Near-Minimax Optimal Estimation With Shallow ReLU Neural Networks

Rahul Parhi, *Member, IEEE*, and Robert D. Nowak, *Fellow, IEEE*

***Abstract*— We study the problem of estimating an unknown function from noisy data using shallow ReLU neural networks. The estimators we study minimize the sum of squared data-fitting errors plus a regularization term proportional to the squared Euclidean norm of the network weights. This minimization corresponds to the common approach of training a neural network with weight decay. We quantify the performance (mean-squared error) of these neural network estimators when the data-generating function belongs to the second-order Radon-domain bounded variation space. This space of functions was recently proposed as the natural function space associated with shallow ReLU neural networks. We derive a minimax lower bound for the estimation problem for this function space and show that the neural network estimators are minimax optimal up to logarithmic factors. This minimax rate is immune to the curse of dimensionality. We quantify an explicit gap between neural networks and linear methods (which include kernel methods) by deriving a linear minimax lower bound for the estimation problem, showing that linear methods necessarily suffer the curse of dimensionality in this function space. As a result, this paper sheds light on the phenomenon that neural networks seem to break the curse of dimensionality.**

***Index Terms*— Neural networks, ridge functions, sparsity, function approximation, nonparametric function estimation.**

I. Introduction

THE fundamental building blocks of neural networks are *ridge functions*. A ridge function is a multivariate function mapping $\mathbb{R}^d \to \mathbb{R}$ of the form

$$\boldsymbol{x} \mapsto \rho(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^d,$$

where $\rho : \mathbb{R} \to \mathbb{R}$ is referred to as the *profile* of the ridge function and $\boldsymbol{w} \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$ is referred to as the *direction* of the ridge function.

This paper studies the problem estimating functions from noisy samples using shallow neural networks, which are superpositions of ridge functions, of the form

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} v_k \, \rho(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k), \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{1}$$

where the $\rho : \mathbb{R} \to \mathbb{R}$ is the *activation function*, K is the *width* of the neural network, and, for $k = 1, \dots, K$, $v_k \in \mathbb{R} \setminus \{0\}$ and $\boldsymbol{w}_k \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$ are the *weights* of the neural network and $b_k \in \mathbb{R}$ are the *biases* of the neural network. Throughout the paper, we will focus on the rectified linear unit (ReLU) activation function, $\rho(x) = \max\{0, x\}$, which is widely used in practice [28].

We consider the problem of nonparametric function estimation where the goal is to estimate an unknown function $f : \Omega \to \mathbb{R}$, where $\Omega \subset \mathbb{R}^d$ is a bounded domain, from the noisy samples

$$y_n = f(\boldsymbol{x}_n) + \varepsilon_n, \; n = 1, \dots, N, \tag{2}$$

where the noise $\{\varepsilon_n\}_{n=1}^N$ are i.i.d. Gaussian random variables and $\{\boldsymbol{x}_n\}_{n=1}^N \subset \Omega$ are the design points. We study the performance of neural network estimators of the form in (1) that minimize the objective of the sum of squared data-fitting errors plus a regularization term proportional to the squared Euclidean norm of the network weights. This minimization corresponds to the common approach of gradient-based training of a neural network with *weight decay* [26]. That is, training a neural network using gradient descent with weight decay is simply gradient descent applied to this objective.

In order to quantify the performance of such estimators, we consider cases in which f is an unknown function within a known function space. To this end, we will consider functions mapping $\Omega \to \mathbb{R}$ which belong to the Banach space of functions of second-order bounded variation in the Radon domain, denoted $\mathscr{R}\,\mathrm{BV}^2(\Omega)$. Our recent work in [35], [37] proposed this Banach space as the "natural" function space associated with shallow ReLU networks. This space contains several classical multivariate function spaces including certain Sobolev spaces as well as certain *spectral Barron spaces*, pioneered in the seminal work of Barron on approximation and estimation using shallow sigmoidal networks [2].

It was first observed in [2] that neural network estimators can be *immune to the curse of dimensionality*. This paper sheds light on this phenomenon. $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ contains classical multivariate function spaces including the L^1- and L^2-Sobolev spaces of order $d+1$, where d is the ambient dimension of

Manuscript received 20 September 2021; revised 26 May 2022; accepted 12 September 2022. Date of publication 22 September 2022; date of current version 20 January 2023. This work was supported in part by NSF under Grant DMS-2134140, in part by the Office of Naval Research Multidisciplinary University Research Initiative (ONR MURI) under Grant N00014-20-1-2787, in part by the Air Force Office of Scientific Research/Air Force Research Laboratory (AFOSR/AFRL) under Grant FA9550-18-1-0166, and in part by NSF Graduate Research Fellowship Program under Grant DGE-1747503. *(Corresponding author: Rahul Parhi.)*

Rahul Parhi was with the Department of Electrical and Computer Engineering, University of Wisconsin–Madison, Madison, WI 53706 USA. He is now with the Biomedical Imaging Group, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland (e-mail: rahul.parhi@epfl.ch).

Robert D. Nowak is with the Department of Electrical and Computer Engineering, University of Wisconsin–Madison, Madison, WI 53706 USA (e-mail: rdnowak@wisc.edu).

Communicated by S. Jalali, Associate Editor for Machine Learning and Statistics.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TIT.2022.3208653.

Digital Object Identifier 10.1109/TIT.2022.3208653

the domain $\Omega \subset \mathbb{R}^d$. It is classically known that this sort of Sobolev-regularity is sufficient to overcome the curse of dimensionality. On the other hand, $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ also contains functions that are much less regular. In particular, functions with significant variation and irregularity, but only in a few directions, also belong to $\mathscr{R}\,\mathrm{BV}^2(\Omega)$. For example, any ridge function with a profile that has just its first two weak derivatives in $L^2(\Omega)$ is included in $\mathscr{R}\,\mathrm{BV}^2(\Omega)$. This shows that $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ may be regarded as a *mixed variation* space [12], since it contains functions that are more regular in some directions and less in others. This makes $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ a compelling framework for high-dimensional estimation. Moreover, the neural network estimators we study are *locally adaptive* to such mixed variation.

Our past work [35], [37] derives a *neural network representer theorem* which proves that shallow ReLU networks are solutions to data-fitting problems in $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$, the space of functions defined on $\mathbb{R}^d$ of second-order bounded variation in the Radon domain. Remarkably, this variational problem can be recast as a finite-dimensional neural network training problem where the regularization corresponds to training a shallow ReLU network with weight decay. This is the reason we view these spaces as the natural function space of shallow ReLU networks. This connection is reminiscent of the classical reproducing kernel Hilbert space (RKHS) representer theorem which says that kernel machines are solutions to data-fitting variational problems over the associated RKHS, although the neural network variational problem is posed over a (non-Hilbertian) Banach space.

We summarize the contributions of this paper below.

1) We first discuss how to define $\mathscr{R}\,\mathrm{BV}^2(\Omega)$, where $\Omega \subset \mathbb{R}^d$ is a *bounded domain*, while preserving a representer theorem for shallow ReLU networks. This implies that data-fitting with functions in $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ can be recast as a finite-dimensional neural network training problem that may be solved using gradient-descent with weight decay. This result sets the stage for discussing approximation and estimation error for functions in $\mathscr{R}\,\mathrm{BV}^2(\Omega)$.
2) We relate $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ spaces to previously studied function spaces related to shallow neural networks. In particular, we show that $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ is exactly the same (in the sense of equivalent Banach spaces) as the so-called *variation space* associated to shallow ReLU networks that has been studied by a number of authors [1], [27], [31], [43]. This provides a novel analytic characterization of this space. Using this characterization, we can apply previously derived optimal approximation rates for functions from the variation space [1], [44] to characterize the optimal approximation rates for functions in $\mathscr{R}\,\mathrm{BV}^2(\Omega)$. The approximation rate (with respect to the $L^\infty(\Omega)$-norm) is $K^{-\frac{d+3}{2d}}$, where K is the number of neurons in the approximant. Remarkably, this rate is *immune to the curse of dimensionality*, as it tends to $K^{-1/2}$ as $d \to \infty$. We also show that $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ is *larger* than the second-order spectral Barron space.
3) We show that a shallow ReLU network that minimizes the sum of squared data-fitting errors plus a regularization term proportional to the sum of squared weights (i.e., training a shallow ReLU network with weight decay to a global minimizer) is a minimax optimal (up to logarithmic factors) estimator when the data are generated according to (2), where $f \in \mathscr{R}\,\mathrm{BV}^2(\Omega)$. The minimax rate of the mean-squared error is, up to logarithmic factors, $N^{-\frac{d+3}{2d+3}}$. Remarkably, this rate is *immune to the curse of dimensionality*, as it tends to $N^{-1/2}$ as $d \to \infty$.
4) Using the results of this paper, we show that there is a fundamental gap between neural networks and more classical linear methods (which include kernel methods). In particular, we use ridgelet analysis to derive a minimax lower bound for the estimation problem when restricted to linear estimators. We find that the linear minimax lower bound is $N^{-\frac{3}{d+3}}$, which suffers the curse of dimensionality as $d \to \infty$. This result says that linear methods are suboptimal for estimating functions in $\mathscr{R}\,\mathrm{BV}^2(\Omega)$. We also show this gap qualitatively via numerical experiments.

A. Related Work

There is a large body of work regarding the problem of statistical estimation with ridge functions, under many different names, including projection pursuit regression [17], ridgelet shrinkage [8], and, of course, estimation with neural networks [2]. The last few years have led to a number of related works that consider the problem of minimax estimation with neural networks [18], [20], [24], [41], [47]. These works fall into two categories: 1) they consider the problem of estimating a function that is *explicitly synthesized* from a dictionary of neurons; 2) they consider the problem of estimating a function from a particular (classical) space of functions (e.g., Hölder, Sobolev, Besov, etc.). Moreover, the procedures for actually constructing the estimators in these works usually involve greedy algorithms and do not correspond to how neural networks are actually trained in practice. The work of this paper is different from these past works in that we consider the problem of estimating functions from a new, not classical, function space, $\mathscr{R}\,\mathrm{BV}^2(\Omega)$, and study the performance of estimators that correspond to solutions to problem of training shallow ReLU networks with weight decay, a common regularization scheme used when training neural networks in practice.

B. Roadmap

In Section II we introduce notation used in the remainder of the paper. In Section III we introduce relevant results from our previous work [35], [37]. In Section IV we discuss how to define $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ where $\Omega \subset \mathbb{R}^d$ is a bounded domain and derive a new representer theorem for shallow ReLU networks by considering variational problems over $\mathscr{R}\,\mathrm{BV}^2(\Omega)$. In Section V we relate $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ to previously studied function spaces associated to shallow networks. In Section VI we derive optimal approximation rates for functions in $\mathscr{R}\,\mathrm{BV}^2(\Omega)$, where the approximants are shallow ReLU networks. In Section VII we show that shallow ReLU network estimators are minimax optimal (up to logarithmic factors) for estimating functions in $\mathscr{R}\,\mathrm{BV}^2(\Omega)$. In Section VIII we show that there is a fundamental gap between neural networks and linear methods (including kernel methods).

II. Preliminaries and Notation

Let $L^p(\Omega)$ denote the usual Lebesgue space, where Ω is a domain (either bounded or unbounded). This space is a Banach space when equipped with the norm

$$\|f\|_{L^p(\Omega)} := \left(\int_\Omega |f(\boldsymbol{x})|^p \,\mathrm{d}\boldsymbol{x}\right)^{1/p}, \quad 1 \le p < \infty,$$
$$\|f\|_{L^\infty(\Omega)} := \operatorname*{ess\,sup}_{\boldsymbol{x}\in\Omega} |f(\boldsymbol{x})|, \quad p = \infty.$$

When we do not specify the underlying measure, it will correspond to the Haar measure of Ω (e.g., Lebesgue measure when $\Omega = \mathbb{R}^d$ or the surface measure when $\Omega = \mathbb{S}^{d-1}$, the surface of the Euclidean sphere in $\mathbb{R}^d$). When we do specify a particular measure, say μ, we will write $L^p(\Omega; \mu)$.

We will also work with the Banach space of finite Radon measures on Ω, denoted $\mathcal{M}(\Omega)$. The norm $\|\cdot\|_{\mathcal{M}(\Omega)}$ is exactly the *total variation norm* (in the sense of measures). We can view this space as a subspace of distributions (generalized functions) on Ω. The space $\mathcal{M}(\Omega)$ may be regarded as a "generalization" of $L^1(\Omega)$ in the sense that if $f \in L^1(\Omega)$, $\|f\|_{L^1(\Omega)} = \|f\|_{\mathcal{M}(\Omega)}$, but $\mathcal{M}(\Omega)$ is a strictly larger space that also contains the shifted Dirac impulses $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \Omega$, such that $\|\delta(\cdot - \boldsymbol{x}_0)\|_{\mathcal{M}(\Omega)} = 1$. We also remark that the $\mathcal{M}$-norm is the continuous-domain analogue of the ℓ^1-norm. We refer the reader to [15, Chapter 7] for more details about this space.

We will also use the notation $a_N \lesssim b_N$ to mean there exists a constant C (independent of N) such that $a_N \le C\, b_N$, $a_N \gtrsim b_N$ to mean $b_N \lesssim a_N$, and $a_N \asymp b_N$ to mean $a_N \lesssim b_N$ and $a_N \gtrsim b_N$. We will also subscript $\lesssim$, $\gtrsim$, and $\asymp$ with any parameters that the implicit constant depends on.

