ON RIDGE SPLINES, NEURAL NETWORKS, AND VARIATIONAL PROBLEMS IN RADON-DOMAIN BV SPACES

by

Rahul Parhi

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Electrical Engineering)

at the

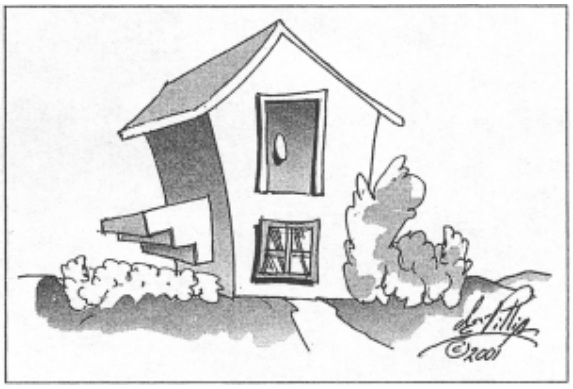
UNIVERSITY OF WISCONSIN–MADISON

2022

Date of final oral examination: July 18, 2022

The dissertation is approved by the following members of the Final Oral Committee:

Robert D. Nowak, Professor, Electrical & Computer Engineering, UW–Madison
Ronald A. DeVore, Professor, Mathematics, Texas A&M University
Nicolás García Trillos, Assistant Professor, Statistics, UW–Madison
Dimitris Papailiopoulos, Associate Professor, Electrical & Computer Engineering, UW–Madison
Betsy Stovall, Professor, Mathematics, UW–Madison
Michael Unser, Professor, Biomedical Imaging Group, École Polytechnique Fédérale de Lausanne



It is virtually impossible to get anything exactly right.
—Carl de Boor, University of Wisconsin–Madison

Acknowledgments

I am deeply indebted to my research advisor Robert D. Nowak for his continuous advice, support, and encouragement over the last four years. Rob always encouraged me to pursue research directions that I valued, even if they were different than what he was currently researching. I am also thankful that Rob allowed me to take (a lot of) time away from research to pursue my interest in teaching by hiring me as a teaching assistant for many of his courses as well as letting me help develop a new advanced topics course in Spring 2021.

I am also grateful to Ronald A. DeVore, Nicolás García Trillos, Dimitris Papailiopoulos, Betsy Stovall, and Michael Unser for serving on my committee. I would like to thank Ron for many useful discussions about topics from functional analysis and approximation theory, Nicolás for teaching me optimal transport and showing me that topics from functional analysis such as the calculus of variations and partial differential equations have applications in data science, Dimitris for useful feedback about my research, and Betsy for teaching me harmonic analysis. Finally, I would like to thank Michael—his research introduced me to the spline and its mathematical beauty, inspiring a large part of this dissertation. I look forward to moving to Lausanne, Switzerland to join his lab as a postdoctoral researcher.

The research carried out in this dissertation was supported by the National Science Foundation (NSF) through an NSF graduate research fellowship under grant DGE-1747503 and through the LUCID NSF research traineeship under grant DGE-1545481.

Contents

List of Figures

Abstract

Deep neural networks are not well understood mathematically and their success in many science and engineering applications is usually only backed by empirical evidence. In this dissertation, we study neural networks from first principles, beginning with the simplest architecture of shallow feedforward neural networks. We use tools from variational spline theory to mathematically understand neural networks. In particular, we view neural networks as a type of spline. We propose and study a new family of Banach spaces, which are bounded variation (BV) spaces defined via the Radon transform. These are the "native spaces" for neural networks. We show that finite-width neural networks are solutions to data-fitting variational problems over these spaces. Moreover, these variational problems can be recast as finite-dimensional neural network training problems with regularization schemes related to weight decay and path-norm regularization, giving theoretical insight into these common regularization schemes as well as providing several new, principled forms of regularization for (deep) neural networks. The Radon-domain BV spaces are also interesting from the perspective of functional analysis and statistical estimation. The best approximation and estimation error rates of these spaces are (essentially) independent of the input dimension, while the best linear approximation and estimation error rates suffer the curse of dimensionality. The Radon-domain BV spaces contain functions that are very smooth in all directions except (perhaps) a few directions. The anisotropic nature of these spaces distinguishes them from classical function spaces. This dissertation provides a first step towards a mathematical theory of neural networks through the lens of spline theory and functional analysis.

Bibliographic Note

This dissertation is based on the following papers:

- Rahul Parhi and Robert D. Nowak. 2020. The role of neural network activation functions. *IEEE Signal Processing Letters* 27:1779–1783.
- Rahul Parhi and Robert D. Nowak. 2021. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research* 22(43):1–40.
- Rahul Parhi and Robert D. Nowak. 2022. What kinds of functions do deep neural networks learn? Insights from variational spline theory. *SIAM Journal on Mathematics of Data Science* 4(2):464–489.
- Rahul Parhi and Robert D. Nowak. 2022. Near-minimax optimal estimation with shallow ReLU neural networks. *Submitted*. https://arxiv.org/abs/2109.08844.
- Rahul Parhi and Robert D. Nowak. 2022. On continuous-domain inverse problems with sparse superpositions of decaying sinusoids as solutions. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5603–5607.

Many of the arguments and proofs from these papers have been greatly simplified and made more concise in this dissertation. To this end, this dissertation often departs from the presentation in the papers this work is based upon, and corrects technical errors that arose in the aforementioned papers, although, to quote Carl de Boor, it is virtually impossible to get anything exactly right.

Chapter 1

Introduction

A fundamental problem in many data science[1] problems is to *reconstruct* an unknown object (signal, image, function, etc.) from possibly noisy measurements. Objects are typically modeled as functions mapping $\mathbb{R}^d \to \mathbb{R}$ and the problem is typically formulated as a *linear inverse problem*. Given $f : \mathbb{R}^d \to \mathbb{R}$ such that $f \in \mathcal{X}$, where $\mathcal{X}$ is some function space[2], we wish to *estimate* f from a vector of M measurements

$$\boldsymbol{y} = \mathrm{H}\{f\} + \boldsymbol{\varepsilon} \in \mathbb{R}^M, \tag{1.1}$$

where $\mathrm{H} : \mathcal{X} \to \mathbb{R}^M$ symbolizes the *known* linear measurement process, $\boldsymbol{\varepsilon} \in \mathbb{R}^M$ denotes a perturbation or noise term, typically assumed to be a vector of independent and identically distributed (i.i.d.) zero-mean random variables, $\boldsymbol{y} \in \mathbb{R}^M$ denotes the measured data, and M denotes the total number of measurements.

Given a vector of measurements $\boldsymbol{y} \in \mathbb{R}^M$, the goal is to construct an estimate of the data-generating object $f \in \mathcal{X}$. This type of problem is referred to as a linear inverse problem since solving this problem amounts to inverting the linear measurement operator H. Moreover, this problem is *ill-posed* since $f : \mathbb{R}^d \to \mathbb{R}$ is a continuous-domain object being reconstructed from a finite number of measurements.

[1] Data science is an interdisciplinary field interested in extracting meaningful information from data and subsumes many classical fields such as signal processing, machine learning, and statistics.

[2] Typical examples of $\mathcal{X}$ are L^2-Sobolev spaces, Besov spaces, or bounded variation spaces.

If we write $\mathrm{H}\{f\} = (\langle h_1, f\rangle, \ldots, \langle h_M, f\rangle)$, where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $\mathcal{X}$ and $\mathcal{X}'$, we see that this problem formulation captures many settings of interest. Indeed, in magnetic resonance imaging (MRI), the goal is to reconstruct an image $f : \mathbb{R}^2 \to \mathbb{R}$, from frequency domain measurements, i.e., $h_m(\boldsymbol{x}) = e^{-\mathrm{j}\boldsymbol{\omega}_m^{\mathsf{T}}\boldsymbol{x}}$ (the complex exponential) so that $\mathrm{H}\{f\} = (\widehat{f}(\boldsymbol{\omega}_1), \ldots, \widehat{f}(\boldsymbol{\omega}_M))$, where $\widehat{f}$ denotes the Fourier transform of f and $\{\boldsymbol{\omega}_m\}_{m=1}^M \subset \mathbb{R}^2$ denotes the frequency domain sampling locations. In statistics and (supervised) machine learning, the goal is to reconstruct a function $f : \mathbb{R}^d \to \mathbb{R}$ from *point evaluation* (or *ideal sampling*) measurements, i.e., $h_m(\boldsymbol{x}) = \delta(\boldsymbol{x} - \boldsymbol{x}_m)$ (the Dirac impulse) so that $\mathrm{H}\{f\} = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_M))$, where $\{\boldsymbol{x}_m\}_{m=1}^M \subset \mathbb{R}^d$ denotes the (spatial domain) sampling locations.

Classically, this type of problem was solved using spline, wavelet, or kernel based approaches, which are are well understood mathematically. The advent of the deep learning era in the last decade, has shown that deep neural network based approaches have outpeformed many state-of-the-art methods in a variety of signal processing and machine learning tasks such as image classification (Krizhevsky et al., 2012), speech recognition (Hinton et al., 2012a), and inverse problems in imaging (Jin et al., 2017; Ongie et al., 2020b).

Unfortunately, deep neural network based approaches are not well understood mathematically and usually only backed by empirical validation. In this dissertation, we study neural networks from first principles, beginning with the simplest architecture of shallow feedforward neural networks, which are superpositions of *ridge functions*. The goal of this dissertation is to develop a mathematical theory for neural networks. We show that neural networks can be viewed as a type of spline, what we call a (deep) *ridge spline*, and use tools from spline theory to mathematically understand neural networks. As a result, we propose and study a new family of Banach spaces, which are bounded variation (BV) spaces defined in the Radon domain, which are the native spaces of (sparse) ridge splines. The remainder of this chapter recounts the history of the problem of reconstructing an object from measurements leading up to the era of deep learning. At the end of this chapter, we provide a roadmap of the remaining chapters and summarize the contributions of this dissertation.

1.1 Atomic Decompositions

The problem of efficiently reconstructing an object $f \in \mathcal{X}$ hinges on choosing a *representation system* for $\mathcal{X}$. This is equivalent to finding a dictionary of atoms $\mathcal{D} \coloneqq \{\psi_n\}_{n\in\mathbb{Z}}$, where the atoms $\psi_n \in \mathcal{X}$, so that we have the *exact decompostion* of an object $f \in \mathcal{X}$ as

$$f = \sum_{n\in\mathbb{Z}} a_n \psi_n,$$

for some sequence of coefficients $\{a_n\}_{n\in\mathbb{Z}}$, or an *approximate decomposition*

$$f = \sum_{n\in\mathcal{J}} a_n \psi_n + e_N,$$

where $\mathcal{J} \subset \mathbb{Z}$ is an index set such that $|\mathcal{J}| = N$ and e_N is the error or residual which decays at some rate as $N \to \infty$ in some suitable norm. Choosing the right representation system allows for more accurate estimation and reconstruction of the data-generating object (Chen et al., 2001).

This is a classical problem in signal processing, dating back to early work in speech and radar signal processing. Such signals have naturally occurring oscillatory behavior. As a result, decomposing these signals into superpositions of sinusoids, e.g., $\mathcal{D} = \left\{x \mapsto e^{\mathrm{j}2\pi n x}\right\}_{n\in\mathbb{Z}}$, the traditional Fourier dictionary for 1-periodic signals, was widely successful. Unfortunately, Fourier-type dictionaries fail to efficiently capture objects that are spatially inhomogeneous or exhibit singularities, which are properties that arise in many real-world objects (e.g., the edges in an image). This issue can be explicitly seen via the Gibbs phenomenon.

To this end many new representation systems were proposed in the 1980s and 1990s based on wavelets (Rioul and Vetterli, 1991; Daubechies, 1992), multiresolution analysis (Mallat, 1989), and splines (Unser, 1999). The common goal of these representation systems is to efficiently decompose objects with certain *regularity properties*[3]. More generally, the field of *applied harmonic analysis* studies representation systems which provide decompositions for functions with certain regularity (Kutyniok, 2008).

[3]The regularity of an object is governed by what function space it lives in.

1.2 The Sparsity Revolution

Suppose $\mathcal{D} = \{\psi_n\}_{n\in\mathbb{Z}}$ is a representation system for $\mathcal{X}$. A common approach to reconstructing an object $f \in \mathcal{X}$ from a vector of measurements as in (1.1) is to consider the following regularized least-squares problem

$$\min_{\boldsymbol{a}\in\ell^p(\mathbb{Z})} \underbrace{\left\| \boldsymbol{y} - \mathrm{H}\left\{ \sum_{n\in\mathbb{Z}} a_n \psi_n \right\} \right\|_2^2}_{\text{data fidelity}} + \underbrace{\lambda \left\| \boldsymbol{a} \right\|_p^p}_{\text{regularization}}, \tag{1.2}$$

where $\lambda > 0$ is an adjustable hyperparameter which controls the strength of the *regularization term*[4]. The purpose of the regularization term is to ensure that the problem is *well-posed*. The choice of $p = 2$ in (1.2) corresponds to the well-known Tikhonov regularization, first proposed by Tikhonov (1963) in the context of regularizing solutions to integral equations. While this form of regularization was state-of-the art in the 20th century for many object reconstruction tasks, the dawn of the 21st century revealed that the idea of *sparsity*, i.e., $p = 1$, plays a key role in object reconstruction (Bruckstein et al., 2009; Elad, 2010). It has been seen that many real-world objects are sparse in certain dictionaries, e.g., natural images are sparse in certain wavelet dictionaries, which is the key idea behind JPEG2000 compression (Taubman and Marcellin, 2012) and magnetic resonance images are sparse in some transform domain, which is the key idea behind sparse MRI (Lustig et al., 2007).

By leveraging the idea of sparsity, many real-world objects can be reconstructed with much better accuracy than classical, Tikhonov-type, techniques. This paradigm is supported by the theory of *compressed sensing* (Candès et al., 2006; Donoho, 2006; Candès and Romberg, 2007). The problem in (1.2) with $p = 1$ is a *discrete-domain* notion of sparsity in that the sparsity is enforced on the discrete sequence of expansion coefficients. Moreover, the formulation in (1.2) referred to as the *synthesis*

[4]For large λ, the solutions to the problem in (1.2) will favor more regular solutions with a smaller ℓ^p-norm of coefficients, while for small λ, the solutions to the problem in (1.2) will favor solutions that better match the data.

formulation of the problem since the solution is explicitly *synthesized* from the dictionary $\mathcal{D} = \{\psi_n\}_{n\in\mathbb{Z}}$. In particular, when $p = 1$ and $\{\psi_n\}_{n\in\mathbb{Z}}$ corresponds to an orthogonal wavelet basis, the solutions to the problem in (1.2) correspond to the well-known *wavelet shrinkage estimator* of Donoho and Johnstone (1998), which is a minimax optimal estimator when the data-generating object lies in the scale of Besov spaces and the measurements are noisy point evaluations.

An alternative formulation of reconstruction problem is the so-called *analysis formulation*, where we consider solutions to the following regularized least-squares problem

$$\min_{f\in\mathcal{X}} \|\boldsymbol{y} - \mathrm{H}\{f\}\|_2^2 + \lambda |f|_{\mathcal{X}}^p, \tag{1.3}$$

where $\lambda > 0$ and $p \in [1, \infty)$ are adjustable hyperparameters and $|\cdot|_{\mathcal{X}}$ is a (semi)norm that defines the *native space* $\mathcal{X}$. The regularization term in (1.3) typically takes the form

$$|f|_{\mathcal{X}}^p = \|\mathrm{L}\, f\|_{\mathcal{M}}, \tag{1.4}$$

where L is usually a pseudodifferential operator and the $\mathcal{M}$-norm denotes the total variation norm (in the sense of measures) and is the *continuous-domain analogue* of the ℓ^1-norm[5]. The last few years has led to a line of work characterizing the solutions to the problem in (1.3). In particular, it has been shown that solutions to (1.3) can be expanded in terms of a dictionary matched to the regularization term (Boyer et al., 2019; Bredies and Carioni, 2020; Unser, 2021; Unser and Aziznejad, 2022). When the regularization term takes the form in (1.4), the operator L *analyzes* the function f into its coefficients in the dictionary, and so the formulation in (1.3) is referred to as the analysis formulation of the problem. This formulation is also referred to as the *variational formulation* of the problem since the optimization problem being studied is a variational problem (in the sense of the calculus of variations).

[5]The space $(\mathcal{M}(\mathbb{R}^d), \|\cdot\|_{\mathcal{M}})$ is the Banach space of finite Radon measures on $\mathbb{R}^d$. See Chapter 2, Section 2.1 for its precise definition.

For example, when $\mathcal{X}$ is the bounded variation (BV) space on $\mathbb{R}$, defined by[6]

$$\mathrm{BV}(\mathbb{R}) \coloneqq \{f \in \mathcal{S}'(\mathbb{R}) : \ \|\mathrm{D}\, f\|_{\mathcal{M}} < \infty\},$$

where $\mathcal{S}'(\mathbb{R})$ denotes the space of tempered distributions on $\mathbb{R}$ and D denotes the distributional derivative operator. It can be shown under mild conditions[7] on the measurement operator that the solution set to the variational problem

$$\mathcal{V} \coloneqq \underset{f \in \mathrm{BV}(\mathbb{R})}{\arg\min} \|\boldsymbol{y} - \mathrm{H}\{f\}\|_2^2 + \lambda \|\mathrm{D}\, f\|_{\mathcal{M}}, \tag{1.5}$$

where $\lambda > 0$ is an adjustable hyperparameter, is nonempty, convex, and weak* compact. The extreme points of $\mathcal{V}$ are given by *piecewise constant* functions of the form

$$s(x) = \sum_{n=1}^{N} a_n u(x - t_n) + c, \tag{1.6}$$

where u is the unit step function[8], $\{a_n\}_{n=1}^N \subset \mathbb{R} \setminus \{0\}$, $\{t_n\}_{n=1}^N \subset \mathbb{R}$, $c \in \mathbb{R}$, and $N < M$. The convex hull of these extreme points is the full solution set (Fisher and Jerome, 1975; Unser et al., 2017). What is remarkable about this result is that the solution set to the variational problem in (1.5) is completely characterized by piecewise constant functions where the number of jumps (or knots) N is strictly less than the number of measurements M. This is due to the sparsity-promoting nature of the $\mathcal{M}$-norm. Moreover, the regularization term $\|\mathrm{D}\, f\|_{\mathcal{M}} =: \mathrm{TV}(f)$ is exactly the *total variation* (TV) of the function f, and the problem in (1.5) corresponds to well-known technique of TV denoising (Rudin et al., 1992). In this setting we see

[6]Note that the space $\mathrm{BV}(\mathbb{R})$ we consider is not the usual space of bounded variation functions studied by mathematicians. The typical bounded variation space on $\mathbb{R}$ is the space $\overline{\mathrm{BV}}(\mathbb{R}) \coloneqq \mathrm{BV}(\mathbb{R}) \cap L^1(\mathbb{R})$, which is slightly smaller since it does not contain constant functions. We use this non-standard notation for notational convenience.

[7]In particular, that the measurement operator is weak* continuous on $\mathrm{BV}(\mathbb{R})$.

[8]The unit step function $u(x)$ is 0 for $x < 0$ and 1 for $x \geq 0$.

that the operator D analyzes functions of the form in (1.6) since

$$\mathrm{D}\, s = \sum_{n=1}^{N} a_n \delta(\cdot - t_n), \tag{1.7}$$

where δ is the Dirac impulse. From (1.6) and (1.7), we see that the dictionary matched to the regularization term $f \mapsto \|\mathrm{D}\, f\|_{\mathcal{M}}$ is exactly $\{u(\cdot - t)\}_{t\in\mathbb{R}}$, the dictionary of shifted *Green's functions* of D. The additional constant $c \in \mathbb{R}$ that appears in (1.6) lies in the null space of D. Another way to see that sparsity-promoting nature of the $\mathcal{M}$-norm is to notice that

$$\|\mathrm{D}\, s\|_{\mathcal{M}} = \sum_{n=1}^{N} |a_n| = \|\boldsymbol{a}\|_1.$$

Thus, the effect of the regularization term in (1.5) imposes sparsity in the expansion coefficients with respect to the dictionary $\{u(\cdot - t)\}_{t\in\mathbb{R}}$, similar to the synthesis formulation of the problem.

If we replace the native space in (1.5) with the kth-order variant of the BV space defined by

$$\mathrm{BV}^k(\mathbb{R}) \coloneqq \big\{ f \in \mathcal{S}'(\mathbb{R}) : \ \|\mathrm{D}^k f\|_{\mathcal{M}} < \infty \big\}, \tag{1.8}$$

where $k \in \mathbb{N}$, D^k is the kth-order derivative operator, and replace the regularization term by $\mathrm{TV}^k(f) \coloneqq \|\mathrm{D}^k f\|_{\mathcal{M}}$, the kth-order total variation of f, the solution set is completely characterized by splines[9] of degree $k-1$ (sometimes referred to as splines of order k), where the number of knots is strictly less than the number of measurements M (Fisher and Jerome, 1975; Unser et al., 2017). These spline solutions are the well-known *locally adaptive* (or *sparse adaptive*) splines of Mammen and van de Geer (1997), which are minimax optimal estimators when the data-generating object lies in the kth-order BV space and the measurements are noisy point evaluations.

[9] See Section 1.5 for the precise definition of a spline.

1.2.1 The Curse of Dimensionality

Although sparsity-promoting methods, in both the discrete- and continuous-domains, were revolutionary in both theory and practice, many techniques suffer from the *curse of dimensionality*. In particular, sparsity-promoting techniques based on splines and wavelets were fruitful for processing and reconstructing signals, images, and videos (i.e., low-dimensional objects). However, many problems in data science are inherently high-dimensional and such methods cannot efficiently estimate high-dimensional objects. Indeed, many multivariate generalizations of splines and wavelets hinge on tensor product constructions which become computationally intractable for dimensions larger than, say, ten. Moreover, the number of measurements needed to estimate an ε-close approximation to the data-generating object grows exponentially with the input dimension[10]. As a result, these sparsity-promoting techniques that were fruitful in many signal processing (low-dimensional) problems did not gain popularity for high-dimensional data science problems.

1.3 Hilbert Spaces and Kernel Methods

In order to circumvent the curse of dimensionality, the 1990s led to a (re)emergence of Hilbert space techniques (i.e., Tikhonov regularization methods) from the machine learning community. The fundamental result is the so-called *representer theorem* which characterizes the solutions to variational problems as in (1.3) when $\mathcal{X}$ is a Hilbert space and $|\cdot|_{\mathcal{X}}^p$ is the squared Hilbert norm (Wendland, 2004, Theorem 16.1). In particular, let $\mathcal{H}$ be a Hilbert space and let $\mathrm{H} : f \mapsto (\langle h_1, f\rangle, \ldots, \langle h_M, f\rangle) \in \mathbb{R}^M$ be a continuous measurement operator on $\mathcal{H}$. Then, the representer theorem states that there exists a *unique solution* $s \in \mathcal{H}$ to the variational problem

$$s = \operatorname*{arg\,min}_{f \in \mathcal{H}} \|\boldsymbol{y} - \mathrm{H}\{f\}\|_2^2 + \lambda \|f\|_{\mathcal{H}}^2, \tag{1.9}$$

[10]This can be seen explicitly via the approximation and estimation error rates with wavelets for multivariate function spaces (see, e.g., Candès, 2003, Equation 4.3).

that admits the *representation*

$$s = \sum_{m=1}^{M} a_m h_m^\sharp,$$

where $h_m^\sharp$ is the unique Riesz representer of h_m, i.e., $\langle h_m, f \rangle = \langle h_m^\sharp, f \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product of $\mathcal{H}$.

Of particular interest to the machine learning community is the setting in which $\mathcal{H}$ is a Hilbert space on $\mathbb{R}^d$ and H is the point evaluation (or ideal sampling) operator, i.e., $\mathrm{H} : f \mapsto (\langle \delta(\cdot - \boldsymbol{x}_1), f \rangle, \ldots, \langle \delta(\cdot - \boldsymbol{x}_M), f \rangle) = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_M)) \in \mathbb{R}^M$, for some set of sampling locations $\{\boldsymbol{x}_m\}_{m=1}^M \subset \mathbb{R}^d$. In this case, we are interested in Hilbert spaces in which the point evaluation operator is continuous. This corresponds to $\mathcal{H}$ being a so-called *reproducing kernel Hilbert space* (RKHS) (Aronszajn, 1950). In this case, we have that the Riesz representer of $\delta(\cdot - \boldsymbol{x}_m)$ is the so-called *reproducing kernel* of $\mathcal{H}$ given by a function $k(\cdot, \boldsymbol{x}_m) \in \mathcal{H}$. With the property

$$\langle \delta(\cdot - \boldsymbol{x}_m), f \rangle = \langle k(\cdot, \boldsymbol{x}_m), f \rangle_{\mathcal{H}} = f(\boldsymbol{x}_m)$$

for all $f \in \mathcal{H}$. The property in the above display is the so-called *reproducing property* of $k(\cdot, \cdot)$. The reproducing kernel is also symmetric and positive semidefinite (Schölkopf and Smola, 2002). In this setting, we have that the unique solution to (1.9) admits the representation

$$s = \sum_{m=1}^{M} a_m k(\cdot, \boldsymbol{x}_m).$$

In other words, the solution is a linear expansion of the reproducing kernel centered at the sampling locations. The utility of the RKHS representer theorem is that the infinite-dimensional variational problem over $\mathcal{H}$ can be recast as a finite-dimensional optimization problem by plugging in the representation in the above display into the variational problem. The resulting finite-dimensional optimization problem is

$$\min_{\boldsymbol{a} \in \mathbb{R}^M} \|\boldsymbol{y} - \mathbf{K}\boldsymbol{a}\|_2^2 + \lambda \boldsymbol{a}^\mathsf{T} \mathbf{K} \boldsymbol{a},$$

were $\mathbf{K} \in \mathbb{R}^{M \times M}$ is the so-called *kernel Gram matrix*, where $[\mathbf{K}]_{m,n} = k(\boldsymbol{x}_m, \boldsymbol{x}_n)$. This optimization problem is of size M, independent of the input dimension d, and can be easily solved numerically. This technique is the so-called *kernel trick* of machine learning.

Remark 1.1. When $\lambda \downarrow 0$, the RKHS representer theorem recovers many classical results from signal processing including the sampling theorem for bandlimited signals as well as more general sampling theorems in shift-invariant spaces (Unser, 2000).

Remark 1.2. Hilbert space techniques and, in particular, kernel methods also hold when the norm in (1.9) is replaced with a seminorm (Unser and Aziznejad, 2022). In fact, the earliest instances of kernel methods are due to Schoenberg (1964); de Boor and Lynch (1966), which consider the minimization of L^2-Sobolev seminorms subject to interpolation constraints and shows that the unique solution is a spline with knots at the sampling locations. These splines are the well-known smoothing splines popularized in the statistics community by Kimeldorf and Wahba (1970a,b, 1971).

1.3.1 Drawback of Kernel Methods

Although kernel methods are able to circumvent the curse of dimensionality from a computational perspective, they suffer from the same phenomenon observed in the signal processing community before the sparsity revolution: kernel expansions fail to capture objects that are spatially inhomogeneous or exhibit singularities. In other words, they only "work" when the data-generating function is very regular in all dimensions and does not exhibit anisoptropy. This is due to the fact that kernel methods are special cases of Tikhonov regularization since every Hilbert space is (topologically) isomorphic to an L^2-space.

To illustrate this phenomenon explicitely, we consider the problem of estimating (or reconstructing) a signal in $\mathrm{BV}^2[0,1]$, a non-Hilbertian Banach space, from noisy point evaluation measurements. The space $\mathrm{BV}^2[0,1]$ is defined as

$$\mathrm{BV}^2[0,1] \coloneqq \left\{ f \in \mathcal{D}'[0,1] : \ \|\mathrm{D}^2 f\|_{\mathcal{M}} < \infty \right\},$$

where $\mathcal{D}'[0,1]$ is the space of distributions on $[0,1]$ and the $\mathcal{M}$-norm is on $[0,1]$. This space is simply the restriction of $\mathrm{BV}^2(\mathbb{R})$ (defined in (1.8)) to $[0,1]$. We are restricting our setting to a domain since we will later quantify the estimation (or reconstruction) error of a signal with respect to the L^2-norm on $[0,1]$. In particular, consider the problem of estimating the triangular waveform in blue in Figure 1.1 from the noisy point evaluation measurements seen in red in Figure 1.1.

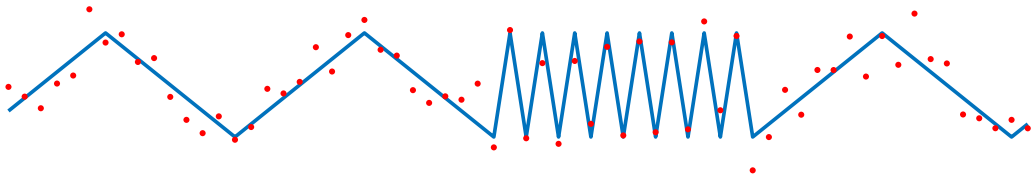


Figure 1.1: Triangular waveform (blue) contained in $\mathrm{BV}^2[0,1]$ and noisy point evaluation measurements (red).

If the triangular waveform is the function $f \in \mathrm{BV}^2[0,1]$, our measurements take the form

$$y_m = f(x_m) + \varepsilon_m,\ m = 1, \dots, M,$$

where $\{x_m\}_{m=1}^M \subset [0,1]$ and $\{\varepsilon_m\}_{m=1}^M$ are i.i.d. $\mathcal{N}(0,1)$ random variables. It is easy to see that this function is in $\mathrm{BV}^2[0,1]$ since $\mathrm{D}^2 f$ takes the form of a finite impulse train. We consider three approaches to estimate this signal:

1. The cubic smoothing spline, a kernel method;
2. The Daubechies 3 wavelet shrinkage estimator, a sparsity-promoting technique;
3. The linear locally adaptive spline, a sparsity-promoting technique.

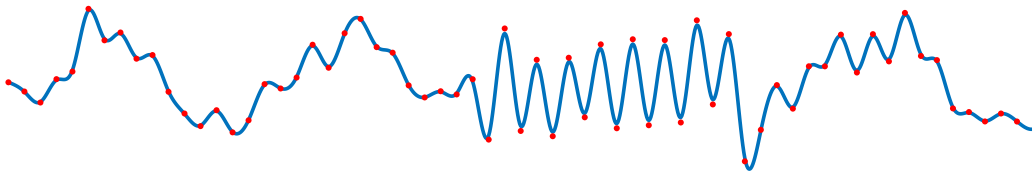
Smoothing splines. The cubic smoothing spline is the unique solution to the variational problem

$$\min_{f \in H^2[0,1]} \sum_{m=1}^{M} |y_m - f(x_m)|^2 + \lambda \|\mathrm{D}^2 f\|_{L^2}^2, \tag{1.10}$$

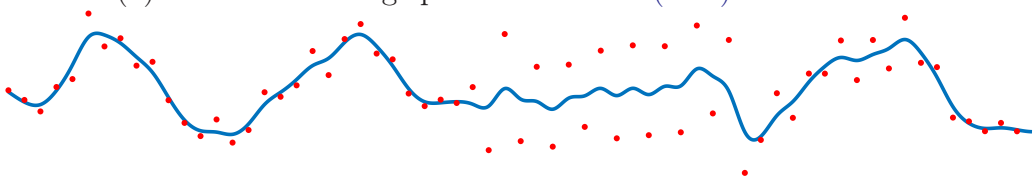
where the native space $H^2[0,1]$ is the second-order L^2-Sobolev space defined by[11]

$$H^2[0,1] \coloneqq \left\{ f \in \mathcal{D}'[0,1] : \ \|\mathrm{D}^2 f\|_{L^2}^2 < \infty \right\}.$$

Clearly the triangular waveform is not in $H^2[0,1]$ (since the Dirac impulse is not square integrable). Therefore, intuitively, we would expect the cubic smoothing spline to stuggle at estimating the triangular waveform from measurements. Indeed, this is illustrated in Figure 1.2. We see that even if we try to adjust the hyperparameter λ in (1.10), the cubic smoothing spline struggles to estimate the data generating function. For small λ, we see in Figure 1.2(a) that the cubic smoothing spline undersmooths the low variation portion of the data. For large λ, we see in Figure 1.2(b) that the cubic smoothing spline oversmooths the high variation portion of the data. The underlying issue is that the triangular waveform is spatially inhomogeneous and the smoothing spline estimate cannot adapt to spatial inhomogeneity of the data-generating function.



(a) Cubic smoothing spline solution to (1.10) with small λ.



(b) Cubic smoothing spline solution to (1.10) with large λ.

Figure 1.2: Cubic smoothing spline estimation of triangular waveform.

[11] The usual definition of $H^2[0,1]$ is all $f \in \mathcal{D}'[0,1]$ such that $f, \mathrm{D}\, f, \mathrm{D}^2 f \in L^2[0,1]$. One can easily verify that the definitions are equivalent.

Wavelet shrinkage. The Daubechies 3 wavelet shrinkage estimator is synthesized from any solution to the optimization problem

$$\min_{\boldsymbol{a}\in\ell^1(\mathbb{Z})} \sum_{m=1}^{M} \left| y_m - \sum_{n\in\mathbb{Z}} a_n \psi_n(x_m) \right|^2 + \lambda \|\boldsymbol{a}\|_1,$$

where $\{\psi_n\}_{n\in Z}$ is an ordering of the Daubechies 3 wavelet basis, which essentially corresponds to translates and dilates of the Daubechies 3 mother wavelet shown in Figure 1.3. The number 3 refers to the number of vanishing moments of the mother wavelet, which must be larger than 2 since the goal is to estimate a function in $\mathrm{BV}^2[0,1]$. We also remark that there are many technical nuances when working with wavelet systems on an interval, which we do not mention and refer the reader to Cohen et al. (1993) for more details.

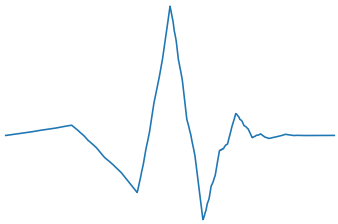


Figure 1.3: The Daubechies 3 mother wavelet.

In Figure 1.4 we see the result of estimating the triangular waveform from measurements. In particular, we see that the estimate is able to automatically adapt to the spatial inhomogeneity of the data-generating function.

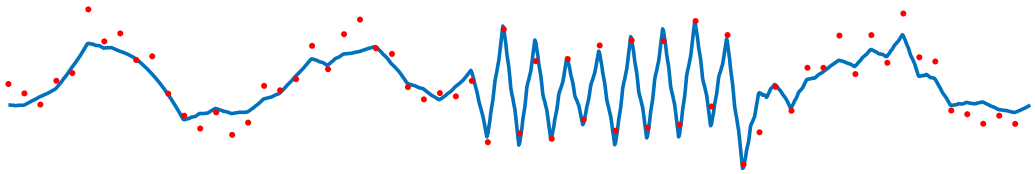


Figure 1.4: Daubechies 3 wavelet shrinkage estimation of triangular waveform.

Locally adaptive splines. The linear locally adaptive spline is any linear spline solution to the variational problem

$$\min_{f \in \mathrm{BV}^2[0,1]} \sum_{m=1}^{M} |y_m - f(x_m)|^2 + \lambda \|\mathrm{D}^2 f\|_{\mathcal{M}}.$$

Since the native space for linear locally adaptive splines is $\mathrm{BV}^2[0,1]$, we expect that the linear locally adaptive spline will automatically adapt to the spatial inhomogeneity of the data-generating function and estimate the triangular waveform well. Indeed, this is illustrated in Figure 1.5.

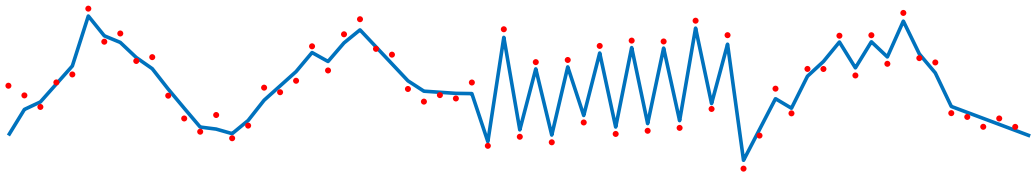


Figure 1.5: Linear locally adaptive spline estimation of triangular waveform.

Minimax rates. We can explicitly quantify the mean-squared error of the three approaches used to estimate the triangular waveform. Suppose that the sampling locations $\{x_m\}_{m=1}^M$ are nicely distributed in $[0,1]$ (e.g., uniformly distributed or equally spaced) and let $f_{M,\mathrm{sspl}}$, $f_{M,\mathrm{wav}}$, $f_{M,\mathrm{laspl}}$ denote the cubic smoothing spline, Daubechies 3 wavelet shrinkage, and linear locally adaptive spline estimators for the triangular waveform $f \in \mathrm{BV}^2[0,1]$ computed from M point evaluation measurements. Assuming that $\|\mathrm{D}^2 f\|_{\mathcal{M}} = \mathrm{TV}^2(f) \leq C$ for some universal constant C, It can be shown that:

- $\mathbb{E}\|f - f_{M,\mathrm{sspl}}\|_{L^2}^2 \lessapprox M^{-3/4}$;
- $\mathbb{E}\|f - f_{M,\mathrm{wav}}\|_{L^2}^2 \lessapprox M^{-4/5}$;
- $\mathbb{E}\|f - f_{M,\mathrm{laspl}}\|_{L^2}^2 \lessapprox M^{-4/5}$,

where $\mathbb{E}$ is the expectation operator, $\lessapprox$ hides universal constants and logarithmic (in M) factors, and the L^2-norm is on $[0,1]$. We refer the reader to Mammen and van de

Geer (1997); Donoho and Johnstone (1998); Tibshirani (2014) for these mean-squared error rates.

Moreover, it can be shown that no estimator can achieve a mean-squared error rate better (up to logarithmic factors) than that of the wavelet shrinkage and locally adaptive spline estimators for functions in $\mathrm{BV}^2[0,1]$. This is quantified via the so-called *minimax rate* for $\mathrm{BV}^2[0,1]$ (Donoho and Johnstone, 1998), which says

$$\inf_{f_M} \sup_{\substack{f \in \mathrm{BV}^2[0,1] \\ \mathrm{TV}^2(f) \leq C}} \mathbb{E}\|f - f_M\|_{L^2}^2 \asymp M^{-4/5},$$

where the inf is over all functions of the M data and $\asymp$ denotes equivalence up to universal constants.

Therefore, we see that the wavelet shrinkage and locally adaptive spline estimators are (up to logarithmic factors) *minimax optimal* for estimating functions in $\mathrm{BV}^2[0,1]$ from noisy point evaluation measurements. Intuitively, the smoothing spline estimator fails to be minimax optimal since its native space is *strictly* smaller than $\mathrm{BV}^2[0,1]$. Indeed, an application of Hölder's inequality shows $H^2[0,1] \overset{\text{c.}}{\hookrightarrow} \mathrm{BV}^2[0,1]$, where $\overset{\text{c.}}{\hookrightarrow}$ denotes a continuous embedding. On the other hand, the wavelet shrinkage estimator is designed for estimating functions in Besov spaces, and it is well-known that $B^2_{1,1}[0,1] \overset{\text{c.}}{\hookrightarrow} \mathrm{BV}^2[0,1] \overset{\text{c.}}{\hookrightarrow} B^2_{1,\infty}[0,1]$, where $B^s_{p,q}[0,1]$ is the usual Besov space on $[0,1]$ (see, e.g., Peetre, 1976). Finally, the native space for the locally adaptive spline estimator is exactly $\mathrm{BV}^2[0,1]$, so it is unsurprising it performs well at estimating functions from $\mathrm{BV}^2[0,1]$.

The fundamental difference between the smoothing spline estimator and the wavelet shrinkage or locally adaptive spline estimator is that the smoothing spline estimator is a *linear* function of the data[12], while the wavelet shrinkage and locally adaptive spline estimators are *nonlinear* functions of the data. All kernel (and more generally Hilbert space) methods are linear methods and it is well-known from the wavelet literature that linear methods cannot estimate functions that are spatially

[12]A linear method is an the estimator constructed via a *linear* mapping $T : \mathbb{R}^M \to \mathrm{BV}^2[0,1] : (y_1, \ldots, y_M) \mapsto f_{\text{linear}}$, which can depend on the sampling locations $\{x_m\}_{m=1}^M$ in an arbitrary way.

inhomogeneous or exhibit singularities, while nonlinear methods can. As we will later see in this dissertation, neural network methods are nonlinear methods.

