Neural Networks, Ridge Splines, and TV Regularization in the Radon Domain

Rahul Parhi
Robert D. Nowak

Department of Electrical and Computer Engineering
University of Wisconsin-Madison
Madison, WI 53706, USA

Abstract

We develop a variational framework to understand the properties of the functions learned by neural networks fit to data. We propose and study of a family of continuous-domain linear inverse problems with total variation-like regularization in the Radon domain subject to data fitting constraints. We derive a representer theorem showing that finite-width, single-hidden layer neural networks are solutions to these inverse problems. We draw on many techniques from variational spline theory and so we propose the notion of a ridge spline, which corresponds to fitting data with a single-hidden layer neural network. The representer theorem is reminiscent of the classical Reproducing Kernel Hilbert space representer theorem, but the neural network problem is set in a non-Hilbertian Banach space. Although the learning problems are posed in the continuous-domain, similar to kernel methods, the problems can be recast as finite-dimensional neural network training problems. These neural network training problems have regularizers which are related to the well-known weight decay and path-norm regularizers. Thus, our result gives insight into functional characteristics of trained neural networks and also into the design neural network regularizers. We also show that these regularizers promote neural network solutions with desirable generalization properties.

Keywords: neural networks, splines, inverse problems, regularization, sparsity

1. Introduction

Single-hidden layer neural networks are superpositions of ridge functions. A ridge function is any multivariate function mapping $\mathbb{R}^d \to \mathbb{R}$ of the form

$$x \mapsto \rho(w^T x),$$

where $\rho : \mathbb{R} \to \mathbb{R}$ is a univariate real-valued function and $w \in \mathbb{R}^d \setminus \{0\}$. Single-hidden layer neural networks, in particular, are superpositions of the form

$$x \mapsto \sum_{k=1}^K v_k \rho(w_k^T x - b_k),$$

where $\rho : \mathbb{R} \to \mathbb{R}$ is a fixed activation function, $K$ is the width of the network, and for $k = 1, \ldots, K$, $v_k \in \mathbb{R}$ and $w_k \in \mathbb{R}^d \setminus \{0\}$ are the weights of the neural network and $b_k \in \mathbb{R}$ are the biases or offsets. Variants of the well-known universal approximation theorem state that any continuous function can be approximated arbitrarily well by a superposition of the
form in (1.2), under extremely mild conditions on the activation function (Cybenko, 1989; Hornik et al., 1989; Funahashi, 1989; Barron, 1993; Leshno et al., 1993).

Ridge functions are ubiquitous in mathematics and engineering, especially due to the popularity of neural networks, and we refer to the book of Pinkus (2015) and the survey of Konyagin et al. (2018) for a fairly up-to-date treatment on the current state of research regarding ridge functions. One of the most popular areas of research has been regarding approximation theory with superpositions of ridge functions (i.e., single-hidden layer neural networks), in which there are many papers establishing optimal or near-optimal approximation rates for various function spaces (Maiorov, 2010; Klusowski and Barron, 2016a; Mhaskar, 2020). Another, less popular (though practically more interesting), research area studies what happens when you fit data with a single-hidden layer neural network. This question has been viewed from both a statistical perspective, where risk bounds are established (Klusowski and Barron, 2016b), and more recently, in the univariate case, from a functional analytic perspective, where connections to classical spline theory are established (Savarese et al., 2019; Parhi and Nowak, 2019; Williams et al., 2019). We also remark that these questions have also been studied in the context of deep neural networks. See, for example, Shaham et al. (2018); Grohs et al. (2019) for approximation theory, Barron and Klusowski (2019) for statistical properties, and Balestriero and Baraniuk (2018); Unser (2019a); Ergen and Pilanci (2020) for connections to splines.

Although the term ridge function is rather modern, it is important to note that such functions have been studied for many years under the name plane waves. Much of the early work with plane waves revolves around representing solutions to partial differential equations (PDE), e.g., the wave equation, as a superposition of plane waves. We refer the reader to the classic book of John (2013) for a full treatment of this subject. The key analysis tool used in these PDE problems is the Radon transform. Since a ridge function as in (1.1) is constant along the hyperplanes \( w^T x = c, c \in \mathbb{R} \), analysis of such functions becomes convenient in the Radon domain. More modern applications of ridge functions arise in computerized tomography following the seminal paper of Logan and Shepp (1975), where they coined the term "ridge function", and the development of ridgelets in the 1990s, a wavelet-like system inspired by neural networks, independently proposed by Murata (1996), Rubin (1998), and Candès (1998, 1999). Many refinements to the ridgelet transform have been made recently (Kostadinova et al., 2014; Sonoda and Murata, 2017). As one might expect, the main analysis tool used in these applications is the Radon transform. Thus, we see that ridge functions and the Radon transform are intrinsically connected.

Recent work from the machine learning community has used this connection to understand what kinds of functions can be represented by infinite-width (continuum-width) single-hidden layer neural networks with Rectified Linear Unit (ReLU) activation functions, where the “size” of the network weights is bounded (Ongie et al., 2020). In particular, they prove that the “size” of the network weights (specifically the Euclidean norm of the network weights) corresponds to a seminorm in the Radon domain of the function represented by a infinite-width network.

This paper focuses on the practical problem of fitting a finite-width neural network to finite-dimensional data, with an eye towards characterizing the properties of the resulting functions. We view this problem as a recovery problem where we wish to recover an unknown function from linear measurements. We deviate from the usual finite-dimensional recovery
paradigm and pose the problem in the continuous-domain, allowing us to use techniques from the theory of variational methods. We show that continuous-domain linear inverse problems with total variation regularization in the Radon domain admit sparse atomic solutions, with the atoms being the familiar neurons of a neural network.

1.1 Contributions

Let $\mathcal{F}$ be a real vector space of multidimensional functions, $\mathcal{V} : \mathcal{F} \to \mathbb{R}^N$ a continuous linear sensing or measurement operator (N can be viewed as the number of measurements or data), and let $f : \mathbb{R}^d \to \mathbb{R}$ be a multidimensional function such that $f \in \mathcal{F}$. Consider the continuous-domain inverse problem

$$
\min_{f \in \mathcal{F}} G(\mathcal{V}f) + \|f\|,
$$

where $\|\cdot\| : \mathcal{F} \to \mathbb{R}_{\geq 0}$ is a (semi)norm or regularizer and $G : \mathbb{R}^N \to \mathbb{R}$ is a convex data fitting term.

We summarize the contributions of this paper below.

1. Our main result is the development of a family of seminorms $\|\cdot\|_{(m)}$ (indexed by $m \in 2\mathbb{Z}$, $m \geq 2$) of total variation seminorms in the Radon domain so that the solutions to generalized scattered data approximation problem (1.3) with $\|\cdot\| := \|\cdot\|_{(m)}$ take the form

$$
x \mapsto \sum_{k=1}^K v_k \rho_m(u_k^T x - b_k) + c(x),
$$

where $\rho_m = \max\{0, \cdot\}^{m-1}/(m-1)!$ is the $m$th order truncated power function, $c(\cdot)$ is a polynomial of degree strictly less than $m$, and $K \leq N$. These seminorms are inspired by the seminorm proposed in Ongie et al. (2020), which is equivalent to $\|\cdot\|_{(m)}$ with $m = 2$. Specifically, the seminorm $\|f\|_{(m)}$ is the total variation (TV) norm (in the sense of measures) of $\Lambda^{d-1} \mathcal{R} \Delta^{m/2} f$, where $\Delta$ is the Laplacian operator, $\mathcal{R}$ is the Radon transform, and $\Lambda^{d-1}$ is a kind of ramp filter. In other words, our main result is the derivation of a neural network representer theorem. Our result says that single-hidden layer neural networks are solutions to continuous-domain linear inverse problems with TV regularization in the Radon domain. We also remark when $m = 2$, the problem corresponds to ReLU networks.

2. We propose the notion of a ridge spline by noticing that our problem formulation in (1.3) is similar to those studied in variational spline theory (Prenter, 2013; Duchon, 1977; Unser et al., 2017), with the key twist being that our family of seminorms are in the Radon domain. Thus, we refer to the solutions to (1.4) with our family of seminorms as ridge splines to emphasize that the solutions are superpositions of ridge functions. We view our notion of a ridge spline as a kind of spline inbetween a univariate spline and a traditional multivariate spline. Ridge splines are a piecewise polynomial approximation of multivariate scattered data. Moreover, by specializing

---

1. For example, $\mathcal{V}f = (f(x_1), \ldots, f(x_N)) \in \mathbb{R}^N$, for some data $\{x_n\}_{n=1}^N \subset \mathbb{R}^d$.
2. The evenness requirement, $m \in 2\mathbb{Z}$, arises due to the symmetries of the Radon domain.
our result to the univariate case, our notion of a ridge spline exactly coincides with the notion of a univariate polynomial spline.

3. By specializing our main result to setting in which \( V \) corresponds to \textit{ideal sampling}, i.e., point evaluations, the generality of (1.3) allows us to consider the machine learning problem of approximating the scattered data \( \{(x_n, y_n)\}_{n=1}^{N} \subset \mathbb{R}^d \times \mathbb{R} \) with both \textit{interpolation constraints} in the case of noise-free data as well as \textit{regularized problems} where we have soft-constraints in the case of noisy data. Thus, a direct consequence of our representer theorem result says the infinite-dimensional problem in (1.3) can be re-cast as a \textit{finite-dimensional neural network training problem} with various regularizers that are related to weight decay (Krogh and Hertz, 1992) and path-norm (Neyshabur et al., 2015) regularizers, which are used in practice. In other words, a neural network trained to fit data with an appropriate regularizer is “optimal” in the sense of the seminorm \( \| \cdot \|_{(m)} \). We also note that in these neural network training problems, it is sufficient that the width \( K \) of the network be \( N \), the size of the data.

4. Specializing our results to the supervised learning problem of binary classification shows that neural network solutions with small seminorm make good predictions on new data. Binary classification corresponds to the ideal sampling setting, restricting ourselves to \( y_n \in \{-1, +1\}, n = 1, \ldots, N \), and predicting these by the sign of the function that solves (1.3) (this can be done with an appropriate data fitting term). We derive statistical \textit{generalization bounds} for the class of neural networks with uniformly bounded seminorm \( \| \cdot \|_{(m)} \). In particular, we show that the seminorm bounds the \textit{Rademacher complexity} of these neural networks and use standard results from machine learning theory to relate this to the generalization error. This says that a small seminorm implies good generalization properties.

1.2 Related work

The closest work to this paper is that of Ongie et al. (2020), where the authors consider infinite-width (continuum-width) single-hidden layer ReLU networks where the Euclidean norm of the network weights are bounded. They ask the question about what functions can be represented by such infinite-width, but bounded norm, networks. They show that a TV seminorm in the Radon domain exactly captures the Euclidean norm of the network weights, but do not address the optimization problem of fitting neural networks to data. Inspired by this seminorm, we develop and study a family of TV seminorms in the Radon domain and consider the problem of scattered data approximation. We show that single-hidden layer neural networks, with fewer neurons than data, are solutions to the problem of minimizing these seminorms over the space of all functions of bounded seminorms, subject to data fitting constraints.

Although our main result might seem obvious on a surface level, actually proving it is quite delicate. The problem of learning from a continuous dictionary of atoms with TV-like regularization has been studied before, both in the context of splines (Fisher and Jerome, 1975; Mammen and van de Geer, 1997) and machine learning (Rosset et al., 2007; Bach, 2017). It is extremely important to note that all of these prior works (Fisher and Jerome, 1975; Mammen and van de Geer, 1997; Rosset et al., 2007; Bach, 2017) make the
vital assumption that the relevant spaces are bounded and hence (after taking the closure) compact. This allows appealing to standard compactness arguments, which are useful for proving, e.g., that minimizers to their problem even exist.

Since the Radon domain is an unbounded domain, we cannot appeal to these types of arguments for the problem we study. Thus, the first question we ask, and subsequently answer, regards existence of solutions to (1.3) with our family of seminorms. To this end, we draw on techniques from the recently developed variational framework of L-splines (Unser et al., 2017). We also remark that we cannot directly apply the results from this framework since the fundamental assumption about splines is that spline atoms are translates of a single function. Meanwhile, neural network atoms as in (1.2) are parameterized by both a direction $w_k$ and a translation $b_k$. We also draw on several recent results from variational methods (Boyer et al., 2019; Bredies and Carioni, 2020). Thus, the results of this paper provide a very general variational framework as well as novel insights into understanding the properties of functions learned by neural networks fit to data.

1.3 Roadmap
In Section 2 we state our main results and highlight some of the technical challenges and novelties in proving our results. In Section 3 we introduce the notation and mathematical formulation used throughout the paper. In Section 4 we prove our main result, the representer theorem. In Section 5 we discuss connections between ridge splines and classical spline theory. In Section 6 we discuss applications of the representer theorem to neural network training and regularization.

2. Main Results
Our main contribution is a representer theorem for problems of the form in (1.3) with our proposed family of seminorms. Our other contributions are (rather straightforward) corollaries to this result. In this section we will state the main results of this paper along with relevant historical remarks.

2.1 The representer theorem
The notion of a representer theorem is a fundamental result regarding kernel methods (Kimeldorf and Wahba, 1971; Schölkopf et al., 2001; Schölkopf and Smola, 2002). In particular, let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be any real-valued Hilbert space on $\mathbb{R}^d$ and consider the scattered data $\{(x_n, y_n)\}_{n=1}^N \subset \mathbb{R}^d \times \mathbb{R}$. Then, the classical representer theorem considers

$$\hat{f} = \arg \min_{f \in \mathcal{H}} \sum_{n=1}^N \ell(f(x_n), y_n) + \lambda \|f\|^2_{\mathcal{H}}, \quad (2.1)$$

where $\ell(\cdot, \cdot)$ is a convex loss function and $\lambda > 0$ is an adjustable regularization parameter. The representer theorem then states that the solution $\hat{f}$ is unique and $\hat{f} \in \text{span}\{k(\cdot, x_n)\}_{n=1}^N$, 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 (Unser et al., 2017; Boyer et al., 2019) as a way to designate a parametric formulation of solutions to an optimization problem, ideally being a linear combination from some dictionary of atoms. This has allowed much more general problems to be considered than ones like (2.1), which are restricted to regularizers which are Hilbertian (semi)norms. In particular, some of the recent theory is able to to consider problems where the search space is a locally convex topological vector space and the regularizers being a seminorm defined on that space (Bredies and Carioni, 2020). The main utility of these more general representer theorems arise in understanding sparsity-promoting regularizers such as the \(\ell_1\)-norm or its continuous-domain analogue, the \(M(\mathbb{R}^d)\)-norm (the total variation norm in the sense of measures), of which the structural properties of the solutions are still not completely understood, though a theory is beginning to emerge. The generality of these kinds of representer theorems has been especially useful in some of the recent development of the notion of reproducing kernel Banach spaces (Zhang et al., 2009; Xu and Ye, 2019) and of 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).

We build off of these recent results, and propose a family of seminorms\(^3\) (indexed by \(m \in 2\mathbb{Z}, \ m \geq 2\))

\[
\|f\|_{(m)} := c_d \left\| \Lambda^{d-1} \mathcal{R} \Delta^{m/2} f \right\|_{\mathcal{M}(\mathbb{P}^d)},
\]

where, \(c_d\) is a dimension dependent constant defined in (3.9), \(\Lambda^{d-1}\) is a kind of ramp filter in the Radon domain defined in (3.6), \(\mathcal{R}\) is the Radon transform defined in (3.2), \(\Delta = \sum_{k=1}^d \partial^2 / \partial x_k^2\) is the Laplacian operator, \(\mathbb{P}^d\) denotes the Radon domain as defined in Theorem 3.3 and Remark 3.4, and \(\mathcal{M}(\mathbb{P}^d)\) denotes the total variation norm (in the sense of measures) on the Radon domain.