III. Shallow Neural Networks, Splines, and Variational Methods

In this section we will discuss relevant results from our prior work in [35] and [37], making connections between shallow neural networks, splines, and variational methods. 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

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

where s denotes the $(d-1)$-dimensional 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

$$\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d) = \{f \in L^{\infty,1}(\mathbb{R}^d) : \ \mathscr{R}\,\mathrm{TV}^2(f) < \infty\}, \tag{3}$$

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

$$\mathscr{R}\,\mathrm{TV}^2(f) = c_d \left\|\partial_t^2 \Lambda^{d-1} \mathscr{R} f\right\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} \tag{4}$$

denotes the second-order total variation of a function in the offset variable of the (filtered) 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 (4) must be understood in the distributional sense. We refer the reader to [35, Section 3] for more details.

The $\mathscr{R}\,\mathrm{TV}^2$-seminorm was first proposed in [33] and studied in extensive detail in [35] and [37]. When equipped with the norm

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

where $\{\boldsymbol{e}_k\}_{k=1}^d$ denotes the canonical basis of $\mathbb{R}^d$, $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ is a Banach space [37, Lemma 2.4]. In particular, it is a Banach space with a sparsity-promoting norm as $\mathscr{R}\,\mathrm{TV}^2(\cdot)$ is defined via an $\mathcal{M}$-norm. The terms $|f(\mathbf{0})| + \sum_{k=1}^d |f(\boldsymbol{e}_k) - f(\mathbf{0})|$ that appear in the above display impose a norm on the null space of $\mathscr{R}\,\mathrm{TV}^2(\cdot)$, which corresponds to affine functions on $\mathbb{R}^d$, and is an upper bound on the Lipschitz constant of the affine portion of f.

Intuitively, the $\mathscr{R}\,\mathrm{TV}^2$-seminorm measures sparsity of second derivatives in the Radon domain. The Radon transform naturally arises when working with ridge functions. In particular, the second derivative of the (filtered) Radon transform of a ReLU ridge function is essentially a Dirac impulse located at the weight and bias of the ReLU ridge function [35, Lemma 17]. This arises due to the fact that in the univariate case, the second derivative of the ReLU is a Dirac impulse. Thus, the seminorm in (4) favors ReLU ridge functions and so functions in $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ with small $\mathscr{R}\,\mathrm{TV}^2$-seminorm will typically take the form of a sparse superposition of ReLU ridge functions. We now state the main result of [35].

Proposition 1 (Special Case of [35, Theorem 1]): Let $\ell(\cdot,\cdot) : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ be a strictly convex, coercive, and lower-semicontinuous in its second argument loss function and let $\lambda > 0$ be an adjustable regularization parameter. Then, for any data $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N \subset \mathbb{R}^d \times \mathbb{R}$, there exists a solution to the variational problem

$$\min_{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} \sum_{n=1}^N \ell(y_n, f(\boldsymbol{x}_n)) + \lambda\, \mathscr{R}\,\mathrm{TV}^2(f) \tag{5}$$

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

that takes the form of a shallow ReLU network plus an affine function. In particular, it takes the form

$$s(\boldsymbol{x}) = \sum_{k=1}^{K} v_k\, \rho(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k) + \boldsymbol{c}^\mathsf{T}\boldsymbol{x} + c_0, \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{6}$$

where $K \leq N - (d+1)$, ρ is the ReLU, $\boldsymbol{w}_k \in \mathbb{S}^{d-1}$, $v_k \in \mathbb{R} \setminus \{0\}$, $b_k \in \mathbb{R}$, $\boldsymbol{c} \in \mathbb{R}^d$, and $c_0 \in \mathbb{R}$.

We remark that the affine function that appears in (6) is known as a *skip connection* in neural network parlance [19]. In other words, (6) is a shallow ReLU network with a skip connection.

A. Shallow Neural Networks and Splines

When $d = 1$, the space $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ is the classical second-order bounded variation space

$$\mathrm{BV}^2(\mathbb{R}) \coloneqq \{f : \mathbb{R} \to \mathbb{R} :\ \mathrm{TV}^2(f) < \infty\},$$

where

$$\mathrm{TV}^2(f) \coloneqq \left\|\mathrm{D}^2 f\right\|_{\mathcal{M}(\mathbb{R})}$$

is the second-order total variation of a function $f : \mathbb{R} \to \mathbb{R}$, where D is the (distributional) derivative operator [35, Section 5.1]. In this case, the result of Proposition 1 recovers the classical representer theorem for locally adaptive linear splines, which dates back to the 1970s [14], [29], [48]. Moreover, we also have that $\mathscr{R}\mathrm{TV}^2(f) = \mathrm{TV}^2(f)$ [35, Section 5.1].

B. Connections to Neural Network Training

We view $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ as the natural function space associated with shallow ReLU networks since the problem in (5) can be recast as a finite-dimensional neural network training problem that corresponds to training a sufficiently wide shallow ReLU network (with a skip connection) with weight decay or with the so-called "path-norm" regularizer. In particular, consider the shallow ReLU network with a skip connection:

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

where $\boldsymbol{\theta}$ denotes the parameters of the neural network, i.e., $\{v_k, \boldsymbol{w}_k, b_k\}_{k=1}^K$, $\boldsymbol{c}$ and c_0. Then, it was shown in [35, Theorem 8] that, the solutions to either of the following (equivalent) finite-dimensional neural network training problems

$$\min_{\boldsymbol{\theta}\in\Theta} \sum_{n=1}^{N} \ell(y_n, f_{\boldsymbol{\theta}}(\boldsymbol{x}_n)) + \frac{\lambda}{2}\sum_{k=1}^{K} |v_k|^2 + \|\boldsymbol{w}_k\|_2^2 \tag{7}$$

$$\min_{\boldsymbol{\theta}\in\Theta} \sum_{n=1}^{N} \ell(y_n, f_{\boldsymbol{\theta}}(\boldsymbol{x}_n)) + \lambda\sum_{k=1}^{K} |v_k|\|\boldsymbol{w}_k\|_2 \tag{8}$$

where $\Theta = \mathbb{R}^M$ is the parameter space and M is the total number of scalar parameters of network, are solutions to the variational problem in (5), so long as $K \geq N - (d+1)$. The problem in (7) corresponds to training a shallow ReLU network with weight decay [26] and the problem in (8) corresponds to training a neural network with path-norm regularization [32]. Therefore, the above says that trained[2] shallow ReLU networks are "optimal" with respect to the space $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$. This result follows from the fact that

$$\mathscr{R}\mathrm{TV}^2(f_{\boldsymbol{\theta}}) = \sum_{k=1}^{K} |v_k|\|\boldsymbol{w}_k\|_2, \tag{9}$$

which can be viewed as a kind of ℓ^1-norm, giving insight into the sparsity promoting nature of the $\mathscr{R}\mathrm{TV}^2$-seminorm on neural network parameters.[3] Moreover, this result also gives insight into the sparsity-promoting nature of training a shallow ReLU network with weight decay. We refer the reader to [35] for more details about recasting the problem in (5) as the problems in (7) and (8), the equivalence of (7) and (8), and the derivation of the equality in (9).

This result also says, in the univariate case, that the function learned by training a sufficiently wide ReLU network with weight decay or with path-norm regularization on data is a locally adaptive linear spline [34], [40].

IV. The $\mathscr{R}\mathrm{BV}^2$-Space on a Bounded Domain

In approximation theory and nonparametric function estimation it is common to quantify error with respect to the $L^p(\Omega)$-norm, $1 \leq p \leq \infty$, where $\Omega \subset \mathbb{R}^d$ is a bounded domain. Therefore, we are interested in working with the $\mathscr{R}\mathrm{BV}^2$-space defined on a *bounded domain*. In this section we will define the $\mathscr{R}\mathrm{BV}^2$-space on a bounded domain while still maintaining a similar representer theorem as in $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$.

We can define the $\mathscr{R}\mathrm{BV}^2$-space on a bounded domain $\Omega \subset \mathbb{R}^d$ using the standard approach of considering restrictions of functions in $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$. This provides the following definition:

$$\mathscr{R}\mathrm{BV}^2(\Omega) \coloneqq \{f \in \mathscr{D}'(\Omega) :\ \exists g \in \mathscr{R}\mathrm{BV}^2(\mathbb{R}^d) \text{ s.t. } g|_\Omega = f\},$$

where $\mathscr{D}'(\Omega)$ denotes the space of distributions (generalized functions) on Ω. Similarly, we can define the second-order total variation in the Radon domain of a function f defined on a bounded domain $\Omega \subset \mathbb{R}^d$:

$$\mathscr{R}\mathrm{TV}^2_\Omega(f) \coloneqq \inf_{g\in\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)} \mathscr{R}\mathrm{TV}^2(g) \quad \text{s.t.} \quad g|_\Omega = f. \tag{10}$$

This gives an alternative characterization of $\mathscr{R}\mathrm{BV}^2(\Omega)$ as

$$\mathscr{R}\mathrm{BV}^2(\Omega) = \{f \in \mathscr{D}'(\Omega) :\ \mathscr{R}\mathrm{TV}^2_\Omega(f) < \infty\}.$$

We also remark that since $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ is a Banach space, $\mathscr{R}\mathrm{BV}^2(\Omega)$ is also a Banach space. In particular, it is a Banach space when equipped with the norm

$$\|f\|_{\mathscr{R}\mathrm{BV}^2(\Omega)} \coloneqq \inf_{g\in\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)} \|g\|_{\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)} \quad \text{s.t.} \quad g|_\Omega = f.$$

[2] Assuming that the network is trained to a global minimizer.

[3] The equality in (9) assumes that the neural network is written in reduced form, i.e., the weight bias pairs $(\boldsymbol{w}_k, b_k)$ $k = 1, \ldots, K$ are unique up to certain symmetries. See [35] for more details.

A. Extensions From $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ to $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$

In this section we will discuss how to identify functions in $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ with functions in $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$, where $\Omega \subset \mathbb{R}^d$ is a bounded domain.

Lemma 2: Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. Given $f \in \mathscr{R}\,\mathrm{BV}^2(\Omega)$, there exists an extension $f_{\mathsf{ext}} \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ that admits an integral representation

$$f_{\mathsf{ext}}(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \rho(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)\,\mathrm{d}\mu(\boldsymbol{w}, b) + \boldsymbol{c}^\mathsf{T}\boldsymbol{x} + c_0,$$

such that $\operatorname{supp}\mu \subset Z_\Omega$, where Z_Ω is the set

$$\overline{\{\boldsymbol{z} = (\boldsymbol{w}, b) \in \mathbb{S}^{d-1}\times\mathbb{R} :\ \{\boldsymbol{x} :\ \boldsymbol{w}^\mathsf{T}\boldsymbol{x} = b\} \cap \Omega \neq \varnothing\}}, \tag{11}$$

where $\overline{A}$ denotes the closure of the set A. This extension has the property that $f_{\mathsf{ext}}|_\Omega = f$ and

$$\mathscr{R}\,\mathrm{TV}^2_\Omega(f) = \mathscr{R}\,\mathrm{TV}^2(f_{\mathsf{ext}}) = \|\mu\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} = \|\,\mu_{Z_\Omega}\|_{\mathcal{M}(Z_\Omega)}.$$

The set Z_Ω simply excludes ReLU functions that are linear functions (no activation threshold) when restricted to Ω. The proof of Lemma 2 relies on several properties of the space $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ from our previous work in [35]. We introduce the relevant background and then prove Lemma 2 in Appendix A.

Remark 3: When

$$\Omega = \mathbb{B}^d_1 \coloneqq \{\boldsymbol{x} \in \mathbb{R}^d :\ \|\boldsymbol{x}\|_2 \leq 1\}, \tag{12}$$

the Euclidean unit ball in $\mathbb{R}^d$, we have that Z_Ω from (11) is exactly

$$Z_\Omega = \mathbb{S}^{d-1} \times [-1, 1].$$

Therefore, from Lemma 2, we can identify functions in $f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}^d_1)$ with integral representations of the form

$$f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} \rho(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)\,\mathrm{d}\mu(\boldsymbol{w}, b) + \boldsymbol{c}^\mathsf{T}\boldsymbol{x} + c_0,$$

where $\boldsymbol{x} \in \mathbb{B}^d_1$.

Remark 4: Similar to the discussion in Section III-A, when $d = 1$, the space $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}^d_1)$ is exactly the classical second-order bounded variation spaces defined on $[-1, 1]$:

$$\mathrm{BV}^2[-1, 1] \coloneqq \{f : [-1, 1] \to \mathbb{R} :\ \mathrm{TV}^2_{[-1,1]}(f) < \infty\},$$

where

$$\mathrm{TV}^2_{[-1,1]}(f) \coloneqq \left\|\mathrm{D}^2 f\right\|_{\mathcal{M}[-1,1]},$$

where we recall that D is the (distributional) derivative operator. Moreover, we also have that $\mathscr{R}\,\mathrm{TV}^2_{[-1,1]}(f) = \mathrm{TV}^2_{[-1,1]}(f)$.

B. A Representer Theorem in $\mathscr{R}\,\mathrm{BV}^2(\Omega)$

We will now discuss a representer theorem for functions in $\mathscr{R}\,\mathrm{BV}^2(\Omega)$, where $\Omega \subset \mathbb{R}^d$ is a bounded domain. For simplicity we will suppose that $\Omega = \mathbb{B}^d_1$ as defined in (12). Similar results as those stated in the sequel can be derived for more general bounded domains $\Omega \subset \mathbb{R}^d$. We have the following new representer theorem for data-fitting variational problems over $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}^d_1)$.

