\mathrm{TV}^2(f) \quad s.t. \quad f(\boldsymbol{x}_i) = \boldsymbol{y}_i, \quad i = 1, \ldots, N, \tag{2.23}$$

is the weak-closure of the convex hull of its extreme points, which all take the form*

$$f_{\mathrm{ReLU}}(\boldsymbol{x}) = \sum_{k=1}^{K} \boldsymbol{v}_k (\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k)_+ + \mathbf{C}\boldsymbol{x} + \boldsymbol{c}_0, \tag{2.24}$$

where $K \leq ND$, $(\cdot)_+$ denotes the ReLU, $\boldsymbol{v}_k \in \mathbb{R}^D \setminus \{\mathbf{0}\}$, $\boldsymbol{w}_k \in \mathbb{S}^{d-1}$, $b_k \in \mathbb{R}$, $\mathbf{C} \in \mathbb{R}^{D \times d}$, and $\boldsymbol{c}_0 \in \mathbb{R}^D$. In particular, there always exists a solution with $K \leq \min\{N^2, ND\}$.

The proof of Theorem 2.7 appears in Appendix D. What is remarkable here is the existence of a solution that always satisfies the bound $K \leq N^2$, independent of the output dimension D. As discussed in Remark 2.3, the fact that $\boldsymbol{w}_k \in \mathbb{S}^{d-1}$ does not constrain the network is due to the positive homogeneity of the ReLU. Indeed, given any shallow ReLU network with $\boldsymbol{w}_k \in \mathbb{R}^d \setminus \{\mathbf{0}\}$, we can rewrite the implementation as

$$f_{\mathrm{ReLU}}(\boldsymbol{x}) = \sum_{k=1}^{K} \boldsymbol{v}_k \|\boldsymbol{w}_k\|_2 (\widetilde{\boldsymbol{w}}_k^\mathsf{T}\boldsymbol{x} - \widetilde{b}_k)_+ + \mathbf{C}\boldsymbol{x} + \boldsymbol{c}_0, \tag{2.25}$$

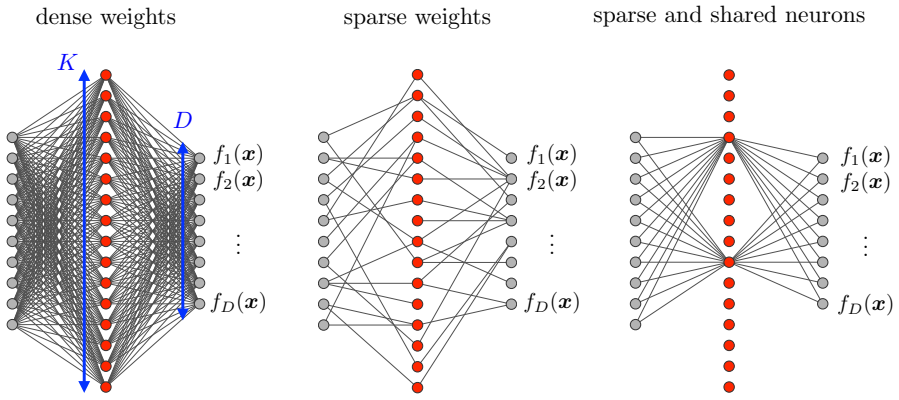

Fig. 1 *Three neural networks with different weight-sparsity patterns. Since weight decay minimizes the* $\mathscr{R}\,\mathrm{TV}^2$ *of the network, it favors the rightmost architecture. This architecture exhibits both neuron sparsity and neuron sharing. Each output depends on the same few neurons. This observation also gives insight into the regularity of the optimal functions: They favor functions that only vary in a few directions across all outputs. This is in contrast with the middle network, where each output has variation in a small number of directions, but this set of directions can be different for each output. This neuron sharing phenomenon is rigorously quantified in* [62, Theorem 9].

where $\widetilde{\boldsymbol{w}}_k := \boldsymbol{w}_k/\|\boldsymbol{w}_k\|_2 \in \mathbb{S}^{d-1}$ and $\widetilde{b}_k := b_k/\|\boldsymbol{w}_k\|_2 \in \mathbb{R}$. Thanks to the definition of the vector-valued $\mathscr{R}\,\mathrm{TV}^2$ (2.19) when the network is in reduced form, we have that

$$\mathscr{R}\,\mathrm{TV}^2(f_{\mathrm{ReLU}}) = \sum_{k=1}^{K} \|\boldsymbol{v}_k\|_2 \|\boldsymbol{w}_k\|_2. \tag{2.26}$$

Due to the equivalence of path-norm and weight-decay regularization [53], we see that, analogous to the scalar-valued case, weight-decay regularization favors functions that are regular in the sense of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$. What is even more interesting is that this shows that weight-decay regularization favors networks that *share neurons.* This arises due to the fact that the path-norm in (2.26) is essentially a multitask lasso regularizer [62]. This is summarized in Figure 1.

3. A Representer Theorem for Deep ReLU Networks. In this section, we prove our representer theorem for deep ReLU networks. We consider functions that are compositions of functions from the Banach space defined in Lemma 2.5. Let

$$\begin{aligned} &\mathscr{R}\,\mathrm{BV}^2_{\mathsf{deep}}(\mathbb{R}^{d_0};\ldots;\mathbb{R}^{d_L}) \\ &\quad := \{f = f^{(L)} \circ \cdots \circ f^{(1)} : \ f^{(\ell)} \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell}), \ell = 1,\ldots,L\} \end{aligned} \tag{3.1}$$

denote the space of all such functions.

For brevity, we write $\mathscr{R}\,\mathrm{BV}^2_{\mathsf{deep}}(L)$ for $\mathscr{R}\,\mathrm{BV}^2_{\mathsf{deep}}(\mathbb{R}^{d_0};\ldots;\mathbb{R}^{d_L})$. This specification reflects two standard architectural hyperparameters for deep neural networks: the number of hidden layers L and the functional "widths," d_ℓ, of each layer. That is, each function in the composition ultimately corresponds to a layer in a deep neural network.

LEMMA 3.1. *Let* $f = f^{(L)} \circ \cdots \circ f^{(1)} \in \mathscr{R}\,\mathrm{BV}^2_{\mathsf{deep}}(L)$. *Then,* f *is Lipschitz continuous and satisfies the Lipschitz bound*

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\|_2 \leq \left(\prod_{\ell=1}^{L} \|f^{(\ell)}\|_{\mathscr{R}\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})} \right) \|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{3.2}$$

Proof. The result follows by repeated application of Lemma 2.6. ∎

This lemma reveals that the "norm" on the compositional space controls the Lipschitz regularity of the underlying function. We now state our representer theorem for deep ReLU networks.

THEOREM 3.2. *Let L be a positive integer corresponding to the depth of a deep ReLU network and let $d_0, \ldots, d_L$ be positive integers corresponding to the intermediate dimensions of the network. Consider the problem of learning from the data $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^N \subset \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$. Let $\mathcal{L}(\cdot,\cdot)$ be an arbitrary loss function and let $\lambda > 0$ be a regularization parameter. If a solution exists to the problem*

$$\min_{\substack{f^{(\ell)} \in \mathscr{R}\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell}) \\ \ell=1,\ldots,L \\ f = f^{(L)} \circ \cdots \circ f^{(1)}}} \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{y}_i, f(\boldsymbol{x}_i)) + \lambda \sum_{\ell=1}^{L} \mathscr{R}\,\mathrm{TV}^2(f^{(\ell)}), \tag{3.3}$$

then there exists a solution of the form

$$f_{\mathrm{deep}}(\boldsymbol{x}) = \boldsymbol{x}^{(L)}, \tag{3.4}$$

where $\boldsymbol{x}^{(L)}$ is computed recursively via

$$\begin{cases} \boldsymbol{x}^{(0)} := \boldsymbol{x}, \\ \boldsymbol{x}^{(\ell)} := \mathbf{V}^{(\ell)} \boldsymbol{\rho}(\mathbf{W}^{(\ell)} \boldsymbol{x}^{(\ell-1)} - \boldsymbol{b}^{(\ell)}) + \mathbf{C}^{(\ell)} \boldsymbol{x}^{(\ell-1)} + \boldsymbol{c}_0^{(\ell)}, \quad \ell = 1, \ldots, L, \end{cases} \tag{3.5}$$

where $\boldsymbol{\rho}$ applies the ReLU $(\cdot)_+$ componentwise and, for $\ell = 1, \ldots, L$, $\mathbf{V}^{(\ell)} \in \mathbb{R}^{d_\ell \times K^{(\ell)}}$, $\mathbf{W}^{(\ell)} \in \mathbb{R}^{K^{(\ell)} \times d_{\ell-1}}$, $\boldsymbol{b}^{(\ell)} \in \mathbb{R}^{K^{(\ell)}}$, $\mathbf{C}^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, and $\boldsymbol{c}_0^{(\ell)} \in \mathbb{R}^{d_\ell}$, where $K^{(\ell)} \leq \min\{N^2, N d_\ell\}$.

