Journal of Machine Learning Research 22 (2021) 1-40 Submitted 6/20; Revised 11/20; Published 2/21

Banach Space Representer Theorems for Neural Networks and Ridge Splines

Rahul Parhi RAHUL@ECE.WISC.EDU
Robert D. Nowak RDNOWAK@WISC.EDU
Department of Electrical and Computer Engineering
University of Wisconsin–Madison
Madison, WI 53706, USA

Editor: Lorenzo Rosasco

Abstract

We develop a variational framework to understand the properties of the functions learned by neural networks fit to data. We propose and study 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 polynomial ridge splines, which correspond to single-hidden layer neural networks with truncated power functions as the activation function. The representer theorem is reminiscent of the classical reproducing kernel Hilbert space representer theorem, but we show that the neural network problem is posed over a non-Hilbertian Banach space. While 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

$$\boldsymbol{x} \mapsto \rho(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}), \tag{1}$$

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

$$\boldsymbol{x} \mapsto \sum_{k=1}^{K} v_k \, \rho(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k), \tag{2}$$

where $\rho : \mathbb{R} \to \mathbb{R}$ is the *activation function*, K is the *width* of the network, and for $k = 1, \dots, K$, $v_k \in \mathbb{R}$ and $\boldsymbol{w}_k \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$ are the *weights* of the neural network and $b_k \in \mathbb{R}$ are the *biases* or *offsets*.

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 function 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 topological vector space of multivariate 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)[1], and let $f : \mathbb{R}^d \to \mathbb{R}$ be a multivariate function such that $f \in \mathcal{F}$. Consider the continuous-domain inverse problem

$$\min_{f \in \mathcal{F}} G(\mathcal{V}f) + \|f\|, \tag{3}$$

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 $\|\cdot\|_{(m)}$, where $m \geq 2$ is an integer, so that the solutions to the problem (3) with $\|\cdot\| \coloneqq \|\cdot\|_{(m)}$ take the form

$$\boldsymbol{x} \mapsto \sum_{k=1}^{K} v_k \, \rho_m(\boldsymbol{w}_k^{\mathsf{T}} \boldsymbol{x} - b_k) + c(\boldsymbol{x}), \tag{4}$$

   where $\rho_m = \max\{0, \cdot\}^{m-1}/(m-1)!$ is the mth-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 $\partial_t^m \Lambda^{d-1} \mathscr{R} f$, where $\mathscr{R}$ is the Radon transform, Λ^{d-1} is a "ramp" filter, and ∂_t^m is the mth partial derivative with respect to the "offset" variable of the Radon domain. 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. When $m = 2$, the solutions correspond to ReLU networks.

2. We propose the notion of a *ridge spline* by noticing that our problem formulation in (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 (4) with our family of seminorms as *mth-order polynomial ridge splines* to emphasize that the solutions are superpositions

1. For example, $\mathcal{V}f = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_N)) \in \mathbb{R}^N$, for some data $\{\boldsymbol{x}_n\}_{n=1}^N \subset \mathbb{R}^d$.

of ridge functions. We view our notion of a ridge spline as a kind of spline in-between a univariate spline and a traditional multivariate spline. Unlike polyharmonic splines, the usual multivariate analogue of univariate polynomial splines, ridge splines are multivariate piecewise polynomial functions. Moreover, by specializing 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 $\mathcal{V}$ corresponds to *ideal sampling*, i.e., point evaluations, the generality of (3) allows us to consider the machine learning problem of approximating the scattered data $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N \subset \mathbb{R}^d \times \mathbb{R}$ with both *interpolation constraints* in the case of noise-free data as well as *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 (3) can be recast as a *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 $\|\cdot\|_{(m)}$, characterizing a key property of the learned function. 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 (3) (this can be done with an appropriate data fitting term). We derive statistical *generalization bounds* for the class of neural networks with uniformly bounded seminorm $\|\cdot\|_{(m)}$. In particular, we show that the seminorm bounds the *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

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). 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 (2), under extremely mild conditions on the activation function (Cybenko, 1989; Hornik et al., 1989; Funahashi, 1989; Barron, 1993; Leshno et al., 1993). There are also 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 variational spline theory are established (Savarese et al., 2019; Williams et al., 2019; Parhi and Nowak, 2020). 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 (2019); 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) is constant along the hyperplanes $\boldsymbol{w}^\mathsf{T}\boldsymbol{x} = c$, $c \in \mathbb{R}$, 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 "norm" of the network weights is bounded (Ongie et al., 2020). 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 in which the seminorms are well-defined, subject to data fitting constraints. A side effect of our analysis also provides an understanding of the topological structure, specifically a non-Hilbertian Banach space structure, of the spaces defined by these seminorms.