Theorem 5: Let $\ell(\cdot,\cdot) : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ be a strictly convex, coercive, and lower-semicontinuous loss function and let $\lambda > 0$ be an adjustable regularization parameter. Then, for any data $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N \subset \mathbb{B}^d_1 \times \mathbb{R}$, there exists a solution to the variational problem

$$\min_{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}^d_1)} \sum_{n=1}^N \ell(y_n, f(\boldsymbol{x}_n)) + \lambda\,\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}^d_1}(f) \tag{13}$$

that takes the form of a shallow ReLU network with a skip connection. In particular, it takes the form

$$s(\boldsymbol{x}) = \sum_{k=1}^K v_k\,\rho(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k) + \boldsymbol{c}^\mathsf{T}\boldsymbol{x} + c_0, \quad \boldsymbol{x} \in \mathbb{B}^d_1, \tag{14}$$

where $K \leq N - (d+1)$, ρ is the ReLU, $\boldsymbol{w}_k \in \mathbb{S}^{d-1}$, $v_k \in \mathbb{R}\setminus\{0\}$, $b_k \in [-1, 1]$, $\boldsymbol{c} \in \mathbb{R}^d$ and $c_0 \in \mathbb{R}$.

Just as in Section III-B, we view $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}^d_1)$ is the natural function space associated with shallow ReLU networks since the problem in (13) can also be recast as a finite-dimensional neural network training problem that corresponds to training a sufficiently wide shallow ReLU network (with a skip connection) with weight decay or with path-norm regularization as in (7) and (8) with the additional restriction that the activation thresholds of the neurons stay within $\mathbb{B}^d_1$. Moreover, similar to (9) we have in this case that[4]

$$\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}^d_1}(f_{\boldsymbol{\theta}}) = \sum_{k=1}^K |v_k|\|\boldsymbol{w}_k\|_2. \tag{15}$$

V. $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ and Previously Studied Spaces

Understanding the properties of shallow neural networks has received much attention since the 1990s starting with the seminal work of Barron [2] in which he studied the approximation properties of shallow sigmoidal networks in the so-called first-order spectral Barron space. The fundamental idea is to consider functions that are *synthesized* from continuously many neurons. Such functions can be expressed as an integral of a neural activation function against a finite (Radon) measure. This idea was adopted by a number of authors in the study of the so-called *variation spaces* of shallow neural networks [1], [27], [31], [43].

In this section we will discuss how $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ is related to previously studied function spaces, including the variation spaces. For simplicity we will suppose that $\Omega = \mathbb{B}^d_1$ as defined in (12). Similar results as those stated in the sequel can be derived for more general bounded domains $\Omega \subset \mathbb{R}^d$.

A. Variation Spaces

Following the setup from [43], in the case of shallow ReLU networks, the associated variation space for functions defined

[4] Just as in (9), the equality in (15) holds assuming the neural network is written in reduced form.

on $\mathbb{B}_1^d$ is defined as

$$\mathscr{V}^2(\mathbb{B}_1^d) := \left\{ f : \mathbb{B}_1^d \to \mathbb{R} : \right. \\ \left. f = \int_{\mathbb{S}^{d-1}\times[-2,2]} \rho(\boldsymbol{w}^\mathsf{T}(\cdot) - b)\,\mathrm{d}\mu(\boldsymbol{w}, b) \right\},$$

where ρ is the ReLU and $\mu \in \mathcal{M}(\mathbb{S}^{d-1}\times[-2,2])$. The reason for integrating the b variable over $[-2,2]$ is so that affine functions can be captured by this space (see [43, Section 3] for more details). This space is known to be a Banach space (see [43]) when equipped with the norm

$$\begin{aligned} \|f\|_{\mathscr{V}^2(\mathbb{B}_1^d)} &:= \inf_{\mu\in\mathcal{M}(\mathbb{S}^{d-1}\times[-2,2])} \|\mu\|_{\mathcal{M}(\mathbb{S}^{d-1}\times[-2,2])} \\ &\quad \text{s.t.} \quad f = \int_{\mathbb{S}^{d-1}\times[-2,2]} \rho(\boldsymbol{w}^\mathsf{T}(\cdot) - b)\,\mathrm{d}\mu(\boldsymbol{w}, b). \end{aligned}$$

We will now show that $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ and $\mathscr{V}^2(\mathbb{B}_1^d)$ are in fact the same space, providing more evidence that $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ is the natural function space associated to shallow ReLU networks.

Theorem 6: $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ and $\mathscr{V}^2(\mathbb{B}_1^d)$ are equivalent Banach spaces (i.e., Banach spaces with equivalent norms).

Proof: Given $f \in \mathscr{V}^2(\mathbb{B}_1^d)$, we have the representation

$$f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times[-2,2]} \rho(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)\,\mathrm{d}\mu(\boldsymbol{w}, b). \tag{16}$$

Given $g \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$, we have from Remark 3 the representation

$$g(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} \rho(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)\,\mathrm{d}\mu(\boldsymbol{w}, b) + \boldsymbol{c}^\mathsf{T}\boldsymbol{x} + c_0.$$

Clearly we can represent any function in $\mathscr{V}^2(\mathbb{B}_1^d)$ with the representation of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ and vice-versa. Therefore, $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d) = \mathscr{V}^2(\mathbb{B}_1^d)$. To see why the norms are equivalent, note that the only difference between the norms is how they handle the null space of the $\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(\cdot)$ seminorm. Since this null space is the space of affine functions, which is finite-dimensional combined with the fact that all norms are equivalent on finite-dimensional spaces, we have that the norms $\|\cdot\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)}$ and $\|\cdot\|_{\mathscr{V}^2(\mathbb{R}^d)}$ are equivalent. □

B. Spectral Barron Spaces

The spectral Barron spaces were first studied by Barron in [2]. These spaces are defined by

$$\mathscr{B}^s(\mathbb{B}_1^d) := \left\{ f : \mathscr{D}'(\mathbb{B}_1^d) : \inf_{\substack{g\in L^1(\mathbb{R}^d) \\ g|_{\mathbb{B}_1^d} = f}} \left\| \widehat{\Delta^{s/2} g} \right\|_{L^1(\mathbb{R}^d)} < \infty \right\},$$

where $\mathscr{D}'(\mathbb{B}_1^d)$ denotes the space of distributions (generalized functions) on $\mathbb{R}^d$, $\widehat{\cdot}$ denotes the (generalized) Fourier transform and Δ denotes the (weak) Laplacian operator where fractional powers are defined in terms of Riesz potentials.

Barron studied the first-order spectral Barron space, $\mathscr{B}^1(\mathbb{B}_1^d)$ in his seminal work about approximation and estimation with shallow sigmoidal networks in [2]. The higher-order variants were studied by a number of authors [25], [36], [43], [52]. In particular, it was shown in [25] that $\mathscr{B}^2(\mathbb{B}_1^d) \subset \mathscr{V}^2(\mathbb{B}_1^d)$. Therefore, by Theorem 6, we have that $\mathscr{B}^2(\mathbb{B}_1^d) \subset \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$.

C. Sobolev Spaces

$\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ also contains the classical L^1- and L^2-Sobolev spaces of order $d+1$. Let $\Omega \subset \mathbb{R}^d$ be a domain (either bounded or unbounded) and recall the Sobolev space $W^{k,p}(\Omega)$ of functions in $L^p(\Omega)$ with all (weak) derivatives up to and including order k also in $L^p(\Omega)$. This is a Banach space when equipped with the norm

$$\|f\|_{W^{k,p}(\Omega)} := \left(\sum_{|\boldsymbol{\alpha}|\le k} \|\partial^{\boldsymbol{\alpha}} f\|^p_{L^p(\Omega)} \right)^{1/p},$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$, $|\boldsymbol{\alpha}| = \alpha_1 + \cdots + \alpha_d$, and $\partial^{\boldsymbol{\alpha}}$ is the usual multi-index notation for mixed partial derivatives. When $p = 2$, $W^{k,2}(\Omega)$ is a Hilbert space and we write $H^k(\Omega)$ for $W^{k,2}(\Omega)$. The following theorem summarizes the relationship between Sobolev spaces and $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$.

Theorem 7: Given $f \in H^{d+1}(\mathbb{B}_1^d)$,

$$\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \lesssim_d \|f\|_{W^{d+1,1}(\mathbb{B}_1^d)} \lesssim_d \|f\|_{H^{d+1}(\mathbb{B}_1^d)},$$

where we recall that $\lesssim_d$ means the implicit constant depends on d. In particular, the above display says that $H^{d+1}(\mathbb{B}_1^d) \subset W^{d+1,1}(\mathbb{B}_1^d) \subset \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$.

The proof of Theorem 7 appears in Appendix C. We also remark that in order to generalize Theorem 7 to more general bounded domains $\Omega \subset \mathbb{R}^d$ requires that the boundary of Ω is sufficiently nice. It suffices that Ω has Lipschitz boundary.

D. Observations

The result of Theorem 7 says that very regular functions (those with $d+1$ derivatives in either $L^1(\mathbb{B}_1^d)$ or $L^2(\mathbb{B}_1^d)$) are contained in $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$. On the other hand, functions that are not very regular are also in $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$. For example, take any univariate function $g \in H^2(\mathbb{R})$ and use it as the profile of a ridge function

$$f(\boldsymbol{x}) = g(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{B}_1^d, \tag{17}$$

where $\boldsymbol{w} \in \mathbb{S}^{d-1}$. If g has only has two weak derivatives, then the function f is in $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ and $H^2(\mathbb{B}_1^d)$, but not in $H^{d+1}(\mathbb{B}_1^d)$. Although this function may not be very regular, it only varies in the direction $\boldsymbol{w} \in \mathbb{S}^{d-1}$. This shows that $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ can be viewed as a *mixed variation* space [12] in that it includes highly regular functions that are very isotropic, e.g., functions from the Sobolev space $H^{d+1}(\mathbb{B}_1^d)$ or less regular functions that are highly anisotropic, e.g., the ridge function in (17).

VI. Approximation Rates in $\mathscr{R}\,\mathrm{BV}^2(\Omega)$

A well-known result in approximation theory, first due to Maurey and Pisier [38], is that given a dictionary of atoms contained in a Hilbert space $\mathcal{H}$, the closure (with respect to the topology of $\mathcal{H}$) of the convex, symmetric hull of the dictionary is *immune to the curse of dimensionality* [2], [3], [11], [22],

[38]. This means that given a function f in the closure of the convex, symmetric hull of the dictionary, there exists a K-term superposition of atoms from the dictionary f_K such that $\|f - f_K\|_{\mathcal{H}} \lesssim K^{-1/2}$, which does not depend on the input dimension of the function. This fact was fundamental to the approximation rates (which do not grow with the input dimension) derived for functions belonging to the spectral Barron spaces (first studied by Barron in [2]).

It turns out that the unit-ball in the variation spaces of shallow neural networks can be characterized by the closure of the convex, symmetric hull of a dictionary of neural activation functions and are therefore also immune to the curse of dimensionality [1], [43]. We will use results from [1], [43] to readily derive approximation rates for functions in $\mathscr{R}\mathrm{BV}^2(\Omega)$ that are immune to the curse of dimensionality. For simplicity we will suppose that $\Omega = \mathbb{B}_1^d$ as defined in (12). Similar results as those stated in the sequel can be derived for more general bounded domains $\Omega \subset \mathbb{R}^d$.

Theorem 8: Given $f \in \mathscr{R}\mathrm{BV}^2(\mathbb{B}_1^d)$, there exists a shallow ReLU network (with a skip connection) with K neurons of the form in (14), denoted f_K, such that

$$\|f - f_K\|_{L^\infty(\mathbb{B}_1^d)} \lesssim_d \mathscr{R}\mathrm{TV}^2_{\mathbb{B}_1^d}(f)\, K^{-\frac{d+3}{2d}}.$$

Proof: Given $f \in \mathscr{R}\mathrm{BV}^2(\mathbb{B}_1^d)$, we have from Remark 3 the representation

$$f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} \rho(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)\,\mathrm{d}\mu(\boldsymbol{w}, b) + \boldsymbol{c}^\mathsf{T}\boldsymbol{x} + c_0.$$

It is known that the integral in the above display can be approximated in $L^\infty(\mathbb{B}_1^d)$ by a superposition of K ReLU neurons of the form $\boldsymbol{x} \mapsto \rho(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)$, $\boldsymbol{w} \in \mathbb{S}^{d-1}$ and $b \in [-1,1]$, denoted $\widetilde{f}_K$, with an approximation rate of

$$\left\| \int_{\mathbb{S}^{d-1}\times[-1,1]} \rho(\boldsymbol{w}^\mathsf{T}(\cdot) - b)\,\mathrm{d}\mu(\boldsymbol{w}, b) - \widetilde{f}_K \right\|_{L^\infty(\mathbb{B}_1^d)} \lesssim_d \|\mu\|_{\mathcal{M}(\mathbb{S}^{d-1}\times[-1,1])}\, K^{-\frac{d+3}{2d}},$$

We refer the reader to [30] and [1, Proposition 1] for this fact. Next, since $\|\mu\|_{\mathcal{M}(\mathbb{S}^{d-1}\times[-1,1])} = \mathscr{R}\mathrm{TV}^2_{\mathbb{B}_1^d}(f)$, the result follows by choosing $f_K(\boldsymbol{x}) \coloneqq \widetilde{f}_K(\boldsymbol{x}) + \boldsymbol{c}^\mathsf{T}\boldsymbol{x} + c_0$. □

Remark 9: The approximation rate in Theorem 8 cannot be improved. We refer the reader to [44] for approximation lower bounds in the variation spaces of shallow neural networks. We also remark that since Theorem 8 holds in $L^\infty(\mathbb{B}_1^d)$, it also holds for any $L^p(\mathbb{B}_1^d)$, $1 \le p < \infty$, where the implicit constant will depend on d and p.