Remark 3.3. Note that the search space in (3.3) is over the Cartesian product

$$\mathscr{R}\mathrm{BV}^2(\mathbb{R}^{d_0};\mathbb{R}^{d_1}) \times \cdots \times \mathscr{R}\mathrm{BV}^2(\mathbb{R}^{d_{L-1}};\mathbb{R}^{d_L}) \tag{3.6}$$

rather than $\mathscr{R}\mathrm{BV}^2_{\mathsf{deep}}(L)$. This is because, given a function $f \in \mathscr{R}\mathrm{BV}^2_{\mathsf{deep}}(L)$, there could be many decompositions such that $f = f^{(L)} \circ \cdots \circ f^{(1)}$. Therefore, in order for the regularization term in (3.3) to be well defined, we formulate the problem over (3.6). Note that since each term in the Cartesian product is a Banach space, the search space for this problem is itself a Banach space when equipped with the norm

$$\|(f^{(1)}, \ldots, f^{(L)})\|_{\mathscr{R}\mathrm{BV}^2_{\mathsf{deep}}(L)} := \sum_{\ell=1}^{L} \|f^{(\ell)}\|_{\mathscr{R}\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})}, \tag{3.7}$$

where we identify $\mathscr{R}\mathrm{BV}^2_{\mathsf{deep}}(L)$ with the Cartesian product (3.6). This type of Banach space has recently been referred to as a *neural RKBS* [13].

The neural network architecture that appears in (3.5) can be seen in Figure 2. Moreover, this exact architecture was studied empirically in [26]. It has also been studied in [31] from the statistical perspective under the name "accordion networks".

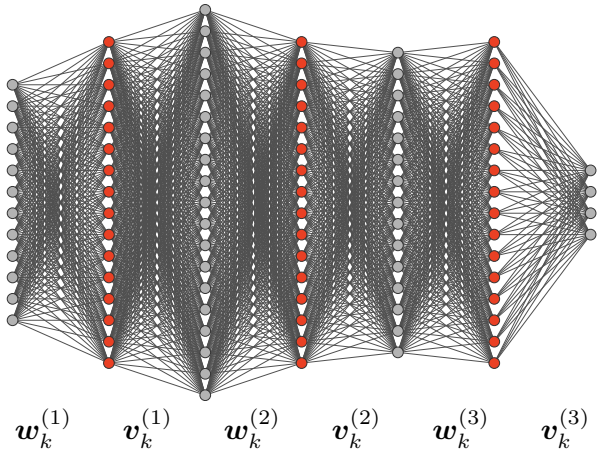

Fig. 2 *This figure shows the architecture of the deep neural network in* (3.5) *in the case of* $L = 3$ *hidden layers. The orange nodes denote ReLU nodes and the gray nodes denote linear nodes. Skip connection nodes are omitted for clarity.*

Remark 3.4. The regularizer that appears in (3.3) can be replaced by

$$\psi_0 \left(\sum_{\ell=1}^{L} \psi_\ell(\mathscr{R}\,\mathrm{TV}^2(f^{(\ell)})) \right), \tag{3.8}$$

where $\psi_\ell : [0, \infty) \to \mathbb{R}$, $\ell = 0, \ldots, L$, is a strictly increasing and convex function, and it can still have solutions that take the form of a deep neural network as in (3.4). Thus, there are many choices of regularization that result in a representer theorem for deep ReLU networks.

Remark 3.5. Notice that (3.4) is precisely the standard L-hidden layer deep ReLU network architecture with *rank-controlled weight matrices* and *skip connections.* Indeed, the weight matrix of the ℓth layer is $\mathbf{A}^{(\ell)} := \mathbf{W}^{(\ell+1)}\mathbf{V}^{(\ell)}$. More specifically, by dropping biases and skip connections for clarity, we see that $f_{\text{deep}}(\boldsymbol{x})$ in (3.4) can be computed recursively as

$$\begin{cases} \widetilde{\boldsymbol{x}}^{(0)} := \boldsymbol{x}, \\ \widetilde{\boldsymbol{x}}^{(\ell)} := \boldsymbol{\rho}(\mathbf{A}^{(\ell-1)}\widetilde{\boldsymbol{x}}^{(\ell-1)}), \quad \ell = 1, \ldots, L, \\ s(\boldsymbol{x}) := \mathbf{A}^{(L)}\widetilde{\boldsymbol{x}}^{(L)}, \end{cases} \tag{3.9}$$

where

$$\begin{cases} \mathbf{A}^{(0)} := \mathbf{W}^{(1)}, \\ \mathbf{A}^{(\ell)} := \mathbf{W}^{(\ell+1)}\mathbf{V}^{(\ell)}, \quad \ell = 1, \ldots, L-1, \\ \mathbf{A}^{(L)} := \mathbf{V}^{(L)}. \end{cases} \tag{3.10}$$

From the dimensions of $\mathbf{V}^{(\ell)}$ and $\mathbf{W}^{(\ell)}$ in Theorem 3.2, we see that for $\ell = 0, \ldots, L$, $\operatorname{rank}(\mathbf{A}^{(\ell)}) \leq \min\{N^2, Nd_{\ell-1}, d_\ell\}$ and $\operatorname{rank}(\mathbf{A}^{(L)}) \leq d_L$.

Proof of Theorem 3.2. Let $\widetilde{f} = \widetilde{f}^{(L)} \circ \cdots \circ \widetilde{f}^{(1)}$ be a (not necessarily unique) solution to (3.3). By applying $\widetilde{f}$ to each data point $\boldsymbol{x}_i$, $i = 1, \ldots, N$, we can recursively compute the intermediate vectors $\boldsymbol{z}_{i,\ell} \in \mathbb{R}^{d_\ell}$ as follows:

- Initialize $\boldsymbol{z}_{i,0} := \boldsymbol{x}_i$.
- For each $\ell = 1, \dots, L$, recursively update $\boldsymbol{z}_{i,\ell} := \widetilde{f}^{(\ell)}(\boldsymbol{z}_{i,\ell-1})$.

The solution $\widetilde{f}$ must satisfy

$$\widetilde{f}^{(\ell)} \in \operatorname*{arg\,min}_{f \in \mathscr{R}\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})} \mathscr{R}\mathrm{TV}^2(f) \quad \text{s.t.} \quad f(\boldsymbol{z}_{i,\ell-1}) = \boldsymbol{z}_{i,\ell}, \quad i = 1, \dots, N, \tag{3.11}$$

for $\ell = 1, \dots, L$. To see this, note that if (3.11) did not hold, it would contradict the optimality of $\widetilde{f}$. By Theorem 2.7, there always exists a solution to (3.11) that enforces the form of the solution in (3.4). ∎

Remark 3.6. Existence of minimizers is assumed *a priori* in Theorem 3.2. It turns out that, under mild conditions (e.g., lower semicontinuity of the loss function), this is always guaranteed. For the sake of pedagogy, we do not delve into those details in the present paper, and instead refer the reader to the original version of the article [52, Theorem 3.2].

4. Applications to Deep Network Training and Regularization. In this section we discuss applications of the representer theorem Theorem 3.2 to the training and regularization of deep ReLU networks. Since Theorem 3.2 ensures that solutions to the function-space problem in (3.3) are realizable by deep ReLU networks (3.4), one can find a solution to (3.3) by finding a solution to a finite-dimensional deep network training problem. A direct corollary of (2.26) is the following lemma.

LEMMA 4.1. *Given a deep neural network $f = f^{(L)} \circ \cdots \circ f^{(1)}$ as in* (3.4)*, where each $f^{(\ell)}$ is written in reduced form, it holds that*

$$\sum_{\ell=1}^{L} \mathscr{R}\mathrm{TV}^2(f^{(\ell)}) = \sum_{\ell=1}^{L} \sum_{k=1}^{K^{(\ell)}} \|\boldsymbol{v}_k^{(\ell)}\|_2 \|\boldsymbol{w}_k^{(\ell)}\|_2, \tag{4.1}$$

where $\boldsymbol{v}_k^{(\ell)}$ is the kth column of $\mathbf{V}^{(\ell)}$ and $\boldsymbol{w}_k^{(\ell)}$ is the kth row of $\mathbf{W}^{(\ell)}$.