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 make the assumption that the relevant spaces are compact. This allows appealing to standard arguments which

are useful for proving, e.g., that minimizers to their problem even exist. We also remark that some of these prior works simply assume, without proof, existence of minimizers.

Since the Radon domain is an unbounded domain, we cannot appeal to these types of arguments for the problem we study. Thus, a very important question we ask, and subsequently answer, regards existence of solutions to (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 (2) are parameterized by both a direction $\boldsymbol{w}_k$ and a translation b_k. We also draw on recent results from variational methods (Bredies and Carioni, 2020). Thus, the results of this paper provide a 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 polynomial splines. In Section 6 we discuss applications of the representer theorem to neural network training, regularization, and generalization.

2. Main Results

Our main contribution is a representer theorem for problems of the form in (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 $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N \subset \mathbb{R}^d \times \mathbb{R}$. The classical representer theorem considers the variational problem

$$\bar{f} = \operatorname*{arg\,min}_{f \in \mathcal{H}} \sum_{n=1}^{N} \ell(f(\boldsymbol{x}_n), y_n) + \lambda \|f\|_{\mathcal{H}}^2, \tag{5}$$

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 $\bar{f}$ is unique and $\bar{f} \in \operatorname{span}\{k(\cdot, \boldsymbol{x}_n)\}_{n=1}^N$, 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; Unser, 2020) as a way to designate a parametric formulation of solutions to a variational optimization problem, ideally being a linear combination from some dictionary of atoms. This has allowed more much more general problems to be considered than ones like (5), 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 ℓ^1-norm or its continuous-domain analogue, the $\mathcal{M}$-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 (indexed by an integer $m \geq 2$)

$$\|f\|_{(m)} \coloneqq c_d \left\| \partial_t^m \Lambda^{d-1} \mathscr{R} f \right\|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})}, \tag{6}$$

where $\mathscr{R}$ is the Radon transform defined in (24), Λ^{d-1} is a ramp filter in the Radon domain defined in (28), ∂_t^m is the mth partial derivative with respect to t, the offset variable in the Radon domain discussed in Section 3.4, and c_d is a dimension dependent constant defined in (31), $\|\cdot\|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})}$ denotes the total variation norm (in the sense of measures) on the Radon domain. We remark that the $\mathcal{M}$-norm can be viewed, as a "generalization" of the L^1-norm, with the key property that we can apply the $\mathcal{M}$-norm to distributions that are also "absolutely integrable" such as the Dirac impulse (i.e., distributions that can be associated with a finite Radon measure). The space $\mathbb{S}^{d-1} \times \mathbb{R}$ denotes the Radon domain; in particular, the Radon transform computes integrals over hyperplanes in $\mathbb{R}^d$. Since every hyperplane can be written as $\left\{ \boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{x} = t \right\}$ for $\boldsymbol{\gamma} \in \mathbb{S}^{d-1}$, the surface of the ℓ^2-sphere in $\mathbb{R}^d$, and $t \in \mathbb{R}$, the Radon domain is $\mathbb{S}^{d-1} \times \mathbb{R}$. Finally, the space $\mathcal{M}(X)$ is the Banach space of finite Radon measures on X.