1.4 Neurons and Neural Networks

Artificial neural networks are a representation system inspired by how biological brains process information. The fundamental building block of an artificial neural network is an artificial neuron, inspired by the biological neuron (McCulloch and Pitts, 1943; Rosenblatt, 1958). A biological neuron is depicted in Figure 1.6.

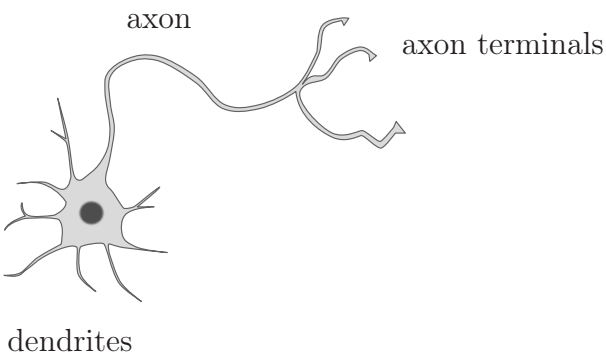


Figure 1.6: A biological neuron.

In a biological neuron, electrical signals are recieved at the dendrites and weighted according to their importance. Once the sum of the weighted input signals surpass some threshold (or bias), the neuron fires in which a firing signal sent down the axon. This firing signal can then be fed to other neurons from the axon terminals. Mathematically, if we have d input signals, the input to the neuron can be viewed as a vector $\boldsymbol{x} \in \mathbb{R}^d$. We can take a weighted sum of these input signals by taking the inner product of $\boldsymbol{x} \in \mathbb{R}^d$ with a weight vector $\boldsymbol{w} \in \mathbb{R}^d$ since

$$\boldsymbol{w}^\mathsf{T}\boldsymbol{x} = w_1 x_1 + \cdots + w_d x_d.$$

Once the quantity $\boldsymbol{w}^\mathsf{T}\boldsymbol{x}$ surpasses some threshold or bias $b \in \mathbb{R}$, the neuron can fire.

Therefore, we can model a biological neuron as the function

$$\boldsymbol{x} \mapsto \begin{cases} 1, & \text{if } \boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} \geq b \\ 0, & \text{else,} \end{cases}$$

where an output of 1 indicates that the neuron has fired. Written more compactly, an artificial neuron is the function $\boldsymbol{x} \mapsto u(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} - b)$, where u is the unit step function used in (1.6). More generally, we can replace u with an arbitrary function $\rho : \mathbb{R} \to \mathbb{R}$ and so an artificial neuron is a function mapping $\mathbb{R}^d \to \mathbb{R}$ that takes the form

$$\boldsymbol{x} \mapsto \rho(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x} - b), \tag{1.11}$$

where ρ is referred to as the *activation function*. Functions that take the form in (1.11) are referred to as *ridge functions*. In the remainder of this dissertation, we refer to artificial neurons simply as neurons and artificial neural networks simply as neural networks.

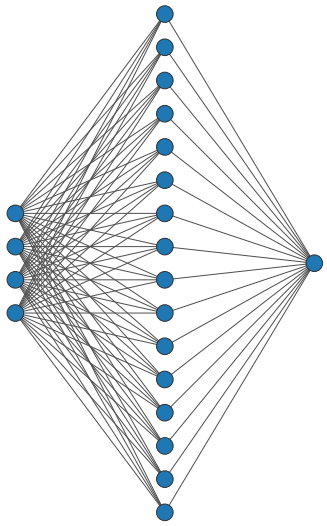


Figure 1.7: A shallow neural network.

The so-called single-layer perceptron, or *shallow neural network*, is a superposition

of neurons as in (1.11) and is a function mapping $\mathbb{R}^d \to \mathbb{R}$ that takes the form

$$\boldsymbol{x} \mapsto \sum_{n=1}^{N} v_n \rho(\boldsymbol{w}_n^{\mathsf{T}} \boldsymbol{x} - b_n), \tag{1.12}$$

where $\{v_n\}_{n=1}^N \subset \mathbb{R}$ are the weights of the outputs of the neurons, $\{\boldsymbol{w}_n\}_{n=1}^N \subset \mathbb{R}^d$ are the weights of the inputs of the neurons, $\{b_n\}_{n=1}^N \subset \mathbb{R}$ are the biases of the neurons, and the number $N \in \mathbb{N}$ is the number of neurons in the network and also corresponds to the *width* of the network. Such neural networks are often depicted with diagrams as in Figure 1.7. The nodes in Figure 1.7 represent either inputs, outputs, or neurons in the neural network and the edges between nodes represent the weights.

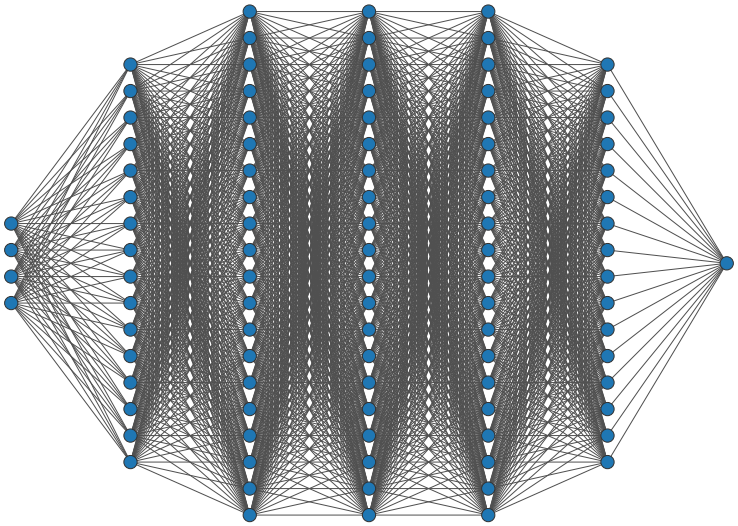


Figure 1.8: A deep neural network.

The so-called multi-layer perceptron, or *deep neural network*, corresponds to compositions of function as in (1.12), and is a mapping $\mathbb{R}^d \to \mathbb{R}$ typically written in the form

$$\begin{cases} \boldsymbol{x}^{(0)} \coloneqq \boldsymbol{x}, \\ \boldsymbol{x}^{(\ell)} \coloneqq \boldsymbol{\rho}(\mathbf{A}^{(\ell-1)} \boldsymbol{x}^{(\ell-1)} - \boldsymbol{b}^{(\ell-1)}), \quad \ell = 1, \dots, L, \\ x^{(L)} \coloneqq \boldsymbol{a}^{(L)\mathsf{T}} \boldsymbol{x}^{(L)}, \end{cases} \tag{1.13}$$

where $\boldsymbol{\rho}$ denotes applying the activation function ρ component-wise, $\mathbf{A}^{(0)} \in \mathbb{R}^{N^{(1)} \times d}$, $\mathbf{A}^{(\ell)} \in \mathbb{R}^{N^{(\ell+1)} \times N^{(\ell)}}$, $\ell = 1, \ldots, L-1$, $\boldsymbol{a}^{(L)} \in \mathbb{R}^{N^{(L)}}$, and $\boldsymbol{b}^{(\ell)} \in \mathbb{R}^{N^{(\ell+1)}}$, $\ell = 0, \ldots, L-1$. The functional mapping of a deep neural network is then the function $\boldsymbol{x} \mapsto x^{(L)}$. The matrices $\left\{\mathbf{A}^{(\ell)}\right\}_{\ell=1}^{L}$ corresponds to the weights of the inputs and outputs of the neurons, the vectors $\left\{\boldsymbol{b}^{(\ell)}\right\}_{\ell=1}^{L-1}$ correspond to the biases of the neurons, and the numbers $\left\{N^{(\ell)}\right\}_{\ell=1}^{L}$ denote the widths of the *layers*. The number of layers is referred to as the *depth* of the neural network. A deep neural network diagram is depicted in Figure 1.8, using the same conventions as in Figure 1.7. One could easily make (1.13) vector-valued by replacing the vector ${\boldsymbol{a}^{(L)}}^{\mathsf{T}}$ with a matrix $\mathbf{A}^{(L)}$.

The choice of activation function plays an important role in the efficacy of neural networks. Traditionally, the sigmoid function defined by

$$\sigma_c(x) \coloneqq \frac{1}{1 + e^{-cx}},$$

where $c > 0$, was the standard choice of activation function. This function is a smooth approximation of the unit step function. In particular, we have that

$$\lim_{c \to \infty} \sigma_c(x) = u(x), \quad x \in \mathbb{R} \setminus \{0\}.$$

Recently, however, the *rectified linear unit* (ReLU) activation function defined by

$$\mathrm{ReLU}(x) \coloneqq x_+ \coloneqq \begin{cases} x, & \text{if } x \geq 0, \\ 0, & \text{else.} \end{cases} \tag{1.14}$$

has become the preferred choice. The initial motivation of the ReLU activation function was to promote sparsity in the sense of decreasing the number of active neurons (Glorot et al., 2011). It has also been empirically observed that the training of neural networks is much faster with ReLU activations over the traditional sigmoid activation function (LeCun et al., 2015).

1.4.1 Breaking the Curse of Dimensionality

There are many intriguing properties of neural networks, particularly that they appear to break the curse of dimensionality. This was, perhaps, first observed in the seminal work of Barron (1993, 1994) studying the approximation (and estimation) properties of shallow sigmoid neural networks. The fundamental result he showed is that functions mapping $\mathbb{R}^d \to \mathbb{R}$ that satisfy certain decay properties of their Fourier transform can be approximated by a shallow sigmoidal network with N neurons with an L^2-error (on the unit ball in $\mathbb{R}^d$) at a rate that decays like $N^{-1/2}$, *independent* of the input dimension d. This result hinges on results due to Maurey and Pisier, written down by Pisier (1981), regarding dimension-free approximation rates in certain Banach spaces.

The seminal work of Barron (1993, 1994) started a line of research studying the approximation rates of functions in so-called *variation spaces* by shallow neural networks with a variety of different activation functions (Kurková and Sanguineti, 2001; Mhaskar, 2004; Bach, 2017; Siegel and Xu, 2021a). The property that these spaces seem to break the curse of dimensionality makes them interesting from a data science perspective. In particular, these spaces are *mixed variation* spaces, a term coined by Donoho (2000) to refer to function spaces that contain functions that are isotropic and very regular in all directions as well as functions that are anisotropic and very unregular in only a few directions. The fact that the approximation rates in these spaces do not grow with the input dimension says that they are "small" in comparison to classical function spaces such as Sobolev, Besov, or Triebel–Lizorkin spaces whose approximation rates grow exponentially with dimension. We refer the reader to the survey of DeVore et al. (2021) for an up to date summary on the approximation theory with neural networks, including deep neural networks.

1.4.2 Neural Network Training

Fitting data with a (deep) neural network is typically an ill-posed problem, particularly when the neural network is *overparameterized*, i.e., there are more parameters than data. In order to circumvent this, some form of regularization is typically imposed when fitting data with a neural network. Let $f_{\boldsymbol{\theta}}$ denote a deep neural network as

in (1.13), where $\boldsymbol{\theta}$ denotes all the neural network parameters. Fitting a data set $\{(\boldsymbol{x}_m, y_m)\}_{m=1}^M \subset \mathbb{R}^d \times \mathbb{R}$ with a deep neural network amounts to finding a solution to the optimization problem

$$\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{m=1}^{M} |y_m - f_{\boldsymbol{\theta}}(\boldsymbol{x}_m)|^2 + \lambda C(\boldsymbol{\theta}), \tag{1.15}$$

where $\lambda > 0$ is an adjustable hyperparameter, $\boldsymbol{\Theta}$ denotes the space of all neural network parameters, and the regularizer $C : \boldsymbol{\Theta} \to [0, \infty)$ denotes a measure of the *capacity* or *complexity* of the neural network parameterized by $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Typical choices of $C(\cdot)$ correspond to norms of the parameter vector $\boldsymbol{\theta}$. The act of *training* a neural network corresponds to a numerical method, typically a gradient based method, that attempts to find a solution to the optimization in (1.15). A common choice for $C(\cdot)$ is to consider the squared ℓ^2-norm (i.e., the squared Euclidean norm) of all the weights in the network. This form of regularization corresponds to training a neural network with *weight decay* (Krogh and Hertz, 1992).

1.5 Splines: A Perfect Fit for Data Science

The spline was invented by Schoenberg (1946). In their simplest form, splines are piecewise polynomial functions with pieces that are smoothly connected together. The joining points of the pieces are called *knots*. For a spline of order $k \in \mathbb{N}$, each piece is a polynomial of degree $k-1$, and the spline and its derivatives are continuous up to order $k-2$ at the knots. We see that a function $f : \mathbb{R} \to \mathbb{R}$ is a polynomial spline of order k if and only if

$$\mathrm{D}^k f = \sum_{n=1}^{N} a_n \delta(\cdot - t_n), \tag{1.16}$$

where D^k is the kth-order distributional derivative operator, $\{a_n\}_{n=1}^N \subset \mathbb{R}$ is a sequence of weights, and $\{t_n\}_{n=1}^N \subset \mathbb{R}$ are the knots of the spline. The function $\mathrm{D}^k f$ is referred to as the *innovation* of the spline. We see that splines are continuous-domain functions

that are intrinsically sparse and, in particular, have a *finite rate of innovation* (Vetterli et al., 2002). Since the order of the differential operator that uncovers the innovation of the spline is of order k, a spline of degree $k-1$ is referred to as a spline of order k. By associating a spline to an operator as in (1.16), we see that we can define generalized splines via the following definition.

Definition 1.3 (see, e.g., Unser et al. (2017, Definition 2))**.** *A function $f : \mathbb{R}^d \to \mathbb{R}$ of slow growth*[13] *is said to be an* L*-spline if*

$$\mathrm{L}\, f = \sum_{n=1}^{N} a_n \delta(\cdot - \boldsymbol{t}_n),$$

where L *is a pseudodifferential operator with a finite-dimensional null space, $\{a_n\}_{n=1}^N \subset \mathbb{R}$ is a sequence of weights, and $\{\boldsymbol{t}_n\}_{n=1}^N \subset \mathbb{R}^d$ are the locations of the knots of the spline. The function* $\mathrm{L}\, f$ *is referred to as the innovation of f.*

We see that Definition 1.3 recovers the polynomial splines when $\mathrm{L} = \mathrm{D}^k$. In this dissertation, we will mostly be interested D^k-splines and variants. Definition 1.3 implies that a D^k-spline admits the representation

$$x \mapsto \sum_{n=1}^{N} a_n \rho_k(x - t_n) + \underbrace{\sum_{\ell=0}^{k-1} c_\ell x^\ell}_{=: q_k(x)},$$

where ρ_k is a *Green's function* (or fundamental solution) of D^k, e.g., choose

$$\rho_k(x) \coloneqq \frac{1}{2} \operatorname{sgn}(x) \frac{x^{k-1}}{(k-1)!}, \tag{1.17}$$

[13] The growth restriction of the function is to ensure that the null space of L is finite-dimensional. This is a non-issue in the univariate case, but in the multivariate case, it is difficult to ensure that the null space of an operator is finite-dimensional without a growth restriction. We refer the reader to Unser et al. (2017) for more details.

or

$$\rho_k(x) \coloneqq \frac{x_+^{k-1}}{(k-1)!},$$

and q_k lies in the null space of D^k (polynomials of degree strictly less than k). The functions in the above display are referred to as *truncated power functions*. Splines have always been fundamentally connected to data science, ranging from classical techniques such as kernel methods, to more modern techniques such as wavelets and neural networks. In the remainder of this section, we summarize how splines provide a unifying framework for data science techniques and are therefore a perfect fit for data science.

1.5.1 Splines and Kernel Methods

As mentioned in Remark 1.2 in Section 1.3, the precursor to kernel methods were smoothing splines, which are the unique solution to the data-fitting variational problem

$$\min_{f \in \mathrm{BL}^k(\mathbb{R})} \sum_{m=1}^{M} |y_m - f(x_m)|^2 + \lambda \|\mathrm{D}^k f\|_{L^2}^2, \tag{1.18}$$

where the native space $\mathrm{BL}^k(\mathbb{R})$ denotes the kth-order Beppo-Levi space defined by[14]

$$\mathrm{BL}^k(\mathbb{R}) \coloneqq \left\{ f \in \mathcal{S}'(\mathbb{R}) : \ \|\mathrm{D}^k f\|_{L^2}^2 < \infty \right\}.$$

The unique solution to (1.18) is a D^{2k}-spline with knots at the sampling locations $\{x_m\}_{m=1}^M$. This particular spline is known as a smoothing spline (Unser and Blu, 2005). Kernel methods consider problems as in (1.18), but instead consider data-fitting variational problems over abstract (semi)reproducing kernel Hilbert spaces.

[14] The Beppo-Levi space $\mathrm{BL}^k(\mathbb{R})$ is sometimes mistakenly referred to as the homogeneous Sobolev sapce $\dot{H}^k(\mathbb{R})$ since both are defined as functions in which the homogeneous Sobolev seminorm $f \mapsto \|\mathrm{D}^k f\|_{L^2}$ is finite, but we remark that homogeneous function spaces are typically viewed as subspaces of quotient space $\mathcal{S}'(\mathbb{R})/\mathcal{P}(\mathbb{R})$, where $\mathcal{P}(\mathbb{R})$ is the space of polynomials on $\mathbb{R}$. If we consider the restriction of functions in $\mathrm{BL}^k(\mathbb{R})$ to an interval, say, $[0,1]$ the resulting space is the usual Sobolev space $H^k[0,1]$. We refer the reader to Wendland (2004, Chapter 10) for more details about Beppo-Levi spaces.

1.5.2 Splines and Wavelets

Wavelets were revolutionary in data science, particularly in sparse signal processing, due to the idea of wavelet shrinkage by Donoho and Johnstone (1998). Splines and wavelets have always been fundamentally connected due to the self-similarity and multiresolution properties that arise in spline and wavelet atoms. In particular, wavelet systems can be constructed from spline functions resulting in systems of spline wavelets, which have many unique properties over more conventional wavelet systems. We refer the reader to the article of Unser (1997) for reasons to use spline wavelets. Moreover, the core of every wavelet system is in fact a spline function since every scaling function can be expressed as the convolution of a B-spline and a distribution (Unser and Blu, 2003).

1.5.3 Splines and Neural Networks

As we saw in Section 1.4, a (deep) neural network is defined via compositions of affine functions and nonlinearities via the activation function, where the standard choice of activation function is the ReLU defined in (1.14). The ReLU activation function is exactly the second-order truncated power function, which is the fundamental building block for linear splines. Therefore we immediately see a connection between (linear) splines and neural networks.

A very special property of deep ReLU networks is that their input-output relation is continuous piecewise-linear (CPwL) (Montufar et al., 2014). The reverse is also true in that any CPwL function can be represented with a sufficiently wide and deep ReLU network (Arora et al., 2018). Thus, one can interpret a deep ReLU network as a multivariate linear spline. This connection between deep neural networks and splines has been observed by a number of authors (Poggio et al., 2015; Unser, 2019; Balestriero and Baraniuk, 2020). In particular, one can view a deep neural network as a hierarchical or deep spline to emphasize the compositional nature of deep neural networks.

Another remarkable observation, perhaps first made by Savarese et al. (2019) is

that shallow univariate ReLU networks trained with weight decay[15] are linear locally adaptive splines. This observation can be deduced quite easily. Indeed, consider fitting the data set $\{(x_m, y_m)\}_{m=1}^M \subset \mathbb{R} \times \mathbb{R}$ and recall that a linear locally adaptive spline is any spline solution to the variational problem

$$\min_{f \in \mathrm{BV}^2(\mathbb{R})} \sum_{m=1}^M |y_m - f(x_m)|^2 + \lambda \|\mathrm{D}^2 f\|_{\mathcal{M}}.$$

Since the extreme points of the solution set to the above display are linear splines where the number of knots is strictly less than the number of data, we know there exists a solution to the above display of the form

$$f_{\mathrm{laspl}}(x) = \sum_{n=1}^{N_0} a_n \, \mathrm{ReLU}(x - t_n) + c_1 x + c_0,$$

for some $N_0 < M$. A direct calculation shows that

$$\|\mathrm{D}^2 f_{\mathrm{laspl}}\|_{\mathcal{M}} = \sum_{n=1}^{N_0} |a_n| = \|\boldsymbol{a}\|_1.$$

Therefore, if we consider data-fitting over all linear splines with at least N_0 knots, where we regularize the quantity $\|\boldsymbol{a}\|_1$, any solution will be a linear locally adaptive spline. Next, notice that a shallow ReLU network can be written as

$$f_{\mathrm{nn}}(x) = \sum_{n=1}^{N} v_n \, \mathrm{ReLU}(w_n x - b_n) + \underbrace{c_1 x + c_0}_{(*)},$$

where the additional affine function that appears in $(*)$ is referred to as a *skip connection* in neural network parlance (He et al., 2016). From Definition 1.3 we see

[15] In particular, any global minimizer to the optimization problem which corresponds to training a shallow univariate ReLU network with weight decay.

that f_{nn} is a linear spline since

$$\mathrm{D}^2 f_{\text{nn}} = \sum_{n=1}^{N} v_n |w_n| \delta\left(\cdot - \frac{b_n}{w_n} \right).$$

Therefore, if we consider data-fitting over all shallow univariate ReLU networks (with a skip connection) with $N \geq N_0$ neurons and regularize the quantity

$$\|\mathrm{D}^2 f_{\text{nn}}\|_{\mathcal{M}} = \sum_{n=1}^{N} |v_n||w_n|, \tag{1.19}$$

then, any solution will be a locally adaptive spline. Finally, we will later show in Theorem 3.15 that the solutions to the regularized neural network training problem with the regularizer in the above display are equivalent to the solutions when regularizing the quantity

$$\frac{1}{2} \sum_{n=1}^{N} |v_n|^2 + |w_n|^2, \tag{1.20}$$

which corresponds to training a shallow ReLU network (with a skip connection) with weight decay. The equivalence of regularizers in (1.19) and (1.20) goes back to the work of Grandvalet (1998) and was then rediscovered by Neyshabur et al. (2015b). What is remarkable about this result is that the regularizer in (1.20) appears to be a Tikhonov-type regularizer, but is in fact it is a sparsity-promoting regularizer. Thus, we see, for a variety of reasons, that (deep) neural networks are splines.

1.6 Roadmap and Contributions

This dissertation is organized as follows.

Chapter 2: In this chapter we introduce the relevant background and notation from functional analysis used extensively in this dissertation. The expert reader can simply glance at this chapter to familiarize themselves with the notation used in this dissertation.

Chapter 3: In this chapter we propose and study a new family of Banach spaces which are multivariate generalizations of the $\mathrm{BV}^k(\mathbb{R})$ spaces defined in (1.8), inspired by the Radon-domain seminorm proposed by Kurková et al. (1997); Ongie et al. (2020a), with the property that the extreme points of the solution set to data-fitting variational problems over these spaces correspond to shallow neural networks (with skip connections) with less neurons than data. Since in the univariate case, these extreme points correspond to kth-order splines, we call such functions kth-order ridge splines to emphasize that they are superpositions of ridge functions. We also refer to these spaces as the native spaces for (sparse) ridge splines. These spaces are, in essence, kth-order BV spaces defined in the Radon domain, so we call these spaces $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$. We then show that the solutions to optimization problem that corresponds to training a shallow ReLU network with weight decay are solutions to data-fitting variational problem over $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$, providing a multivariate generalization of the result of Savarese et al. (2019) discussed in Section 1.5.3. We then extend this characterization to vector-valued, compositional function spaces and deep neural networks, providing several new, principled forms of regularization for deep neural networks.

Chapter 4: In this chapter we consider the restriction of $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ spaces to a bounded domain $\Omega \subset \mathbb{R}^d$. We then derive (nonlinear) approximation rates in $L^2(\Omega)$ for functions in $\mathscr{R}\,\mathrm{BV}^k(\Omega)$ and show that these rates cannot be improved. These results readily follow from showing that the spaces $\mathscr{R}\,\mathrm{BV}^k(\Omega)$ are equivalent (in the sense of Banach spaces) to the variation spaces for shallow neural networks with activation functions given by ρ_k defined in (1.17), and invoking the approximation rates derived by Siegel and Xu (2021b). In the special case of $\mathscr{R}\,\mathrm{BV}^2(\Omega)$, we are also able to derive approximation rates in $L^\infty(\Omega)$ and use this result to show that the solutions to the problem of training a shallow ReLU network with weight decay are (up to logarithmic factors) minimax optimal estimators for estimating functions in $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ from noisy point evaluation measurements, providing a multivariate generalization of the

minimax results of Mammen and van de Geer (1997) for functions in $\mathrm{BV}^2[0,1]$. The approximation and estimation error rates do not grow with the input dimension, providing insight into the phenomenon that neural networks seem to break the curse of dimensionality. We also derive a *linear* minimax lower bound for estimation of functions in $\mathscr{R}\,\mathrm{BV}^2(\Omega)$, showing that linear methods (which include kernel methods) are suboptimal at estimating functions in $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ and necessarily suffer the curse of dimensionality.

Chapter 5: In this chapter we discuss several directions for future work including understanding the $\mathscr{R}\,\mathrm{BV}^k$-spaces via wavelet-like atomic decompositions, defining new Banach spaces via generalized Radon transforms, and alternative definitions for for vector-valued $\mathscr{R}\,\mathrm{BV}^k$-spaces different from the simple Cartesian product of scalar-valued $\mathscr{R}\,\mathrm{BV}^k$-spaces.

Chapter 2

Elements of Functional Analysis

In this chapter, we review the relevant background from functional analysis used throughout this dissertation. We will mostly be interested in function spaces of functions mapping $\mathbb{R}^d \to \mathbb{R}$ (or $\mathbb{C}$) and functions mapping $\mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$ (or $\mathbb{C}$), where

$$\mathbb{S}^{d-1} \coloneqq \left\{ \boldsymbol{x} \in \mathbb{R}^d : \ \|\boldsymbol{x}\|_2 = 1 \right\}$$

denotes the Euclidean sphere in $\mathbb{R}^d$. We use the term function space to mean a topological vector space whose elements are functions.

2.1 Spaces of Functions, Measures, and Distributions

L^p spaces. For $1 \leq p \leq \infty$, let $L^p(\mathbb{R}^d)$ denote the Lebesgue space on $\mathbb{R}^d$, which is a Banach space when equipped with the norm

$$\|f\|_{L^p} \coloneqq \begin{cases} \left(\int_{\mathbb{R}^d} |f(\boldsymbol{x})|^p \, \mathrm{d}\boldsymbol{x} \right)^{1/p}, & 1 \leq p < \infty, \\ \operatorname*{ess\,sup}_{\boldsymbol{x} \in \mathbb{R}^d} |f(\boldsymbol{x})|, & p = \infty. \end{cases}$$

The space $L^p(\mathbb{S}^{d-1} \times \mathbb{R})$ is defined analogously, except that the underlying measure is the product measure of the surface measure on $\mathbb{S}^{d-1}$ and the univariate Lebesgue measure on $\mathbb{R}$. The space L^p is a Banach space for $1 \leq p \leq \infty$ and a Hilbert space if and only if $p = 2$.

The spaces $\mathcal{S}$ and $\mathcal{S}'$. Let $\mathcal{S}(\mathbb{R}^d)$ denote the Schwartz space of smooth and rapidly decaying test functions on $\mathbb{R}^d$. These are functions $\varphi \in C^\infty(\mathbb{R}^d)$ such that

$$p_{\boldsymbol{\alpha},\boldsymbol{\beta}}(\varphi) \coloneqq \sup_{\boldsymbol{x} \in \mathbb{R}^d} \left| \boldsymbol{x}^{\boldsymbol{\alpha}} (\partial^{\boldsymbol{\beta}} \varphi)(\boldsymbol{x}) \right| < \infty, \quad \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}_0^d,$$

where $\boldsymbol{x}^{\boldsymbol{\alpha}} = x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ and $\partial^{\boldsymbol{\beta}} = \partial_{x_1}^{\beta_1} \cdots \partial_{x_d}^{\beta_d}$ is the usual multi-index notation, and $C^\infty(\mathbb{R}^d)$ denotes the space of infinitely differentiable functions on $\mathbb{R}^d$. We endow $\mathcal{S}(\mathbb{R}^d)$ with the topology induced by by the family of seminorms $\{p_{\boldsymbol{\alpha},\boldsymbol{\beta}}\}_{\boldsymbol{\alpha},\boldsymbol{\beta} \in \mathbb{N}_0^d}$, making it a Fréchet space (Rudin, 1991, Chapter 7). Let $\mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R})$ denote the Schwartz space of smooth and rapidly decaying test functions on $\mathbb{S}^{d-1} \times \mathbb{R}$, defined as $\mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R}) \coloneqq C^\infty(\mathbb{S}^{d-1}) \mathbin{\widehat{\otimes}} \mathcal{S}(\mathbb{R})$, where $\mathbin{\widehat{\otimes}}$ denotes the topological tensor product (Trèves, 1967, Chapter 43).

The continuous dual of $\mathcal{S}(\mathbb{R}^d)$ (resp. $\mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R})$) is the space of tempered distributions on $\mathbb{R}^d$ (resp. $\mathbb{S}^{d-1} \times \mathbb{R}$), denoted $\mathcal{S}'(\mathbb{R}^d)$ (resp. $\mathcal{S}'(\mathbb{S}^{d-1} \times \mathbb{R})$). This is the space of continuous linear functionals on $\mathcal{S}(\mathbb{R}^d)$ (resp. $\mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R})$) (Rudin, 1991, Chapter 7). In particular, a tempered distribution $u \in \mathcal{S}'(\mathbb{R}^d)$ defines a continuous linear functional on the space of Schwartz functions $\mathcal{S}(\mathbb{R}^d)$ via $u : \varphi \mapsto \langle u, \varphi \rangle$, where $\langle \cdot, \cdot \rangle$ denotes the *duality pairing* between $\mathcal{S}(\mathbb{R}^d)$ and $\mathcal{S}'(\mathbb{R}^d)$. The duality pairing between $\psi \in \mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R})$ and $v \in \mathcal{S}'(\mathbb{S}^{d-1} \times \mathbb{R})$ will be denoted by $[v, \psi]$. We abuse notation and also let $\langle \cdot, \cdot \rangle$ (resp. $[\cdot, \cdot]$) denote the pairing between any dual pair of spaces on $\mathbb{R}^d$ (resp. $\mathbb{S}^{d-1} \times \mathbb{R}$), where the exact pairing will be clear from context. We also use $\langle \cdot, \cdot \rangle$ and $[\cdot, \cdot]$ to denote the corresponding L^2-inner products.

The spaces C_0 and $\mathcal{M}$. Let $C_0(\mathbb{R}^d)$ denote the space of continuous functions on $\mathbb{R}^d$ vanishing at infinity. This space is a Banach space when equipped with the L^∞-norm. By the Riesz–Markov–Kakutani representation theorem, the continuous

dual of $C_0(\mathbb{R}^d)$ is the space $\mathcal{M}(\mathbb{R}^d)$ of finite Radon measures on $\mathbb{R}^d$, which is a Banach space when equipped with the norm

$$\|u\|_{\mathcal{M}} := \sup_{\substack{\varphi \in C_0(\mathbb{R}^d) \\ \|\varphi\|_{L^\infty}=1}} \langle u, \varphi \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the pairing between $C_0(\mathbb{R}^d)$ and $\mathcal{M}(\mathbb{R}^d)$. It is well-known that $C_0(\mathbb{R}^d) = \overline{(\mathcal{S}(\mathbb{R}^d), \|\cdot\|_{L^\infty})}$ (the closure of $\mathcal{S}(\mathbb{R}^d)$ with respect to $\|\cdot\|_{L^\infty}$). Therefore, we can alternatively define $\mathcal{M}(\mathbb{R}^d) = \big(C_0(\mathbb{R}^d)\big)'$ as

$$\mathcal{M}(\mathbb{R}^d) = \left\{ u \in \mathcal{S}'(\mathbb{R}^d) : \ \|u\|_{\mathcal{M}} = \sup_{\substack{\varphi \in \mathcal{S}(\mathbb{R}^d) \\ \|\varphi\|_{L^\infty}=1}} \langle u, \varphi \rangle < \infty \right\}.$$

The $\mathcal{M}$-norm is exactly the *total variation norm* in the sense of measures (Folland, 1999, Chapter 7). The definition in the above display allows us to view $\mathcal{M}(\mathbb{R}^d)$ as a subspace of $\mathcal{S}'(\mathbb{R}^d)$. The Banach space $(\mathcal{M}(\mathbb{R}^d), \|\cdot\|_{\mathcal{M}})$ can be viewed as a "generalization" of the Banach space $(L^1(\mathbb{R}^d), \|\cdot\|_{L^1})$. Indeed, this is due to the following three properties:

1. $L^1(\mathbb{R}^d) \subset \mathcal{M}(\mathbb{R}^d)$, where the containment is strict;
2. The shifted Dirac impulse $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, is not contained in $L^1(\mathbb{R}^d)$, but $\delta(\cdot - \boldsymbol{x}_0) \in \mathcal{M}(\mathbb{R}^d)$ with $\|\delta(\cdot - \boldsymbol{x}_0)\|_{\mathcal{M}} = 1$;
3. For every $f \in L^1(\mathbb{R}^d)$, we have that $\|f\|_{L^1} = \|f\|_{\mathcal{M}}$.

Working in this formalism allows us to work rigorously with (tempered) distributions such as the Dirac impulse, which is often overlooked in engineering textbooks (see, e.g., Feichtinger (2017); Unser (2020) for more details). The spaces $C_0(\mathbb{S}^{d-1} \times \mathbb{R})$ and $\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})$ are defined analogously, where the same properties hold.

2.2 Linear Operators

We will primarily be interested in linear operators mapping between functions mapping $\mathbb{R}^d \to \mathbb{R}$ (or $\mathbb{C}$) and functions mapping $\mathbb{R}^d \to \mathbb{R}$ (or $\mathbb{C}$) and linear operators mapping from functions mapping functions mapping $\mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$ (or $\mathbb{C}$) and functions mapping $\mathbb{R}^d \to \mathbb{R}$ (or $\mathbb{C}$). In particular, we will want to apply these operators to functions that are tempered distributions.

Definition 2.1. *Let* $\mathrm{L} : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ *be a continuous linear operator. The* adjoint *of* L *is the unique continuous linear operator* $\mathrm{L}^* : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ *such that*

$$\langle \mathrm{L}\{\varphi\}, \phi \rangle = \langle \mathrm{L}^*\{\phi\}, \varphi \rangle,$$

for all $\varphi, \phi \in \mathcal{S}(\mathbb{R}^d)$*, where the duality pairing* $\langle \cdot, \cdot \rangle$ *is the pairing between* $\mathcal{S}(\mathbb{R}^d)$ *and* $\mathcal{S}'(\mathbb{R}^d)$*.*

Definition 2.2. *Let* $\mathrm{T} : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{S}^{d-1} \times \mathbb{R})$ *be a continuous linear operator. The* adjoint *of* T *is the unique continuous linear operator* $\mathrm{T}^* : \mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R}) \to \mathcal{S}'(\mathbb{R}^d)$ *such that*

$$[\mathrm{T}\{\varphi\}, \psi] = \langle \mathrm{T}^*\{\psi\}, \varphi \rangle,$$

for all $\varphi \in \mathcal{S}(\mathbb{R}^d)$ *and* $\psi \in \mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R})$*, where the duality pairing* $\langle \cdot, \cdot \rangle$ *(resp.* $[\cdot, \cdot]$*) is the pairing between* $\mathcal{S}(\mathbb{R}^d)$ *and* $\mathcal{S}'(\mathbb{R}^d)$ *(resp.* $\mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R})$ *and* $\mathcal{S}'(\mathbb{S}^{d-1} \times \mathbb{R})$*).*

Remark 2.3. The definition of the adjoint of an operator mapping $\mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R}) \to \mathcal{S}'(\mathbb{R}^d)$ is analogous to Definition 2.2.

One can quickly verify that for both Definitions 2.1 and 2.2, the double adjoint of an operator is itself. Indeed, we illustrate this explicitly for a continuous linear operator $\mathrm{L} : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$. By Definition 2.1 we have for all $\varphi, \phi \in \mathcal{S}(\mathbb{R}^d)$ the following two equalities: (i) $\langle \mathrm{L}^*\{\phi\}, \varphi \rangle = \langle \mathrm{L}\{\varphi\}, \phi \rangle$; (ii) $\langle \mathrm{L}^*\{\phi\}, \varphi \rangle = \langle \mathrm{L}^{**}\{\varphi\}, \phi \rangle$. Subtracting these two equalities yields $\langle \mathrm{L}\{\varphi\}, \phi \rangle - \langle \mathrm{L}^{**}\{\varphi\}, \phi \rangle = 0$, i.e., for all $\varphi, \phi \in \mathcal{S}(\mathbb{R}^d)$, $\langle \mathrm{L}\{\varphi\} - \mathrm{L}^{**}\{\varphi\}, \phi \rangle = 0$. Therefore, $\mathrm{L}^{**}\{\varphi\} = \mathrm{L}\{\varphi\}$ for all $\varphi \in \mathcal{S}(\mathbb{R}^d)$.

The Schwartz kernel theorem. Linear operators can be completely characterized by their (Schwartz) kernel. This is summarized by the Schwartz kernel theorem.

Theorem 2.4 (Schwartz kernel theorem)**.** *Let* $\mathrm{L} : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$ *be a continuous linear operator. Then, there exists a unique tempered distribution* $h \in \mathcal{S}'(\mathbb{R}^d \times \mathbb{R}^d)$ *such that*

$$\langle \mathrm{L}\{\varphi\}, \phi \rangle = (h, \varphi \otimes \phi), \tag{2.1}$$

for all $\varphi, \phi \in \mathcal{S}(\mathbb{R}^d)$*, where* $(\cdot, \cdot)$ *denotes the pairing between* $\mathcal{S}(\mathbb{R}^d \times \mathbb{R}^d)$ *and* $\mathcal{S}'(\mathbb{R}^d \times \mathbb{R}^d)$ *and* $(\varphi \otimes \phi)(\boldsymbol{x}, \boldsymbol{y}) = \varphi(\boldsymbol{x})\phi(\boldsymbol{y})$ *is the tensor product between* φ *and* ϕ*.*

Remark 2.5. When $\mathrm{L}\{\varphi\}$ and h are both locally integrable functions, (2.1) can be rewritten as

$$\mathrm{L}\{\varphi\}(\boldsymbol{x}) = \int_{\mathbb{R}^d} h(\boldsymbol{x}, \boldsymbol{y})\varphi(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}.$$

When $\mathrm{L}\{\varphi\}$ and h are not both locally integral functions, we occasionally abuse notation and write $\mathrm{L}\{\varphi\}$ as in the above display.

Remark 2.6. When the operator L is shift-invariant, its kernel $h(\boldsymbol{x}, \boldsymbol{y})$ satisfies $h(\boldsymbol{x}, \boldsymbol{y}) = h(\boldsymbol{x} - \boldsymbol{y})$ for some $h \in \mathcal{S}'(\mathbb{R}^d)$. Then L is a convolution operator with $\mathrm{L}\{\varphi\} = h * \varphi$. In this case, $h = \mathrm{L}\{\delta\}$ is known as the *impulse response* of L.

The Schwartz kernel theorem is intimately linked to the *nuclearity* of $\mathcal{S}(\mathbb{R}^d)$ (Trèves, 1967, Chapters 50 and 51). While the form of the theorem in Theorem 2.4 can be proved using elementary techniques (see, e.g., Simon, 1971, Theorem 5), the result actually holds more generally on locally convex nuclear spaces via more advanced techniques (Grothendieck, 1955). For example, consider the following variant of the Schwartz kernel theorem.

