SIAM J. MATH. DATA SCI.
Vol. 4, No. 2, pp. 464–489

What Kinds of Functions Do Deep Neural Networks Learn? Insights from Variational Spline Theory*

Rahul Parhi† and Robert D. Nowak†

Abstract. We develop a variational framework to understand the properties of functions learned by fitting deep neural networks with rectified linear unit (ReLU) activations to data. We propose a new function space, which is related to classical bounded variation-type spaces, that captures the compositional structure associated with deep neural networks. We derive a representer theorem showing that deep ReLU networks are solutions to regularized data-fitting problems over functions from this space. The function space consists of compositions of functions from the Banach space of second-order bounded variation in the Radon domain. This Banach space has a sparsity-promoting norm, giving insight into the role of sparsity in deep neural networks. The neural network solutions have skip connections and rank-bounded weight matrices, providing new theoretical support for these common architectural choices. The variational problem we study can be recast as a finite-dimensional neural network training problem with regularization schemes related to the notions of weight decay and path-norm regularization. Finally, our analysis builds on techniques from variational spline theory, providing new connections between deep neural networks and splines.

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

AMS subject classifications. 46E27, 47A52, 68T05, 82C32, 94A12

DOI. 10.1137/21M1418642

1. Introduction.

A fundamental problem in signal processing, machine learning, and statistics is estimating an unknown function from possibly noisy measurements. Specifically, in supervised learning, the goal is to find a mapping $f : \mathbb{R}^d \to \mathbb{R}^D$ that agrees (in some sense) with a scattered data set $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N \subset \mathbb{R}^d \times \mathbb{R}^D$, i.e., $\boldsymbol{y}_n \approx f(\boldsymbol{x}_n)$, $n = 1, \dots, N$. As there are infinitely many functions that can agree with any given data set, this problem is inherently ill-posed. To circumvent this, some form of *regularization* is imposed on the learning problem. This problem was classically solved via kernel methods, which are solutions to regularized variational problems over reproducing kernel Hilbert spaces [1, 49]. While these variational problems are infinite-dimensional, the reproducing kernel Hilbert space representer theorem [21, 42] says there exists a unique, parametric solution to the problem, allowing the problem to be recast as a finite-dimensional optimization. Kernel methods (even before the term "kernel methods" was coined) have had widespread success dating all the way back to

*Received by the editors May 10, 2021; accepted for publication (in revised form) December 13, 2021; published electronically April 13, 2022.
https://doi.org/10.1137/21M1418642
Funding: This research was partially supported by NSF grant DMS-2134140, 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.

†Department of Electrical and Computer Engineering, University of Wisconsin–Madison, Madison, WI 53706 USA (rahul@ece.wisc.edu, rdnowak@wisc.edu).

the 1960s, especially due to the tight connections between kernels, reproducing kernel Hilbert spaces, and splines [11, 27, 49].

However, the last decade has shown that deep neural networks often outperform kernel methods in a wide variety of tasks, ranging from speech recognition [17] to image classification [22] to solving inverse problems in imaging [19]. Thus, there is great interest in understanding the properties of functions learned from data by neural networks, particularly with the rectified linear unit (ReLU) activation function, which is widely used in practice [25].

Our prior work in [34, 35] has proven *Banach space representer theorems* for single-hidden-layer neural networks with ReLU activations by considering variational problems over certain Banach spaces. In the univariate case, this space is the classical Banach space of second-order bounded variation functions, and the neural network solutions are exactly the well-known locally adaptive linear splines [12, 26, 48]. In the multivariate case, this space is the Banach space of second-order bounded variation functions in the *Radon domain*. It is shown in [34, 35] that these variational problems 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 variational problems. This form of neural network regularization corresponds to the commonly used technique of *weight decay* [23] in gradient descent methods for training neural networks. Due to the similarities of the variational problems studied in [34, 35] with those studied in variational spline theory, we refer to the neural networks in the multivariate case as *ridge splines* of degree one since single-hidden-layer neural networks are simply superpositions of ridge functions, and the functions are multivariate continuous piecewise-linear functions.

This paper extends this characterization to deep (multilayer) neural networks with ReLU activation functions. We also remark that a special property of deep ReLU networks is that their input-output relation is continuous piecewise-linear [28]. 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 [37, 6, 7, 34, 35, 45, 3, 10, 36]. In particular, one can view a deep neural network as a hierarchical or deep spline [37, 6, 7, 45, 3, 10] to emphasize the compositional nature of deep neural networks. Due to this special property, we will work exclusively with ReLU activation functions in this paper, though all of our results are straightforward to extend to any truncated power activation function.

1.1. Contributions. This paper develops a new variational framework to understand the properties of functions learned by deep neural networks fit to data. In particular, we derive a *representer theorem* for the standard fully connected feedforward deep ReLU network architecture. We show that there exist solutions to a certain variational problem that are realizable by a deep ReLU network. Moreover, these deep ReLU networks have skip connections and rank-bounded weight matrices. The number of hidden layers and the rank bounds of the weight matrices are hyperparameters to the variational problem and are therefore controllable a priori. We refer to the neural network solutions as *deep ridge splines* of degree one due to the similarity of the variational problem studied in this paper with the variational

problems studied in variational spline theory. This paper contributes the following new results:

1. We propose a new function space, which is related to classical bounded variation-type spaces, that captures the compositional structure associated with deep neural networks by considering functions that are compositions of functions from the Banach space studied in our previous work [35].
2. We prove a representer theorem that shows that deep ReLU networks with skip connections and rank-bounded weight matrices are solutions to regularized data-fitting problems over functions from this compositional function space.
3. The regularizer in the variational problem corresponds to the sum of the Banach norms of each function in the composition. These are sparsity-promoting norms. Moreover, these regularizers can be expressed in terms of neural network parameters, suggesting several new, principled forms of regularization for deep ReLU networks that promote sparse (in the sense of the number of active neurons) solutions. These regularizers are related to the notion of weight decay in neural network training as well as path-norm regularization.

1.2. Connections to empirical studies in deep learning. Our results provide new theoretical support and insight for a number of empirical findings in deep learning. We show that the common neural network regularization method of "weight decay" [31] corresponds to Radon domain total variation regularization. This characterizes the functional properties of neural networks trained with weight decay—the functions they represent are "smooth" in a precise sense. The optimal solutions to the variational problem require "skip connections" between layers, which provides a new theoretical explanation for the benefits skip connections provide in practice [16]. The sparse nature of our solutions sheds new light on the roles of sparsity and redundancy in deep learning, ranging from "drop-out" [18] to the "lottery ticket hypothesis" [14]. And finally, low-rank weight matrices are a natural by-product of our variational theory that has precedent in practical studies of deep neural networks; it has been empirically observed that low-rank weight matrices can speed up learning [4] and improve accuracy [15], robustness [40], and computational efficiency [50] of deep neural networks.

1.3. Related work. Viewing regularized neural network training problems as variational problems over certain function spaces has received a lot of interest in the last few years [5, 41, 34, 35, 45, 3, 10], although many of the techniques used in these works are quite classical and rooted in variational spline theory and the study of continuous-domain inverse problems [52, 12, 26]. A common theme in these works is to leverage the sparsifying nature of total variation (TV) regularization to learn *sparse solutions*. Our previous work in [34, 35] proves representer theorems for both univariate and multivariate single-hidden-layer neural networks by considering such sparsity-promoting TV regularization. The key analysis tool used in [35] was the Radon transform due to its tight connections with the analysis of ridge functions. This is because single-hidden-layer neural networks are superpositions of ridge functions (neurons). While the connections between ridge functions and the Radon transform are classical, dating back to early work in the representation of solutions to certain partial differential equations as superpositions of ridge functions [20], working with single-hidden-layer ReLU networks in the Radon domain was first studied by [33].

Another line of related work is concerned with the "optimal shaping" of the activation functions in a deep neural network [45, 3, 10]. In particular, [45] proves a representer theorem regarding the optimal shaping of the activation functions. They consider the standard fully connected feedforward deep neural network architecture but allow the activation functions to be learnable. They impose a second-order TV penalty on the activation functions, and so the optimal shaping of the activation functions corresponds to linear splines with adaptive knot locations. We remark that we use several techniques developed in [45, 3] to prove our representer theorem in this paper, particularly in proving existence of solutions to the variational problem we study. Finally, there is a line of work regarding "deep kernel learning" [9], in which they derive a representer theorem for compositions of kernel machines. They consider a construction similar to ours regarding the function space they study, but they consider compositions of reproducing kernel Hilbert spaces, and so the resulting solutions to their variational problem do not take the form of a deep neural network.

1.4. Roadmap. In section 2 we introduce the notation and mathematical formulation used in the remainder of the paper as well as extend the results of [35] in preparation for proving our deep ReLU network representer theorem. 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. Preliminaries. Let $\mathscr{S}(\mathbb{R}^d)$ be the Schwartz space of smooth and rapidly decaying test functions on $\mathbb{R}^d$. Its continuous dual, $\mathscr{S}'(\mathbb{R}^d)$, is the space of tempered distributions on $\mathbb{R}^d$. We are also interested in these spaces on $\mathbb{S}^{d-1} \times \mathbb{R}$, where $\mathbb{S}^{d-1}$ denotes the surface of the Euclidean sphere in $\mathbb{R}^d$. We say $\psi \in \mathscr{S}(\mathbb{S}^{d-1} \times \mathbb{R})$ when ψ is smooth and satisfies the decay condition

$$\sup_{\substack{\boldsymbol{\gamma} \in \mathbb{S}^{d-1} \\ t \in \mathbb{R}}} \left| (1 + |t|^k) \frac{\mathrm{d}^\ell}{\mathrm{d}t^\ell} (\mathrm{D}\, \psi)(\boldsymbol{\gamma}, t) \right| < \infty$$

for all integers $k, \ell \geq 0$ and for all differential operators of all orders D in $\boldsymbol{\gamma}$ [44, Chapter 6]. Since the Schwartz spaces are nuclear, it follows that the above definition is equivalent to saying $\mathscr{S}(\mathbb{S}^{d-1} \times \mathbb{R}) = \mathcal{D}(\mathbb{S}^{d-1}) \mathbin{\widehat{\otimes}} \mathscr{S}(\mathbb{R})$, where $\mathcal{D}(\mathbb{S}^{d-1})$ is the space of smooth functions on $\mathbb{S}^{d-1}$ and $\widehat{\otimes}$ is the *topological* tensor product [51, Chapter III]. We can then define the space of tempered distributions on $\mathbb{S}^{d-1} \times \mathbb{R}$ as its continuous dual, $\mathscr{S}'(\mathbb{S}^{d-1} \times \mathbb{R})$.