Lemma 4.1 implies the following corollary to Theorem 3.2.

COROLLARY 4.2. *Let $\boldsymbol{\theta}$ denote the parameters of a deep neural network as in* (3.4)*, and let Θ denote the space of all such parameters. Write $f_{\boldsymbol{\theta}}$ to denote a deep neural network parameterized by $\boldsymbol{\theta}$. Then, the solutions to the finite-dimensional neural network training problem*

$$\min_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{y}_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) + \lambda \sum_{\ell=1}^{L} \sum_{k=1}^{K^{(\ell)}} \|\boldsymbol{v}_k^{(\ell)}\|_2 \|\boldsymbol{w}_k^{(\ell)}\|_2 \tag{4.2}$$

are solutions to (3.3) *as long as $K^{(\ell)} \geq \min\{N^2, N d_\ell\}$.*

The weight-decay regularizer is actually equivalent to the regularizer in Corollary 4.2. This yields the following corollary about weight-decay regularization.

COROLLARY 4.3. *The solutions to*

$$\min_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{y}_i, f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)) + \lambda \sum_{\ell=1}^{L} \frac{\|\mathbf{V}^{(\ell)}\|_{\mathsf{F}}^2 + \|\mathbf{W}^{(\ell)}\|_{\mathsf{F}}^2}{2} \tag{4.3}$$

are also solutions to (4.2)*, where $\|\cdot\|_{\mathsf{F}}$ denotes the Frobenius norm of a matrix. Moreover, the solutions to* (4.3) *satisfy the property that $\|\boldsymbol{v}_k^{(\ell)}\|_2 = \|\boldsymbol{w}_k^{(\ell)}\|_2$, $\ell = 1, \dots, L$, $k = 1, \dots, K^{(\ell)}$.*

Proof. The kth neuron in the ℓth layer of a deep neural network as in (3.4) takes the form $\boldsymbol{x} \mapsto \boldsymbol{v}_k^{(\ell)}({\boldsymbol{w}_k^{(\ell)}}^{\mathsf{T}}\boldsymbol{x} - b_k^{(\ell)})_+$. Due to the positive homogeneity of the ReLU, $\boldsymbol{v}_k^{(\ell)}$ and $\boldsymbol{w}_k^{(\ell)}$ can be rescaled so that $\|\boldsymbol{v}_k^{(\ell)}\|_2 = \|\boldsymbol{w}_k^{(\ell)}\|_2$ without altering the function of the network. Therefore, minimization of $\|\boldsymbol{v}_k^{(\ell)}\|_2^2 + \|\boldsymbol{w}_k^{(\ell)}\|_2^2$ is achieved when $\|\boldsymbol{v}_k^{(\ell)}\|_2 = \|\boldsymbol{w}_k^{(\ell)}\|_2$. The result then follows from the fact that when $\|\boldsymbol{v}_k^{(\ell)}\|_2 = \|\boldsymbol{w}_k^{(\ell)}\|_2$ we have that

$$\frac{\|\boldsymbol{v}_k^{(\ell)}\|_2^2 + \|\boldsymbol{w}_k^{(\ell)}\|_2^2}{2} = \|\boldsymbol{v}_k^{(\ell)}\|_2 \|\boldsymbol{w}_k^{(\ell)}\|_2. \tag{4.4}$$

∎

Remark 4.4. Due to the sparsity-promoting and neuron sharing nature of the $\mathscr{R}\,\mathrm{TV}^2$, the regularizers that appear in (4.2) and (4.3) promote sparse (in the sense of the number of active neurons) deep ReLU network solutions (cf. Figure 1).

Remark 4.5. Corollaries 4.2 and 4.3 reveal that training sufficiently wide deep ReLU networks with linear bottlenecks favors functions that are regular in the sense of the compositional space $\mathscr{R}\,\mathrm{BV}^2_{\mathsf{deep}}(L)$.

4.1. Connections to Existing Deep Network Regularization Schemes. The regularizers that appear in (4.2) and (4.3) are *principled* regularizers for training deep ReLU networks. In this section we will show how these regularizers are related to other regularizers that are studied in deep learning.

A common regularization scheme for deep ReLU networks is the path-norm regularizer. In particular, several works [10, 45, 46, 47] consider deep ReLU networks with no biases or skip connections mapping $\mathbb{R}^d \to \mathbb{R}$ of the form $s(\boldsymbol{x}) = x^{(L)}$, where $x^{(L)}$ is computed via

$$\begin{cases} \boldsymbol{x}^{(0)} := \boldsymbol{x}, \\ \boldsymbol{x}^{(\ell)} := \boldsymbol{\rho}(\mathbf{A}^{(\ell-1)}\boldsymbol{x}^{(\ell-1)}), \quad \ell = 1, \dots, L, \\ x^{(L)} := {\boldsymbol{a}^{(L)}}^{\mathsf{T}}\boldsymbol{x}^{(L)}, \end{cases} \tag{4.5}$$

where $\boldsymbol{\rho}$ denotes applying the ReLU componentwise, $\mathbf{A}^{(0)} \in \mathbb{R}^{K^{(1)} \times d}$, $\mathbf{A}^{(\ell)} \in \mathbb{R}^{K^{(\ell+1)} \times K^{(\ell)}}$, $\ell = 1, \dots, L-1$, and $\boldsymbol{a}^{(L)} \in \mathbb{R}^{K^{(L)}}$. Note that (4.5) is almost the same as the architecture in our framework if we drop biases and skip connections (see (3.9) in Remark 3.5). These works then consider path-norm regularization of the form

$$\sum_{k_L=1}^{K^{(L)}} \sum_{k_{L-1}=1}^{K^{(L-1)}} \cdots \sum_{k_1=1}^{K^{(1)}} \sum_{k_0=1}^{d} |*|a_{k_0,k_1}|*|a_{k_1,k_2} \cdots |*|a_{k_{L-1},k_L}|*|a_{k_L}, \tag{4.6}$$

where $a_{k_\ell,k_{\ell+1}}$ denotes the $(k_\ell, k_{\ell+1})$th entry in $\mathbf{A}^{(\ell)}$ and a_{k_L} denotes the k_Lth entry in $\boldsymbol{a}^{(L)}$.

Consider regularizing a deep ReLU network (with no biases or skip connections) from our framework with the following regularizer, which arises with a particular choice of $\{\psi_\ell\}_{\ell=0}^{L}$ in Remark 3.4:

$$\prod_{\ell=1}^{L} \sum_{k=1}^{K^{(\ell)}} \|\boldsymbol{v}_k^{(\ell)}\|_2 \|\boldsymbol{w}_k^{(\ell)}\|_2. \tag{4.7}$$

Observe that (4.7) is an upper bound on something that looks very similar to the path-norm in (4.6). Indeed, first notice that if we write the deep ReLU network from our framework in the form in (3.9), we have

$$|a_{k_\ell,k_{\ell+1}}| = |{\boldsymbol{v}_k^{(\ell)}}^{\mathsf{T}} \boldsymbol{w}_k^{(\ell+1)}| \leq \|\boldsymbol{v}_k^{(\ell)}\|_2 \|\boldsymbol{w}_k^{(\ell+1)}\|_2, \tag{4.8}$$

where $a_{k_\ell,k_{\ell+1}}$ denotes the $(k_\ell, k_{\ell+1})$th entry in $\mathbf{A}^{(\ell)}$ as defined in Remark 3.5. Therefore,

$$\begin{aligned}\prod_{\ell=1}^{L} \sum_{k=1}^{K^{(\ell)}} \|\boldsymbol{v}_k^{(\ell)}\|_2 \|\boldsymbol{w}_k^{(\ell)}\|_2 &= \sum_{k_L=1}^{K^{(L)}} \cdots \sum_{k_1=1}^{K^{(1)}} \|\boldsymbol{w}_{k_1}^{(1)}\|_2 \|\boldsymbol{v}_{k_1}^{(1)}\|_2 \|\boldsymbol{w}_{k_2}^{(2)}\|_2 \|\boldsymbol{v}_{k_2}^{(2)}\|_2 \cdots \|\boldsymbol{w}_{k_L}^{(L)}\|_2 \|\boldsymbol{v}_{k_L}^{(L)}\|_2 \\ &\geq \sum_{k_L=1}^{K^{(L)}} \cdots \sum_{k_1=1}^{K^{(1)}} \|\boldsymbol{w}_{k_1}^{(1)}\|_2 \, |a_{k_1,k_2}| \cdots |a_{k_{L-1},k_L}| \, \|\boldsymbol{v}_{k_L}^{(L)}\|_2, \end{aligned} \tag{4.9}$$

where the last line holds from (4.8). We see that the last line in the above equation is the same as the path-norm in (4.6), apart from how it treats weights in the first and last layers. We also remark that the work in [10] shows that the path-norm in (4.6) controls the Rademacher and Gaussian complexity of deep ReLU networks.