Theorem 2.7. *Let* $\mathrm{T} : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{S}^{d-1} \times \mathbb{R})$ *be a continuous linear operator. Then, there exists a unique tempered distribution* $h \in \mathcal{S}'(\mathbb{R}^d \times \mathbb{S}^{d-1} \times \mathbb{R})$ *such that*

$$[\mathrm{T}\{\varphi\}, \psi] = (h, \varphi \otimes \psi),$$

for all $\varphi \in \mathcal{S}(\mathbb{R}^d)$ *and* $\psi \in \mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R})$*, where* $(\cdot, \cdot)$ *denotes the pairing between* $\mathcal{S}(\mathbb{R}^d \times \mathbb{S}^{d-1} \times \mathbb{R})$ *and* $\mathcal{S}'(\mathbb{R}^d \times \mathbb{S}^{d-1} \times \mathbb{R})$ *and* $(\varphi \otimes \psi)(\boldsymbol{x}, \boldsymbol{z}) = \varphi(\boldsymbol{x})\psi(\boldsymbol{z})$ *is the tensor product between* φ *and* ϕ*, where* $\boldsymbol{x} \in \mathbb{R}^d$ *and* $\boldsymbol{z} \in \mathbb{S}^{d-1} \times \mathbb{R}$*.*

Remark 2.8. Another variant of the Schwartz kernel theorem holds for continuous linear operators mapping $\mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R}) \to \mathcal{S}'(\mathbb{R}^d)$.

Extension of operators by duality. Continuous linear operators that map $\mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{R}^d)$ can be extended to map $\mathcal{S}'(\mathbb{R}^d) \to \mathcal{S}'(\mathbb{R}^d)$. Indeed, suppose that $\mathrm{L} : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{R}^d)$ and $\mathrm{L}^* : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{R}^d)$ are both continuous linear operators. Then, for $u \in \mathcal{S}'(\mathbb{R}^d)$, we *define* $\mathrm{L}\{u\}$ as the tempered distribution such that

$$\langle \mathrm{L}\{u\}, \varphi \rangle = \langle u, \mathrm{L}^*\{\varphi\} \rangle$$

for all $\varphi \in \mathcal{S}(\mathbb{R}^d)$. This technique can be applied more generally, e.g., for continuous linear operators mapping $\mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R})$ and continuous linear operators mapping $\mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R}) \to \mathcal{S}(\mathbb{R}^d)$.

2.3 Two Topologies of a Dual Banach Space

Given a Banach space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, we will be interested in two topologies for its continuous dual $\mathcal{X}'$, which is a Banach space when equipped with the (dual) norm

$$\|u\|_{\mathcal{X}'} \coloneqq \sup_{\substack{v \in \mathcal{X} \\ \|v\|_{\mathcal{X}} = 1}} \langle u, v \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the duality pairing between $\mathcal{X}$ and $\mathcal{X}'$.

Definition 2.9. *A sequence* $\{u_n\}_{n=1}^{\infty} \subset \mathcal{X}'$ *is said to converge to* f *in the* strong topology *if*

$$\lim_{n \to \infty} \|u_n - u\|_{\mathcal{X}'} = 0.$$

In other words, the strong topology of $\mathcal{X}'$ *is the topology induced by the norm* $\|\cdot\|_{\mathcal{X}'}$ *and is the usual topology implicitly assumed when working with the Banach space* $\mathcal{X}'$*.*

Definition 2.10. *A sequence* $\{u_n\}_{n=1}^{\infty} \subset \mathcal{X}'$ *is said to converge to* u *in the weak* topology if*

$$\lim_{n\to\infty} \langle u_n - u, v\rangle = 0,$$

for all $v \in \mathcal{X}$.

Remark 2.11. The weak* topology is *coarser* than the strong topology of $\mathcal{X}'$.

We will be interested in the continuity of linear functionals on dual Banach spaces with respect to the two topologies in Definitions 2.9 and 2.10. In particular, the space of all linear functionals on $\mathcal{X}'$ which are continuous with respect to the strong topology is (by definition) its continuous dual $\mathcal{X}''$. We refer to these linear functionals as being continuous on $\mathcal{X}'$.

Proposition 2.12 (Reed and Simon (1972, Theorem IV.20, pg. 114))**.** *The space of weak* continuous linear functionals on* $\mathcal{X}'$ *is the space* $\mathcal{X}$.

From Proposition 2.12 we see that weak* continuity is actually a stronger notion of continuity than the standard notion. Indeed, this is due to the fact that a Banach space $\mathcal{X}$ is isometrically isomorphic to a closed subspace of its bidual $\mathcal{X}''$ (Rudin, 1991). In particular, we can view $\mathcal{X} \overset{c.}{\hookrightarrow} \mathcal{X}''$ via the canonical embedding of a Banach space into its bidual. When $\mathcal{X}$ is a reflexive Banach space, the inclusion is actually an equality, i.e., $\mathcal{X} = \mathcal{X}''$, and therefore the space of continuous linear functionals on $\mathcal{X}'$ is the same as the space of weak* continuous linear functionals on $\mathcal{X}'$. On the other hand, when $\mathcal{X}$ is a non-reflexive Banach space, the inclusion $\mathcal{X} \subset \mathcal{X}''$ is strict. In this dissertation, we will mostly be interested in non-reflexive spaces, e.g., $\mathcal{X}' = \big(C_0(\mathbb{R}^d)\big)' = \mathcal{M}(\mathbb{R}^d)$.

The Banach–Alaoglu theorem. It is well-known that closed balls are not compact in infinite-dimensional spaces with respect to the topology induced by the norm, i.e., the strong topology of a dual space. The utility of working with the weak* topology is that closed balls are compact in the weak* topology by the Banach–Alaoglu theorem (Rudin, 1991, Chapter 3).

Theorem 2.13 (Banach–Alaoglu theorem)**.** *The closed unit ball*

$$B \coloneqq \{f \in \mathcal{X}' : \|f\|_{\mathcal{X}'} \leq 1\}$$

is weak compact.*

Remark 2.14. This result allows us to use compactness arguments to prove that solutions exist to certain variational problems.

2.4 Direct-Sum Decompositions and Projectors

There are two ways of working with direct-sums of topological vector spaces: (i) explicitly, via projectors; (ii) abstractly, via quotient spaces and equivalence classes. These two methods are equivalent whenever one can identify abstract quotient space as a *concrete subspace* of the original space. In other words, by selecting a concrete representer from each coset (which is an equivalence class). In this dissertation, we will work with direct-sums explicitly via projectors rather than abstractly.

Let $\mathcal{X}$ be a topological vector space. A continuous linear operator $\mathrm{P} : \mathcal{X} \to \mathcal{X}$ with the property that $\mathrm{P}^2 = \mathrm{P}$ on $\mathcal{X}$ is called a *projection operator* or a *projector* (Dunford and Schwartz, 1988, pg. 140). When $\mathcal{X}$ is a Fréchet space, then $\mathcal{U} \coloneqq \mathrm{P}(\mathcal{X})$ is a closed subspace of $\mathcal{X}$. In this case, P is the projector of $\mathcal{X}$ onto $\mathcal{U}$, i.e., $\mathrm{P} = \mathrm{Proj}_{\mathcal{U}}$, and we have the direct-sum decomposition $\mathcal{X} = \mathcal{U} \oplus \mathcal{V}$, where $\mathcal{V}$ is the null space of P. Said differently, the projector of $\mathcal{X}$ onto $\mathcal{V}$ is the *complementary projector* of P, i.e., $\mathrm{Proj}_{\mathcal{V}} = \mathrm{Id} - \mathrm{P}$, where Id is the identity operator. The space $\mathcal{U} = \mathrm{P}(\mathcal{X})$ is also a topological vector space with the topology induced by the topology of $\mathcal{X}$. Let $(\mathcal{X}, \mathcal{X}')$ be a dual pair of topological spaces, the $(\mathcal{U}, \mathcal{U}')$ is also a dual pair of topological spaces where $\mathcal{U}' = \mathrm{P}^*(\mathcal{X}')$, where $\mathrm{P}^* : \mathcal{X}' \to \mathcal{X}'$ is the dual (adjoint) projector.

2.5 The Fourier, Hilbert, and Radon Transforms

The Fourier transform. The Fourier transform of $\varphi \in \mathcal{S}(\mathbb{R}^d)$ is given by

$$\widehat{\varphi}(\boldsymbol{\omega}) = \mathscr{F}\{\varphi\}(\boldsymbol{\omega}) = \int_{\mathbb{R}^d} \varphi(\boldsymbol{x}) e^{-\mathrm{j}\boldsymbol{\omega}^\mathsf{T}\boldsymbol{x}} \, \mathrm{d}\boldsymbol{x}, \quad \boldsymbol{\omega} \in \mathbb{R}^d,$$

where $\mathrm{j}^2 = -1$. The Fourier transform $\mathscr{F} : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{R}^d)$ is a continuous linear bijection whose inverse $\mathscr{F}^{-1} : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{R}^d)$ is also continuous (Rudin, 1991, Chapter 7). The inverse Fourier transform of $\widehat{\varphi} \in \mathcal{S}(\mathbb{R}^d)$ is given by

$$\mathscr{F}^{-1}\{\widehat{\varphi}\}(\boldsymbol{x}) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \widehat{\varphi}(\boldsymbol{\omega}) e^{\mathrm{j}\boldsymbol{\omega}^\mathsf{T}\boldsymbol{x}} \, \mathrm{d}\boldsymbol{\omega}, \quad \boldsymbol{x} \in \mathbb{R}^d. \tag{2.2}$$

We can extend $\mathscr{F}$ and $\mathscr{F}^{-1}$ to act on $\mathcal{S}'(\mathbb{R}^d)$ by duality.

An important property of the Fourier transform is Plancherel's theorem, which states that $\mathscr{F} : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ is an isometry. More specifically, given $\varphi \in \mathcal{S}(\mathbb{R}^d)$, we have the equality

$$(2\pi)^d \|\varphi\|_{L^2}^2 = \|\widehat{\varphi}\|_{L^2}^2,$$

and the operator $\mathscr{F} : \mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{R}^d)$ admits a unique extension $\mathscr{F} : L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ since $L^2(\mathbb{R}^d) = \overline{(\mathcal{S}(\mathbb{R}^d), \|\cdot\|_{L^2})}$.

The Hilbert transform. The Hilbert transform of $\varphi \in \mathcal{S}(\mathbb{R})$ is given by

$$\mathscr{H}\{\varphi\}(x) = \frac{1}{\pi} \mathrm{p.v.} \int_{\mathbb{R}} \frac{f(x-y)}{y} \, \mathrm{d}y \coloneqq \frac{1}{\pi} \lim_{\varepsilon \to 0} \int_{|y|>\varepsilon} \frac{f(x-y)}{y} \, \mathrm{d}y, \quad x \in \mathbb{R},$$

where p.v. denotes that the integral is understood in the Cauchy principle value sense, as defined above. Although the Hilbert transform of a Schwartz function is not a Schwartz function, one may quickly verify that $\mathscr{H}$ maps $\mathcal{S}(\mathbb{R})$ to $L^2(\mathbb{R})$. In particular, for $\varphi \in \mathcal{S}(\mathbb{R})$, the Hilbert transform satisfies

$$\widehat{\mathscr{H}\varphi}(\omega) = -\mathrm{j}\,\mathrm{sgn}(\omega)\widehat{\varphi}(\omega), \tag{2.3}$$

where $\omega \mapsto -\mathrm{j}\,\mathrm{sgn}(\omega)$ is the *frequency response* (or Fourier symbol/multiplier) of $\mathscr{H}$, where the Fourier transform in the above display is understood as the Fourier transform of an $L^2(\mathbb{R})$ function, i.e., defined via density by Plancheral's theorem. From (2.3), we see that the Hilbert transform is skew-adjoint on $L^2(\mathbb{R})$, i.e., $\mathscr{H}^* = -\mathscr{H}$; in particular, $\mathscr{H}$ is unitary. Although the Hilbert transform *cannot* be extended to $\mathcal{S}'(\mathbb{R}^d)$ by duality, it can be extended to a large class of distributions, which suffices for our purposes (Pandey, 2011, Chapter 3).

The Radon transform. The Radon transform of $\varphi \in \mathcal{S}(\mathbb{R}^d)$ is given by

$$\mathscr{R}\{\varphi\}(\boldsymbol{\alpha}, t) = \int_{\mathbb{R}^d} \varphi(\boldsymbol{x})\delta(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x} - t)\,\mathrm{d}\boldsymbol{x}, \quad (\boldsymbol{\alpha}, t) \in \mathbb{S}^{d-1} \times \mathbb{R},$$

where δ is the univariate Dirac impulse. The Radon domain is the hypercylinder $\mathbb{S}^{d-1} \times \mathbb{R}$, with a *direction* variable $\boldsymbol{\alpha} \in \mathbb{S}^{d-1}$ and an *offset* variable $t \in \mathbb{R}$. Note that the Radon transform of φ evaluated at $(\boldsymbol{\alpha}, t)$ is precisely the integral of φ over the hyperplane given by

$$P_{(\boldsymbol{\alpha},t)} \coloneqq \big\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x} = t\big\}.$$

Since $P_{(\boldsymbol{\alpha},t)} = P_{(-\boldsymbol{\alpha},-t)}$, we see that the Radon transform is always an even function. The Radon transform maps $\mathcal{S}(\mathbb{R}^d)$ to a subspace of $\mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R})$. This subspace is characterized by the following range theorem for the Radon transform.

Theorem 2.15 (Ludwig (1966, Theorem 2.1))**.** *A function ψ is the Radon transform of a function $\varphi \in \mathcal{S}(\mathbb{R}^d)$ if and only if*

1. *$\psi \in \mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R})$;*
2. *ψ is even, i.e., $\psi(\boldsymbol{\alpha}, t) = \psi(-\boldsymbol{\alpha}, -t)$;*
3. *$\Psi_k(\boldsymbol{\alpha}) \coloneqq \int_{\mathbb{R}} \psi(\boldsymbol{\alpha}, t)t^k\,\mathrm{d}t$ is a homogeneous polynomial (in $\boldsymbol{\alpha}$) for all $k \in \mathbb{N}_0$.*

In other words, the range of the Radon transform $\mathcal{S}_{\mathscr{R}} \coloneqq \mathscr{R}\big(\mathcal{S}(\mathbb{R}^d)\big)$ is the subspace of $\mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R})$ satisfying the properties in Theorem 2.15. The conditions in Item 3 are often referred to as the moment (or Cavalieri) conditions of the Radon transform.

The Radon transform is invertible on $\mathcal{S}(\mathbb{R}^d)$ via the so-called *filtered backprojection operator*. The Radon transform itself is sometimes referred to as the *projection operator* since given $\varphi \in \mathcal{S}(\mathbb{R}^d)$, for a fixed direction $\boldsymbol{\alpha}_0 \in \mathbb{S}^{d-1}$, the function $\mathscr{R}\{\varphi\}(\boldsymbol{\alpha}_0, \cdot)$ is a univariate function which corresponds to the projection[1] of φ in the direction specified by $\boldsymbol{\alpha}_0$. The adjoint of the Radon transform is the so-called *backprojection operator* and is given by

$$\mathscr{R}^*\{\psi\}(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}} \psi(\boldsymbol{\alpha}, \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x}) \, \mathrm{d}\sigma(\boldsymbol{\alpha}),$$

for sufficiently nice functions $\psi : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$, where σ denotes the surface measure on $\mathbb{S}^{d-1}$. Given a function $\varphi \in \mathcal{S}(\mathbb{R}^d)$, a calculation shows that

$$\widehat{\mathscr{R}^*\mathscr{R}\varphi}(\boldsymbol{\omega}) = \frac{\widehat{\varphi}(\boldsymbol{\omega})}{c_d \|\boldsymbol{\omega}\|_2^{d-1}},$$

where $c_d \coloneqq 1/(2(2\pi)^{d-1})$. In other words, the result of applying the projection followed by the backprojection to φ results in a "blurring" (attenuation of high frequencies) of φ. We can "deblur" the projected backprojection by applying a *ramp filter* to amplify high frequencies. The exact (spatial domain) filter is given by $(\mathscr{R}^*\mathscr{R})^{-1}$ whose frequency response is

$$\boldsymbol{\omega} \mapsto c_d \|\boldsymbol{\omega}\|_2^{d-1}. \tag{2.4}$$

This is realized by the operator

$$(\mathscr{R}^*\mathscr{R})^{-1} = c_d(-\Delta)^{\frac{d-1}{2}}, \tag{2.5}$$

where $\Delta = \partial_{x_1}^2 + \cdots + \partial_{x_d}^2$ is the d-dimensional Laplacian operator. Therefore, for $\varphi \in \mathcal{S}(\mathbb{R}^d)$, we have the following inversion formula[2] for the Radon transform:

$$c_d(-\Delta)^{\frac{d-1}{2}}\mathscr{R}^*\mathscr{R}\varphi = \varphi.$$

[1] Technically speaking, this is not truly projection.

[2] This inversion formula also holds for $\varphi \in L^1(\mathbb{R}^d)$.

The operator $c_d(-\Delta)^{\frac{d-1}{2}}\mathscr{R}^*$ is known as the *filtered backprojection operator*.

Since the frequency response in (2.4) is a radial function, by the *intertwining properties* of the Radon transform (Helgason, 2011, Lemma 2.1), the filtering can also be carried out in the Radon domain via the filter

$$\mathrm{K}^{d-1} \coloneqq c_d(-\partial_t^2)^{\frac{d-1}{2}} = \begin{cases} c_d(-1)^{\frac{d-1}{2}}\partial_t^{d-1}, & d \text{ is odd} \\ c_d(-1)^{\frac{d-2}{2}}\,\mathscr{H}_t\,\partial_t^{d-1}, & d \text{ is even,} \end{cases}$$

where $\mathscr{H}_t$ denotes the Hilbert transform with respect to the t variable. The frequency response of this operator is

$$\widehat{K}^{d-1}(\omega) = c_d|\omega|^{d-1},$$

where the Fourier transform is the univariate Fourier transform with respect to $t \to \omega$. Due to the Hilbert transform in K^{d-1} that arises when d is even, we see that when d is even, K^{d-1} is a *global operator*. When applied to sufficiently nice functions, we have the equality

$$c_d(-\Delta)^{\frac{d-1}{2}}\mathscr{R}^* = \mathscr{R}^*\,\mathrm{K}^{d-1}\,.$$

We summarize this in the following theorem regarding the continuity and invertibility of the Radon transform on $\mathcal{S}(\mathbb{R}^d)$.

Theorem 2.16 (see, e.g., Ludwig (1966); Helgason (2011))**.** *The operator $\mathscr{R}$ continuously maps $\mathcal{S}(\mathbb{R}^d) \to \mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R})$. Moreover,*

$$\mathscr{R}^*\,\mathrm{K}^{d-1}\,\mathscr{R} = c_d(-\Delta)^{\frac{d-1}{2}}\mathscr{R}^*\mathscr{R} = c_d\mathscr{R}^*\mathscr{R}(-\Delta)^{\frac{d-1}{2}} = \mathrm{Id}$$

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

Just like the Fourier transform, the Radon transform also admits a kind of Plancheral's theorem (Ludwig, 1966, Theorem 1.3). In particular, it states that $\mathrm{K}^{\frac{d-1}{2}}\mathscr{R} : L^2(\mathbb{R}^d) \to L^2(\mathbb{S}^{d-1} \times \mathbb{R})$ is an isometry, where $\mathrm{K}^{\frac{d-1}{2}}$ is defined via the frequency response

$$\widehat{K}^{\frac{d-1}{2}}(\omega) = \sqrt{c_d}\,|\omega|^{\frac{d-1}{2}}.$$

Indeed, given $f \in \mathcal{S}(\mathbb{R}^d)$ we have

$$\|\mathrm{K}^{\frac{d-1}{2}} \mathscr{R} f\|_{L^2}^2 = \left[\mathrm{K}^{\frac{d-1}{2}} \mathscr{R} f, \mathrm{K}^{\frac{d-1}{2}} \mathscr{R} f\right] = \left\langle f, \mathscr{R}^* \mathrm{K}^{\frac{d-1}{2}} \mathscr{R} f\right\rangle = \langle f, f\rangle = \|f\|_{L^2}^2.$$

The operator $\mathrm{K}^{\frac{d-1}{2}} \mathscr{R}$ admits a unique extension $\mathrm{K}^{\frac{d-1}{2}} \mathscr{R} : L^2(\mathbb{R}^d) \to L^2(\mathbb{S}^{d-1} \times \mathbb{R})$ since $L^2(\mathbb{R}^d) = \overline{(\mathcal{S}(\mathbb{R}^d), \|\cdot\|_{L^2})}$. We refer to this operator as the *half-filtered projection operator*. Moreover, this operator is invertible on $L^2_{\mathrm{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$, the subspace of even functions in $L^2(\mathbb{S}^{d-1} \times \mathbb{R})$, by its adjoint operator. Indeed, we have that $\mathrm{K}^{\frac{d-1}{2}} \mathscr{R} : L^2(\mathbb{R}^d) \to L^2_{\mathrm{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$ with inverse given by $\mathscr{R}^* \mathrm{K}^{\frac{d-1}{2}} : L^2_{\mathrm{even}}(\mathbb{S}^{d-1} \times \mathbb{R}) \to L^2(\mathbb{R}^d)$.

Unlike the Fourier transform, extending the Radon transform to (tempered) distributions is a delicate matter. Indeed, the naïve approach to define the operators $\mathscr{R}$, $\mathrm{K}^{d-1} \mathscr{R}$, and $\mathscr{R}^*$ would be via duality in a similar manner used to extend the Fourier transform to $\mathcal{S}'(\mathbb{R}^d)$. This is summarized in the following definition.

Definition 2.17 (Unser (2022b, Definition 4))**.** *The distribution $g \in \mathcal{S}'(\mathbb{S}^{d-1} \times \mathbb{R})$ is the* formal Radon transform *(or* formal projection*) of the distribution $f \in \mathcal{S}'(\mathbb{S}^{d-1} \times \mathbb{R})$ if*

$$[g, \psi] = \langle f, \mathscr{R}^*\{\psi\}\rangle, \tag{2.6}$$

for all $\psi \in \mathrm{K}^{d-1} \mathscr{R}\big(\mathcal{S}(\mathbb{R}^d)\big)$. Likewise, $g \in \mathcal{S}'(\mathbb{S}^{d-1} \times \mathbb{R})$ is a formal filtered projection *of $f \in \mathcal{S}'(\mathbb{R}^d)$ if*

$$[g, \psi] = \big\langle f, \mathscr{R}^* \mathrm{K}^{d-1}\{\psi\}\big\rangle, \tag{2.7}$$

for all $\psi \in \mathscr{R}\big(\mathcal{S}(\mathbb{R}^d)\big)$. Finally, $f \in \mathcal{S}'(\mathbb{R}^d)$ is the backprojection *of $g \in \mathcal{S}'(\mathbb{S}^{d-1} \times \mathbb{R})$ if*

$$\langle f, \varphi\rangle = [g, \mathscr{R}\{\varphi\}], \tag{2.8}$$

for all $\varphi \in \mathcal{S}(\mathbb{R}^d)$.

The issue that arises with the formal definitions in Definition 2.17 is that the definitions are not unique. In particular, for (2.6) and (2.7), there are infinitely many distributions $g \in \mathcal{S}'(\mathbb{S}^{d-1} \times \mathbb{R})$ that satisfy (2.6) or (2.7). On the other hand, the definition in (2.8) does provide a unique definition and therefore we *did not* refer

to f as the formal backprojection of g. We refer the reader to Unser (2022b) for more details. The fundamental issue boils down to the fact that the null space of the operator $\mathscr{R}^* : \mathcal{S}'(\mathbb{S}^{d-1} \times \mathbb{R}) \to \mathcal{S}'(\mathbb{R}^d)$ contains many exotic functions (Ludwig, 1966, Theorem 4.2). To this end, we can understand the Radon transform and related operators for a variety of distributions by working with the so-called *Radon-compatible Banach spaces* of Unser (2022b).

Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ be a Banach space such that $\mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R}) \overset{\text{d.}}{\hookrightarrow} \mathcal{X} \overset{\text{d.}}{\hookrightarrow} \mathcal{S}'(\mathbb{S}^{d-1} \times \mathbb{R})$, where $\overset{\text{d.}}{\hookrightarrow}$ denotes a dense embedding. Then, define the dual pair of Radon-compatible Banach spaces as $(\mathcal{X}_{\mathscr{R}} \coloneqq \overline{(\mathcal{S}_{\mathscr{R}}, \|\cdot\|_{\mathcal{X}})}, \mathcal{X}'_{\mathscr{R}})$. This dual pair satisfies many important properties, which we summarize in the following proposition.

Proposition 2.18 (Unser (2022b, Theorem 7))**.** *If there exists a complementary Banach space $\mathcal{X}^{\mathsf{c}}_{\mathscr{R}}$ such that $\mathcal{X} = \mathcal{X}_{\mathscr{R}} \oplus \mathcal{X}^{\mathsf{c}}_{\mathscr{R}}$, then*

1. *The dual space is decomposable as $\mathcal{X}' = \mathcal{X}'_{\mathscr{R}} \oplus (\mathcal{X}^{\mathsf{c}}_{\mathscr{R}})'$.*
2. *The complement space $\mathcal{X}^{\mathsf{c}}_{\mathscr{R}}$ is the null space of $\mathscr{R}^* \, \mathrm{K}^{d-1} : \mathcal{X} \to \mathscr{R}^* \, \mathrm{K}^{d-1}(\mathcal{X}_{\mathscr{R}}) =: \mathcal{Y}$.*
3. *The dual complement space $(\mathcal{X}^{\mathsf{c}}_{\mathscr{R}})'$ is the null space of $\mathscr{R}^* : \mathcal{X}' \to \mathscr{R}^*(\mathcal{X}'_{\mathscr{R}}) = \mathcal{Y}'$.*
4. *$\mathrm{P}_{\mathscr{R}} \coloneqq \mathscr{R}\mathscr{R}^* \, \mathrm{K}^{d-1} : \mathcal{X} \to \mathcal{X}_{\mathscr{R}}$ and $\mathrm{P}^*_{\mathscr{R}} = \mathrm{K}^{d-1} \, \mathscr{R}\mathscr{R}^* : \mathcal{X}' \to \mathcal{X}'_{\mathscr{R}}$ form a dual pair of projectors with $\mathrm{P}_{\mathscr{R}}(\mathcal{X}) = \mathcal{X}_{\mathscr{R}}$ and $\mathrm{P}^*_{\mathscr{R}}(\mathcal{X}') = \mathcal{X}'_{\mathscr{R}}$.*

This proposition allows us to have a unique definition of the Radon transform of functions that live in the Banach space $\mathcal{X}'$ as above. Of particular interest is the distributional definition of the Radon transform ridge distributions.

Theorem 2.19 (Unser (2022b, Proposition 10 & Corollary 11))**.** *Let $(\boldsymbol{\alpha}_0, t_0) \in \mathbb{S}^{d-1} \times \mathbb{R}$ and let $r \in \mathcal{S}'(\mathbb{R})$. Define the ridge distribution*

$$r_{(\boldsymbol{\alpha}_0, t_0)}(\boldsymbol{x}) \coloneqq r(\boldsymbol{\alpha}_0^{\mathsf{T}} \boldsymbol{x} - t_0).$$

Furthermore, suppose that $\delta(\cdot - \boldsymbol{\alpha}_0)\, r(\cdot - t_0) \in \mathcal{X}'$*, where* $(\mathcal{X}, \mathcal{X}')$ *is a dual pair of Banach spaces as above. Then,*

$$\begin{aligned} \mathrm{K}^{d-1}\, \mathscr{R}\big\{r_{(\boldsymbol{\alpha}_0, t_0)}\big\} &= \mathrm{P}^*_{\mathscr{R}}\{\delta(\cdot - \boldsymbol{\alpha}_0)\, r(\cdot - t_0)\} \\ \mathscr{R}\big\{r_{(\boldsymbol{\alpha}_0, t_0)}\big\} &= \mathrm{P}^*_{\mathscr{R}}\{\delta(\cdot - \boldsymbol{\alpha}_0)\, (q_{d-1} * r)(\cdot - t_0)\}, \end{aligned}$$

where $q_{d-1}(t) = c_d\, \mathscr{F}^{-1}_{\omega}\Big\{1/|\omega|^{d-1}\Big\}(t)$ *is the univariate impulse response of the Radon domain inverse filtering operator* $(\mathrm{K}^{d-1})^{-1}$*. In particular, when* $\mathcal{X}_{\mathscr{R}} = \mathcal{X}_{\mathrm{even}}$*, the subspace of even functions in* $\mathcal{X}$*, then* $\mathrm{P}^*_{\mathscr{R}} = \mathrm{P}_{\mathrm{even}}$*, the even projector, defined as*

$$\mathrm{P}_{\mathrm{even}}\{f\} \coloneqq \frac{f + f^{\vee}}{2},$$

where $f^{\vee}(\boldsymbol{z}) = f(-\boldsymbol{z})$ *is the reflection of* f *and* $\boldsymbol{z} = (\boldsymbol{\alpha}, t) \in \mathbb{S}^{d-1} \times \mathbb{R}$*.*

Another important result about the Radon transform is regarding its connection with the Fourier transform via the so-called *Fourier slice theorem*. In particular, for any $f \in \mathcal{S}'(\mathbb{R}^d)$, the Fourier slice theorem states that

$$\widehat{\mathscr{R}\{f\}}(\boldsymbol{\alpha}, \omega) = \widehat{f}(\omega \boldsymbol{\alpha}),$$

where the Fourier transform on the left-hand side is the univariate Fourier transform with respect to $t \to \omega$ and the Fourier transform on the right-hand side is the usual multivariate Fourier transform of f. We refer the reader to Ramm and Katsevich (1996, Chapter 10) for the version of the Fourier slice theorem that applies to $f \in \mathcal{S}'(\mathbb{R}^d)$.

Chapter 3

Representer Theorems for Sparse Ridge Splines

Shallow neural networks are superpositions of *ridge functions*. A ridge function is any function mapping $\mathbb{R}^d \to \mathbb{R}$ that can be written in the form

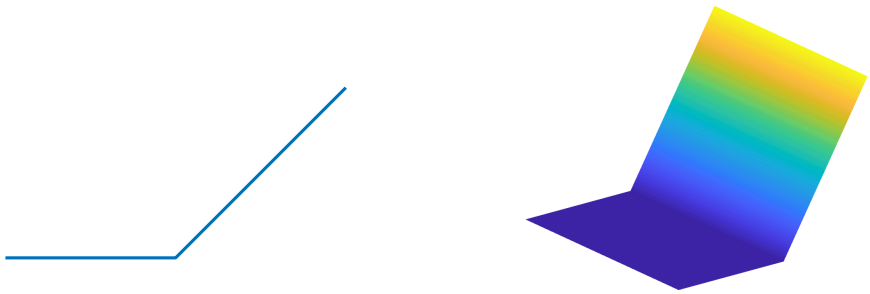
$$\boldsymbol{x} \mapsto r(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x}),$$

where $r : \mathbb{R} \to \mathbb{R}$ is referred to as the *ridge profile* and $\boldsymbol{\alpha} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ is referred to as the *ridge direction*. A ridge function is, in essense, a univariate function since it is constant along the hyperplanes $\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x} = c$, where $c \in \mathbb{R}$ (Pinkus, 2015). We illustrate a ridge function in Figure 3.1. Ridge functions are ubiquitous in mathematics, science, and engineering. For example,

- plane waves are time-varying functions of the form

  $$(\boldsymbol{x}, t) \mapsto r_t(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x}), \quad (\boldsymbol{x}, t) \in \mathbb{R}^d \times \mathbb{R},$$

  where $\boldsymbol{\alpha} \in \mathbb{S}^{d-1}$, which arise as solutions to many partial differential equations (PDEs), e.g., the wave equation (John, 1981). Plane waves are ridge functions with a *time-varying* profile $r_t : \mathbb{R} \to \mathbb{R}$, where $t \in \mathbb{R}$, and unit-norm direction $\boldsymbol{\alpha} \in \mathbb{S}^{d-1}$.



(a) $\mathrm{ReLU}(x) = x_+$.

(b) A ReLU ridge function.

Figure 3.1: The rectified linear unit (ReLU) and a ReLU ridge function.

- the complex exponential $\boldsymbol{x} \mapsto e^{\mathrm{j}\boldsymbol{\omega}^\mathsf{T}\boldsymbol{x}}$, where $\boldsymbol{\omega} \in \mathbb{R}^d$, is the fundamental building block for representing a function as a superposition of pure frequencies via the Fourier inversion formula in (2.2). These building blocks are precisely ridge functions with profiles given by $e^{\mathrm{j}(\cdot)}$ and direction given by the frequency variable $\boldsymbol{\omega} \in \mathbb{R}^d$.

- shallow neural networks are superpositions of ridge functions of the form

$$\boldsymbol{x} \mapsto \sum_{n=1}^{N} v_n \rho(\boldsymbol{w}_n^\mathsf{T}\boldsymbol{x} - b_n), \quad \boldsymbol{x} \in \mathbb{R}^d,$$

  where we are using the same notation as in (1.12). Each term in this superposition is a ridge function with profile given by the shifted activation function $\rho(\cdot - b_n)$ and direction given by the weight vector $\boldsymbol{w}_n \in \mathbb{R}^d$.

Ridge functions are intimately tied to the Radon transform. This observation goes back to classical work regarding representing solutions to PDEs as superpositions of plane waves, in which the PDEs are analyzed in the Radon domain (John, 1981; Evans, 2010). The term "ridge function" is rather modern and was coined by Logan and Shepp (1975) in their seminal work of computerized tomography (CT), in which images are reconstructed from their Radon transform via ridge functions. Moreover, a precursor to sparse signal approximation was the theory of ridgelets, which are a wavelet-like representation system where the atoms are ridge functions (Murata,

1996; Rubin, 1998; Candès, 1998, 1999). In fact, the continuous ridgelet transform is a univariate wavelet transform in the offset variable of the Radon domain (Candès, 1998, 1999; Kostadinova et al., 2014; Sonoda and Murata, 2017).

Recently, the connection between shallow ReLU networks and the Radon transform was recently exploited by Ongie et al. (2020a), although similar results exist in the case of unit step activation functions in Kurková et al. (1997). In Ongie et al. (2020a), the authors propose an operator that *sparsifies* ReLU neurons. In particular, implicit in the calculation in Ongie et al. (2020a, Example 1) is that

$$\partial_t^2 \, \mathrm{K}^{d-1} \mathscr{R}\big\{\rho_2(\boldsymbol{\alpha}_0^\mathsf{T}(\cdot) - t_0)\big\}(\boldsymbol{\alpha}, t) = \frac{\delta(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)\delta(t - t_0) + \delta(\boldsymbol{\alpha} + \boldsymbol{\alpha}_0)\delta(t + t_0)}{2}, \quad (3.1)$$

where we recall that $\mathrm{K}^{d-1}\mathscr{R}$ is the filtered projection operator, ρ_2 is any Green's function of D^2 (e.g., the ReLU), $(\boldsymbol{\alpha}_0, t_0) \in \mathbb{S}^{d-1} \times \mathbb{R}$, and δ is the Dirac impulse in the appropriate domain. The reason that (3.1) is an even impulse is due to the symmetries of the Radon domain. The property in (3.1) is analogous to the property that

$$\mathrm{D}^2\{\rho_2(\cdot - t_0)\}(t) = \delta(t - t_0),$$

which gave rise to the definition of the linear spline as in Definition 1.3. Moreover, the largest space of functions in which the seminorm $f \mapsto \|\mathrm{D}^2 f\|_{\mathcal{M}}$ is finite, defines $\mathrm{BV}^2(\mathbb{R})$, the native space for locally adaptive linear splines. To this end, we propose and study the family of function spaces defined by the seminorm $f \mapsto \|\partial_t^k \, \mathrm{K}^{d-1} \mathscr{R} f\|_{\mathcal{M}}$, where $k \in \mathbb{N}$. These seminorms are, in particular, total variation seminorms in the (filtered) Radon domain.

In this chapter, we prove several properties about these spaces, including that they are non-reflexive Banach spaces. We derive a *representer theorem* for these Banach spaces, showing that functions that are realizable by the sum of a shallow neural network and a polynomial term are universal solutions to variational inverse problems with total variation regularization in the Radon domain. These functions can be viewed as multivariate generalizations of splines and so we refer to these functions as *ridge splines*, emphasizing that they are superpositions of ridge functions. Finally, we

discuss applications of this representer theorem to learning with both shallow and deep neural networks.

3.1 Representer Theorems Beyond Hilbert Spaces

In this section, we review the relevant background and historical remarks about representer theorems. As discussed in Chapter 1, Section 1.3, the notion of a *representer theorem* is a fundamental result regarding kernel methods. In particular, let $\mathcal{H}$ be any real-valued Hilbert space on $\mathbb{R}^d$ and consider the data set $\{(\boldsymbol{x}_m, y_m)\}_{m=1}^M \subset \mathbb{R}^d \times \mathbb{R}$. The RKHS representer theorem considers the variational problem

$$f_{\text{RKHS}} = \arg\min_{f \in \mathcal{H}} \sum_{m=1}^{M} \ell(y_m, f(\boldsymbol{x}_m)) + \lambda \|f\|_{\mathcal{H}}^2, \tag{3.2}$$

where $\ell(\cdot, \cdot)$ is a convex and lower semicontinuous (in its second argument) loss function and $\lambda > 0$ is an adjustable hyperparameter. The representer theorem then states that the solution f_{RKHS} is unique and $f_{\text{RKHS}} \in \text{span}\{k(\cdot, \boldsymbol{x}_m)\}_{m=1}^M$, where $k(\cdot, \cdot)$ is the reproducing kernel of $\mathcal{H}$. Kernel methods (even before the term "kernel methods" was coined) have received much success dating all the way back to the 1960s, especially due to the tight connections between kernels, reproducing kernel Hilbert spaces, and splines (de Boor and Lynch, 1966; Micchelli, 1984; Wahba, 1990).

Recently, the term "representer theorem" has started being used for general problems of convex regularization (Boyer et al., 2019; Bredies and Carioni, 2020; Unser, 2021; Unser and Aziznejad, 2022) as a way to designate a parametric formulation of solutions to a data-fitting variational problem, ideally being a linear combination from some dictionary of atoms. This has allowed more general problems to be considered than ones like (3.2), which are restricted to regularizers which are Hilbertian (semi)norms. The main utility of these more general representer theorems arises in understanding *sparsity-promoting* regularizers such as the ℓ^1-norm or its continuous-domain analogue, the $\mathcal{M}$-norm, of which the structural properties of the solutions are still not completely understood, though a theory is emerging. The generality

of these kinds of representer theorems have been especially useful in some of the recent developments of *reproducing kernel Banach spaces* (Zhang et al., 2009; Xu and Ye, 2019), an infinite-dimensional theory of compressed sensing (Adcock and Hansen, 2016; Adcock et al., 2017), as well as other inverse problems set in the continuous-domain (Bredies and Pikkarainen, 2013).