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 space of continuous functions vanishing at infinity [13, Chapter 7]. Since $C_0(X)$ is a Banach space when equipped with the uniform norm, we have

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

The norm $\|\cdot\|_{\mathcal{M}(X)}$ is exactly the TV norm (in the sense of measures). As $\mathscr{S}(X)$ is dense in $C_0(X)$, we can associate every measure in $\mathcal{M}(X)$ with a tempered distribution and view $\mathcal{M}(X) \subset \mathscr{S}'(X)$, providing the description

$$\mathcal{M}(X) \coloneqq \{u \in \mathscr{S}'(X) : \ \|u\|_{\mathcal{M}(X)} < \infty\},$$

and so the duality pairing $\langle \cdot, \cdot \rangle$ in (2.1) can be viewed, formally, as the integral

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

where u is viewed as an element of $\mathscr{S}'(X)$. The space $\mathcal{M}(X)$ 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. In this paper, we will mostly work with $X = \mathbb{S}^{d-1} \times \mathbb{R}$.

2.1. Scalar-valued single-hidden-layer ReLU networks and variational problems. Our work in [35] proved a representer theorem for single-hidden-layer ReLU networks with scalar outputs by considering variational problems over the space of functions of second-order bounded variation in the Radon domain. The Radon transform of a function $f : \mathbb{R}^d \to \mathbb{R}$ is given by

$$\mathscr{R}\{f\}(\boldsymbol{\gamma}, t) \coloneqq \int_{\{\boldsymbol{x} : \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{x} = t\}} f(\boldsymbol{x}) \, \mathrm{d}s(\boldsymbol{x}), \quad (\boldsymbol{\gamma}, t) \in \mathbb{S}^{d-1} \times \mathbb{R},$$

where s denotes the Lebesgue measure on the hyperplane $\{\boldsymbol{x} : \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{x} = t\}$. The Radon domain is parameterized by a *direction* $\boldsymbol{\gamma} \in \mathbb{S}^{d-1}$ and an *offset* $t \in \mathbb{R}$. When working with the Radon transform of functions defined on $\mathbb{R}^d$, the following *ramp filter* arises in the Radon inversion formula:

$$\Lambda^{d-1} = (-\partial_t^2)^{\frac{d-1}{2}},$$

where ∂_t denotes the partial derivative with respect to the offset variable, t, of the Radon domain and fractional powers are defined in terms of Riesz potentials. The space of functions of second-order bounded variation in the Radon domain is then given by

$$\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d) = \{f \in L^{\infty,1}(\mathbb{R}^d) : \ \mathscr{R}\,\mathrm{TV}^2(f) < \infty\}, \tag{2.2}$$

where $L^{\infty,1}(\mathbb{R}^d)$ is the Banach space[1] of functions mapping $\mathbb{R}^d \to \mathbb{R}$ of at most linear growth and

$$\mathscr{R}\,\mathrm{TV}^2(f) = c_d \left\| \partial_t^2 \Lambda^{d-1} \mathscr{R} f \right\|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})} \tag{2.3}$$

denotes the second-order TV of a function in the offset variable of the Radon domain, where $c_d^{-1} = 2(2\pi)^{d-1}$ is a dimension-dependant constant that arises when working with the Radon transform. Note that all the operators that appear in (2.3) must be understood in the distributional sense. We refer the reader to [35, section 3] for more details. We now state the main result of [35].

Proposition 2.1 (special case of [35, Theorem 1]). *Consider the problem of interpolating the scattered data* $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ *with* $N > d + 1$. *Then, under the hypothesis of feasibility (i.e.,* $y_n = y_m$ *whenever* $\boldsymbol{x}_n = \boldsymbol{x}_m$*), there exists a solution to the variational problem*

$$\min_{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} \mathscr{R}\,\mathrm{TV}^2(f) \quad s.t. \quad f(\boldsymbol{x}_n) = y_n, \ n = 1, \ldots, N, \tag{2.4}$$

[1] It is a Banach space when equipped with the norm $\|f\|_{\infty,1} \coloneqq \operatorname{ess\,sup}_{\boldsymbol{x} \in \mathbb{R}^d} |f(\boldsymbol{x})| (1 + \|\boldsymbol{x}\|_2)^{-1}$.

of the form

$$s(\boldsymbol{x}) = \sum_{k=1}^{K} v_k \, \rho(\boldsymbol{w}_k^{\mathsf{T}} \boldsymbol{x} - b_k) + \boldsymbol{c}^{\mathsf{T}} \boldsymbol{x} + c_0, \tag{2.5}$$

where $K \leq N - (d+1)$, $\rho = \max\{0, \cdot\}$, $v_k \in \mathbb{R}$, $\boldsymbol{w}_k \in \mathbb{S}^{d-1}$, $b_k \in \mathbb{R}$, $\boldsymbol{c} \in \mathbb{R}^d$, *and* $c_0 \in \mathbb{R}$.

Remark 2.2. Proposition 2.1 says that there always exists a solution to the variational problem in (2.4) that can be realized by a single-hidden-layer ReLU network with a *skip connection* [16], which is the affine term in (2.5). In other words, Proposition 2.1 is a *representer theorem* for single-hidden-layer ReLU networks.

Remark 2.3. As discussed in [35, Remark 3], the fact that $\boldsymbol{w}_k \in \mathbb{S}^{d-1}$ in (2.5) does not restrict the single-hidden-layer neural network due to the positive homogeneity of the ReLU. Indeed, given any single-hidden-layer neural 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 network as

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

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

Given a single-hidden-layer ReLU network, we can explicitly compute its $\mathscr{R}\,\mathrm{TV}^2$-seminorm in terms of network parameters. This is summarized in the following proposition.

Proposition 2.4 (special case of [35, Lemma 25]). *Given a single-hidden-layer neural network*

$$s(\boldsymbol{x}) = \sum_{k=1}^{K} v_k \, \rho(\boldsymbol{w}_k^{\mathsf{T}} \boldsymbol{x} - b_k) + \boldsymbol{c}^{\mathsf{T}} \boldsymbol{x} + c_0,$$

where $\rho = \max\{0, \cdot\}$, $v_k \in \mathbb{R}$, $\boldsymbol{w}_k \in \mathbb{R}^d$, $b_k \in \mathbb{R}$, $\boldsymbol{c} \in \mathbb{R}^d$, *and* $c_0 \in \mathbb{R}$,

$$\mathscr{R}\,\mathrm{TV}^2(s) = \sum_{k=1}^{K} |v_k| \|\boldsymbol{w}_k\|_2. \tag{2.6}$$

We remark that (2.6) is sometimes referred to as the *path-norm* of the network [29]. Moreover, we see that (2.6) 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 network weights.

Note 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 was proven in [35, Theorem 22] that $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ can be turned into a bona fide Banach space when equipped with an appropriate norm.

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

$$\|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} \coloneqq \mathscr{R}\,\mathrm{TV}^2(f) + |f(\boldsymbol{0})| + \sum_{k=1}^{d} |f(\boldsymbol{e}_k) - f(\boldsymbol{0})|, \tag{2.7}$$

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. *For any* $\boldsymbol{x}_0 \in \mathbb{R}^d$*, the Dirac impulse* $\delta(\cdot - \boldsymbol{x}_0) : f \mapsto f(\boldsymbol{x}_0)$ *is weak* continuous on* $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$.

The proof of Lemma 2.5 appears in Appendix A. We remark that item 1 is a corollary of [35, Theorem 22] and item 2 is a new result. In particular, item 2 plays a crucial role in proving the existence of solutions to the variational problem studied in our deep ReLU network representer theorem. The $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$-norm is a sparsity-promoting norm since $\mathscr{R}\,\mathrm{TV}^2(\cdot)$ is defined via an $\mathcal{M}$-norm, the continuous-domain analogue of the ℓ^1-norm.

Remark 2.6. Lemma 2.5 implies that the result of Proposition 2.1 also holds for regularized problems of the form

$$\min_{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} \sum_{n=1}^N \ell(y_n, f(\boldsymbol{x}_n)) + \lambda\, \mathscr{R}\,\mathrm{TV}^2(f),$$

where $\lambda > 0$ is an adjustable regularization parameter and the loss function $\ell(\cdot, \cdot)$ is convex, coercive, and lower semicontinuous. Note that these are slightly weaker conditions on the loss function than in [35, Theorem 1]. This version of the result holds due to the weak* continuity of the Dirac impulse $\delta(\cdot - \boldsymbol{x}_0) : f \mapsto f(\boldsymbol{x}_0)$ on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ combined with [47, Theorem 3] for the conditions on the loss function.

While Proposition 2.1 provides a powerful representer theorem result for single-hidden-layer neural networks, the affine component of any solution is unregularized due to the null space of $\mathscr{R}\,\mathrm{TV}^2(\cdot)$ being the space of affine functions on $\mathbb{R}^d$. Therefore, we modify the problem in (2.4) in order to explicitly regularize the affine component of the functions. This results in the following new representer theorem for single-hidden-layer ReLU networks.

Theorem 2.7. *Consider the problem of interpolating the scattered data* $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ *with* $N > 0$*. Then, under the hypothesis of feasibility (i.e.,* $y_n = y_m$ *whenever* $\boldsymbol{x}_n = \boldsymbol{x}_m$*), there exists a solution to the variational problem*

$$\min_{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} \|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} \quad s.t. \quad f(\boldsymbol{x}_n) = y_n,\ n = 1, \ldots, N, \tag{2.8}$$

of the form

$$s(\boldsymbol{x}) = \sum_{k=1}^K v_k\, \rho(\boldsymbol{w}_k^\mathsf{T} \boldsymbol{x} - b_k) + \boldsymbol{c}^\mathsf{T} \boldsymbol{x} + c_0, \tag{2.9}$$

where $K \leq N$*,* $\rho = \max\{0, \cdot\}$*,* $v_k \in \mathbb{R}$*,* $\boldsymbol{w}_k \in \mathbb{S}^{d-1}$*,* $b_k \in \mathbb{R}$*,* $\boldsymbol{c} \in \mathbb{R}^d$*, and* $c_0 \in \mathbb{R}$*.*

The proof of Theorem 2.7 appears in Appendix B. The key difference between Theorem 2.7 and Proposition 2.1 is that in Theorem 2.7, we are minimizing the $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$-norm rather than the $\mathscr{R}\,\mathrm{TV}^2$-seminorm as in Proposition 2.1. This results in the sparsity of the number of neurons in the solution being N rather than $N - (d+1)$. Additionally, Theorem 2.7 explicitly regularizes the skip connection that appears in (2.9).