5. Conclusion. In this paper we have proven a representer theorem for deep ReLU networks. We have shown that deep ReLU networks with L-hidden layers, skip connections, and rank-bounded weight matrices are solutions to a variational problem over compositional $\mathscr{R}\mathrm{BV}^2$-spaces. This function-space problem can be recast as a finite-dimensional neural network training problem with various choices of regularization on the network parameters. We have therefore derived several new, principled regularizers for deep ReLU networks. Moreover, these regularizers promote sparse solutions. We have shown that these new regularizers are related to the well-known weight decay and path-norm regularization schemes commonly used in the training of deep ReLU networks. The main follow-up question revolves around sharper understanding of the compositional space $\mathscr{R}\mathrm{BV}^2_{\mathsf{deep}}(L)$. This first requires sharper understanding of the $\mathscr{R}\mathrm{BV}^2$-spaces. The function spaces studied in this paper are not classical, and understanding their analytic properties will play a key role in understanding the kinds of functions that deep neural networks learn from data.

Appendix A. Topological Properties of $\mathscr{R}\mathbf{BV}^{\mathbf{2}}(\mathbb{R}^{\boldsymbol{d}})$. In this section we will prove Lemma 2.4. We rely on many results developed in [51]. While the definition of $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ given in (2.5) is convenient from an intuitive perspective, it does not lend itself to analysis due to $\mathscr{R}\mathrm{TV}^2(\cdot)$ being a seminorm with null space $\mathcal{P}_1(\mathbb{R}^d)$, the space of polynomials of degree at most one, i.e., affine functions on $\mathbb{R}^d$. Thus, we use the result of [51, Theorem 22] to characterize $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ as a Banach space. In particular, [51, Theorem 22] considers an arbitrary *biorthogonal system* of $\mathcal{P}_1(\mathbb{R}^d)$ in order to equip $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ with a bona fide norm.

Definition A.1. *Let $\mathcal{N}$ be a finite-dimensional space with $N_0 := \dim \mathcal{N}$. The pair $(\boldsymbol{\phi}, \boldsymbol{p}) = \{(\phi_n, p_n)\}_{n=0}^{N_0-1}$ is called a* biorthogonal system *for $\mathcal{N}$ if $\boldsymbol{p} = \{p_n\}_{n=0}^{N_0-1}$ is a basis of $\mathcal{N}$ and the "boundary" functionals $\boldsymbol{\phi} = \{\phi_n\}_{n=0}^{N_0-1}$ with $\phi_n \in \mathcal{N}'$ (the continuous dual of $\mathcal{N}$) satisfy the biorthogonality condition $\langle \phi_k, p_n \rangle = \delta[k-n]$, $k, n = 0, \ldots, N_0-1$, where $\delta[\cdot]$ is the Kronecker impulse.*

PROPOSITION A.2 (see [51, Theorem 22, item 3]). *Let $(\boldsymbol{\phi}, \boldsymbol{p})$ be a biorthogonal system for $\mathcal{P}_1(\mathbb{R}^d)$. Then $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ equipped with the norm*

$$\|f\|_{\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)} = \mathscr{R}\,\mathrm{TV}^2(f) + \|\boldsymbol{\phi}(f)\|_1, \tag{A.1}$$

where $\boldsymbol{\phi}(f) = (\langle \phi_0, f\rangle, \ldots, \langle \phi_d, f\rangle) \in \mathbb{R}^{d+1}$, is a Banach space.

Proof of Lemma 2.4*, item* 1. By Proposition A.2 it suffices to find a biorthogonal system $(\boldsymbol{\phi}, \boldsymbol{p})$ of $\mathcal{P}_1(\mathbb{R}^d)$ such that for every $f \in \mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ we have

$$\|\boldsymbol{\phi}(f)\|_1 = |f(\mathbf{0})| + \sum_{k=1}^{d} |f(\boldsymbol{e}_k) - f(\mathbf{0})|. \tag{A.2}$$

Put $p_0(\boldsymbol{x}) := 1$ and $p_k(\boldsymbol{x}) := x_k$, $k = 1, \ldots, d$. Clearly, $\boldsymbol{p}$ is a basis for $\mathcal{P}_1(\mathbb{R}^d)$. Put $\phi_0 := \delta$ and $\phi_k := \delta(\cdot - \boldsymbol{e}_k) - \delta$, $k = 1, \ldots, d$, where δ denotes the Dirac impulse on $\mathbb{R}^d$ and $\boldsymbol{e}_k$ denotes the kth canonical basis vector of $\mathbb{R}^d$. Then, $(\boldsymbol{\phi}, \boldsymbol{p})$ is a biorthogonal system for $\mathcal{P}_1(\mathbb{R}^d)$. Indeed, we have $\langle \phi_0, p_0\rangle = 1$ and $\langle \phi_k, p_k\rangle = p_k(\boldsymbol{e}_k) - p_k(\mathbf{0}) = 1 - 0 = 1$, $k = 1, \ldots, d$. We also have

$$\langle \phi_0, p_k\rangle = p_k(\mathbf{0}) = 0, \quad k = 1, \ldots, d, \tag{A.3}$$

$$\langle \phi_k, p_0\rangle = p_0(\boldsymbol{e}_k) - p_0(\mathbf{0}) = 1 - 1 = 0, \quad k = 1, \ldots, d, \tag{A.4}$$

$$\langle \phi_k, p_n\rangle = p_n(\boldsymbol{e}_k) - p_n(\mathbf{0}) = 0 + 0 = 0, \quad k, n = 1, \ldots, d, \quad k \neq n. \tag{A.5}$$

A computation shows that (A.2) holds with this choice of biorthogonal system. ∎

In order to prove item 2 of Lemma 2.4, we must show that $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ has a predual (so that it can be endowed with a weak* topology) and consequently show that the point evaluation functional is weak* continuous. In other words, we must show that the Dirac impulse, $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, lies in the predual.

Before we can prove this, we require an important result from [51]. Recall from (2.6) that

$$\mathscr{R}\,\mathrm{TV}^2(f) = c_d \|\partial_t^2 \Lambda^{d-1} \mathscr{R} f\|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})}. \tag{A.6}$$

Put $\mathrm{R} := c_d\, \partial_t^2 \Lambda^{d-1} \mathscr{R}$. As discussed in [51], for every $f \in \mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$, $u := \mathrm{R}\, f \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})$ is always even, i.e., $u(\boldsymbol{w}, b) = u(-\boldsymbol{w}, -b)$. This means we have

$$\mathscr{R}\,\mathrm{TV}^2(f) = \|\mathrm{R}\, f\|_{\mathcal{M}(\mathbb{P}^d)}, \tag{A.7}$$

where $\mathbb{P}^d$ denotes the manifold of hyperplanes on $\mathbb{R}^d$. In particular, we can view $\mathcal{M}(\mathbb{P}^d)$ as the subspace of $\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})$ with only even finite Radon measures. Indeed, this is due to the fact that every hyperplane takes the form $h_{(\boldsymbol{w}, b)} := \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{w}^\mathsf{T} \boldsymbol{x} = b\}$ for some $(\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$ and $h_{(\boldsymbol{w}, b)} = h_{(-\boldsymbol{w}, -b)}$.

PROPOSITION A.3 (see [51, Lemma 21 and Theorem 22]). *Let $(\boldsymbol{\phi}, \boldsymbol{p})$ be a biorthogonal system for $\mathcal{P}_1(\mathbb{R}^d)$. Then, every $f \in \mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ has the unique direct-sum decomposition*

$$f = \mathrm{R}_{\boldsymbol{\phi}}^{-1}\{u\} + q, \tag{A.8}$$

where $u = \mathrm{R}f \in \mathcal{M}(\mathbb{P}^d)$, $q = \sum_{k=0}^{d} \langle \phi_k, f \rangle p_k \in \mathcal{P}_1(\mathbb{R}^d)$, *and*

$$\mathrm{R}_{\boldsymbol{\phi}}^{-1} : u \mapsto \int_{\mathbb{S}^{d-1}\times\mathbb{R}} g_{\boldsymbol{\phi}}(\cdot, \boldsymbol{z}) \, \mathrm{d}u(\boldsymbol{z}). \tag{A.9}$$

Also,

$$g_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{z}) = r_{\boldsymbol{z}}(\boldsymbol{x}) - \sum_{k=0}^{d} p_k(\boldsymbol{x}) q_k(\boldsymbol{z}), \tag{A.10}$$

where $r_{\boldsymbol{z}} = r_{(\boldsymbol{w},b)} = \rho(\boldsymbol{w}^{\mathsf{T}}(\cdot) - b)$ *and* $q_k(\boldsymbol{z}) := \langle \phi_k, r_{\boldsymbol{z}} \rangle$, *where* $\boldsymbol{z} = (\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$ *and* $\rho = |\cdot|/2$.