The family of seminorms in (6) are thus exactly total variation seminorms in the Radon domain. For brevity, we will write

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

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 discussed in Unser et al. (2017) and in the L^2-theory of radial basis functions and polyharmonic splines (Wendland, 2010, Chapter 10), constructing operators acting on multivariate functions with finite-dimensional null spaces is nearly impossible[2]. To bypass this technicality, we use a common technique

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

from variational spline theory (see, e.g., 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^∞-norm

$$\|f\|_{\infty,n_0} \coloneqq \operatorname*{ess\,sup}_{\boldsymbol{x}\in\mathbb{R}^d} |f(\boldsymbol{x})|(1+\|\boldsymbol{x}\|_2)^{-n_0},$$

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 $\mathbb{R}^d \to \mathbb{R}$ with algebraic growth rate n_0. We will later see in (34) that the appropriate choice of algebraic growth rate for the operator R_m is $n_0 \coloneqq m-1$. This allows us to define the (growth restricted) *null space* of R_m as

$$\mathcal{N}_m \coloneqq \Big\{ q \in L^{\infty,m-1}(\mathbb{R}^d) : \ \mathrm{R}_m\, q = 0 \Big\} \tag{7}$$

and the (growth restricted) *native space* of R_m as

$$\mathcal{F}_m \coloneqq \Big\{ f \in L^{\infty,m-1}(\mathbb{R}^d) : \ \mathrm{R}_m\, f \in \mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R}) \Big\}. \tag{8}$$

We prove in Lemma 19 that $\mathcal{N}_m$ is indeed finite-dimensional.

We now state our representer theorem, in which we show that there exists a sparse solution to the inverse problem in (3) when the seminorm takes the form in (6) and the search space is $\mathcal{F}_m$. In particular, we show that that the sparse solution takes the form of the sum of a single-hidden layer neural network as in (2) and a low degree polynomial. In this context, sparse means that the width of the neural network and the degree of the polynomial are, a priori, bounded from above.

Theorem 1 *Assume the following:*

1. *The function $G : \mathbb{R}^N \to \mathbb{R}$ is a strictly convex, coercive, and lower semi-continuous.*
2. *The operator $\mathcal{V} : \mathcal{F}_m \to \mathbb{R}^N$ is continuous[3], linear, and surjective.*
3. *The inverse problem is well-posed over the null space $\mathcal{N}_m$ 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 variational problem

$$\min_{f\in\mathcal{F}_m} \ G(\mathcal{V}f) + \|\mathrm{R}_m\, f\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} \tag{9}$$

that takes the form

$$s(\boldsymbol{x}) = \sum_{k=1}^{K} v_k\, \rho_m(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{x} - b_k) + c(\boldsymbol{x}), \tag{10}$$

where $K \leq N - \dim\mathcal{N}_m$, $\rho_m = \max\{0,\cdot\}^{m-1}/(m-1)!$, $\boldsymbol{w}_k \in \mathbb{S}^{d-1}$, $v_k \in \mathbb{R}$, $b_k \in \mathbb{R}$, and $c(\cdot)$ is a polynomial of degree strictly less than m.

3. In order to define continuity, $\mathcal{F}_m$ needs to be a *topological* vector space. We prove in Theorem 22 that $\mathcal{F}_m$ is a Banach space, which provides it a topology, allowing continuity to be defined.

Proving Theorem 1 hinges on several technical results, the most important being the topological structure of the native space $\mathcal{F}_m$. In order to do any kind of analysis (e.g., proving that minimizers of (9) even exist), we require the native space $\mathcal{F}_m$ to have some "nice" topological structure. We prove in Theorem 22 that $\mathcal{F}_m$, 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 21. 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, which causes some nuances compared to prior work of Unser et al. (2017); Unser and Fageot (2019).

Theorem 1 shows that while the problem is posed in the continuum, it admits *parametric* solutions in terms of a finite number of parameters. This demonstrates the sparsifying effect of the $\mathcal{M}$-norm, similar to its discrete analogue, the ℓ^1-norm. We also remark that although the problem in Theorem 1 admits a sparse solution, it is important to note that the solution may not be unique and there may also exist non-sparse solutions.