We remark that the Radon transform computes integrals over hyperplanes and \(\mathbb{P}^d\) denotes the space of hyperplanes in \(\mathbb{R}^d\). \(\mathbb{P}^d\) can be viewed as “half” of \(S^{d-1} \times \mathbb{R}\), where \(S^{d-1}\) denotes the surface of the \(\ell^2\)-unit-sphere in \(\mathbb{R}^d\), in the sense that \(S^{d-1} \times \mathbb{R}\) is a double covering of \(\mathbb{P}^d\). This follows from the fact that any hyperplane in \(\mathbb{R}^d\) can be written as \(\gamma^T x = t\) where \((\gamma, t) \in S^{d-1} \times \mathbb{R}\), but the exact same hyperplane can also be defined with \((-\gamma, -t)\). In other words, we can associate functions defined on \(\mathbb{P}^d\) with even functions defined on \(S^{d-1} \times \mathbb{R}\). This is discussed further in Remark 3.4.

We also remark that the total variation norm \(\|\cdot\|_{\mathcal{M}(\mathbb{P}^d)}\) can be viewed, formally, as the \(L^1(\mathbb{P}^d)\)-norm. In particular, we have the containment \(L^1(\mathbb{P}^d) \subset \mathcal{M}(\mathbb{P}^d)\), but \(\mathcal{M}(\mathbb{P}^d)\) also includes “absolutely integrable tempered distributions”, such as the Dirac impulse. We also remark that we will also consider the space \(\mathcal{M}(\mathbb{R}^d)\), and these spaces will be defined more carefully in Section 3.

The seminorms in (2.2) are thus exactly total variation seminorms in the Radon domain. For brevity, we will write

\[
R_m := c_d \Lambda^{d-1} \mathcal{R} \Delta^{m/2}.
\]

\(^3\) The operator that appears in the seminorms in (2.2) is understood in the *distributional sense.*

\(^4\) The requirement that \(m\) is even is discussed in Remarks 3.7 and 4.5.
Before stating our representer theorem, we remark that our result requires that the null space of the operator $R_m$ is small, i.e., finite-dimensional. As mentioned in Unser et al. (2017) and in the $L^2$ theory of radial basis functions and polyharmonic splines (Wendland, 2010, Chapter 10), making sure the null space of an operator acting on multivariate functions nearly impossible\textsuperscript{5}. To bypass this technicality, we use the same technique as in Unser et al. (2017) and impose a growth restriction to the functions of interest via the weighted Lebesgue space $L^{\infty,n_0}(\mathbb{R}^d)$ (not to be confused with the Lorentz spaces), defined via the weighted $L^{\infty}(\mathbb{R}^d)$-norm

$$
\|f\|_{L^{\infty,n_0}(\mathbb{R}^d)} = \text{ess sup}_{x \in \mathbb{R}^d} |f(x)|(1 + \|x\|_2)^{-n_0},
$$

where $n_0 \in \mathbb{Z}$ is the algebraic growth rate. In other words, the space $L^{\infty,n_0}(\mathbb{R}^d)$ is the space of functions mapping $\mathbb{R}^d \rightarrow \mathbb{R}$ with algebraic growth rate $n_0$. We will later see in (4.4) that the appropriate choice of algebraic growth rate for the operator $R_m$ is $n_0 := m - 1$. This allows us to define the (growth restricted) null space of $R_m$ as

$$
\mathcal{N}_m := \{ q \in L^{\infty,m-1}(\mathbb{R}^d) : R_m q = 0 \} \tag{2.3}
$$

and the (growth restricted) native space $R_m$ as

$$
\mathcal{F}_m := \{ f \in L^{\infty,m-1}(\mathbb{R}^d) : \|R_m f\|_{M(\mathbb{R}^d)} < \infty \}. \tag{2.4}
$$

We prove in Lemma 4.6 that $\mathcal{N}_m$ is indeed finite-dimensional. We now state our representer theorem.

**Theorem 2.1** Assume the following:

1. The data fitting term $G : \mathbb{R}^N \rightarrow \mathbb{R}$ is a proper strictly convex function that is coercive, and lower semi-continuous with respect to the topology of $\mathbb{R}^N$, which is the standard topology on finite-dimensional spaces.

2. The measurement operator $\mathcal{V} : \mathcal{F}_m \rightarrow \mathbb{R}^N$ is a continuous linear operator.

3. The recovery problem is well-posed over the null space of $R_m$, i.e., $\mathcal{V}q_1 = \mathcal{V}q_2$ if and only if $q_1 = q_2$, for any $q_1, q_2 \in \mathcal{N}_m$.

Then, there exists a sparse minimizer to the general minimization problem

$$
\min_{f \in \mathcal{F}_m} G(\mathcal{V}f) + \|R_m f\|_{M(\mathbb{R}^d)} \tag{2.5}
$$

that is necessarily a single-hidden layer neural network of the form

$$
s(x) = \sum_{k=1}^{K} v_k \rho_m(w_k^T x - b_k) + c(x), \tag{2.6}
$$

where the number of neurons is $K \leq N - \dim \mathcal{N}_m$, $\rho_m = \max\{0, \cdot\}^{m-1}/(m-1)!$ is the $m$th order truncated power function, $w_k \in \mathbb{S}^{d-1}$, $v_k \in \mathbb{R}$, $b_k \in \mathbb{R}$, and $c(\cdot)$ is a polynomial of degree strictly less than $m$.

\textsuperscript{5} Consider $\Delta$, the Laplacian operator in $\mathbb{R}^d$. Its null space is the space of harmonic functions which is infinite-dimensional. On the other hand, the univariate Laplacian operator, $d^2/dx^2$, has a finite-dimensional null space which is simply span$\{1, x\}$. 

Remark 2.2 Theorem 2.1 shows that while the problem is posed in the continuum, it admits parametric solutions in terms of a finite number of neurons. This demonstrates the sparsifying effect of the $M(P_d)$-norm, similar to its discrete analogue, the $\ell^1$-norm.

Remark 2.3 The fact that $w_k \in S^{d-1}$ does not restrict the single-hidden layer neural network in anyway whatsoever. Indeed, given any single-hidden layer neural network with $w_k \in \mathbb{R}^d \setminus \{0\}$, we can always use the fact that $\rho_m$ is homogeneous of degree $m-1$ to rewrite the network as

$$x \mapsto \sum_{k=1}^K v_k \|w_k\|^{m-1} \rho_m(\tilde{w}_k^T x - \tilde{b}_k) + c(x),$$

where $\tilde{w}_k := w_k/\|w_k\|_2 \in S^{d-1}$ and $\tilde{b}_k := b_k/\|w_k\|_2 \in \mathbb{R}$. We use this fact to prove Proposition 2.13 which, considers finite-dimensional neural network training problems (with no constraints on the input layer weights) that are equivalent to the problem in (2.5).

Remark 2.4 The polynomial term $c(x)$ that appears in (2.6) corresponds to a term in the null space $N_m$. When $m = 2$, the network in (1.1) is a ReLU network and $c(x)$ takes the form

$$c(x) = u^T x + s,$$

where $u \in \mathbb{R}^d$ and $s \in \mathbb{R}$, i.e., when $m = 2$, (2.6) corresponds to a ReLU network with skip connections (He et al., 2016).

Proving Theorem 2.1 hinges on several technical results, the most important being the topological structure of the native space $F_m$. In order to do any kind of analysis (e.g., proving that minimizers of (2.5) even exist), we require the native space $F_m$ to have some “nice” topological structure. We prove in Theorem 4.9 that $F_m$, when equipped with a proper direct-sum topology, is a Banach space. This key result hinges on being able to construct a stable right inverse of the operator $R_m$, which we outline in Lemma 4.8. We remark that exhibiting a Banach space structure of the native space of an operator is common in variational inverse problems, e.g., in the theory of L-splines (Unser et al., 2017; Unser and Fageot, 2019). We do remark, however, our result is, to the best of our knowledge, the first time exhibiting this structure on a non-Euclidean domain. Due to the fact that the Radon domain $P^d$ does not carry the topology of the Euclidean space $\mathbb{R}^d$, we run into some nuances compared to the prior work in Unser et al. (2017); Unser and Fageot (2019).

2.2 Ridge splines

Splines and variational problems are tightly connected (Duchon, 1977; Prenter, 2013; Unser et al., 2017). In the framework of L-splines (Unser et al., 2017), a pseudodifferential operator, $L : \mathcal{S}'(\mathbb{R}^d) \rightarrow \mathcal{S}'(\mathbb{R}^d)$, where $\mathcal{S}'(\mathbb{R}^d)$ denotes the space of tempered distributions on $\mathbb{R}^d$, is associated with a spline, and variational problems of the form

$$\min_{f \in F_m} G(Vf) + \|Lf\|_{\mathcal{M}(\mathbb{R}^d)},$$

(2.7)

where $G$ is a data fitting term, $V$ is a measurement operator, and $F_m$ is the native space of $L$. The key result from Unser et al. (2017) is a representer theorem for the above continuous-domain inverse problem which states that there exists a sparse solution which is a so-called L-spline.
Remark 2.5 In this paper we will be working with both Dirac impulses in $\mathcal{M}(\mathbb{P}^d)$ and in $\mathcal{M}(\mathbb{R}^d)$. To make the domain clear, we will subscript “$\delta$” with the appropriate domain, i.e., $\delta_{\mathbb{R}^d} \in \mathcal{M}(\mathbb{R}^d)$ and $\delta_{\mathbb{P}^d} \in \mathcal{M}(\mathbb{P}^d)$.

Definition 2.6 (nonuniform L-spline (Unser et al., 2017, Definition 2)) A function $s: \mathbb{R}^d \to \mathbb{R}$ (of slow growth) is said to be a nonuniform L-spline if

$$L\{s\} = \sum_{k=1}^{K} v_k \delta_{\mathbb{R}^d}(\cdot - x_k),$$

where $\{v_k\}_{k=1}^{K}$ is a sequence of weights and the locations of Dirac impulses are at the spline knots $\{x_k\}_{k=1}^{K}$.

Remark 2.7 When $d = 1$ and $L$ is the $m$th derivative operator, the corresponding L-splines are the well-known polynomial splines of order $m$ (degree $m - 1$).

\[\text{ Figure 1: In the left plot we have a cubic spline with 7 knots. After applying } D^4, \text{ the fourth derivative operator, we are left with 7 Dirac impulses as seen in the right plot.}\]

Definition 2.8 (nonuniform polynomial ridge spline) A function $s: \mathbb{R}^d \to \mathbb{R}$ (of slow growth) is said to be a nonuniform polynomial ridge spline of order $m$ if

$$R_m\{s\} = \sum_{k=1}^{K} v_k \delta_{\mathbb{P}^d}(\cdot - z_k),$$

where $\{v_k\}_{k=1}^{K}$ is a sequence of weights and the locations of the Dirac impulses are at $z_k = (w_k, b_k) \in \mathbb{P}^d$. The collection $\{z_k\}_{k=1}^{K}$ can be viewed as a collection of a kind of Radon domain spline knot.

Remark 2.9 The term ridge spline seems to have been used once before in Klusowski and Barron (2016a) to refer to a finite-width single-hidden layer neural network, though Klusowski and Barron (2016a) makes no connections to scattered data approximation, which is the usual setting for splines.
Figure 2: On the left plot we have a two-dimensional cubic ridge spline with 7 neurons. After applying $\Delta^2$, we get an “impulse sheet”, i.e., a mapping of the form $x \mapsto \delta_{\mathbb{R}}(w^T x - b)$, for each neuron, designated by the black lines in the top down view of the cubic ridge spline in the middle plot. Then, after applying the Radon transform and ramp filter to the middle plot, we arrive with 7 Dirac impulses in the Radon domain, which are designated by the dots in the left plot. We have parameterized the directions in the Radon domain by $\theta \in [0, \pi]$. This parameterization of the two-dimensional Radon domain is known as a sinogram.

Remark 2.10 When $s$ takes the form of a single-hidden layer neural network as in (2.6), (2.8) holds. In other words, single-hidden layer neural networks with truncated power function activation functions are polynomial ridge splines. The way to understand this is that the atoms of the single-hidden layer neural network as in (1.1) are “sparsified” by $R_m$ in the sense that

$$R_m r_{(w,b)}^{(m)} = \delta_{\mathbb{P}^d}(\cdot - (w, b)),$$

where $r_{(w,b)}^{(m)}(x) := \rho_m(w^T x - b), w \in S^{d-1}, b \in \mathbb{R}$. We show that this is true in Lemma 4.2. In other words, $r_{(w,b)}^{(m)}$ can be viewed as a kind of translated Green’s function of $\Lambda^{d-1} R \Delta^{m/2}$, where the translation is in the Radon domain.

We illustrate the sparsifying effect of the operator $L$ in the case of cubic splines, i.e., $L = D^4$, the fourth derivative operator, in Figure 1. We also illustrate the sparsifying effect of the operator $R_m$ in the case of cubic ridge splines, i.e., $m = 4$, in Figure 2.

Remark 2.11 In the univariate case ($d = 1$), our notion of a polynomial ridge spline of order $m$ exactly coincides with the classical notion of a univariate polynomial spline of order $m$. We show this in Section 5.1.

Remark 2.12 Notice that the operator $R_m = c_d \Lambda^{d-1} R \Delta^{m/2}$ is an $m$th order derivative operator, followed by the Radon transform then the ramp filter. The Radon transform and ramp filter simply map the problem to the Radon domain. Theorem 2.1 says the solutions to such a problem are polynomial ridge splines of order $m$. Thus, we see that the problem we are studying is essentially the $L$-spline problem posed in the Radon domain. We explore this viewpoint in Section 5.2. This is unsurprising and often seen when generalizing
notions of univariate functions to multivariate functions via a ridge function type construction. Indeed, the ridgelet transform is just a univariate wavelet transform in the Radon domain (Kostadinova et al., 2014; Sonoda and Murata, 2017).

2.3 Scattered data approximation and neural network training

Since Theorem 2.1 says that a single-hidden layer neural network as in (2.6) is a solution to the continuous-domain inverse problem in (2.5), we can recast the continuous-domain problem in (2.5) as the equivalent finite-dimensional neural network training problem

\[
\min_{\theta \in \Theta} G(\mathcal{V} f_{\theta}) + \|R_m f_{\theta}\|_{\mathcal{M}(\mathbb{P}^d)},
\]

so long as the number of neurons \(K\) is large enough\(^6\) \((K \geq N\) suffices), where

\[
f_{\theta}(x) := \sum_{k=1}^{K} v_k \rho_m(w_k^T x - b_k) + c(x),
\]

where \(\theta = (w_1, \ldots, w_K, v_1, \ldots, v_K, b_1, \ldots, b_K, c)\) contains the neural network parameters and \(\Theta\) is the collection of all \(\theta\) such that \(v_k \in \mathbb{R}\), \(w_k \in \mathbb{R}^d\), and \(b_k \in \mathbb{R}\) for \(k = 1, \ldots, K\), and where \(c\) is a polynomial of degree strictly less than \(m\). We show in Lemma 6.1 that

\[
\|R_m f_{\theta}\|_{\mathcal{M}(\mathbb{P}^d)} = \sum_{k=1}^{K} |v_k| \|w_k\|_2^{m-1},
\]

and then use this fact to show that (2.9) is equivalent to two neural network training problems with variants of well-known neural network regularizers in the following proposition.

**Proposition 2.13** The continuous-domain problem in (2.5) is equivalent to the finite-dimensional neural network training problem in (2.9) which is subsequently equivalent to

\[
\min_{\theta \in \Theta} G(\mathcal{V} f_{\theta}) + \sum_{k=1}^{K} |v_k| \|w_k\|_2^{m-1}
\]

(2.10)

and is also equivalent to

\[
\min_{\theta \in \Theta} G(\mathcal{V} f_{\theta}) + \frac{1}{2} \sum_{k=1}^{K} |v_k|^2 + \|w_k\|_2^{2m-2},
\]

(2.11)

so long as the number of neurons \(K \geq N - \dim \mathcal{N}_m\).

**Remark 2.14** Said differently, the infinite-dimensional problem in (2.5), the finite-dimensional problem in (2.10), and the finite-dimensional problem in (2.11) are all equivalent optimizations.