Lemma 2.8. *Given a single-hidden-layer neural network*

$$s(\boldsymbol{x}) = \sum_{k=1}^{K} v_k \, \rho(\boldsymbol{w}_k^\mathsf{T} \boldsymbol{x} - b_k) + \boldsymbol{c}^\mathsf{T} \boldsymbol{x} + c_0,$$

where $\rho = \max\{0, \cdot\}$, $v_k \in \mathbb{R}$, $\boldsymbol{w}_k \in \mathbb{R}^d$, $b_k \in \mathbb{R}$, $\boldsymbol{c} \in \mathbb{R}^d$, *and* $c_0 \in \mathbb{R}$,

$$\|s\|_{\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)} = \sum_{k=1}^{K} |v_k| \|\boldsymbol{w}_k\|_2 + |s(\boldsymbol{0})| + \sum_{n=1}^{d} |s(\boldsymbol{e}_n) - s(\boldsymbol{0})|. \tag{2.10}$$

Proof. The result follows from Proposition 2.4 and Lemma 2.5. ∎

2.2. Vector-valued single-hidden-layer ReLU networks and variational problems. Since a deep neural network is the composition of vector-valued single-hidden-layer neural networks, we require a representer theorem for vector-valued single-hidden-layer ReLU networks as a precursor to our representer theorem for deep ReLU networks. Extending Theorem 2.7 for vector-valued functions follows standard techniques. In particular, we follow the technique of [43] which derives a representer theorem for vector-valued smoothing splines.

Lemma 2.9. *Define the vector-valued analogue of* $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ *by the Cartesian product*

$$\underbrace{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d) \times \cdots \times \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)}_{D \text{ times}}.$$

This space can be viewed as the Bochner space $\ell^1([D]; \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d))$, *where* $[D] = \{1, \ldots, D\}$, *and can therefore be equipped with the norm*

$$\|f\|_{\ell^1([D];\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d))} = \sum_{m=1}^{D} \|f_m\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)},$$

where $f = (f_1, \ldots, f_D)$. *For brevity, write* $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d; \mathbb{R}^D)$ *for* $\ell^1([D]; \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d))$. *This space has the following properties:*

1. *It is a Banach space.*
2. *For any* $\boldsymbol{x}_0 \in \mathbb{R}^d$, *the point evaluation operator*

$$\widetilde{\boldsymbol{x}}_0 : f \mapsto f(\boldsymbol{x}_0) = \begin{bmatrix} \langle \delta(\cdot - \boldsymbol{x}_0), f_1 \rangle \\ \vdots \\ \langle \delta(\cdot - \boldsymbol{x}_0), f_D \rangle \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{x}_0) \\ \vdots \\ f_D(\boldsymbol{x}_0) \end{bmatrix}$$

is componentwise weak continuous.*

Proof. Item 1 follows by construction since $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ is itself a Banach space from item 1 in Lemma 2.5. Item 2 follows from item 2 in Lemma 2.5. ∎

Remark 2.10. We can define different (but equivalent) norms on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d; \mathbb{R}^D)$ via the $\ell^p([D]; \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d))$-norms, where $1 \leq p < \infty$. We focus on the case of $p = 1$ in this paper for clarity.

Lemma 2.11. *Let $f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$. Then, f is Lipschitz continuous and satisfies the Lipschitz bound*

$$\|f(\boldsymbol{x}) - f(\boldsymbol{y})\|_1 \leq \|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)}\, \|\boldsymbol{x} - \boldsymbol{y}\|_1.$$

The proof of Lemma 2.11 appears in Appendix D.

Theorem 2.12. *Consider the problem of interpolating the scattered data $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N \subset \mathbb{R}^d \times \mathbb{R}^D$ with $N > 0$. Then, under the hypothesis of feasibility (i.e., $\boldsymbol{y}_n = \boldsymbol{y}_m$ whenever $\boldsymbol{x}_n = \boldsymbol{x}_m$), there exists a solution to the variational problem*

$$\min_{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)} \|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)} \quad s.t. \quad f(\boldsymbol{x}_n) = \boldsymbol{y}_n,\ n = 1, \ldots, N, \tag{2.11}$$

of the form

$$s(\boldsymbol{x}) = \sum_{k=1}^K \boldsymbol{v}_k\, \rho(\boldsymbol{w}_k^\mathsf{T} \boldsymbol{x} - b_k) + \mathbf{C}\boldsymbol{x} + \boldsymbol{c}_0, \tag{2.12}$$

where $K \leq ND$, $\rho = \max\{0, \cdot\}$, $\boldsymbol{v}_k \in \mathbb{R}^D$, $\boldsymbol{w}_k \in \mathbb{S}^{d-1}$, $b_k \in \mathbb{R}$, $\mathbf{C} \in \mathbb{R}^{D \times d}$, and $\boldsymbol{c}_0 \in \mathbb{R}^D$. Moreover, there always exists a solution of the form in (2.12) in which $\boldsymbol{v}_k$ is 1-sparse.

The proof of Theorem 2.12 appears in Appendix C. We also remark that the tightness of the bound $K \leq ND$ is an open question.

Remark 2.13. As discussed in Remark 2.3, the fact that $\boldsymbol{w}_k \in \mathbb{S}^{d-1}$ in (2.12) does not restrict the single-hidden-layer neural network due to the positive homogeneity of the ReLU.

Lemma 2.14. *Given a vector-valued single-hidden-layer neural network*

$$s(\boldsymbol{x}) = \sum_{k=1}^K \boldsymbol{v}_k\, \rho(\boldsymbol{w}_k^\mathsf{T} \boldsymbol{x} - b_k) + \mathbf{C}\boldsymbol{x} + \boldsymbol{c}_0,$$

where $\rho = \max\{0, \cdot\}$, $\boldsymbol{v}_k \in \mathbb{R}^D$, $\boldsymbol{w}_k \in \mathbb{R}^d$, $b_k \in \mathbb{R}$, $\mathbf{C} \in \mathbb{R}^{D \times d}$, and $\boldsymbol{c}_0 \in \mathbb{R}^D$,

$$\|s\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)} = \sum_{k=1}^K \|\boldsymbol{v}_k\|_1 \|\boldsymbol{w}_k\|_2 + \sum_{m=1}^D \left(|s_m(\mathbf{0})| + \sum_{n=1}^d |s_m(\boldsymbol{e}_n) - s_m(\mathbf{0})|\right). \tag{2.13}$$

Proof. For $m = 1, \ldots, D$, we can write

$$s_m(\boldsymbol{x}) = \sum_{k=1}^K v_{k,m}\, \rho(\boldsymbol{w}_k^\mathsf{T} \boldsymbol{x} - b_k) + \boldsymbol{c}_m^\mathsf{T} \boldsymbol{x} + c_{0,m},$$

where s_m is the mth component of s, $\boldsymbol{c}_m$ is the mth row of $\mathbf{C}$, and $c_{0,m}$ is the mth component of $\boldsymbol{c}_0$. The result follows from Lemma 2.8 and the definition of the $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$-norm. ∎

3. A representer theorem for deep ReLU networks. In this section, we will prove our representer theorem for deep ReLU networks. We consider functions that are compositions of functions from the Banach spaces defined in Lemma 2.9. Let

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

denote the space of all such functions.

For brevity, we will write $\mathscr{R}\,\mathrm{BV}^2_{\mathsf{deep}}(L)$ for $\mathscr{R}\,\mathrm{BV}^2_{\mathsf{deep}}(\mathbb{R}^{d_0};\cdots;\mathbb{R}^{d_L})$. This definition reflects two standard architectural specifications 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 will ultimately correspond to a layer in a deep neural network in our representer theorem.

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*

$$\|f(\boldsymbol{x})-f(\boldsymbol{y})\|_1\le\left(\prod_{\ell=1}^{L}\left\|f^{(\ell)}\right\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})}\right)\|\boldsymbol{x}-\boldsymbol{y}\|_1.$$

Proof. The result follows by repeatedly applying Lemma 2.11. ∎

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 $d_0,\dots,d_L$ be positive integers corresponding to the intermediate dimensions of a deep neural network. Consider the problem of approximating the scattered data $\{(\boldsymbol{x}_n,\boldsymbol{y}_n)\}_{n=1}^N\subset\mathbb{R}^{d_0}\times\mathbb{R}^{d_L}$ with $N>0$ denoting the number of data. Let $\ell(\cdot,\cdot)$ be an arbitrary nonnegative lower semicontinuous loss function, and let $\lambda>0$ be a regularization parameter. Then, there exists a solution to the variational problem*

$$\min_{\substack{f^{(\ell)}\in\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})\\ \ell=1,\dots,L\\ f=f^{(L)}\circ\cdots\circ f^{(1)}}}\sum_{n=1}^{N}\ell(\boldsymbol{y}_n,f(\boldsymbol{x}_n))+\lambda\sum_{\ell=1}^{L}\left\|f^{(\ell)}\right\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})}\tag{3.1}$$

of the form

$$s(\boldsymbol{x})=\boldsymbol{x}^{(L)},\tag{3.2}$$

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

$$\begin{cases}\boldsymbol{x}^{(0)}:=\boldsymbol{x},\\ \boldsymbol{x}^{(\ell)}:=\mathbf{V}^{(\ell)}\boldsymbol{\rho}(\mathbf{W}^{(\ell)}\boldsymbol{x}^{(\ell-1)}-\boldsymbol{b}^{(\ell)})+\mathbf{C}^{(\ell)}\boldsymbol{x}^{(\ell-1)}+\boldsymbol{c}_0^{(\ell)},\quad \ell=1,\dots,L,\end{cases}\tag{3.3}$$

where $\boldsymbol{\rho}$ applies $\rho=\max\{0,\cdot\}$ componentwise and, for $\ell=1,\dots,L$, $\mathbf{V}^{(\ell)}\in\mathbb{R}^{d_\ell\times K^{(\ell)}}$, $\mathbf{W}^{(\ell)}\in\mathbb{R}^{K^{(\ell)}\times d_{\ell-1}}$, $\boldsymbol{b}^{(\ell)}\in\mathbb{R}^{K^{(\ell)}}$, $\mathbf{C}^{(\ell)}\in\mathbb{R}^{d_\ell\times d_{\ell-1}}$, and $\boldsymbol{c}_0^{(\ell)}\in\mathbb{R}^{d_\ell}$, where $K^{(\ell)}\le Nd_\ell$.