Remark 10: As $d \to \infty$, Theorem 8 and Remark 9 says that the approximation rate is $K^{-1/2}$ and is therefore immune to the curse of dimensionality.

VII. Function Estimation in $\mathscr{R}\mathrm{BV}^2(\Omega)$

In this section we will consider the usual setup of non-parametric regression in the *fixed design* setting. Consider the problem of estimating a function $f \in \mathscr{R}\mathrm{BV}^2(\Omega)$ from the noisy samples

$$y_n = f(\boldsymbol{x}_n) + \varepsilon_n,\ n = 1, \ldots, N,$$

where $\{\varepsilon_n\}_{n=1}^N$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables and $\{\boldsymbol{x}_n\}_{n=1}^N \subset \Omega$ are fixed, but *scattered*, design points. For simplicity we will suppose that $\Omega = \mathbb{B}_1^d$ as defined in (12). Similar results as those stated in the sequel can be derived for more general bounded domains $\Omega \subset \mathbb{R}^d$.

Theorem 11: Consider the problem of estimating a function $f \in \mathscr{R}\mathrm{BV}^2(\mathbb{B}_1^d)$ such that $\mathscr{R}\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \le C$ from the noisy samples

$$y_n = f(\boldsymbol{x}_n) + \varepsilon_n,\ n = 1, \ldots, N,$$

where $\{\varepsilon_n\}_{n=1}^N$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables and $\{\boldsymbol{x}_n\}_{n=1}^N \subset \mathbb{B}_1^d$ are fixed design points. Then, any solution to the variational problem

$$\widehat{f} \in \operatorname*{arg\,min}_{f \in \mathscr{R}\mathrm{BV}^2(\mathbb{B}_1^d)} \sum_{n=1}^N |y_n - f(\boldsymbol{x}_n)|^2 \text{ s.t. } \mathscr{R}\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \le C \tag{18}$$

has a mean-squared error bound of

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^N \left|f(\boldsymbol{x}_n) - \widehat{f}(\boldsymbol{x}_n)\right|^2\right] \lesssim_d \widetilde{O}\left(C^{\frac{2d}{2d+3}}\left(\frac{N}{\sigma^2}\right)^{-\frac{d+3}{2d+3}}\right), \tag{19}$$

where $\widetilde{O}(\cdot)$ hides universal constants and logarithmic factors, where the only random variables in the expectation above are the noise terms $\{\varepsilon_n\}_{n=1}^N$.

Remark 12: Notice that as $d \to \infty$, we have that $C^{\frac{2d}{2d+3}} \to C$ and so the bound scales linearly with the constant C.

The proof of Theorem 11 follows standard techniques (see, e.g., [49, Chapter 9] or [51, Chapter 13]) based on the metric entropy of the model class

$$\{f \in \mathscr{R}\mathrm{BV}^2(\mathbb{B}_1^d) :\ \mathscr{R}\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \le C\} \tag{20}$$

with respect to the *empirical* L^2-norm defined with respect to the sampling locations $\{\boldsymbol{x}_n\}_{n=1}^N$

$$\|f\|_N^2 \coloneqq \frac{1}{N}\sum_{n=1}^N |f(\boldsymbol{x}_n)|^2. \tag{21}$$

We use our approximation rate Section VI to upper bound this metric entropy. The proof of Theorem 11 appears in Appendix D.

Remark 13: Computing an estimator that satisfies the bound in (19) requires finding a solution to the variational problem in (18). By Theorem 5, one can find a solution to the variational problem by training a sufficiently wide shallow ReLU network via gradient descent with weight decay (to a global minimizer). This is the same as finding a solution to the non-convex neural network training problem in (7), where, by Lagrange calculus, the choice of λ depends on C and the data through the data-fitting term. An alternative approach would be to the use greedy algorithms (also known as Frank–Wolfe algorithms) [1], [16], [22], [45].

Remark 14: Since when $d = 1$, $\mathscr{R}\mathrm{BV}^2(\mathbb{B}_1^d)$ is exactly the space $\mathrm{BV}^2[-1,1]$ (see the discussion in Section IV), the result of Theorem 11 recovers the well-known mean-squared error rate of $N^{-4/5}$ of locally adaptive linear spline estimators [29].

The result of Theorem 11 can be extended from the fixed design setting to the random design setting using standard techniques (see, e.g., [51, Chapter 14]). In particular, assuming the design points $\{\boldsymbol{x}_n\}_{n=1}^N \subset \mathbb{B}_1^d$ are i.i.d. uniform random variables on $\mathbb{B}_1^d$, we can use the techniques outlined in [51, Chapter 14] to derive the same mean-squared error rate (for sufficiently large N) with respect to $\|\cdot\|_{L^2(\mathbb{B}_1^d;\mathbb{P}_X)}$, where $\mathbb{P}_X$ denotes the uniform probability measure on $\mathbb{B}_1^d$. This follows from the fact that the empirical norm $\|\cdot\|_N$ concentrates to the population norm $\|\cdot\|_{L^1(\mathbb{B}_1^d;\mathbb{P}_X)}$ at the same rate as the right-hand side of (19) [51, Chapter 14, Corollary 14.15]. Therefore, we have the following corollary to Theorem 11.

Corollary 15: Consider the problem of estimating a function $f : \mathbb{B}_1^d \to \mathbb{R}$ satisfying $\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \leq C$

$$y_n = f(\boldsymbol{x}_n) + \varepsilon_n,\ n = 1, \ldots, N,$$

where $\{\varepsilon_n\}_{n=1}^N$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables and $\{\boldsymbol{x}_n\}_{n=1}^N \subset \mathbb{B}_1^d$ are i.i.d. uniform random variables on $\mathbb{B}_1^d$. Then, for sufficiently large N, any solution to the variational problem

$$\widehat{f} \in \mathop{\arg\min}_{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)} \sum_{n=1}^N |y_n - f(\boldsymbol{x}_n)|^2 \text{ s.t. } \mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \leq C$$

has a mean-squared error bound of

$$\mathbb{E}\left\|f - \widehat{f}\right\|^2_{L^2(\mathbb{B}_1^d;\mathbb{P}_X)} \lesssim_d \widetilde{O}\left(C^{\frac{2d}{2d+3}}\left(\frac{N}{\sigma^2}\right)^{-\frac{d+3}{2d+3}} + \left(\frac{N}{C'}\right)^{-\frac{d+3}{2d+3}}\right),$$

where $\widetilde{O}(\cdot)$ hides universal constants and logarithmic factors, $C' > 0$ is a constant that depends on C, and $\mathbb{P}_X$ denotes the uniform probability measure on $\mathbb{B}_1^d$.

Remark 16: Corollary 15 also provides an upper bound on the *sampling number* for the $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ model class when $\sigma \to 0$. We refer the reader to [4] for a precise definition of sampling numbers for model classes.

The following theorem shows that this mean-squared error rate cannot be improved. In other words, the rate in Theorem 11 is (up to logarithmic factors) minimax optimal.

Theorem 17: Consider the problem of estimating a function $f : \mathbb{B}_1^d \to \mathbb{R}$ satisfying $\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \leq C$ from the noisy samples

$$y_n = f(\boldsymbol{x}_n) + \varepsilon_n,\ n = 1, \ldots, N,$$

where $\{\varepsilon_n\}_{n=1}^N$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables. Then, we have the following minimax lower bound

$$\inf_{\widehat{f}} \sup_{\substack{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d) \\ \mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \leq C}} \mathbb{E}\left\|f - \widehat{f}\right\|^2_{L^2(\mathbb{B}_1^d;\mathbb{P}_X)} \gtrsim_d \left(\frac{N}{\sigma^2}\right)^{-\frac{d+3}{2d+3}},$$

where the $\inf$ is over all functions of the data and $\mathbb{P}_X$ denotes the uniform probability measure on $\mathbb{B}_1^d$.

The proof of Theorem 17 invokes a general result of Yang and Barron [53] regarding minimax rates over model classes. Invoking the result involves bounds on the $L^2(\mathbb{B}_1^d;\mathbb{P}_X)$-metric entropy of the model class in (20). We can readily bound this metric entropy due to recent results which tightly bound the metric entropy of model classes in the variation space $\mathscr{V}^2(\mathbb{B}_1^d)$ from [44]. The proof of Theorem 17 appears in Appendix E.

A. Breaking the Curse of Dimensionality

When $d = 1$, Theorems 11 and 17 recovers (up to logarithmic factors) the well-known minimax rate of $N^{-4/5}$ for $\mathrm{BV}^2[-1,1]$ model classes [13]. On the other hand, when $d \to \infty$, the rate approaches (up to logarithmic factors) $N^{-1/2}$, and is therefore immune to the curse of dimensionality. To understand why this is happening, we recall from Section V-D that $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ can be viewed as a mixed variation space.

Classical folklore in nonparametric statistics says that the minimax rate for $H^k(\mathbb{B}_1^d)$ model classes is $N^{-\frac{2k}{2k+d}}$. From Theorem 7, we have that $H^{d+1}(\mathbb{B}_1^d) \subset \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$. The minimax rate for $H^{d+1}(\mathbb{B}_1^d)$ model classes is then $N^{-\frac{2d+2}{3d+2}}$. As $d \to \infty$, this rate is $N^{-2/3}$. Therefore, we see that the space $H^{d+1}(\mathbb{B}_1^d)$ is also immune to the curse of dimensionality, but estimating functions in $H^{d+1}(\mathbb{B}_1^d)$ is strictly easier than estimating functions in the larger $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ space. This is due to the fact that $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ is a mixed variation space that contains highly isotropically regular functions that belong to the Sobolev space $H^{d+1}(\mathbb{B}_1^d)$ as well as anistropic less regular functions such as the ridge function defined in (17), which may only have two weak derivatives.

These observations about $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ make it a compelling framework for high-dimensional nonparametric estimation. Moreover, the connections with shallow ReLU networks could also shed light on the empirical success of neural networks in practice: neural networks learn functions in spaces that are immune to the curse of dimensionality.

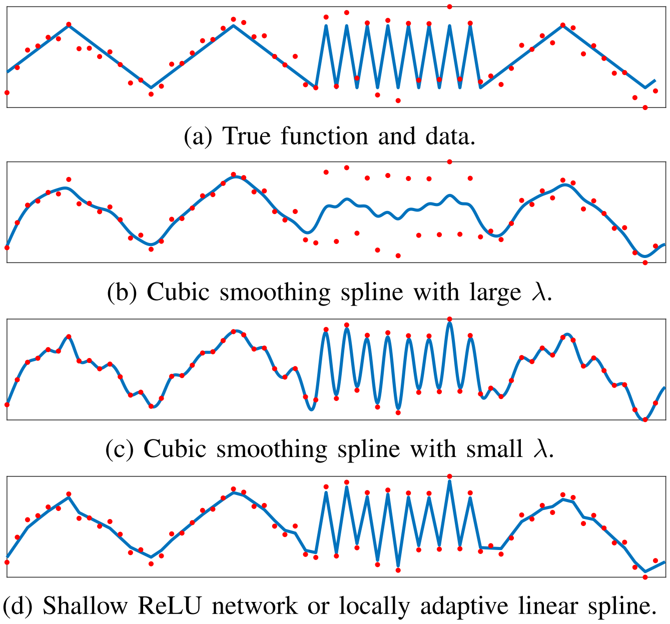
VIII. Neural Networks vs. Linear Methods

In this section we will illustrate the idea that the estimator studied in Section VII is *locally adaptive* (a term coined by Donoho and Johnstone in [13]) unlike more classical *linear methods* (which include kernel methods [42]). We will illustrate this both quantitatively via rates for function estimation as well as qualitatively via numerical experiments. For the problem of function estimation, a linear method is a method in which the estimator is a *linear* function of the data $(y_1, \ldots, y_N)$, i.e., the estimator is computed via a linear map $T : \mathbb{R}^N \to \mathscr{F}$, where $\mathscr{F}$ is some model class and T can depend on the design points $\{\boldsymbol{x}_n\}_{n=1}^N$ in an arbitrary way. Due to the sparsity-promoting nature of the $\mathcal{M}$-norm used to define $\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(\cdot)$, the estimator in Theorem 11 is a *nonlinear* function of the data. This is analogous to LASSO-type estimators arising from ℓ^1-norm regularized problems, which are nonlinear estimators for discrete-domain problems.

A. The Univariate Case

In the univariate case, we have from Remark 4 that the variational problem in (18) reduces to the (regularized) variational problem

$$\min_{f \in \mathrm{BV}^2[-1,1]} \sum_{n=1}^N |y_n - f(x_n)|^2 + \lambda \left\|\mathrm{D}^2 f\right\|_{\mathcal{M}[-1,1]}, \qquad (22)$$



(a) True function and data.

(b) Cubic smoothing spline with large λ.

(c) Cubic smoothing spline with small λ.

(d) Shallow ReLU network or locally adaptive linear spline.