---

6. We characterize what large enough means in Proposition 2.13
Remark 2.15 When $m = 2$, which coincides with neural networks with ReLU activation functions, (2.10) and (2.11) correspond to previously studied training problems. The regularizer in (2.10) exactly coincides with the notion of the $\ell^1$-path-norm regularization as proposed in Neyshabur et al. (2015) and the regularizer in (2.11) exactly coincides with the notion of training a neural network with weight decay as proposed in Krogh and Hertz (1992). Thus, our result shows that these notions of regularization are intrinsically tied to the ReLU activation function, and perhaps variants such as the regularizers that appear in (2.10) and (2.11) should be used in practice for non-ReLU activation functions, where $m - 1$ could corresponds to the algebraic growth rate of the activation function.

In machine learning, the measurement model is usually taken to be ideal sampling, i.e., the measurement operator $\mathcal{V}$ acts on a function $f : \mathbb{R}^d \to \mathbb{R}$ via

$$
\mathcal{V} : f \mapsto \begin{bmatrix}
\langle \delta(\cdot - x_1), f \rangle \\
\vdots \\
\langle \delta(\cdot - x_N), f \rangle 
\end{bmatrix} = 
\begin{bmatrix}
f(x_1) \\
\vdots \\
f(x_N)
\end{bmatrix} \in \mathbb{R}^N,
$$

so machine learning problem is to approximate the scattered data $\{(x_n, y_n)\}_{n=1}^N \subset \mathbb{R}^d \times \mathbb{R}$. The generality of our main result in Theorem 2.1 says that the solutions problems with interpolation constraints in the case of noise-free data

$$
\min_{f \in \mathcal{F}} \|R_m f\|_{\mathcal{M}(\mathbb{R}^d)} \quad \text{s.t.} \quad f(x_n) = y_n, \ n = 1, \ldots, N
$$

and to regularized problems where we have soft-constraints in the case of noisy data

$$
\min_{f \in \mathcal{F}} \sum_{n=1}^N \ell(f(x_n), y_n) + \lambda \|R_m f\|_{\mathcal{M}(\mathbb{R}^d)}, \quad (2.12)
$$

where $\lambda > 0$ is an adjustable regularization parameter and $\ell(\cdot, \cdot)$ is an appropriate loss function, e.g., the squared error loss, are single-hidden layer neural networks. We can then invoke Proposition 2.13 to recast the continuous-domain problem in (2.12) with either of the equivalent finite-dimensional neural network training problems:

$$
\min_{\theta \in \Theta} \sum_{n=1}^N \ell(f_\theta(x_n), y_n) + \lambda \left( \sum_{k=1}^K |v_k| \|w_k\|_{2}^{m-1} \right) \\
\min_{\theta \in \Theta} \sum_{n=1}^N \ell(f_\theta(x_n), y_n) + \lambda \left( \frac{1}{2} \sum_{k=1}^K |v_k|^2 + \|w_k\|_{2}^{2m-2} \right),
$$

so long as the number of neurons $K$ is large enough as stated in Proposition 2.13. The two problems in the above display correspond to how neural network training problems are actually setup in the machine learning problem.

2.4 Statistical generalization bounds

Neural networks are widely used for pattern classification. In the ideal sampling scenario, the generality of our main result in Theorem 2.1 allows us to consider optimizations of the
\begin{align}
\min_{f \in \mathcal{F}} \sum_{n=1}^{N} \ell(y_n f(x_n)) \quad \text{s.t.} \quad \|R_m f\|_{\mathcal{M}(\mathcal{P}^d)} \leq B, \label{eq:1.3}
\end{align}

for some constant \( B < \infty \), where \( \ell(\cdot) \) is an appropriate \( L \)-Lipschitz loss function of a scalar quantity. If we assume that \( \{(x_n, y_n)^{N}_{n=1}\} \) are drawn independently and identically from some unknown underlying probability distribution, \( y_n \in \{-1, +1\} \), \( n = 1, \ldots, N \), and the loss function assigns positive losses when \( \text{sgn}(f(x_n)) \neq y_n \) (or equivalently when \( y_n f(x_n) < 0 \)), this is the binary classification setting. Given this set up, it is natural to examine if solutions to \( \eqref{eq:1.3} \) predict well on new random examples \( (x, y) \) drawn independently from the same underlying distribution.

We can invoke Proposition \ref{prop:2.13} and consider optimization over neural network parameters by considering the equivalent optimization to \( \eqref{eq:1.3} \)
\begin{align}
\min_{\theta \in \Theta} \sum_{n=1}^{N} \ell(y_n f_{\theta}(x_n)) \quad \text{s.t.} \quad \sum_{k=1}^{K} |v_k| \|w_k\|_{2}^{m-1} \leq B, \label{eq:1.4}
\end{align}

where
\[
f_{\theta}(x) := \sum_{k=1}^{K} v_k \rho_m(w_k^T x - b_k) + c(x).
\]

This allows us to bound the error probability of binary classifiers that solve the optimization in the above display. Let \( \bar{f} \) be a minimizer of the optimization in \( \eqref{eq:1.4} \) and let
\[
\mathcal{F}_{m,B} := \left\{ f_{\theta} : \theta \in \Theta, \sum_{k=1}^{K} |v_k| \|w_k\|_{2}^{m-1} \leq B, K \geq 0 \right\}.
\]

We show that \( B \) directly controls the error probability of \( \bar{f} \), i.e., \( \mathbb{P}(y \bar{f}(x) < 0) \), where \( (x, y) \) is an independent sample from the underlying distribution. This is referred to as the generalization error in machine learning parlance. We follow the standard approach based on Rademacher complexity (Bartlett and Mendelson, 2002; Shalev-Shwartz and Ben-David, 2014). For every \( f \in \mathcal{F}_{m,B} \), define its risk and empirical risk
\[
R(f) := \mathbb{E}[\ell(y f(x))] \quad \text{and} \quad \hat{R}_N(f) := \frac{1}{N} \sum_{n=1}^{N} \ell(y_n f(x_n)),
\]

and assume the loss function satisfies \( 0 \leq \ell(y_n f(x_n)) \leq C_0 \) almost surely, for \( n = 1, \ldots, N \) and some constant \( C_0 < \infty \). Then, we have the following generalization bound for the minimizer \( \bar{f} \). With probability at least \( 1 - \delta \),
\[
R(\bar{f}) \leq \hat{R}_N(\bar{f}) + L \mathfrak{R}(\mathcal{F}_{m,B}) + C_0 \sqrt{\frac{\log(1/\delta)}{2N}},
\]

where we use the fact that the loss \( \ell \) is \( L \)-Lipschitz and \( \mathfrak{R}(\mathcal{F}_{m,B}) \) is the Rademacher complexity of the class \( \mathcal{F}_{m,B} \) defined via
\[
\mathfrak{R}(\mathcal{F}_{m,B}) := 2 \mathbb{E} \left[ \sup_{f \in \mathcal{F}_{m,B}} \frac{1}{N} \sum_{n=1}^{N} \sigma_n f(x_n) \right],
\]

13
where \( \{ \sigma_n \}_{n=1}^N \) are independent and identically distributed Rademacher random variables. In particular, if the expected loss is an upper bound on the probability of error (e.g., squared error or hinge loss), then we may use this to bound the probability of error \( \mathbb{P}(y \bar{f}(x) < 0) \leq R(\bar{f}) \). We bound the Rademacher complexity of \( F_{m,B} \) in the following theorem.

**Theorem 2.16** Assume that \( \|x_n\|_2 \leq C/2 \) almost surely for \( n = 1, \ldots, N \) and some constant \( C < \infty \). Then,

\[
\mathcal{R}(F_{m,B}) \leq \frac{2BC^{m-1}}{\sqrt{N(m-1)!}} + \mathcal{R}(c),
\]

where \( \mathcal{R}(c) \) denotes the Rademacher complexity of the polynomial terms \( c(x) \) that appear in the solutions to the optimization in (2.14).

**Remark 2.17** Theorem 2.16 shows that the Rademacher complexity, and hence the generalization error, is controlled by bounding the seminorm \( \|R_m f\|_{(m)} \leq B \). In practice, neural networks are typically implemented without the polynomial term \( c(x) \), in which case \( \mathcal{R}(c) \) can be ignored. It is also reasonable to expect that the Rademacher complexity of the non-polynomial portion of the network will dominate this term.

### 3. Preliminaries & Notation

#### 3.1 Spaces of functions and distributions

Let \( \mathcal{S}(\mathbb{R}^d) \) be the Schwartz space of smooth and rapidly decaying test functions on \( \mathbb{R}^d \). Its continuous dual, \( \mathcal{S}'(\mathbb{R}^d) \), is the space of tempered distributions on \( \mathbb{R}^d \). Since we are interested in the Radon domain, we are also interested in these spaces on \( S^{d-1} \times \mathbb{R} \), where \( S^{d-1} \) is the surface of the \( \ell^2 \)-unit sphere in \( \mathbb{R}^d \). We say \( \psi \in \mathcal{S}(S^{d-1} \times \mathbb{R}) \) when \( \psi \) is smooth and satisfies the decay condition (Stein and Shakarchi, 2011, Chapter 6)

\[
\sup_{\gamma \in S^{d-1}} \sup_{t \in \mathbb{R}} \left| \frac{1 + \|t\|^k}{(1 + \|t\|^k)^{\frac{d}{2}}} \frac{d^\ell}{dt^\ell}(D \psi)(\gamma, t) \right| < \infty
\]

for all integers \( k, \ell \geq 0 \) and for all differential operators \( D \) in \( \gamma \). Since the Schwartz spaces are nuclear, it follows that the above definition is equivalent to saying \( \mathcal{S}(S^{d-1} \times \mathbb{R}) = \mathcal{D}(S^{d-1}) \hat{\otimes} \mathcal{S}(\mathbb{R}) \), where \( \mathcal{D}(S^{d-1}) \) is the space of smooth functions on \( S^{d-1} \) and \( \hat{\otimes} \) is the topological tensor product (Wolff and Schaefer, 2012, Chapter III). We can then define the space of tempered distributions on \( S^{d-1} \times \mathbb{R} \) as its continuous dual, \( \mathcal{S}'(S^{d-1} \times \mathbb{R}) \).

We will later see that in order to define the Radon transform of distributions, we will be interested in the *Lizorkin test functions* \( \mathcal{S}_0(\mathbb{R}^d) \) of highly time-frequency localized functions over \( \mathbb{R}^d \) (Holschneider, 1995). This is a closed subspace of \( \mathcal{S}(\mathbb{R}^d) \) consisting of functions with all moments equal to 0. In other words, \( \varphi \in \mathcal{S}_0(\mathbb{R}^d) \) when \( \varphi \in \mathcal{S}(\mathbb{R}^d) \) and

\[
\int_{\mathbb{R}^d} x^\alpha \varphi(x) \, dx = 0
\]

for every multi-index \( \alpha \). We can then define the space of *Lizorkin distributions*, \( \mathcal{S}'_0(\mathbb{R}^d) \), the continuous dual of the Lizorkin test functions. The space of Lizorkin distributions can
be viewed as being topologically isomorphic to the quotient space of tempered distributions by the space of polynomials, i.e., if \( P(\mathbb{R}^d) \) is the space of polynomials on \( \mathbb{R}^d \), then \( \mathcal{S}'(\mathbb{R}^d) \cong \mathcal{S}(\mathbb{R}^d)/P(\mathbb{R}^d) \) (Yuan et al., 2010, Chapter 8). Just as above, we can define the Lizorkin test functions \( \mathbb{S}^{d-1} \times \mathbb{R} \) as \( \mathcal{S}_0(\mathbb{S}^{d-1} \times \mathbb{R}) = \mathcal{D}(\mathbb{S}^{d-1}) \otimes \mathcal{S}_0(\mathbb{R}) \) and the space of Lizorkin distributions on \( \mathbb{S}^{d-1} \times \mathbb{R} \) as its continuous dual, \( \mathcal{S}'_0(\mathbb{S}^{d-1} \times \mathbb{R}) \).

The Riesz–Markov–Kakutani representation theorem says that \( \mathcal{M}(\mathbb{R}^d) \), the space of finite Radon measures on \( \mathbb{R}^d \), is the continuous dual of \( C_0(\mathbb{R}^d) \), the space of continuous functions vanishing at infinity (Folland, 1999, Chapter 7). Since \( C_0(\mathbb{R}^d) \) is a Banach space when equipped with the uniform norm, we have the dual norm

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

where the last equality holds since \( \mathcal{S}(\mathbb{R}^d) \) is dense in \( C_0(\mathbb{R}^d) \), and the duality pairing \( \langle \cdot, \cdot \rangle \) can be viewed, formally, as the integral

\[
\langle u, \varphi \rangle = \int_{\mathbb{R}^d} \varphi(x) u(x) \, dx.
\]

The norm \( \|\cdot\|_{\mathcal{M}(\mathbb{R}^d)} \) is exactly the total variation norm (in the sense of measures). We can always view \( \mathcal{M}(\mathbb{R}^d) \subset \mathcal{S}'(\mathbb{R}^d) \), providing the description

\[
\mathcal{M}(\mathbb{R}^d) := \left\{ u \in \mathcal{S}'(\mathbb{R}^d) : \|u\|_{\mathcal{M}(\mathbb{R}^d)} < \infty \right\}.
\]

In other words, \( \mathcal{M}(\mathbb{R}^d) \) is the set of all tempered distributions with finite total variation. We will also be interested in the Lizorkin distributions with finite total variation. Since polynomials do not decay and hence do not have finite \( \|\cdot\|_{\mathcal{M}(\mathbb{R}^d)} \)-norm, it follows that the subspace of Lizorkin distributions on \( \mathbb{R}^d \) with finite total variation is the entire space \( \mathcal{M}(\mathbb{R}^d) \).

We can analogously define \( \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R}) \) as the continuous dual of \( C_0(\mathbb{S}^{d-1} \times \mathbb{R}) \), the space of continuous functions vanishing at infinity (in the variable on \( \mathbb{R} \)) with the norm

\[
\|u\|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})} := \sup_{\psi \in C_0(\mathbb{S}^{d-1} \times \mathbb{R})} \langle u, \psi \rangle = \sup_{\psi \in \mathcal{S}(\mathbb{S}^{d-1} \times \mathbb{R})} \langle u, \psi \rangle, \quad \|\psi\|_{L^\infty(\mathbb{S}^{d-1} \times \mathbb{R})} = 1
\]

and the description

\[
\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R}) := \left\{ u \in \mathcal{S}'(\mathbb{S}^{d-1} \times \mathbb{R}) : \|u\|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})} < \infty \right\}.
\]

As before, the subspace of Lizorkin distributions on \( \mathbb{S}^{d-1} \times \mathbb{R} \) with finite total variation is the entire space \( \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R}) \).

**Remark 3.1** We are interested in the spaces \( \mathcal{M}(X) \supset L^1(X) \) as they are spaces slightly larger than \( L^1(X) \) that also include tempered distributions. We are specifically interested in the fact that \( \mathcal{M}(X) \) contains the evaluation functionals, i.e., the translated Dirac impulses \( \delta(\cdot-x_0) \), \( x_0 \in X \), which satisfy \( \delta(\cdot-x_0) \in \mathcal{M}(X) \) with \( \|\delta(\cdot-x_0)\|_{\mathcal{M}(X)} = 1 \), but \( \delta(\cdot-x_0) \not\in L^1(X) \). Additionally, every \( f \in L^1(X) \) satisfies \( \|f\|_{L^1(X)} = \|f\|_{\mathcal{M}(X)} \).

---

7. The argument is essentially that \( \mathcal{M}(\mathbb{R}^d) \cap P(\mathbb{R}^d) = \{0\} \) and so \( \mathcal{M}(\mathbb{R}^d)/P(\mathbb{R}^d) \) is really \( \mathcal{M}(\mathbb{R}^d)/\{0\} \cong \mathcal{M}(\mathbb{R}^d) \).
3.2 The Fourier transform

The Fourier transform $\mathcal{F}$ of $f : \mathbb{R}^d \to \mathbb{C}$ and inverse Fourier transform $\mathcal{F}^{-1}$ of $F : \mathbb{R}^d \to \mathbb{C}$ are given by

$$
\mathcal{F}\{f\}(\xi) := \int_{\mathbb{R}^d} f(x) e^{-ix^T \xi} \, dx, \quad \xi \in \mathbb{R}^d
$$

$$
\mathcal{F}^{-1}\{F\}(x) := \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{ix^T \xi} F(\xi) \, d\xi, \quad x \in \mathbb{R}^d,
$$

where $i^2 = -1$. We will usually write $\hat{\cdot}$ for $\mathcal{F}\{\cdot\}$.