The operator $\mathrm{R}_{\boldsymbol{\phi}}^{-1}$ defined in (A.9) has several useful properties (see [51, Theorem 22, items 1 and 2]). In particular, it is a bounded right-inverse of R and, when restricted to

$$\mathscr{R}\mathrm{BV}_{\boldsymbol{\phi}}^2(\mathbb{R}^d) := \{ f \in \mathscr{R}\mathrm{BV}^2(\mathbb{R}^d) : \ \boldsymbol{\phi}(f) = \mathbf{0} \}, \tag{A.11}$$

it is the *bona fide* inverse of R. The space $\mathscr{R}\mathrm{BV}_{\boldsymbol{\phi}}^2(\mathbb{R}^d)$ is also a concrete transcription of the abstract quotient $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)/\mathcal{P}_1(\mathbb{R}^d)$. We have that $\mathrm{R} : \mathscr{R}\mathrm{BV}_{\boldsymbol{\phi}}^2(\mathbb{R}^d) \to \mathcal{M}(\mathbb{P}^d)$ is an isometric isomorphism with the inverse given by $\mathrm{R}_{\boldsymbol{\phi}}^{-1}$. Additionally. we have from Proposition A.3 that $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d) \cong \mathscr{R}\mathrm{BV}_{\boldsymbol{\phi}}^2(\mathbb{R}^d) \oplus \mathcal{P}_1(\mathbb{R}^d)$, where $\mathscr{R}\mathrm{BV}_{\boldsymbol{\phi}}^2(\mathbb{R}^d)$ is a Banach space when equipped with the norm $f \mapsto \|\mathrm{R}f\|_{\mathcal{M}(\mathbb{P}^d)}$ and $\mathcal{P}_1(\mathbb{R}^d)$ is a Banach space when equipped with the norm $f \mapsto \|\boldsymbol{\phi}(f)\|_1$. These properties will be important in proving item 2 of Lemma 2.4.

Proof of Lemma 2.4*, item* 2. Let $(\boldsymbol{\phi}, \boldsymbol{p})$ be the biorthogonal system constructed in the proof of Lemma 2.4, item 1. Since $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d) \cong \mathscr{R}\mathrm{BV}_{\boldsymbol{\phi}}^2(\mathbb{R}^d) \oplus \mathcal{P}_1(\mathbb{R}^d)$, the following diagram immediately reveals that $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ has a predual.

$$\begin{array}{ccc} \mathscr{R}\mathrm{BV}_{\boldsymbol{\phi}}^2(\mathbb{R}^d) & \underset{\mathrm{R}_{\boldsymbol{\phi}}^{-1}}{\overset{\mathrm{R}}{\rightleftarrows}} & \mathcal{M}(\mathbb{P}^d) \\ \mathrm{dual} \uparrow & & \uparrow \mathrm{dual} \\ \mathcal{X} & \underset{\mathrm{R}_{\boldsymbol{\phi}}^{-1*}}{\overset{\mathrm{R}^*}{\rightleftarrows}} & C_0(\mathbb{P}^d) \end{array} \tag{A.12}$$

Thus, showing that $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, is weak* continuous on $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ is equivalent to showing that it is weak* continuous on both $\mathscr{R}\mathrm{BV}_{\boldsymbol{\phi}}^2(\mathbb{R}^d)$ and $\mathcal{P}_1(\mathbb{R}^d)$.

Clearly, $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, is continuous on $\mathcal{P}_1(\mathbb{R}^d)$ (since every element of $\mathcal{P}_1(\mathbb{R}^d)$ is a continuous function). Then, since $\mathcal{P}_1(\mathbb{R}^d)$ is finite-dimensional, the spaces of continuous linear functionals and weak* continuous linear functionals are the same. Thus, $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, is weak* continuous on $\mathcal{P}_1(\mathbb{R}^d)$. It remains to show that $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, is weak* continuous on $\mathscr{R}\mathrm{BV}_{\boldsymbol{\phi}}^2(\mathbb{R}^d)$. From (A.12), we have that $\mathcal{X}$ is a predual of $\mathscr{R}\mathrm{BV}_{\boldsymbol{\phi}}^2(\mathbb{R}^d)$, i.e., $\mathcal{X}' = \mathscr{R}\mathrm{BV}_{\boldsymbol{\phi}}^2(\mathbb{R}^d)$. We must show that $\delta(\cdot - \boldsymbol{x}_0) \in \mathcal{X}$, $\boldsymbol{x}_0 \in \mathbb{R}^d$.

Observe that $\delta(\cdot - \boldsymbol{x}_0) \in \mathcal{X}$ if and only if $\mathrm{R}_{\boldsymbol{\phi}}^{-1*}\{\delta(\cdot - \boldsymbol{x}_0)\} \in C_0(\mathbb{P}^d)$. From Proposition A.3 we see that $\mathrm{R}_{\boldsymbol{\phi}}^{-1*}\{\delta(\cdot - \boldsymbol{x}_0)\} = g_{\boldsymbol{\phi}}(\boldsymbol{x}_0, \cdot)$ as defined in (A.10). Since $\rho = |\cdot|/2$ in (A.10), we have that

$$g_{\boldsymbol{\phi}}(\boldsymbol{x}_0,(\boldsymbol{w},b)) = \frac{|\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_0 - b|}{2} - \sum_{k=0}^{d} p_k(\boldsymbol{x}_0)\left\langle \phi_k, \frac{|\boldsymbol{w}^\mathsf{T}(\cdot) - b|}{2}\right\rangle$$
$$\overset{(*)}{=} \frac{|\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_0 - b|}{2} - \left[\frac{|-b|}{2} + \sum_{k=1}^{d} x_{0,k}\left(\frac{|w_k - b|}{2} - \frac{|-b|}{2}\right)\right]$$
$$= \frac{|\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_0 - b|}{2} - \frac{|b|}{2}\left(1 - \sum_{k=1}^{d} x_{0,k}\right) - \sum_{k=1}^{d} x_{0,k}\frac{|w_k - b|}{2}, \tag{A.13}$$

where $(*)$ follows by substituting in the biorthogonal system $(\boldsymbol{\phi},\boldsymbol{p})$ constructed in the proof of Lemma 2.4, item 1. Clearly, $g_{\boldsymbol{\phi}}(\boldsymbol{x}_0,\cdot)$ is continuous and $g_{\boldsymbol{\phi}}(\boldsymbol{x}_0,(\boldsymbol{w},b)) = g_{\boldsymbol{\phi}}(\boldsymbol{x}_0,(-\boldsymbol{w},-b))$, so $g_{\boldsymbol{\phi}}(\boldsymbol{x}_0,\cdot)$ is an even function on $\mathbb{S}^{d-1}\times\mathbb{R}$ and therefore a continuous function on $\mathbb{P}^d$. It remains to check that $g_{\boldsymbol{\phi}}(\boldsymbol{x}_0,\cdot)$ is vanishing at infinity. Certainly, this is true. Indeed, for sufficiently large b we have

$$g_{\boldsymbol{\phi}}(\boldsymbol{x}_0,(\boldsymbol{w},b)) = \frac{-\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_0 + b}{2} - \frac{b}{2}\left(1 - \sum_{k=1}^{d} x_{0,k}\right) - \sum_{k=1}^{d} x_{0,k}\frac{-w_k + b}{2} = 0, \tag{A.14}$$

and for sufficiently small b we have

$$g_{\boldsymbol{\phi}}(\boldsymbol{x}_0,(\boldsymbol{w},b)) = \frac{\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_0 - b}{2} - \frac{-b}{2}\left(1 - \sum_{k=1}^{d} x_{0,k}\right) - \sum_{k=1}^{d} x_{0,k}\frac{w_k - b}{2} = 0. \tag{A.15}$$

Therefore, $g_{\boldsymbol{\phi}}(\boldsymbol{x}_0,\cdot)$ is compactly supported on $\mathbb{P}^d$, and so $g_{\boldsymbol{\phi}}(\boldsymbol{x}_0,\cdot)\in C_0(\mathbb{P}^d)$. Thus, the Dirac impulse $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0\in\mathbb{R}^d$, is weak* continuous on $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$. ∎