Remark 2 *The polynomial term $c(\boldsymbol{x})$ that appears in* (10) *corresponds to a term in the null space $\mathcal{N}_m$. When $m = 2$, the network in* (10) *is a ReLU network and $c(\boldsymbol{x})$ takes the form*

$$c(\boldsymbol{x}) = \boldsymbol{u}^{\mathsf{T}}\boldsymbol{x} + s,$$

where $\boldsymbol{u} \in \mathbb{R}^d$ and $s \in \mathbb{R}$. Thus, when $m = 2$, (10) *corresponds to a ReLU network with a* skip connection *(He et al., 2016).*

Remark 3 *The fact that $\boldsymbol{w}_k \in \mathbb{S}^{d-1}$ does not restrict the single-hidden layer neural network due to the homogeneity of the truncated power functions. Indeed, given any single-hidden layer neural network with $\boldsymbol{w}_k \in \mathbb{R}^d \setminus \{\boldsymbol{0}\}$, we can use the fact that ρ_m is homogeneous of degree $m-1$ to rewrite the network as*

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

where $\widetilde{\boldsymbol{w}}_k \coloneqq \boldsymbol{w}_k/\|\boldsymbol{w}_k\|_2 \in \mathbb{S}^{d-1}$ and $\widetilde{b}_k \coloneqq b_k/\|\boldsymbol{w}_k\|_2 \in \mathbb{R}$. We use this fact to prove Theorem 8 which recasts the variational problem in (9) *as a finite-dimensional neural network training problem (with no constraints on the input layer weights).*

The proof of Theorem 1 appears in Section 4.

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, $\mathrm{L} : \mathscr{S}'(\mathbb{R}^d) \to \mathscr{S}'(\mathbb{R}^d)$, where $\mathscr{S}'(\mathbb{R}^d)$ denotes the space of tempered distributions on $\mathbb{R}^d$, is associated with a spline, and variational problems of the form

$$\min_{f \in \mathcal{F}_m} \; G(\mathcal{V}f) + \|\mathrm{L}\, f\|_{\mathcal{M}(\mathbb{R}^d)} \tag{11}$$

are studied, where G is a data fitting term, $\mathcal{V}$ is a measurement operator, $\mathcal{F}_m$ is the native space of L, and $\mathcal{M}(\mathbb{R}^d)$ is the space of finite Radon measures on $\mathbb{R}^d$ which forms a Banach space when equipped with $\|\cdot\|_{\mathcal{M}(\mathbb{R}^d)}$, the total variation norm in the sense of measures. The key result from Unser et al. (2017) is a representer theorem for the above problem which states that there exists a sparse solution which is a so-called L-spline. By associating an operator to a spline, we have a simple way to characterize what functions are splines via the following definition.

Definition 4 (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*

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

where $\delta_{\mathbb{R}^d}$ denotes the Dirac impulse on $\mathbb{R}^d$, $\{v_k\}_{k=1}^K$ is a sequence of weights and the locations of Dirac impulses are at the spline knots $\{\boldsymbol{x}_k\}_{k=1}^K$.

Due to the similarities between the variational problem in (11) and our variational problem in (9), we can similarly define the notion of a (polynomial) *ridge spline*. Before stating this definition, we remark that in this paper we will be working with Dirac impulses on different domains. For clarity, we will subscript the "δ" with the appropriate domain, e.g., $\delta_{\mathbb{R}^d}$ denotes the Dirac impulse on $\mathbb{R}^d$ and $\delta_{\mathbb{S}^{d-1}\times\mathbb{R}}$ denotes the Dirac impulse on $\mathbb{S}^{d-1} \times \mathbb{R}$.

Definition 5 (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*

$$\mathrm{R}_m\{s\} = \sum_{k=1}^{K} v_k \left[\frac{\delta_{\mathbb{S}^{d-1}\times\mathbb{R}}(\cdot - \boldsymbol{z}_k) + (-1)^m \delta_{\mathbb{S}^{d-1}\times\mathbb{R}}(\cdot + \boldsymbol{z}_k)}{2} \right], \tag{12}$$

where $\{v_k\}_{k=1}^K$ is a sequence of weights and the locations of the Dirac impulses are at $\boldsymbol{z}_k = (\boldsymbol{w}_k, b_k) \in \mathbb{S}^{d-1} \times \mathbb{R}$. The collection $\{\boldsymbol{z}_k\}_{k=1}^K$ can be viewed as a collection of Radon domain spline knots.