Fig. 1. In (a) we generate data from noisy samples of a function in $\mathrm{BV}^2[-1,1]$ but not in $H^2[-1,1]$. In (b) and (c) we fit the data using a cubic smoothing spline with both large and small λ. In (d) we fit the data using a locally adaptive linear spline which corresponds to training a shallow ReLU network (to a global minimizer) with weight decay (or path-norm regularization).

where $\lambda > 0$ is the regularization parameter. The solutions are locally adaptive linear spline estimators [29]. It is known that the minimax rate for $\mathrm{BV}^2[-1,1]$ model classes is $N^{-4/5}$ [13], which is achieved by the locally adaptive linear spline estimator [29]. Moreover, when restricted to *linear estimators*, the linear minimax rate is known to be $N^{-3/4}$ [13], which is achieved (up to logarithmic factors) by the cubic smoothing spline estimator [10], [23]. The cubic smoothing spline is a solution to the variational problem

$$\min_{f \in H^2[-1,1]} \sum_{n=1}^{N} |y_n - f(x_n)|^2 + \lambda \left\| \mathrm{D}^2 f \right\|_{L^2[-1,1]}^2, \tag{23}$$

where

$$H^2[-1,1] \coloneqq \{f \in \mathscr{D}'[-1,1] : \|\mathrm{D}^2 f\|_{L^2[-1,1]} < \infty\},$$

is the second-order L^2-Sobolev space and $\mathscr{D}'[-1,1]$ denotes the space of distributions (generalized functions) on $[-1,1]$. Moreover, we have the strict containment $H^2[-1,1] \subset \mathrm{BV}^2[-1,1]$. The key difference between the problem in (22) and the problem in (23) is the difference between the *sparsity-promoting* $\mathcal{M}$-norm regularization in (22) and the L^2-norm regularization in (23). This is analogous to the difference between ℓ^1-norm and ℓ^2-norm regularization in discrete-domain problems.

The main takeaway message here is that this difference *quantifies* a fundamental gap between neural network estimators and any linear/kernel estimator; the gap between the rates $N^{-4/5}$ and $N^{-3/4}$. The reason for this gap is that functions in $\mathrm{BV}^2[-1,1]$ are *spatially inhomogeneous*, while functions in $H^2[-1,1]$ are *spatially homogeneous*. Neural network estimators are able to adapt to the inhomogeneities of the data-generating function (and are therefore *locally adaptive*), while linear methods cannot. This shows that even the simplest neural networks (shallow, univariate) *outperform* linear methods when the data-generating function is spatially inhomogeneous. We illustrate this phenomenon in Fig. 1, where we consider the problem of fitting data generated from a spatially inhomogenous function in $\mathrm{BV}^2[-1,1]$ that is not in $H^2[-1,1]$ using a shallow ReLU network and a cubic smoothing spline. As these results are qualitative, we manually adjusted the regularization parameter λ in the experiments in order to find solutions that visually capture the phenomenon described above. The code to generate Fig. 1 is publicly available.[5]

In Fig. 1(a) we plot a function (in blue) and generate a data set by taking noisy samples (in red) of the function plus i.i.d. Gaussian noise. Clearly this function is in $\mathrm{BV}^2[-1,1]$ but not in $H^2[-1,1]$ since taking two (distributional) derivatives of this function is an impulse train. This function is spatially inhomogeneous since it is highly oscillatory in some regions and less oscillatory in others.

In Fig. 1(b) and Fig. 1(c), we plot the cubic smoothing spline fit to the data for large and small λ, respectively. This illustrates that the cubic smoothing spline (which is a kernel method) *cannot* adapt to the spatial inhomogeneity of the underlying function. Even by adjusting the regularization parameter λ, the solution cannot adapt to the spatial inhomogeneity of the underlying function. Indeed, we see for large λ in Fig. 1(b) that the cubic smoothing spline oversmooths the high variation portion of the data and we see for small λ in Fig. 1(c) that the cubic smoothing spline undersmooths (overfits) the low variation portion of the data.

In Fig. 1(d) we plot a solution to the variational problem in (22), which is a locally adaptive linear spline which can be computed by training a shallow ReLU network (to a global minimizer) with weight decay or path-norm regularization. In this case, we see that the locally adaptive linear spline is able to adapt to the spatial inhomogeneities of the underlying function.

We also remark that wavelet shrinkage estimators, in which the mother wavelet is sufficiently regular, are also a minimax optimal estimators for nonparametric estimation of $\mathrm{BV}^2[-1,1]$ functions [13]. This shows that in the simplest setting, shallow ReLU networks trained with weight decay (to a global minimizer) perform exactly the same as classical techniques such as locally adaptive spline estimators and wavelet shrinkage estimators.

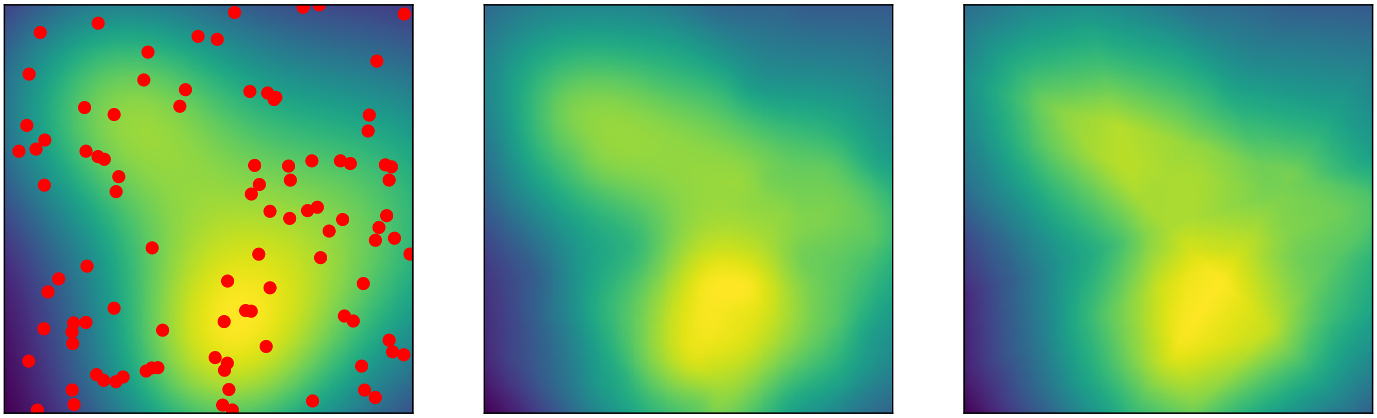
B. The Multivariate Case

In the multivariate case, we see a similar gap from the univariate case. In particular, we derive the following linear minimax lower bound for the estimation problem over $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$.

Theorem 18: Consider the problem of estimating a function $f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ satisfying $\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \leq C$ from the noisy samples

$$y_n = f(\boldsymbol{x}_n) + \varepsilon_n,\ n = 1, \ldots, N,$$

[5] https://github.com/rp/estimation-shallow-relu



(a) True function and data. (b) Thin-plate spline. (c) Shallow ReLU network.

Fig. 2. In (a) we generate noisy samples of a function in both $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^2)$ and $H^2(\mathbb{B}_1^2)$. In (b) we fit the data using a thin-plate spline. In (c) we fit the data with a shallow ReLU network trained with weight decay.

where $\{\varepsilon_n\}_{n=1}^N$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables and $\{\boldsymbol{x}_n\}_{n=1}^N \subset \mathbb{B}_1^d$ are i.i.d. uniform random variables on $\mathbb{B}_1^d$. Then, for sufficiently large N, we have the following linear minimax lower bound

$$\inf_{\widehat{f} \text{ linear}} \sup_{\substack{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d) \\ \mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \leq C}} \mathbb{E}\left\| f - \widehat{f} \right\|^2_{L^2(\mathbb{B}_1^d; \mathbb{P}_X)} \gtrsim_d \left(\frac{N}{\sigma^2} \right)^{-\frac{3}{d+3}},$$

where the inf is over all *linear* functions of the data and $\mathbb{P}_X$ denotes the uniform probability measure on $\mathbb{B}_1^d$.

The proof of Theorem 18 appears in Appendix F and hinges on several results from ridgelet analysis developed by Candès [7], [8]. Just as in the univariate case, the takeaway message here is that this lower bound quantifies a fundamental gap between neural network estimators and any linear/kernel estimator. The minimax rates for nonlinear and linear estimation are $N^{-\frac{d+3}{2d+3}}$ and $N^{-\frac{3}{d+3}}$, respectively. As $d \to \infty$, the nonlinear estimation rate tends to $N^{-1/2}$, which is immune to the curse of dimensionality, while the linear estimation rate suffers the curse of dimensionality. Moreover, these rates recover the univariate ($d = 1$) rates of $N^{-4/5}$ and $N^{-3/4}$. The reason for the gap between the nonlinear and linear minimax rates is that functions in $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ are *spatially inhomogeneous* since it is a mixed variation space and neural network estimators are able to adapt to the inhomogeneities of the data-generating function (and are therefore *locally adaptive*), while linear methods cannot.

We illustrate this phenomenon by considering the problem of estimating a two-dimensional function and compare solutions to the variational problem in (18) with the thin-plate spline estimator [50], which is a linear method and a special case of a kernel method. The thin-plate spline is a solution to the variational problem

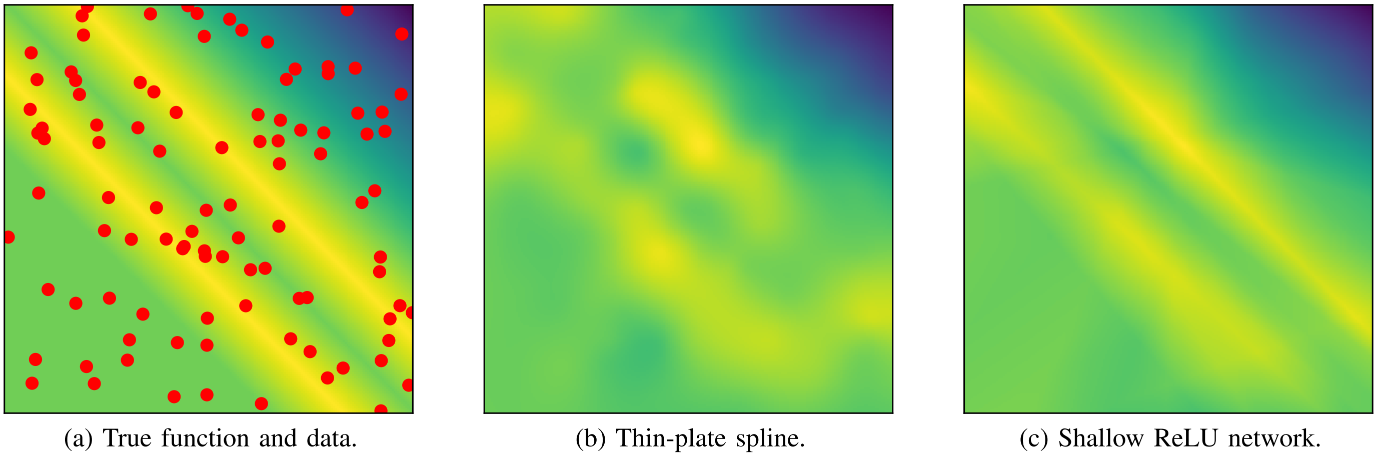
$$\min_{f \in H^2(\mathbb{B}_1^2)} \sum_{n=1}^N \left| y_n - f(\boldsymbol{x}_n) \right|^2 + \lambda \Big(\|\partial_{x_1}^2 f\|^2_{L^2(\mathbb{B}_1^2)} + 2\|\partial_{x_2}\partial_{x_1} f\|^2_{L^2(\mathbb{B}_1^2)} + \|\partial_{x_2}^2 f\|^2_{L^2(\mathbb{B}_1^d)} \Big),$$

where $H^2(\mathbb{B}_1^2)$ is the second-order L^2-Sobolev space, which is defined as the space of all functions where the regularizer in the above display is finite. Notice that the problem in the above is a generalization of the cubic smoothing spline problem in (23). We compare the shallow ReLU network estimator to the thin-plate spline estimator for two functions, one that is in both $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^2)$ and $H^2(\mathbb{B}_1^2)$, and one that is only in $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^2)$. In all the experiments, we manually adjusted the regularization parameter λ to obtain the best results for each method. Thus, the results (visually) compare the best performance of each method.

In Fig. 2 we consider a function that is a superposition of three Gaussians. This function is infinitely differentiable and therefore in both $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ and $H^2(\mathbb{B}_1^2)$. In Fig. 2(a), we plot the function with a heatmap where lighter colors correspond to larger values and darker colors correspond to smaller values. We then generate a data set by taking noisy samples (in red) of the function plus i.i.d. Gaussian noise. In Fig. 2(b), we plot the heatmap of the thin-plate spline fit to the data. We see that the thin-plate spline estimates the original function quite well. In Fig. 2(c), we plot the heatmap of the shallow ReLU network. We also see that the shallow ReLU network estimates the original function quite well.

In Fig. 3 we consider a function that is a ridge function in a random direction where the profile is a continuous piecewise-linear function, a triangular waveform. This function does not have two weak derivatives and is therefore not in $H^2(\mathbb{B}_1^2)$, but is in $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$. In Fig. 3(a), we plot the heatmap of the function. We then generate a data set by taking noisy samples (in red) of the function plus i.i.d. Gaussian noise. In Fig. 3(b), we plot the heatmap of the thin-plate spline fit to the data. We see that the thin-plate spline struggles to estimate the original function. In Fig. 3(c), we plot the heatmap of the shallow ReLU network. We see that the shallow ReLU network estimates the original function quite well.

The main takeaway message here is that the shallow ReLU network is able to *locally adapt* to the mixed variation of the data-generating function, whether it be a highly isotropically regular function or a anistropically less regular function,


(a) True function and data.