Proof of Lemma 2.4*, item* 3. This immediately follows from the direct-sum decomposition of $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ specified in the proof of item 2. Indeed, since $\mathscr{R}\mathrm{BV}^2_{\boldsymbol{\phi}}(\mathbb{R}^d)$ is isometrically isomorphic to $\mathcal{M}(\mathbb{P}^d)$, we deduce that the extreme points of the unit ball

$$\{f\in\mathscr{R}\mathrm{BV}^2_{\boldsymbol{\phi}}(\mathbb{R}^d):\ \|\mathrm{R}f\|_{\mathcal{M}}\le 1\} \tag{A.16}$$

take the form $\{\pm g_{\boldsymbol{\phi}}(\cdot,(\boldsymbol{w},b))\}_{(\boldsymbol{w},b)\in\mathbb{S}^{d-1}\times\mathbb{R}}$ (since the extreme points of the unit ball of $\mathcal{M}(\mathbb{P}^d)$ take the form $\pm(\delta_{\boldsymbol{z}} + \delta_{-\boldsymbol{z}})/2$, $\boldsymbol{z}\in\mathbb{S}^{d-1}\times\mathbb{R}$). Modulo affine functions, $g_{\boldsymbol{\phi}}(\cdot,(\boldsymbol{w},b))$ is a ReLU neuron and so the result follows. ∎

Appendix B. Proof of Lemma 2.5.

Proof. Observe that the vectorized operator $\underline{\partial_t^2\mathrm{K}\mathscr{R}}$ maps $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$ to $\mathcal{M}(\mathbb{P}^d;\mathbb{R}^D)$ and its null space is the space of D-variate affine functions, $\mathcal{P}_1(\mathbb{R}^d;\mathbb{R}^D)$. Thus, any choice of norm on the Cartesian product of $\mathcal{M}(\mathbb{P}^d;\mathbb{R}^D)\times\mathcal{P}_1(\mathbb{R}^d;\mathbb{R}^D)$ will induce an isometric isomorphism to $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$ (cf. Appendix A).

Thus, when equipped with the norm

$$\|f\|_{\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)} := \mathscr{R}\mathrm{TV}^2(f) + \sum_{j=1}^{D}\left(|f_j(\boldsymbol{0})| + \sum_{k=1}^{d}|f_j(\boldsymbol{e}_k) - f_j(\boldsymbol{0})|\right), \tag{B.1}$$

$\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$ is isometrically isomorphic to $\mathcal{M}(\mathbb{P}^d;\mathbb{R}^D)\times\mathcal{P}_1(\mathbb{R}^d;\mathbb{R}^D)$ and hence Banach, where we equip $\mathcal{M}(\mathbb{P}^d;\mathbb{R}^D)$ with the $(2,\mathcal{M})$-norm.

Since the latter is clearly a dual Banach space, $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$ is also a dual Banach space and the weak* continuity of the componentwise point evaluation functional boils down to $g_{\boldsymbol{\phi}}(\boldsymbol{x}_0,\cdot)\boldsymbol{e}_j\in C_0(\mathbb{P}^d;\mathbb{R}^D)$, $j\in[D]$ (with $g_{\boldsymbol{\phi}}$ from (A.13)). This holds

since $g_{\boldsymbol{\phi}}(\boldsymbol{x}_0, \cdot) \in C_0(\mathbb{P}^d)$ from the proof of Lemma 2.4, item 2. To complete the proof, we note that the extreme points of the unit ball of $\mathcal{M}(\mathbb{P}^d; \mathbb{R}^D)$ take the form $\boldsymbol{u}(\delta_{\boldsymbol{z}} + \delta_{-\boldsymbol{z}})/2$, where $\boldsymbol{u} \in \mathbb{S}^{D-1}$ and $\boldsymbol{z} \in \mathbb{S}^{d-1} \times \mathbb{R}$. Therefore, modulo affine functions, the extreme points of the unit $\mathscr{R}\,\mathrm{TV}^2$ ball are of the form $\boldsymbol{u} g_{\boldsymbol{\phi}}(\cdot, (\boldsymbol{w}, b))$, which is equal modulo an affine function to $\boldsymbol{u}(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)_+$. ∎

Appendix C. Proof of Lemma 2.6. By construction, every $f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d; \mathbb{R}^D)$ admits the decomposition

$$f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} g_{\boldsymbol{\phi}}(\boldsymbol{x}, (\boldsymbol{w}, b))\, \mathrm{d}\boldsymbol{\nu}(\boldsymbol{w}, b) + \mathbf{C}\boldsymbol{x} + \boldsymbol{c}_0, \tag{C.1}$$

with $g_{\boldsymbol{\phi}}$ as in (A.13) and $\boldsymbol{\nu} \in \mathcal{M}(\mathbb{P}^d; \mathbb{R}^D)$. In particular, it holds that

$$\|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)} = \|\boldsymbol{\nu}\|_{2,\mathcal{M}} + \|\mathbf{C}\|_{1,1} + \|\boldsymbol{c}_0\|_1, \tag{C.2}$$

where the $(1,1)$-norm of $\mathbf{C}$ denotes the 1-norm of the vectorization of $\mathbf{C}$.

Proof of Lemma 2.6. We will first bound the Lipschitz constant of $g_{\boldsymbol{\phi}}(\cdot, \boldsymbol{z})$ defined in (A.13), where $\boldsymbol{z} = (\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\begin{aligned}
|g_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{z}) - g_{\boldsymbol{\phi}}(\boldsymbol{y}, \boldsymbol{z})| &= \left| \frac{|\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b|}{2} - \frac{|\boldsymbol{w}^\mathsf{T}\boldsymbol{y} - b|}{2} \right. \\
&\quad \left. - \frac{|b|}{2}\left[\left(1 - \sum_{k=1}^d x_k\right) - \left(1 - \sum_{k=1}^d y_k\right)\right] - \sum_{k=1}^d (x_k - y_k)\frac{|w_k - b|}{2} \right| \\
&\le \frac{\big|\, |\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b| - |\boldsymbol{w}^\mathsf{T}\boldsymbol{y} - b| \,\big|}{2} \\
&\quad + \left| \sum_{k=1}^d (x_k - y_k)\frac{|b|}{2} - \sum_{k=1}^d (x_k - y_k)\frac{|w_k - b|}{2} \right| \\
&\le \frac{\big|\, |\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b| - |\boldsymbol{w}^\mathsf{T}\boldsymbol{y} - b| \,\big|}{2} + \sum_{k=1}^d |x_k - y_k| \frac{\big|\, |b| - |w_k - b| \,\big|}{2} \\
&\overset{(*)}{\le} \frac{|\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - \boldsymbol{w}^\mathsf{T}\boldsymbol{y}|}{2} + \sum_{k=1}^d |x_k - y_k| \frac{|w_k|}{2} \\
&\overset{(\S)}{\le} \frac{\|\boldsymbol{w}\|_2 \|\boldsymbol{x} - \boldsymbol{y}\|_2 + \|\boldsymbol{w}\|_2 \|\boldsymbol{x} - \boldsymbol{y}\|_2}{2} \\
&\overset{(\dagger)}{=} \|\boldsymbol{x} - \boldsymbol{y}\|_2,
\end{aligned} \tag{C.3}$$

where $(*)$ holds from the reverse triangle inequality, $(\S)$ holds from the Cauchy–Schwarz inequality, and $(\dagger)$ holds since $\|\boldsymbol{w}\|_2 = 1$.

Next, from (C.1) we have, for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\begin{aligned}
\|f(\boldsymbol{x}) - f(\boldsymbol{y})\|_2 &\le \int_{\mathbb{S}^{d-1}\times\mathbb{R}} |g(\boldsymbol{x}, (\boldsymbol{w}, b)) - g(\boldsymbol{y}, (\boldsymbol{w}, b))|\, \mathrm{d}|\boldsymbol{\nu}|_{2,\mathcal{M}}(\boldsymbol{w}, b) + \|\mathbf{C}(\boldsymbol{x} - \boldsymbol{y})\|_2 \\
&\le \|\boldsymbol{x} - \boldsymbol{y}\|_2 \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \mathrm{d}|\boldsymbol{\nu}|_{2,\mathcal{M}}(\boldsymbol{w}, b) + \|\mathbf{C}\|_\mathsf{F} \|\boldsymbol{x} - \boldsymbol{y}\|_2 \\
&\le \|\boldsymbol{\nu}\|_{2,\mathcal{M}} \|\boldsymbol{x} - \boldsymbol{y}\|_2 + \|\mathbf{C}\|_{1,1} \|\boldsymbol{x} - \boldsymbol{y}\|_2 \\
&\le \|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)} \|\boldsymbol{x} - \boldsymbol{y}\|_2,
\end{aligned} \tag{C.4}$$

where $|\boldsymbol{\nu}|_{2,\mathcal{M}} \in \mathcal{M}(\mathbb{P}^d)$ denotes the *TV measure* (which is a scalar-valued measure). ∎

Appendix D. Proof of Theorem 2.7.