The earliest instance of a representer theorem is, perhaps, due to Zuhovickiĭ (1948), where the classical problem of *Radon measure recovery* is studied. In the Dirac recovery literature, this problem is also referred to as the *Beurling LASSO* problem (De Castro and Gamboa, 2012) and can be viewed as the proper continuous-domain analogue of the finite-dimensional compressed sensing problem. This problem is posed over the non-reflexive Banach space of finite Radon measures and studies the variational problem

$$\min_{u \in \mathcal{M}(\mathbb{R}^d)} \|u\|_{\mathcal{M}} \quad \text{s.t.} \quad \mathrm{H}\{u\} = \boldsymbol{z} \in \mathbb{R}^M,$$

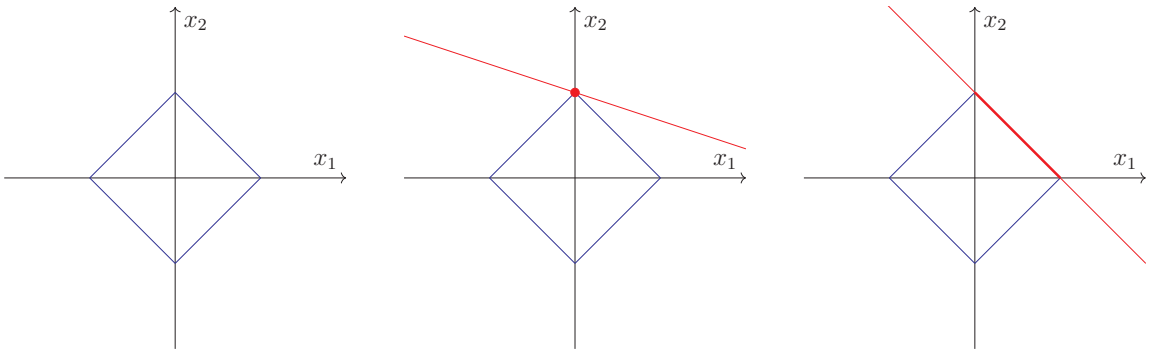
where $\mathrm{H} : \mathcal{M}(\mathbb{R}^d) \to \mathbb{R}^M$ is (component-wise) weak* continuous on $\mathcal{M}(\mathbb{R}^d)$. Zuhovickiĭ (1948) derives a representer theorem for this problem (long before the term "representer theorem" was coined). In particular, the solution set to this variational problem is nonempty, convex, and weak* compact and the extreme points are given by superpositions of Dirac impulses of the form

$$\boldsymbol{x} \mapsto \sum_{n=1}^{N} a_n \delta(\cdot - \boldsymbol{t}_n),$$

where $\{a_n\}_{n=1}^N \subset \mathbb{R} \setminus \{0\}$, $\{\boldsymbol{t}_n\}_{n=1}^N \subset \mathbb{R}^d$, and $N \leq M$. The key idea behind this representer theorem is that the extreme points of the unit ball of $\mathcal{M}(\mathbb{R}^d)$ are the Dirac impulses $\pm\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$. To see why the extreme points of the solution set take the form of a sparse superposition of the extreme points of the unit ball, consider the following compressed sensing (ℓ^1-norm minimization) problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^N} \|\boldsymbol{x}\|_1 \quad \text{s.t.} \quad \mathbf{H}\boldsymbol{x} = \boldsymbol{z}, \tag{3.3}$$

where the operator $\mathbf{H} \in \mathbb{R}^{M \times N}$ is now a matrix. We illustrate this intuition in Figure 3.2. From Figure 3.2(a), we see that the extreme points of the unit ℓ^1-ball are the Kronecker impulses $\pm\delta[\cdot - k]$, $k = 1, \ldots, N$. In Figure 3.2(b) we illustrate a scenario where the solution to (3.3) is unique in which case it is a constant scaling of a single Kronecker impulse. In Figure 3.2(b) we illustrate a scenario where the solution to (3.3) is nonunique in which case the extreme points of the solution set take the form of a constant scaling of a single Kroncker impulse, and the full solution set is the convex hull of these extreme points.



(a) (b) (c)

Figure 3.2: Illustration of the compressed sensing optimization problem when $N = 2$ and $M = 1$. The blue diamond denotes the ℓ^1-ball. The red lines denote the signals $\boldsymbol{x} \in \mathbb{R}^N$ consistent with the measurements $\mathbf{H}\boldsymbol{x} = \boldsymbol{z}$. The solutions to the compressed sensing problem in (3.3) are highlighted. In (a) we illustrate the ℓ^1-ball. In (b) we illustrate the situation of a unique solution. In (c) we illustrate the situation of nonunique solutions.

In the continuous-domain formulation of this problem, the same phenomenon arises. To see why the extreme points of the unit ball in $\mathcal{M}(\mathbb{R}^d)$ take the form of $\pm\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, we reproduce the following standard proof adapted from Bredies and Carioni (2020, Proposition 4.1).

Proposition 3.1 (see Bredies and Carioni (2020, Proposition 4.1))**.** *The extreme points of the unit ball*

$$B_{\mathcal{M}(\mathbb{R}^d)} = \left\{ u \in \mathcal{M}(\mathbb{R}^d) : \ \|u\|_{\mathcal{M}} \leq 1 \right\}$$

take the form $\pm\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, *where* δ *is the Dirac impulse.*

Proof. In this proof we view elements of $\mathcal{M}(\mathbb{R}^d)$ as measures (instead of tempered distributions). We first show that $\sigma\delta_{\boldsymbol{x}}$, $\sigma \in \{-1, +1\}$, where $\delta_{\boldsymbol{x}}$ denotes the Dirac measure supported at $\boldsymbol{x} \in \mathbb{R}^d$, is an extreme point. Suppose

$$\sigma\delta_{\boldsymbol{x}} = tu_1 + (1-t)u_2, \tag{3.4}$$

for some $u_1, u_2 \in B_{\mathcal{M}(\mathbb{R}^d)}$ and some $t \in (0,1)$. We want to show that $u_1 = u_2 = \sigma\delta_{\boldsymbol{x}}$. Let $|\cdot|$ denote the total variation measure. Clearly $|u_1|$ and $|u_2|$ must be probability measures. Indeed, if not, then $1 = \|\sigma\delta_{\boldsymbol{x}}\|_{\mathcal{M}} \leq t\|u_1\|_{\mathcal{M}} + (1-t)\|u_2\|_{\mathcal{M}} < t + (1-t) = 1$, a contradiction. Next,

$$\delta_{\boldsymbol{x}} = |\sigma\delta_{\boldsymbol{x}}| \leq t|u_1| + (1-t)|u_2| =: u,$$

where the inequality is understood pointwise. Since $|u_1|$ and $|u_2|$ are probability measures, u must be a probabilty measure. The inequality in the above display must be an equality. Indeed, given a measurable set E, if $\boldsymbol{x} \in E$ then

$$1 = \delta_{\boldsymbol{x}}(E) \leq u(E) \leq 1.$$

On the other hand, if $\boldsymbol{x} \notin E$ then

$$1 = \delta_{\boldsymbol{x}}(\mathbb{R}^d \setminus E) \leq u(\mathbb{R}^d \setminus E) \leq 1$$

and so $u(\mathbb{R}^d \setminus E) = 1$. Therefore, $u(E) = u(\mathbb{R}^d) - u(\mathbb{R}^d \setminus E) = 1 - 1 = 0$. Hence $u = \delta_{\boldsymbol{x}}$. This implies $|u_1| = |u_2| = \delta_{\boldsymbol{x}}$ which implies $u_1 = \sigma\delta_{\boldsymbol{x}}$ and $u_2 = \sigma\delta_{\boldsymbol{x}}$. Therefore, $\sigma\delta_{\boldsymbol{x}}$ is an extreme point of $B_{\mathcal{M}(\mathbb{R}^d)}$.

We now show that these are the only extreme points. We proceed by contradiction. Let u be an extreme point that is not a Dirac measure. Then, $\|u\|_{\mathcal{M}} = 1$. For any measurable set A, let $u \llcorner E(A) \coloneqq u(E \cap A)$ denote the restriction of u to the

measurable set E. For every measurable E it is always true that

$$u = u \llcorner E + u \llcorner (\mathbb{R}^d \setminus E) = \underbrace{|u|(E)}_{=: t} \underbrace{\left[\frac{u \llcorner E}{|u|(E)}\right]}_{=: u_1} + |u|(\mathbb{R}^d \setminus E) \underbrace{\left[\frac{u \llcorner (\mathbb{R}^d \setminus E)}{|u|(\mathbb{R}^d \setminus E)}\right]}_{=: u_2}.$$

Since the above display holds for every measurable E combined with the fact that u is not a Dirac impulse, we can always find an E such that $u \neq u_1$ and $u \neq u_2$. Thus, the above display implies $u = tu_1 + (1-t)u_2$ with $u \neq u_1$ and $u \neq u_2$, a contradiction. Therefore, u cannot be an extreme point. □

It turns out that many variational problems that hinge on sparsity-promoting regularization with the $\mathcal{M}$-norm can be reduced to the problem of Radon measure recovery (e.g., the locally adaptive spline problems). This boils down to establishing an isomorphism between the native space of the variational problem and a space of finite Radon measures.

In the last few years, the neural network community has also been interested in sparsity-promoting regularization for neural networks. In particular, many authors consider the problem of learning with *continuum-width* shallow neural networks by considering functions that take the form of a neuronal activation function integrated against a finite Radon measure (see, e.g., Rosset et al., 2007; Bach, 2017, and references therein). While this synthesis formulation of learning is insightful, there is a strong incentive to make the connection with regularization theory in direct analogy with the classical theory of inverse problems and machine learning that follows the analysis/variational formulation of the problem, which is the viewpoint adopted in this chapter.

3.2 $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$, $k \in \mathbb{N}$: the Sparse Ridge Spline Native Spaces

Recall from Chapter 1 that the native space for the kth-order locally adaptive splines is the kth-order BV space defined as

$$\mathrm{BV}^k(\mathbb{R}) = \big\{ f \in \mathcal{S}'(\mathbb{R}) : \ \mathrm{TV}^k(f) < \infty \big\},$$

for $k \in \mathbb{N}$, where $\mathrm{TV}^k(f) = \|\mathrm{D}^k f\|_{\mathcal{M}}$ is the kth-order total variation of f. A key property about the $\mathrm{TV}^k(\cdot)$ seminorm is that its null space, which is the null space of the operator D^k, defined by

$$\mathcal{N}(\mathrm{D}^k) \coloneqq \big\{ q \in \mathrm{BV}^k(\mathbb{R}) : \ \mathrm{D}^k q = 0 \big\}$$

is finite-dimensional. It is, in particular, the space of polynomials of degree at most $k-1$. Since we will be studying the seminorms

$$f \mapsto \|\partial_t^k \, \mathrm{K}^{d-1} \, \mathscr{R} f\|_{\mathcal{M}}, \tag{3.5}$$

for $k \in \mathbb{N}$, we must carefully define the native space so that the null space of the operator

$$\mathrm{D}_{\mathscr{R}}^k \coloneqq \partial_t^k \, \mathrm{K}^{d-1} \, \mathscr{R},$$

which can be viewed as a "Radonized" kth-order derivative operator, is finite-dimensional. To this end, we impose a *growth restriction* when defining the native space for the seminorm in (3.5). Thus, we define the family of native spaces

$$\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d) \coloneqq \left\{ f \in \mathcal{S}'(\mathbb{R}^d) : \begin{array}{l} \|\mathrm{D}_{\mathscr{R}}^k f\|_{\mathcal{M}} < \infty \\ \operatorname*{ess\,sup}_{\boldsymbol{x} \in \mathbb{R}^d} |f(\boldsymbol{x})| (1 + \|\boldsymbol{x}\|_2)^{-k+1} \end{array} \right\}, \tag{3.6}$$

for $k \in \mathbb{N}$. The second constraint in (3.6) imposes that functions in $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ do not grow faster than the algebraic growth rate of $k-1$. This growth restriction

ensures that the null space of the operator $\mathrm{D}^k_{\mathscr{R}}$, defined by

$$\mathcal{N}(\mathrm{D}^k_{\mathscr{R}}) \coloneqq \left\{q \in \mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d) : \ \mathrm{D}^k_{\mathscr{R}}\, q = 0\right\}$$

is finite-dimensional, while also being non-empty. We later show in Lemma 3.4 that $\mathcal{N}(\mathrm{D}^k_{\mathscr{R}})$ is, in particular, the space of polynomials of degree at most $k-1$. Imposing such a growth restriction on the native space is a common technique in multivariate scattered data approximation, in particular, in the L^2-theory of radial basis functions and polyharmonic splines (Wendland, 2004, Chapter 10). This is because constructing operators acting on multivariate functions with finite-dimensional null spaces is nearly impossible[1]. Since the seminorm $f \mapsto \|\mathrm{D}^k_{\mathscr{R}}\, f\|_{\mathcal{M}}$ is exactly the kth-order total variation of f in the (filtered) Radon domain, we write $\mathscr{R}\,\mathrm{TV}^k(f) \coloneqq \|\mathrm{D}^k_{\mathscr{R}}\, f\|_{\mathcal{M}}$.

We can view $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ as a subspace of $L^\infty(\mathbb{R}^d; n)$, which is the weighted L^∞-space defined by

$$L^\infty(\mathbb{R}^d; n) \coloneqq \left\{f \in \mathcal{S}'(\mathbb{R}^d) : \ \|f\|_{L^\infty,n} \coloneqq \operatorname*{ess\,sup}_{\boldsymbol{x}\in\mathbb{R}^d} |f(\boldsymbol{x})|(1+\|\boldsymbol{x}\|_2)^{-n} < \infty\right\}, \quad (3.7)$$

when $n = k-1$. This space is a Banach space whose predual is given by the weighted L^1-space defined by

$$L^1(\mathbb{R}^d; -n) \coloneqq \left\{f \in \mathcal{S}'(\mathbb{R}^d) : \ \|f\|_{L^1,-n} \coloneqq \int_{\mathbb{R}^d} |f(\boldsymbol{x})|(1+\|\boldsymbol{x}\|_2)^n \,\mathrm{d}\boldsymbol{x} < \infty\right\}. \quad (3.8)$$

We refer the reader to Unser et al. (2017, Section 4) for more details about the dual pair $(L^1(\mathbb{R}^d; -n), L^\infty(\mathbb{R}^d; n))$.

3.2.1 The Representer Theorem

Theorem 3.2. *Consider the following setting:*

[1]For example, consider Δ, the Laplacian operator in $\mathbb{R}^d$. Its null space is the space of harmonic functions which is infinite-dimensional for $d \geq 2$. On the other hand, the univariate Laplacian operator, $\mathrm{d}^2/\mathrm{d}x^2$, has a finite-dimensional null space which is simply $\operatorname{span}\{1, x\}$.

1. *The loss function* $\ell(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ *is convex, coercive, and lower semicontinuous in its second argument.*

2. *The linear measurement functionals* $h_m : \mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d) \to \mathbb{R} : f \mapsto \langle h_m, f \rangle$*, where* $m = 1, \ldots, M$*, are linearly independent and weak* continuous.*

3. *The number of measurements* M *is strictly greater than the dimension of the null space* $\mathcal{N}(\mathrm{D}^k_{\mathscr{R}})$*.*

4. *The regularization hyperparameter* $\lambda > 0$ *is fixed.*

Then, for any fixed $\boldsymbol{y} \in \mathbb{R}^M$*, the solution set to the data-fitting variational problem*

$$\mathcal{V} \coloneqq \operatorname*{arg\,min}_{f \in \mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)} \sum_{m=1}^{M} \ell(y_m, \langle h_m, f \rangle) + \lambda\, \mathscr{R}\,\mathrm{TV}^k(f)$$

is nonempty, convex, and weak compact. If* $\ell(\cdot,\cdot)$ *is strictly convex (or if it imposes the equality* $y_m = \langle h_m, f \rangle$*, for* $m = 1, \ldots, M$*), then the solution set* $\mathcal{V}$ *is the weak* closure of the convex hull of its extreme points, which can all be expressed as*

$$f_{\mathrm{ridge}}(\boldsymbol{x}) = \sum_{n=1}^{N_0} v_n \rho_k(\boldsymbol{w}_n^{\mathsf{T}} \boldsymbol{x} - b_n) + c(\boldsymbol{x}),$$

where $\{v_n\}_{n=1}^{N_0} \subset \mathbb{R} \setminus \{0\}$*,* $\{\boldsymbol{w}_n\}_{n=1}^{N_0} \subset \mathbb{S}^{d-1}$*,* $\{b_n\}_{n=1}^{N_0} \subset \mathbb{R}$*,* $c(\cdot)$ *is a polynomial of degree at most* $k-1$*, and* $N_0 < M$*. The corresponding regularization cost, which is common to all solutions, is* $\mathscr{R}\,\mathrm{TV}^k(f_{\mathrm{ridge}}) = \sum_{n=1}^{N_0} |v_n| = \|\boldsymbol{v}\|_1$*.*

The key takeaway of Theorem 3.2 is that the solution set to data-fitting variational problems over $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ is *completely characterized* by functions that are realizable by a shallow neural network plus a polynomial term (i.e., a ridge spline) with less neurons than measurements. The fact that the number of neurons is strictly less than the number of measurements illustrates the sparsifying effect of the $\mathcal{M}$-norm.

In order to prove Theorem 3.2, we require several intermediary results. In particular, we must understand the topological properties of the family of spaces

$\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$. We first provide a definition for ridge splines analogous to the operator-theoretic definition of a spline in Definition 1.3. Then, we show that, when equipped with the proper direct-sum topology, $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ is a non-reflexive Banach space. Establishing such a direct-sum decomposition is a common technique in spline theory, going back to early work on smoothing splines (de Boor and Lynch, 1966; Kimeldorf and Wahba, 1970a,b, 1971). Finally, we use these results to prove the representer theorem.

3.2.2 An Operator-Theoretic Definition of a Ridge Spline

The key idea in the definition of a spline in Definition 1.3 was the Green's function property of the spline atoms. A similar property holds for ridge functions whose profiles given by ρ_k defined in (1.17).

Lemma 3.3. *Let $\boldsymbol{z}_0 = (\boldsymbol{\alpha}_0, t_0) \in \mathbb{S}^{d-1} \times \mathbb{R}$. Then, it holds that*

$$\mathrm{D}^k_{\mathscr{R}}\big\{\rho_k(\boldsymbol{\alpha}_0^{\mathsf{T}}(\cdot) - t_0)\big\}(\boldsymbol{z}) = \frac{\delta(\boldsymbol{z} - \boldsymbol{z}_0) + (-1)^k \delta(\boldsymbol{z} + \boldsymbol{z}_0)}{2},$$

where $k \in \mathbb{N}$ and δ is the Dirac impulse on $\mathbb{S}^{d-1} \times \mathbb{R}$.

Proof. First note that the $\delta(\cdot - \boldsymbol{\alpha}_0)\,\rho_k(\cdot - t_0) \in \mathcal{X}'$, where

$$\mathcal{X} = C(\mathbb{S}^{d-1}) \mathbin{\widehat{\otimes}} L^1(\mathbb{R}; -k+1),$$

where $C(\mathbb{S}^{d-1})$ is the space of continuous functions on $\mathbb{S}^{d-1}$ and $L^1(\mathbb{R}; -k+1)$ is the weighted L^1-space defined in (3.8). This is because

$$\mathcal{X}' = \mathcal{M}(\mathbb{S}^{d-1}) \mathbin{\widehat{\otimes}} L^\infty(\mathbb{R}; k-1),$$

where $\mathcal{M}(\mathbb{S}^{d-1}) = \big(C(\mathbb{S}^{d-1})\big)'$ is the space of finite Radon measures on $\mathbb{S}^{d-1}$ and $L^\infty(\mathbb{R}; k-1)$ is the weighted L^∞-space defined in (3.7). Consider the pair of Radon-compatible subspaces $(\mathcal{X}_{\mathscr{R}}, \mathcal{X}'_{\mathscr{R}})$ as in Section 2.5 with this choice of $(\mathcal{X}, \mathcal{X}')$. One can

then verify that in this case we have that $\mathcal{X}_{\mathscr{R}} = \mathcal{X}_{\text{even}}$[2] and so the dual projector $\mathrm{P}^*_{\mathscr{R}} : \mathcal{X}' \to \mathcal{X}'_{\mathscr{R}}$ from Proposition 2.18 is the even projector, P_{even}. Next, since $\mathrm{D}^k_{\mathscr{R}} = \partial_t^k \, \mathrm{K}^{d-1} \, \mathscr{R}$, we have by Theorem 2.19 that

$$\begin{aligned}
\mathrm{D}^k_{\mathscr{R}}\big\{\rho_k(\boldsymbol{\alpha}_0^\mathsf{T}(\cdot) - t_0)\big\}(\boldsymbol{\alpha}, t) &= \partial_t^k \, \mathrm{P}^*_{\mathscr{R}}\{\delta(\cdot - \boldsymbol{\alpha}_0) \, \rho_k(\cdot - t_0)\}(\boldsymbol{\alpha}, t) \\
&= \partial_t^k \, \mathrm{P}_{\text{even}}\{\delta(\cdot - \boldsymbol{\alpha}_0) \, \rho_k(\cdot - t_0)\}(\boldsymbol{\alpha}, t) \\
&= \partial_t^k \left\{ \frac{\delta(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)\rho_k(t - t_0) + \delta(-\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)\rho_k(-t - t_0)}{2} \right\} \\
&= \frac{\delta(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)\delta(t - t_0) + (-1)^k \delta(-\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)\delta(-t - t_0)}{2} \\
&= \frac{\delta(\boldsymbol{\alpha} - \boldsymbol{\alpha}_0)\delta(t - t_0) + (-1)^k \delta(\boldsymbol{\alpha} + \boldsymbol{\alpha}_0)\delta(t + t_0)}{2} \\
&= \frac{\delta(\boldsymbol{z} - \boldsymbol{z}_0) + (-1)^k \delta(\boldsymbol{z} + \boldsymbol{z}_0)}{2}.
\end{aligned}$$

Note that this result holds when ρ_k is replaced with *any* Green's function of D^k. □

The other important property of the definition of a spline in Definition 1.3 is that the null space of the sparsifying operator L was finite-dimensional. This is also true for the null space of $\mathrm{D}^k_{\mathscr{R}}$.

Lemma 3.4. *The null space $\mathcal{N}(\mathrm{D}^k_{\mathscr{R}})$ is exactly the space of polynomials of degree at most $k-1$ on $\mathbb{R}^d$, denoted $\mathcal{P}_{k-1}(\mathbb{R}^d)$, which is a finite-dimensional space of dimension $\binom{d+k-1}{k-1}$.*

Proof. Let $f \in \mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$. By the Fourier slice theorem,

$$\widehat{\mathrm{D}^k_{\mathscr{R}}\{f\}}(\boldsymbol{\alpha}, \omega) = c_d(\mathrm{j}\omega)^k |\omega|^{d-1} \widehat{\mathscr{R}\{f\}}(\boldsymbol{\alpha}, \omega) = c_d(\mathrm{j}\omega)^k |\omega|^{d-1} \widehat{f}(\omega\boldsymbol{\alpha}). \tag{3.9}$$

In order for $f \in \mathcal{N}(\mathrm{D}^k_{\mathscr{R}})$, we require that (3.9) is 0 for all $(\boldsymbol{\alpha}, \omega) \in \mathbb{S}^{d-1} \times \mathbb{R}$. From the right-hand side of (3.9), we see that when $\omega = 0$, (3.9) is 0. Therefore, $f \in \mathcal{N}(\mathrm{D}^k_{\mathscr{R}})$

[2]This follows from the fact that the so-called Lizorkin space $\mathcal{S}_\infty(\mathbb{R}) \subset \mathcal{S}(\mathbb{R})$ of Schwartz functions with all moments vanishing is dense in $L^p(\mathbb{R})$ for $1 \leq p < \infty$ (Samko, 1995). In particular, it can be seen that the even functions in the Lizorkin space on $\mathbb{S}^{d-1} \times \mathbb{R}$ (appropriately defined) form a subspace of $\mathcal{S}_{\mathscr{R}}$ from the moment conditions in Theorem 2.15. See Unser (2022a, Lemma 2) for more details.

if and only if $\widehat{f}$ is supported only at $\mathbf{0}$. Therefore, f must be a polynomial, and, in particular, from the growth restriction in the definition of $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$, a polynomial of degree at most $k-1$. Therefore, $\mathcal{N}(\mathrm{D}^k_{\mathscr{R}}) \subset \mathcal{P}_{k-1}(\mathbb{R}^d)$.

Next, we have by the intertwining properties of the Radon transform and the Laplacian that

$$\mathrm{D}^k_{\mathscr{R}} = \partial_t^k\,\mathrm{K}^{d-1}\,\mathscr{R} = \mathrm{L}_t\,\mathrm{K}^{d-1}\,\mathscr{R}\Delta^{k/2}, \tag{3.10}$$

where L_t is a Fourier multiplier in the t variable defined via the univariate frequency response from $t \to \omega$

$$\widehat{L}_t(\omega) = \mathrm{sgn}(\omega)^k.$$

When k is an even integer, $\mathrm{L}_t = \mathrm{Id}$ and when k is an odd integer, L_t is proportional to the Hilbert transform $\mathscr{H}_t$. From (3.10), we see that $\mathcal{N}(\mathrm{D}^k_{\mathscr{R}})$ is at least as large as $\mathcal{N}(\Delta^{k/2})$, the growth restricted null space of $\Delta^{k/2}$. Finally, it is well-known that the growth restricted null space of $\Delta^{k/2}$ is the space of polynomials of degree at most $k-1$. The argument for this claim is that the only harmonic functions of slow growth are polynomials. Therefore, $\mathcal{N}(\mathrm{D}^k_{\mathscr{R}}) \supset \mathcal{P}_{k-1}(\mathbb{R}^d)$. Thus, we have shown that $\mathcal{N}(\mathrm{D}^k_{\mathscr{R}}) = \mathcal{P}_{k-1}(\mathbb{R}^d)$. The dimension of $\mathcal{P}_{k-1}(\mathbb{R}^d)$ follows from a counting argument. □

With Lemmas 3.3 and 3.4, we have the following definition of a ridge spline.

Definition 3.5. *A function $f : \mathbb{R}^d \to \mathbb{R}$ of slow growth*[3] *is said to be a ridge spline of order $k \in \mathbb{N}$ if*

$$\mathrm{D}^k_{\mathscr{R}} f = \partial_t^k\,\mathrm{K}^{d-1}\,\mathscr{R} f = \sum_{n=1}^{N} a_n \delta_k(\cdot - \boldsymbol{z}_n),$$

where

$$\delta_k(\boldsymbol{z}) \coloneqq \frac{\delta(\boldsymbol{z}) + (-1)^k \delta(-\boldsymbol{z})}{2}, \tag{3.11}$$

$\{a_n\}_{n=1}^N \subset \mathbb{R}$ is a sequence of weights, and $\{\boldsymbol{z}_n\}_{n=1}^N \subset \mathbb{S}^{d-1} \times \mathbb{R}$ are the directions and offsets of the neurons of the ridge spline. The function $\mathrm{D}^k_{\mathscr{R}} f$ is referred to as the innovation of f.

[3] i.e., in the space $L^\infty(\mathbb{R}^d; k-1)$.

3.2.3 Direct-Sum Decomposition of $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$

In this section we equip $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ with an appropriate direct-sum topology, showing that these spaces are non-reflexive Banach spaces. We use the techniques developed in Unser et al. (2017) which established a direct-sum decomposition for the native spaces for sparse L-splines. Since the null space $\mathcal{N}(\mathrm{D}^k_{\mathscr{R}})$ is finite-dimensional, we are guaranteed the existence of a *biorthogonal system* for $\mathcal{N}(\mathrm{D}^k_{\mathscr{R}})$.

Definition 3.6. *Let $\mathcal{N}$ be a finite-dimensional space of dimension $D_0 \coloneqq \dim \mathcal{N}$. The pair $(\boldsymbol{\phi}, \boldsymbol{p}) = \{(\phi_\ell, p_\ell)\}_{\ell=0}^{D_0-1}$ is called a* biorthogonal system *for $\mathcal{N}$ if $\boldsymbol{p} = \{p_\ell\}_{\ell=0}^{D_0-1}$ is a basis of $\mathcal{N}$ and the "boundary" functionals $\boldsymbol{\phi} = \{\phi_\ell\}_{\ell=0}^{D_0-1}$ with $\phi_\ell \in \mathcal{N}'$ (the continuous dual of $\mathcal{N}$) satisfy the biorthogonality condition $\langle \phi_\ell, p_n \rangle = \delta[\ell - n]$, $\ell, n = 0, \dots, D_0 - 1$, where $\delta[\cdot]$ is the Kronecker impulse and $\langle \cdot, \cdot \rangle$ is the duality pairing between $\mathcal{N}$ and $\mathcal{N}'$.*

Put $D_0 \coloneqq \dim \mathcal{N}(\mathrm{D}^k_{\mathscr{R}}) = \binom{d+k-1}{k-1}$ and let $\boldsymbol{\eta} = (\boldsymbol{\phi}, \boldsymbol{p})$ be a biorthogonal system for $\mathcal{N}(\mathrm{D}^k_{\mathscr{R}})$. Definition 3.6 implies that every $q \in \mathcal{N}(\mathrm{D}^k_{\mathscr{R}})$ admits the *unique representation*

$$q = \sum_{\ell=0}^{D_0-1} \langle \phi_\ell, q \rangle p_\ell.$$

We can define the projector $\mathrm{P}_{\mathcal{N}(\mathrm{D}^k_{\mathscr{R}}),\boldsymbol{\eta}} : \mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d) \to \mathcal{N}(\mathrm{D}^k_{\mathscr{R}})$ as

$$\mathrm{P}_{\mathcal{N}(\mathrm{D}^k_{\mathscr{R}}),\boldsymbol{\eta}}\{f\} = \sum_{\ell=0}^{D_0-1} \langle \phi_\ell, f \rangle p_\ell.$$

Moreover, since $\mathcal{N}(\mathrm{D}^k_{\mathscr{R}})$ is finite-dimensional, it is a Banach space when equipped with the norm

$$\|q\|_{\mathcal{N}(\mathrm{D}^k_{\mathscr{R}}),\boldsymbol{\eta}} \coloneqq \sum_{\ell=0}^{D_0-1} |\langle \phi_\ell, q \rangle|.$$

One could impose any finite-dimensional norm on the coefficients $\{\langle \phi_\ell, q \rangle\}_{\ell=0}^{D_0-1}$ to define an equivalent norm (since all norms are equivalent in finite dimensions).

Next, as a precursor to establishing the direct-sum decomposition of $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$, we construct a stable (i.e., bounded) right-inverse of the operator $\mathrm{D}^k_{\mathscr{R}}$. This inverse

will be later used to establish an isomorphism between $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ and a space of finite Radon measures, which will be used to establish the direct-sum decomposition of $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ in Theorem 3.8 and prove Theorem 3.2, the representer theorem. The space of finite Radon measures we are interested is the subspace $\mathcal{M}_k(\mathbb{S}^{d-1}\times\mathbb{R})\subset\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})$ defined by

$$\mathcal{M}_k(\mathbb{S}^{d-1}\times\mathbb{R}) \coloneqq \left\{\frac{u+(-1)^k u^\vee}{2} \;:\; u\in\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})\right\},$$

where $u^\vee$ is the reflection of u. If we define the projector

$$\mathrm{P}_k\{u\} \coloneqq \frac{u+(-1)^k u^\vee}{2}, \tag{3.12}$$

we have that $\mathcal{M}_k(\mathbb{S}^{d-1}\times\mathbb{R}) = \mathrm{P}_k\big(\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})\big)$.

Lemma 3.7. *Let $\boldsymbol{\eta}=(\boldsymbol{\phi},\boldsymbol{p})$ be a biorthogonal system for $\mathcal{N}(\mathrm{D}^k_{\mathscr{R}})\subset\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)\subset L^\infty(\mathbb{R}^d;k-1)$. Then, there exists a unique operator $\mathrm{D}^{-k}_{\mathscr{R},\boldsymbol{\eta}}:\mathcal{M}_k(\mathbb{S}^{d-1}\times\mathbb{R})\to L^\infty(\mathbb{R}^d;k-1)$ with the property that*

$$\begin{aligned}\mathrm{D}^k_{\mathscr{R}}\,\mathrm{D}^{-k}_{\mathscr{R},\boldsymbol{\eta}}\,u &= u && \textit{(right-inverse property)}\\ \mathrm{P}_{\mathcal{N}(\mathrm{D}^k_{\mathscr{R}}),\boldsymbol{\eta}}\big\{\mathrm{D}^{-k}_{\mathscr{R},\boldsymbol{\eta}}\,u\big\} &= 0 && \textit{(boundary conditions)}\end{aligned}$$

for all $u\in\mathcal{M}_k(\mathbb{S}^{d-1}\times\mathbb{R})$. The kernel of this operator is

$$g_{k,\boldsymbol{\eta}}(\boldsymbol{x},\boldsymbol{z}=(\boldsymbol{\alpha},t)) = \rho_k(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x}-t) - \mathrm{P}_{\mathcal{N}(\mathrm{D}^k_{\mathscr{R}}),\boldsymbol{\eta}}\big\{\rho_k(\boldsymbol{\alpha}^\mathsf{T}(\cdot)-t)\big\}(\boldsymbol{x}),$$

where $\boldsymbol{x}\in\mathbb{R}^d$ and $\boldsymbol{z}\in\mathbb{S}^{d-1}\times\mathbb{R}$. Moreover, this kernel satisfies the stability/continuity bound

$$C_{\boldsymbol{\eta}} \coloneqq \sup_{\substack{\boldsymbol{x}\in\mathbb{R}^d\\ \boldsymbol{z}\in\mathbb{S}^{d-1}\times\mathbb{R}}} |g_{k,\boldsymbol{\eta}}(\boldsymbol{x},\boldsymbol{z})|(1+\|\boldsymbol{x}\|_2)^{-k+1} < \infty.$$

Proof. For the proof of the continuity bound, we refer the reader to Unser et al. (2017, Theorem 3), which establishes such a bound for generic linear operators mapping

from finite Radon measures to $L^\infty(\mathbb{R}^d; n)$, $n \in \mathbb{N}$. This ensures that we can specify linear operators by their kernels by the Schwartz kernel theorem.

Next, we see that the stability condition implies that $\|g_{k,\boldsymbol{\eta}}(\boldsymbol{x}, \cdot)\|_{L^\infty} < \infty$ and so $\mathrm{D}_{\mathscr{R},\boldsymbol{\eta}}^{-k} u$ is well-defined. Therefore, the right-inverse property fo $\mathrm{D}_{\mathscr{R},\boldsymbol{\eta}}^{-k}$ holds by a direct calculation and noting that $\mathrm{D}_{\mathscr{R}}^{k} : \mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d) \to \mathcal{M}_k(\mathbb{S}^{d-1} \times \mathbb{R})$. The boundary conditions hold since the kernel of $\mathrm{D}_{\mathscr{R},\boldsymbol{\eta}}^{-k}$ subtracts off the null space component (this can also be verified by a direct calculation). Finally, uniqueness of this operator holds due to the uniqueness of the representations of elements of $\mathcal{N}(\mathrm{D}_{\mathscr{R}}^{k})$. □

Theorem 3.8. *Let $\boldsymbol{\eta} = (\boldsymbol{\phi}, \boldsymbol{p})$ be a biorthogonal system for $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$. Then, the following equivalent conditions hold:*

1. *Every $f \in \mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ admits a unique direct-sum decomposition as*

$$f = \mathrm{D}_{\mathscr{R},\boldsymbol{\eta}}^{-k} u + q_{\boldsymbol{\eta}}, \tag{3.13}$$

   where $u = \mathrm{D}_{\mathscr{R}}^{k} f \in \mathcal{M}_k(\mathbb{S}^{d-1} \times \mathbb{R})$ and $q_{\boldsymbol{\eta}} = \mathrm{P}_{\mathcal{N}(\mathrm{D}_{\mathscr{R}}^{k}),\boldsymbol{\eta}} f \in \mathcal{N}(\mathrm{D}_{\mathscr{R}}^{k})$.

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

$$\|f\|_{\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d),\boldsymbol{\eta}} \coloneqq \mathscr{R}\,\mathrm{TV}^k(f) + \|\mathrm{P}_{\mathcal{N}(\mathrm{D}_{\mathscr{R}}^{k}),\boldsymbol{\eta}} f\|_{\mathcal{N}(\mathrm{D}_{\mathscr{R}}^{k}),\boldsymbol{\eta}}$$

   is a Banach space.

Proof. Define the following subspace of $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$:

$$\mathscr{R}\,\mathrm{BV}_{\boldsymbol{\eta}}^{k}(\mathbb{R}^d) = \Big\{ f \in \mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d) : \mathrm{P}_{\mathcal{N}(\mathrm{D}_{\mathscr{R}}^{k}),\boldsymbol{\eta}} f = 0 \Big\}. \tag{3.14}$$

We have the equality $\mathscr{R}\,\mathrm{BV}_{\boldsymbol{\eta}}^{k}(\mathbb{R}^d) = (\mathrm{Id} - \mathrm{P}_{\mathcal{N}(\mathrm{D}_{\mathscr{R}}^{k}),\boldsymbol{\eta}})\big(\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)\big)$. In particular, $\mathscr{R}\,\mathrm{BV}_{\boldsymbol{\eta}}^{k}(\mathbb{R}^d)$ is a concrete transcription of the abstract quotient $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)/\mathcal{N}(\mathrm{D}_{\mathscr{R}}^{k})$ (i.e., we are working with a concrete representer from the equivalence class/coset members of the abstract quotient). Thus, we see that $\mathrm{D}_{\mathscr{R}}^{k} : \mathscr{R}\,\mathrm{BV}_{\boldsymbol{\eta}}^{k}(\mathbb{R}^d) \to \mathcal{M}_k(\mathbb{S}^{d-1} \times \mathbb{R})$ is a bijective isometry. By the bounded inverse theorem, there exists a bounded

inverse $\mathrm{D}_{\mathscr{R}}^{-k} : \mathcal{M}_k(\mathbb{S}^{d-1} \times \mathbb{R}) \to \mathscr{R}\,\mathrm{BV}_{\boldsymbol{\eta}}^k(\mathbb{R}^d)$. This inverse is exactly $\mathrm{D}_{\mathscr{R}}^{-k} = \mathrm{D}_{\mathscr{R},\boldsymbol{\eta}}^{-k}$, the unique operator constructed in Lemma 3.7 satisfying the conditions in (3.14).

1. *Direct-sum decomposition.* The discussion above immediately implies the direct-sum decomposition in (3.13).

2. *The Banach norm.* The discussion above specifies the structural property that $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d) = \mathscr{R}\,\mathrm{BV}_{\boldsymbol{\eta}}^k(\mathbb{R}^d) \oplus \mathcal{N}(\mathrm{D}_{\mathscr{R}}^k)$. Since $\mathscr{R}\,\mathrm{BV}_{\boldsymbol{\eta}}^k(\mathbb{R}^d)$ is isometrically isomorphic to $\mathcal{M}_k(\mathbb{S}^{d-1} \times \mathbb{R})$ via $\mathrm{D}_{\mathscr{R}}^k$, we can equip $\mathscr{R}\,\mathrm{BV}_{\boldsymbol{\eta}}^k(\mathbb{R}^d)$ with the norm $f \mapsto \|\mathrm{D}_{\mathscr{R}}^k f\|_{\mathcal{M}} = \mathscr{R}\,\mathrm{TV}^k(f)$, making it a Banach space. We can also equip $\mathcal{N}(\mathrm{D}_{\mathscr{R}}^k)$ with the norm $q \mapsto \|q\|_{\mathcal{N}(\mathrm{D}_{\mathscr{R}}^k),\boldsymbol{\eta}}$. Therefore, from the direct-sum decomposition in (3.13), we can make $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ a Banach space when we equip it with the composite norm

$$\|f\|_{\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d),\boldsymbol{\eta}} := \|u\|_{\mathcal{M}} + \|q_{\boldsymbol{\eta}}\|_{\mathcal{N}(\mathrm{D}_{\mathscr{R}}^k),\boldsymbol{\eta}} = \mathscr{R}\,\mathrm{TV}^k(f) + \|\mathrm{P}_{\mathcal{N}(\mathrm{D}_{\mathscr{R}}^k),\boldsymbol{\eta}}\, f\|_{\mathcal{N}(\mathrm{D}_{\mathscr{R}}^k)}.$$

□

Remark 3.9. Theorem 3.8 says, in particular, that $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ is isometrically isomorphic to $\mathcal{M}_k(\mathbb{S}^{d-1} \times \mathbb{R}) \times \mathcal{N}(\mathrm{D}_{\mathscr{R}}^k)$. Since $\mathcal{M}_k(\mathbb{S}^{d-1} \times \mathbb{R})$ is non-reflexive, this implies that $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ is non-reflexive. Moreover, this says that $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ does not admit an unconditional basis.