3.3 The Hilbert transform

The Hilbert transform $\mathcal{H}$ of $f : \mathbb{R} \to \mathbb{C}$ is given by

$$
\mathcal{H}\{f\}(x) := \frac{i}{\pi} \mathrm{p.v.} \int_{-\infty}^{\infty} \frac{f(y)}{x-y} \, dy, \quad x \in \mathbb{R},
$$

where p.v. denotes understanding the integral in the Cauchy principal value sense. The prefactor was chosen so that

$$
\widehat{\mathcal{H}\{f\}}(\xi) = (\text{sgn} \xi) \widehat{f}(\xi) \quad \text{and} \quad \mathcal{H}\mathcal{H} f = f.
$$

3.4 The Radon transform

The Radon transform $\mathcal{R}$ of $f : \mathbb{R}^d \to \mathbb{R}$ and the dual Radon transform $\mathcal{R}^*$ of $\Phi : \mathbb{S}^{d-1} \times \mathbb{R} \to \mathbb{R}$ are given by

$$
\mathcal{R}\{f\}(\gamma,t) := \int_{\{x : \gamma^T x = t\}} f(x) \, ds(x), \quad \gamma \in \mathbb{S}^{d-1}, t \in \mathbb{R} \tag{3.2}
$$

$$
\mathcal{R}^*\{\Phi\}(x) := \int_{\mathbb{S}^{d-1}} \Phi(\gamma, \gamma^T x) \, d\sigma(\gamma), \quad x \in \mathbb{R}^d,
$$

where $s$ denotes the surface measure on the plane $\{x : \gamma^T x = t\}$, and $\sigma$ denotes the surface measure on $\mathbb{S}^{d-1}$. We will sometimes write $z = (\gamma, t)$ as the variable in the Radon domain. Note that the Radon transform a function is always even, i.e., $\Phi(\gamma, t) = \Phi(-\gamma, -t)$. This is why, as discussed in Section 2.1, the Radon domain is in fact $\mathbb{P}^d$. We will make this more explicit in Remark 3.4.

**Remark 3.2** Another way to view the Radon transform is to consider, formally, integral

$$
\mathcal{R}\{f\}(\gamma,t) = \int_{\mathbb{R}^d} f(x) \delta_R(\gamma^T x - t) \, dx, \quad \gamma \in \mathbb{S}^{d-1}, t \in \mathbb{R}. \tag{3.3}
$$
Similarly, another way to view the dual Radon transform is to consider, formally, the integral

$$R^* \{ \Phi \}(x) = \int_{S^{d-1} \times \mathbb{R}} \delta_{\mathbb{R}}(\gamma^T x - t) \Phi(\gamma, t) \, d(\sigma \times \lambda)(\gamma, t), \quad x \in \mathbb{R}^d, \quad (3.4)$$

where $\lambda$ denotes the univariate Lebesgue measure.

The fundamental theorem of the Radon transform is the Radon inversion formula, which states for any $f \in \mathcal{S}(\mathbb{R}^d)$

$$2(2\pi)^{d-1} f = R^* \Lambda^{d-1} R f, \quad (3.5)$$

where the ramp filter $\Lambda^d$ of a function $\Phi(\gamma, t)$ is given by

$$\Lambda^d \{ \Phi \} (\gamma, t) := \begin{cases} \partial_t^d \Phi(\gamma, t), & d \text{ even} \\ \mathcal{H}_t \partial_t^d \Phi(\gamma, t), & d \text{ odd} \end{cases}, \quad (3.6)$$

where $\mathcal{H}_t$ is the Hilbert transform (in the variable $t$) and $\partial_t$ is short-hand for $\partial/\partial t$. It’s easier to see that $\Lambda^d$ is indeed a ramp filter by looking at its frequency response with respect to the $t$ variable. We have

$$\hat{\Lambda}^d \Phi(\gamma, \xi) = i^d |\xi|^d \hat{\Phi}(\gamma, \xi).$$

We will be interested in Radon transforms of distributions. As usual, this proceeds via duality, though some care has to be taken. It is easy to verify that if $\varphi \in \mathcal{S}(\mathbb{R}^d)$, then $R \{ \varphi \} \in \mathcal{S}(S^{d-1} \times \mathbb{R})$. This is not true about the dual transform. Indeed, if $\psi \in \mathcal{S}(S^{d-1} \times \mathbb{R})$, then it may not be true that $R^* \{ \psi \} \in \mathcal{S}(\mathbb{R}^d)$. Due to recent developments in ridgelet analysis (Kostadinova et al., 2014), specifically regarding the continuity of the Radon transform of Lizorkin test functions, we have the following result.

**Theorem 3.3** ((Helgason, 2014, Cor. 2.5), (Kostadinova et al., 2014, Cor. 6.1))

The transforms

$$R : \mathcal{S}_0(\mathbb{R}^d) \rightarrow \mathcal{S}_0(\mathbb{P}^d)$$

$$R^* : \mathcal{S}_0(\mathbb{P}^d) \rightarrow \mathcal{S}_0(\mathbb{R}^d)$$

are continuous bijections, where $\mathcal{S}_0(\mathbb{P}^d)$ is the subspace of even functions in $\mathcal{S}_0(S^{d-1} \times \mathbb{R})$.

**Remark 3.4** Technically, $\mathbb{P}^d$ is the space of all hyperplanes in $\mathbb{R}^d$. Each hyperplane $h \in \mathbb{P}^d$ can be written as $h = \{ x \in \mathbb{R}^d : \gamma^T x = t \}$. Notice that the pairs $(\gamma, t)$ and $(-\gamma, -t)$ give the same hyperplane, i.e., the map $(\gamma, t) \mapsto h$ is a double covering of $S^{d-1} \times \mathbb{R}$ onto $\mathbb{P}^d$. We can thus identify functions on $\mathbb{P}^d$ with functions $\psi$ on $S^{d-1} \times \mathbb{R}$ satisfying the symmetry condition $\psi(\gamma, t) = \psi(-\gamma, -t)$. This is why the space $\mathcal{S}_0(\mathbb{P}^d)$ is the subspace of even functions in $\mathcal{S}_0(S^{d-1} \times \mathbb{R})$.

**Remark 3.5** We also use the domain $\mathbb{P}^d$ for the previously defined spaces, e.g., $C_0(\mathbb{P}^d)$ is the subspace of even functions in $C_0(S^{d-1} \times \mathbb{R})$, and its continuous dual is $\mathcal{M}(\mathbb{P}^d)$, the subspace of even finite Radon measures in $\mathcal{M}(S^{d-1} \times \mathbb{R})$. The norms on these subspaces are defined analogously to the full spaces.
Theorem 3.3 allows us to define the Radon transform and dual Radon transform of distributions by duality by restricting ourselves to the Lizorkin test functions and distributions, i.e., the action of the Radon transform of \( f \in \mathcal{S}'(\mathbb{R}^d) \) on \( \psi \in \mathcal{S}_0^d(\mathbb{R}) \) is defined to be \( \langle \mathcal{R} f, \psi \rangle := \langle f, \mathcal{R}^* \psi \rangle \), and the action of the dual Radon transform of \( \Phi \in \mathcal{S}_0^d(\mathbb{P}^d) \) on \( \varphi \in \mathcal{S}_0^d(\mathbb{R}^d) \) is defined to be \( \langle \mathcal{R}^* \Phi, \varphi \rangle := \langle \Phi, \mathcal{R} \varphi \rangle \). This means we have the following corollary to Theorem 3.3

**Corollary 3.6** The transforms

\[
\mathcal{R} : \mathcal{S}'_0(\mathbb{R}^d) \rightarrow \mathcal{S}_0^d(\mathbb{P}^d)
\]

\[
\mathcal{R}^* : \mathcal{S}'_0(\mathbb{P}^d) \rightarrow \mathcal{S}_0^d(\mathbb{R}^d)
\]

are continuous bijections.

**Remark 3.7** The reason we need to consider \( \mathcal{S}_0^d(\mathbb{P}^d) \) and \( \mathcal{S}_0^d(\mathbb{P}^d) \) instead of \( \mathcal{S}_0^d(\mathbb{S}^{d-1} \times \mathbb{R}) \) and \( \mathcal{S}_0^d(\mathbb{S}^{d-1} \times \mathbb{R}) \) is because the dual Radon transform \( \mathcal{R}^* \) annihilates odd functions. Another way to see this is from (3.2), the definition of the Radon transform, which says, in particular, that the Radon transform of a function is necessarily even. With this in hand, we see that the Radon domain is actually \( \mathbb{P}^d \) and not \( \mathbb{S}^{d-1} \times \mathbb{R} \). We will abuse notation and write \( (\gamma, t) \in \mathbb{P}^d \) when we mean the equivalence class \([\gamma, t]\) defined by the equivalence relation \( (\gamma, t) \sim (\tilde{\gamma}, \tilde{t}) \) if and only if \( \gamma = \sigma \tilde{\gamma} \) and \( t = \sigma \tilde{t} \) where \( \sigma \in \{-1, +1\} \) and \( (\gamma, t) \in \mathbb{S}^{d-1} \times \mathbb{R} \).

**Remark 3.8** We will be interested in the translated Dirac impulses in \( \mathcal{M}(\mathbb{P}^d) \). It is important to note that these are not the same as the translated Dirac impulses in \( \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R}) \). To understand this, we recall that the Dirac impulse is defined to be the evaluation function. In particular, if \( \delta_{\mathbb{S}^{d-1} \times \mathbb{R}}(\cdot - z_0) \) is a Dirac impulse in \( \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R}) \) translated to \( z_0 \in \mathbb{S}^{d-1} \times \mathbb{R} \), then, it is easy to verify that the Dirac impulse in \( \mathcal{M}(\mathbb{P}^d) \) translated to \( z_0 \in \mathbb{P}^d \) must be

\[
\delta_{\mathbb{P}^d}(\cdot - z_0) = \frac{\delta_{\mathbb{S}^{d-1} \times \mathbb{R}}(\cdot - z_0) + \delta_{\mathbb{S}^{d-1} \times \mathbb{R}}(\cdot + z_0)}{2}.
\]

We also have the following inversion formula for the dual Radon transform (Helgason, 2014, Theorem 3.7). For any \( \Phi \in \mathcal{S}_0^d(\mathbb{P}^d) \)

\[
2(2\pi)^{d-1} \Phi = \Lambda^{d-1} \mathcal{R} \mathcal{R}^* \Phi. \tag{3.7}
\]

**Remark 3.9** The inversion formulas in (3.5) and (3.7) can be rewritten in many ways using the intertwining relations of the Radon transform and its dual with the Laplacian operator (Helgason, 2014, Lemma 2.1). We have

\[
(-\Delta)^{\frac{d-1}{2}} \mathcal{R}^* = \mathcal{R}^* \Lambda^{d-1} \quad \text{and} \quad \mathcal{R}(-\Delta)^{\frac{d-1}{2}} = \Lambda^{d-1} \mathcal{R}. \tag{3.8}
\]

**Remark 3.10** Since the constant \( 2(2\pi)^{d-1} \) arises often when working with the Radon transform, put

\[
c_d := \frac{1}{2(2\pi)^{d-1}}. \tag{3.9}
\]
Remark 3.11 Here and in the rest of the paper, we use the pairing $\langle \cdot, \cdot \rangle$ to generically denote the duality pairing between a space and its continuous dual. We will not use a different notation for different pairings to reduce clutter and the exact pairings should be clear from context.

4. The Representer Theorem

In this section we will prove Theorem 2.1, our representer theorem. The general strategy will be to reduce the problem in (2.5) to one that is similar to the classical problem of Radon measure recovery, which has been studied since as early as the 1930s (Beurling, 1938; Krein, 1938). Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. The prototypical Radon measure recovery problem studies optimizations of the form

$$
\min_{u \in \mathcal{M}(\Omega)} \|u\|_{\mathcal{M}(\Omega)} \quad \text{s.t.} \quad Vu = y, \quad (4.1)
$$

where $V : \mathcal{M}(\Omega) \rightarrow \mathbb{R}^N$ is a linear continuous measurement operator and $y \in \mathbb{R}^N$. The first “representer theorem” for (4.1) is from Zuhovickiĭ (1948). This representer theorem essentially states that there exists a sparse solution to (4.1) of the form

$$
\sum_{k=1}^{K} v_k \delta_{\mathbb{R}^d}(\cdot - z_k),
$$

with $K \leq N$. Refinements to this result have been made over the years, e.g., (Fisher and Jerome, 1975, Theorem 1), including very modern results, e.g., (Unser et al., 2017, Theorem 7), (Boyer et al., 2019, Section 4.2.3), (Bredies and Carioni, 2020, Theorem 4.2). For our problem, we reduce (2.5) to the following Radon measure recovery problem.

Lemma 4.1 Assume the following:

1. The data fitting term $G : \mathbb{R}^N \rightarrow \mathbb{R}$ is a proper convex function that is coercive, and lower semi-continuous with respect to the topology of $\mathbb{R}^N$, which is the standard topology on finite-dimensional spaces.

2. The measurement operator $V : \mathcal{M}(\mathbb{P}^d) \rightarrow \mathbb{R}^N$ is a continuous linear operator.

Then, there exists a sparse minimizer to the generalized Radon measure recovery problem

$$
\min_{u \in \mathcal{M}(\mathbb{P}^d)} G(Vu) + \|u\|_{\mathcal{M}(\mathbb{P}^d)} \quad (4.2)
$$

of the form

$$
\bar{u} = \sum_{k=1}^{K} v_k \delta_{\mathbb{P}^d}(\cdot - z_k),
$$

with $K \leq N$, $v_k \in \mathbb{R}$ and $z_k = (w_k, b_k) \in \mathbb{P}^d$, $k = 1, \ldots, K$. 

19
Lemma 4.1 is technically a new result given that its a generalized Radon measure recovery problem posed in the Radon domain, though the proof follows as an extremely simply corollary to recent developments about the solutions to variational inverse problems with finite-dimensional data (Bredies and Carioni, 2020). We include a proof of Lemma 4.1 in Appendix A for completeness.

In order to reduce (2.5) to (4.2), we need to understand which functions are translated Green’s functions of $\mathbb{R}_m$. As claimed in Remark 2.10, these are exactly the atoms of a single-hidden layer neural network as in (2.6).

**Lemma 4.2** The translated Green’s functions of $\mathbb{R}_m = c_d \lambda^{d-1} \mathcal{R} \Delta^{m/2}$ are

$$r^{(m)}_{(w,b)}(x) := \rho_m(w^\top x - b),$$

where $(w, b) \in \mathbb{P}^d$. In other words,

$$c_d \lambda^{d-1} \mathcal{R} \Delta^{m/2} r^{(m)}_{(w,b)} = \delta_{\mathbb{P}^d} (\cdot - (w, b)).$$

**Proof** We need to check that

$$c_d \left< \lambda^{d-1} \mathcal{R} \Delta^{m/2} r^{(m)}_{(w,b)}, \psi \right> = \psi(w, b)$$

for all even Lizorkin test functions $\psi \in \mathcal{D}(\mathbb{P}^d)$. Recall that $m \in 2\mathbb{Z}$ and $m \geq 2$. Next, notice by direct calculations that when $m \geq 2$, $\Delta r^{(m)}_{(w,b)} = \|w\|_2^2 r^{(m-2)}_{(w,b)} = r^{(m-2)}_{(w,b)}$, and when $m = 2$, $\Delta \left< r^{(2)}_{(w,b)} \right>(x) = \|w\|_2^2 \delta_{\mathbb{R}}(w^\top x - b) = \delta_{\mathbb{R}}(w^\top x - b).