Remark 6 *The reason*

$$\frac{\delta_{\mathbb{S}^{d-1}\times\mathbb{R}}(\cdot - \boldsymbol{z}_k) + (-1)^m \delta_{\mathbb{S}^{d-1}\times\mathbb{R}}(\cdot + \boldsymbol{z}_k)}{2} \tag{13}$$

appears in (12) *rather than $\delta_{\mathbb{S}^{d-1}\times\mathbb{R}}(\cdot - \boldsymbol{z}_k)$ is due to the fact that the operator R_m maps functions $f \in \mathcal{F}_m$ to even (respectively odd) elements of $\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})$ when m is even (respectively odd). Thus,* (13) *can be viewed as an even or odd version of the normal translated Dirac impulse in the sense that when acting on even or odd test functions defined on $\mathbb{S}^{d-1} \times \mathbb{R}$, it is the point evaluation operator.*

Remark 7 *When s is a neural network as in* (10)*, we have that* (12) *holds. The way to understand this is that the neurons in* (1) *are "sparsified" by R_m in the sense that*

$$\mathrm{R}_m \, r^{(m)}_{(\boldsymbol{w},b)} = \frac{\delta_{\mathbb{S}^{d-1}\times\mathbb{R}}(\cdot - \boldsymbol{z}_k) + (-1)^m \delta_{\mathbb{S}^{d-1}\times\mathbb{R}}(\cdot + \boldsymbol{z}_k)}{2},$$

where $r^{(m)}_{(\boldsymbol{w},b)}(\boldsymbol{x}) := \rho_m(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)$, $(\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$. *We show that this is true in Lemma 17. In other words,* $r^{(m)}_{(\boldsymbol{w},b)}$ *can be viewed as a translated* Green's function *of* R_m*, where the translation is in the Radon domain.*

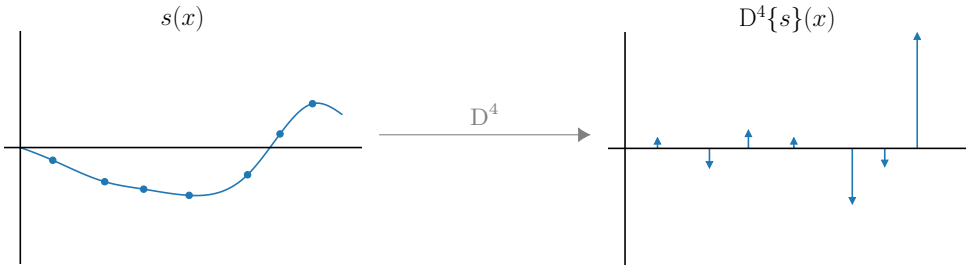

Figure 1: In the left plot we have a cubic spline with 7 knots. After applying D^4, the fourth derivative operator, we are left with 7 Dirac impulses as seen in the right plot.

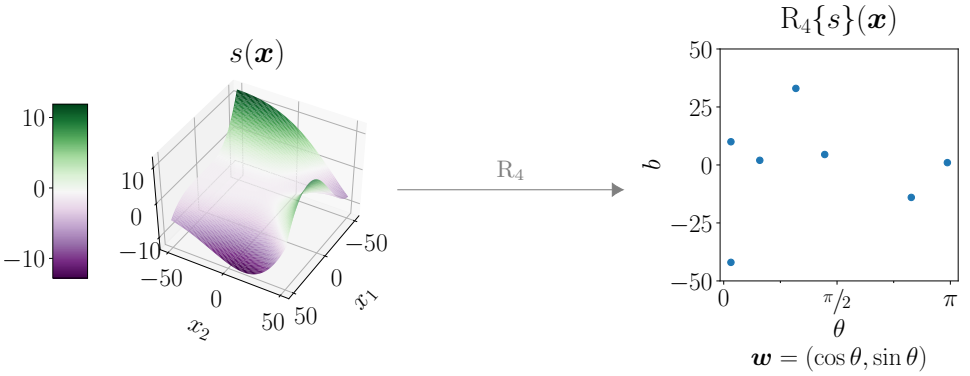

Figure 2: In the left plot we have a two-dimensional cubic ridge spline with 7 neurons. After applying R_4, we are left with 7 Dirac impulses in the Radon domain, which are designated by the dots in the right plot. We have parameterized the directions in the Radon domain by $\theta \in [0, \pi)$. This parameterization of the two-dimensional Radon domain is known as a *sinogram*. This parameterization of the Radon domain eliminates the two impulses per neuron we see in (12) as $\theta \in [0, \pi)$ is only "half" of the unit circle $\mathbb{S}^1$.