(b) Thin-plate spline.

(c) Shallow ReLU network.

Fig. 3. In (a) we generate noisy samples of a function in $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^2)$ but not in $H^2(\mathbb{B}_1^2)$. In (b) we fit the data using a thin-plate spline. In (c) we fit the data with a shallow ReLU network trained with weight decay.

while linear/kernel methods cannot. The code to generate Figs. 2 and 3 is publicly available.[6]

Remark 19: We believe that the results of Sections VIII-A and VIII-B provide compelling evidence that trying to understand neural networks via linearization schemes such as the neural tangent kernel [21] do not properly capture what neural networks are actually doing in practice. The key idea being that neural networks are able to locally adapt to the mixed variation of the underlying data-generating function.

IX. Conclusion

In this paper we studied the problem of estimating an unknown function defined on a bounded domain $\Omega \subset \mathbb{R}^d$ from $\mathscr{R}\,\mathrm{BV}^2(\Omega)$, the natural function space of shallow ReLU networks, from noisy samples. We studied the estimators that correspond to training a shallow ReLU network with weight decay (or path-norm regularization) to a global minimizer. We showed that these estimators provide (up to logarithmic factors) minimax optimal rates of convergence for $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ model classes. Moreover, these rates were immune to the curse of dimensionality. We showed that $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ contains highly isotropically regular functions that belong to the Sobolev space $H^{d+1}(\Omega)$ as well as anisotropic less regular functions, and therefore can be viewed as mixed variation spaces, giving insight into why shallow ReLU network estimators are immune to the curse of dimensionality. In particular, we quantify an explicit gap between linear and nonlinear methods and show that linear methods are suboptimal for estimating functions in $\mathscr{R}\,\mathrm{BV}^2(\Omega)$.

There are a number of open questions that may be asked. For example, considering higher-order variants of $\mathscr{R}\,\mathrm{BV}^2(\Omega)$. Our previous work in [35] also studied the higher-order variants defined on $\mathbb{R}^d$, $\mathscr{R}\,\mathrm{BV}^m(\mathbb{R}^d)$, where $m \geq 2$ is an integer. These higher-order spaces are defined by the seminorm $\mathscr{R}\,\mathrm{TV}^m(\cdot)$, which corresponds to replacing ∂_t^2 with ∂_t^m in (4) and considering a different growth restriction than in (3). These higher-order spaces correspond to shallow neural networks with activation functions that are the $(m-1)$th power of the ReLU. Although many of the results in this paper are straightforward to generalize to $\mathscr{R}\,\mathrm{BV}^m$-spaces, some of the results are also very specific to $\mathscr{R}\,\mathrm{BV}^2$-spaces. In particular, it is currently an open question on whether or not similar approximation rates as in Theorem 8 can be derived in $L^\infty(\mathbb{B}_1^d)$. Using results from [44], we can derive similar optimal approximation rates in $L^2(\mathbb{B}_1^d)$, but the mean-squared error rates hinged on the $L^\infty(\mathbb{B}_1^d)$ approximation rates. Finally, perhaps the most important open question regards estimation with deep ReLU networks fit to data. Our prior work in [37] developed a deep variant of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$, and derived a representer theorem for deep ReLU networks. This deep $\mathscr{R}\,\mathrm{BV}^2$-space could provide the right framework for nonparametric estimation with deep ReLU networks.

Appendix A
Proof of Lemma 2

The proof of Lemma 2 relies on the direct-sum decomposition of the space $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ from our previous work in [35].

A. The Direct-Sum Decomposition of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$

It was shown in [35, Theorem 22] that $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ is a non-reflexive Banach space, in particular, it is a Banach space with a sparsity-promoting norm. In this section we will summarize the relevant results from [35] about the Banach structure of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$. We first remark that the space $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ as defined in (3) is defined by a *seminorm* $\mathscr{R}\,\mathrm{TV}^2(\cdot)$. The null space of this seminorm on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ is the space of affine functions, i.e., polynomials of degree strictly less than 2, on $\mathbb{R}^d$, denoted by $\mathcal{P}_1(\mathbb{R}^d)$. In [35], we equip $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ with a *bona fide* norm by considering an arbitrary *biorthogonal system* for $\mathcal{P}_1(\mathbb{R}^d)$.

Definition 20: defn]defn:biorthogonal-system Let $\mathcal{N}$ be a finite-dimensional space with $N_0 := \dim \mathcal{N}$. The pair $(\boldsymbol{\phi}, \boldsymbol{p}) = \{(\phi_n, p_n)\}_{n=1}^{N_0}$ is called a *biorthogonal system* for $\mathcal{N}$ if $\boldsymbol{p} = \{p_n\}_{n=1}^{N_0}$ is a basis of $\mathcal{N}$ and the "boundary" functionals $\boldsymbol{\phi} = \{\phi_n\}_{n=1}^{N_0}$ with $\phi_n \in \mathcal{N}'$ (the continuous dual of $\mathcal{N}$)

[6] https://github.com/rp/estimation-shallow-relu

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

Recall from (4) that

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

For brevity, put

$$\mathrm{R} \coloneqq c_d\, \partial_t^2 \Lambda^{d-1} \mathscr{R},$$

i.e., $\mathscr{R}\,\mathrm{TV}^2(f) = \|\mathrm{R}\, f\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})}$. Also, note that $\dim \mathcal{P}_1(\mathbb{R}^d) = d+1$.

Proposition 21 (See [35, Lemma 21 and Theorem 22]): prop]prop:direct-sum-inverse 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

$$f = \mathrm{R}_{\boldsymbol{\phi}}^{-1}\, \mu + q, \tag{24}$$

where $\mu = \mathrm{R}_m\, f \in \mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})$ is an even measure,[7] $q = \sum_{k=1}^{d+1} \langle \phi_k, f \rangle p_k \in \mathcal{P}_1(\mathbb{R}^d)$, and

$$\mathrm{R}_{\boldsymbol{\phi}}^{-1} : \mu \mapsto \int_{\mathbb{S}^{d-1}\times\mathbb{R}} g_{\boldsymbol{\phi}}(\cdot, \boldsymbol{z})\, \mathrm{d}\mu(\boldsymbol{z}), \tag{25}$$

where

$$g_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{z}) = r_{\boldsymbol{z}}(\boldsymbol{x}) - \sum_{k=1}^{d+1} p_k(\boldsymbol{x}) q_k(\boldsymbol{z}), \tag{26}$$

where $r_{\boldsymbol{z}} = r_{(\boldsymbol{w},b)} = \rho(\boldsymbol{w}^\mathsf{T}(\cdot) - b)$, where ρ is the ReLU, and $q_k(\boldsymbol{z}) \coloneqq \langle \phi_k, r_{\boldsymbol{z}} \rangle$, where $\boldsymbol{z} = (\boldsymbol{w}, b) \in \mathbb{S}^{d-1}\times\mathbb{R}$.

The operator $\mathrm{R}_{\boldsymbol{\phi}}^{-1}$ defined in (25) 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

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

it is the *bona fide* inverse of R when restricted to the subspace of even measures in $\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})$. The space $\mathscr{R}\,\mathrm{BV}_{\boldsymbol{\phi}}^2(\mathbb{R}^d)$ is a concrete transcription of the abstract quotient $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)/\mathcal{P}_1(\mathbb{R}^d)$. Additionally we have from Proposition 21 the direct-sum decomposition $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d) \cong \mathscr{R}\,\mathrm{BV}_{\boldsymbol{\phi}}^2(\mathbb{R}^d) \oplus \mathcal{P}_1(\mathbb{R}^d)$, where $\mathscr{R}\,\mathrm{BV}_{\boldsymbol{\phi}}^2(\mathbb{R}^d)$ is a Banach space when equipped with the norm $f \mapsto \|\mathrm{R}\, f\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})}$ and $\mathcal{P}_1(\mathbb{R}^d)$ is a Banach space when equipped with the norm $f \mapsto \|\boldsymbol{\phi}(f)\|_1$. We also remark that the construction of $\mathrm{R}_{\boldsymbol{\phi}}^{-1}$ guarantees orthogonality of the two components in (24) and the biorthogonal system $(\boldsymbol{\phi}, \boldsymbol{p})$ guarantees unicity. This leads the following result equipping $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ with a norm to provide a Banach space structure.

Proposition 22 (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

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

where $\boldsymbol{\phi}(f) = (\langle \phi_1, f \rangle, \dots, \langle \phi_{d+1}, f \rangle) \in \mathbb{R}^{d+1}$, is a Banach space.

With these results we can now prove Lemma 2.

Proof of Lemma 2: Given $f \in \mathscr{R}\,\mathrm{BV}^2(\Omega)$ suppose there exists an extension $\widetilde{f}_{\mathsf{ext}}$ such that $\widetilde{f}_{\mathsf{ext}}\Big|_\Omega = f$ and $\mathscr{R}\,\mathrm{TV}_\Omega^2(f) = \mathscr{R}\,\mathrm{TV}^2(\widetilde{f}_{\mathsf{ext}})$ with direct-sum decomposition

$$\widetilde{f}_{\mathsf{ext}} = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} g_{\boldsymbol{\phi}}(\cdot, \boldsymbol{z})\, \mathrm{d}\widetilde{\mu}(\boldsymbol{z}) + \widetilde{q}, \tag{27}$$

such that $\operatorname{supp} \widetilde{\mu} \not\subset Z_\Omega$. Next, notice that given $g_{\boldsymbol{\phi}}(\cdot, \boldsymbol{z})$, where $\boldsymbol{z} \notin Z_\Omega$, we have that $g_{\boldsymbol{\phi}}(\cdot, \boldsymbol{z})\big|_\Omega$ is an affine function. Therefore, we can find another extension f_{ext} such that $f_{\mathsf{ext}}|_\Omega = f$ where $\mathscr{R}\,\mathrm{TV}^2(f_{\mathsf{ext}}) < \mathscr{R}\,\mathrm{TV}^2(\widetilde{f}_{\mathsf{ext}}) = \|\widetilde{\mu}\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})}$ by absorbing every $g_{\boldsymbol{\phi}}(\cdot, \boldsymbol{z})$ where $\boldsymbol{z} \notin Z_\Omega$ in the integrand of (27) into the affine term in the direct-sum decomposition so that the restriction to Ω stays the same, a contradiction. Therefore, there exists an extension $f_{\mathsf{ext}} \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ that admits an integral representation

$$f_{\mathsf{ext}}(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} g_{\boldsymbol{\phi}}(\boldsymbol{x}, (\boldsymbol{w}, b))\, \mathrm{d}\mu(\boldsymbol{w}, b) + q(\boldsymbol{x}) \tag{28}$$

such that $\operatorname{supp} \mu \subset Z_\Omega$, where μ is an *even* measure and q is an affine function.

Next, since $\Omega \subset \mathbb{R}^d$ is a bounded domain, $Z_\Omega \subset \mathbb{S}^{d-1}\times\mathbb{R}$ is also a bounded domain. Therefore, since $\operatorname{supp} \mu \subset Z_\Omega$, we can write

$$f_{\mathsf{ext}}(\boldsymbol{x}) = \int_{Z_\Omega} \rho(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)\, \mathrm{d}\mu(\boldsymbol{w}, b) + \widetilde{q}(\boldsymbol{x}), \tag{29}$$

where we combine the affine terms from $g_{\boldsymbol{\phi}}$ (defined in (26)) and q into the new affine function $\widetilde{q}$. Moreover, with the above representation we have that $\mathscr{R}\,\mathrm{TV}_\Omega^2(f) = \|\mu\|_{\mathcal{M}(Z_\Omega)}$. We also remark that although μ is even from Proposition 21, we can replace μ with a generic, i.e., not restricted to being even, measure $\widetilde{\mu} \in \mathcal{M}(Z_\Omega)$ by noting that integrating against an even measure in (28) corresponds to integrating against a generic measure by considering the activation function $\rho = |\cdot|$. Then, since $|\cdot|$ and $\max\{0, \cdot\}$ only differ by an affine function, we can absorb this difference for every neuron in the integrand with and the affine function $\widetilde{q}$ into a new affine function $\widetilde{\widetilde{q}}$. Finally, this generic, i.e., not even, measure has the same $\mathcal{M}$-norm as the even measure. □

Appendix B
Proof of Theorem 5

The proof of Theorem 5 relies on notation introduced in Appendix A.