Hence, we have that $\Delta^{m/2} \left< r^{(m)}_{(w,b)} \right>(x) = \delta_{\mathbb{R}}(w^\top x - b)$. Next,

$$c_d \left< \lambda^{d-1} \mathcal{R} \Delta^{m/2} r^{(m)}_{(w,b)}, \psi \right> = c_d \left< \Delta^{m/2} r^{(m)}_{(w,b)}, \mathcal{R}^* \lambda^{d-1} \psi \right> = c_d \left< \delta_{\mathbb{R}}(w^\top (\cdot) - b), \mathcal{R}^* \lambda^{d-1} \psi \right> = c_d \left[ \mathcal{R} \mathcal{R}^* \lambda^{d-1} \psi \right](w, b) \overset{(\S)}{=} \psi(w, b),$$

where $(\S)$ holds by (3.3) and $(\overset{(*)}{\S})$ holds by the dual Radon transform inversion formula (3.7) combined with the intertwining relations (3.8).

Since the translated Green’s functions of $\mathbb{R}_m$ are $r^{(m)}_{(w,b)}$ as defined in (4.3), we choose to define the growth restriction previously discussed in Section 2.1 for the null space and native space of $\mathbb{R}_m$ via the algebraic growth rate

$$n_0 := \inf \left\{ n \in \mathbb{N} : r^{(m)}_{(w,b)} \in L^{\infty,n}(\mathbb{R}^d) \right\} = m - 1.$$  

(4.4)

Before we can reduce (2.5) to (4.2), we must establish that minimizers to (2.5) in Theorem 2.1 even exist. As is usual for variational problems, to show that minimizers to exist, we require the space that is being optimized over has some “nice” topological properties. In (2.5), we are optimizing over the native space $\mathcal{F}_m$. Although $\mathcal{F}_m$ is defined by the seminorm $\|R_m f\|_{\mathcal{M}(\mathbb{P}^d)}$, we show in Theorem 4.9 if we equip $\mathcal{F}_m$ with the proper direct-sum topology, it forms a bona fide Banach space. In order to prove that $\mathcal{F}_m$ is a Banach space, we require three intermediary results.
Lemma 4.3 For each function \( f \in \mathcal{F}_m \), there exists \( \mu \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R}) \) such that
\[
f(x) = \int_{\mathbb{S}^{d-1} \times \mathbb{R}} r^{(m)}_{(w,b)}(x) \, d\mu(w,b) + c(x),
\]
where \( c(\cdot) \) is a polynomial of degree strictly less than \( m \). In particular, \( \mu \) is unique and an even measure, i.e., \( \mu \in \mathcal{M}(\mathbb{P}^d) \).

Remark 4.4 Since \( \mu \) in Lemma 4.3 depends on \( f \), it follows that the integral representation in (4.5) is a kind of Calderón-type reproducing formula (Calderón, 1964). The integral representation in (4.5) can be viewed as an infinite-width (continuum-width) neural network as has been studied in several recent works (Ongie et al., 2020; Mhaskar, 2020; Bach, 2017). We also remark that (4.5) shares many similarities with the dual Ridgelet transform (Murata, 1996; Rubin, 1998; Candès, 1998, 1999).

Remark 4.5 Since \( \mu \) in Lemma 4.3 is even, we see that the lemma statement could instead be phrased that \( f \) can be written as an integral of functions of the form
\[
r^{(m)}_{(w,b)}(x) + r^{(m)}_{(-w,-b)}(x)
\]
where the integration is against a generic \( \mu \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R}) \). When \( m \) is even, this corresponds to functions of the form \( \tau_m(w^T x - b) \) where \( \tau_m = |\cdot|^{m-1}/2!(m-1)! \). When \( m \) is odd, \( \tau_m(w^T x - b) \) is a polynomial of degree strictly less than \( m \) in \( x \), and so it gets annihilated by \( R_m \). Thus, we see the symmetry of \( \mathbb{P}^d \), the Radon domain, enforces that \( m \) be even. Another way to see this is to notice that when \( m \) is even, both \( \rho_m \) and \( \tau_m \) are Green’s functions of the univariate \( m \)th derivative operator, while when \( m \) is odd, only \( \rho_m \) is a Green’s function of the univariate \( m \)th derivative operator.

Lemma 4.6 The null space \( N_m \) of \( R_m \) defined in (2.3) is finite-dimensional. In particular, it is a subset of polynomials of degree strictly less than \( m \).

Lemma 4.6 says, in particular, that we can find a biorthogonal system for \( N_m \).

Definition 4.7 Consider a finite-dimensional space \( \mathcal{N} \) with \( N_0 := \dim \mathcal{N} \). The pair \( (\phi, p) = \{(\phi_n, p_n)\}_{n=1}^{N_0} \) is called a biorthogonal system for \( \mathcal{N} \) if \( \{p_n\}_{n=1}^{N_0} \) is a basis of \( \mathcal{N} \) and the vector of “boundary” functionals \( \phi = (\phi_1, \ldots, \phi_{N_0}) \) with \( \phi_n \in \mathcal{N}' \) (the continuous dual of \( \mathcal{N} \)) satisfy the biorthogonality condition \( \langle \phi_k, p_n \rangle = \delta[k-n], k,n = 1, \ldots, N_0 \), where \( \delta[\cdot] \) is the Kronecker impulse.

Put \( N_0 := \dim N_m \) and let \( (\phi, p) \) be a biorthogonal system for \( N_m \). Definition 4.7 says, in particular, that any \( q \in N_m \) has the unique representation
\[
q = \sum_{n=1}^{N_0} \langle \phi_n, q \rangle p_n.
\]
We will sometimes write \( \phi(f) \) to denote the vector \( (\langle \phi_1, f \rangle, \ldots, \langle \phi_{N_0}, f \rangle) \).
Lemma 4.8 Let \((\phi, p)\) be a biorthogonal system for \(N_m \subset F_m \subset L^\infty,m-1(R^d)\). Then, there exists a unique operator

\[
R_{m,\phi}^{-1} : \psi \mapsto R_{m,\phi}^{-1} \psi = \int_{S^{d-1} \times R} g_{\phi}(\cdot, z) \psi(z) \, d(\sigma \times \lambda)(z),
\]

where we recall that \(\sigma\) is the surface measure on \(S^{d-1}\) and \(\lambda\) is the univariate Lebesgue measure. The operator \(R_{m,\phi}^{-1}\) satisfies

\[
R_m R_{m,\phi}^{-1} \psi = \psi \quad \text{(right-inverse property)}
\]

\[
\phi(R_{m,\phi}^{-1} \psi) = 0 \quad \text{(boundary conditions)}
\]

(4.6)

for all \(\psi \in S_0(\mathcal{P}_d)\). The kernel of this operator is

\[
g_{\phi}(x, z) = r^{(m)}_{\phi}(x) - \sum_{n=1}^{N_0} p_n(x) q_n(z),
\]

where \(r^{(m)}_{\phi}\) is defined as in (4.3) and \(q_n(z) := \langle \phi_n, r_{\phi}\rangle\). Moreover, \(R_{m,\phi}^{-1}\) admits a continuous extension \(M(\mathcal{P}_d) \to L^\infty,m-1(R^d)\) with (4.6) holding for all \(\psi \in M(\mathcal{P}_d)\).

With these three results, we can now establish the Banach space structure of \(F_m\).

Theorem 4.9 Let \((\phi, p)\) be a biorthogonal system for the null space \(N_m\) of \(R_m\) as defined in (2.3) and let \(F_m\) be the native space of \(R_m\) as defined in (2.4). Then, the following hold:

1. Every \(f \in F_m\) admits a unique representation

\[
f = R_{m,\phi}^{-1} u + q,
\]

(4.7)

where \(u = R_m f \in M(\mathcal{P}_d)\) and \(q = \sum_{n=1}^{N_0} \langle \phi_n, f \rangle p_n \in N_m\). In particular, this specifies the structural property \(F_m \cong F_{m,\phi} \oplus N_m\), where

\[
F_{m,\phi} := \{ f \in F : \phi(f) = 0 \}.
\]

2. \(F_m\) is a Banach space when equipped with the norm

\[
\|f\|_{F_m} := \|R_m f\|_{M(\mathcal{P}_d)} + \|\phi(f)\|_2.
\]

The proofs of Lemmas 4.3, 4.6, and 4.8 and Theorem 4.9 appear in Appendix A.

Remark 4.10 Item 1. in Theorem 4.9 establishes existence and uniqueness of distributional solutions to a kind of polyharmonic PDE where the boundary conditions are governed by \(\phi\). Specifically,

\[
\begin{cases}
R_m f = u \\
\phi(f) = (c_1, \ldots, c_{N_0})
\end{cases}
= \begin{cases}
\Delta^{m/2} f = \mathcal{F}^\ast \{u\} \\
\phi(f) = (c_1, \ldots, c_{N_0})
\end{cases}.
\]

This is vital in proving that single-hidden layer neural networks are solutions to (2.5) in Theorem 2.1. Item 2. in Theorem 4.9 is vital in proving existence of minimizers to (2.5) in Theorem 2.1.
We will now prove Theorem 2.1. Before doing so, we state the following recent result from the variational problems literature in order to establish existence of solutions to (2.5) in Theorem 2.1.

**Theorem 4.11 (Bredies and Carioni (2020, Theorem 3.3))** Assume the following:

1. $\mathcal{F}$ is a locally convex topological vector space.
2. $H$ is a finite-dimensional Hilbert space.
3. $G : H \to \mathbb{R}$ is a proper convex function that is coercive, and lower semi-continuous with respect to the topology of $H$, which is the standard topology on finite-dimensional spaces.
4. $\mathcal{V} : \mathcal{F} \to H$ is a continuous linear operator.
5. $\|\cdot\|$ is a seminorm.

Then, there exists a sparse minimizer to the variational problem

$$\inf_{f \in \mathcal{F}} G(\mathcal{V} f) + \|f\|$$

of the form

$$\tilde{f}(x) = \sum_{k=1}^{K} v_k f_k(x) + c(x),$$

where $K \leq N$, $f_k + N \in \text{Ext}(B + N)$, and $c \in N$. Here, $N$ is the null space of the seminorm $\|\cdot\|$, $B$ is the closed unit ball with respect to the seminorm $\|\cdot\|$, and $\text{Ext}(\cdot)$ denotes the extreme points its input set.

**Remark 4.12** It is often difficult to find the extreme points of $B + N$. Thus, we simply use the existence result of Theorem 4.11 and manually construct a sparse minimizer in the proof of Theorem 2.1.

### 4.1 Proof of Theorem 2.1

**Proof** We first remark that a solution to (2.5) exists by Theorem 4.11. Next, the general strategy is to recast the problem in (2.5) as one with interpolation constraints. To do this use a technique from Unser (2019b). We use the fact that $G$ is a strictly convex function. In particular, this says for any two solutions $\tilde{f}, \bar{f}$ of (2.5), it must be the case that $\mathcal{V}\tilde{f} = \mathcal{V}\bar{f}$ (since otherwise, it would contradict the strict convexity of $G$). Hence, there exists $z \in \mathbb{R}^N$ such that $z = \mathcal{V}\tilde{f} = \mathcal{V}\bar{f}$. Although $z \in \mathbb{R}^N$ is not usually known before hand, this property provides us with the parametric characterization of the solution set to (2.5) as

$$S_z := \arg \min_{f \in \mathcal{F}} \|R_m f\|_{\mathcal{M}(\mathbb{P}d)} \quad \text{s.t.} \quad \mathcal{V}f = z \quad (4.8)$$

for some $z \in \mathbb{R}^N$. Hence, it suffices to show that there exists a solution to (4.8) of the form in (2.6). We will now show this for a fixed $z \in \mathbb{R}^N$. 

23
Consider the $N \times N_0$ matrix

$$A := [V_{p_1} \cdots V_{p_{N_0}}],$$

where $\{p_n\}_{n=1}^{N_0}$ is a basis for $N_m$. Every $q \in N_m$ has an expansion $\sum_{n=1}^{N_0} c_n p_n$. The condition in Item 3. in Theorem 2.1 is equivalent to saying $Ac = \mathcal{V}q$, where $c = (c_1, \ldots, c_{N_0})$. Another consequence of Item 3. in Theorem 2.1, is that this system is overdetermined, i.e., $N > N_0$. Hence, it is solvable if and only if $A^T A$ is invertible and the solution is given by

$$c = (A^T A)^{-1} A^T (\mathcal{V}q).$$

The invertibility of $A^T A$ says $\lambda_{\min}(A^T A) > 0$, where $\lambda_{\min}(\cdot)$ denotes the minimal eigenvalue. This is equivalent to saying $\sigma_{\min}^2(A) > 0$, where $\sigma_{\min}(\cdot)$ denotes the minimal singular value. This says, in particular, that span $\{a_n\}_{n=1}^{N_0} = \mathbb{R}^{N_0}$, where $a_n^T$ is the $n$th row of $A$. Thus, there exists a subset of $N_0$ rows of $A$ that span $\mathbb{R}^{N_0}$. Without loss of generality, suppose this subset is $\{a_n\}_{n=1}^{N_0}$. This says the submatrix $A_0$ of $A$ defined by

$$A_0 := \begin{bmatrix} a_1^T \\ \vdots \\ a_{N_0}^T \end{bmatrix}$$

is invertible. Consider the components $(\nu_1, \ldots, \nu_N)$ of $\mathcal{V}$ via $\mathcal{V}: f \mapsto (\langle \nu_1, f \rangle, \ldots, \langle \nu_N, f \rangle)$ Then, we can write $a_n$ as the vector $(\langle \nu_n, p_1 \rangle, \ldots, \langle \nu_n, p_{N_0} \rangle)$, $n = 1, \ldots, N_0$. Hence the reduced subset of measurements $(\nu_1, \ldots, \nu_{N_0})$ are linearly independent with respect to $N_m$. Let $\mathcal{V}_0$ denote this reduced set of measurements and let $\mathcal{V}_1$ denote the remaining set of measurements, i.e., $\mathcal{V} = (\mathcal{V}_0, \mathcal{V}_1)$.

Next, notice that $\mathcal{V}_0 p_n$ is the $n$th column of $A_0$. Let $e_n$ denote the $n$th canonical basis vector. Then, we have the equality $\mathcal{V}_0 p_n = A_0 e_n$. Using the invertibility of $A_0$, this says

$$A_0^{-1} (\mathcal{V}_0 p_n) = e_n.$$ 

If we put $\phi_0 := A_0^{-1} \circ \mathcal{V}_0$, the above display is exactly the biorthogonality property and hence $(\phi_0, \mathcal{P})$ form a biorthogonal system for $N_m$.

By Theorem 4.9, every $f \in \mathcal{F}_m$ admits a unique representation $f = R_{m,\phi_0}^{-1} u + q$, where $u \in \mathcal{M}(\mathbb{P}^d)$ and $q \in N_m$. This says $R_m f = u$. Hence, we can rewrite the problem in (4.8) as

$$\min_{u \in \mathcal{M}(\mathbb{P}^d)} \|u\|_{\mathcal{M}(\mathbb{P}^d)}$$

subject to

$$\mathcal{V}f = z,$$

$$f = R_{m,\phi_0}^{-1} u + q,$$

$$q \in N_m.$$ 

(4.9)

Write $z_0 = (z_1, \ldots, z_{N_0})$ and $z_1 = (z_{N_0+1}, \ldots, z_N)$. Then, the constraint $\mathcal{V}f = z$ can be written as the two constraints $\mathcal{V}_0 f = z_0$ and $\mathcal{V}_1 f = z_1$. By the boundary conditions in (4.6) we have that $\phi_0(f) = \phi_0(R_{m,\phi_0}^{-1} u + q) = \phi_0(R_{m,\phi_0}^{-1} u) + \phi_0(q) = \phi_0(q)$. Thus, by definition
of \( \phi_0 \) we have that \( z_0 = \mathcal{V}_0 f = \mathcal{V}_0 q \). Since \((\phi_0, p)\) is a biorthogonal system for \( \mathcal{N}_m \), there exists a unique \( q := q_0 \) such that \( \mathcal{V}_0 q_0 = z_0 \). Hence, (4.9) can be rewritten as

\[
\begin{align*}
\min_{u \in \mathcal{M}(\mathbb{P}^d)} \|u\|_{\mathcal{M}(\mathbb{P}^d)} &= \min_{u \in \mathcal{M}(\mathbb{P}^d)} \|u\|_{\mathcal{M}(\mathbb{P}^d)} \\
\text{s.t. } \mathcal{V}_1 f &= z_1, \\
f &= \mathcal{R}_{m, \phi}^{-1} u + q_0
\end{align*}
\]

This says there exists a solution to the above display of the form \( \bar{f} = \mathcal{R}_{m, \phi}^{-1} \bar{u} + q_0 \), where \( \bar{u} \in \arg \min_{u \in \mathcal{M}(\mathbb{P}^d)} \|u\|_{\mathcal{M}(\mathbb{P}^d)} \) s.t. \( \mathcal{V}_1 (\mathcal{R}_{m, \phi}^{-1} u) = z_1 - \mathcal{V}_1 q_0 \).