3.2.4 Proof of the Representer Theorem

Proof of Theorem 3.2. The proof of the representer theorem follows directly from the abstract representer theorem for direct-sums in Unser and Aziznejad (2022, Theorem 3). This abstract result gives the generic form of the extreme points of $\mathcal{V}$ as

$$f_{\text{extreme}}(\boldsymbol{x}) = \sum_{n=1}^{N_0} v_n e_n(\boldsymbol{x}) + c(\boldsymbol{x}),$$

where $\{v_n\}_{n=1}^{N_0} \subset \mathbb{R} \setminus \{0\}$, $c \in \mathcal{N}(\mathrm{D}_{\mathscr{R}}^k)$, $N_0 < M$, and $\{e_n\}_{n=1}^{N_0}$ are the extreme points of of the unit ball

$$B \coloneqq \big\{ f \in \mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d) : \ \mathscr{R}\,\mathrm{TV}^k(f) \leq 1 \big\}.$$

First note that from Lemma 3.4, we have that $\mathcal{N}(\mathrm{D}_{\mathscr{R}}^k) = \mathcal{P}_{k-1}(\mathbb{R}^d)$ and so c is a polynomial of degree at most $k-1$. Next, note that the extreme points of the unit ball $\big\{ u \in \mathcal{M}_k(\mathbb{S}^{d-1} \times \mathbb{R}) : \ \|u\|_{\mathcal{M}} \leq 1 \big\}$ take the form $\pm\delta_k(\cdot - \boldsymbol{z}_0)$, $\boldsymbol{z}_0 \in \mathbb{S}^{d-1} \times \mathbb{R}$, where δ_k is defined in (3.11). Indeed, this follows from the fact that the extreme points of the unit ball $\big\{ u \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R}) : \ \|u\|_{\mathcal{M}} \leq 1 \big\}$ take the form $\pm\delta(\cdot - \boldsymbol{z}_0)$, $\boldsymbol{z}_0 \in \mathbb{S}^{d-1} \times \mathbb{R}$ (see, e.g., Bredies and Carioni, 2020, Proposition 4.1), combined with the fact that $\mathcal{M}_k(\mathbb{S}^{d-1} \times \mathbb{R}) = \mathrm{P}_k\big(\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})\big)$, where P_k is the unit norm projector defined in (3.12). In particular, since P_k is a unit norm projector, we have that the extreme points of $\big\{ u \in \mathcal{M}_k(\mathbb{S}^{d-1} \times \mathbb{R}) : \ \|u\|_{\mathcal{M}} \leq 1 \big\}$ take the form $\mathrm{P}_k\{\pm\delta(\cdot - \boldsymbol{z}_0)\} = \pm\delta_k(\cdot - \boldsymbol{z}_0)$, $\boldsymbol{z}_0 \in \mathbb{S}^{d-1} \times \mathbb{R}$ (Neumayer and Unser, 2022, Proposition 5).

Next, let $\boldsymbol{\eta} = (\boldsymbol{\phi}, \boldsymbol{p})$ be a biorthogonal system for $\mathcal{N}(\mathrm{D}_{\mathscr{R}}^k)$. Since $\mathrm{D}_{\mathscr{R},\boldsymbol{\eta}}^{-k}$ constructed in Lemma 3.7 is an isometric isomorphism from $\mathcal{M}_k(\mathbb{S}^{d-1} \times \mathbb{R})$ to $\mathscr{R}\,\mathrm{BV}_{\boldsymbol{\eta}}^k(\mathbb{R}^d)$ (defined in (3.14)), we have that $\mathrm{D}_{\mathscr{R},\boldsymbol{\eta}}^{-k}$ maps extreme points of the unit ball in $\mathcal{M}_k(\mathbb{S}^{d-1} \times \mathbb{R})$ to the extreme points of the unit ball in $\mathscr{R}\,\mathrm{BV}_{\boldsymbol{\eta}}^k(\mathbb{R}^d)$, i.e., we have that the extreme points of the unit ball

$$\big\{ f \in \mathscr{R}\,\mathrm{BV}_{\boldsymbol{\eta}}^k(\mathbb{R}^d) : \ \mathscr{R}\,\mathrm{TV}^k(f) \leq 1 \big\}$$

take the form $\mathrm{D}_{\mathscr{R},\boldsymbol{\eta}}^{-k}\{\pm\delta_k(\cdot - \boldsymbol{z}_0)\}$, $\boldsymbol{z}_0 \in \mathbb{S}^{d-1} \times \mathbb{R}$. From the definition of $\mathrm{D}_{\mathscr{R},\boldsymbol{\eta}}^{-k}$ in Lemma 3.7, we have that these extreme points take the form

$$\boldsymbol{x} \mapsto \pm\left(\frac{\rho_k(\boldsymbol{\alpha}_0^\mathsf{T}\boldsymbol{x} - t_0) + (-1)^k \rho_k(-\boldsymbol{\alpha}_0^\mathsf{T}\boldsymbol{x} + t_0)}{2} \right) = \pm\rho_k(\boldsymbol{\alpha}_0^\mathsf{T}\boldsymbol{x} - t_0)$$

where $\boldsymbol{z}_0 = (\boldsymbol{\alpha}_0, t_0) \in \mathbb{S}^{d-1} \times \mathbb{R}$ and the equality holds due the symmetry/antisymmetry of ρ_k defined in (1.17). Therefore, the extreme points of B take the form $\boldsymbol{x} \mapsto \pm\rho_k(\boldsymbol{\alpha}_0^\mathsf{T}\boldsymbol{x} - t_0) + q(\boldsymbol{x})$, for some $q \in \mathcal{N}(\mathrm{D}_{\mathscr{R}}^k)$, which proves the theorem. □

3.2.5 Discussion

The $\mathscr{R}\,\mathrm{TV}^k$-seminorm satisfies many useful properties. It is rotation-, scale-, and translation-invariant, summarized in the following theorem.

Theorem 3.10. *Given $f \in \mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$, define $g(\boldsymbol{x}) = f(\gamma \mathbf{U}\boldsymbol{x} - \boldsymbol{b})$, where $\gamma > 0$ is a scaling factor, $\mathbf{U} \in \mathbb{R}^{d\times d}$ is an orthogonal matrix (i.e., $\mathbf{U}^\mathsf{T}\mathbf{U} = \mathbf{I}$), and $\boldsymbol{b} \in \mathbb{R}^d$ is a translation. Then,*

$$\mathscr{R}\,\mathrm{TV}^k(f) = \gamma^{k-1}\,\mathscr{R}\,\mathrm{TV}^k(g).$$

The proof of this theorem is a direct calculation. We refer the reader to Ongie et al. (2020a) for an explicit calculation in the case that $k = 2$. The $\mathscr{R}\,\mathrm{TV}^k$-seminorm also recovers the TV^k-seminorm in the univariate case. Indeed, this follows by noticing that in the univariate case, the Radon transform is simply the operator

$$\mathscr{R}\{f\}(\alpha, t) = f\left(\frac{t}{\alpha}\right),$$

where $\alpha \in \{\pm 1\}$ and $t \in \mathbb{R}$, and the ramp filter K^{d-1} is simply multiplication by the constant $1/2$. Therefore, in the univariate case we have

$$\mathscr{R}\,\mathrm{TV}^k(f) = \frac{1}{2}\|\partial_t^k \mathscr{R} f\|_{\mathcal{M}} = \frac{1}{2}\sum_{\alpha\in\{\pm 1\}} \left\|\mathrm{D}^k f\left(\frac{\cdot}{\alpha}\right)\right\|_{\mathcal{M}} = \|\mathrm{D}^k f\|_{\mathcal{M}} = \mathrm{TV}^k(f),$$

where the second to last equality holds since $f(\cdot/\alpha)$, $\alpha \in \{\pm 1\}$, is either f or its reflection, both of which have the same $\|\mathrm{D}^k\{\cdot\}\|_{\mathcal{M}}$ value. Thus, in the univariate case, we have $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}) = \mathrm{BV}^k(\mathbb{R})$. Therefore, the representer theorem in Theorem 3.2 can be viewed as a multivariate generalization of the representer theorem for locally adaptive splines, and so sparse ridge splines can be viewed as a multivariate generalization of locally adaptive splines.

Since when $d = 1$, $\mathscr{R}\,\mathrm{TV}^1(\cdot) = \mathrm{TV}^1(\cdot) = \mathrm{TV}(\cdot)$, which is the usual notation of total variation in the univariate case, for $d > 1$ and when $k = 1$, $\mathscr{R}\,\mathrm{TV}(\cdot) \coloneqq \mathscr{R}\,\mathrm{TV}^1(\cdot)$ can be viewed as a new notion of multivariate total variation, different from the usual notion $\mathrm{TV}(f) = \|\nabla f\|_{\mathcal{M}}$ and deserves further study.

3.2.6 Fractional Ordered Spaces

We can also define fractional (non-integer) ordered variants of the $\mathscr{R}\,\mathrm{BV}$-spaces. Given a real number $s \geq 1$, there are two possible definitions which make sense in our setting, the *symmetric* space

$$\mathscr{R}\,\mathrm{BV}^s_{\mathrm{sym}}(\mathbb{R}^d) \coloneqq \left\{ f \in \mathcal{S}'(\mathbb{R}^d) : \begin{array}{l} \|\partial^s_{t*}\,\mathrm{K}^{d-1}\,\mathscr{R} f\|_{\mathcal{M}} < \infty \\ \operatorname*{ess\,sup}_{\boldsymbol{x}\in\mathbb{R}^d} |f(\boldsymbol{x})|(1+\|\boldsymbol{x}\|_2)^{-\lceil s-1\rceil} \end{array} \right\} \tag{3.15}$$

and the *antisymmetric* space

$$\mathscr{R}\,\mathrm{BV}^s_{\mathrm{antisym}}(\mathbb{R}^d) \coloneqq \left\{ f \in \mathcal{S}'(\mathbb{R}^d) : \begin{array}{l} \|\partial^s_{t*}\,\mathscr{H}_t\,\mathrm{K}^{d-1}\,\mathscr{R} f\|_{\mathcal{M}} < \infty \\ \operatorname*{ess\,sup}_{\boldsymbol{x}\in\mathbb{R}^d} |f(\boldsymbol{x})|(1+\|\boldsymbol{x}\|_2)^{-\lceil s-1\rceil} \end{array} \right\}, \tag{3.16}$$

where ∂^s_{t*} denotes the self-adjoint fractional derivative operator defined by the univariate frequency response $\omega \mapsto |\omega|^s$. The reason for the growth restriction of order $\lceil s-1\rceil$ in (3.15) and (3.16) is due to the fact that the null space of the fractional derivative operator D^s defined by the frequency response $\omega \mapsto (\mathrm{j}\omega)^s$ is the space of polynomials of degree at most $\lceil s-1\rceil$ (Unser and Blu, 2000).

We refer to the space defined in (3.15) as the symmetric version of the space since the resulting activation functions in the representer theorem over $\mathscr{R}\,\mathrm{BV}^s_{\mathrm{sym}}(\mathbb{R}^d)$ will be proportional to $|\cdot|^{s-1}$ (i.e., they are symmetric about the origin), while the resulting activation functions in the representer theorem over $\mathscr{R}\,\mathrm{BV}^s_{\mathrm{antisym}}(\mathbb{R}^d)$ will be proportional to $\mathrm{sgn}(\cdot)|\cdot|^{s-1}$ (i.e., they are antisymmetric about the origin). These are exactly the building blocks of the symmetric and antisymmetric fractional splines of Unser and Blu (2000).

3.3 Applications to Learning with Neural Networks

While Theorem 3.2 is a powerful representer theorem result for general inverse problems, the problem of learning is interested in the setting where the measurement

operator corresponds to point evaluations. Additionally, the ReLU activation (truncated linear function) is of particular interest to the neural network community. In this section, we discuss applications of Theorem 3.2 to the problem of learning with ReLU neural networks, i.e., the $k = 2$ case. To this end, the first step is to establish that the point evaluation operator is weak* continuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$.

In order to equip $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ with an explicit direct-sum topology, we consider an explicit biorthogonal system $\boldsymbol{\eta} = (\boldsymbol{\phi}, \boldsymbol{p})$ for the null space $\mathcal{N}(\mathrm{D}^2_{\mathscr{R}}) = \mathcal{P}_1(\mathbb{R}^d)$, the space of affine functions on $\mathbb{R}^d$. Put $p_0(\boldsymbol{x}) \coloneqq 1$ and $p_\ell(\boldsymbol{x}) \coloneqq x_\ell$, $\ell = 1, \dots, d$. Clearly $\boldsymbol{p}$ is a basis for $\mathcal{P}_1(\mathbb{R}^d)$. Put $\phi_0 \coloneqq \delta$ and $\phi_\ell \coloneqq \delta(\cdot - \boldsymbol{e}_\ell) - \delta$, $\ell = 1, \dots, d$, where δ denotes the Dirac impulse on $\mathbb{R}^d$ and $\boldsymbol{e}_\ell$ denotes the ℓth canonical basis vector of $\mathbb{R}^d$. Then, $(\boldsymbol{\phi}, \boldsymbol{p})$ is a biorthogonal system for $\mathcal{P}_1(\mathbb{R}^d)$. Indeed, we have $\langle \phi_0, p_0 \rangle = 1$ and $\langle \phi_\ell, p_\ell \rangle = p_\ell(\boldsymbol{e}_\ell) - p_\ell(\boldsymbol{0}) = 1 - 0 = 1$, $\ell = 1, \dots, d$. We also have

$$
\begin{aligned}
&\langle \phi_0, p_\ell \rangle = p_\ell(\boldsymbol{0}) = 0, \quad \ell = 1, \dots, d, \\
&\langle \phi_\ell, p_0 \rangle = p_0(\boldsymbol{e}_\ell) - p_0(\boldsymbol{0}) = 1 - 1 = 0, \quad \ell = 1, \dots, d, \\
&\langle \phi_\ell, p_n \rangle = p_n(\boldsymbol{e}_\ell) - p_n(\boldsymbol{0}) = 0 + 0 = 0, \quad \ell, n = 1, \dots, d, \quad \ell \neq n.
\end{aligned}
$$

Lemma 3.11. *The space $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ equipped with the norm*

$$
\|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} \coloneqq \mathscr{R}\,\mathrm{TV}^2(f) + |f(\boldsymbol{0})| + \sum_{\ell=1}^{d} |f(\boldsymbol{e}_\ell) - f(\boldsymbol{0})|, \tag{3.17}
$$

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

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

Proof. Consider the biorthogonal system $\boldsymbol{\eta} = (\boldsymbol{\phi}, \boldsymbol{p})$ defined above. The proof of Item 1 follows from substituting $\boldsymbol{\eta}$ into Theorem 3.8. Next, from Theorem 3.8 $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d) \cong \mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\eta}}(\mathbb{R}^d) \oplus \mathcal{P}_1(\mathbb{R}^d)$, showing that $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, is weak*

continuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ is equivalent to showing that it is weak* continuous on both $\mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\eta}}(\mathbb{R}^d)$ and $\mathcal{P}_1(\mathbb{R}^d)$.

Clearly $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, is continuous on $\mathcal{P}_1(\mathbb{R}^d)$ (since every element of $\mathcal{P}_1(\mathbb{R}^d)$ is a continuous function). Then, since $\mathcal{P}_1(\mathbb{R}^d)$ is finite-dimensional, the space of continuous linear functionals and weak* continuous linear functionals are the same. Thus, $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, is weak* continuous on $\mathcal{P}_1(\mathbb{R}^d)$. It remains to show that $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, is weak* continuous on $\mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\eta}}(\mathbb{R}^d)$. Let $\mathcal{X}$ be the predual of $\mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\eta}}(\mathbb{R}^d)$, i.e., $\mathcal{X}' = \mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\eta}}(\mathbb{R}^d)$. We must show that $\delta(\cdot - \boldsymbol{x}_0) \in \mathcal{X}$, $\boldsymbol{x}_0 \in \mathbb{R}^d$. The Riesz–Markov–Kakutani representation theorem states that the predual of $\mathcal{M}_2(\mathbb{S}^{d-1} \times \mathbb{R}) = \mathcal{M}_{\mathrm{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$ is $C_{0,\mathrm{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$, the subspace of even functions in $C_0(\mathbb{S}^{d-1} \times \mathbb{R})$. From Theorem 3.8, we have that the following relations of all these spaces.

$$\begin{array}{ccc}
\mathscr{R}\,\mathrm{BV}^2_{\boldsymbol{\eta}}(\mathbb{R}^d) & \underset{\mathrm{D}^{-2}_{\mathscr{R},\boldsymbol{\eta}}}{\overset{\mathrm{D}^2_{\mathscr{R}}}{\rightleftarrows}} & \mathcal{M}_{\mathrm{even}}(\mathbb{S}^{d-1} \times \mathbb{R}) \\
\mathrm{dual}\ \uparrow\ \downarrow & & \uparrow\ \downarrow\ \mathrm{dual} \\
\mathcal{X} & \underset{\left(\mathrm{D}^{-2}_{\mathscr{R},\boldsymbol{\eta}}\right)^*}{\overset{\left(\mathrm{D}^2_{\mathscr{R}}\right)^*}{\rightleftarrows}} & C_{0,\mathrm{even}}(\mathbb{S}^{d-1} \times \mathbb{R})
\end{array}$$

The above diagram shows that $\delta(\cdot - \boldsymbol{x}_0) \in \mathcal{X}$ if and only if $\left(\mathrm{D}^{-2}_{\mathscr{R},\boldsymbol{\eta}}\right)^*\{\delta(\cdot - \boldsymbol{x}_0)\} \in C_{0,\mathrm{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$. From Theorem 3.8 we see that $\left(\mathrm{D}^{-2}_{\mathscr{R},\boldsymbol{\eta}}\right)^*\{\delta(\cdot - \boldsymbol{x}_0)\} = g_{2,\boldsymbol{\eta}}(\boldsymbol{x}_0, \cdot)$ defined in Lemma 3.7. By choosing $\rho_2 = |\cdot|/2$ in Lemma 3.7, we have

$$\begin{aligned}
g_{2,\boldsymbol{\eta}}(\boldsymbol{x}_0, (\boldsymbol{\alpha}, t)) &= \frac{|\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{x}_0 - t|}{2} - \sum_{k=0}^{d} p_\ell(\boldsymbol{x}_0)\left\langle \phi_\ell, \frac{|\boldsymbol{\alpha}^{\mathsf{T}}(\cdot) - t|}{2} \right\rangle \\
&\overset{(*)}{=} \frac{|\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{x}_0 - t|}{2} - \left[\frac{|-t|}{2} + \sum_{k=1}^{d} x_{0,k}\left(\frac{|\alpha_k - t|}{2} - \frac{|-t|}{2}\right)\right] \\
&= \frac{|\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{x}_0 - t|}{2} - \frac{|t|}{2}\left(1 - \sum_{k=1}^{d} x_{0,k}\right) - \sum_{k=1}^{d} x_{0,k}\frac{|\alpha_k - t|}{2},
\end{aligned}$$

where $(*)$ follows by substituting in the biorthogonal system $\boldsymbol{\eta} = (\boldsymbol{\phi}, \boldsymbol{p})$. Clearly $g_{2,\boldsymbol{\eta}}(\boldsymbol{x}_0, \cdot)$ is continuous and $g_{2,\boldsymbol{\eta}}(\boldsymbol{x}_0, (\boldsymbol{\alpha}, t)) = g_{2,\boldsymbol{\eta}}(\boldsymbol{x}_0, (-\boldsymbol{\alpha}, -t))$, so $g_{2,\boldsymbol{\eta}}(\boldsymbol{x}_0, \cdot)$ is an even function on $\mathbb{S}^{d-1} \times \mathbb{R}$. It remains to check that $g_{2,\boldsymbol{\eta}}(\boldsymbol{x}_0, \cdot)$ is vanishing at infinity. Certainly this is true. Indeed, for sufficiently large t we have

$$g_{2,\boldsymbol{\eta}}(\boldsymbol{x}_0, (\boldsymbol{\alpha}, t)) = \frac{-\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{x}_0 + t}{2} - \frac{t}{2}\left(1 - \sum_{k=1}^{d} x_{0,k}\right) - \sum_{k=1}^{d} x_{0,k}\frac{-\alpha_k + t}{2} = 0,$$

and for sufficiently small t we have

$$g_{2,\boldsymbol{\eta}}(\boldsymbol{x}_0, (\boldsymbol{\alpha}, t)) = \frac{\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{x}_0 - t}{2} - \frac{-t}{2}\left(1 - \sum_{k=1}^{d} x_{0,k}\right) - \sum_{k=1}^{d} x_{0,k}\frac{\alpha_k - t}{2} = 0.$$

Therefore, $g_{2,\boldsymbol{\eta}}(\boldsymbol{x}_0, \cdot) \in C_{0,\mathrm{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$. Thus, the Dirac impulse $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \mathbb{R}^d$, is weak* continuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$. □

The norm defined in Lemma 3.11 is also an upper bound on the Lipschitz constant of the function.

Lemma 3.12. *Let $f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$. Then, f is Lipschitz continuous and satisfies the Lipschitz bound*

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

for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$.

Proof. To prove this, we appeal to the direct-sum decomposition of $f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ in Theorem 3.8. Let $\boldsymbol{\eta} = (\boldsymbol{\phi}, \boldsymbol{p})$ be the biorthogonal system used to define the $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$-norm. Then, from (3.13) in Theorem 3.8, f admits the direct-sum decomposition

$$f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} g_{2,\boldsymbol{\eta}}(\boldsymbol{x}, (\boldsymbol{\alpha}, t))\, \mathrm{d}\mu(\boldsymbol{\alpha}, t) + \boldsymbol{c}^{\mathsf{T}}\boldsymbol{x} + c_0,$$

where $\mu = \mathrm{D}^k_{\mathscr{R}} f \in \mathcal{M}_{\mathrm{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$ and $\boldsymbol{c}^{\mathsf{T}}\boldsymbol{x} + c_0 = \mathrm{P}_{\mathcal{P}_1(\mathbb{R}^d),\boldsymbol{\eta}}\{f\}(\boldsymbol{x})$. From Theo-

rem 3.8, we have that

$$\|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} = \|\mu\|_{\mathcal{M}} + \|\boldsymbol{c}\|_1 + |c_0|. \tag{3.18}$$

We first bound the Lipschitz constant of $g_{2,\boldsymbol{\eta}}(\cdot, \boldsymbol{z} = (\boldsymbol{\alpha}, t))$. For any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\begin{aligned}
|g_{2,\boldsymbol{\eta}}(\boldsymbol{x}, \boldsymbol{z}) - g_{2,\boldsymbol{\eta}}(\boldsymbol{y}, \boldsymbol{z})| = \Bigg| & \frac{\left|\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x} - t\right|}{2} - \frac{\left|\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{y} - b\right|}{2} \\
& - \frac{|t|}{2}\left[\left(1 - \sum_{k=1}^d x_k\right) - \left(1 - \sum_{k=1}^d y_k\right)\right] - \sum_{k=1}^d (x_k - y_k)\frac{|\alpha_k - t|}{2}\Bigg| \\
\leq & \frac{\left|\,\left|\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x} - t\right| - \left|\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{y} - t\right|\,\right|}{2} \\
& + \left|\sum_{k=1}^d (x_k - y_k)\frac{|t|}{2} - \sum_{k=1}^d (x_k - y_k)\frac{|\alpha_k - t|}{2}\right| \\
\leq & \frac{\left|\,\left|\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x} - t\right| - \left|\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{y} - t\right|\,\right|}{2} + \sum_{k=1}^d |x_k - y_k|\frac{|\,|t| - |\alpha_k - t|\,|}{2} \\
\overset{(*)}{\leq} & \frac{\left|\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x} - \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{y}\right|}{2} + \sum_{k=1}^d |x_k - y_k|\frac{|\alpha_k|}{2} \\
\overset{(\S)}{\leq} & \frac{\|\boldsymbol{\alpha}\|_\infty \|\boldsymbol{x} - \boldsymbol{y}\|_1 + \|\boldsymbol{\alpha}\|_\infty \|\boldsymbol{x} - \boldsymbol{y}\|_1}{2} \overset{(\dagger)}{\leq} \|\boldsymbol{x} - \boldsymbol{y}\|_1
\end{aligned}$$

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

Next, we have for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$,

$$\begin{aligned}
|f(\boldsymbol{x}) - f(\boldsymbol{y})| &\leq \int_{\mathbb{S}^{d-1}\times\mathbb{R}} |g_{\boldsymbol{2},\boldsymbol{\eta}}(\boldsymbol{x}, (\boldsymbol{\alpha}, t)) - g_{2,\boldsymbol{\eta}}(\boldsymbol{y}, (\boldsymbol{\alpha}, t))|\,\mathrm{d}|\mu|(\boldsymbol{\alpha}, t) + |\boldsymbol{c}^\mathsf{T}(\boldsymbol{x} - \boldsymbol{y})| \\
&\leq \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \|\boldsymbol{x} - \boldsymbol{y}\|_1\,\mathrm{d}|\mu|(\boldsymbol{\alpha}, t) + \|\boldsymbol{c}\|_\infty \|\boldsymbol{x} - \boldsymbol{y}\|_1 \\
&\leq \|\mu\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} \|\boldsymbol{x} - \boldsymbol{y}\|_1 + \|\boldsymbol{c}\|_1 \|\boldsymbol{x} - \boldsymbol{y}\|_1 \\
&\leq \|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)}\, \|\boldsymbol{x} - \boldsymbol{y}\|_1,
\end{aligned}$$

where the third line follows from the fact that $\|\cdot\|_\infty \leq \|\cdot\|_1$ in finite-dimensional spaces and the fourth line follows from (3.18). □

3.3.1 Learning with Shallow Neural Networks

From Lemma 3.11, we immediately have the following representer theorem for learning with shallow ReLU networks as a corollary to the general representer theorem in Theorem 3.2.

Corollary 3.13. *Let $\ell(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a strictly convex, coercive, and lower semicontinuous loss function in its second argument, $\{(\boldsymbol{x}_m, y_m)\}_{m=1}^M \subset \mathbb{R}^d \times \mathbb{R}$ be a given set of distinct data points with $M > d+1$, and let the regularization hyperparameter $\lambda > 0$ be fixed. Then, the solution set to the learning problem*

$$\mathcal{V} \coloneqq \operatorname*{arg\,min}_{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} \sum_{m=1}^M \ell(y_m, f(\boldsymbol{x}_m)) + \lambda\, \mathscr{R}\,\mathrm{TV}^2(f) \tag{3.19}$$

is nonempty, convex, and weak compact. The solution set $\mathcal{V}$ is the weak* closure of the convex hull of its extreme points, which can all be expressed as*

$$f_{\mathrm{ReLU}}(\boldsymbol{x}) = \sum_{n=1}^{N_0} v_n\, \mathrm{ReLU}(\boldsymbol{w}_n^\mathsf{T}\boldsymbol{x} - b_n) + \boldsymbol{c}^\mathsf{T}\boldsymbol{x} + c_0,$$

where $\{v_n\}_{n=1}^{N_0} \subset \mathbb{R}\setminus\{0\}$, $\{\boldsymbol{w}_n\}_{n=1}^{N_0} \subset \mathbb{S}^{d-1}$, $\{b_n\}_{n=1}^{N_0} \subset \mathbb{R}$, $\boldsymbol{c} \in \mathbb{R}^d$, $c_0 \in \mathbb{R}$, and $N_0 < M$. The corresponding regularization cost, which is common to all solutions, is $\mathscr{R}\,\mathrm{TV}^2(f_{\mathrm{ReLU}}) = \sum_{n=1}^{N_0} |v_n| = \|\boldsymbol{v}\|_1$.

Remark 3.14. In fact, one can show that the Dirac impulse is weak* continuous on $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ for $k \geq 2$ with an appropriate choice of biorthogonal system for the polynomial null space.

What is remarkable about this result is that it implies that the solution set to the variational problem in (3.19) is completely characterized by shallow ReLU networks with a skip connection. Moreover, since the ReLU is positively homogeneous of degree

1, if we consider the shallow ReLU network

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{n=1}^{N} v_n \, \mathrm{ReLU}(\boldsymbol{w}_n^{\mathsf{T}} \boldsymbol{x} - b_n) + \boldsymbol{c}^{\mathsf{T}} \boldsymbol{x} + c_0,$$

where $\boldsymbol{\theta}$ contains all the neural network parameters (i.e., weights and biases), and we do not constrain $\boldsymbol{w}_n \in \mathbb{S}^{d-1}$ and instead let $\boldsymbol{w}_n \in \mathbb{R}^d$, we can reparameterize the network as

$$f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sum_{n=1}^{N} v_n \|\boldsymbol{w}_n\|_2 \, \mathrm{ReLU}(\widetilde{\boldsymbol{w}}_n^{\mathsf{T}} \boldsymbol{x} - \widetilde{b}_n) + \boldsymbol{c}^{\mathsf{T}} \boldsymbol{x} + c_0,$$

where $\widetilde{\boldsymbol{w}}_n = \boldsymbol{w}_n / \|\boldsymbol{w}_n\|_2 \in \mathbb{S}^{d-1}$ and $\widetilde{b}_n = b_n / \|\boldsymbol{w}_n\|_2 \in \mathbb{R}$. Therefore, given a shallow ReLU network with a skip connection, parameterized by $\boldsymbol{\theta}$, we have that[4]

$$\mathscr{R}\,\mathrm{TV}^2(f_{\boldsymbol{\theta}}) = \sum_{n=1}^{N} |v_n| \|\boldsymbol{w}_n\|_2.$$

The quantity in the above display, is sometimes referred to as the *path-norm* of the neural network (Neyshabur et al., 2015a). Therefore, we can recast the variational problem in (3.19) as the finite-dimensional neural network training problem

$$\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \sum_{m=1}^{M} \ell(y_m, f_{\boldsymbol{\theta}}(\boldsymbol{x}_m)) + \lambda \sum_{n=1}^{N} |v_n| \|\boldsymbol{w}_n\|_2,$$

so long as $N \geq N_0$, where $\boldsymbol{\Theta} = \mathbb{R}^K$ is the total number of scalar neural network parameters. In particular, the solutions to the problem in the above display will always be solutions to the variational problem in (3.19). It turns out that the training problem in the above display is equivalent to the training problem that corresponds to training a shallow ReLU network with weight decay. This is summarized in the following theorem.

[4] Assuming the neural network is in "reduced form", i.e., the weight bias pairs are unique up to certain symmetries that arise due to the symmetries of the Radon domain.

Theorem 3.15. *Let $\boldsymbol{\Theta}$ denote the space of neural network parameters for a shallow ReLU network with a skip connection with N neurons. Then, the following two optimization problems are equivalent (in the sense that their solution sets are the same):*

1. $\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{m=1}^{M} \ell(y_m, f_{\boldsymbol{\theta}}(\boldsymbol{x}_m)) + \lambda \sum_{n=1}^{N} |v_n| \|\boldsymbol{w}_n\|_2$;

2. $\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{m=1}^{M} \ell(y_m, f_{\boldsymbol{\theta}}(\boldsymbol{x}_m)) + \frac{\lambda}{2} \sum_{n=1}^{N} |v_n|^2 + \|\boldsymbol{w}_n\|_2^2$.

Proof. The key idea behind the proof of this claim hinges on the fact that the ReLU activation is positively homogeneous of degree 1. In particular, consider the nth neuron $\boldsymbol{x} \mapsto v_n \operatorname{ReLU}(\boldsymbol{w}_n^\mathsf{T}\boldsymbol{x} - b_n)$. Due to the homogeneity of the ReLU, the weights can be rescaled so that $|v_n| = \|\boldsymbol{w}_n\|_2$, without changing the functional mapping of the neuron. Therefore, minimizing $|v_n|^2 + \|\boldsymbol{w}_n\|_2^2$ is achieved when $|v_n| = \|\boldsymbol{w}_n\|_2$, and so at any solution to the second optimization in the theorem statement we have

$$\frac{|v_n|^2 + \|\boldsymbol{w}_n\|_2^2}{2} = |v_n| \|\boldsymbol{w}_n\|_2,$$

and so the two optimization problems are equivalent. □

The main takeaway from this result is that training a shallow ReLU network with a skip connection with weight decay (to a global minimizer) results in a solution to the variational problem in (3.19).

Remark 3.16. The connection between the neural network training problems in Theorem 3.15 and the variational problem in (3.19) hiinges on the fact that the biases remain *unregularized*. One could also consider neural network training problems where the biases were also regularized and they would be related to variational problems over a different Banach space.

3.3.2 Learning with Deep Neural Networks

We can establish similar results connecting neural network training problems with variational problems in the case of deep neural networks. The construction essentially boils down to considering functions that are compositions of functions in $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ in order to impose the compositional structure that arises in deep neural networks.

We remark that there are a few lines of related work concerning similar variational formulations of learning with deep neural network. One line of work is concerned with the "optimal shaping" of the activation functions in a deep neural network (Unser, 2019; Aziznejad et al., 2020; Bohra et al., 2020). In particular, Unser (2019) proves a representer theorem regarding the optimal shaping of the activation functions. They consider the standard fully-connected feedforward deep neural network architecture, but allow the activation functions to be learnable. They impose a second-order total variation penalty on the activation functions and so the optimal shaping of the activation functions corresponds to linear splines with adaptive knot locations. We remark that we use several techniques developed in Unser (2019); Aziznejad et al. (2020) to prove our representer theorem in this paper, particularly in proving existence of solutions to the variational problem we study. Finally, there is a line of work regarding "deep kernel learning" (Bohn et al., 2019), in which they derive a representer theorem for compositions of kernel machines. They consider a construction similar to ours regarding the function space they study, but they consider compositions of reproducing kernel Hilbert spaces and so the resulting solutions to their variational problem do not take the form of a deep neural network.

First, define the vector-valued analogue of $\mathscr{R}\,\mathrm{BV}^2$ as the Cartesian product

$$\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D) := \underbrace{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d) \times \cdots \times \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)}_{D \text{ times}}. \tag{3.20}$$

This space can be viewed as the Bochner space $\ell^1([D];\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d))$, where $[D] = \{1,\dots,D\}$, and can therefore be turned into a Banach space when equipped with

the norm

$$\|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)} \coloneqq \|f\|_{\ell^1([D];\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d))} = \sum_{k=1}^{D} \|f_k\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)}, \tag{3.21}$$

where $f = (f_1, \dots, f_D)$. We could equip any finite-dimensional norm on the output vector to define an equivalent norm, but we work with the ℓ^1-norm for simplicity. Next, consider the "deep" analogue of $\mathscr{R}\,\mathrm{BV}^2$, defined as

$$\mathscr{R}\,\mathrm{BV}^2_{\mathrm{deep}}(\mathbb{R}^{d_0};\cdots;\mathbb{R}^{d_L}) \coloneqq \left\{ f = f^{(L)} \circ \cdots \circ f^{(1)} \,:\, \begin{array}{l} f^{(\ell)} \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell}), \\ \ell = 1, \dots, L \end{array} \right\}.$$

An immediate corollary to Lemma 3.12 is that functions in $\mathscr{R}\,\mathrm{BV}^2_{\mathrm{deep}}(\mathbb{R}^{d_0};\cdots;\mathbb{R}^{d_L})$ are Lipschitz continuous.

Corollary 3.17. *Let $f = f^{(L)} \circ \cdots \circ f^{(1)} \in \mathscr{R}\,\mathrm{BV}^2_{\mathrm{deep}}(\mathbb{R}^{d_0};\cdots;\mathbb{R}^{d_L})$. Then, f is Lipschitz continuous and satisfies the Lipschitz bound*

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

To this end, we have the folllowing representer theorem for learning with deep ReLU networks.

Theorem 3.18. *Let $\ell(\cdot,\cdot) : \mathbb{R}^{d_L} \times \mathbb{R}^{d_L} \to \mathbb{R}_{\geq 0}$ be a lower semicontinuous loss function in its second argument and let $\{(\boldsymbol{x}_m, \boldsymbol{y}_m)\}_{m=1}^M \subset \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ be a given set of distinct data points with $M > 0$, and let the regularization hyperparameter $\lambda > 0$ be fixed. Then, there exists a solution the learning problem*

$$\min_{\substack{f^{(\ell)} \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell}) \\ \ell = 1,\dots,L \\ f = f^{(L)} \circ \cdots \circ f^{(1)}}} \sum_{m=1}^{M} \ell(\boldsymbol{y}_m, f(\boldsymbol{x}_m)) + \lambda \sum_{\ell=1}^{L} \|f^{(\ell)}\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})} \tag{3.22}$$

of the form

$$f_{\mathrm{deep}}(\boldsymbol{x}) = \boldsymbol{x}^{(L)}, \tag{3.23}$$

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

$$\begin{cases} \boldsymbol{x}^{(0)} \coloneqq \boldsymbol{x}, \\ \boldsymbol{x}^{(\ell)} \coloneqq \mathbf{V}^{(\ell)} \boldsymbol{\rho}(\mathbf{W}^{(\ell)} \boldsymbol{x}^{(\ell-1)} - \boldsymbol{b}^{(\ell)}) + \mathbf{C}^{(\ell)} \boldsymbol{x}^{(\ell-1)} + \boldsymbol{c}_0^{(\ell)}, \quad \ell = 1, \dots, L, \end{cases} \tag{3.24}$$

where $\boldsymbol{\rho}$ *applies* $\rho = \max\{0, \cdot\}$ *component-wise and for* $\ell = 1, \dots, L$, $\mathbf{V}^{(\ell)} \in \mathbb{R}^{d_\ell \times N^{(\ell)}}$, $\mathbf{W}^{(\ell)} \in \mathbb{R}^{N^{(\ell)} \times d_{\ell-1}}$, $\boldsymbol{b}^{(\ell)} \in \mathbb{R}^{N^{(\ell)}}$, $\mathbf{C}^{(\ell)} \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$, *and* $\boldsymbol{c}_0^{(\ell)} \in \mathbb{R}^{d_\ell}$, *where* $N^{(\ell)} \leq M d_\ell$.

Remark 3.19. In (3.22), we regularize the full Banach norms of the functions rather than the $\mathscr{R}\,\mathrm{TV}^2$-seminorms of each component to simplify the proof that there exist solutions to the variational problem. Similar results hold when only considering $\mathscr{R}\,\mathrm{TV}^2$-seminorms in the regularizer, though more care has to be taken to prove that solutions exist.

The neural network architecture that appears in (3.24) can be seen in Figure 3.3. Moreover, this exact architecture was recently studied in the empirical work in Golubeva et al. (2021), and is referred to as a deep ReLU network with *linear bottlenecks*. Since the variational problem in (3.22) is reminiscent of the variational problems studied in variational spline theory and since the resulting deep ReLU network solution in (3.23) is a continuous piecewise-linear function, in a similar vein to Unser (2019), we refer to such functions as *deep ridge splines* of degree one.

Remark 3.20. Since the regularizer in (3.22) directly controls the $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}}; \mathbb{R}^{d_\ell})$-norm of each layer, we see from Corollary 3.17, that the variational problem is essentially regularizing a bound on the Lipschitz constant of the function.

Remark 3.21. The regularizer that appears in (3.22) can be replaced by

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

where $\psi_\ell : [0, \infty) \to \mathbb{R}$, $\ell = 0, \dots, L$ is a strictly increasing and convex function, and still admit a solution that takes the form of a deep neural network as in (3.23). Thus,

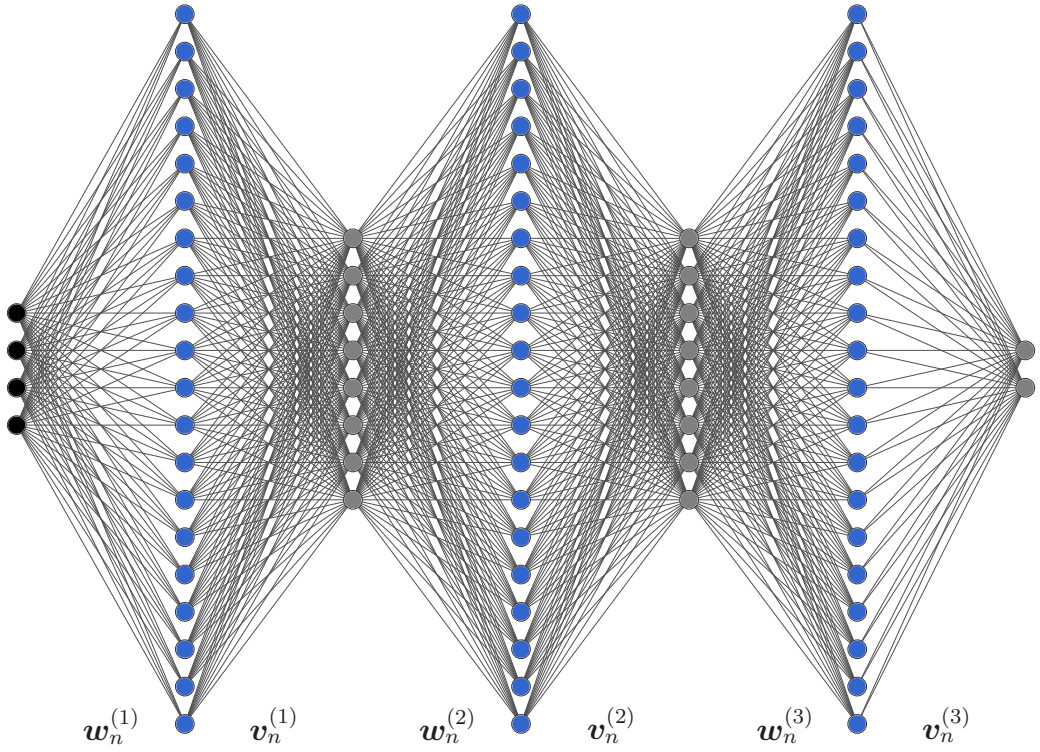


Figure 3.3: The architecture of the deep neural network in (3.24) in the case of $L = 3$ hidden layers. The black nodes denote input nodes, the blue nodes denote ReLU nodes, and the gray nodes denote linear nodes. Skip connection nodes are omitted for clarity.

there are many choices of regularization that result in a representer theorem for deep ReLU networks.