Proof. By item 2 in Lemma 2.5, we can endow $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$ with a weak* topology. The remainder of the proof relies on recent characterizations of abstract representer theorems (see [16, 17, 68, 70]). From the weak*-continuity of the componentwise point evaluation functional on $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$ (item 2 in Lemma 2.5), our setting coincides with the hypotheses of [70, Theorem 3] with ND scalar measurements. First, this abstract result ensures that the solution set is nonempty, convex, and weak*-compact. Second, it ensures that it is the weak*-closure of the convex hull of its extreme points, which can all be expressed as

$$f_{\text{extreme}} = \sum_{k=1}^{K} v_k e_k + q, \tag{D.1}$$

where the number K of atoms satisfies $K \leq ND$, q is in the null space of the regularizer (i.e., it is an affine function), and, for $k = 1, \ldots, K$, $v_k \in \mathbb{R} \setminus \{0\}$ and e_k is an extreme point of the unit regularization ball. The characterization of extreme points in item 3 in Lemma 2.5 characterizes the form of the solution in (2.24). To complete the proof, we note that since (2.26) holds, we can recast the problem as a multitask lasso problem and invoke the sparsity bound from [62, Theorem 10] (see also [62, Step (iv), Proof of Theorem 5, Appendix C]). ∎

REFERENCES

[1] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404, https://doi.org/10.1090/S0002-9947-1950-0051437-7. (Cited on p. 128)

[2] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, *Understanding deep neural networks with rectified linear units*, in International Conference on Learning Representations, 2018, pp. 1–17. (Cited on p. 129)

[3] S. Aziznejad, H. Gupta, J. Campos, and M. Unser, *Deep neural networks with trainable activations and controlled Lipschitz constant*, IEEE Trans. Signal Process., 68 (2020), pp. 4688–4699, https://doi.org/10.1109/TSP.2020.3014611. (Cited on p. 129)

[4] L. J. Ba and R. Caruana, *Do deep nets really need to be deep?*, in Proceedings of the 27th International Conference on Neural Information Processing Systems, Vol. 2, ACM, 2014, pp. 2654–2662. (Cited on p. 130)

[5] F. Bach, *Breaking the curse of dimensionality with convex neural networks*, J. Mach. Learn. Res., 18 (2017), pp. 629–681, https://jmlr.org/papers/v18/14-546.html. (Cited on p. 130)

[6] R. Balestriero and R. Baraniuk, *A spline theory of deep learning*, in Proceedings of the 35th International Conference on Machine Learning, Proc. Mach. Learn. Res. 80, PMLR, 2018, pp. 374–383. (Cited on p. 129)

[7] R. Balestriero and R. G. Baraniuk, *Mad Max: Affine spline insights into deep learning*, Proc. IEEE, 109 (2021), pp. 704–727, https://doi.org/10.1109/JPROC.2020.3042100. (Cited on p. 129)

[8] A. R. Barron, *Universal approximation bounds for superpositions of a sigmoidal function*, IEEE Trans. Inform. theory, 39 (1993), pp. 930–945, https://doi.org/10.1109/18.256500. (Cited on p. 130)

[9] A. R. Barron, *Approximation and estimation bounds for artificial neural networks*, Mach. Learn., 14 (1994), pp. 115–133, https://doi.org/10.1023/A:1022650905902. (Cited on p. 130)

[10] A. R. Barron and J. M. Klusowski, *Complexity, Statistical Risk, and Metric Entropy of Deep Nets Using Total Path Variation*, preprint, arXiv:1902.00800, 2019. (Cited on pp. 140, 141)

[11] F. Bartolucci, M. Carioni, J. A. Iglesias, Y. Korolev, E. Naldi, and S. Vigogna, *A Lipschitz Spaces View of Infinitely Wide Shallow Neural Networks*, preprint, arXiv:2410.14591, 2024. (Cited on p. 130)

[12] 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. (Cited on p. 130)

[13] F. Bartolucci, E. De Vito, L. Rosasco, and S. Vigogna, *Neural Reproducing Kernel Banach Spaces and Representer Theorems for Deep Networks*, preprint, arXiv:2403.08750, 2024. (Cited on pp. 130, 137)

[14] B. Bohn, C. Rieger, and M. Griebel, *A representer theorem for deep kernel learning*, J. Mach. Learn. Res., 20 (2019), pp. 1–32, https://jmlr.org/papers/v20/17-621.html. (Cited on p. 130)

[15] P. Bohra, J. Campos, H. Gupta, S. Aziznejad, and M. Unser, *Learning activation functions in deep (spline) neural networks*, IEEE Open J. Signal Process., 1 (2020), pp. 295–309, https://doi.org/10.1109/OJSP.2020.3039379. (Cited on p. 129)

[16] 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. (Cited on pp. 133, 146)

[17] K. Bredies and M. Carioni, *Sparsity of solutions for variational inverse problems with finite-dimensional data*, Calc. Var. Partial Differential Equations, 59 (2020), art. 14, https://doi.org/10.1007/s00526-019-1658-1. (Cited on pp. 133, 146)

[18] Z. Chen, *Neural Hilbert ladders: Multi-layer neural networks in function space*, J. Mach. Learn. Res., 25 (2024), pp. 1–65, https://jmlr.org/papers/v25/23-1225.html. (Cited on p. 130)

[19] 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), art. 101713, https://doi.org/10.1016/j.acha.2024.101713. (Cited on p. 130)

[20] J. Diestel and J. Uhl, *Vector Measures*, Math. Surv. Monogr., AMS, 1977. (Cited on p. 134)

[21] W. E and S. Wojtowytsch, *On the Banach spaces associated with multi-layer ReLU networks: Function representation, approximation theory and gradient descent dynamics*, CSIAM Trans. Appl. Math., 1 (2020), pp. 387–440, https://doi.org/10.4208/csiam-am.20-211. (Cited on p. 130)

[22] W. E and S. Wojtowytsch, *Representation formulas and pointwise properties for Barron functions*, Calc. Var. Partial Differential Equations, 61 (2022), art. 46, https://doi.org/10.1007/s00526-021-02156-6. (Cited on pp. 128, 130)

[23] 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. (Cited on pp. 128, 130)

[24] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed., John Wiley & Sons, New York, 1999. (Cited on p. 131)

[25] J. Frankle and M. Carbin, *The lottery ticket hypothesis: Finding sparse, trainable neural networks*, in International Conference on Learning Representations, 2019, pp. 1–42. (Cited on p. 129)

[26] A. Golubeva, B. Neyshabur, and G. Gur-Ari, *Are wider nets better given the same number of parameters?*, in International Conference on Learning Representations, 2021, pp. 1–19. (Cited on pp. 130, 137)

[27] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2016, pp. 770–778. (Cited on pp. 129, 132)

[28] T. J. Heeringa, L. Spek, and C. Brune, *Deep Networks Are Reproducing Kernel Chains*, preprint, arXiv:2501.03697, 2025. (Cited on p. 130)

[29] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, *Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups*, IEEE Signal Process. Mag., 29 (2012), pp. 82–97, https://doi.org/10.1109/MSP.2012.2205597. (Cited on p. 128)

[30] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, *Improving Neural Networks by Preventing Co-adaptation of Feature Detectors*, preprint, arXiv:1207.0580, 2012. (Cited on p. 129)

[31] A. Jacot, S. H. Choi, and Y. Wen, *How DNNs break the curse of dimensionality: Compositionality and symmetry learning*, in International Conference on Learning Representations, 2025, pp. 1–35. (Cited on p. 137)

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *ImageNet classification with deep convolutional neural networks*, in Advances in Neural Information Processing Systems 25, Curran Associates, 2012, pp. 1097–1105. (Cited on p. 128)

[33] A. Krogh and J. A. Hertz, *A simple weight decay can improve generalization*, in Advances in Neural Information Processing Systems, Morgan Kaufmann, 1992, pp. 950–957. (Cited on p. 129)