By Lemma 4.1, there exists a sparse minimizer to the above display that is a linear combination of at most \( N - N_0 \) Dirac impulses in \( \mathcal{M}(\mathbb{P}^d) \). By Lemma 4.8, the solution \( \bar{f} = \mathcal{R}_{m, \phi}^{-1} \bar{u} + q_0 \) takes the form in (2.6), so we’re done.

5. Ridge Splines and Classical Spline Theory

5.1 Univariate ridge splines are univariate polynomial splines

Univariate ridge splines and classical univariate splines are in fact the same object. As mentioned in Remark 2.11, when \( d = 1 \), our notion of a polynomial ridge spline of order \( m \) exactly coincides with the notion of a univariate polynomial spline of order \( m \). To see this, by Remark 2.7, it suffices to verify that

\[
\|\mathcal{R}_m f\|_{\mathcal{M}(\mathbb{R})} = \frac{1}{2} \left\| \mathcal{D}^{m/2} f \right\|_{\mathcal{M}(\mathbb{R})}
\]

when \( d = 1 \), where \( \mathcal{D} \) is the univariate derivative operator. Certainly this is true. Indeed, when \( d = 1 \), we have that \( c_d = 1/2 \) and \( \Delta^{m/2} = \mathcal{D}^m \). We also have from (3.3) that the univariate Radon transform is simply

\[
\mathcal{R}\{f\}(\gamma, t) = \int_{\mathbb{R}} f(x) \delta(\gamma x - t) \, dx = \frac{1}{|\gamma|} \int_{\mathbb{R}} f(x) \delta(x - \frac{t}{\gamma}) \, dx = \int_{\mathbb{R}} f(x) \delta(x - \frac{t}{\gamma}) \, dx = f\left(\frac{t}{\gamma}\right),
\]

where the second equality holds since the Dirac impulse is homogeneous of degree \(-1\) and the third equality holds since \( \gamma \in \mathbb{S}^0 = \{-1, +1\} \). Thus,

\[
\left\| \mathcal{R}^{d-1} \mathcal{R}^{m/2} f \right\|_{\mathcal{M}(\mathbb{P}^d)} \bigg|_{d=1} = \frac{1}{2} \left\| \mathcal{D}^{m} f \right\|_{\mathcal{M}(\mathbb{R})}
\]

where the last equality holds since \( f(\cdot/\gamma) \) is either \( f \) or its reflection, both of which will have the same \( \|\mathcal{D}^{m} \{\cdot/\gamma\}\|_{\mathcal{M}(\mathbb{R})} \) value. Thus, by Remark 2.7 and the main result from the
framework of L-splines (Unser et al., 2017), we see that univariate polynomial ridge splines of order \( m \) are exactly the same as classical univariate polynomial splines of order \( m \). This connection between regularized univariate single-hidden layer neural networks and classical notions of univariate splines have been recently explored in Savarese et al. (2019); Parhi and Nowak (2019). This says, by Proposition 2.13, that training a wide enough univariate neural network with either an appropriate path-norm regularizer or an appropriate weight decay regularizer on data results in a polynomial spline approximation of the data. Moreover, these splines are in fact the well-known locally adaptive regression splines of Mammen and van de Geer (1997).

5.2 Ridge splines correspond to univariate splines in the Radon domain

Another way to view a ridge spline is as a continuum of univariate polynomial splines in the Radon domain, where the continuum is indexed by directions \( \gamma \in S^{d-1} \). Suppose \( V \) corresponds to the ideal sampling setting where the sampling locations are located at \( \{x_n\}_{n=1}^N \subset \mathbb{R}^d \). Then, using the same technique we did in the proof of Theorem 2.1, we can recast the continuous-domain inverse problem in (2.5) as one with interpolation constraints:

\[
\min_{f \in F_m} \|R_m f\|_{\mathcal{M}(\mathbb{P}^d)} \quad \text{s.t.} \quad f(x_n) = z_n, \ n = 1, \ldots, N, \tag{5.1}
\]

for some \( z \in \mathbb{R}^N \). By (3.5), the Radon inversion formula, we can always write \( f = c_d R^* \Lambda^{d-1} f \) for any \( f \in F_m \), where the operators are understood in the distributional sense via Corollary 3.6. Thus, using the definition of \( R_m \), we can rewrite the above optimization as

\[
\min_{f \in F_m} c_d \left\| \Lambda^{d-1} \Delta^{m/2} f \right\|_{\mathcal{M}(\mathbb{P}^d)} \quad \text{s.t.} \quad (c_d R^* \Lambda^{d-1} R f)(x_n) = z_n, \ n = 1, \ldots, N.
\]

By (3.8), the intertwining relations of the Radon transform and the Laplacian, combined with the fact that \( m \) is even by assumption, we see that the above optimization can be rewritten as

\[
\min_{f \in F_m} c_d \left\| \partial_t^m \Lambda^{d-1} R f \right\|_{\mathcal{M}(\mathbb{P}^d)} \quad \text{s.t.} \quad (c_d R^* \Lambda^{d-1} R f)(x_n) = z_n, \ n = 1, \ldots, N.
\]

If we put \( \Phi := c_d \Lambda^{d-1} R f \), then the above optimization is

\[
\min_{\Phi \in \tilde{F}_m} \| \partial_t^m \Phi \|_{\mathcal{M}(\mathbb{P}^d)} \quad \text{s.t.} \quad \mathcal{R}^* \{ \Phi \}(x_n) = \int_{S^{d-1}} \Phi(\gamma, \gamma^T x_n) d\sigma(\gamma) = z_n, \ n = 1, \ldots, N, \tag{5.2}
\]

where \( \tilde{F}_m \) is the image of \( c_d \Lambda^{d-1} R \) applied to \( F_m \). Since \( \mathcal{R}^* \) is a continuous linear operator, it follows from classical spline theory (e.g., Unser et al. (2017)) that there exists a sparse minimizer \( \Phi \) to (5.3) that is a univariate polynomial spline of order \( m \) in the offset variable \( t \), i.e., for a fixed direction \( \gamma \in S^{d-1} \), the function \( \Phi(\gamma, \cdot) : \mathbb{R} \rightarrow \mathbb{R} \) is an \( m \)th order polynomial spline. More specifically, this follows by considering an optimization for each \( \gamma \in S^{d-1} \):

\[
\min_{\Phi(\gamma, \cdot)} \| \partial_t^m \Phi(\gamma, \cdot) \|_{\mathcal{M}(\mathbb{R}^d)} \quad \text{s.t.} \quad \Phi(\gamma, \gamma^T x_n) = z_n(\gamma), \ n = 1, \ldots, N, \tag{5.3}
\]
where
\[
\int_{\mathbb{S}^{d-1}} z_n(\gamma) \, d\sigma(\gamma) = z_n, \quad n = 1, \ldots, N
\]
and noting that by finding a solution for each fixed \(\gamma \in \mathbb{S}^{d-1}\), we can find a \(\Phi\) that attains a lower bound for (5.2)\(^8\), but this \(\Phi\) is clearly feasible for (5.2) and is hence a solution to (5.2). It then follows from, e.g., Unser et al. (2017), that \(\Phi(\gamma, \cdot)\) is a polynomial spline of order \(m\). In particular, due to the structure of the interpolation constraints in (5.3), we see that \(\Phi(\gamma, \cdot)\) is a spline where the sampling points are at \(\{\gamma^T x_n\}_{n=1}^N \subset \mathbb{R}\). This viewpoint allows us to understand additional structural information about the sparse (i.e., single-hidden layer neural network) solutions to (5.1). In particular, its classically known\(^9\) that the univariate spline \(\Phi(\gamma, \cdot)\) has some set of adaptive knot locations \(\{t^\ell(\gamma)\}_{\gamma \in \mathbb{S}^{d-1}, \ell=1}^{K_\gamma} \subset \mathbb{R}\) with \(K_\gamma \leq N - m\) and there are no knots outside the sampling locations\(^10\), i.e.,
\[
|t^\ell(\gamma)| \leq \max_{n=1, \ldots, N} |\gamma^T x_n|, \quad \ell = 1, \ldots, K_\gamma.
\]

It is then clear that for \(\Phi\) to be a sparse minimizer of (2.5), it must satisfy Definition 2.8. This implies that the biases in a ridge spline solution to (2.5) exactly correspond to these knot locations. Thus, we can see that for a ridge spline solution to (2.5) as in (2.6) with \(K\) neurons, we have the additional information about a bound on the bias terms
\[
|b_k| \leq \sup_{\gamma \in \mathbb{S}^{d-1}} \max_{\ell = 1, \ldots, K_\gamma} |t^\ell(\gamma)| \leq \sup_{\gamma \in \mathbb{S}^{d-1}} \max_{n=1, \ldots, N} |\gamma^T x_n| \leq \max_{n=1, \ldots, N} \|x_n\|_2, \quad k = 1, \ldots, K,
\]
when we are in the ideal sampling scenario.

6. Applications to Neural Networks

6.1 Finite-dimensional neural network training problems

In this section we will prove Proposition 2.13.

Lemma 6.1 Consider the single-hidden layer neural network
\[
f_\theta(x) := \sum_{k=1}^K v_k \rho_m(w_k^T x - b_k) + c(x),
\]
where \(\theta = (w_1, \ldots, w_K, v_1, \ldots, v_K, b_1, \ldots, b_K, c)\) contains the neural network parameters such that \(v_k \in \mathbb{R}, w_k \in \mathbb{R}^d,\) and \(b_k \in \mathbb{R}\) for \(k = 1, \ldots, K,\) and where \(c\) is a polynomial of degree strictly less than \(m\). Then,
\[
\|R_m f_\theta\|_{M(\mathbb{P}^d)} = \sum_{k=1}^K |v_k|\|w_k\|_{2}^{m-1}.
\]

---

8. Since the integral of a min is less than or equal to the min of an integral.
9. Since we can always explicitly construct spline solutions with the spline knots bounded by the data.
10. Notice that the number of knots and the knot locations depend on the direction \(\gamma \in \mathbb{S}^{d-1}\).
Proof This proof is a direct calculation. Write
\[ R_m f_\theta = \sum_{k=1}^{K} v_k R_m \rho_m(w_k^T(\cdot) - b_k) \]
\[ = \sum_{k=1}^{K} v_k \|w_k\|_2^{m-1} R_m \rho_m(\tilde{w}_k^T(\cdot) - \tilde{b}_k) \]
\[ = \sum_{k=1}^{K} v_k \|w_k\|_2^{m-1} \delta_{p_d}(\cdot - (\tilde{w}_k, \tilde{b}_k)), \]
where the second line follows from the substitution \( \tilde{w}_k := w_k / \|w_k\|_2 \in S^{d-1} \) and \( \tilde{b}_k := b_k / \|w_k\|_2 \in \mathbb{R} \) combined with the homogenity of degree \( m - 1 \) of \( \rho_m \) and the third line follows from Lemma 4.2. Taking the \( \|\cdot\|_{M(\mathbb{P}^d)} \)-norm proves the lemma.

6.1.1 Proof of Proposition 2.13

Proof Equivalence of the problem in (2.5) and (2.9) follows from Theorem 2.1. Equivalence of the problem in (2.9) and (2.10) follows from Lemma 6.1. Thus, we just need to show that the problems in (2.10) and (2.11) are equivalent.

Put
\[ I[\theta] := G(V f_\theta) + \sum_{k=1}^{K} v_k \|w_k\|_2^{m-1} \quad \text{and} \quad J[\theta] := G(V f_\theta) + \frac{1}{2} \sum_{k=1}^{K} |v_k|^2 + \|w_k\|_2^{2m-2}. \]

Let \( \bar{\theta} \) be an optimal solution to (2.10) (i.e., \( \bar{\theta} \) minimizes \( I[\cdot] \)) and let \( \tilde{\theta} \) be an optimal solution to (2.11) (i.e., \( \tilde{\theta} \) minimizes \( J[\cdot] \)). We will show that \( I[\bar{\theta}] = J[\tilde{\theta}] \). Clearly,
\[ I[\bar{\theta}] \leq I[\tilde{\theta}] \leq J[\tilde{\theta}], \quad (6.1) \]
where the second inequality holds by the inequality of arithmetic and geometric means.

Next, define \( \theta \) via the parameters
\[ w_k := \frac{|\tilde{v}_k|^{1/(2m-2)}}{\|w_k\|_2^{1/2}} \tilde{w}_k, \quad v_k := \text{sgn}(\tilde{v}_k)|\tilde{v}_k|^{1/2}\|\tilde{w}_k\|_2^{(m-1)/2}, \quad b_k := \frac{|\tilde{v}_k|^{1/(2m-2)}}{\|w_k\|_2^{1/2}} \tilde{b}_k, \quad \text{and} \quad c := \tilde{c}, \]
for \( k = 1, \ldots, K \). Notice that \( \theta \) and \( \bar{\theta} \) implement the same function, i.e., \( f_\theta = f_{\bar{\theta}} \). Also notice that
\[ \frac{1}{2} \sum_{k=1}^{K} |v_k|^2 + \|w_k\|_2^{2m-2} = \frac{1}{2} \sum_{k=1}^{K} |\tilde{v}_k|^2 \|w_k\|_2^{m-1} + |\tilde{v}_k| \|\tilde{w}_k\|_2^{(m-1)/2} = \sum_{k=1}^{K} |\tilde{v}_k|^2 \|w_k\|_2^{m-1}. \]

This says \( J[\theta] = I[\tilde{\theta}] \). Hence,
\[ J[\tilde{\theta}] \leq J[\theta] = I[\tilde{\theta}], \quad (6.2) \]
Combining (6.1) and (6.2) completes the proof.
6.2 Generalization bounds for binary classification

In this section we will prove Theorem 2.16. We first remark that by (5.4) in Section 5.2, we know that the bias terms $b_k$, $k = 1, \ldots, K$ in a solution to (2.14) satisfy

$$|b_k| \leq \max_{n=1,\ldots,N} \|x_n\|_2.$$ 

Thus, to prove Theorem 2.16, it suffices to consider the class of neural networks

$$F_{m,B,C} := \{ f \in F_{m,B} : |b_k| \leq \frac{C}{2}, k = 1, \ldots, K \}.$$ 

6.2.1 Proof of Theorem 2.16

**Proof** Using the rescaling technique discussed in Remark 2.3, without loss of generality, we may assume that $\|w_k\|_2 = 1$ (since we can absorb the norm of $w_k$ into the magnitude of $v_k$). In this case,

$$\|R_m f_{\theta}\|_{\mathcal{M}(\mathbb{P}^d)} = \sum_{k=1}^K |v_k| \leq B.$$

To begin we bound the empirical Rademacher complexity of $F_{m,B,C}$. The so-called empirical Rademacher complexity, denoted by $\hat{R}(F_{m,B,C})$, is computed by taking the conditional expectation, conditioning on $\{x_n\}_{n=1}^N$, i.e., the only random variables are $\{\sigma_n\}_{n=1}^N$.