.

We illustrate the sparsifying effect of the operator L in the case of cubic splines, i.e., $\mathrm{L} = \mathrm{D}^4$, 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. We also remark that 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.

We finally remark that when m is even, we have the equality

$$\mathrm{R}_m = c_d \, \Lambda^{d-1} \, \mathscr{R} \, \Delta^{m/2},$$

by the intertwining relations of the Radon transform and the Laplacian, which we later discuss in (30). This provides another way to understand how R_m sparsifies ridge splines. We illustrate this in the $m = 2$ (i.e., ReLU network) case in Figure 3.

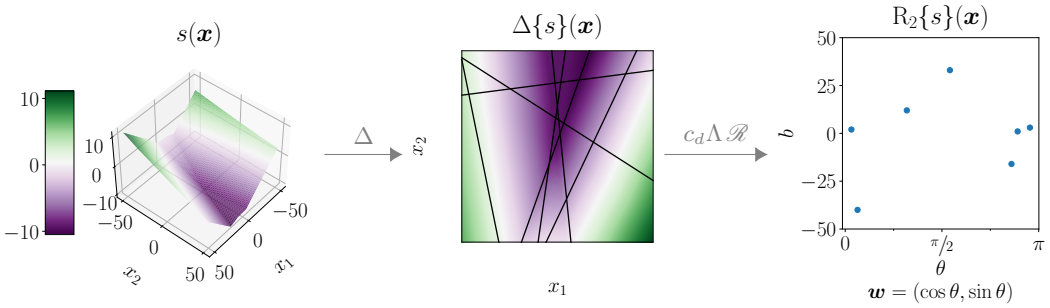

Figure 3: On the left plot we have a two-dimensional linear ridge spline (single-hidden layer ReLU network) with 7 neurons. After applying Δ, we get an "impulse sheet", i.e., a mapping of the form $\boldsymbol{x} \mapsto \delta_{\mathbb{R}}(\boldsymbol{w}_k^{\mathsf{T}} \boldsymbol{x} - b_k)$, for each neuron, designated by the black lines in the top down view of the linear 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. Just as in Figure 2, we have parameterized the directions in the Radon domain by $\theta \in [0, \pi)$, eliminating the two impulses per neuron we see in (12).

2.3 Scattered data approximation and neural network training

Since Theorem 1 says that a single-hidden layer neural network as in (10) is a solution to the continuous-domain inverse problem in (9), we can recast the continuous-domain problem in (9) as the *finite-dimensional neural network training* problem

$$\min_{\boldsymbol{\theta} \in \Theta} \; G(\mathcal{V} f_{\boldsymbol{\theta}}) + \|\mathrm{R}_m \, f_{\boldsymbol{\theta}}\|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})}, \tag{14}$$

so long as the number of neurons K is large enough[4] ($K \geq N$ suffices, giving insight into the efficacy of overparameterization in neural network models), where

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

4. We characterize what large enough means in Theorem 8.

where $\boldsymbol{\theta} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K, v_1, \ldots, v_K, b_1, \ldots, b_K, c)$ contains the neural network parameters and Θ is the collection of all $\boldsymbol{\theta}$ such that $v_k \in \mathbb{R}$, $\boldsymbol{w}_k \in \mathbb{R}^d$, 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 25 that

$$\|\mathrm{R}_m f_{\boldsymbol{\theta}}\|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})} = \sum_{k=1}^{K} |v_k| \|\boldsymbol{w}_k\|_2^{m-1},$$

and then use this fact to show that (14) is equivalent to two neural network training problems, with variants of well-known neural network regularizers, in the following proposition.