Proof: Let $(\boldsymbol{\phi}, \boldsymbol{p})$ be a biorthogonal system for $\mathcal{P}_1(\mathbb{R}^d)$. From the proof of Lemma 2, we can identify functions in $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ with integral representations as in (28). Therefore, we can instead consider the variational problem

$$\min_{\substack{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d) \\ f = \mathrm{R}_{\boldsymbol{\phi}}^{-1}\, \mu + q \\ \operatorname{supp} \mu \subset \mathbb{S}^{d-1}\times[-1,1]}} \quad \sum_{n=1}^N \ell(y_n, f(\boldsymbol{x}_n)) + \lambda\, \|\mathrm{R}\, f\|_{\mathcal{M}(\mathbb{S}^{d-1}\times[-1,1])}.$$

The restrictions of the functions in the solution set of the above display to $\mathbb{B}_1^d$ will then correspond to the solution set of the problem in (13). Next, we remark that the proof is identical to the proof of Proposition 1 (which is a special case of our

[7] i.e., $\mathrm{d}\mu(\boldsymbol{z}) = \mathrm{d}\mu(-\boldsymbol{z})$.

prior work in [35, Theorem 1]). This is because the proof of [35, Theorem 1] boiled down to the fact that $\mathbb{S}^{d-1} \times \mathbb{R}$ is locally compact. Since $\mathbb{S}^{d-1} \times [-1, 1]$ is also locally compact, the same proof holds. □

Appendix C
Proof of Theorem 7

Proof: Since $\mathbb{B}_1^d$ has a Lipschitz boundary, there exists a bounded extension operator

$$\mathcal{E} : W^{d+1,1}(\mathbb{B}_1^d) \to W^{d+1,1}(\mathbb{R}^d),$$

where we refer the reader to [6] or [46, Chapter VI] for explicit constructions of this operator. Therefore, for $f \in W^{d+1,1}(\mathbb{B}_1^d)$,

$$\|\mathcal{E} f\|_{W^{d+1,1}(\mathbb{R}^d)} \lesssim_d \|f\|_{W^{d+1,1}(\mathbb{B}_1^d)}.$$

Given $f \in W^{d+1,1}(\mathbb{R}^d)$, it was shown in [33] that

$$\mathscr{R}\,\mathrm{TV}^2(f) \lesssim_d \|f\|_{W^{d+1,1}(\mathbb{R}^d)}.$$

Next, we have from the definition of $\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(\cdot)$ in (10) that given any $g \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$,

$$\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}\left(g|_{\mathbb{B}_1^d}\right) \leq \mathscr{R}\,\mathrm{TV}^2(g).$$

Therefore, for any $f \in W^{d+1,1}(\mathbb{B}_1^d)$,

$$\begin{aligned}\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) &\leq \mathscr{R}\,\mathrm{TV}^2(\mathcal{E} f) \\ &\lesssim_d \|\mathcal{E} f\|_{W^{d+1,1}(\mathbb{R}^d)} \\ &\lesssim_d \|f\|_{W^{d+1,1}(\mathbb{B}_1^d)}.\end{aligned}$$

The result then follows from the fact that $L^2(\mathbb{B}_1^d)$ is continuously embedded in $L^1(\mathbb{B}_1^d)$. □

Appendix D
Proof of Theorem 11

To prove Theorem 11, we will use the general result regarding nonparametric least squares estimators from [51, Chapter 13]. This general result follows from Theorem 13.5 and the remarks following, the discussion on pg. 424, and Corollary 13.7 in [51, Chapter 13]. We summarize this general result in the following proposition.

Proposition 23 (See [51, Chapter 13]): Let $\mathscr{F}$ be a *convex* model class that contains the constant function, i.e., $f \equiv 1 \in \mathscr{F}$. Given $f \in \mathscr{F}$, consider the problem of estimating f from the noisy samples

$$y_n = f(\boldsymbol{x}_n) + \varepsilon_n,\ n = 1, \dots, N,$$

where $\{\varepsilon_n\}_{n=1}^N$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables and $\{\boldsymbol{x}_n\}_{n=1}^N$ are fixed design points in the domain of f. Then, assuming a solution exists, any solution to the nonparametric least-squares problem

$$\widehat{f} \in \operatorname*{arg\,min}_{f \in \mathscr{F}} \sum_{n=1}^N |y_n - f(\boldsymbol{x}_n)|^2$$

has a mean-squared error bound of

$$\mathbb{E}\left\|f - \widehat{f}\right\|_N^2 \lesssim \delta_N^2,$$

where $\|\cdot\|_N$ is defined in (21) and $\delta_N = \delta$ satisfies the inequality

$$\frac{16}{\sqrt{N}} \int_{\frac{\delta^2}{2\sigma^2}}^{\delta} \sqrt{\log \mathcal{N}(t, \partial\mathscr{F}, \|\cdot\|_N)}\,\mathrm{d}t \leq \frac{\delta^2}{4\sigma}, \tag{30}$$

where $\mathcal{N}(t, \partial\mathscr{F}, \|\cdot\|_N)$ denotes the t-covering number of the metric space $(\partial\mathscr{F}, \|\cdot\|_N)$ and

$$\partial\mathscr{F} = \mathscr{F} - \mathscr{F} = \{f_1 - f_2 : \ f_1, f_2 \in \mathscr{F}\}.$$

We will now use Proposition 23 to prove Theorem 11.

Proof of Theorem 11: In Theorem 11, our model class is

$$\mathscr{F}_C \coloneqq \{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d) : \ \mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \leq C\}. \tag{31}$$

Since $\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(\cdot)$ is a seminorm on a Banach space, $\mathscr{F}_C$ is convex. The constant function is contained in $\mathscr{F}_C$ since the null space of $\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(\cdot)$ is the space of affine functions.

Notice that

$$\partial\mathscr{F}_C = \mathscr{F}_C - \mathscr{F}_C = 2\mathscr{F}_C \subset \mathscr{F}_{2C},$$

so it suffices to upper bound the metric entropy of $\mathscr{F}_{2C}$ to find a δ_N that satisfies (30). By noticing that $\|\cdot\|_N \leq \|\cdot\|_{L^\infty(\mathbb{B}_1^d)}$, we can use the approximation rate from Theorem 8 to upper bound (up to logarithmic factors) the metric entropy

$$\log \mathcal{N}(t, \mathscr{F}_{2C}, \|\cdot\|_N) \lessapprox_d \left(\frac{C}{t}\right)^{\frac{2d}{d+3}}$$

where $\lessapprox$ hides constant and logarithmic factors. The subscript d denotes that the implicit constant depends on d. The connection between approximation rates and metric entropy can be viewed as a variant of Carl's inequality [9] (also see [44, Theorem 10])

Next,

$$\begin{aligned}&\frac{1}{\sqrt{N}} \int_{\frac{\delta^2}{2\sigma^2}}^{\delta} \sqrt{\log \mathcal{N}(t, \partial\mathscr{F}, \|\cdot\|_N)}\,\mathrm{d}t \\ &\leq \frac{1}{\sqrt{N}} \int_0^{\delta} \sqrt{\log \mathcal{N}(t, \partial\mathscr{F}, \|\cdot\|_N)}\,\mathrm{d}t \\ &\lessapprox_d \frac{1}{\sqrt{N}} \int_0^{\delta} \left(\frac{C}{t}\right)^{\frac{d}{d+3}} \mathrm{d}t \\ &= \frac{C^{\frac{d}{d+3}}}{\sqrt{N}} \left[t^{\frac{3}{d+3}}\right]_0^{\delta} \\ &= C^{\frac{d}{d+3}} \frac{\delta^{\frac{3}{d+3}}}{\sqrt{N}}.\end{aligned}$$

From (30), we want to find $\delta_N = \delta$ that satisfies

$$C^{\frac{d}{d+3}} \frac{\delta^{\frac{3}{d+3}}}{\sqrt{N}} \lessapprox_d \frac{\delta^2}{\sigma}. \tag{32}$$

We have (up to logarithmic factors) that

$$\delta_N^2 \asymp_d C^{\frac{2d}{2d+3}} \left(\frac{N}{\sigma^2}\right)^{-\frac{d+3}{2d+3}}$$

satisfies (32). □

APPENDIX E
PROOF OF THEOREM 17

To prove Theorem 17 we will use the general result of Yang and Barron (see [53, Proposition 1] and [51, Chapter 15]) regarding minimax rates over model classes. We summarize this result in the following proposition.

Proposition 24 (See [53, Proposition 1] and [51, Chapter 15]): Let $\mathscr{F}$ be a model class. Given $f \in \mathscr{F}$, consider the problem of estimating f from the noisy samples

$$y_n = f(\boldsymbol{x}_n) + \varepsilon_n,\ n = 1, \dots, N,$$

where $\{\varepsilon_n\}_{n=1}^N$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables and $\{\boldsymbol{x}_n\}_{n=1}^N$ are i.i.d. from some probability measure $\mathbb{P}_X$ supported on $\mathbb{B}_1^d$. Then, if functions in $\mathscr{F}$ are uniformly bounded and the metric entropy is of the form

$$\log \mathcal{N}(t, \mathscr{F}, \|\cdot\|_{L^2(\mathbb{B}_1^d;\mathbb{P}_X)}) \asymp \left(\frac{1}{t}\right)^r, \quad r > 0,$$

where $\|\cdot\|_{L^2(\mathbb{B}_1^d;\mathbb{P}_X)}$ denotes the L^2-norm with respect to the measure $\mathbb{P}_X$ on $\mathbb{B}_1^d$, we have the minimax rate

$$\inf_{\widehat{f}} \sup_{f \in \mathscr{F}} \mathbb{E}\left\| f - \widehat{f} \right\|_{L^2(\mathbb{B}_1^d;\mathbb{P}_X)}^2 \asymp t_N^2,$$

where $t_N^2 = t^2$ satisfies

$$t^2 \asymp \frac{\log \mathcal{N}(t, \mathscr{F}, \|\cdot\|_{L^2(\mathbb{B}_1^d;\mathbb{P}_X)})}{N}.$$

We will use the result of Proposition 24 to derive the minimax rate for the model class

$$\mathscr{G}_C \coloneqq \{f \in \mathscr{V}^2(\mathbb{B}_1^d) : \|f\|_{\mathscr{V}^2(\mathbb{B}_1^d)} \leq C\}, \tag{33}$$

where $\mathscr{V}^2(\mathbb{B}_1^d)$ is the variation space defined in Section V. We will then use this minimax rate to derive a minimax lower bound for the model class in (20).

Lemma 25: Consider the problem of estimating $f \in \mathscr{G}_C$ (defined in (33)) from the noisy samples

$$y_n = f(\boldsymbol{x}_n) + \varepsilon_n,\ n = 1, \dots, N,$$

where $\{\varepsilon_n\}_{n=1}^N$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables and $\{\boldsymbol{x}_n\}_{n=1}^N$ are i.i.d. uniform random variables on $\mathbb{B}_1^d$. The minimax rate for this model class is

$$\inf_{\widehat{f}} \sup_{f \in \mathscr{G}_C} \mathbb{E}\left\| f - \widehat{f} \right\|_{L^2(\mathbb{B}_1^d;\mathbb{P}_X)}^2 \asymp_d N^{-\frac{d+3}{2d+3}},$$

where the $L^2(\mathbb{B}_1^d; \mathbb{P}_X)$-norm is the L^2-norm with respect to the uniform probability measure on $\mathbb{B}_1^d$.

Proof: We are interested in applying Proposition 24 with $\mathbb{P}_X$ being the uniform probability measure on $\mathbb{B}_1^d$. Since the Lebesgue measure is just a constant scaling of the uniform measure (where the constant is the volume of $\mathbb{B}_1^d$), it suffices to know the metric entropy with respect to the $L^2(\mathbb{B}_1^d)$-norm. The model class in (33) was extensively studied in [44] and it is known that

$$\log \mathcal{N}(t, \mathscr{G}_C, \|\cdot\|_{L^2(\mathbb{B}_1^d)}) \asymp_d \left(\frac{1}{t}\right)^{\frac{2d}{d+3}}.$$

We refer the reader to [44, Theorem 4 and Equation (68)] for the upper bound and [44, Theorem 8] for the lower bound. We also remark that the model class $\mathscr{G}_C$ is uniformly bounded since the functions in $\mathscr{V}^2(\mathbb{B}_1^d)$ can be written as a superposition of $L^\infty(\mathbb{B}_1^d)$-bounded atoms. With the metric entropy in the above display, we immediately have the minimax rate in the lemma statement by applying Proposition 24. □

We will now use Lemma 25 to derive a minimax lower bound for the model class in (20).

Proof of Theorem 17: It suffices to show that $\mathscr{G}_C \subset \mathscr{F}_C$, where $\mathscr{F}_C$ is defined in (31). Given $f \in \mathscr{V}^2(\mathbb{B}_1^d)$ (or in $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$, since they are the same space by Theorem 6), we can find an integral representation as in (16) such that

$$\|f\|_{\mathscr{V}^2(\mathbb{B}_1^d)} = \|\mu\|_{\mathcal{M}(\mathbb{S}^{d-1}\times[-2,2])}.$$

Next, if we let $\nu := \mu|_{\mathbb{S}^{d-1}\times[-1,1]}$, we can write f as an integral representation as in Remark 3 such that

$$\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \leq \|\nu\|_{\mathcal{M}(\mathbb{S}^{d-1}\times[-1,1])}.$$

The previous two displays imply $\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \leq \|f\|_{\mathscr{V}^2(\mathbb{B}_1^d)}$. Therefore, $\mathscr{G}_C \subset \mathscr{F}_C$. □

APPENDIX F
PROOF OF THEOREM 18

To prove Theorem 18, we will require several results from ridgelet analysis. It was shown in [7, Theorem 7] that we have the continuous embedding

$$R_{1,1}^{(d+3)/2}(\mathbb{B}_1^d) \subset \mathscr{V}^2(\mathbb{B}_1^d)$$

where we recall that $\mathscr{V}^2(\mathbb{B}_1^d)$ is the variation space for shallow ReLU networks, and $R_{p,q}^s(\mathbb{B}_1^d)$ denotes the *ridgelet space* of Candès [7]. Ridgelet spaces were proposed as a generalization of Besov spaces, and in the univariate case, the ridgelet space $R_{p,q}^s(\mathbb{B}_1^d)$ coincides with the Besov space $B_{p,q}^s[-1, 1]$.

Next, recall that we showed in the proof of Theorem 17 that $\mathscr{G}_C \subset \mathscr{F}_C$, where $\mathscr{G}_C$ and $\mathscr{F}_C$ are the model classes defined in (33) and (20), respectively. Combining this fact with the above display, we see that to prove Theorem 18, it suffices to show the linear minimax lower bound for the model class

$$\mathscr{H}_C \coloneqq \{f \in R_{1,1}^{(d+3)/2}(\mathbb{B}_1^d) : \|f\|_{R_{1,1}^{(d+3)/2}(\mathbb{B}_1^d)} \leq C\}.$$

We will make use of the following generic result.