Remark 3.22. Notice that (3.23) is precisely the standard L-hidden layer deep ReLU network architecture with *rank-bounded weight matrices* and *skip connections*. Indeed, the weight matrix of the ℓth layer is $\mathbf{A}^{(\ell)} \coloneqq \mathbf{W}^{(\ell+1)}\mathbf{V}^{(\ell)}$. More specifically, by dropping biases and skip connections for clarity, we see that $f_{\text{deep}}(\boldsymbol{x})$ in (3.23) can be computed recursively as

$$\begin{cases} \widetilde{\boldsymbol{x}}^{(0)} \coloneqq \boldsymbol{x}, \\ \widetilde{\boldsymbol{x}}^{(\ell)} \coloneqq \boldsymbol{\rho}(\mathbf{A}^{(\ell-1)}\widetilde{\boldsymbol{x}}^{(\ell-1)}), \quad \ell = 1, \dots, L, \\ f_{\text{deep}}(\boldsymbol{x}) \coloneqq \mathbf{A}^{(L)}\widetilde{\boldsymbol{x}}^{(L)}, \end{cases} \tag{3.25}$$

where

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

From the dimensions of $\mathbf{V}^{(\ell)}$ and $\mathbf{W}^{(\ell)}$ in Theorem 3.18, we see that for $\ell = 0, \dots, L$, $\operatorname{rank}(\mathbf{A}^{(\ell)}) \leq \min\{Md_{\ell+1}, d_\ell\}$ and $\operatorname{rank}(\mathbf{A}^{(L)}) \leq d_L$. In a typical scenario, where the $\{d_\ell\}_{\ell=1}^{L}$ are of the same order, this implies that $\operatorname{rank}(\mathbf{A}^{(\ell)}) \leq d_\ell$.

Remark 3.23. The architecture of the network in (3.24) is not restrictive of what functions can be represented by such a network. In particular, the architecture in (3.24) is as expressive as the standard deep ReLU network architecture with hidden layer widths of $d_1, \dots, d_L$.

Before proving Theorem 3.18, we require two intermediary results.

Lemma 3.24. *Consider the problem of interpolating the scattered data* $\{(\boldsymbol{x}_m, y_m)\}_{m=1}^{M} \subset \mathbb{R}^d \times \mathbb{R}$ *with* $M > 0$. *Then, under the hypothesis of feasibility (i.e.,* $y_m = y_n$ *whenever* $\boldsymbol{x}_m = \boldsymbol{x}_n$*), there exists a solution to the variational problem*

$$\min_{f \in \mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)} \|f\|_{\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)} \quad \text{s.t.} \quad f(\boldsymbol{x}_m) = y_m,\ m = 1, \dots, M \tag{3.26}$$

of the form

$$f_{\mathrm{ReLU}}(\boldsymbol{x}) = \sum_{n=1}^{N} v_n \operatorname{ReLU}(\boldsymbol{w}_n^{\mathsf{T}}\boldsymbol{x} - b_n) + \boldsymbol{c}^{\mathsf{T}}\boldsymbol{x} + c_0, \tag{3.27}$$

where $N \leq M$, $v_n \in \mathbb{R}$, $\boldsymbol{w}_n \in \mathbb{S}^{d-1}$, $b_n \in \mathbb{R}$, $\boldsymbol{c} \in \mathbb{R}^d$, *and* $c_0 \in \mathbb{R}$.

Proof. Since the Dirac impulse is weak* continuous on $\mathscr{R}\mathrm{BV}^2(\mathbb{R}^d)$ from Lemma 3.11, by the abstract representer theorem in Unser (2021, Theorem 2), there exists a solution to the variational problem (3.26). Let s be a (not necessarily unique) solution

to (3.26). This solution must be a minimizer of

$$\min_{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} \mathscr{R}\,\mathrm{TV}^2(f) \quad \text{s.t.} \quad \begin{cases} f(\boldsymbol{x}_m) = y_m, & m = 1, \dots, M, \\ f(\mathbf{0}) = s(\mathbf{0}), & \\ f(\boldsymbol{e}_\ell) = s(\boldsymbol{e}_\ell), & \ell = 1, \dots, d. \end{cases}$$

By Corollary 3.13, there exists a solution to the above display that takes the form in (3.27) with $N \leq M$ neurons, so we can always find a solution to the original problem in (3.26) of the form in (3.27). □

Lemma 3.25. *Consider the problem of interpolating the scattered data* $\{(\boldsymbol{x}_m, \boldsymbol{y}_m)\}_{m=1}^M \subset \mathbb{R}^d \times \mathbb{R}^D$ *with* $M > 0$. *Then, under the hypothesis of feasibility (i.e.,* $\boldsymbol{y}_m = \boldsymbol{y}_n$ *whenever* $\boldsymbol{x}_m = \boldsymbol{x}_n$*), there exists a solution to the variational problem*

$$\min_{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)} \|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)} \quad \text{s.t.} \quad f(\boldsymbol{x}_m) = \boldsymbol{y}_m,\ m = 1, \dots, M \tag{3.28}$$

of the form

$$f_{\mathrm{ReLU}}(\boldsymbol{x}) = \sum_{n=1}^N \boldsymbol{v}_n\,\mathrm{ReLU}(\boldsymbol{w}_n^\mathsf{T}\boldsymbol{x} - b_n) + \mathbf{C}\boldsymbol{x} + \boldsymbol{c}_0, \tag{3.29}$$

where $N \leq MD$, $\boldsymbol{v}_n \in \mathbb{R}^D$, $\boldsymbol{w}_n \in \mathbb{S}^{d-1}$, $b_n \in \mathbb{R}$, $\mathbf{C} \in \mathbb{R}^{D \times d}$, *and* $\boldsymbol{c}_0 \in \mathbb{R}^D$. *Moreover, there always exists a solution of the form in (3.29) in which* $\boldsymbol{v}_n$ *is* 1*-sparse.*

Proof. By Lemma 3.11, the point evaluation operator is component-wise weak* continuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$. Therefore, the measurement functionals

$$\langle \nu_{m,k}, f \rangle \coloneqq f_k(\boldsymbol{x}_m),\ m = 1, \dots, M,\ k = 1, \dots, D,$$

where $f = (f_1, \dots, f_D) \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$ and $\langle \cdot, \cdot \rangle$ denotes the pairing of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$ and its continuous dual, are contained in the predual of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$, i.e., they are weak* continuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d;\mathbb{R}^D)$. Moreover, these functionals are linearly independent[5]. Therefore, the problem in (3.28) satisfies the hypotheses of Unser

[5]Assuming that $\boldsymbol{x}_m \neq \boldsymbol{x}_n$ for $m \neq n$.

(2021, Theorem 2) and so a solution to (3.28) exists. Next, note that we can rewrite the problem in (3.28) as

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

where $\boldsymbol{y}_m = (y_{m,1},\dots,y_{m,D}) \in \mathbb{R}^D$. Let $\widetilde{s} = (\widetilde{s}_1,\dots,\widetilde{s}_D)$ be a (not necessarily unique) solution to (3.28). From the above display we see that this solution must satisfy

$$\widetilde{s}_k \in \operatorname*{arg\,min}_{f\in\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} \|f\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} \quad \text{s.t.} \quad f(\boldsymbol{x}_m) = y_{m,k},\ m=1,\dots,M, \tag{3.30}$$

for $k=1,\dots,D$. To see this, note that if the above display did not hold, it would contradict the optimality of $\widetilde{s}$. By Lemma 3.24, there exists a solution to the above display that takes the form in (3.27) with $N_k \leq M$ neurons. By combining these solutions into a single vector-valued function with potential combining of neurons[6] we see that there exists a solution to the original problem in (3.28) that takes the form in (3.29) with $N \leq N_1 + \cdots + N_D \leq MD$ neurons. If no neurons combine, each $\boldsymbol{v}_n$ is 1-sparse. □

Remark 3.26. One could also write a solution of (3.28) such that each output is completely independent of any other output, i.e., the outputs are completely decoupled. This corresponds to fitting the data with D separate single-hidden layer ReLU networks. This follows from the fact that s_k is a minimizer to the problem in (3.30). This corresponds to the representation in (3.29) having each $\boldsymbol{v}_n$ being 1-sparse.

Proof of Theorem 3.18. Given $f = f^{(L)} \circ \cdots \circ f^{(1)}$ such that $f^{(\ell)} \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})$, $\ell = 1,\dots,L$, write

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

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

for the objective value of f. Next, consider an arbitrary $g = g^{(L)} \circ \cdots \circ g^{(1)}$ such that $g^{(\ell)} \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}}; \mathbb{R}^{d_\ell})$, $\ell = 1, \ldots, L$, with objective value $C \coloneqq \mathcal{J}(g)$. We may transform the unconstrained problem in (3.22) into the equivalent constrained problem

$$\min_{\substack{f^{(\ell)} \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}}; \mathbb{R}^{d_\ell}) \\ \ell = 1, \ldots, L \\ f = f^{(L)} \circ \cdots \circ f^{(1)}}} \mathcal{J}(f) \quad \text{s.t.} \quad \|f^{(\ell)}\|_{\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}}; \mathbb{R}^{d_\ell})} \leq C/\lambda,\ \ell = 1, \ldots, L. \tag{3.31}$$

This transformation is valid since any function that does not satisfy the constraints in (3.31) has a strictly larger objective value than g, and is therefore not in the solution set. For $f_0 = f_0^{(L)} \circ \cdots \circ f_0^{(1)}$, $f_0^{(\ell)} \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}}; \mathbb{R}^{d_\ell})$, $\ell = 1, \ldots, L$, we show that the map $f_0^{(\widetilde{\ell})} \mapsto \mathcal{J}(f_0)$, for a fixed $\widetilde{\ell} \in \{1, \ldots, L\}$, is weak* lower semi-continuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\widetilde{\ell}-1}}; \mathbb{R}^{d_{\widetilde{\ell}}})$. First notice that the map $f_0^{(\widetilde{\ell})} \mapsto f_0(\boldsymbol{x}_0)$, for any $\boldsymbol{x}_0 \in \mathbb{R}^d$, is component-wise weak* continuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\widetilde{\ell}-1}}; \mathbb{R}^{d_{\widetilde{\ell}}})$. Indeed, since each $f_0^{(\ell)}$, $\ell = 1, \ldots, L$, is component-wise continuous by Lemma 3.12 and since the point evaluation is component-wise weak* continuous by Lemma 3.11, the map $f_0^{(\widetilde{\ell})} \mapsto f_0^{(L)} \circ \cdots \circ f_0^{(1)}(\boldsymbol{x}_0)$ is made up of compositions of component-wise continuous and component-wise weak* continuous functions, and is therefore itself component-wise weak* continuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\widetilde{\ell}-1}}; \mathbb{R}^{d_{\widetilde{\ell}}})$. Next, since the loss function is lower semi-continuous and every norm is weak* continuous on its corresponding Banach space, we have that $f_0^{(\widetilde{\ell})} \mapsto \mathcal{J}(f_0)$ is weak* lower semi-continuous on $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\widetilde{\ell}-1}}; \mathbb{R}^{d_{\widetilde{\ell}}})$. Therefore, $(f_0^{(1)}, \ldots, f_0^{(L)}) \mapsto \mathcal{J}(f)$ is weak* lower semi-continuous over the Cartesian product search space[7] in (3.31). Finally, by the Banach–Alaoglu theorem (Rudin, 1991, Chapter 3), the feasible set in (3.31) is weak* compact. Thus, there exists a solution to (3.31) (and subsequently (3.22)) by the Weierstrass extreme value theorem on general topological spaces (Kurdila and Zabarankin, 2006, Chapter 5).

Let $\widetilde{s} = \widetilde{s}^{(L)} \circ \cdots \circ \widetilde{s}^{(1)}$ be a (not necessarily unique) solution to (3.22). By applying $\widetilde{s}$ to each data point $\boldsymbol{x}_m$, $m = 1, \ldots, M$, we can recursively compute the intermediate vectors $\boldsymbol{z}_{m,\ell} \in \mathbb{R}^{d_\ell}$ as follows

[7] The search space is the Cartesian product $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_0}; \mathbb{R}^{d_1}) \times \cdots \times \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{L-1}}; \mathbb{R}^{d_L})$.

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

The solution $\widetilde{s}$ must satisfy

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

for $\ell = 1, \dots, L$. To see this, note that if the above display did not hold, it would contradict the optimality of $\widetilde{s}$. By Lemma 3.25, there always exists a solution to the above display that enforces the form of the solution in (3.23). □

3.3.3 New Regularization Methods for Neural Networks

With the results developed in this section, we have several new, principled forms of regularization for deep neural networks. As we saw in Section 3.3.1, the solutions to the problem of training a shallow ReLU network with weight decay or path-norm regularization are minimum $\mathscr{R}\,\mathrm{TV}^2$-seminorm solutions to data-fitting variational problems over $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$. This variational formulation of learning, particularly the results about deep neural networks developed in Section 3.3.2, provides several extensions/modifications of weight decay and path-norm regularization as well as provides new theoretical support and insight for a number of empirical findings in deep learning. In particular, these results characterize the functional properties of neural networks trained with weight decay—the functions they represent are regular in a precise (i.e., $\mathscr{R}\,\mathrm{BV}^2$) sense. The optimal solutions to the variational problem require skip connections between layers, which provides a new theoretical explanation for the benefits skip connections provide in practice (He et al., 2016). The sparse nature of our solutions sheds new light on the roles of sparsity and redundancy in deep learning, ranging from "drop-out" (Hinton et al., 2012b) to the "lottery ticket hypothesis" (Frankle and Carbin, 2018). And finally, low-rank weight matrices are a natural by-product of our variational formulation that has precedent in practical studies of deep neural networks; it has been empirically observed that

low-rank weight matrices can speed up learning (Ba and Caruana, 2014) and improve accuracy (Golubeva et al., 2021), robustness (Sanyal et al., 2019), and computational efficiency (Wang et al., 2021) of deep neural networks.

Given a deep neural network $s = s^{(L)} \circ \cdots \circ s^{(1)}$ as in (3.23), by a direct calculation we have

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

where $\boldsymbol{v}_n^{(\ell)}$ is the nth column of $\mathbf{V}^{(\ell)}$ and $\boldsymbol{w}_n^{(\ell)}$ is the nth row of $\mathbf{W}^{(\ell)}$. Therefore, the solutions to the finite-dimensional neural network training problem

$$\min_{\boldsymbol{\theta}\in\Theta} \sum_{m=1}^{M} \ell(\boldsymbol{y}_m, f_{\boldsymbol{\theta}}(\boldsymbol{x}_m)) + \lambda \sum_{\ell=1}^{L} \left(\sum_{n=1}^{N^{(\ell)}} \|\boldsymbol{v}_n^{(\ell)}\|_1 \|\boldsymbol{w}_n^{(\ell)}\|_2 + \sum_{k=1}^{D} \left(|f_{\boldsymbol{\theta},k}^{(\ell)}(\mathbf{0})| + \sum_{m=1}^{d} |f_{\boldsymbol{\theta},k}^{(\ell)}(\boldsymbol{e}_m) - f_{\boldsymbol{\theta},k}^{(\ell)}(\mathbf{0})| \right) \right) \tag{3.32}$$

are solutions to the variational problem in (3.22) so long as $N^{(\ell)} \geq M d_\ell$, where $\{(\boldsymbol{x}_m, \boldsymbol{y}_m)\}_{m=1}^{M} \subset \mathbb{R}^{d_0} \times \mathbb{R}^{d_L}$ is a scattered data set, $\ell(\cdot,\cdot)$ is an arbitrary non-negative lower semi-continuous loss function, and $\lambda > 0$ is an adjustable regularization parameter. By the same argument as in the proof of Theorem 3.15, the problem in the above display is equivalent to the problem

$$\min_{\boldsymbol{\theta}\in\Theta} \sum_{m=1}^{M} \ell(\boldsymbol{y}_m, f_{\boldsymbol{\theta}}(\boldsymbol{x}_m)) + \lambda \sum_{\ell=1}^{L} \left(\frac{\|\mathbf{V}^{(\ell)}\|_{1,2}^2 + \|\mathbf{W}^{(\ell)}\|_{\mathsf{F}}^2}{2} + \sum_{k=1}^{D} \left(|f_{\boldsymbol{\theta},k}^{(\ell)}(\mathbf{0})| + \sum_{m=1}^{d} |f_{\boldsymbol{\theta},k}^{(\ell)}(\boldsymbol{e}_m) - f_{\boldsymbol{\theta},k}^{(\ell)}(\mathbf{0})| \right) \right), \tag{3.33}$$

where

$$\|\mathbf{V}^{(\ell)}\|_{1,2}^2 \coloneqq \sum_{n=1}^{N^{(\ell)}} \|\boldsymbol{v}_n^{(\ell)}\|_1^2$$

is the mixed $\ell^1\ell^2$-norm of $\mathbf{V}^{(\ell)}$ and $\|\cdot\|_{\mathsf{F}}$ is the usual Frobenius norm of a matrix. The problems in (3.32) and (3.33) take the form of neural network training problems with new, principled forms of regularization. Moreover, due to the sparsity-promoting nature of the $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^{d_{\ell-1}};\mathbb{R}^{d_\ell})$-norms, the regularizers that appear in (3.32) and

(3.33) promote sparse (in the sense of the number of active neurons) deep ReLU network solutions.

One could also consider a different variational problem from (3.22) in which only the $\mathscr{R}\,\mathrm{TV}^2$-seminorms are penalized rather than the $\mathscr{R}\,\mathrm{BV}^2$-norms. In that case, the resulting (equivalent) regularizers would be

$$\sum_{\ell=1}^{L}\sum_{k=1}^{K^{(\ell)}}\|\boldsymbol{v}_k^{(\ell)}\|_1\|\boldsymbol{w}_k^{(\ell)}\|_2 \tag{3.34}$$

and

$$\frac{1}{2}\sum_{\ell=1}^{L}\|\mathbf{V}^{(\ell)}\|_{1,2}^2+\|\mathbf{W}^{(\ell)}\|_{\mathsf{F}}^2. \tag{3.35}$$

In this setting, with a particular choice of $\{\psi_\ell\}_{\ell=0}^{L}$ in Remark 3.21, we may also consider the regularizer

$$\prod_{\ell=1}^{L}\sum_{k=1}^{K^{(\ell)}}\|\boldsymbol{v}_k^{(\ell)}\|_1\|\boldsymbol{w}_k^{(\ell)}\|_2. \tag{3.36}$$

We can view (3.35) as a modification/extension of the regularizer that corresponds to training a deep neural network with weight decay. We can also view (3.34) and (3.36) as modificatins/extensions of the well-known path-norm regularizer for deep neural networks. In fact, the regularizer that appears in (3.36) is essentially an upper bound on the standard path-norm for deep neural networks. Indeed, consider a deep neural network mapping $\mathbb{R}^d\to\mathbb{R}$ as in (1.13). The usual path-norm of such a neural network takes the form

$$\sum_{n_L=1}^{N^{(L)}}\sum_{n_{L-1}=1}^{N^{(L-1)}}\cdots\sum_{n_1=1}^{N^{(1)}}\sum_{n_0=1}^{d}|a_{n_0,n_1}||a_{n_1,n_2}|\cdots|a_{n_{L-1},n_L}||a_{n_L}|, \tag{3.37}$$

where $a_{n_\ell,n_{\ell+1}}$ denotes the $(n_\ell,n_{\ell+1})$th entry in $\mathbf{A}^{(\ell)}$ and a_{n_L} denotes the n_Lth entry in $\boldsymbol{a}^{(L)}$. We refer the reader to Neyshabur et al. (2015a, 2017, 2015c); Barron and Klusowski (2019) for more details about the path-norm for deep neural networks. If

we consider the parameterization of a deep neural network as in (3.23), we have that

$$|a_{n_\ell, n_{\ell+1}}| = |\boldsymbol{v}_n^{(\ell)\mathsf{T}} \boldsymbol{w}_n^{(\ell+1)}| \leq \|\boldsymbol{v}_n^{(\ell)}\|_2 \|\boldsymbol{w}_n^{(\ell+1)}\|_2 \leq \|\boldsymbol{v}_n^{(\ell)}\|_1 \|\boldsymbol{w}_n^{(\ell+1)}\|_2. \tag{3.38}$$

Therefore,

$$\begin{aligned}\prod_{\ell=1}^{L} \sum_{n=1}^{N^{(\ell)}} \|\boldsymbol{v}_N^{(\ell)}\|_1 \|\boldsymbol{w}_N^{(\ell)}\|_2 &= \sum_{n_L=1}^{N^{(L)}} \cdots \sum_{n_1=1}^{N^{(1)}} \|\boldsymbol{w}_{n_1}^{(1)}\|_2 \|\boldsymbol{v}_{n_1}^{(1)}\|_1 \|\boldsymbol{w}_{n_2}^{(2)}\|_2 \|\boldsymbol{v}_{n_2}^{(2)}\|_1 \cdots \|\boldsymbol{w}_{n_L}^{(L)}\|_2 \|\boldsymbol{v}_{n_L}^{(L)}\|_1 \\ &\geq \sum_{n_L=1}^{N^{(L)}} \cdots \sum_{n_1=1}^{N^{(1)}} \|\boldsymbol{w}_{n_1}^{(1)}\|_2 \, |a_{n_1,n_2}| \cdots \left|a_{n_{L-1},n_L}\right| \, \|\boldsymbol{v}_{n_L}^{(L)}\|_1,\end{aligned}$$

where the last line holds from (3.38). We see that the last line in the above display is the same as the path-norm in (3.37), apart from how it treats weights in the first and last layers. We also remark that the work of Barron and Klusowski (2019), the authors show that the path-norm in (3.37) controls the Rademacher and Gaussian complexity of deep ReLU networks.

Chapter 4

Approximation and Estimation with Ridge Splines

A fundamental problem in approximation theory (resp. nonparametric statistics) is to understand the best[1] approximation (resp. estimation error) rate for functions that lie in certain function spaces, with respect to some kind of dictionary. These rates are typically measured with respect to the L^2- or L^∞-norm. In order to quantify these rates with respect to these norms, we require the Banach spaces of interest to be continuously embedded in an L^2- or L^∞-space. In this dissertation, we are interested in the Banach spaces $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$, $k \in \mathbb{N}$. Clearly these spaces are not continuously embedded in $L^2(\mathbb{R}^d)$ or $L^\infty(\mathbb{R}^d)$. Indeed, polynomials of degree strictly less than k are included in $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$, which are not in $L^2(\mathbb{R}^d)$ or $L^\infty(\mathbb{R}^d)$. Thus, in order to discuss L^2- and L^∞-approximation (resp. estimation error) rates for such functions, we must consider the restriction of these functions to a bounded domain $\Omega \subset \mathbb{R}^d$ so that it makes sense to discuss the $L^2(\Omega)$ and $L^\infty(\Omega)$-norms of a function in $\mathscr{R}\,\mathrm{BV}^k(\Omega)$.

In this chapter, we first study the approximation rates for functions in the appropriately defined spaces $\mathscr{R}\,\mathrm{BV}^k(\Omega)$, $k \in \mathbb{N}$, where $\Omega \subset \mathbb{R}^d$ is a bounded domain. We show that the approximation rates do not grow with the input dimension, shedding light

[1] Best in that no other approximation (resp. estimation) scheme can result in a better rate.

on the phenomenon that neural networks seem to break the curse of dimensionality. These results follow by showing that $\mathscr{R}\,\mathrm{BV}^k(\Omega)$ is equivalent (as a Banach space) to the so-called kth-order variation space and invoking previously derived approximation rates for the variation spaces.

We also discuss what kinds of functions lie in $\mathscr{R}\,\mathrm{BV}^k(\Omega)$. In particular, we show that these spaces are a *mixed variation* space, a term coined by Donoho (2000) to refer to function spaces that contain functions that are isotropic and very regular in all directions as well as functions that are anisotropic and very unregular in only a few directions. Using these results, we study the problem of estimating (i.e., learning) a function from noisy point evaluation measurements with shallow ReLU networks, and show that the mean-squared error (i.e., expected L^2-error) of the learned function from the data-generating function also does not grow with the input dimension. Finally, we show that linear methods (which include kernel methods) are suboptimal for the estimation problem by quantifying an explicit gap between linear and nonlinear estimation method via minimax rates.

4.1 $\mathscr{R}\,\mathrm{BV}^k(\Omega)$: Restricting $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ to a Bounded Domain $\Omega \subset \mathbb{R}^d$

In this section we define the $\mathscr{R}\,\mathrm{BV}^k$-spaces, $k \in \mathbb{N}$, on a bounded domain. We define the $\mathscr{R}\,\mathrm{BV}^k$-spaces on a bounded domain $\Omega \subset \mathbb{R}^d$ using the standard approach of considering restrictions of functions in $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$. This provides the following definition:

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

where $\mathcal{D}'(\Omega)$ denotes the space of distributions on Ω. Similarly, we can define the kth-order total variation in the (filtered) Radon domain of a function f defined on a

bounded domain $\Omega \subset \mathbb{R}^d$:

$$\mathscr{R}\,\mathrm{TV}^k_\Omega(f) \coloneqq \inf_{g \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)} \mathscr{R}\,\mathrm{TV}^k(g) \quad \text{s.t.} \quad g|_\Omega = f. \tag{4.1}$$

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

$$\mathscr{R}\,\mathrm{BV}^k(\Omega) = \big\{ f \in \mathcal{D}'(\Omega) : \ \mathscr{R}\,\mathrm{TV}^k_\Omega(f) < \infty \big\}.$$

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

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

4.1.1 Extensions From $\mathscr{R}\,\mathrm{BV}^k(\Omega)$ to $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$

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

Lemma 4.1. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. Given $f \in \mathscr{R}\,\mathrm{BV}^k(\Omega)$, there exists an extension $f_{\mathsf{ext}} \in \mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ that admits an integral representation*

$$f_{\mathsf{ext}}(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \rho_k(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x} - t)\,\mathrm{d}\mu(\boldsymbol{\alpha}, t) + q(\boldsymbol{x}),$$

where $q(\cdot)$ is a polynomial of degree $< k$, such that $\operatorname{supp}\mu \subset Z_\Omega$, where

$$Z_\Omega \coloneqq \overline{\{\boldsymbol{z} = (\boldsymbol{\alpha}, t) \in \mathbb{S}^{d-1}\times\mathbb{R} : \ \{\boldsymbol{x} \in \mathbb{R}^d : \ \boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x} = t\} \cap \Omega \neq \varnothing\}}, \tag{4.2}$$

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

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

Remark 4.2. The set Z_Ω simply excludes activation functions that are polynomial functions (no activation threshold) when restricted to Ω.

Proof. Let $\boldsymbol{\eta} = (\boldsymbol{\phi}, \boldsymbol{p})$ be a biorthogonal system for $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$. Given $f \in \mathscr{R}\,\mathrm{BV}^k(\Omega)$, suppose there exists an extension $\widetilde{f}_{\mathsf{ext}}$ such that $\widetilde{f}_{\mathsf{ext}}\big|_\Omega = f$ and $\mathscr{R}\,\mathrm{TV}^k_\Omega(f) = \mathscr{R}\,\mathrm{TV}^k(\widetilde{f}_{\mathsf{ext}})$ whose direct-sum decomposition from Theorem 3.8 takes the form

$$\widetilde{f}_{\mathsf{ext}} = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} g_{k,\boldsymbol{\eta}}(\cdot, \boldsymbol{z})\,\mathrm{d}\widetilde{\mu}(\boldsymbol{z}) + \widetilde{q}, \tag{4.3}$$

and $\operatorname{supp}\widetilde{\mu} \not\subset Z_\Omega$. Next, notice that given $g_{k,\boldsymbol{\eta}}(\cdot, \boldsymbol{z})$, where $\boldsymbol{z} \notin Z_\Omega$, we have that $g_{k,\boldsymbol{\eta}}(\cdot, \boldsymbol{z})|_\Omega$ is a polynomial of degree $< k$. 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 $g_{k,\boldsymbol{\eta}}(\cdot, \boldsymbol{z})$ where $\boldsymbol{z} \notin Z_\Omega$ in the integrand of (4.3) into the polynomial term in the direct-sum decomposition, a contradiction. Therefore, there exists an extension $f_{\mathsf{ext}} \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d)$ that admits an integral representation

$$f_{\mathsf{ext}}(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} g_{k,\boldsymbol{\eta}}(\boldsymbol{x}, (\boldsymbol{\alpha}, t))\,\mathrm{d}\mu(\boldsymbol{\alpha}, t) + q(\boldsymbol{x}) \tag{4.4}$$

such that $\operatorname{supp}\mu \subset Z_\Omega$, where μ is an even (resp. odd) measure when k is even (resp. odd) and q is a polynomial of degree $< k$.

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

$$f_{\mathsf{ext}}(\boldsymbol{x}) = \int_{Z_\Omega} \rho_k(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x} - t)\,\mathrm{d}\mu(\boldsymbol{\alpha}, t) + \widetilde{q}(\boldsymbol{x}), \tag{4.5}$$

where we combine the polynomial terms from $g_{k,\boldsymbol{\eta}}$ (defined in Lemma 3.7) and q into the new polynomial $\widetilde{q}$. Moreover, with the above representation we have that $\mathscr{R}\,\mathrm{TV}^k_\Omega(f) = \|\mu \llcorner Z_\Omega\|_{\mathcal{M}(Z_\Omega)}$. We also remark that although μ is even (resp. odd) when k is even (resp. odd), we can replace μ with a generic, i.e., not restricted to being even/odd, measure $\widetilde{\mu} \in \mathcal{M}(Z_\Omega)$ by noting that integrating against an even/odd measure in (4.4) is exactly the same as integrating against a generic measure due to

the symmetry/antisymmetry of ρ_k defined in (1.17). This generic, i.e., not even/odd, measure has the same $\mathcal{M}$-norm as the even/odd measure. □

Remark 4.3. When

$$\Omega = \mathbb{B}_1^d \coloneqq \big\{\boldsymbol{x} \in \mathbb{R}^d : \ \|\boldsymbol{x}\|_2 \leq 1\big\}, \tag{4.6}$$

the Euclidean unit ball in $\mathbb{R}^d$, we have that Z_Ω from (4.2) is exactly

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

Therefore, from Theorem 4.1, we can identify functions in $f \in \mathscr{R}\,\mathrm{BV}^k(\mathbb{B}_1^d)$ with integral representations of the form

$$f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} \rho_k(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x} - t)\,\mathrm{d}\mu(\boldsymbol{\alpha}, t) + c(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{B}_1^d.$$

Remark 4.4. When $d = 1$, the space $\mathscr{R}\,\mathrm{BV}^k(\mathbb{B}_1^d)$ is exactly the classical kth-order bounded variation spaces defined on $[-1, 1]$:

$$\mathrm{BV}^k[-1, 1] \coloneqq \big\{f \in \mathcal{D}[-1, 1] : \ \mathrm{TV}^k_{[-1,1]}(f) < \infty\big\},$$

where

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

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

4.1.2 Representer Theorems over $\mathscr{R}\,\mathrm{BV}^k(\Omega)$

It turns out that data-fitting variational problems over $\mathscr{R}\,\mathrm{BV}^k(\Omega)$ also admit representer theorems similar to Theorem 3.2. This is summarized in the following theorem.

Theorem 4.5. *Consider the following setting:*

1. *The loss function* $\ell(\cdot,\cdot):\mathbb{R}\times\mathbb{R}\to\mathbb{R}_{\geq 0}$ *is convex, coercive, and lower semicontinuous in its second argument.*

2. *The linear measurement functionals* $h_m : \mathscr{R}\,\mathrm{BV}^k(\Omega)\to\mathbb{R} : f\mapsto\langle h_m, f\rangle$*, where* $m=1,\ldots,M$*, are linearly independent and weak* continuous.*

3. *The number of measurements* M *is strictly greater than the dimension of the null space* $\mathcal{P}_{k-1}(\Omega)$*, the space of polynomials of degree* $k-1$*.*

4. *The regularization hyperparameter* $\lambda>0$ *is fixed.*

Then, for any fixed $\boldsymbol{y}\in\mathbb{R}^M$*, the solution set to the data-fitting variational problem*

$$\mathcal{V} := \operatorname*{arg\,min}_{f\in\mathscr{R}\,\mathrm{BV}^k(\Omega)} \sum_{m=1}^{M} \ell(y_m,\langle h_m, f\rangle) + \lambda\,\mathscr{R}\,\mathrm{TV}^k_\Omega(f)$$

is nonempty, convex, and weak compact. If* $\ell(\cdot,\cdot)$ *is strictly convex (or if it imposes the equality* $y_m=\langle h_m,f\rangle$*, for* $m=1,\ldots,M$*), then the solution set* $\mathcal{V}$ *is the weak* closure of the convex hull of its extreme points, which can all be expressed as*

$$f_{\mathrm{ridge}}(\boldsymbol{x}) = \sum_{n=1}^{N_0} v_n \rho_k(\boldsymbol{w}_n^{\mathsf{T}}\boldsymbol{x} - b_n) + c(\boldsymbol{x}),$$

where $\{v_n\}_{n=1}^{N_0}\subset\mathbb{R}\setminus\{0\}$*,* $\{(\boldsymbol{w}_n,b_n)\}_{n=1}^{N_0}\subset Z_\Omega$ *as defined in (4.2),* $c(\cdot)\in\mathcal{P}_{k-1}(\Omega)$*, and* $N_0<M$*. The corresponding regularization cost, which is common to all solutions, is* $\mathscr{R}\,\mathrm{TV}^k(f_{\mathrm{ridge}}) = \sum_{n=1}^{N_0}|v_n| = \|\boldsymbol{v}\|_1$*.*

Proof. The proof is identical to the proof of Theorem 3.2 except we use the direct-sum decomposition from (4.4), which establishes an isometric isomorphism from $\mathscr{R}\,\mathrm{BV}^k(\Omega)$ to $\mathcal{M}_k(Z_\Omega)\times\mathcal{P}_{k-1}(\Omega)$, where $\mathcal{M}_k(Z_\Omega)$ is the subspace of even (resp. odd) measures when k is even (resp. odd). □

4.1.3 Applications to Learning with Shallow Neural Networks

Just as in Corollary 3.13, we can establish the weak* continuity of the Dirac impulse $\delta(\cdot - \boldsymbol{x}_0)$, $\boldsymbol{x}_0 \in \Omega$. In fact, on a bounded domain, the argument is even simpler than the proof of Lemma 3.11 since it boils down to the continuity[2] of the function $g_{k,\boldsymbol{\eta}}(\boldsymbol{x}_0, \cdot)|_\Omega$ since $\mathcal{M}(Z_\Omega) = (C(Z_\Omega))'$, which is clearly true for $k \geq 2$. We instantiate this result explicitly in the case of the ReLU in the following corollary.

Corollary 4.6. *Let $\ell(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ be a strictly convex, coercive, and lower semicontinuous loss function in its second argument, $\{(\boldsymbol{x}_m, y_m)\}_{m=1}^M \subset \Omega \times \mathbb{R}$, where $\Omega \subset \mathbb{R}^d$ is a bounded domain, be a given set of distinct data points with $M > d + 1$, and let the regularization hyperparameter $\lambda > 0$ be fixed. Then, the solution set to the learning problem*

$$\mathcal{V} \coloneqq \operatorname*{arg\,min}_{f \in \mathscr{R}\,\mathrm{BV}^2(\Omega)} \sum_{m=1}^M \ell(y_m, f(\boldsymbol{x}_m)) + \lambda\, \mathscr{R}\,\mathrm{TV}^2_\Omega(f) \tag{4.7}$$

is nonempty, convex, and weak compact. The solution set $\mathcal{V}$ is the weak* closure of the convex hull of its extreme points, which can all be expressed as*

$$f_{\mathrm{ReLU}}(\boldsymbol{x}) = \sum_{n=1}^{N_0} v_n\, \mathrm{ReLU}(\boldsymbol{w}_n^\mathsf{T} \boldsymbol{x} - b_n) + \boldsymbol{c}^\mathsf{T} \boldsymbol{x} + c_0,$$

where $\{v_n\}_{n=1}^{N_0} \subset \mathbb{R} \setminus \{0\}$, $\{(\boldsymbol{w}_n, b_n)\}_{n=1}^{N_0} \subset Z_\Omega$ as defined in (4.2), $\boldsymbol{c} \in \mathbb{R}^d$, $c_0 \in \mathbb{R}$, and $N_0 < M$. The corresponding regularization cost, which is common to all solutions, is $\mathscr{R}\,\mathrm{TV}^2(f_{\mathrm{ReLU}}) = \sum_{n=1}^{N_0} |v_n| = \|\boldsymbol{v}\|_1$.

Similar to the variational problem in Corollary 3.13, the variational problem in (4.7) can also be recast as a finite-dimensional neural network training problem with regularization terms corresponding to path-norm regularization and training a neural network with weight decay. For simplicity, suppose that $\Omega = \mathbb{B}_1^d$ as defined

[2] Recall that in the proof of Lemma 3.11, we had to established that the function $g_{2,\boldsymbol{\eta}}(\boldsymbol{x}_0, \cdot)$ was not only continuous, but also vanished at $\pm\infty$.

in (4.6). We can recast the variational problem in (4.7) as the following equivalent finite-dimensional neural network training problems

1. $\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{m=1}^{M} \ell(y_m, f_{\boldsymbol{\theta}}(\boldsymbol{x}_m)) + \lambda \sum_{n=1}^{N} |v_n| \|\boldsymbol{w}_n\|_2;$

2. $\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{m=1}^{M} \ell(y_m, f_{\boldsymbol{\theta}}(\boldsymbol{x}_m)) + \frac{\lambda}{2} \sum_{n=1}^{N} |v_n|^2 + \|\boldsymbol{w}_n\|_2^2,$

so long as $N \geq N_0$, where we are using the same notation as in Theorem 3.15, where $\boldsymbol{\Theta}$ is now constrained so that $\left\{\boldsymbol{x} \in \mathbb{R}^d : \ \boldsymbol{w}_n^\mathsf{T}\boldsymbol{x} - b_n\right\}$ intersects $\mathbb{B}_1^d$ for all $n = 1, \ldots, N$. The argument for this is identical to the argument in Theorem 3.15 and discussion that precedes the theorem statement.

4.2 $\mathscr{R}\,\mathrm{BV}^k(\Omega)$ and Previously Studied Function Spaces

Understanding the properties of shallow neural networks and their associated function spaces has received much attention since the 1990s starting with the seminal work of Barron (1993) 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 (Kurková and Sanguineti, 2001; Mhaskar, 2004; Bach, 2017; Siegel and Xu, 2021b).

In this section we discuss how $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ is related to previously studied neural function spaces, including the variation spaces. For simplicity we suppose that $\Omega = \mathbb{B}_1^d$ as defined in (4.6). Similar results as those stated in the sequel can be derived for more general bounded domains $\Omega \subset \mathbb{R}^d$, provided they have a sufficiently nice boundary.