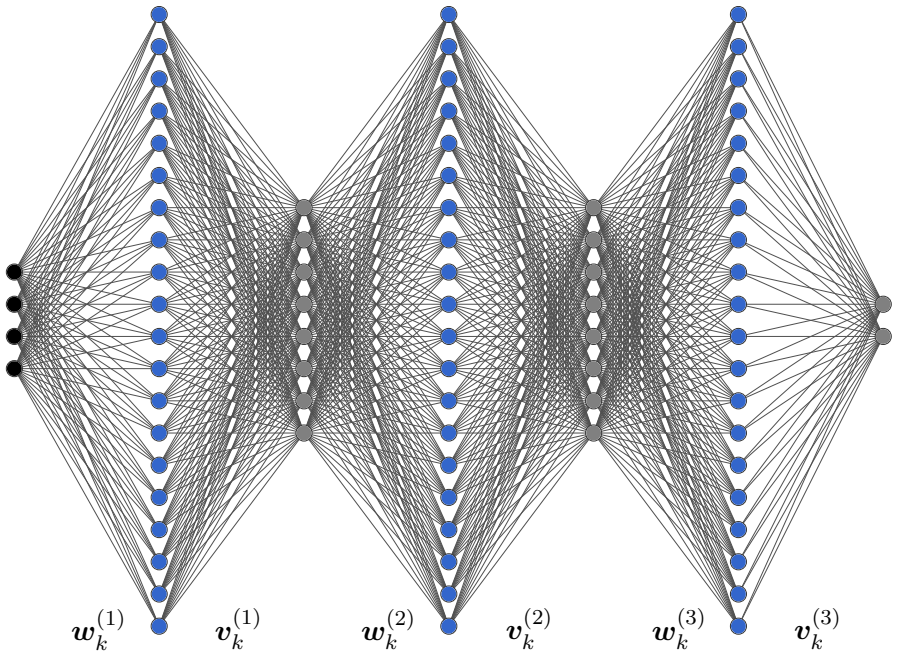

Figure 1. *This figure shows the architecture of the deep neural network in (3.3) in the case of $L = 3$ hidden layers. The black nodes denote input nodes, the blue nodes denote ReLU nodes, and the gray nodes denote linear nodes. Skip connection nodes are omitted for clarity.*

Remark 3.3. Note that the search space in (3.1) is over the Cartesian product

$$\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_0};\mathbb{R}^{d_1}) \times \cdots \times \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{L-1}};\mathbb{R}^{d_L}) \tag{3.4}$$

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 (3.1) to be well defined, we formulate the problem over (3.4).

Remark 3.4. Theorem 3.2 also holds for the problem of interpolating scattered data.

The neural network architecture that appears in (3.3) can be seen in Figure 1. Moreover, this exact architecture was recently studied in the empirical work in [15] and is referred to as a deep ReLU network with *linear bottlenecks*. Since the variational problem in (3.1) is reminiscent of the variational problems studied in variational spline theory, and since the resulting deep ReLU network solution in (3.2) is a continuous piecewise-linear function, in a similar vein to [45, 3, 10], we refer to such functions as *deep ridge splines* of degree one.

Remark 3.5. Since the regularizer in (3.1) directly controls the $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})$-norm of each layer, we see from Lemma 2.11 that the variational problem is essentially regularizing a bound on the Lipschitz constant of the function.

Remark 3.6. The regularizer that appears in (3.1) can be replaced by

$$\psi_0\left(\sum_{\ell=1}^{L} \psi_\ell\left(\left\|f^{(\ell)}\right\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})}\right)\right),$$

where $\psi_\ell : [0,\infty) \to \mathbb{R}$, $\ell = 0,\dots,L$, is a strictly increasing and convex function, and still admit a solution that takes the form of a deep neural network as in (3.2). Thus, there are many choices of regularization that result in a representer theorem for deep ReLU networks.

Remark 3.7. Notice that (3.2) is precisely the standard L-hidden layer deep ReLU network architecture with *rank-bounded weight matrices* and *skip connections*. Indeed, the weight matrix of the ℓth layer is $\mathbf{A}^{(\ell)} \coloneqq \mathbf{W}^{(\ell+1)}\mathbf{V}^{(\ell)}$. More specifically, by dropping biases and skip connections for clarity, we see that $s(\boldsymbol{x})$ in (3.2) can be computed recursively as

$$\begin{cases} \widetilde{\boldsymbol{x}}^{(0)} \coloneqq \boldsymbol{x}, \\ \widetilde{\boldsymbol{x}}^{(\ell)} \coloneqq \boldsymbol{\rho}(\mathbf{A}^{(\ell-1)}\widetilde{\boldsymbol{x}}^{(\ell-1)}), \quad \ell = 1,\dots,L, \\ s(\boldsymbol{x}) \coloneqq \mathbf{A}^{(L)}\widetilde{\boldsymbol{x}}^{(L)}, \end{cases} \tag{3.5}$$

where

$$\begin{cases} \mathbf{A}^{(0)} \coloneqq \mathbf{W}^{(1)}, \\ \mathbf{A}^{(\ell)} \coloneqq \mathbf{W}^{(\ell+1)}\mathbf{V}^{(\ell)}, \quad \ell = 2,\dots,L-1, \\ \mathbf{A}^{(L)} \coloneqq \mathbf{V}^{(L)}. \end{cases}$$

From the dimensions of $\mathbf{V}^{(\ell)}$ and $\mathbf{W}^{(\ell)}$ in Theorem 3.2, we see that for $\ell = 0,\dots,L$, $\operatorname{rank}(\mathbf{A}^{(\ell)}) \le \min\{Nd_{\ell+1}, d_\ell\}$, and $\operatorname{rank}(\mathbf{A}^{(L)}) \le d_L$. In a typical scenario, where the $\{d_\ell\}_{\ell=1}^{L}$ are of the same order, this says that $\operatorname{rank}(\mathbf{A}^{(\ell)}) \le d_\ell$.

Remark 3.8. The architecture of the network in (3.3) is not restrictive of what functions can be represented by such a network. In particular, the architecture in (3.3) is as expressive as the standard deep ReLU network architecture with hidden layer widths of $d_1,\dots,d_L$.

Proof of Theorem 3.2. Given $f = f^{(L)} \circ \cdots \circ f^{(1)}$ such that $f^{(\ell)} \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})$, $\ell = 1,\dots,L$, write

$$\mathcal{J}(f) \coloneqq \mathcal{J}(f^{(1)},\dots,f^{(L)}) \coloneqq \sum_{n=1}^{N} \ell(\boldsymbol{y}_n, f(\boldsymbol{x}_n)) + \lambda \sum_{\ell=1}^{L} \left\| f^{(\ell)} \right\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})}$$

for the objective value of f. Next, consider an arbitrary $g = g^{(L)} \circ \cdots \circ g^{(1)}$ such that $g^{(\ell)} \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})$, $\ell = 1,\dots,L$, with objective value $C \coloneqq \mathcal{J}(g)$. We may transform the unconstrained problem in (3.1) into the equivalent constrained problem

$$\min_{\substack{f^{(\ell)} \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell}) \\ \ell=1,\dots,L \\ f = f^{(L)} \circ \cdots \circ f^{(1)}}} \mathcal{J}(f) \quad \text{s.t.} \quad \left\| f^{(\ell)} \right\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})} \le C/\lambda,\ \ell = 1,\dots,L. \tag{3.6}$$

This transformation is valid since any function that does not satisfy the constraints in (3.6) has a strictly larger objective value than g and is therefore not in the solution set.

For $f_0 = f_0^{(L)} \circ \cdots \circ f_0^{(1)}$, $f_0^{(\ell)} \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})$, $\ell = 1,\dots,L$, we will show that the map $f_0^{(\widetilde{\ell})} \mapsto \mathcal{J}(f_0)$, for a fixed $\widetilde{\ell} \in \{1,\dots,L\}$, is weak* lower semicontinuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\widetilde{\ell}-1}};\mathbb{R}^{d_{\widetilde{\ell}}})$. First notice that the map $f_0^{(\widetilde{\ell})} \mapsto f_0(\boldsymbol{x}_0)$, for any $\boldsymbol{x}_0 \in \mathbb{R}^d$, is componentwise weak* continuous

on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\tilde{\ell}-1}};\mathbb{R}^{d_{\tilde{\ell}}})$. Indeed, since each $f_0^{(\ell)}$, $\ell = 1,\dots,L$, is componentwise continuous by Lemma 2.11, and since the point evaluation is componentwise weak* continuous by Lemma 2.9, the map $f_0^{(\tilde{\ell})} \mapsto f_0^{(L)} \circ \cdots \circ f_0^{(1)}(\boldsymbol{x}_0)$ is made up of compositions of componentwise continuous and componentwise weak* continuous functions and is therefore itself componentwise weak* continuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\tilde{\ell}-1}};\mathbb{R}^{d_{\tilde{\ell}}})$. Next, since the loss function is lower semicontinuous and every norm is weak* continuous on its corresponding Banach space, we have that $f_0^{(\tilde{\ell})} \mapsto \mathcal{J}(f_0)$ is weak* lower semicontinuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\tilde{\ell}-1}};\mathbb{R}^{d_{\tilde{\ell}}})$. Therefore, $(f_0^{(1)},\dots,f_0^{(L)}) \mapsto \mathcal{J}(f)$ is weak* lower semicontinuous over the Cartesian product in (3.4). Finally, by the Banach–Alaoglu theorem [39, Chapter 3], the feasible set in (3.6) is weak* compact. Thus, there exists a solution to (3.6) (and subsequently (3.1)) by the Weierstrass extreme value theorem on general topological spaces [24, Chapter 5].

Let $\widetilde{s} = \widetilde{s}^{(L)} \circ \cdots \circ \widetilde{s}^{(1)}$ be a (not necessarily unique) solution to (3.1). By applying $\widetilde{s}$ to each data point $\boldsymbol{x}_n$, $n = 1,\dots,N$, we can recursively compute the intermediate vectors $\boldsymbol{z}_{n,\ell} \in \mathbb{R}^{d_\ell}$ as follows:

- Initialize $\boldsymbol{z}_{n,0} \coloneqq \boldsymbol{x}_n$.
- For each $\ell = 1,\dots,L$, recursively update $\boldsymbol{z}_{n,\ell} \coloneqq \widetilde{s}^{(\ell)}(\boldsymbol{z}_{n,\ell-1})$.

The solution $\widetilde{s}$ must satisfy

$$\widetilde{s}^{(\ell)} \in \underset{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})}{\arg\min} \|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})} \quad \text{s.t.} \quad f(\boldsymbol{z}_{n,\ell-1}) = \boldsymbol{z}_{n,\ell},\ n = 1,\dots,N,$$

for $\ell = 1,\dots,L$. To see this, note that if the above display did not hold, it would contradict the optimality of $\widetilde{s}$. By Theorem 2.12, there always exists a solution to the above display that enforces the form of the solution in (3.2). ∎

4. Applications to deep network training and regularization. In this section we will discuss applications of the representer theorem in Theorem 3.2 to the training and regularization of deep ReLU networks. Since Theorem 3.2 guarantees existence of a solution to the variational problem in (3.1) that is realizable by a deep ReLU network as in (3.2), one can find a solution to (3.1) by finding a solution to a finite-dimensional deep network training problem.