Proposition 26 (See [8, Proof of Theorem 4.1]): Let $\mathscr{F} \subset L^2(\mathbb{B}_1^d)$ be a convex model class and consider the problem of estimating $f \in \mathscr{F}$ from the continuous white noise model

$$\mathrm{d}Y_\varepsilon(\boldsymbol{x}) = f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \varepsilon\,\mathrm{d}W(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{B}_1^d,$$

where ε is the noise level and $\mathrm{d}W(\boldsymbol{x})$ is a standard d-dimensional Wiener process. Furthermore, suppose that for any $\delta > 0$, there exists $\gtrsim_d K_\delta$ orthogonal elements $\{g_k\}_{k=1}^K \subset \mathscr{F}$ such that $\|g_k\|_{L^2(\mathbb{B}_1^d)} = \delta$, $k = 1, \dots, K$. Then, the linear minimax rate is lower-bounded by

$$\inf_{\widehat{f}\text{ linear}} \sup_{f \in \mathscr{F}} \mathbb{E}\left\| f - \widehat{f} \right\|_{L^2(\mathbb{B}_1^d)}^2 \gtrsim_d \delta_\varepsilon^2,$$

where $\delta_\varepsilon = \delta$ solves

$$\delta^2 = \varepsilon^2 K_\delta.$$

Proposition 27 (See [7, Theorem 11] and [8, Lemmas A.1, A.2, and A.3]):

For any integer $j \geq 2$, There exists a set $\{g_k\}_{k=1}^K$ of orthogonal elements with $K \gtrsim_d 2^{jd}$ contained in

$$\{f \in R_{1,1}^s(\mathbb{B}_1^d) : \|f\|_{R_{1,1}^s(\mathbb{B}_1^d)} \leq C\},$$

where $C > 0$ is a constant, such that

$$\|g_k\|_{L^2(\mathbb{B}_1^d)} = 2^{j(s-d/2)},\ k = 1, \ldots, K.$$

If we choose $\delta = 2^{j(s-d/2)}$, we see that $K \gtrsim_d \delta^{-2d/(2s-d)}$ and so the linear minimax lower bound is δ_ε^2, where $\delta_\varepsilon = \delta$ solves

$$\delta^2 = \varepsilon^2 \delta^{-2d/(2s-d)},$$

i.e.,

$$\delta_\varepsilon^2 = (\varepsilon^2)^{(2s-d)/2s}.$$

With these results, we will now prove Theorem 18.

Proof of Theorem 18: The linear minimax lower bound for the model class $\mathscr{H}_C$ corresponds to the case when $s = (d+3)/2$ and so the linear minimax lower bound for this model class (in the continuous white noise setting) will be

$$(\varepsilon^2)^{3/(d+3)}$$

By a standard sampling argument,[8] we have that the continuous white noise model is asymptotically equivalent to the estimation problem with discrete samples drawn uniformly on $\mathbb{B}_1^d$, where $\varepsilon = \sigma/\sqrt{N}$, for sufficiently large N, so we get the linear minimax lower bound of

$$\left(\frac{N}{\sigma^2}\right)^{-\frac{3}{d+3}}.$$

□

Acknowledgment

The authors would like to thank Jonathan Siegel for answering many questions about his recent work in [43], [44]. The authors would also like to thank Ronald DeVore for helpful discussions.

References

[1] F. Bach, "Breaking the curse of dimensionality with convex neural networks," *J. Mach. Learn. Res.*, vol. 18, no. 19, pp. 629–681, 2017.

[2] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inf. Theory*, vol. 39, no. 3, pp. 930–945, May 1993.

[3] A. R. Barron, A. Cohen, W. Dahmen, and R. A. DeVore, "Approximation and learning by greedy algorithms," *Ann. Statist.*, vol. 36, no. 1, pp. 64–94, Feb. 2008.

[4] P. Binev, A. Bonito, R. DeVore, and G. Petrova, "Optimal learning," 2022, *arXiv:2203.15994*.

[5] L. D. Brown and M. G. Low, "Asymptotic equivalence of nonparametric regression and white noise," *Ann. Statist.*, vol. 24, no. 6, pp. 2384–2398, Dec. 1996.

[6] A. Calderón, "Lebesgue spaces of differentiable functions," in *Proc. Sympos. Pure Math.*, vol. 4, 1961, pp. 33–49.

[7] E. J. Candès, "Ridgelets: Theory and applications," Ph.D. dissertation, Stanford University, Stanford, CA, USA, 1998.

[8] E. J. Candès, "Ridgelets: Estimating with ridge functions," *Ann. Statist.*, vol. 31, no. 5, pp. 1561–1599, Oct. 2003.

[9] B. Carl, "Entropy numbers, s-numbers, and eigenvalue problems," *J. Funct. Anal.*, vol. 41, no. 3, pp. 290–306, May 1981.

[10] C. de Boor and R. E. Lynch, "On splines and their minimum properties," *J. Math. Mech.*, vol. 15, no. 6, pp. 953–969, 1966.

[11] R. A. DeVore and V. N. Temlyakov, "Some remarks on greedy algorithms," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 173–187, Dec. 1996.

[12] D. L. Donoho *et al.*, "High-dimensional data analysis: The curses and blessings of dimensionality," *AMS Math. Challenges Lect.*, vol. 1, p. 32, Aug. 2000.

[13] D. L. Donoho and I. M. Johnstone, "Minimax estimation via wavelet shrinkage," *Ann. Statist.*, vol. 26, no. 3, pp. 879–921, 1998.

[14] S. D. Fisher and J. W. Jerome, "Spline solutions to L^1 extremal problems in one and several variables," *J. Approximation Theory*, vol. 13, no. 1, pp. 73–83, 1975.

[15] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed. New York, NY, USA: Wiley, 1999.

[16] M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval Res. Logistics Quart.*, vol. 3, nos. 1–2, pp. 95–110, 1956.

[17] J. H. Friedman and W. Stuetzle, "Projection pursuit regression," *J. Amer. Statist. Assoc.*, vol. 76, no. 376, pp. 817–823, Dec. 1981.

[18] S. Hayakawa and T. Suzuki, "On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces," *Neural Netw.*, vol. 123, pp. 343–361, Mar. 2020.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1–9.

[20] M. Imaizumi and K. Fukumizu, "Deep neural networks learn non-smooth functions effectively," in *Proc. 22nd Int. Conf. Artif. Intell. Statist.*, 2019, pp. 869–878.

[21] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018.

[22] L. K. Jones, "A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training," *Ann. Statist.*, vol. 20, no. 1, pp. 608–613, Mar. 1992.

[23] G. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," *J. Math. Anal. Appl.*, vol. 33, pp. 82–95, Sep. 1971.

[24] J. M. Klusowski and A. R. Barron, "Minimax lower bounds for ridge combinations including neural nets," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1376–1380.

[25] J. M. Klusowski and A. R. Barron, "Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls," *IEEE Trans. Inf. Theory*, vol. 64, no. 12, pp. 7649–7656, Dec. 2018.

[26] A. Krogh and J. A. Hertz, "A simple weight decay can improve generalization," in *Proc. Adv. Neural Inf. Process. Syst.*, 1992, pp. 950–957.

[27] V. Kůrková and M. Sanguineti, "Bounds on rates of variable-basis and neural-network approximation," *IEEE Trans. Inf. Theory*, vol. 47, no. 6, pp. 2659–2665, Sep. 2001.

[28] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[29] E. Mammen and S. van de Geer, "Locally adaptive regression splines," *Ann. Statist.*, vol. 25, no. 1, pp. 387–413, Feb. 1997.

[30] J. Matoušek, "Improved upper bounds for approximation by zonotopes," *Acta Mathematica*, vol. 177, no. 1, pp. 55–73, 1996.

[31] H. N. Mhaskar, "On the tractability of multivariate integration and approximation by neural networks," *J. Complex.*, vol. 20, no. 4, pp. 561–590, Aug. 2004.

[32] B. Neyshabur, R. R. Salakhutdinov, and N. Srebro, "Path-SGD: Path-normalized optimization in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 2422–2430.

[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 *Proc. Int. Conf. Learn. Represent.*, 2020, pp. 1–24.

[34] R. Parhi and R. D. Nowak, "The role of neural network activation functions," *IEEE Signal Process. Lett.*, vol. 27, pp. 1779–1783, 2020.

[35] R. Parhi and R. D. Nowak, "Banach space representer theorems for neural networks and ridge splines," *J. Mach. Learn. Res.*, vol. 22, no. 43, pp. 1–40, 2021.

[36] R. Parhi and R. D. Nowak, "On continuous-domain inverse problems with sparse superpositions of decaying sinusoids as solutions," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 5603–5607.

[8] See [5] where this argument was first rigorously formalized in the univariate case, and see [39] where this idea was rigorously formalized in the multivariate case, which applies to our setting.

[37] 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.*, vol. 4, no. 2, pp. 464–489, Jun. 2022.

[38] G. Pisier, "Remarques sur un résultat non publié de B. Maurey," *Séminaire Analyse Fonctionnelle (Dit)*, pp. 1–12, Apr. 1981.

[39] M. Reiß, "Asymptotic equivalence for nonparametric regression with multivariate and random design," *Ann. Statist.*, vol. 36, no. 4, pp. 1957–1982, 2008.

[40] P. H. P. Savarese, I. Evron, D. Soudry, and N. Srebro, "How do infinite width bounded norm networks look in function space?" in *Proc. 32nd Conf. Learn. Theory*, 2019, pp. 2667–2690.

[41] J. Schmidt-Hieber, "Nonparametric regression using deep neural networks with ReLU activation function," *Ann. Statist.*, vol. 48, no. 4, pp. 1875–1897, 2020.

[42] B. Schölkopf and A. J. Smola, *Learning With Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2002.

[43] J. W. Siegel and J. Xu, "Characterization of the variation spaces corresponding to shallow neural networks," 2021, *arXiv:2106.15002*.

[44] J. W. Siegel and J. Xu, "Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks," 2021, *arXiv:2101.12365v9*.

[45] J. W. Siegel and J. Xu, "Optimal convergence rates for the orthogonal greedy algorithm," *IEEE Trans. Inf. Theory*, vol. 68, no. 5, pp. 3354–3361, May 2022.

[46] E. M. Stein, *Singular Integrals Differentiability Properties Functions*, vol. 2. Princeton, NJ, USA: Princeton Univ. Press, 1970.

[47] T. Suzuki, "Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: Optimal rate and curse of dimensionality," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–26.

[48] M. Unser, J. Fageot, and J. P. Ward, "Splines are universal solutions of linear inverse problems with generalized TV regularization," *SIAM Rev.*, vol. 59, no. 4, pp. 769–793, 2017.

[49] S. van de Geer, *Empirical Processes M-Estimation*, vol. 6. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[50] G. Wahba, *Spline Models for Observational Data*, vol. 59. Philadelphia, PA, USA: SIAM, 1990.

[51] M. J. Wainwright, *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, vol. 48. Cambridge, U.K.: Cambridge Univ. Press, 2019.

[52] J. Xu, "Finite neuron method and convergence analysis," *Commun. Comput. Phys.*, vol. 28, no. 5, pp. 1707–1745, Jun. 2020.

[53] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Ann. Statist.*, vol. 27, no. 5, pp. 1564–1599, Oct. 1999.

Rahul Parhi (Member, IEEE) received the B.S. degree in mathematics and the B.S. degree in computer science from the University of Minnesota–Twin Cities in 2018 and the M.S. and Ph.D. degrees in electrical engineering from the University of Wisconsin–Madison in 2019 and 2022, respectively. During his Ph.D., he was supported by an NSF Graduate Research Fellowship. He is currently a Post-Doctoral Researcher with the Biomedical Imaging Group, École Polytechnique Fédérale de Lausanne, Switzerland. He is primarily interested in the applications of functional and harmonic analysis to problems in signal processing and data science.

Robert D. Nowak (Fellow, IEEE) received the Ph.D. degree in electrical engineering from the University of Wisconsin–Madison in 1995. He was a Post-Doctoral Fellow with Rice University from 1995 to 1996, an Assistant Professor with Michigan State University from 1996 to 1999, and held assistant and associate professorship positions at Rice University from 1999 to 2003. Since 2003, he has been with the University of Wisconsin–Madison, where he currently holds the Keith and Jane Morgan Nosbusch Professorship in electrical and computer engineering. His research focuses on signal processing, machine learning, optimization, and statistics. His work on sparse signal recovery and compressed sensing has received several awards, including the 2014 IEEE W.R.G. Baker Award. He has held visiting positions at INRIA, Sophia-Antipolis, in 2001; and Trinity College, Cambridge, in 2010. He has served as an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING and the ACM TRANSACTIONS ON SENSOR NETWORKS and as the Secretary of the SIAM Activity Group on Imaging Science. He was the General Chair of the 2007 IEEE Statistical Signal Processing Workshop and the Technical Program Chair of the 2003 IEEE Statistical Signal Processing Workshop, the 2004 IEEE/ACM International Symposium on Information Processing in Sensor Networks, and the inaugural IEEE GlobalSIP Conference in 2013. He is currently a Section Editor of the *SIAM Journal on Mathematics of Data Science* and a Senior Editor of the IEEE JOURNAL ON SELECTED AREAS IN INFORMATION THEORY.