4.2.1 Variation Spaces

Following the setup from Siegel and Xu (2021b), in the case of shallow neural networks with truncated power activation functions, the associated variation space for functions defined on $\mathbb{B}_1^d$ is defined as

$$\mathscr{V}^k(\mathbb{B}_1^d) \coloneqq \left\{ f \in \mathcal{D}'(\mathbb{B}_1^d) : \ f = \int_{\mathbb{S}^{d-1}\times[-2,2]} (\boldsymbol{\alpha}^\mathsf{T}(\cdot) - t)_+^{k-1} \, \mathrm{d}\mu(\boldsymbol{\alpha}, t) \right\},$$

where $k \in \mathbb{N}$ and $\mu \in \mathcal{M}(\mathbb{S}^{d-1} \times [-2,2])$. The reason for integrating the t variable over $[-2,2]$ is so that polynomials of degree at most $k-1$ are included in this space (see Siegel and Xu (2021b, Section 3) for more details). This space is known to be a Banach space when equipped with the norm

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

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

Theorem 4.7. *$\mathscr{R}\,\mathrm{BV}^k(\mathbb{B}_1^d)$ and $\mathscr{V}^k(\mathbb{B}_1^d)$ are equivalent Banach spaces (i.e., Banach spaces with equivalent norms).*

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

$$f(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times[-2,2]} (\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x} - t)_+^{k-1} \, \mathrm{d}\mu(\boldsymbol{\alpha}, t), \tag{4.8}$$

where $\|f\|_{\mathscr{V}^k} = \|\mu\|_{\mathcal{M}}$ (the inf in the definition of the $\mathscr{V}^k$-norm is achievable due to Korolev (2021, Proposition 3.21)). Next, given $g \in \mathscr{R}\,\mathrm{BV}^k(\mathbb{B}_1^d)$, we have from

Remark 4.3 the representation

$$g(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times[-1,1]} \rho_k(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x} - t)\,\mathrm{d}\nu(\boldsymbol{\alpha}, t) + c(\boldsymbol{x}),$$

where $\mathscr{R}\,\mathrm{TV}^k_{\mathbb{B}^d_1}(g) = \|\nu\|_{\mathcal{M}}$. Since ρ_k and $(\cdot)_+^{k-1}$ only differ by a multiplicative factor and a polynomial of degree at most $k-1$, it is clear that we can represent any function in $\mathscr{V}^2(\mathbb{B}^d_1)$ with the representation of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}^d_1)$ and vice-versa. Therefore, $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}^d_1) = \mathscr{V}^2(\mathbb{B}^d_1)$ (as sets). To see why the norms are equivalent, note that the only difference between the norms is how they handle the null space of the $\mathscr{R}\,\mathrm{TV}^k_{\mathbb{B}^d_1}(\cdot)$ seminorm. Since this null space is the space of polynomials of degree at most $k-1$, which is finite-dimensional, the norms are equivalent since all norms are equivalent on finite-dimensional spaces. Therefore, the norms $\mathscr{R}\,\mathrm{BV}^k(\mathbb{B}^d_1)$ and $\mathscr{V}^k(\mathbb{B}^d_1)$ have equivalent norms. □

4.2.2 Spectral Barron and Sobolev Spaces

The spectral Barron spaces were first studied by Barron (1993). On $\mathbb{R}^d$, these spaces are defined by

$$\mathscr{B}^s(\mathbb{B}^d_1) \coloneqq \Big\{ f \in \mathcal{S}'(\mathbb{R}^d) : \ \|(1 + \|\cdot\|_2)^s \widehat{f}(\cdot)\|_{\mathcal{M}} < \infty \Big\}.$$

In fact, one can show that these are Banach spaces when equipped with the norm

$$\|f\|_{\mathscr{B}^s} \coloneqq \|(1 + \|\cdot\|_2)^s \widehat{f}(\cdot)\|_{\mathcal{M}}$$

that are isometrically isomorphic to $\mathcal{M}(\mathbb{R}^d)$ (Parhi and Nowak, 2022b). On a bounded domain $\Omega \subset \mathbb{R}^d$, these spaces are defined by

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

which are Banach spaces when equipped with the norm

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

where $s \geq 0$ is any real number. In particular, Barron (1993) studied the first-order spectral Barron space on the Euclidean ball $\mathbb{B}_1^d$, $\mathscr{B}^1(\mathbb{B}_1^d)$, in his seminal work about approximation and estimation with shallow sigmoidal networks. The higher-order variants were studied by a number of authors (Klusowski and Barron, 2018; Xu, 2020; Siegel and Xu, 2021b). In particular, it was shown by Klusowski and Barron (2018); Xu (2020); Siegel and Xu (2021b) that $\mathscr{B}^k(\mathbb{B}_1^d) \overset{\text{c.}}{\hookrightarrow} \mathscr{V}^k(\mathbb{B}_1^d)$. Therefore, by Theorem 4.7, we have that $\mathscr{B}^k(\mathbb{B}_1^d) \overset{\text{c.}}{\hookrightarrow} \mathscr{R}\,\mathrm{BV}^k(\mathbb{B}_1^d)$. Moreover, it was shown by Xu (2020, Lemma 2.5) that

$$H^{d/2+k+\varepsilon}(\mathbb{B}_1^d) \overset{\text{c.}}{\hookrightarrow} \mathscr{B}^k(\mathbb{B}_1^d),$$

where $\varepsilon > 0$ and $H^s(\mathbb{B}_1^d)$ denotes the (fractional) sth-order L^2-Sobolev space. Therefore, we have the continuous embeddings

$$H^{d/2+k+\varepsilon}(\mathbb{B}_1^d) \overset{\text{c.}}{\hookrightarrow} \mathscr{B}^k(\mathbb{B}_1^d) \overset{\text{c.}}{\hookrightarrow} \mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d).$$

4.2.3 Discussion

The previously stated results say that very regular functions (those with order d derivatives in $L^2(\mathbb{B}_1^d)$) are contained in $\mathscr{R}\,\mathrm{BV}^k(\mathbb{B}_1^d)$. On the other hand, functions that are not very regular are also in $\mathscr{R}\,\mathrm{BV}^k(\mathbb{B}_1^d)$. For example, take any univariate function $g \in \mathrm{BV}^k(\mathbb{R})$ and use it as the profile of a ridge function

$$f(\boldsymbol{x}) = g(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{B}_1^d, \tag{4.9}$$

where $\boldsymbol{\alpha} \in \mathbb{S}^{d-1}$. The function g barely has k derivatives, yet $f \in \mathscr{R}\,\mathrm{BV}^k(\mathbb{B}_1^d)$ while $f \notin H^{d/2+k+\varepsilon}(\mathbb{B}_1^d)$. Although this function may not be very regular, it only varies in the single direction $\boldsymbol{\alpha} \in \mathbb{S}^{d-1}$. This shows that $\mathscr{R}\,\mathrm{BV}^k(\mathbb{B}_1^d)$ can be viewed as a *mixed variation* space in the sense of Donoho (2000) in that it includes highly regular

functions that are very isotropic, e.g., functions from the Sobolev space $H^{d/2+k+\varepsilon}(\mathbb{B}_1^d)$ or less regular functions that are highly anisotropic, e.g., the ridge function in (4.9).

4.3 Nonlinear Approximation with Ridge Splines

A well-known result in approximation theory, first due to Maurey and Pisier (1981), 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* (Pisier, 1981; Jones, 1992; Barron, 1993; DeVore and Temlyakov, 1996; Barron et al., 2008). This means that given a function f in the closure of the convex, symmetric hull of the dictionary, there exists an N-term superposition of atoms from the dictionary f_N such that $\|f - f_N\|_{\mathcal{H}} \lesssim N^{-1/2}$, 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.

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 (Bach, 2017; Siegel and Xu, 2021b). We use results from (Bach, 2017; Siegel and Xu, 2021b) to readily derive approximation rates for functions in $\mathscr{R}\,\mathrm{BV}^k(\Omega)$ that are immune to the curse of dimensionality. Again, for simplicity we suppose that $\Omega = \mathbb{B}_1^d$ as defined in (4.6).

Theorem 4.8. *Given $f \in \mathscr{R}\,\mathrm{BV}^k(\mathbb{B}_1^d)$, there exists a kth-order ridge spline with N neurons, denoted f_N, such that*

$$\|f - f_N\|_{L^2} \lesssim_d \mathscr{R}\,\mathrm{TV}^k_{\mathbb{B}_1^d}(f)\, N^{-\frac{1}{2}-\frac{2k-1}{2d}}.$$

Moreover, this rate cannot be improved.

Proof. Given $f \in \mathscr{R}\,\mathrm{BV}^k(\mathbb{B}_1^d)$, we have from Remark 4.3 the representation

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

It is known from the work of Siegel and Xu (2021b) that best L^2 approximation rate for the integral in the above display by a superposition of N neurons of the form $\boldsymbol{x} \mapsto \rho_k(\boldsymbol{\alpha}^\mathsf{T}\boldsymbol{x} - t)$, $\boldsymbol{\alpha} \in \mathbb{S}^{d-1}$ and $t \in [-1, 1]$, denoted $\widetilde{f}_N$, is

$$\left\| \int_{\mathbb{S}^{d-1}\times[-1,1]} \rho_k(\boldsymbol{\alpha}^\mathsf{T}(\cdot) - t)\,\mathrm{d}\mu(\boldsymbol{\alpha}, t) - \widetilde{f}_N \right\|_{L^2} \lesssim_d \|\mu\|_{\mathcal{M}}\, N^{-\frac{1}{2}-\frac{2k-1}{2d}}.$$

Next, since $\|\mu\|_{\mathcal{M}} = \mathscr{R}\,\mathrm{TV}^k_{\mathbb{B}_1^d}(f)$, the result follows by choosing $f_N \coloneqq \widetilde{f}_N + c$. □

Remark 4.9. As $d \to \infty$, Theorem 4.8 say that the approximation rate is $N^{-1/2}$ and is therefore immune to the curse of dimensionality.

In the special case of $k = 2$, using results regarding the approximation of zonoids by zonotopes from Matoušek (1996) (see also Bach (2017, Proposition 1)), we have that the result of Theorem 4.8 also holds with respect to the $L^\infty(\mathbb{B}_1^d)$-norm[3]. This is summarized in the following theorem.

Theorem 4.10. *Given $f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$, there exists a second-order ridge spline (i.e., shallow ReLU network with a skip connection) with N neurons, denoted f_N, such that*

$$\|f - f_N\|_{L^\infty} \lesssim_d \mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f)\, N^{-\frac{1}{2}-\frac{3}{2d}}.$$

Moreover, this rate cannot be improved.

Remark 4.11. The rates in Theorems 4.8 and 4.10 are *nonlinear approximation rates*. Using results from Siegel and Xu (2021b) bounding the Kolmogorov N-widths of the variation spaces $\mathscr{V}^k(\mathbb{B}_1^d)$, we have that the best linear approximation rates for $\mathscr{R}\,\mathrm{BV}^k(\mathbb{B}_1^d)$ scale like $N^{-\frac{2k-1}{2d}}$, which suffers the curse of dimensionality.

[3]Technically speaking, the approximation rates from Matoušek (1996) only hold for $d \geq 4$, but hold up to logarithmic factors for any $d \geq 1$.

4.4 Nonparametric Function Estimation with Shallow Neural Networks

In this section we consider the usual setup of nonparametric regression in the fixed design setting, and then extend the results to the random design setting. Consider the problem of estimating a function $f \in \mathscr{R}\,\mathrm{BV}^2(\Omega)$ from the noisy samples

$$y_m = f(\boldsymbol{x}_m) + \varepsilon_m,\ m = 1, \dots, M,$$

where $\{\varepsilon_m\}_{m=1}^M$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables and $\{\boldsymbol{x}_m\}_{m=1}^M \subset \Omega$ are fixed, but scattered, design points. There is a large body of work regarding the problem of statistical estimation with ridge functions, under many different names, including projection pursuit regression (Friedman and Stuetzle, 1981), ridgelet shrinkage (Candès, 2003), and, of course, estimation with neural networks (Barron, 1993, 1994). The last few years have led to a number of related works that consider the problem of minimax estimation with neural networks (Klusowski and Barron, 2017; Imaizumi and Fukumizu, 2019; Suzuki, 2019; Schmidt-Hieber, 2020; Hayakawa and Suzuki, 2020). 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 section 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.

We specifically focus on $\mathscr{R}\,\mathrm{BV}^2(\Omega)$ since our results rely on the L^∞ approximation rates from Theorem 4.10, which may or may not hold for the higher-order spaces. Once again, for simplicity we suppose that $\Omega = \mathbb{B}_1^d$ as defined in (4.6). Similar

results as those stated in the sequel can be derived for more general bounded domains $\Omega \subset \mathbb{R}^d$.

Theorem 4.12. *Consider the problem of estimating a function $f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ such that $\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \leq C$ from the noisy samples*

$$y_m = f(\boldsymbol{x}_m) + \varepsilon_m,\ m = 1, \dots, M,$$

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

$$f_{M,\mathrm{ReLU}} \in \operatorname*{arg\,min}_{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)} \sum_{m=1}^{M} |y_m - f(\boldsymbol{x}_m)|^2 \text{ s.t. } \mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \leq C \tag{4.10}$$

has a mean-squared error bound of

$$\mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} |f(\boldsymbol{x}_m) - f_{M,\mathrm{ReLU}}(\boldsymbol{x}_m)|^2\right] \lessapprox_d C^{\frac{2d}{2d+3}} \left(\frac{N}{\sigma^2}\right)^{-\frac{d+3}{2d+3}}, \tag{4.11}$$

where $\lessapprox$ hides universal constants and logarithmic factors, where the only random variables in the expectation above are the noise terms $\{\varepsilon_m\}_{m=1}^M$.

Remark 4.13. As $d \to \infty$, we have that $C^{\frac{2d}{2d+3}} \to C$ and so the bound (asymptotically) scales linearly with the constant C.

The proof of Theorem 4.12 follows standard techniques from nonparametric statistics (see, e.g., van de Geer (2000, Chapter 9) or Wainwright (2019, Chapter 13)). In particular, we use the following general result from Theorem 13.5 and the remarks following, the discussion on pg. 424, and Corollary 13.7 in Wainwright (2019, Chapter 13). We summarize this general result in the following proposition.

Proposition 4.14 (see Wainwright (2019, Chapter 13))**.** *Let $\mathcal{F}$ be a convex model class that contains the constant function, i.e., $f \equiv 1 \in \mathcal{F}$. Given $f \in \mathcal{F}$, consider the*

problem of estimating f from the noisy samples

$$y_m = f(\boldsymbol{x}_m) + \varepsilon_m,\ m = 1, \dots, M,$$

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

$$f_M \in \arg\min_{f \in \mathcal{F}} \sum_{m=1}^M |y_m - f(\boldsymbol{x}_m)|^2$$

has a mean-squared error bound of

$$\mathbb{E}\|f - f_M\|_M^2 \lesssim \delta_M^2,$$

where $\|\cdot\|_M$ is the empirical L^2-norm defined by

$$\|f\|_M^2 \coloneqq \frac{1}{M} \sum_{m=1}^M |f(\boldsymbol{x}_m)|^2.$$

and $\delta_M = \delta$ satisfies the inequality

$$\frac{16}{\sqrt{M}} \int_{\frac{\delta^2}{2\sigma^2}}^{\delta} \sqrt{\log \mathcal{N}(t, \partial\mathcal{F}, \|\cdot\|_M)}\, \mathrm{d}t \leq \frac{\delta^2}{4\sigma}, \tag{4.12}$$

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

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

We now use Proposition 4.14 to prove Theorem 4.12.

Proof of Theorem 4.12. In Theorem 4.12, our model class is

$$\mathcal{F}_C \coloneqq \Big\{ f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d) :\ \mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \leq C \Big\}. \tag{4.13}$$

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

Notice that

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

so it suffices to upper bound the metric entropy of $\mathcal{F}_{2C}$ to find a δ_M that satisfies (4.12). By noticing that $\|\cdot\|_M \le \|\cdot\|_{L^\infty(\mathbb{B}_1^d)}$, we can use the approximation rate from Theorem 4.10 to upper bound (up to logarithmic factors) the metric entropy

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

Next,

$$\begin{aligned}
&\frac{1}{\sqrt{M}} \int_{\frac{\delta^2}{2\sigma^2}}^{\delta} \sqrt{\log \mathcal{N}(t, \partial\mathcal{F}, \|\cdot\|_M)}\,\mathrm{d}t \\
&\le \frac{1}{\sqrt{M}} \int_{0}^{\delta} \sqrt{\log \mathcal{N}(t, \partial\mathcal{F}, \|\cdot\|_M)}\,\mathrm{d}t \\
&\lessapprox_d \frac{1}{\sqrt{M}} \int_0^\delta \left(\frac{C}{t}\right)^{\frac{d}{d+3}} \mathrm{d}t \\
&= \frac{C^{\frac{d}{d+3}}}{\sqrt{M}} \left[t^{\frac{3}{d+3}} \right|_0^\delta \\
&= C^{\frac{d}{d+3}} \frac{\delta^{\frac{3}{d+3}}}{\sqrt{M}}.
\end{aligned}$$

From (4.12), we want to find $\delta_M = \delta$ that satisfies

$$C^{\frac{d}{d+3}} \frac{\delta^{\frac{3}{d+3}}}{\sqrt{M}} \lessapprox_d \frac{\delta^2}{\sigma}. \tag{4.14}$$

We have (up to logarithmic factors) that

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

satisfies (4.14). □

Remark 4.15. By Corollary 4.6 and the discussion thereafter, one can compute the estimator $f_{M,\mathrm{ReLU}}$ that satisfies the bound in (4.11) by training a sufficiently wide shallow ReLU network (to a global minimizer) with weight decay or with path-norm regularization where, by Lagrange calculus, the choice of the regularization parameter λ depends on C and the data through the data-fitting term.

Remark 4.16. Since when $d = 1$, $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ is exactly the space $\mathrm{BV}^2[-1,1]$, the result of Theorem 4.12 recovers the well-known mean-squared error rate of $N^{-4/5}$ of locally adaptive linear spline estimators (Mammen and van de Geer, 1997).

The result of Theorem 4.12 can be extended from the fixed design setting to the random design setting using standard techniques (see, e.g., Wainwright (2019, Chapter 14)). In particular, assuming the design points $\{\boldsymbol{x}_m\}_{m=1}^M \subset \mathbb{B}_1^d$ are i.i.d. uniform random variables on $\mathbb{B}_1^d$, we can use the techniques outlined in Wainwright (2019, Chapter 14) to derive the same mean-squared error rate (for sufficiently large N) with respect to $\|\cdot\|_{L^2(\mathbb{B}_1^d;\mathrm{P}_X)}$ instead of $\|\cdot\|_M$, where P_X denotes the uniform probability measure on $\mathbb{B}_1^d$. This results in the following corollary to Theorem 4.12.

Corollary 4.17. *Consider the problem of estimating a function $f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ such that $\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \leq C$ from the noisy samples*

$$y_m = f(\boldsymbol{x}_m) + \varepsilon_m,\ m = 1, \dots, M,$$

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

variational problem

$$f_{M,\mathrm{ReLU}} \in \operatorname*{arg\,min}_{f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)} \sum_{m=1}^{M} |y_m - f(\boldsymbol{x}_m)|^2 \text{ s.t. } \mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \leq C$$

has a mean-squared error bound of

$$\mathbb{E}\|f - f_{M,\mathrm{ReLU}}\|^2_{L^2(\mathbb{B}_1^d;\mathrm{P}_X)} \lessapprox_d C^{\frac{2d}{2d+3}} \left(\frac{M}{\sigma^2}\right)^{-\frac{d+3}{2d+3}},$$

where the only random variables in the expectation above are the noise terms $\{\varepsilon_m\}_{m=1}^M$.

The following theorem shows that this mean-squared error rate cannot be improved. In other words, the rate in Theorem 4.12 is (up to logarithmic factors) minimax optimal.

Theorem 4.18. *Consider the problem of estimating a function* $f \in \mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ *such that* $\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}_1^d}(f) \leq C$ *from the noisy samples*

$$y_m = f(\boldsymbol{x}_m) + \varepsilon_m,\ m = 1, \ldots, M,$$

where $\{\varepsilon_m\}_{m=1}^M$ *are i.i.d.* $\mathcal{N}(0, \sigma^2)$ *random variables and* $\{\boldsymbol{x}_m\}_{m=1}^M \subset \mathbb{B}_1^d$ *are i.i.d. uniform random variables on* $\mathbb{B}_1^d$. *Then, we have the following minimax lower bound*

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

where the inf *is over all functions of the data and the only random variables in the expectation are the noise terms* $\{\varepsilon_n\}_{n=1}^N$.

The proof of Theorem 4.18 follows from the general result of Yang and Barron (see Yang and Barron (1999, Proposition 1) and Wainwright (2019, Chapter 15)) regarding minimax rates over model classes. We summarize this result in the following proposition.

Proposition 4.19 (see Yang and Barron (1999, Proposition 1) and Wainwright (2019, Chapter 15))**.** *Let $\mathcal{F}$ be a model class. Given $f \in \mathcal{F}$, consider the problem of estimating f from the noisy samples*

$$y_m = f(\boldsymbol{x}_m) + \varepsilon_m,\ m = 1, \dots, M,$$

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

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

we have the minimax rate

$$\inf_{f_M} \sup_{f \in \mathcal{F}} \mathbb{E}\|f - f_M\|^2_{L^2(\mathbb{B}_1^d; \mathrm{P}_X)} \asymp t_M^2,$$

where the only random variables in the expectation are the noise terms $\{\varepsilon_m\}_{m=1}^M$, and $t_M^2 = t^2$ satisfies

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

We first use the result of Theorem 4.19 to derive the minimax rate for the model class

$$\mathscr{G}_C \coloneqq \left\{ f \in \mathscr{V}^2(\mathbb{B}_1^d) : \ \|f\|_{\mathscr{V}^2(\mathbb{B}_1^d)} \le C \right\}, \tag{4.15}$$

where $\mathscr{V}^2(\mathbb{B}_1^d)$ is the second-order variation space defined in Section 4.2.1. We then use this minimax rate to prove Theorem 4.18.

Lemma 4.20. *Consider the problem of estimating $f \in \mathscr{G}_C$ (defined in (4.15)) from the noisy samples*

$$y_m = f(\boldsymbol{x}_m) + \varepsilon_m,\ m = 1, \dots, M,$$

where $\{\varepsilon_m\}_{m=1}^M$ are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables and $\{\boldsymbol{x}_m\}_{m=1}^M$ are i.i.d. uniform

random variables on $\mathbb{B}_1^d$. *The minimax rate for this model class is*

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

where the $L^2(\mathbb{B}_1^d;\mathrm{P}_X)$*-norm is the* L^2*-norm with respect to the uniform probability measure measure on* $\mathbb{B}_1^d$ *and the only random variables in the expectation are the noise terms* $\{\varepsilon_m\}_{m=1}^M$.

Proof. We are interested in applying Theorem 4.19 with 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 (4.15) was extensively studied by Siegel and Xu (2021b) 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 Siegel and Xu (2021b, Theorem 4 and Equation (68)) for the upper bound and Siegel and Xu (2021b, 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 Theorem 4.19. □

We now use Theorem 4.20 to derive a minimax lower bound for the model class in (4.13).

Proof of Theorem 4.18. It suffices to show that $\mathscr{G}_C \subset \mathcal{F}_C$, where $\mathcal{F}_C$ is defined in (4.13). Given $f \in \mathscr{V}^2(\mathbb{B}_1^d)$ (or in $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$, since they equivalent spaces by Theorem 4.7), we can find an integral representation as in (4.8) such that

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

Next, if we let $\nu \coloneqq \mu \llcorner (\mathbb{S}^{d-1} \times [-1,1])$, we can write f as an integral representation as in Remark 4.3 such that

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

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

4.4.1 Breaking the Curse of Dimensionality

When $d = 1$, Theorems 4.12 and 4.18 recovers (up to logarithmic factors) the well-known minimax rate of $M^{-4/5}$ for $\mathrm{BV}^2[-1,1]$ model classes (Donoho and Johnstone, 1998). On the other hand, when $d \to \infty$, the rate approaches $M^{-1/2}$, and is therefore immune to the curse of dimensionality. To understand why this is happening, we recall from Section 4.2.3 that $\mathscr{R}\,\mathrm{BV}^2(\mathbb{B}_1^d)$ can be viewed as a mixed variation space. In particular, it contains highly isotropically regular functions that belong to the Sobolev space $H^{d/2+2+\varepsilon}(\mathbb{B}_1^d)$, $\varepsilon > 0$, as well as anistropic less regular functions such as the ridge function defined in (4.9), which barely has two 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.

4.4.2 Neural Networks vs. Linear Methods

In this section we illustrate the idea that the neural network estimator studied in Section 4.4 is *locally adaptive* (a term coined by Donoho and Johnstone (1998)) unlike more classical *linear methods* (which include kernel methods (Schölkopf and Smola, 2002)). We illustrate this both quantitatively via linear rates for function estimation as well as qualitatively via numerical experiments. For the problem of function estimation, recall that a linear method is a method in which the estimator is a *linear* function of the data $(y_1, \dots, y_M)$, i.e., the estimator is computed via a linear map $T : \mathbb{R}^M \to \mathcal{F}$, where $\mathcal{F}$ is some model class and T can depend on the design points

$\{\boldsymbol{x}_m\}_{m=1}^M$ in an arbitrary way. Due to the sparsity-promoting nature of the $\mathcal{M}$-norm used to define $\mathscr{R}\,\mathrm{TV}^2_{\mathbb{B}^d_1}(\cdot)$, the estimator in Theorem 4.12 is a *nonlinear* function of the data. This is analagous to LASSO-type estimators arising from ℓ^1-norm regularized problems, which are nonlinear estimators for discrete-domain problems.

The Univariate Case

We first recall what happens in the univariate case, discussed in Chapter 1. In the univariate case, we have that the variational problem in (4.10) reduces to the (regularized) variational problem

$$\min_{f \in \mathrm{BV}^2[-1,1]} \sum_{m=1}^M |y_m - f(x_m)|^2 + \lambda \, \|\mathrm{D}^2 f\|_{\mathcal{M}}, \tag{4.16}$$

where $\lambda > 0$ is the regularization parameter. The solutions are locally adaptive linear spline estimators (Mammen and van de Geer, 1997). The minimax rate for $\mathrm{BV}^2[-1,1]$ model classes is $M^{-4/5}$, which is achieved (up to logarithmic factors) by the locally adaptive linear spline estimator. Moreover, when restricted to *linear estimators*, the linear minimax rate is known to be $N^{-3/4}$ (Donoho and Johnstone, 1998), which is achieved (up to logarithmic factors) by the cubic smoothing spline estimator (de Boor and Lynch, 1966). The cubic smoothing spline is a solution to the variational problem

$$\min_{f \in H^2[-1,1]} \sum_{n=1}^N |y_n - f(x_n)|^2 + \lambda \, \|\mathrm{D}^2 f\|_{L^2}^2, \tag{4.17}$$

where we recall that

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

is the second-order L^2-Sobolev space and $\mathcal{D}'[-1,1]$ denotes the space of distributions 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 (4.16) and the problem in (4.17) is the difference between the *sparsity-promoting* $\mathcal{M}$-norm regularization in (4.16) and the L^2-norm

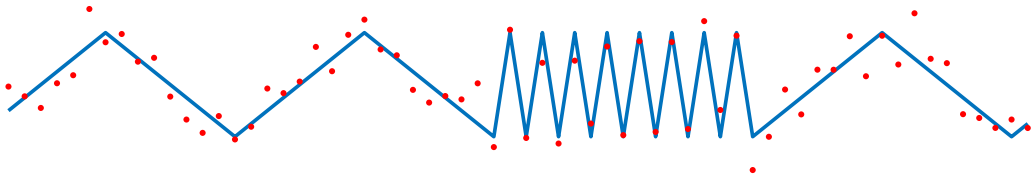
regularization in (4.17). This is analogous to the difference between ℓ^1-norm and ℓ^2-norm regularization in discrete-domain problems.

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 $M^{-4/5}$ and $M^{-3/4}$. 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 Figure 4.1, 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.

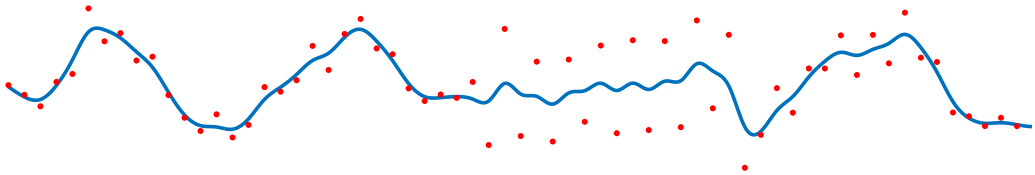
In Figure 4.1(a) we plot a function (in blue) and generate a data set by taking noisy samples (in red) of the function plus i.i.d. Gaussian noise. Clearly this function is in $\mathrm{BV}^2[-1,1]$ but not in $H^2[-1,1]$ since taking two (distributional) derivatives of this function is an impulse train. This function is spatially inhomogeneous since it is highly oscillatory in some regions and less oscillatory in others.

In Figure 4.1(b) and Figure 4.1(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 inhomogenity 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 Figure 4.1(b) that the cubic smoothing spline oversmooths the high variation portion of the data and we see for small λ in Figure 4.1(c) that the cubic smoothing spline undersmooths (overfits) the low variation portion of the data.

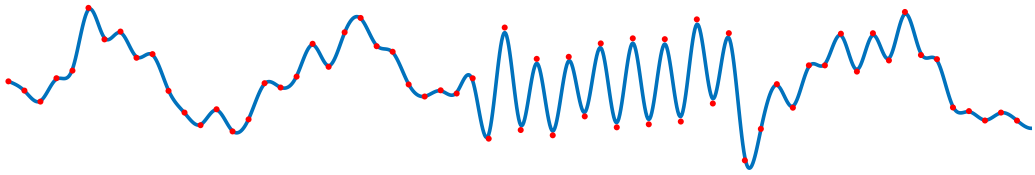
In Figure 4.1(d) we plot a solution to the variational problem in (4.16), which is a locally adaptive linear spline which can be computed by training a shallow ReLU



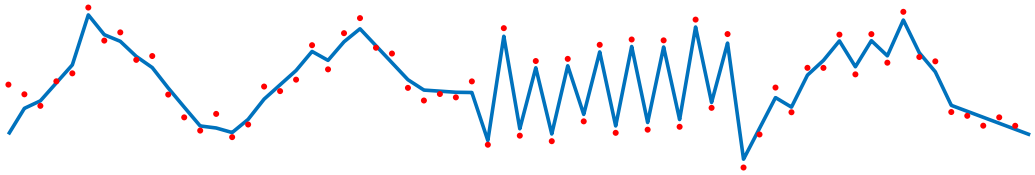
(a) True function and data.



(b) Cubic smoothing spline with large λ.



(c) Cubic smoothing spline with small λ.



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

Figure 4.1: 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).

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 recall 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 as we saw in Chapter 1, Figure 1.4. 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.

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 4.21. *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*

$$y_m = f(\boldsymbol{x}_m) + \varepsilon_m,\ m = 1, \dots, M,$$

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

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

where the inf *is over all* linear *functions of the data and the only random variables in the expectation are the noise terms $\{\varepsilon_n\}_{n=1}^N$*

To prove Theorem 4.21, we require several results from ridgelet analysis. It was shown in Candès (1998, Theorem 7) that we have the continuous embedding

$$R^{(d+3)/2}_{1,1}(\mathbb{B}_1^d) \overset{\text{c.}}{\hookrightarrow} \mathscr{V}^2(\mathbb{B}_1^d),$$

where we recall that $\mathscr{V}^2(\mathbb{B}_1^d)$ is the variation space for shallow ReLU networks, and $R^s_{p,q}(\mathbb{B}_1^d)$ denotes the *ridgelet space* of Candès (1998). Ridgelet spaces were proposed

as a generalization of Besov spaces, and in the univariate case, the ridgelet space $R^s_{p,q}(\mathbb{B}^d_1)$ coincides with the Besov space $B^s_{p,q}[-1,1]$.

Next, recall that we showed in the proof of Theorem 4.18 that $\mathscr{G}_C \subset \mathcal{F}_C$, where $\mathscr{G}_C$ and $\mathcal{F}_C$ are the model classes defined in (4.13) and (4.15), respectively. Combining this fact with the above display, we see that to prove Theorem 4.21, it suffices to show the linear minimax lower bound for the model class

$$\mathscr{H}_C \coloneqq \Big\{ f \in R^{(d+3)/2}_{1,1}(\mathbb{B}^d_1) : \|f\|_{R^{(d+3)/2}_{1,1}(\mathbb{B}^d_1)} \le C \Big\}.$$

We make use of the following generic result.

Proposition 4.22 (see Candès (2003, Proof of Theorem 4.1))**.** *Let $\mathcal{F} \subset L^2(\mathbb{B}^d_1)$ be a convex model class and consider the problem of estimating $f \in \mathcal{F}$ from the continuous white noise model*

$$\mathrm{d}Y_\varepsilon(\boldsymbol{x}) = f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} + \varepsilon\,\mathrm{d}W(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{B}^d_1,$$

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 $\lesssim_d K_\delta$ orthogonal elements $\{g_k\}_{k=1}^K \subset \mathcal{F}$ such that $\|g_k\|_{L^2(\mathbb{B}^d_1)} = \delta$, $k = 1, \dots, K$. Then, the linear minimax rate is lower-bounded by

$$\inf_{f_\varepsilon \text{ linear}} \sup_{f \in \mathcal{F}} \mathbb{E}\|f - f_\varepsilon\|^2_{L^2(\mathbb{B}^d_1)} \gtrsim_d \delta^2_\varepsilon,$$

where $\delta_\varepsilon = \delta$ solves

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

Proposition 4.23 (see Candès (1998, Theorem 11) and Candès (2003, Lemmas A.1, A.2, and A.3))**.** *For any integer $j \ge 2$, There exists a set $\{g_k\}_{k=1}^K$ of orthogonal elements with $K \gtrsim_d 2^{jd}$ contained in*

$$\Big\{ f \in R^s_{1,1}(\mathbb{B}^d_1) : \|f\|_{R^s_{1,1}(\mathbb{B}^d_1)} \le C \Big\},$$

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 now prove Theorem 4.21.

Proof of Theorem 4.21. 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[4], 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 $\varepsilon = \sigma/\sqrt{M}$, for sufficiently large M, so we get the linear minimax lower bound of

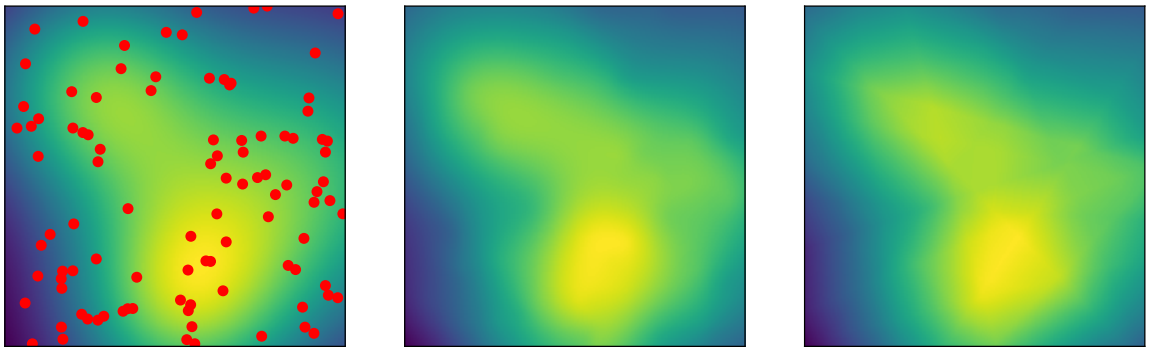
$$\left(\frac{M}{\sigma^2}\right)^{-\frac{3}{d+3}}.$$

□

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 $M^{-\frac{d+3}{2d+3}}$ and

[4]See Brown and Low (1996) where this argument was first rigorously formalized in the univariate case, and see Reiß (2008) where this idea was rigorously formalized in the multivariate case, which applies to our setting.

$M^{-\frac{3}{d+3}}$, respectively. As $d \to \infty$, the nonlinear estimation rate tends to $M^{-1/2}$, which is immune to the curse of dimensionality, while the linear estimation error rate suffers the curse of dimensionality. Moreover, these rates recover the univariate ($d = 1$) rates of $M^{-4/5}$ and $M^{-3/4}$. 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.



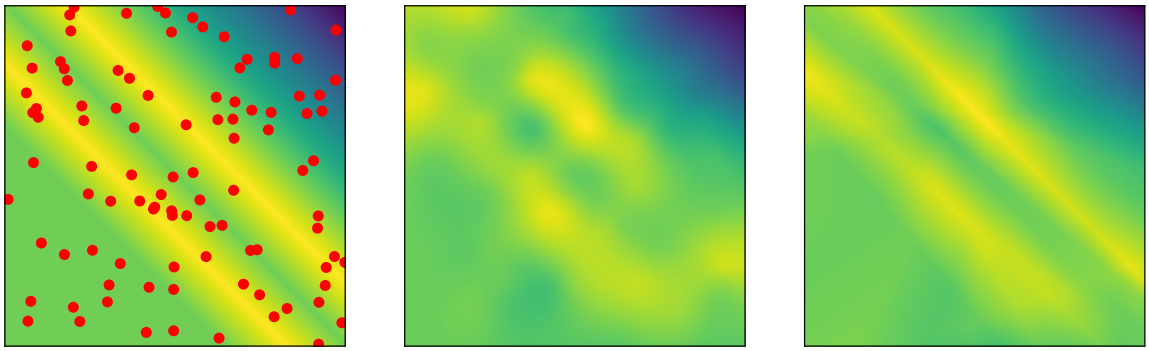
(a) True function and data. (b) Thin-plate spline. (c) Shallow ReLU network.

Figure 4.2: 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.

We illustrate this phenomenon by considering the problem of estimating a two-dimensional function and compare solutions to the variational problem in (4.10) with the thin-plate spline estimator (Wahba, 1990), which is a linear method and a special case of a kernel method. The thin-plate spline is a solution to the variational problem

$$\min_{f \in H^2(\mathbb{B}_1^2)} \sum_{m=1}^{M} |y_m - f(\boldsymbol{x}_m)|^2 + \lambda \Big(\big\| \partial_{x_1}^2 f \big\|_{L^2}^2 + 2 \| \partial_{x_2} \partial_{x_1} f \|_{L^2}^2 + \big\| \partial_{x_2}^2 f \big\|_{L^2}^2 \Big),$$

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



(a) True function and data.

(b) Thin-plate spline.

(c) Shallow ReLU network.

Figure 4.3: 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.

(4.17). 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 Figure 4.2 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 Figure 4.2(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 Figure 4.2(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 Figure 4.2(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 Figure 4.3 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 Figure 4.3(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 Figure 4.3(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 Figure 4.3(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, while linear/kernel methods cannot. We believe that the results of this section provide compelling evidence that trying to understand neural networks via linearization schemes such as the neural tangent kernel (Jacot et al., 2018) 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.

Chapter 5

Concluding Remarks

Inspired by tools from spline theory we proposed and studied a new family of Banach spaces, referred to as Radon-domain bounded variation spaces. These spaces are intrinsically related to neural networks with various activation functions including the popular ReLU. Additionally, the results of this dissertation also provide compelling evidence that neural networks can (and should!) be viewed as a type of spline.

In particular, we derived representer theorems over these spaces, showing that the solution sets to data-fitting variational problems over these spaces are completely characterized by functions that are realizable by neural networks. Moreover, these variational problems can be recast as finite-dimensional neural network training problems with regularization schemes related to weight decay and path-norm regularization, providing new theoretical explanation for these regularization schemes as well as providing several new, principled regularization schemes for deep neural networks.

We also showed that on a bounded domain, these spaces are equivalent (as Banach spaces) to the variation spaces of neural networks. This allowed us to study the approximation properties of these Banach spaces, showing that these spaces are immune to the curse of dimensionality. Using these approximation properties, we were able to show that neural network estimators are near-minimax optimal estimators for functions from these spaces.