[34] A. Kumar, R. Parhi, and M. Belkin, *A gap between the Gaussian RKHS and neural networks: An infinite-center asymptotic analysis*, in Conference on Learning Theory (COLT),

Vol. 291, 2025, pp. 3463–3485, https://proceedings.mlr.press/v291/kumar25b.html. (Cited on p. 130)

[35] 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. (Cited on p. 130)

[36] V. Kůrková and M. Sanguineti, *Comparison of worst case errors in linear and neural network approximation*, IEEE Trans. Inform. Theory, 48 (2002), pp. 264–275, https://doi.org/10.1109/18.971754. (Cited on p. 130)

[37] A. F. Lerma-Pineda, P. Petersen, S. Frieder, and T. Lukasiewicz, *Dimension-Independent Learning Rates for High-Dimensional Classification Problems*, preprint, arXiv:2409.17991, 2024. (Cited on p. 130)

[38] T. Liang, D. Qiao, Y.-X. Wang, and R. Parhi, *Stable minima of ReLU neural networks suffer from the curse of dimensionality: The neural shattering phenomenon*, in Advances in Neural Information Processing Systems, 2025. (Cited on p. 130)

[39] D. Ludwig, *The Radon transform on Euclidean space*, Commun. Pure Appl. Math., 19 (1966), pp. 49–81, https://doi.org/10.1002/cpa.3160190105. (Cited on pp. 131, 132)

[40] E. Mammen and S. van de Geer, *Locally adaptive regression splines*, Ann. Statist., 25 (1997), pp. 387–413, https://doi.org/10.1214/aos/1034276635. (Cited on pp. 128, 130)

[41] T. Mao, J. W. Siegel, and J. Xu, *Approximation Rates for Shallow* ReLU^k *Neural Networks on Sobolev Spaces Via the Radon Transform*, preprint, arXiv:2408.10996, 2024. (Cited on p. 130)

[42] 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. (Cited on p. 130)

[43] G. F. Montufar, R. Pascanu, K. Cho, and Y. Bengio, *On the number of linear regions of deep neural networks*, in Advances in Neural Information Processing Systems 27, MIT Press, 2014, pp. 2924–2932. (Cited on p. 129)

[44] M. Najaf and G. Ongie, *Sampling Theory for Super-Resolution with Implicit Neural Representations*, preprint, arXiv:2506.09949, 2025. (Cited on p. 131)

[45] B. Neyshabur, R. R. Salakhutdinov, and N. Srebro, *Path-SGD: Path-normalized optimization in deep neural networks*, in Advances in Neural Information Processing Systems, Curran Associates, 2015, pp. 2422–2430. (Cited on pp. 133, 140)

[46] B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro, *Geometry of Optimization and Implicit Regularization in Deep Learning*, preprint, arXiv:1705.03071, 2017. (Cited on p. 140)

[47] B. Neyshabur, R. Tomioka, and N. Srebro, *Norm-based capacity control in neural networks*, in Conference on Learning Theory, PMLR, 2015, pp. 1376–1401. (Cited on p. 140)

[48] G. Ongie, R. Willett, D. Soudry, and N. Srebro, *A function space view of bounded norm infinite width ReLU nets: The multivariate case*, in International Conference on Learning Representations, 2020, pp. 1–28. (Cited on p. 130)

[49] S. Paik, M. Celentano, A. Green, and R. J. Tibshirani, *Integral probability metrics meet neural networks: The Radon-Kolmogorov-Smirnov test*, J. Mach. Learn. Res., 26 (2025), pp. 1–57, https://jmlr.org/papers/v26/24-0245.html. (Cited on p. 131)

[50] R. Parhi, P. Bohra, A. El Biari, M. Pourya, and M. Unser, *Random ReLU neural networks as non-Gaussian processes*, J. Mach. Learn. Res., 26 (2025), pp. 1–31, https://jmlr.org/papers/v26/24-0737.html. (Cited on p. 131)

[51] R. Parhi and R. D. Nowak, *Banach space representer theorems for neural networks and ridge splines*, J. Mach. Learn. Res., 22 (2021), pp. 1–40. (Cited on pp. 128, 129, 130, 131, 132, 133, 141, 142, 143)

[52] 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. (Cited on pp. 127, 128, 129, 130, 131, 139)

[53] 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. (Cited on pp. 133, 136)

[54] 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. (Cited on pp. 128, 129, 130)

[55] 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. (Cited on pp. 131, 132)

[56] R. Parhi and M. Unser, *Function-space optimality of neural architectures with multivariate nonlinearities*, SIAM J. Math. Data Sci., 7 (2025), pp. 110–135, https://doi.org/10.1137/23M1620971. (Cited on p. 130)

[57] S. Parkinson, G. Ongie, and R. Willett, *ReLU neural networks with linear layers are biased towards single- and multi-index models*, SIAM J. Math. Data Sci., 7 (2025), pp. 1021–1052, https://doi.org/10.1137/24M1672158. (Cited on p. 130)

[58] T. Poggio, L. Rosasco, A. Shashua, N. Cohen, and F. Anselmi, *Notes on Hierarchical Splines, DCLNs and i-Theory*, Technical report, Center for Brains, Minds and Machines (CBMM), MIT, Boston, 2015. (Cited on p. 129)

[59] A. Sanyal, P. H. Torr, and P. K. Dokania, *Stable rank normalization for improved generalization in neural networks and GANs*, in International Conference on Learning Representations, 2020, pp. 1–30. (Cited on p. 130)

[60] P. H. P. Savarese, I. Evron, D. Soudry, and N. Srebro, *How do infinite width bounded norm networks look in function space?*, in 32nd Annual Conference on Learning Theory, PMLR, 2019, pp. 2667–2690. (Cited on p. 130)

[61] B. Schölkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, MIT Press, 2002. (Cited on p. 128)

[62] 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, https://www.jmlr.org/papers/v25/23-0677.html. (Cited on pp. 133, 135, 136, 146)

[63] J. Shenouda, Y. Zhou, and R. D. Nowak, *ReLUs are sufficient for learning implicit neural representations*, in International Conference on Machine Learning, PMLR, 2024, pp. 44800–44814. (Cited on p. 131)

[64] 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. (Cited on pp. 128, 130)

[65] 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. (Cited on pp. 128, 130)

[66] L. Spek, T. J. Heeringa, F. Schwenninger, and C. Brune, *Duality for neural networks through reproducing kernel Banach spaces*, Appl. Comput. Harmon. Anal., 78 (2025), 101765, https://doi.org/10.1016/j.acha.2025.101765. (Cited on p. 130)

[67] M. Unser, *A representer theorem for deep neural networks*, J. Mach. Learn. Res., 20 (2019), pp. 1–30, https://jmlr.org/papers/v20/18-418.html. (Cited on p. 129)

[68] 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. (Cited on pp. 133, 146)

[69] 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. (Cited on p. 130)

[70] 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. (Cited on pp. 133, 134, 146)

[71] 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. (Cited on p. 128)

[72] G. Wahba, *Spline Models for Observational Data*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 59, SIAM, 1990, https://doi.org/10.1137/1.9781611970128. (Cited on p. 128)

[73] H. Wang, S. Agarwal, and D. Papailiopoulos, *Pufferfish: Communication-efficient models at no extra cost*, Proc. Mach. Learn. Syst., 3 (2021). (Cited on p. 130)

[74] R. Wang, Y. Xu, and M. Yan, *Hypothesis spaces for deep learning*, Neural Netw., 193 (2026), art. 107995, https://doi.org/10.1016/j.neunet.2025.107995. (Cited on p. 130)

[75] R. Wang, Y. Xu, and M. Yan, *Sparse representer theorems for learning in reproducing kernel Banach spaces*, J. Mach. Learn. Res., 25 (2024), pp. 1–45, https://jmlr.org/papers/v25/23-0645.html. (Cited on p. 130)

[76] Y. Yang and D.-X. Zhou, *Optimal rates of approximation by shallow* ReLU^k *neural networks and applications to nonparametric regression*, Constr. Approx., 62 (2025), pp. 329–360, https://doi.org/10.1007/s00365-024-09679-z. (Cited on p. 130)

[77] C. Zeno, G. Ongie, Y. Blumenfeld, N. Weinberger, and D. Soudry, *How do minimum-norm shallow denoisers look in function space?*, in Advances in Neural Information Processing Systems 36, Curran Associates, 2023, pp. 57520–57557. (Cited on p. 131)

[78] S. Zuhovickiĭ, *Remarks on problems in approximation theory*, Mat. Zbirnik KDU, (1948), pp. 169–183. (Cited on p. 130)