Lemma 4.1. *Given a deep neural network $s = s^{(L)} \circ \cdots \circ s^{(1)}$ as in (3.2),*

$$\sum_{\ell=1}^{L} \left\| s^{(\ell)} \right\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})} = \sum_{\ell=1}^{L} \left(\sum_{k=1}^{K^{(\ell)}} \left\| \boldsymbol{v}_k^{(\ell)} \right\|_1 \left\| \boldsymbol{w}_k^{(\ell)} \right\|_2 + \sum_{m=1}^{D} \left(\left| s_m^{(\ell)}(\boldsymbol{0}) \right| + \sum_{n=1}^{d} \left| s_m^{(\ell)}(\boldsymbol{e}_n) - s_m^{(\ell)}(\boldsymbol{0}) \right| \right) \right),$$

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

Proof. The proof follows by invoking Lemma 2.14 on each $s^{(\ell)}$, $\ell = 1,\dots,L$. ∎

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 (3.2), and let $\Theta = \mathbb{R}^M$ denote the space of these parameters, where M is the total number of scalar parameters*

in the network. 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

$$\begin{aligned}\min_{\boldsymbol{\theta}\in\Theta} &\sum_{n=1}^{N} \ell(\boldsymbol{y}_n, f_{\boldsymbol{\theta}}(\boldsymbol{x}_n)) \\ &+ \lambda \sum_{\ell=1}^{L} \left(\sum_{k=1}^{K^{(\ell)}} \left\| \boldsymbol{v}_k^{(\ell)} \right\|_1 \left\| \boldsymbol{w}_k^{(\ell)} \right\|_2 + \sum_{m=1}^{D} \left(\left| f_{\boldsymbol{\theta},m}^{(\ell)}(\mathbf{0}) \right| + \sum_{n=1}^{d} \left| f_{\boldsymbol{\theta},m}^{(\ell)}(\boldsymbol{e}_n) - f_{\boldsymbol{\theta},m}^{(\ell)}(\mathbf{0}) \right| \right) \right),\end{aligned} \tag{4.1}$$

where $\{(\boldsymbol{x}_n, \boldsymbol{y}_n)\}_{n=1}^N \subset \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ is a scattered data set, $\ell(\cdot,\cdot)$ is an arbitrary nonnegative lower semicontinuous loss function, and $\lambda > 0$ is an adjustable regularization parameter, are solutions to (3.1) so long as $K^{(\ell)} \geq N d_\ell$.

We can also consider a different regularizer than the one in Corollary 4.2 that results in a finite-dimensional neural network training problem equivalent to (4.1).

Corollary 4.3. *The solutions to*

$$\begin{aligned}\min_{\boldsymbol{\theta}\in\Theta} &\sum_{n=1}^{N} \ell(\boldsymbol{y}_n, f_{\boldsymbol{\theta}}(\boldsymbol{x}_n)) \\ &+ \lambda \sum_{\ell=1}^{L} \left(\frac{\left\| \mathbf{V}^{(\ell)} \right\|_{1,2}^2 + \left\| \mathbf{W}^{(\ell)} \right\|_{\mathsf{F}}^2}{2} + \sum_{m=1}^{D} \left(\left| f_{\boldsymbol{\theta},m}^{(\ell)}(\mathbf{0}) \right| + \sum_{n=1}^{d} \left| f_{\boldsymbol{\theta},m}^{(\ell)}(\boldsymbol{e}_n) - f_{\boldsymbol{\theta},m}^{(\ell)}(\mathbf{0}) \right| \right) \right)\end{aligned} \tag{4.2}$$

are also solutions to (4.1), where

$$\left\| \mathbf{V}^{(\ell)} \right\|_{1,2}^2 := \sum_{k=1}^{K^{(\ell)}} \left\| \boldsymbol{v}_k^{(\ell)} \right\|_1^2$$

is the mixed $\ell^1\ell^2$-norm of $\mathbf{V}^{(\ell)}$ and $\|\cdot\|_{\mathsf{F}}$ is the usual Frobenius norm of a matrix. Moreover, the solutions to (4.2) satisfy the property that $\|\boldsymbol{v}_k^{(\ell)}\|_1 = \|\boldsymbol{w}_k^{(\ell)}\|_2$, $\ell = 1, \ldots, L$, $k = 1, \ldots, K^{(\ell)}$.

Proof. The kth neuron in the ℓth layer of a deep neural network as in (3.2) takes the form $\boldsymbol{x} \mapsto \boldsymbol{v}_k^{(\ell)} \rho({\boldsymbol{w}_k^{(\ell)}}^{\mathsf{T}} \boldsymbol{x} - b_k^{(\ell)})$. Due to the positive homogeneity of the ReLU, $\boldsymbol{v}_k^{(\ell)}$ and $\boldsymbol{w}_k^{(\ell)}$ can be rescaled so that $\|\boldsymbol{v}_k^{(\ell)}\|_1 = \|\boldsymbol{w}_k^{(\ell)}\|_2$ without altering the function of the network. Therefore, minimizing $\|\boldsymbol{v}_k^{(\ell)}\|_1^2 + \|\boldsymbol{w}_k^{(\ell)}\|_2^2$ is achieved when $\|\boldsymbol{v}_k^{(\ell)}\|_1 = \|\boldsymbol{w}_k^{(\ell)}\|_2$. The result then follows from the fact that when $\|\boldsymbol{v}_k^{(\ell)}\|_1 = \|\boldsymbol{w}_k^{(\ell)}\|_2$ we have

$$\frac{\left\| \boldsymbol{v}_k^{(\ell)} \right\|_1^2 + \left\| \boldsymbol{w}_k^{(\ell)} \right\|_2^2}{2} = \left\| \boldsymbol{v}_k^{(\ell)} \right\|_1 \left\| \boldsymbol{w}_k^{(\ell)} \right\|_2.$$

∎

Remark 4.4. While the problems in (4.1) and (4.2) take the form of neural network training problems with new, principled forms of regularization, it's important to note that the problems are nonconvex, and our results say nothing about algorithms for actually solving the optimization problems.

Remark 4.5. Due to the sparsity-promoting nature of the $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})$-norms, the regularizers that appear in (4.1) and (4.2) promote sparse (in the sense of the number of active neurons) deep ReLU network solutions.

Remark 4.6. The term

$$\sum_{m=1}^{D}\left(\left|f^{(\ell)}_{\boldsymbol{\theta},m}(\mathbf{0})\right|+\sum_{n=1}^{d}\left|f^{(\ell)}_{\boldsymbol{\theta},m}(\boldsymbol{e}_n)-f^{(\ell)}_{\boldsymbol{\theta},m}(\mathbf{0})\right|\right)$$

that appears in (4.1) and (4.2) simply imposes an ℓ^1-norm on the coefficients of the affine part (i.e., the skip connection in the neural network realization) on each layer (see Appendix A). Therefore, one may also consider the regularizers

$$\sum_{\ell=1}^{L}\left(\sum_{k=1}^{K^{(\ell)}}\left\|\boldsymbol{v}_k^{(\ell)}\right\|_1\left\|\boldsymbol{w}_k^{(\ell)}\right\|_2+\left\|\mathbf{C}^{(\ell)}\right\|_{1,1}+\left\|\boldsymbol{c}_0^{(\ell)}\right\|_1\right) \tag{4.3}$$

in place of (4.1) or

$$\sum_{\ell=1}^{L}\left(\frac{\left\|\mathbf{V}^{(\ell)}\right\|_{1,2}^2+\left\|\mathbf{W}^{(\ell)}\right\|_{\mathsf{F}}^2}{2}+\left\|\mathbf{C}^{(\ell)}\right\|_{1,1}+\left\|\boldsymbol{c}_0^{(\ell)}\right\|_1\right), \tag{4.4}$$

in place of (4.2), where

$$\|\mathbf{C}\|_{1,1}\coloneqq\sum_{m=1}^{D}\sum_{k=1}^{d}|c_{m,k}|$$

denotes the mixed $\ell^1\ell^1$-norm of $\mathbf{C}$.

Remark 4.7. It is common in many deep learning papers to consider deep neural networks without biases and skip connections (see, e.g., [29, 30, 32, 8]). Since the term

$$\sum_{m=1}^{D}\left(\left|f^{(\ell)}_{\boldsymbol{\theta},m}(\mathbf{0})\right|+\sum_{n=1}^{d}\left|f^{(\ell)}_{\boldsymbol{\theta},m}(\boldsymbol{e}_n)-f^{(\ell)}_{\boldsymbol{\theta},m}(\mathbf{0})\right|\right) \tag{4.5}$$

that appears in (4.1) and (4.2) simply imposes an ℓ^1-norm on the coefficients of the affine part (i.e., the skip connection in the neural network realization) on each layer (see Appendix A), the following two regularizers naturally arise from our variational framework in the case of a deep neural network with no biases or skip connections:

$$\sum_{\ell=1}^{L}\sum_{k=1}^{K^{(\ell)}}\left\|\boldsymbol{v}_k^{(\ell)}\right\|_1\left\|\boldsymbol{w}_k^{(\ell)}\right\|_2 \tag{4.6}$$

or

$$\frac{1}{2}\sum_{\ell=1}^{L}\left\|\mathbf{V}^{(\ell)}\right\|_{1,2}^2+\left\|\mathbf{W}^{(\ell)}\right\|_{\mathsf{F}}^2, \tag{4.7}$$

where (4.6) and (4.7) are in fact equivalent by the same argument as in the proof of Corollary 4.3.

4.1. Connections to existing deep network regularization schemes. The regularizers that appear in (4.1)–(4.4), (4.6), and (4.7) are *principled* regularizers for training deep ReLU networks. Moreover, we will show in this section that the discussed regularizers are related to the well-known weight decay [23] and path-norm [29] regularizers for deep ReLU networks.

Training a neural network with weight decay is one of the most common regularization schemes for deep ReLU networks. This corresponds to an ℓ^2-norm regularization on all the network weights. The regularizer that appears in (4.7) almost takes the form of an ℓ^2-norm of the network weights except that the regularization on the $\mathbf{V}^{(\ell)}$ is not the Frobenius norm. By considering a slightly different architecture than in (3.3), where it is imposed that the columns of $\mathbf{V}^{(\ell)}$ are 1-sparse, the regularizer in (4.7) exactly corresponds to weight decay (since $\|\mathbf{V}^{(\ell)}\|_{1,2}^2 = \|\mathbf{V}^{(\ell)}\|_{\mathsf{F}}^2$ when the columns of $\mathbf{V}^{(\ell)}$ are 1-sparse). Training this architecture with this regularizer still corresponds to finding a solution to the variational problem in (3.1) since it simply imposes that the outputs of each layer of the deep network are completely *decoupled* (see Remark C.1). The utility of *not* considering such an architecture is to promote *neuron sharing* between the outputs of each layer.