5.1 How Theory Informs Practice

The variational framework developed in this dissertation informs the practical use of neural networks.

- By showing that the solutions to neural network training problems are solutions to variational problems over the Radon-domain BV space $\mathscr{R}\,\mathrm{BV}^k$, we now have a concrete framework for comparing neural networks to more classical data-fitting techniques such as kernel methods (which are optimal solutions to variational problems over an RKHS) by comparing $\mathscr{R}\,\mathrm{BV}^k$ to the RKHS of the particular kernel method.

- Many theoretical results regarding neural networks hold for infinite-width neural networks (Jacot et al., 2018; Wei et al., 2019). The representer theorems in Chapter 3 show that it suffices to only consider (deep) neural networks of finite-width so long as the width is sufficiently wide.

- Skip connections are a common architectural choice in neural networks (He et al., 2016). Many of the reasons for considering skip connections are based on heuristics. Skip connections are a natural by-product of the variational framework developed in this dissertation, providing a principled reason for considering skip connections in neural network architectures.

- It has become folklore in the machine learning community that deep neural networks are simply linear/kernel methods (Monroe, 2022). The results of this dissertation show that neural networks learn functions in the (non-Hilbertian) $\mathscr{R}\,\mathrm{BV}^k$ Banach spaces.

5.2 Open Problems

There are a number of open problems that remain regarding these new function spaces. In the remainder of this chapter we outline several directions for future work.

5.2.1 Approximate Atomic Decomposition of $\mathscr{R}\,\mathrm{BV}^k(\Omega)$

In the univariate case, the Radon-domain BV space $\mathscr{R}\,\mathrm{BV}^k$, reduces to the classical BV^k space. On a bounded domain, it is well-known (see, e.g., Peetre, 1976) that we have the continuous embeddings

$$B^k_{1,1}[0,1] \overset{\text{c.}}{\hookrightarrow} \mathrm{BV}^k[0,1] \overset{\text{c.}}{\hookrightarrow} B^k_{1,\infty}[0,1]. \tag{5.1}$$

Since Besov spaces admit atomic decompositions via wavelets (Triebel, 2008), this result implies that the $\mathrm{BV}^k[0,1]$ spaces approximately admit atomic decompositions since they are tightly sandwiched between two very similar Besov spaces which do admit atomic decompositions, even though the $\mathrm{BV}^k[0,1]$ spaces do not have unconditional bases. It remains an open question whether or not when $d \geq 1$ the $\mathscr{R}\,\mathrm{BV}^k(\Omega)$ spaces admit an approximate atomic decomposition, where $\Omega \subset \mathbb{R}^d$ is a bounded domain.

Using classical function spaces, we cannot tightly sandwich $\mathscr{R}\,\mathrm{BV}^k(\Omega)$ between two similar spaces. Indeed, using L^2-Sobolev spaces, we have from Section 4.2.2 the following sandwiching of $\mathscr{R}\,\mathrm{BV}^k(\Omega)$

$$H^{d/2+k+\varepsilon}(\Omega) \overset{\text{c.}}{\hookrightarrow} \mathscr{R}\,\mathrm{BV}^k(\Omega) \overset{\text{c.}}{\hookrightarrow} H^{k-1}(\Omega),$$

where $\varepsilon > 0$. The gap between the two Sobolev spaces implies that classical function spaces are, perhaps, too coarse to tightly sandwich the new, not classical $\mathscr{R}\,\mathrm{BV}^k(\Omega)$ spaces. To this end, in the remainder of this chapter, we propose a new scale of Banach spaces, which we refer to as Radon-domain Besov spaces, and conjecture how the $\mathscr{R}\,\mathrm{BV}^k(\Omega)$ spaces are related to these new spaces as well as outline some technical difficulties in actually trying to prove the conjecture.

Radon-Domain Besov Spaces

From the L^2-isometries of the Radon transform (see Section 2.5), consider the following 4 parameter ($r, s \geq 0$, $1 \leq p < \infty$, $1 \leq q \leq \infty$) family of function spaces

$$\mathscr{R}B^{r,s}_{p,q}(\mathbb{R}^d) \coloneqq \mathscr{R}^* \, \mathrm{K}^{\frac{d-1}{2}} \big(B^r_{p,q}(\mathbb{S}^{d-1}) \otimes_\alpha B^s_{p,q}(\mathbb{R}) \big), \tag{5.2}$$

where $B^r_{p,q}(\mathbb{S}^{d-1})$ and $B^s_{p,q}(\mathbb{R})$ are the usual Besov spaces on $\mathbb{S}^{d-1}$ and $\mathbb{R}$, respectively, and $\otimes_\alpha$ denotes the completion of the algebraic tensor product with respect to an appropriate tensor norm α which we define explicitly in (5.3). We refer to $\mathscr{R}B^{r,s}_{p,q}(\mathbb{R}^d)$ as a *Radon-domain Besov space* since it is the tensor product of Besov spaces in the (half-filtered) Radon domain. Note that since $\mathscr{R}^* \, \mathrm{K}^{\frac{d-1}{2}}$ is an L^2-isometry combined with the intertwining of Laplacians and the Radon transform, we have that

$$\mathscr{R}B^{0,s}_{2,2}(\mathbb{R}^d) = H^s(\mathbb{R}^d),$$

where $H^s(\mathbb{R}^d)$ is the usual sth-order L^2-Sobolev space on $\mathbb{R}^d$.

This definition captures a kind of anisotropy between the direction and offset variables of the Radon domain. In order to define the tensor norm α, we use the fact that Besov spaces admit atomic decompositions via sequence space representations.

From Narcowich et al. (2006b,a), there exist localized (i.e., wavelet-like) frames on the sphere $\mathbb{S}^{d-1}$. These frames are referred to as *needlets* due to their almost exponential localization and that they look like needles on the sphere. Let $\{\varphi_{\boldsymbol{\eta}}\}_{\boldsymbol{\eta} \in \mathcal{X}}$, denote the needlet system. The index set $\mathcal{X} \subset \mathbb{S}^{d-1}$ is a countable collection of the centers of each needlet function $\varphi_{\boldsymbol{\eta}}$. We can decompose the index set as $\mathcal{X} = \bigcup_{j=0}^{\infty} \mathcal{X}_j$, where the $\mathcal{X}_j$ indexes all needlets at scale j. This system forms a Parseval frame for $L^2(\mathbb{S}^{d-1})$ (i.e., a frame with frame bounds equal to 1) (Narcowich et al., 2006b, Theorem 5.2). Also, let $\{\psi_{j,k}\}_{j \in \mathbb{N}_0, k \in \mathbb{Z}}$ denote the inhomogeneous Meyer wavelet system (Lemarié and Meyer, 1986), where

$$\psi_{0,k}(x) \coloneqq \phi(x - k), \quad k \in \mathbb{Z},$$

where ϕ is the Meyer scaling function and

$$\psi_{j+1,k}(x) := 2^{j/2}\psi(2^j x - k), \quad j \in \mathbb{N}_0, k \in \mathbb{Z},$$

where ψ is the Meyer wavelet function. The system $\{\psi_{j,k}\}_{j\in\mathbb{N}_0,k\in\mathbb{Z}}$ forms an orthobasis for $L^2(\mathbb{R})$.

Next, it is well-known that (see, e.g., Narcowich et al., 2006a, Theorem 5.5) that given $f \in \mathcal{S}'(\mathbb{S}^{d-1})$[1], $f \in B^r_{p,q}(\mathbb{S}^{d-1})$ if and only if

$$\|f\|_{B^r_{p,q}} = \left(\sum_{m=0}^{\infty}\left(2^{m(r+(d-1)/2-(d-1)/p)}\left(\sum_{\boldsymbol{\eta}\in\mathcal{X}_m}|\langle g,\varphi_{\boldsymbol{\eta}}\rangle|^p\right)^{1/p}\right)^q\right)^{1/q} < \infty,$$

where $\{\varphi_{\boldsymbol{\eta}}\}_{\boldsymbol{\eta}\in\mathcal{X}_m}$ are the needlets at scale m, with appropriate modification when $q = \infty$. It is also well-known that (see, e.g., Meyer, 1992, Chapter 6) given $f \in \mathcal{S}'(\mathbb{R})$, $f \in B^s_{p,q}(\mathbb{R})$ if and only if

$$\|f\|_{B^s_{p,q}} = \left(\sum_{j=0}^{\infty}\left(2^{j(s+1/2-1/p)}\left(\sum_{k\in\mathbb{Z}}|\langle f,\psi_{j,k}\rangle|^p\right)^{1/p}\right)^q\right)^{1/q} < \infty,$$

with appropriate modification when $q = \infty$. Note that the pairings $\langle\cdot,\cdot\rangle$ that appear in the above two displays are well-defined since $\varphi_{\boldsymbol{\eta}} \in \mathcal{S}(\mathbb{S}^{d-1})$ and $\psi_{j,k} \in \mathcal{S}(\mathbb{R})$. From these two atomic decompositions, we define norm on $B^r_{p,q}(\mathbb{S}^{d-1}) \otimes_\alpha B^s_{p,q}(\mathbb{R})$ as

$$\begin{aligned}&\|g\|_{B^r_{p,q}(\mathbb{S}^{d-1})\otimes B^s_{p,q}(\mathbb{R})}\\ &= \left(\sum_{m=0}^{\infty}\sum_{j=0}^{\infty}\left(2^{m(r+(d-1)/2-(d-1)/p)+j(s+1/2-1/p)}\left(\sum_{\boldsymbol{\eta}\in\mathcal{X}_m}\sum_{k\in\mathbb{Z}}\left|\left[g,\varphi_{\boldsymbol{\eta}}\otimes\psi_{j,k}\right]\right|^p\right)^{1/p}\right)^q\right)^{1/q}.\end{aligned} \tag{5.3}$$

This is the tensor norm used to complete the tensor product in (5.2). One can check

[1] $\mathcal{S}'(\mathbb{S}^{d-1})$ is the space of distributions on the sphere, which is the continuous dual of the space of test functions on the sphere $\mathcal{S}(\mathbb{S}^{d-1}) := C^\infty(\mathbb{S}^{d-1})$.

that the system $\{\rho_{j,k,\boldsymbol{\eta}}\}_{j\in\mathbb{N}_0,k\in\mathbb{Z},\boldsymbol{\eta}\in\mathcal{X}}$ defined by

$$\rho_{j,k,\boldsymbol{\eta}}(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}} \widetilde{\psi}_{j,k}(\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{x})\varphi_{\boldsymbol{\eta}}(\boldsymbol{\alpha})\,\mathrm{d}\sigma(\boldsymbol{\alpha}),$$

is a Parseval frame for $L^2(\mathbb{R}^d)$, where $\widetilde{\psi}_{j,k} = \mathrm{K}^{\frac{d-1}{2}}\,\psi_{j,k}$, i.e.,

$$\widehat{\mathrm{K}^{\frac{d-1}{2}}\,\psi_{j,k}}(\omega) = \widehat{K}^{\frac{d-1}{2}}(\omega)\widehat{\psi}_{j,k}(\omega) = \sqrt{c_d}|\omega|^{\frac{d-1}{2}}\widehat{\psi}_{j,k}(\omega).$$

This system has the property that given $f \in \mathcal{S}'(\mathbb{R}^d)$,

$$\langle f, \rho_{j,k,\boldsymbol{\eta}}\rangle = \Big[\mathrm{K}^{\frac{d-1}{2}}\,\mathscr{R}f, \varphi_{\boldsymbol{\eta}} \otimes \psi_{j,k}\Big],$$

and so we can equip the space $\mathscr{R}B^{r,s}_{p,q}(\mathbb{R}^d)$ with the norm

$$\|f\|_{\mathscr{R}B^{r,s}_{p,q}} \coloneqq \left(\sum_{m=0}^{\infty}\sum_{j=0}^{\infty}\left(2^{m(s_1+(d-1)/2-(d-1)/p)+j(s_2+1/2-1/p)}\left(\sum_{\boldsymbol{\eta}\in\mathcal{X}_m}\sum_{k\in\mathbb{Z}}|\langle f,\rho_{j,k,\boldsymbol{\eta}}\rangle|^p\right)^{1/p}\right)^q\right)^{1/q}$$

making it a Banach space, with appropriate modification when $q = \infty$.

Remark 5.1. The frame elements $\rho_{j,k,\boldsymbol{\eta}}$ can be viewed as a weighted average of the (half-filtered) wavelet ridge functions

$$\widetilde{\psi}_{j,k}(\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{x}),$$

averaged over all directions $\boldsymbol{\alpha} \in \mathbb{S}^{d-1}$, where the weighting function is the needlet $\varphi_{\boldsymbol{\eta}}$, centered at the direction $\boldsymbol{\eta} \in \mathcal{X} \subset \mathbb{S}^{d-1}$. Therefore, these frame elements are localized functions, parameterized by a location j, scale k, and direction $\boldsymbol{\eta}$. We consider this a localized ridgelet-type tight frame.

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with a sufficiently nice boundary. We can define the space $\mathscr{R}B^{r,s}_{p,q}(\Omega)$ as

$$\mathscr{R}B^{r,s}_{p,q}(\Omega) \coloneqq \big\{f : \mathcal{D}'(\Omega) :\ \exists g \in \mathscr{R}B^{r,s}_{p,q}(\mathbb{R}^d)\ \text{s.t.}\ g|_{\Omega} = f\big\},$$

This is a Banach space when equipped with the norm

$$\|f\|_{\mathscr{R}B^{r,s}_{p,q}(\Omega)} \coloneqq \inf_{g\in\mathscr{R}B^{r,s}_{p,q}(\mathbb{R}^d)} \|g\|_{\mathscr{R}B^{r,s}_{p,q}(\mathbb{R}^d)} \quad \text{s.t.} \quad g|_\Omega = f.$$

Moreover, these spaces admit atomic decompostions since the Besov spaces used to define $\mathscr{R}B^{r,s}_{p,q}(\mathbb{R}^d)$ admit atomic decompositions. From the univariate embeddings in (5.1), we conjecture the following embeddings for the $\mathscr{R}\,\mathrm{BV}^k(\Omega)$ spaces.

Conjecture 5.2. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. We conjecture that we have the following continuous embeddings*

$$\mathscr{R}B^{0,k+\frac{d-1}{2}}_{1,1}(\Omega) \overset{c.}{\hookrightarrow} \mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d) \overset{c.}{\hookrightarrow} \mathscr{R}B^{0,k+\frac{d-1}{2}}_{1,\infty}(\Omega)$$

when Ω has a sufficiently nice boundary.

The main technical difficulties in proving this conjecture arise in the construction of a nice extension operator from $\mathscr{R}B^{r,s}_{p,q}(\Omega) \to \mathscr{R}B^{r,s}_{p,q}(\mathbb{R}^d)$. This problem is very similar to finding an intrinsic definition[2] of $\mathscr{R}\,\mathrm{BV}^k$ on a bounded domain. In the case of Besov spaces, such extension operators exist due to their intrinsic definition via moduli of continuity[3] over the more standard Littlewood–Paley characterization studied in harmonic analysis (DeVore and Sharpley, 1993).

5.2.2 Generalized Radon Transforms and New Representer Theorems

In the representer theorems in Theorems 3.2 and 4.5 and Corollary 3.13, the resulting atoms have normalized singularities along hyperplanes in the sense of Donoho (1993, Definition 5). The singularities are along hyperplanes due to the Radon transform, since the Radon transform of a function is computed via its integral along hyperplanes.

A natural followup question is about defining function spaces via integral transforms along other low-dimensional manifolds, resulting in representer theorems with

[2]Our current definition hinges on the Radon transform, which is a global operator.
[3]The modulus of continuity is a local quantity.

atoms that have different kinds of normalized singularities. We believe that this should be possible through the framework of *generalized Radon transforms* (Quinto, 1980). In particular, we saw in (2.5) that the standard Radon transform $\mathscr{R}$ satisfies

$$(\mathscr{R}^*\mathscr{R})^{-1} = c_d(-\Delta)^{\frac{d-1}{2}}.$$

We believe this problem can be solved by considering generalized Radon transforms $\mathscr{G}$ such that $(\mathscr{G}^*\mathscr{G})^{-1}$ is a pseudodifferential operator, in which case there are some generic results about the inversion of the transform and its dual (Quinto, 1980). The main technical difficulty that arises in proving such a result revolves around developing a distributional theory of these transforms.

5.2.3 Alternative Banach Spaces for Vector-Valued Functions

In Remark 3.9, we saw that the space $\mathscr{R}\,\mathrm{BV}^k(\mathbb{R}^d)$ is isometrically isomorphic to $\mathcal{M}_k(\mathbb{S}^{d-1} \times \mathbb{R}) \times \mathcal{P}_{k-1}(\mathbb{R}^d)$. When defined the vector-valued versions of these spaces (in the case that $k = 2$), our definition of $\mathscr{R}\,\mathrm{BV}^2(\mathbb{R}^d; \mathbb{R}^D)$ from (3.20) and (3.21) is isometrically isomorphic to the space $\ell^1([D]; \mathcal{M}_{\text{even}}(\mathbb{S}^{d-1} \times \mathbb{R})) \times \mathcal{P}_1(\mathbb{R}^d; \mathbb{R}^D)$, where $\mathcal{P}_1(\mathbb{R}^d; \mathbb{R}^D)$ denotes the D-fold Cartesian product of $\mathcal{P}_1(\mathbb{R}^d)$. The main limitation of this construction is that the resulting shallow neural networks in the vector-valued representer theorem in Lemma 3.25 essentially correspond to D separate, decoupled scalar-output neural networks, as opposed to a vector-valued neural network with shared neurons.

Suppose $k = 2$, a natural question to ask is the existence of an analytic characterization of a Banach space isometrically isomorphic to $\mathcal{M}_{\text{even}}(\mathbb{S}^{d-1} \times \mathbb{R}; \ell^p([D])) \times \mathcal{P}_1(\mathbb{R}^d; \mathbb{R}^D)$, for some $1 < p < \infty$. The space $\mathcal{M}_{\text{even}}(\mathbb{S}^{d-1} \times \mathbb{R}; \ell^p([D]))$ is likely to promote coupling between the vector-valued outputs as opposed to the space $\ell^1([D]; \mathcal{M}_{\text{even}}(\mathbb{S}^{d-1} \times \mathbb{R}))$.

References

Adcock, Ben, and Anders C. Hansen. 2016. Generalized sampling and infinite-dimensional compressed sensing. *Foundations of Computational Mathematics* 16(5): 1263–1323.

Adcock, Ben, Anders C. Hansen, Clarice Poon, and Bogdan Roman. 2017. Breaking the coherence barrier: A new theory for compressed sensing. In *Forum of mathematics, sigma*, vol. 5. Cambridge University Press.

Aronszajn, Nachman. 1950. Theory of reproducing kernels. *Transactions of the American mathematical society* 68(3):337–404.

Arora, Raman, Amitabh Basu, Poorya Mianjy, and Anirbit Mukherjee. 2018. Understanding deep neural networks with rectified linear units. In *International conference on learning representations*.

Aziznejad, Shayan, Harshit Gupta, Joaquim Campos, and Michael Unser. 2020. Deep neural networks with trainable activations and controlled Lipschitz constant. *IEEE Transactions on Signal Processing* 68:4688–4699.

Ba, Lei Jimmy, and Rich Caruana. 2014. Do deep nets really need to be deep? In *Proceedings of the 27th international conference on neural information processing systems-volume 2*, 2654–2662.

Bach, Francis. 2017. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research* 18(1):629–681.

Balestriero, Randall, and Richard G. Baraniuk. 2020. Mad max: Affine spline insights into deep learning. *Proceedings of the IEEE*.

Barron, Andrew R. 1993. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory* 39(3):930–945.

———. 1994. Approximation and estimation bounds for artificial neural networks. *Machine Learning* 14(1):115–133.

Barron, Andrew R., Albert Cohen, Wolfgang Dahmen, and Ronald A. DeVore. 2008. Approximation and learning by greedy algorithms. *The Annals of Statistics* 36(1): 64–94.

Barron, Andrew R., and Jason M. Klusowski. 2019. Complexity, statistical risk, and metric entropy of deep nets using total path variation. *arXiv preprint arXiv:1902.00800*.

Bohn, Bastian, Christian Rieger, and Michael Griebel. 2019. A representer theorem for deep kernel learning. *Journal of Machine Learning Research* 20:1–32.

Bohra, Pakshal, Joaquim Campos, Harshit Gupta, Shayan Aziznejad, and Michael Unser. 2020. Learning activation functions in deep (spline) neural networks. *IEEE Open Journal of Signal Processing* 1:295–309.

de Boor, Carl, and Robert E. Lynch. 1966. On splines and their minimum properties. *Journal of Mathematics and Mechanics* 15(6):953–969.

Boyer, Claire, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric De Gournay, and Pierre Weiss. 2019. On representer theorems and convex regularization. *SIAM Journal on Optimization* 29(2):1260–1281.

Bredies, Kristian, and Marcello Carioni. 2020. Sparsity of solutions for variational inverse problems with finite-dimensional data. *Calculus of Variations and Partial Differential Equations* 59(1):1–26.

Bredies, Kristian, and Hanna K. Pikkarainen. 2013. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations* 19(1):190–218.

Brown, Lawrence D., and Mark G. Low. 1996. Asymptotic equivalence of nonparametric regression and white noise. *The Annals of Statistics* 24(6):2384–2398.

Bruckstein, Alfred M., David L. Donoho, and Michael Elad. 2009. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM review* 51(1):34–81.

Candès, Emmanuel J. 1998. Ridgelets: theory and applications. Ph.D. thesis, Stanford University Stanford.

———. 1999. Harmonic analysis of neural networks. *Applied and Computational Harmonic Analysis* 6(2):197–218.

———. 2003. Ridgelets: estimating with ridge functions. *The Annals of Statistics* 31(5):1561–1599.

Candès, Emmanuel J., and Justin Romberg. 2007. Sparsity and incoherence in compressive sampling. *Inverse Problems* 23(3):969.

Candès, Emmanuel J., Justin Romberg, and Terence Tao. 2006. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory* 52(2):489–509.

Chen, Scott Shaobing, David L. Donoho, and Michael A. Saunders. 2001. Atomic decomposition by basis pursuit. *SIAM review* 43(1):129–159.

Cohen, Albert, Ingrid Daubechies, and Pierre Vial. 1993. Wavelets on the interval and fast wavelet transforms. *Applied and computational harmonic analysis*.

Daubechies, Ingrid. 1992. *Ten lectures on wavelets*. SIAM.

De Castro, Yohann, and Fabrice Gamboa. 2012. Exact reconstruction using beurling minimal extrapolation. *Journal of Mathematical Analysis and applications* 395(1): 336–354.

DeVore, Ronald, Boris Hanin, and Guergana Petrova. 2021. Neural network approximation. *Acta Numerica* 30:327–444.

DeVore, Ronald A., and Robert C. Sharpley. 1993. Besov spaces on domains in $\mathbb{R}^d$. *Transactions of the American Mathematical Society* 335(2):843–864.

DeVore, Ronald A., and Vladimir N. Temlyakov. 1996. Some remarks on greedy algorithms. *Advances in Computational Mathematics* 5(1):173–187.

Donoho, David L. 1993. Unconditional bases are optimal bases for data compression and for statistical estimation. *Applied and Computational Harmonic Analysis* 1(1): 100–115.

———. 2000. High-dimensional data analysis: The curses and blessings of dimensionality. *AMS math challenges lecture* 1(2000):32.

———. 2006. Compressed sensing. *IEEE Transactions on Information Theory* 52(4):1289–1306.

Donoho, David L., and Iain M. Johnstone. 1998. Minimax estimation via wavelet shrinkage. *The Annals of Statistics* 26(3):879–921.

Dunford, Nelson, and Jacob T. Schwartz. 1988. *Linear operators, part 1: General theory*. Wiley Classics Library, Wiley.

Elad, Michael. 2010. *Sparse and redundant representations: From theory to applications in signal and image processing*. Springer Science & Business Media.

Evans, Lawrence C. 2010. *Partial differential equations*. Graduate studies in mathematics, American Mathematical Society.

Feichtinger, Hans G. 2017. A novel mathematical approach to the theory of translation invariant linear systems. In *Recent applications of harmonic analysis to function spaces, differential equations, and data science*, 483–516. Springer.

Fisher, Stephen D., and Joseph W. Jerome. 1975. Spline solutions to L^1 extremal problems in one and several variables. *Journal of Approximation Theory* 13(1): 73–83.

Folland, Gerald B. 1999. *Real analysis: Modern techniques and their applications*. 2nd ed. New York: John Wiley & Sons.

Frankle, Jonathan, and Michael Carbin. 2018. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International conference on learning representations*.

Friedman, Jerome H., and Werner Stuetzle. 1981. Projection pursuit regression. *Journal of the American Statistical Association* 76(376):817–823.

Galichon, Alfred. 2018. *Optimal transport methods in economics*. Princeton University Press.

van de Geer, Sara. 2000. *Empirical processes in m-estimation*, vol. 6. Cambridge university press.

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 315–323. JMLR Workshop and Conference Proceedings.

Golubeva, Anna, Behnam Neyshabur, and Guy Gur-Ari. 2021. Are wider nets better given the same number of parameters? *International Conference on Learning Representations*.

Grandvalet, Yves. 1998. Least absolute shrinkage is equivalent to quadratic penalization. In *International conference on artificial neural networks*, 201–206. Springer.

Grothendieck, Alexander. 1955. *Produits tensoriels topologiques et espaces nucléaires*, vol. 16. American Mathematical Society Providence.

Hayakawa, Satoshi, and Taiji Suzuki. 2020. On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. *Neural Networks* 123:343–361.

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition*, 770–778.

Helgason, Sigurdur. 2011. *Integral geometry and radon transforms*. Springer New York.

Hinton, Geoffrey, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury. 2012a. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine* 29(6):82–97.

Hinton, Geoffrey E., Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R. Salakhutdinov. 2012b. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*.

Imaizumi, Masaaki, and Kenji Fukumizu. 2019. Deep neural networks learn non-smooth functions effectively. In *The 22nd international conference on artificial intelligence and statistics*, 869–878. PMLR.

Jacot, Arthur, Franck Gabriel, and Clement Hongler. 2018. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, vol. 31.

Jin, Kyong Hwan, Michael T. McCann, Emmanuel Froustey, and Michael Unser. 2017. Deep convolutional neural network for inverse problems in imaging. *IEEE Transactions on Image Processing* 26(9):4509–4522.

John, Fritz. 1981. *Plane waves and spherical means applied to partial differential equations*. Interscience tracts in pure and applied mathematics, Interscience publishers.

Jones, Lee K. 1992. A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics* 608–613.

Kimeldorf, George S., and Grace Wahba. 1970a. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics* 41(2):495–502.

———. 1970b. Spline functions and stochastic processes. *Sankhyā: The Indian Journal of Statistics, Series A* 173–180.

———. 1971. Some results on Tchebycheffian spline functions. *Journal of mathematical analysis and applications* 33(1):82–95.

Klusowski, Jason M., and Andrew R. Barron. 2017. Minimax lower bounds for ridge combinations including neural nets. In *2017 ieee international symposium on information theory (isit)*, 1376–1380. IEEE.

———. 2018. Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls. *IEEE Transactions on Information Theory* 64(12): 7649–7656.

Korolev, Yury. 2021. Two-layer neural networks with values in a Banach space. *arXiv preprint arXiv:2105.02095*.

Kostadinova, Sanja, Stevan Pilipović, Katerina Saneva, and Jasson Vindas. 2014. The ridgelet transform of distributions. *Integral Transforms and Special Functions* 25(5):344–358.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems* 25:1097–1105.

Krogh, Anders, and John A. Hertz. 1992. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, 950–957.

Kurdila, Andrew J., and Michael Zabarankin. 2006. *Convex functional analysis*. Systems & Control: Foundations & Applications, Birkhäuser Basel.

Kurková, Věra, Paul C Kainen, and Vladik Kreinovich. 1997. Estimates of the number of hidden units and variation with respect to half-spaces. *Neural Networks* 10(6):1061–1068.

Kurková, Vera, and Marcello Sanguineti. 2001. Bounds on rates of variable-basis and neural-network approximation. *IEEE Transactions on Information Theory* 47(6): 2659–2665.

Kutyniok, Gitta. 2008. What is applied harmonic analysis? *Mitteilungen der Deutschen Mathematiker-Vereinigung* 16(2):78–84.

LeCun, Yann, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521(7553):436–444.

Lemarié, Pierre Gilles, and Yves Meyer. 1986. Ondelettes et bases Hilbertiennes. *Revista Matematica Iberoamericana* 2(1):1–18.

Logan, Benjamin F., and Larry A. Shepp. 1975. Optimal reconstruction of a function from its projections. *Duke mathematical journal* 42(4):645–659.

Ludwig, Donald. 1966. The Radon transform on Euclidean space. *Communications on Pure and Applied Mathematics* 19(1):49–81.

Lustig, Michael, David Donoho, and John M. Pauly. 2007. Sparse MRI: The application of compressed sensing for rapid mr imaging. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 58(6):1182–1195.

Mallat, Stéphane G. 1989. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 11(07):674–693.

Mammen, Enno, and Sara van de Geer. 1997. Locally adaptive regression splines. *The Annals of Statistics* 25(1):387–413.

Matoušek, Jiří. 1996. Improved upper bounds for approximation by zonotopes. *Acta Mathematica* 177(1):55–73.

McCulloch, Warren S., and Walter Pitts. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics* 5(4):115–133.

Meyer, Yves. 1992. *Wavelets and operators: Volume 1.* Cambridge Studies in Advanced Mathematics, Cambridge University Press.

Mhaskar, Hrushikesh N. 2004. On the tractability of multivariate integration and approximation by neural networks. *Journal of Complexity* 20(4):561–590.

Micchelli, Charles A. 1984. Interpolation of scattered data: distance matrices and conditionally positive definite functions. In *Approximation theory and spline functions*, 143–145. Springer.

Monroe, Don. 2022. A deeper understanding of deep learning. *Communications of the ACM* 65(6):19–20.

Montufar, Guido F., Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio. 2014. On the number of linear regions of deep neural networks. *Advances in Neural Information Processing Systems* 27:2924–2932.

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

Narcowich, Francis J., Pencho Petrushev, and Joseph D. Ward. 2006a. Decomposition of Besov and Triebel–Lizorkin spaces on the sphere. *Journal of Functional Analysis* 238(2):530–564.

———. 2006b. Localized tight frames on spheres. *SIAM Journal on Mathematical Analysis* 38(2):574–594.

Neumayer, Sebastian, and Michael Unser. 2022. Explicit representations for Banach subspaces of Lizorkin distributions. *arXiv preprint arXiv:2203.05312*.

Neyshabur, Behnam, Russ R. Salakhutdinov, and Nati Srebro. 2015a. Path-SGD: Path-normalized optimization in deep neural networks. In *Advances in neural information processing systems*, 2422–2430.

Neyshabur, Behnam, Ryota Tomioka, Ruslan Salakhutdinov, and Nathan Srebro. 2017. Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071*.

Neyshabur, Behnam, Ryota Tomioka, and Nathan Srebro. 2015b. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *International conference on learning representations (workshop)*.

———. 2015c. Norm-based capacity control in neural networks. In *Conference on learning theory*, 1376–1401. PMLR.

Ongie, Greg, Rebecca Willett, Daniel Soudry, and Nathan Srebro. 2020a. A function space view of bounded norm infinite width ReLU nets: The multivariate case. In *International conference on learning representations*.

Ongie, Gregory, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, and Rebecca Willett. 2020b. Deep learning techniques for inverse problems in imaging. *IEEE Journal on Selected Areas in Information Theory* 1(1): 39–56.

Pandey, J. N. 2011. *The Hilbert transform of Schwartz distributions and applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts, Wiley.

Parhi, Rahul, and Robert D. Nowak. 2020. The role of neural network activation functions. *IEEE Signal Processing Letters* 27:1779–1783.

———. 2021. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research* 22(43):1–40.

———. 2022a. Near-minimax optimal estimation with shallow ReLU neural networks. *Submitted*. https://arxiv.org/abs/2109.08844.

———. 2022b. On continuous-domain inverse problems with sparse superpositions of decaying sinusoids as solutions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5603–5607.

———. 2022c. What kinds of functions do deep neural networks learn? Insights from variational spline theory. *SIAM Journal on Mathematics of Data Science* 4(2): 464–489.

Peetre, Jaak. 1976. *New thoughts on Besov spaces*. Duke University mathematics series, Mathematics Department, Duke University.

Pinkus, Allan. 2015. *Ridge functions*. Cambridge Tracts in Mathematics, Cambridge University Press.

Pisier, Gilles. 1981. Remarques sur un résultat non publié de b. Maurey. *Séminaire Analyse fonctionnelle (dit* 1–12.

Poggio, Tomaso, Lorenzo Rosasco, Amnon Shashua, Nadav Cohen, and Fabio Anselmi. 2015. Notes on hierarchical splines, DCLNs and i-theory. Tech. Rep., Center for Brains, Minds and Machines (CBMM).

Quinto, Eric Todd. 1980. The dependence of the generalized Radon transform on defining measures. *Transactions of the American Mathematical Society* 257(2): 331–346.

Ramm, Alexander G., and Alexander I. Katsevich. 1996. *The Radon transform and local tomography*. Taylor & Francis.

Reed, Michael, and Barry Simon. 1972. *Methods of modern mathematical physics: Functional analysis*. Methods of Modern Mathematical Physics, Academic Press.

Reiß, Markus. 2008. Asymptotic equivalence for nonparametric regression with multivariate and random design. *The Annals of Statistics* 36(4):1957–1982.

Rioul, Olivier, and Martin Vetterli. 1991. Wavelets and signal processing. *IEEE Signal Processing Magazine* 8(4):14–38.

Rosenblatt, Frank. 1958. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological review* 65(6):386.

Rosset, Saharon, Grzegorz Swirszcz, Nathan Srebro, and Ji Zhu. 2007. ℓ_1 regularization in infinite dimensional feature spaces. In *International conference on computational learning theory*, 544–558. Springer.

Rubin, Boris. 1998. The Calderón reproducing formula, windowed X-ray transforms, and Radon transforms in L^p-spaces. *Journal of Fourier Analysis and Applications* 4(2):175–197.

Rudin, Leonid I., Stanley Osher, and Emad Fatemi. 1992. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena* 60(1-4):259–268.

Rudin, Walter. 1991. *Functional analysis*. International Series in Pure and Applied Mathematics, McGraw-Hill.

Samko, Stefan. 1995. Denseness of the spaces Φ_V of Lizorkin type in the mixed $L^{\overline{p}}(\mathbb{R}^n)$-spaces. *Studia Mathematica* 3(113):199–210.

Sanyal, Amartya, Philip H. Torr, and Puneet K. Dokania. 2019. Stable rank normalization for improved generalization in neural networks and GANs. *International Conference on Learning Representations*.

Savarese, Pedro, Itay Evron, Daniel Soudry, and Nathan Srebro. 2019. How do infinite width bounded norm networks look in function space? In *Conference on learning theory*, 2667–2690. PMLR.

Schmidt-Hieber, Johannes. 2020. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics* 48(4):1875–1897.

Schoenberg, Isaac J. 1946. Contribution to the problem of approximation of equidistant data by analytic functions. *Quarterly of Applied Mathematics* 4:45–99.

———. 1964. Spline functions and the problem of graduation. *Proceedings of the National Academy of Sciences of the United States of America* 52(4):947–950.

Schölkopf, Bernhard, and Alexander J. Smola. 2002. *Learning with kernels: Support vector machines, regularization, optimization, and beyond.* Adaptive computation and machine learning, MIT Press.

Siegel, Jonathan W., and Jinchao Xu. 2021a. Characterization of the variation spaces corresponding to shallow neural networks. *arXiv preprint arXiv:2106.15002.*

———. 2021b. Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. *arXiv preprint arXiv:2101.12365v7.*

Simon, Barry. 1971. Distributions and their Hermite expansions. *Journal of Mathematical Physics* 12(1):140–148.

Sonoda, Sho, and Noboru Murata. 2017. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis* 43(2):233–268.

Suzuki, Taiji. 2019. Adaptivity of deep ReLU network for learning in Besov and mixed smooth Besov spaces: optimal rate and curse of dimensionality. In *International conference on learning representations.*

Taubman, David., and Michael Marcellin. 2012. *JPEG2000: Image compression fundamentals, standards and practice.* The Springer International Series in Engineering and Computer Science, Springer US.

Tibshirani, Ryan J. 2014. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics* 42(1):285–323.

Tikhonov, Andrei N. 1963. On the solution of ill-posed problems and the method of regularization. In *Doklady akademii nauk*, vol. 151. No. 3, 501–504. Russian Academy of Sciences.

Trèves, François. 1967. *Topological vector spaces, distributions and kernels*. Pure and Applied Mathematics, Academic Press.

Triebel, Hans. 2008. *Function spaces and wavelets on domains*. EMS tracts in mathematics, European Mathematical Society.

Unser, Michael. 1997. Ten good reasons for using spline wavelets. In *Wavelet applications in signal and image processing V*, vol. 3169, 422–431. International Society for Optics and Photonics.

———. 1999. Splines: A perfect fit for signal and image processing. *IEEE Signal Processing Magazine* 16(6):22–38.

———. 2000. Sampling-50 years after Shannon. *Proceedings of the IEEE* 88(4): 569–587.

———. 2019. A representer theorem for deep neural networks. *Journal of Machine Learning Research* 20(110):1–30.

———. 2020. A note on BIBO stability. *IEEE Transactions on Signal Processing* 68:5904–5913.

———. 2021. A unifying representer theorem for inverse problems and machine learning. *Foundations of Computational Mathematics* 21(4):941–960.

———. 2022a. From kernel methods to neural networks: A unifying variational formulation. *arXiv preprint arXiv:2206.14625*.

———. 2022b. Ridges, neural networks, and the Radon transform. *arXiv preprint arXiv:2203.02543*.

Unser, Michael, and Shayan Aziznejad. 2022. Convex optimization in sums of Banach spaces. *Applied and Computational Harmonic Analysis* 56:1–25.

Unser, Michael, and Thierry Blu. 2000. Fractional splines and wavelets. *SIAM review* 42(1):43–67.

———. 2003. Wavelet theory demystified. *IEEE Transactions on Signal Processing* 51(2):470–483.

———. 2005. Generalized smoothing splines and the optimal discretization of the Wiener filter. *IEEE Transactions on Signal Processing* 53(6):2146–2159.

Unser, Michael, Julien Fageot, and John Paul Ward. 2017. Splines are universal solutions of linear inverse problems with generalized TV regularization. *SIAM Review* 59(4):769–793.

Vetterli, Martin, Pina Marziliano, and Thierry Blu. 2002. Sampling signals with finite rate of innovation. *IEEE Transactions on Signal Processing* 50(6):1417–1428.

Wahba, Grace. 1990. *Spline models for observational data*, vol. 59. SIAM.

Wainwright, Martin J. 2019. *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge University Press.

Wang, Hongyi, Saurabh Agarwal, and Dimitris Papailiopoulos. 2021. Pufferfish: Communication-efficient models at no extra cost. *Proceedings of Machine Learning and Systems* 3.

Wei, Colin, Jason D. Lee, Qiang Liu, and Tengyu Ma. 2019. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. *Advances in Neural Information Processing Systems* 32.

Wendland, Holger. 2004. *Scattered data approximation*. Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press.

Xu, Jinchao. 2020. Finite neuron method and convergence analysis. *Communications in Computational Physics* 28(5):1707–1745.

Xu, Yuesheng, and Qi Ye. 2019. *Generalized Mercer kernels and reproducing kernel Banach spaces*, vol. 258. American Mathematical Society.

Yang, Yuhong, and Andrew Barron. 1999. Information-theoretic determination of minimax rates of convergence. *Annals of Statistics* 1564–1599.

Zhang, Haizhang, Yuesheng Xu, and Jun Zhang. 2009. Reproducing kernel Banach spaces for machine learning. *Journal of Machine Learning Research* 10(Dec):2741–2775.

Zuhovickiĭ, S. 1948. Remarks on problems in approximation theory. *Mat. Zbirnik KDU* 169–183.