The Rademacher complexity is then

$$R(F_{m,B,C}) = \mathbb{E}\left[\hat{R}(F_{m,B,C})\right].$$

We will first consider the empirical Rademacher complexity of a single neuron, i.e., functions of the form $x \mapsto \rho_m(w^T x - b)$, with $\|w\|_2 = 1$ and $|b| \leq C/2$. Write $\mathbb{E}_\sigma[\cdot]$ for $\mathbb{E}[\cdot \mid \{x_n\}_{n=1}^N]$. The empirical Rademacher complexity of a single neuron is defined to be

$$\hat{R}\left(\rho_m(w^T (\cdot) - b)\right) := \mathbb{E}_\sigma\left[\sup_{w: \|w\|_2 = 1, b: |b| \leq C/2} \frac{1}{N} \sum_{n=1}^N \sigma_n \rho_m(w^T x_n - b)\right].$$

Note that since $m$ is even,

$$\rho_m(w^T x - b) = \frac{\rho_2((w^T x - b)^{m-1})}{(m-1)!}.$$ 

Using this and the fact that $\rho_2$ is 1-Lipschitz with the contraction property of the Rademacher complexity (Shalev-Shwartz and Ben-David, 2014, Lemma 26.9), we have

$$\hat{R}\left(\rho_m(w^T (\cdot) - b)\right) = \frac{2}{N(m-1)!} \mathbb{E}_\sigma\left[\sup_{w: \|w\|_2 = 1, b: |b| \leq C/2} \frac{1}{N} \sum_{n=1}^N \sigma_n \rho_2((w^T x_n - b)^{m-1})\right]$$

$$\leq \frac{2}{N(m-1)!} \mathbb{E}_\sigma\left[\sup_{w: \|w\|_2 = 1, b: |b| \leq C/2} \sum_{n=1}^N \sigma_n (w^T x_n - b)^{m-1}\right].$$
Next, by the binomial theorem

\[
\leq \frac{2}{N(m-1)!} \sum_{k=0}^{m-1} \binom{m-1}{k} E_{\sigma} \left[ \sup_{w : \|w\|_2 = 1} \sum_{n=1}^{N} \sigma_n (w^T x_n)^k (-b)^{m-1-k} \right] \\
\leq \frac{2}{N(m-1)!} \sum_{k=0}^{m-1} \binom{m-1}{k} \left( \frac{C}{2} \right)^{m-1-k} E_{\sigma} \left[ \sup_{w : \|w\|_2 = 1} \sum_{n=1}^{N} \sigma_n (w^T x_n)^k \right] \\
= \frac{2}{N(m-1)!} \sum_{k=0}^{m-1} \binom{m-1}{k} \left( \frac{C}{2} \right)^{m-1-k} E_{\sigma} \left[ \left( \sum_{n=1}^{N} \sigma_n \sigma_n^k \right)^T \sigma_n^k \right],
\]

where \((\cdot)^\otimes k\) denotes the \(k\)th order Kronecker product. By Jensen’s inequality we have

\[
E_{\sigma} \left[ \left\| \sum_{n=1}^{N} \sigma_n \sigma_n^k \right\|_2 \right] \leq E_{\sigma} \left[ \left( \sum_{n=1}^{N} \sigma_n \sigma_n^k \right)^2 \right]^{1/2} \leq \sqrt{N} \left( \frac{C}{2} \right)^k,
\]

and so we have

\[
\mathcal{R} \left( \rho_m (w^T (\cdot) - b) \right) \leq \frac{2C^{m-1}}{\sqrt{N(m-1)!}},
\]

Therefore, the empirical Rademacher complexity of \(F_{m,B,C}\) is bounded as follows

\[
\mathcal{R}(F_{m,B,C}) = \sum_{k=1}^{K} |v_k| \mathcal{R} \left( \rho_m (w_k^T (\cdot) - b_k) \right) + \mathcal{R}(c) \\
\leq \frac{2BC^{m-1}}{\sqrt{N(m-1)!}} + \mathcal{R}(c).
\]

Taking the expectation of both sides proves the theorem.

\[\blacksquare\]

**Remark 6.2** An identical proof with an identical bound holds where \(F_{m,B,C}\) is not restricted to neural networks, by appealing to the integral representation in Lemma 4.3 for all functions \(f \in \mathcal{F}_m\).

### 7. Conclusion

In this paper we have developed a variational framework in which we propose and study a family of continuous-domain linear inverse problems in order to understand what happens on the function space level when training a single-hidden layer neural network on data. We have exploited the connection between ridge functions and the Radon transform to show that training a single-hidden layer neural network on data with an appropriate regularizer
results in a function that is optimal with respect to a total variation-like seminorm in the Radon domain. We also show that this seminorm directly controls the generalizability of these neural networks. Our framework encompasses ReLU networks and the appropriate regularizers correspond to the well-known weight decay and path-norm regularizers. Moreover, the variational problems we study are similar to those that are studied in variational spline theory and so we also develop the notion of a ridge spline and make connections between single-hidden layer neural networks and classical spline theory. There are a number of followup research questions that may be asked.

**Computational issues**

Perhaps the most important followup question is how to computationally solve the proposed continuous-domain linear inverse problems. Empirical and theoretical work from the machine learning community has shown that simply running (stochastic) gradient descent on a neural network seems to find global minima (Zhang et al., 2016; Wei et al., 2019; Du et al., 2019a, b), though full theoretical justifications of why these algorithms work currently do not exist. Moreover the theoretical work always operates in the limit of an infinite-width network. Perhaps methods that directly solve the proposed continuous-domain problem can be developed with provable guarantees for finite-width networks.

**Deep networks**

Another important followup question revolves around deep, multilayer networks. Can a variational framework be used to understand what happens when a deep network is trained in data? Answering this question would require posing a continuous-domain inverse problems and deriving a representer theorem showing that deep networks are solutions. We believe answering this question will challenging, due to the compositions of ridge functions that arise in deep networks.

**Appendix A. Auxiliary Proofs for the Representer Theorem**

**A.1 Proof of Lemma 4.1**

**Proof** Consider the unit-ball

\[ B := \left\{ u \in \mathcal{M}(\mathbb{P}^d) : \|u\|_{\mathcal{M}(\mathbb{P}^d)} \leq 1 \right\}. \]

By Theorem 4.11, it suffices to show that the extreme points of \( B \) are of the form \( \sigma \delta_z \), where \( \sigma \in \{-1, +1\} \) and \( \delta_z \) denotes the translated Dirac impulse in \( \mathcal{M}(\mathbb{P}^d) \) supported at \( z \), viewed as a measure. Showing this is a fairly standard exercise, but we go through for the sake of completeness. We will first show that \( \sigma \delta_z \) is an extreme point. Suppose

\[ \sigma \delta_z = tu_1 + (1 - t)u_2, \tag{A.1} \]

for some \( u_1, u_2 \in B \) and some \( t \in [0, 1] \). Let \( |\cdot| \) denote the total-variation measure. Clearly \( |u_1| \) and \( |u_2| \) must be probability measures. Indeed, if not, then \( \|\sigma \delta_z\|_{\mathcal{M}(\mathbb{P}^d)} \leq t \|u_1\|_{\mathcal{M}(\mathbb{P}^d)} + (1 - t) \|u_2\|_{\mathcal{M}(\mathbb{P}^d)} < t + (1 - t) = 1 \), a contradiction. Next,

\[ \delta_z = |\sigma \delta_z| \leq t|u_1| + (1 - t)|u_2| =: u. \]

31
Since \(|u_1|\) and \(|u_2|\) are probability measures, \(u\) must be a probability measure. Hence the inequality in the above display must be an equality. Indeed, given a measurable set \(E\), if \(z \in E\) then
\[
1 = \delta_z(E) \leq u(E) \leq 1.
\]
On the other hand, if \(z \notin E\) then \(u(E) = u(\mathbb{P}^d) - u(\mathbb{P}^d \setminus E) = 1 - 1 = 0\). Hence \(u = \delta_z\). This implies \(|u_1| = |u_2| = \delta_z\). Thus, \(u_1 = a_1\delta_z\) and \(u_2 = a_2\delta_z\), where \(|a_1| = |a_2| = 1\). Thus, (A.1) is equivalent to
\[
\sigma = ta_1 + (1 - t)a_2,
\]
where \(\sigma, a_1, a_2 \in \{-1, +1\}\). This can only hold when \(\sigma = a_1 = a_2\), which implies \(\sigma\delta_z = u_1 = u_2\), which implies \(\sigma\delta_z\) is an extreme point of \(B\).

We will now show that these are the only extreme points. We will proceed by contradiction. Let \(u\) be an extreme point that is not a Dirac impulse. Then, \(\|u\|_{\mathcal{M}(\mathbb{P}^d)} = 1\). Let \(u 1_E\) denote the restriction of \(u\) to the set \(E\). For every measurable \(E\) it is always true that\(^{11}\)
\[
u = u 1_E + u 1_{\mathbb{P}^d \setminus E} = \begin{cases} \frac{u(E)}{|u(E)|} & \text{if } u(E) > 0 \\ \frac{u(E)}{|u(E)|} & \text{if } u(E) < 0 \\ \frac{u(E)}{|u(E)|} & \text{if } u(E) = 0 \end{cases}.
\]
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 says \(u = tu_1 + (1 - t)u_2\) with \(u \neq u_1\) and \(u \neq u_2\), a contradiction. Therefore, \(u\) cannot be an extreme point.

\[\text{A.2 Proof of Lemma 4.3}\]

\[\text{Proof}\] Given \(f \in \mathcal{F}_m\), consider the linear functional
\[
T_f : \psi \mapsto \langle R_m f, \psi \rangle = c_d \langle f, \Delta^{m/2} \mathcal{R}^\ast \Lambda^{d-1} \psi \rangle.
\]
Clearly, \(T_f\) continuously maps \(\mathcal{S}_0(\mathbb{P}^d) \rightarrow \mathbb{R}\). Indeed,
\[
\sup_{\psi \in \mathcal{S}_0(\mathbb{P}^d)} \frac{T_f(\psi)}{\|\psi\|_{L^\infty(\mathbb{P}^d)}} = \sup_{\psi \in \mathcal{C}_0(\mathbb{P}^d)} \frac{T_f(\psi)}{\|\psi\|_{L^\infty(\mathbb{P}^d)}} = \|R_m f\|_{\mathcal{M}(\mathbb{P}^d)} < \infty,
\]
where the first equality holds since \(\mathcal{S}_0(\mathbb{P}^d)\) is dense in \(\mathcal{C}_0(\mathbb{P}^d)\) (cf. Samko (1995)) and the second equality holds by definition of the \(\|\cdot\|_{\mathcal{M}(\mathbb{P}^d)}\)-norm. The above display says \(T_f\) is bounded and hence continuous. The above display also says \(T_f\) continuously maps \(\mathcal{C}_0(\mathbb{P}^d) \rightarrow \mathbb{R}\). By the Riesz–Markov–Kakutani representation, there exists \(\mu \in \mathcal{M}(\mathbb{P}^d)\) such that
\[
T_f(\psi) = \int_{\mathbb{P}^{d-1} \times \mathbb{R}} \psi(w, b) \, d\mu(w, b)
\]
\(^{11}\) We adopt the convention that \(0/0 := 0\) in this equation.
for all $\psi \in \mathcal{S}_0(\mathbb{R}^d)$. Put
\[
g_\mu(x) := \int_{\mathbb{S}^{d-1} \times \mathbb{R}} r^{(m)}_{(w,b)}(x) \, d\mu(w,b),
\]
where $r^{(m)}_{(w,b)}(x)$ is as in (4.3). We now claim that $\Delta^{m/2} f = \Delta^{m/2} g_\mu$ in the sense of Lizorkin distributions. Indeed, first notice that
\[
\Delta^{m/2} g_\mu(x) = \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \Delta^{m/2} r^{(m)}_{(w,b)}(x) \, d\mu(w,b) = \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \delta_\mathbb{R}(w^T x - b) \, d\mu(w,b) = \mathcal{R}^* \mu,
\]
where the second equality holds by Lemma 4.2 and the third equality holds by (3.4). Next, for all $\varphi \in \mathcal{S}_0(\mathbb{R}^d)$
\[
\langle \Delta^{m/2} g_\mu, \varphi \rangle = \langle \mathcal{R}^* \mu, \varphi \rangle = \langle \mu, \mathcal{R} \varphi \rangle = T_f(\mathcal{R} \varphi) = c_d \langle f, \Delta^{m/2} \mathcal{R}^* \mathcal{R}^{d-1} \mathcal{R} \varphi \rangle = \langle f, \Delta^{m/2} \varphi \rangle = \langle \Delta^{m/2} f, \varphi \rangle,
\]
where the penultimate line holds by (3.5), the Radon inversion formula. The equality $\Delta^{m/2} f = \Delta^{m/2} g_\mu$ in the sense of Lizorkin distributions is equivalent to saying $(-\Delta)^{m/2} f = (-\Delta)^{m/2} g_\mu$ in the sense of Lizorkin distributions which is subsequently equivalent to saying
\[
(-\Delta)^{m/2} (f - g_\mu) = p
\]
in the sense of tempered distributions where $p \in \mathcal{P}(\mathbb{R}^d)$ is some polynomial. Taking the Fourier transform of the above display we find
\[
\| \xi \|^m_2 \left[ \hat{f}(\xi) - \hat{g}_\mu(\xi) \right] = \hat{p}(\xi), \quad \xi \in \mathbb{R}^d.
\]
The left-hand side of the above display is 0 at $\xi = 0$. Meanwhile, since $p \in \mathcal{P}(\mathbb{R}^d)$, $\hat{p}$ is supported only at $\{0\}$. Thus, for the equality in the above display to holds we have $p \equiv 0$. Hence,
\[
\| \xi \|^m_2 \left[ \hat{f}(\xi) - \hat{g}_\mu(\xi) \right] = 0, \quad \xi \in \mathbb{R}^d.
\]
For the equality in the above display to hold, $\hat{f} - \hat{g}_\mu$ is supported at $\{0\}$. Hence, $f - g_\mu$ must be a polynomial. Finally, since $f \in \mathcal{F}_m \subset L^{\infty,m-1}(\mathbb{R}^d)$ and $g_\mu \in L^{\infty,m-1}(\mathbb{R}^d)$ by construction (since it is a superposition of functions of growth rate $m - 1$), we have that $f - g_\mu \in L^{\infty,m-1}(\mathbb{R}^d)$. Thus,
\[
f(x) = \int_{\mathbb{S}^{d-1} \times \mathbb{R}} r^{(m)}_{(w,b)}(x) \, d\mu(w,b) + c(x),
\]
where \( c \) is a polynomial of growth rate at most \( m - 1 \), i.e., \( c \) is a polynomial of degree strictly less than \( m \).

### A.3 Proof of Lemma 4.6

**Proof** Given \( f \in \mathcal{F}_m \), by Lemma 4.3, there exists \( \mu \in \mathcal{M}(\mathbb{P}^d) \) and \( p \in \mathcal{P}(\mathbb{R}^d) \) with degree strictly less than \( m \) such that

\[
 f(z) = \int_{\mathbb{R}^{d-1} \times \mathbb{R}} r^{(m)}(w, b) d\mu(w, b) + c(z). \quad (A.3)
\]

First, clearly the (growth restricted) null space of \( R_m = c_d \Delta^{d-1} \mathcal{R} \Delta^{m/2} \) is a subset of the (growth restricted) null space of \( \Delta^{m/2} \), so it suffices to show that the (growth restricted) null space of \( \Delta^{m/2} \) is finite-dimensional. Clearly every polynomial of degree strictly less than \( m \) gets annihilated by \( \Delta^{m/2} \). Next, from a similar calculation as in (A.2) applied to (A.3), we see that \( \Delta^{m/2} f \equiv 0 \) if and only if \( \mu \) is the zero measure. Therefore, \( N_m \) is a subset of polynomials of degree strictly less than \( m \) and is thus finite-dimensional.

### A.4 Proof of Lemma 4.8

Before proving Lemma 4.8, we state a variant of the Schwartz kernel theorem for smooth manifolds which allows us to characterize operators by their kernel.