Another common regularization scheme for deep ReLU networks is the path-norm regularizer. In particular, several works [29, 30, 32, 8] 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

$$\begin{cases} \boldsymbol{x}^{(0)} \coloneqq \boldsymbol{x}, \\ \boldsymbol{x}^{(\ell)} \coloneqq \boldsymbol{\rho}(\mathbf{A}^{(\ell-1)}\boldsymbol{x}^{(\ell-1)}), \quad \ell = 1, \dots, L, \\ x^{(L)} \coloneqq {\boldsymbol{a}^{(L)}}^{\mathsf{T}} \boldsymbol{x}^{(L)}, \end{cases} \tag{4.8}$$

where $\boldsymbol{\rho}$ denotes applying ρ 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 (4.8) is almost the same as the architecture in our framework if we drop biases and skip connections (see (3.5) in Remark 3.7). These works then consider path-norm regularization of the form

$$\sum_{k_L=1}^{K^{(L)}} \sum_{k_{L-1}=1}^{K^{(L-1)}} \cdots \sum_{k_1=1}^{K^{(1)}} \sum_{k_0=1}^{d} |a_{k_0,k_1}||a_{k_1,k_2}| \cdots \left|a_{k_{L-1},k_L}\right||a_{k_L}|, \tag{4.9}$$

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,[2] which arises with a particular choice of $\{\psi_\ell\}_{\ell=0}^{L}$ in Remark 3.6,

$$\prod_{\ell=1}^{L} \sum_{k=1}^{K^{(\ell)}} \left\|\boldsymbol{v}_k^{(\ell)}\right\|_1 \left\|\boldsymbol{w}_k^{(\ell)}\right\|_2. \tag{4.10}$$

[2]Where we drop the term in (4.5) as discussed in Remark 4.7.

We have that (4.10) is an upper bound on something that looks very similar to the path-norm in (4.9). Indeed, first notice that if we write the deep ReLU network from our framework in the form in (3.5), we have

$$|a_{k_\ell,k_{\ell+1}}| = \left|\boldsymbol{v}_k^{(\ell)\mathsf{T}} \boldsymbol{w}_k^{(\ell+1)}\right| \leq \left\|\boldsymbol{v}_k^{(\ell)}\right\|_2 \left\|\boldsymbol{w}_k^{(\ell+1)}\right\|_2 \leq \left\|\boldsymbol{v}_k^{(\ell)}\right\|_1 \left\|\boldsymbol{w}_k^{(\ell+1)}\right\|_2, \tag{4.11}$$

where $a_{k_\ell,k_{\ell+1}}$ denotes the $(k_\ell, k_{\ell+1})$th entry in $\mathbf{A}^{(\ell)}$ as defined in Remark 3.7. Therefore,

$$\begin{aligned}\prod_{\ell=1}^{L} \sum_{k=1}^{K^{(\ell)}} \left\|\boldsymbol{v}_k^{(\ell)}\right\|_1 \left\|\boldsymbol{w}_k^{(\ell)}\right\|_2 &= \sum_{k_L=1}^{K^{(L)}} \cdots \sum_{k_1=1}^{K^{(1)}} \left\|\boldsymbol{w}_{k_1}^{(1)}\right\|_2 \left\|\boldsymbol{v}_{k_1}^{(1)}\right\|_1 \left\|\boldsymbol{w}_{k_2}^{(2)}\right\|_2 \left\|\boldsymbol{v}_{k_2}^{(2)}\right\|_1 \cdots \left\|\boldsymbol{w}_{k_L}^{(L)}\right\|_2 \left\|\boldsymbol{v}_{k_L}^{(L)}\right\|_1 \\ &\geq \sum_{k_L=1}^{K^{(L)}} \cdots \sum_{k_1=1}^{K^{(1)}} \left\|\boldsymbol{w}_{k_1}^{(1)}\right\|_2 |a_{k_1,k_2}| \cdots |a_{k_{L-1},k_L}| \left\|\boldsymbol{v}_{k_L}^{(L)}\right\|_1,\end{aligned}$$

where the last line holds from (4.11). We see that the last line in the above display is the same as the path-norm in (4.9), apart from how it treats weights in the first and last layers. We also remark that the work in [8] shows that the path-norm in (4.9) 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.[3] 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 compositions of functions in $\mathscr{R}\,\mathrm{BV}^2$-spaces. This variational problem can be recast as a finite-dimensional neural network training problem with various choices of regularization. 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 more understanding of the compositional space $\mathscr{R}\,\mathrm{BV}^2_{\mathsf{deep}}(L)$. This entails first having further understanding of the $\mathscr{R}\,\mathrm{BV}^2$-spaces. The function spaces studied in this paper are new and not classical, and future work will be directed at understanding how these new spaces are related to classical function spaces studied in functional analysis.

Appendix A. Topological properties of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$. In this section we will prove Lemma 2.5. We will rely on many results developed in [35]. While the definition of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ given in (2.2) 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 in $\mathbb{R}^d$. Thus, we use the result of [35, Theorem 22] to characterize $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ as a Banach space. In particular, [35, 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*

[3] As stated in the introduction, all our results are straightforward to generalize to any truncated power activation function.

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* $\langle \phi_k, p_n \rangle = \delta[k-n]$*,* $k, n = 0, \dots, N_0 - 1$*, where* $\delta[\cdot]$ *is the Kronecker impulse.*

Proposition A.2 (see [35, 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*

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

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

Proof of Lemma 2.5, item 1. By Proposition A.2 it suffices to find a biorthogonal system $(\boldsymbol{\phi}, \boldsymbol{p})$ of $\mathcal{P}_1(\mathbb{R}^d)$ so that for every $f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ we have

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

Put $p_0(\boldsymbol{x}) \coloneqq 1$ and $p_k(\boldsymbol{x}) \coloneqq x_k$, $k = 1, \dots, d$. Clearly $\boldsymbol{p}$ is a basis for $\mathcal{P}_1(\mathbb{R}^d)$. Put $\phi_0 \coloneqq \delta$ and $\phi_k \coloneqq \delta(\cdot - \boldsymbol{e}_k) - \delta$, $k = 1, \dots, d$, 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 $\langle \phi_k, p_k \rangle = p_k(\boldsymbol{e}_k) - p_k(\mathbf{0}) = 1 - 0 = 1$, $k = 1, \dots, d$. We also have

$$\begin{aligned}
&\langle \phi_0, p_k \rangle = p_k(\mathbf{0}) = 0, \quad k = 1, \dots, d, \\
&\langle \phi_k, p_0 \rangle = p_0(\boldsymbol{e}_k) - p_0(\mathbf{0}) = 1 - 1 = 0, \quad k = 1, \dots, d, \\
&\langle \phi_k, p_n \rangle = p_n(\boldsymbol{e}_k) - p_n(\mathbf{0}) = 0 + 0 = 0, \quad k, n = 1, \dots, d, \quad k \neq n.
\end{aligned}$$

A computation shows that (A.1) holds with this choice of biorthogonal system. ∎

In order to prove item 2 of Lemma 2.5, we must show that the Dirac impulse, $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, is weak* continuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$. The following proposition characterizes the weak* continuous linear functionals on a Banach space.

Proposition A.3 (see [38, Theorem IV.20, page 114]). *Let* $\mathcal{X}$ *be a Banach space. The only weak* continuous linear functionals on* $\mathcal{X}'$ *(the continuous dual of* $\mathcal{X}$*) are elements of* $\mathcal{X}$*.*

Therefore, we must show that the Dirac impulse is contained in the predual of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$. Before we can prove this, we require an important result from [35]. Recall from (2.3) that

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

Put $\mathrm{R} \coloneqq c_d\, \partial_t^2 \Lambda^{d-1}\, \mathscr{R}$. As discussed in [35], for every $f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$, $u \coloneqq \mathrm{R}\, f \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})$ is always even, i.e., $u(\boldsymbol{\gamma}, t) = u(-\boldsymbol{\gamma}, -t)$. This means we have

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

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 $h_{(\boldsymbol{\gamma}, t)} \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{x} = t\}$ for some $(\boldsymbol{\gamma}, t) \in \mathbb{S}^{d-1} \times \mathbb{R}$ and $h_{(\boldsymbol{\gamma}, t)} = h_{(-\boldsymbol{\gamma}, -t)}$.

Proposition A.4 (see [35, 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*

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

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

$$\mathrm{R}_{\boldsymbol{\phi}}^{-1} : u \mapsto \int_{\mathbb{S}^{d-1}\times\mathbb{R}} g_{\boldsymbol{\phi}}(\cdot, \boldsymbol{z}) u(\boldsymbol{z})\, \mathrm{d}(\sigma \times \lambda)(\boldsymbol{z}), \tag{A.2}$$

where σ is the surface measure on $\mathbb{S}^{d-1}$ and λ is the Lebesgue measure on $\mathbb{R}$ and

$$g_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{z}) = r_{\boldsymbol{z}}(\boldsymbol{x}) - \sum_{k=0}^{d} p_k(\boldsymbol{x}) q_k(\boldsymbol{z}), \tag{A.3}$$

where $r_{\boldsymbol{z}} = r_{(\boldsymbol{w},b)} = \rho(\boldsymbol{w}^{\mathsf{T}}(\cdot) - b)$ and $q_k(\boldsymbol{z}) \coloneqq \langle \phi_k, r_{\boldsymbol{z}} \rangle$, where $\boldsymbol{z} = (\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$ and ρ denotes any Green function of D^2, the second derivative operator, e.g., $\rho = \max\{0, \cdot\}$ (the ReLU) or $\rho = |\cdot|/2$.

The operator $\mathrm{R}_{\boldsymbol{\phi}}^{-1}$ defined in (A.2) has several useful properties (see [35, Theorem 22, items 1 and 2]). In particular, it is a stable (i.e., bounded) right-inverse of R, and when restricted to

$$\mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\phi}}(\mathbb{R}^d) \coloneqq \{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d) : \boldsymbol{\phi}(f) = \boldsymbol{0}\},$$

it is the bona fide inverse of R. The space $\mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\phi}}(\mathbb{R}^d)$ 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 $\mathrm{R} : \mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\phi}}(\mathbb{R}^d) \to \mathcal{M}(\mathbb{P}^d)$ is an isometric isomorphism with the inverse given by $\mathrm{R}_{\boldsymbol{\phi}}^{-1}$. Additionally we have from Proposition A.4 that $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d) \cong \mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\phi}}(\mathbb{R}^d) \oplus \mathcal{P}_1(\mathbb{R}^d)$, where $\mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\phi}}(\mathbb{R}^d)$ 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 2.5.

Proof of Lemma 2.5, item 2. Let $(\boldsymbol{\phi}, \boldsymbol{p})$ be the biorthogonal system constructed in the proof of Lemma 2.5, item 1. Since $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d) \cong \mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\phi}}(\mathbb{R}^d) \oplus \mathcal{P}_1(\mathbb{R}^d)$, 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 $\mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\phi}}(\mathbb{R}^d)$ 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 $\mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\phi}}(\mathbb{R}^d)$. Let $\mathcal{X}$ be the predual of $\mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\phi}}(\mathbb{R}^d)$, i.e., $\mathcal{X}' = \mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\phi}}(\mathbb{R}^d)$. We must show that $\delta(\cdot - \boldsymbol{x}_0) \in \mathcal{X}$, $\boldsymbol{x}_0 \in \mathbb{R}^d$. The Riesz–Markov–Kakutani representation theorem

says that the predual of $\mathcal{M}(\mathbb{P}^d)$ is $C_0(\mathbb{P}^d)$. The following diagram shows how these spaces are related.

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

The above diagram shows that $\delta(\cdot - \boldsymbol{x}_0) \in \mathcal{X}$ if and only if $\mathrm{R}^{-1*}_{\boldsymbol{\phi}}\{\delta(\cdot - \boldsymbol{x}_0)\} \in C_0(\mathbb{P}^d)$. From Proposition A.4 we see that $\mathrm{R}^{-1*}_{\boldsymbol{\phi}}\{\delta(\cdot - \boldsymbol{x}_0)\} = g_{\boldsymbol{\phi}}(\boldsymbol{x}_0, \cdot)$ defined in (A.3). By choosing $\rho = |\cdot|/2$ in (A.3) we have

$$\begin{aligned} g_{\boldsymbol{\phi}}(\boldsymbol{x}_0, (\boldsymbol{w}, b)) &= \frac{|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_0 - b|}{2} - \sum_{k=0}^{d} p_k(\boldsymbol{x}_0) \left\langle \phi_k, \frac{|\boldsymbol{w}^{\mathsf{T}}(\cdot) - b|}{2} \right\rangle \\ &\overset{(*)}{=} \frac{|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_0 - b|}{2} - \left[\frac{|-b|}{2} + \sum_{k=1}^{d} x_{0,k} \left(\frac{|w_k - b|}{2} - \frac{|-b|}{2} \right) \right] \\ &= \frac{|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_0 - b|}{2} - \frac{|b|}{2}\left(1 - \sum_{k=1}^{d} x_{0,k}\right) - \sum_{k=1}^{d} x_{0,k} \frac{|w_k - b|}{2}, \end{aligned} \tag{A.4}$$

where $(*)$ follows by substituting in the biorthogonal system $(\boldsymbol{\phi}, \boldsymbol{p})$ constructed in the proof of Lemma 2.5, 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

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

and for sufficiently small b we have

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

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 $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$. ∎

Appendix B. Proof of Theorem 2.7. In order to prove Theorem 2.7, we require that solutions to the variational problem in Theorem 2.7 exist. We will use the following recent result regarding existence of solutions to variational problems over Banach spaces.

Proposition B.1 (special case of [46, Theorem 2]). *Let $(\mathcal{X}, \mathcal{X}')$ be a dual pair of Banach spaces and $\{\nu_n\}_{n=1}^N \subset \mathcal{X}$ be a collection of linearly independent measurement functionals. Then, the solution set to*

$$\operatorname*{arg\,min}_{f \in \mathcal{X}'} \|f\|_{\mathcal{X}'} \quad s.t. \quad \langle \nu_n, f \rangle = y_n,\ n = 1, \dots, N,$$

is nonempty, convex, and weak compact, where $\langle \cdot, \cdot \rangle$ denotes the pairing of $\mathcal{X}'$ and its continuous dual, $\mathcal{X}''$.*[4]

Remark B.2. The result of [46, Theorem 2] is more general than what is stated in Proposition B.1, but we are only interested in the existence result in this paper.

Proof of Theorem 2.7. By Lemma 2.5, we have that $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ is a Banach space and that the functionals $\nu_n := \delta(\cdot - \boldsymbol{x}_n)$, $n = 1, \dots, N$, are weak* continuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ (and are therefore contained in the predual of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$). Moreover, this choice of $\{\nu_n\}_{n=1}^N$ is clearly linearly independent.[5] Therefore, the problem in (2.8) satisfies the hypotheses of Proposition B.1, and so a solution to (2.8) exists. Let $\widetilde{s}$ be a (not necessarily unique) solution to (2.8). This solution must be a minimizer of

$$\min_{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} \mathscr{R}\,\mathrm{TV}^2(f) \quad \text{s.t.} \quad \begin{cases} f(\boldsymbol{x}_n) = y_n, & n = 1, \dots, N, \\ f(\boldsymbol{0}) = \widetilde{s}(\boldsymbol{0}), & \\ f(\boldsymbol{e}_k) = \widetilde{s}(\boldsymbol{e}_k), & k = 1, \dots, d. \end{cases}$$

By Proposition 2.1, there exists a solution to the above display that takes the form in (2.9) with $K \leq N$ neurons, so we can always find a solution to the original problem in (2.8) of the form in (2.9). ∎

Appendix C. Proof of Theorem 2.12.

Proof. By Lemma 2.9, we have that $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d; \mathbb{R}^D)$ is a Banach space and that the point evaluation operator is componentwise weak* continuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d; \mathbb{R}^d)$. Therefore, the functionals

$$\langle \nu_{n,m}, f \rangle = f_m(\boldsymbol{x}_n),\ n = 1, \dots, N,\ m = 1, \dots, D,$$

where $f = (f_1, \dots, f_D) \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d; \mathbb{R}^D)$ and $\langle \cdot, \cdot \rangle$ denotes the pairing of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d; \mathbb{R}^D)$ and its continuous dual, are contained in the predual of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d; \mathbb{R}^D)$. Moreover, these functionals are linearly independent.[6] Therefore, the problem in (2.11) satisfies the hypotheses of Proposition B.1, and so a solution to (2.11) exists. Next, note that we can rewrite the problem in (2.11) as

$$\min_{\substack{f = (f_1, \dots, f_D) \\ f_m \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d) \\ m = 1, \dots, D}} \sum_{m=1}^{D} \|f_m\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} \quad \text{s.t.} \quad f_m(\boldsymbol{x}_n) = y_{n,m},\ \begin{cases} n = 1, \dots, N, \\ m = 1, \dots, D, \end{cases}$$

[4] Note that $\nu_n \in \mathcal{X}$ implies $\nu_n \in \mathcal{X}''$ by the canonical embedding of a Banach space in its bidual.
[5] Assuming that $\boldsymbol{x}_n \neq \boldsymbol{x}_k$ for $n \neq k$.
[6] Assuming that $\boldsymbol{x}_n \neq \boldsymbol{x}_k$ for $n \neq k$.

where $\boldsymbol{y}_n = (y_{n,1}, \dots, y_{n,D}) \in \mathbb{R}^D$. Let $\widetilde{s} = (\widetilde{s}_1, \dots, \widetilde{s}_D)$ be a (not necessarily unique) solution to (2.11). From the above display we see that this solution must satisfy

$$\widetilde{s}_m \in \underset{f \in \mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)}{\arg\min} \|f\|_{\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)} \quad \text{s.t.} \quad f(\boldsymbol{x}_n) = y_{n,m},\ n = 1, \dots, N, \tag{C.1}$$

for $m = 1, \dots, D$. To see this, note that if the above display did not hold, it would contradict the optimality of $\widetilde{s}$. By Theorem 2.7, there exists a solution to the above display that takes the form in (2.9) with $K_m \leq N$ neurons. By combining these solutions into a single vector-valued function with potential combining of neurons[7] we see that there exists a solution to the original problem in (2.11) that takes the form in (2.12) with $K \leq K_1 + \cdots + K_D \leq ND$ neurons. If no neurons combine, each $\boldsymbol{v}_k$ is 1-sparse. ∎

Remark C.1. One could also write a solution of (2.11) such that each output is completely independent of any other output; i.e., the outputs are completely decoupled. This corresponds to fitting the data with D separate single-hidden-layer ReLU networks. This follows from the fact that s_m is a minimizer to the problem in (C.1). This corresponds to the representation in (2.12) having each $\boldsymbol{v}_k$ be 1-sparse.

Appendix D. Proof of Lemma 2.11. Before proving Lemma 2.11, we will first bound the Lipschitz constant of functions in $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$. To do this, we will rely on Proposition A.4 with the biorthogonal system constructed in the proof of Lemma 2.5 given in Appendix A. In particular, Proposition A.4 provides the direct-sum decomposition of $f \in \mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ by

$$f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1} \times \mathbb{R}} g_{\boldsymbol{\phi}}(\boldsymbol{x}, (\boldsymbol{w}, b)) u(\boldsymbol{w}, b) \, \mathrm{d}\sigma(\boldsymbol{w}) \, \mathrm{d}b + \boldsymbol{c}^{\mathsf{T}} \boldsymbol{x} + c_0, \tag{D.1}$$

with $g_{\boldsymbol{\phi}}$ as in (A.4). It can easily be checked that this decomposition has the property that

$$\|f\|_{\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)} = \|u\|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})} + \|\boldsymbol{c}\|_1 + |c_0|, \tag{D.2}$$

and we refer the reader to [35, Theorem 22, item 3] for more details.