**Theorem A.1** *(Hörmander (2015, Theorem 5.2.1))* Let \( X \) and \( Y \) be smooth manifolds. For every continuous linear \( G : C^\infty_{cpct}(Y) \to \mathcal{D}'(X) \), there exists a unique \( g \) (the Schwartz kernel of \( G \)) such that

\[
 \langle G u, v \rangle = \langle g, u \otimes v \rangle,
\]

where \( C^\infty_{cpct}(Y) \) denotes smooth functions with compact support on \( Y \), \( \mathcal{D}'(X) \) denotes the space of distributions on \( X \), and \( \otimes \) is the tensor product. Here, continuity is understood sequentially, i.e., if \( \varphi_k \to \varphi \) in \( C^\infty_{cpct}(Y) \), then \( G \varphi_k \to G \varphi \) weakly in \( \mathcal{D}'(X) \).

**Remark A.2** Since the Radon domain, \( \mathbb{P}^d \), is a smooth manifold, Theorem A.1 says, we can view operators mapping functions from \( \mathbb{P}^d \to \mathbb{R} \) to functions from \( \mathbb{R}^d \to \mathbb{R} \) via their (Schwartz) kernels. We do exactly this in Lemma 4.8.

**Proof** [Proof of Lemma 4.8] Since \( R_m r_z = \delta_{\mathbb{P}^d}(\cdot - z) \) by Lemma 4.2 and \( p_n \in \mathcal{N}_m \), a direct calculation results in

\[
 R_m R_{m, \varphi}^{-1} \psi = \psi
\]

for all \( \psi \in \mathcal{S}_0(\mathbb{P}^d) \). To check the boundary conditions we check for all \( \psi \in \mathcal{S}_0(\mathbb{P}^d) \) that

\[
 \langle \phi_k, R_{m, \varphi}^{-1} \psi \rangle = 0, \quad k = 1, \ldots, N_0.
\]

Write

\[
 \langle \phi_k, R_{m, \varphi}^{-1} \psi \rangle = \langle \phi_k, \int_{\mathbb{R}^{d-1} \times \mathbb{R}} r_z(\cdot) \psi(z) d(\sigma \times \lambda)(z) \rangle - \sum_{n=1}^{N_0} \langle \phi_k, p_n \rangle \langle q_n, \psi \rangle.
\]

\[
 = \langle \psi, \psi \rangle
\]
Next, for any $k = 1, \ldots, N_0$

$$
(q_k, \psi) = \int_{\mathbb{S}^{d-1} \times \mathbb{R}} q_k(z) \psi(z) \, d(\sigma \times \lambda)(z) = \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \langle \phi_k, r_z \rangle \psi(z) \, d(\sigma \times \lambda)(z)
$$

Thus, $\langle \phi_k, R_{m,\psi}^{-1} \psi \rangle = 0$, for $k = 1, \ldots, N_0$. Uniqueness of $R_{m,\phi}^{-1}$ follows from the fact that the biorthogonal system provides a unique representation of elements of $N_m$.

Clearly $R_{m,\phi}^{-1}$ is continuous and hence bounded. Thus we see that this inverse is stable.

By continuity and the denseness of $\mathcal{S}_0(\mathbb{P}^d)$ in $C_0(\mathbb{P}^d)$ (cf. Samko (1995)), we can continuously extend $R_{m,\phi}^{-1}$ to act on $C_0(\mathbb{P}^d)$. It then follows by duality that $R_{m,\phi}^{-1}$ extends continuously to act on $\mathcal{M}(\mathbb{P}^d)$. Finally, it is clear that this extension maps $\mathcal{M}(\mathbb{P}^d) \to L^{\infty,m-1}(\mathbb{P}^d)$ by construction of the (Schwartz) kernel $g_\phi$ of $R_{m,\phi}^{-1}$.

\section*{A.5 Proof of Theorem 4.9}

\textbf{Proof}

1. Recall from the theorem statement the definition

$$
F_{m,\phi} := \{ f \in \mathcal{F} : \phi(f) = 0 \}.
$$

Clearly this is a vector space. Next, since $f \mapsto \|R_m f\|_{\mathcal{M}(\mathbb{P}^d)}$ is a seminorm, it is a norm except for the property that $\|R_m f\|_{\mathcal{M}(\mathbb{P}^d)} = 0$ if and only if $f \equiv 0$ (since every $q \in N_m$ has $\|R_m q\|_{\mathcal{M}(\mathbb{P}^d)} = 0$). By imposing the boundary conditions $\phi(f) = 0$ in the definition of $F_{m,\phi}$, we enforce that every $f \in F_{m,\phi}$ has no null space component. Thus, $F_{m,\phi}$ is a bona fide Banach space when equipped with the norm $f \mapsto \|R_m f\|_{\mathcal{M}(\mathbb{P}^d)}$.

In particular, this says we have the topological isomorphism $F_{m,\phi} \cong F_m / N_m$.

The fact that $f \mapsto \|R_m f\|_{\mathcal{M}(\mathbb{P}^d)}$ is a norm for $F_{m,\phi}$ implies that $R_m$ is a bijection that maps $F_{m,\phi} \to \mathcal{M}(\mathbb{P}^d)$. Since these are both Banach spaces, the bounded inverse theorem (Pollard, 1999, Chapter 5) says there exists a bounded inverse $R_m^{-1} : \mathcal{M}(\mathbb{P}^d) \to F_{m,\phi}$ of $R_m$. This inverse is precisely the operator $R_{m,\phi}^{-1}$ constructed in Lemma 4.8 due to the boundary conditions in (4.6). In other words, $R_{m,\phi}^{-1} : \mathcal{M}(\mathbb{P}^d) \to F_{m,\phi}$ is a bona fide inverse.

Next, given a biorthogonal system $(\phi, p)$ of $N_m$, consider the projection operator

$$
\text{proj}_{N_m} : f \mapsto \sum_{n=1}^{N_0} \langle \phi_n, f \rangle p_n.
$$

Then, for every $f \in F_m$ we can write $f = \tilde{f} + q$, where $q := \text{proj}_{N_m}$ and so $\tilde{f} = f - q$.

By this construction, $\phi(\tilde{f}) = 0$, so $\tilde{f} \in F_{m,\phi}$. Clearly, $\tilde{f} = R_{m,\phi}^{-1} u$, where $u := R_m \tilde{f} = ...$
$R_m f \in \mathcal{M}(\mathbb{P}^d)$. Indeed, this is true since $R_m^{-1} : \mathcal{M}(\mathbb{P}^d) \to F_{m,\phi}$ is a bona fide inverse. Thus, (4.7) holds. Since $F_{m,\phi} \cong F_m/\mathcal{N}_m$, we have $F_{m,\phi} \cap \mathcal{N}_m = \{0\}$. Hence, we have the direct-sum decomposition $F_m = F_{m,\phi} \oplus \mathcal{N}_m$.

2. Since every $q \in \mathcal{N}_m$ has the unique representation

$$q = \sum_{n=1}^{N_0} \langle \phi_n, q \rangle p_n,$$

we can always identify $q$ by its expansion coefficients $\phi(q) \in \mathbb{R}^{N_0}$. Thus, the norm $q \mapsto \|\phi(q)\|_2$ provides $\mathcal{N}_m$ with a Banach space structure. Finally, since both $F_{m,\phi}$ and $\mathcal{N}_m$ can be endowed with norms to provide a Banach space structure, we can use the direct-sum decomposition $F_m = F_{m,\phi} \oplus \mathcal{N}_m$ to equip $F_m$ with the composite norm

$$\|f\|_{F_m} := \|R_m f\|_{\mathcal{M}(\mathbb{P}^d)} + \|\phi(f)\|_2$$

to provide $F_m$ a Banach space structure. 

References

B. Adcock and A. C. Hansen. Generalized sampling and infinite-dimensional compressed sensing. *Foundations of Computational Mathematics*, 16(5):1263–1323, 2016.

B. Adcock, A. C. Hansen, C. Poon, and B. Roman. Breaking the coherence barrier: A new theory for compressed sensing. In *Forum of Mathematics, Sigma*, volume 5. Cambridge University Press, 2017.

F. Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

R. Balestriero and R. Baraniuk. A spline theory of deep learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 374–383, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

A. R. Barron and J. M. Klusowski. Complexity, statistical risk, and metric entropy of deep nets using total path variation. arXiv preprint arXiv:1902.00800, 2019.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

A. Beurling. Sur les intégrales de Fourier absolument convergentes et leur application à une transformation fonctionelle. In *Ninth Scandinavian Mathematical Congress*, pages 345–366, 1938.
C. Boyer, A. Chambolle, Y. D. Castro, V. Duval, F. De Gournay, and P. Weiss. On representer theorems and convex regularization. *SIAM Journal on Optimization*, 29(2):1260–1281, 2019.

K. Bredies and M. Carioni. Sparsity of solutions for variational inverse problems with finite-dimensional data. *Calculus of Variations and Partial Differential Equations*, 59(1):14, 2020.

K. Bredies and H. K. Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.

A. Calderón. Intermediate spaces and interpolation, the complex method. *Studia Mathematica*, 24(2):113–190, 1964.

E. J. Candès. *Ridgelets: theory and applications*. PhD thesis, Stanford University Stanford, 1998.

E. J. Candès. Harmonic analysis of neural networks. *Applied and Computational Harmonic Analysis*, 6(2):197–218, 1999.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

C. de Boor and R. E. Lynch. On splines and their minimum properties. *Journal of Mathematics and Mechanics*, 15(6):953–969, 1966.

S. Du, J. Lee, H. Li, L. Wang, and X. Zhai. Gradient descent finds global minima of deep neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1675–1685, Long Beach, California, USA, 09–15 Jun 2019a. PMLR.

S. S. Du, X. Zhai, B. Póczos, and A. Singh. Gradient descent provably optimizes over-parameterized neural networks. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019b.

J. Duchon. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer, 1977.

T. Ergen and M. Pilanci. Convex duality of deep neural networks. *arXiv preprint arXiv:2002.09773*, 2020.

S. D. Fisher and J. W. Jerome. Spline solutions to $L^1$ extremal problems in one and several variables. *Journal of Approximation Theory*, 13(1):73–83, 1975.

G. B. Folland. *Real analysis: modern techniques and their applications*. John Wiley & Sons, New York, 2nd edition, 1999.

K.-I. Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3):183–192, 1989.

37
P. Grohs, D. Perekrestenko, D. Elbrächter, and H. Bölcskei. Deep neural network approximation theory. *arXiv preprint arXiv:1901.02220*, 2019.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

S. Helgason. *Integral Geometry and Radon Transforms*. Springer New York, 2014. ISBN 9781489994202.

M. Holschneider. *Wavelets: An Analysis Tool*. Oxford mathematical monographs. Clarendon Press, 1995. ISBN 9780198505211.

L. Hörmander. *The analysis of linear partial differential operators I: Distribution theory and Fourier analysis*. Springer, 2015.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

F. John. *Plane Waves and Spherical Means: Applied to Partial Differential Equations*. Springer New York, 2013. ISBN 9781461394532.

G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

J. M. Klusowski and A. R. Barron. Uniform approximation by neural networks activated by first and second order ridge splines. *arXiv preprint arXiv:1607.07819*, 2016a.

J. M. Klusowski and A. R. Barron. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv preprint arXiv:1607.01434*, 2016b.

S. V. Konyagin, A. A. Kuleshov, and V. E. Maiorov. Some problems in the theory of ridge functions. *Proceedings of the Steklov Institute of Mathematics*, 301(1):144–169, 2018.

S. Kostadinova, S. Pilipović, K. Saneva, and J. Vindas. The ridgelet transform of distributions. *Integral Transforms and Special Functions*, 25(5):344–358, 2014.

M. G. Krein. The L-problem in an abstract normed linear space. *Some questions in the theory of moments*, 1938.

A. Krogh and J. A. Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957, 1992.

M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6 (6):861–867, 1993.

B. F. Logan and L. A. Shepp. Optimal reconstruction of a function from its projections. *Duke mathematical journal*, 42(4):645–659, 1975.

V. E. Maiorov. Best approximation by ridge functions in $L_p$-spaces. *Ukrainian Mathematical Journal*, 62(3):452–466, 2010.
E. Mammen and S. van de Geer. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997.

H. N. Mhaskar. Dimension independent bounds for general shallow networks. *Neural Networks*, 123:142–152, 2020.

C. A. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. In *Approximation theory and spline functions*, pages 143–145. Springer, 1984.

N. Murata. An integral representation of functions using three-layered networks and their approximation bounds. *Neural Networks*, 9(6):947–956, 1996.

B. Neyshabur, R. R. Salakhutdinov, and N. Srebro. Path-sgd: Path-normalized optimization in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 2422–2430, 2015.

G. Ongie, R. Willett, D. Soudry, and N. Srebro. A function space view of bounded norm infinite width ReLU nets: The multivariate case. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

R. Parhi and R. D. Nowak. Minimum “norm” neural networks are splines. *arXiv preprint arXiv:1910.02333*, 2019.

A. Pinkus. *Ridge Functions*. Cambridge Tracts in Mathematics. Cambridge University Press, 2015. ISBN 9781316432587.

P. M. Prenter. *Splines and Variational Methods*. Dover Books on Mathematics. Dover Publications, 2013. ISBN 9780486783499.

S. Rosset, G. Swirszcz, N. Srebro, and J. Zhu. $\ell_1$ regularization in infinite dimensional feature spaces. In *International Conference on Computational Learning Theory*, pages 544–558. Springer, 2007.

B. Rubin. 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, 1998.

S. Samko. Denseness of the spaces $\Phi_Y$ of Lizorkin type in the mixed $L^p(\mathbb{R}^n)$-spaces. *Studia Mathematica*, 3(113):199–210, 1995.

P. H. P. Savarese, I. Evron, D. Soudry, and N. Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, pages 2667–2690, 2019.

B. Schölkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001.
U. Shaham, A. Cloninger, and R. R. Coifman. Provable approximation properties for deep neural networks. *Applied and Computational Harmonic Analysis*, 44(3):537–557, 2018.

S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

S. Sonoda and N. Murata. Neural network with unbounded activation functions is universal approximator. *Applied and Computational Harmonic Analysis*, 43(2):233–268, 2017.

E. M. Stein and R. Shakarchi. *Fourier analysis: an introduction*, volume 1. Princeton University Press, 2011.

M. Unser. A representer theorem for deep neural networks. *Journal of Machine Learning Research*, 20(110):1–30, 2019a.

M. Unser. A unifying representer theorem for inverse problems and machine learning. *arXiv preprint arXiv:1903.00687*, 2019b.

M. Unser and J. Fageot. Native banach spaces for splines and variational inverse problems. *arXiv preprint arXiv:1904.10818*, 2019.

M. Unser, J. Fageot, and J. P. Ward. Splines are universal solutions of linear inverse problems with generalized TV regularization. *SIAM Review*, 59(4):769–793, 2017.

G. Wahba. *Spline models for observational data*, volume 59. SIAM, 1990.

C. Wei, J. D. Lee, Q. Liu, and T. Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pages 9709–9721, 2019.

H. Wendland. *Scattered Data Approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2010. ISBN 9780521131018.

F. Williams, M. Trager, D. Panozzo, C. Silva, D. Zorin, and J. Bruna. Gradient dynamics of shallow univariate ReLU networks. In *Advances in Neural Information Processing Systems*, pages 8376–8385, 2019.

M. P. Wolff and H. H. Schaefer. *Topological Vector Spaces*. Graduate Texts in Mathematics. Springer New York, 2012. ISBN 9781461271550.

Y. Xu and Q. Ye. *Generalized Mercer kernels and reproducing kernel Banach spaces*, volume 258. American Mathematical Society, 2019.

W. Yuan, W. Sickel, and D. Yang. *Morrey and Campanato Meet Besov, Lizorkin and Triebel*. Springer, 2010. ISBN 9783642146077.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

H. Zhang, Y. Xu, and J. Zhang. Reproducing kernel Banach spaces for machine learning. *Journal of Machine Learning Research*, 10(Dec):2741–2775, 2009.
S. Zuhovickii. Remarks on problems in approximation theory. *Mat. Zbirnik KDU*, pages 169–183, 1948.