Lemma D.1. *Let $f \in \mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$. Then, f is Lipschitz continuous and satisfies the Lipschitz bound*

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq \|f\|_{\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)} \|\boldsymbol{x} - \boldsymbol{y}\|_1.$$

Proof. We will first bound the Lipschitz constant of $g_{\boldsymbol{\phi}}(\cdot, \boldsymbol{z})$ defined in (A.4), where $\boldsymbol{z} = (\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\begin{aligned} |g_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{z}) - g_{\boldsymbol{\phi}}(\boldsymbol{y}, \boldsymbol{z})| = \Bigg| & \frac{|\boldsymbol{w}^{\mathsf{T}} \boldsymbol{x} - b|}{2} - \frac{|\boldsymbol{w}^{\mathsf{T}} \boldsymbol{y} - b|}{2} \\ & - \frac{|b|}{2} \left[\left(1 - \sum_{k=1}^{d} x_k \right) - \left(1 - \sum_{k=1}^{d} y_k \right) \right] - \sum_{k=1}^{d} (x_k - y_k) \frac{|w_k - b|}{2} \Bigg| \end{aligned}$$

[7] This would happen in the event that $\widetilde{s}_m$ and $\widetilde{s}_\ell$, $m \neq \ell$, shared a common neuron.

$$
\begin{aligned}
&\leq \frac{\big|\,|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}-b|-|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{y}-b|\,\big|}{2} \\
&\quad + \left|\sum_{k=1}^{d}(x_k-y_k)\frac{|b|}{2}-\sum_{k=1}^{d}(x_k-y_k)\frac{|w_k-b|}{2}\right| \\
&\leq \frac{\big|\,|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}-b|-|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{y}-b|\,\big|}{2}+\sum_{k=1}^{d}|x_k-y_k|\frac{\big|\,|b|-|w_k-b|\,\big|}{2} \\
&\overset{(*)}{\leq} \frac{|\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}-\boldsymbol{w}^{\mathsf{T}}\boldsymbol{y}|}{2}+\sum_{k=1}^{d}|x_k-y_k|\frac{|w_k|}{2} \\
&\overset{(\S)}{\leq} \frac{\|\boldsymbol{w}\|_\infty\|\boldsymbol{x}-\boldsymbol{y}\|_1+\|\boldsymbol{w}\|_\infty\|\boldsymbol{x}-\boldsymbol{y}\|_1}{2} \\
&\overset{(\dagger)}{\leq} \|\boldsymbol{x}-\boldsymbol{y}\|_1,
\end{aligned}
$$

where $(*)$ holds from the reverse triangle inequality, $(\S)$ holds from Hölder's inequality, and $(\dagger)$ holds from the fact that $\|\cdot\|_\infty \leq \|\cdot\|_2$ in finite-dimensional spaces combined with $\|\boldsymbol{w}\|_2 = 1$.

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

$$
\begin{aligned}
|f(\boldsymbol{x})-f(\boldsymbol{y})| &\leq \int_{\mathbb{S}^{d-1}\times\mathbb{R}} |g(\boldsymbol{x},(\boldsymbol{w},b))-g(\boldsymbol{y},(\boldsymbol{w},b))||u(\boldsymbol{w},b)|\,\mathrm{d}\sigma(\boldsymbol{w})\,\mathrm{d}b + \left|\boldsymbol{c}^{\mathsf{T}}(\boldsymbol{x}-\boldsymbol{y})\right| \\
&\leq \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \|\boldsymbol{x}-\boldsymbol{y}\|_1|u(\boldsymbol{w},b)|\,\mathrm{d}\sigma(\boldsymbol{w})\,\mathrm{d}b + \|\boldsymbol{c}\|_\infty\|\boldsymbol{x}-\boldsymbol{y}\|_1 \\
&\leq \|u\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})}\|\boldsymbol{x}-\boldsymbol{y}\|_1 + \|\boldsymbol{c}\|_1\|\boldsymbol{x}-\boldsymbol{y}\|_1 \\
&\leq \|f\|_{\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)}\,\|\boldsymbol{x}-\boldsymbol{y}\|_1,
\end{aligned}
$$

where the third line follows from the fact that $\|\cdot\|_\infty \leq \|\cdot\|_1$ in finite-dimensional spaces and the fourth line follows from (D.2). ∎

Proof of Lemma 2.11. Write $f = (f_1, \ldots, f_D)$. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$
\begin{aligned}
\|f(\boldsymbol{x})-f(\boldsymbol{y})\|_1 &= \sum_{m=1}^{D}|f_m(\boldsymbol{x})-f_m(\boldsymbol{y})| \\
&\leq \left(\sum_{m=1}^{D}\|f_m\|_{\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)}\right)\|\boldsymbol{x}-\boldsymbol{y}\|_1 \\
&= \|f\|_{\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)}\|\boldsymbol{x}-\boldsymbol{y}\|_1,
\end{aligned}
$$

where the second line follows from Lemma D.1 and the third line follows from the definition of $\|\cdot\|_{\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)}$ in Lemma 2.9. ∎

Remark D.2. The Lipschitz bounds in Lemmas 2.11 and D.1 are by no means the tightest Lipschitz bounds.

REFERENCES

[1] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[2] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, *Understanding deep neural networks with rectified linear units*, in Proceedings of the 6th International Conference on Learning Representations, International Society of the Learning Sciences, 2018.

[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.

[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-Volume 2, MIT Press, 2014, pp. 2654–2662.

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

[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. (PMLR) 80, JMLR, Cambridge, MA, 2018, pp. 374–383.

[7] R. Balestriero and R. G. Baraniuk, *Mad max: Affine spline insights into deep learning*, Proc. IEEE, 109 (2020), pp. 704–727.

[8] 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, https://arxiv.org/abs/1902.00800.

[9] B. Bohn, C. Rieger, and M. Griebel, *A representer theorem for deep kernel learning*, J. Mach. Learn. Res., 20 (2019), pp. 1–32.

[10] 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.

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

[12] 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.

[13] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed., John Wiley and Sons, New York, 1999.

[14] J. Frankle and M. Carbin, *The lottery ticket hypothesis: Finding sparse, trainable neural networks*, in Proceedings of the 6th International Conference on Learning Representations, International Society of the Learning Sciences, 2018.

[15] A. Golubeva, B. Neyshabur, and G. Gur-Ari, *Are wider nets better given the same number of parameters?*, Proceedings of the 9th International Conference on Learning Representations, International Society of the Learning Sciences, 2021.

[16] 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, 2016, pp. 770–778.

[17] 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.

[18] 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, https://arxiv.org/abs/1207.0580.

[19] K. H. Jin, M. T. McCann, E. Froustey, and M. Unser, *Deep convolutional neural network for inverse problems in imaging*, IEEE Trans. Image Process., 26 (2017), pp. 4509–4522.

[20] F. John, *Plane Waves and Spherical Means: Applied to Partial Differential Equations*, Springer, New York, 2013.

[21] G. Kimeldorf and G. Wahba, *Some results on Tchebycheffian spline functions*, J. Math. Anal. Appl., 33 (1971), pp. 82–95.

[22] 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, Red Hook, NY, 2012, pp. 1097–1105.

[23] A. Krogh and J. A. Hertz, *A simple weight decay can improve generalization*, in Advances in Neural Information Processing Systems 4, Morgan Kaufmann, San Francisco, 1992, pp. 950–957.
[24] A. J. Kurdila and M. Zabarankin, *Convex Functional Analysis*, Systems Control Found. Appl., Birkhäuser, Basel, 2006.
[25] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning*, Nature, 521 (2015), pp. 436–444.
[26] E. Mammen and S. van de Geer, *Locally adaptive regression splines*, Ann. Statist., 25 (1997), pp. 387–413.
[27] C. A. Micchelli, *Interpolation of scattered data: Distance matrices and conditionally positive definite functions*, in Approximation Theory and Spline Functions, Springer, Dordrecht, the Netherlands, 1984, pp. 143–145.
[28] 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, Curran Associates, Red Hook, NY, 2014, pp. 2924–2932.
[29] B. Neyshabur, R. R. Salakhutdinov, and N. Srebro, *Path-SGD: Path-normalized optimization in deep neural networks*, in Advances in Neural Information Processing Systems 28, MIT Press, Cambridge, MA, 2015, pp. 2422–2430.
[30] B. Neyshabur, R. Tomioka, R. Salakhutdinov, and N. Srebro, *Geometry of Optimization and Implicit Regularization in Deep Learning*, preprint, arXiv:1705.03071, 2017, https://arxiv.org/abs/1705.03071.
[31] B. Neyshabur, R. Tomioka, and N. Srebro, *In search of the real inductive bias: On the role of implicit regularization in deep learning*, in Proceedings of the ICLR (Workshop), International Society of the Learning Sciences, 2015.
[32] B. Neyshabur, R. Tomioka, and N. Srebro, *Norm-based capacity control in neural networks*, in Proceedings of the Conference on Learning Theory, Proc. Mach. Learn. Res. (PMLR), JMLR, Cambridge, MA, 2015, pp. 1376–1401.
[33] 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 8th International Conference on Learning Representations, International Society of the Learning Sciences, 2020.
[34] R. Parhi and R. D. Nowak, *The role of neural network activation functions*, IEEE Signal Process. Lett., 27 (2020), pp. 1779–1783.
[35] 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.
[36] R. Parhi and R. D. Nowak, *Near-Minimax Optimal Estimation with Shallow ReLU Neural Networks*, preprint, arXiv:2109.08844, 2021, https://arxiv.org/abs/2109.08844.
[37] 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, 2015.
[38] M. Reed and B. Simon, *Methods of Modern Mathematical Physics: Functional Analysis*, Methods of Modern Mathematical Physics 1, Academic Press, New York, 1972.
[39] W. Rudin, *Functional Analysis*, 2nd ed., Internat. Ser. Pure Appl. Math., McGraw-Hill, New York, 1991.
[40] A. Sanyal, P. H. Torr, and P. K. Dokania, *Stable rank normalization for improved generalization in neural networks and GANs*, in Proceedings of the 7th International Conference on Learning Representations, International Society of the Learning Sciences, 2019.
[41] P. H. 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, JMLR, 2019, pp. 2667–2690.
[42] B. Schölkopf, R. Herbrich, and A. J. Smola, *A generalized representer theorem*, in International Conference on Computational Learning Theory, Springer, Berlin, 2001, pp. 416–426.
[43] G. S. Sidhu and H. L. Weinert, *Vector-valued Lg-splines* I. *Interpolating splines*, J. Math. Anal. Appl., 70 (1979), pp. 505–529.
[44] E. M. Stein and R. Shakarchi, *Fourier Analysis: An Introduction*, Princeton Lectures in Analysis 1, Princeton University Press, Princeton, NJ, 2011.
[45] M. Unser, *A representer theorem for deep neural networks*, J. Mach. Learn. Res., 20 (2019), pp. 1–30.
[46] M. Unser, *A unifying representer theorem for inverse problems and machine learning*, Found. Comput. Math., 21 (2020), pp. 941–960.

[47] M. Unser and S. Aziznejad, *Convex optimization in sums of Banach spaces*, Appl. Comput. Harmon. Anal., 56 (2022), pp. 1–25.
[48] 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.
[49] G. Wahba, *Spline Models for Observational Data*, CBMS-NSF Regional Conf. Ser. Math. 59, SIAM, Philadelphia, 1990.
[50] H. Wang, S. Agarwal, and D. Papailiopoulos, *Pufferfish: Communication-efficient models at no extra cost*, in Proceedings of Machine Learning and Systems 3, Systems and Machine Learning Foundation, Indio, CA, 2021.
[51] M. P. Wolff and H. H. Schaefer, *Topological Vector Spaces*, Grad. Texts in Math., Springer, New York, 2012.
[52] S. Zuhovickiĭ, *Remarks on problems in approximation theory*, Mat. Zbirnik KDU, 1948, pp. 169–183.