Theorem 8 *The solutions to the finite-dimensional optimization in* (14) *are solutions to optimization in* (9) *so long as* $K \geq N - \dim \mathcal{N}_m$*. Additionally, the optimization in* (14) *is equivalent to*

$$\min_{\boldsymbol{\theta} \in \Theta} G(\mathcal{V} f_{\boldsymbol{\theta}}) + \sum_{k=1}^{K} |v_k| \|\boldsymbol{w}_k\|_2^{m-1}, \tag{15}$$

for any $K \in \mathbb{N}$*. Furthermore, the solutions to*

$$\min_{\boldsymbol{\theta} \in \Theta} G(\mathcal{V} f_{\boldsymbol{\theta}}) + \frac{1}{2} \sum_{k=1}^{K} \Big(|v_k|^2 + \|\boldsymbol{w}_k\|_2^{2m-2} \Big) \tag{16}$$

are also solutions to (15) *for any* $K \in \mathbb{N}$*. Finally, for both problems in* (15) *and* (16)*, for* K_1 *and* K_2 *such that* $K_1 > K_2$*, a global minimizer when* $K = K_2$ *will always be a global minimizer when* $K = K_1$*.*

When $m = 2$, which coincides with neural networks with ReLU activation functions, (15) and (16) correspond to previously studied training problems. The regularizer in (15) coincides with the notion of the ℓ^1*-path-norm* regularization as proposed in Neyshabur et al. (2015) and the regularizer in (16) 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 (15) and (16) 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 *ideal sampling*, i.e., the measurement operator $\mathcal{V}$ acts on a function $f : \mathbb{R}^d \to \mathbb{R}$ via

$$\mathcal{V} : \mathcal{F}_m \ni f \mapsto \begin{bmatrix} \langle \delta_{\mathbb{R}^d}(\cdot - \boldsymbol{x}_1), f \rangle \\ \vdots \\ \langle \delta_{\mathbb{R}^d}(\cdot - \boldsymbol{x}_N), f \rangle \end{bmatrix} = \begin{bmatrix} f(\boldsymbol{x}_1) \\ \vdots \\ f(\boldsymbol{x}_N) \end{bmatrix} \in \mathbb{R}^N, \tag{17}$$

so the problem is to approximate the scattered data $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N \subset \mathbb{R}^d \times \mathbb{R}$. For the above $\mathcal{V}$ to be a valid measurement operator for Theorem 1, it must be continuous.

Lemma 9 *The operator* $\mathcal{V} : \mathcal{F}_m \to \mathbb{R}^N$ *defined in* (17) *is continuous.*

The proof of Lemma 9 appears in Appendix A. Lemma 9 also says that $\mathcal{F}_m$ is a *reproducing kernel Banach space*.

By choosing an appropriate data fitting term G, the generality of our main result in Theorem 1 says that the solutions problems with interpolation constraints

$$\min_{f \in \mathcal{F}_m} \|\mathrm{R}_m f\|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})} \quad \text{s.t.} \quad f(\boldsymbol{x}_n) = y_n,\ n = 1, \ldots, N$$

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

$$\min_{f \in \mathcal{F}_m} \sum_{n=1}^{N} \ell(f(\boldsymbol{x}_n), y_n) + \lambda \|\mathrm{R}_m f\|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})}, \tag{18}$$

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 Theorem 8 to recast the problem in (18) with either of the equivalent finite-dimensional neural network training problems:

$$\begin{aligned} &\min_{\boldsymbol{\theta} \in \Theta} \sum_{n=1}^{N} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_n), y_n) + \lambda \sum_{k=1}^{K} |v_k| \|\boldsymbol{w}_k\|_2^{m-1} \\ &\min_{\boldsymbol{\theta} \in \Theta} \sum_{n=1}^{N} \ell(f_{\boldsymbol{\theta}}(\boldsymbol{x}_n), y_n) + \frac{\lambda}{2} \sum_{k=1}^{K} \Big(|v_k|^2 + \|\boldsymbol{w}_k\|_2^{2m-2} \Big), \end{aligned} \tag{19}$$

so long as the number of neurons K is large enough as stated in Theorem 8. The two problems in the above display correspond to how neural network training problems are actually set up.

2.4 Statistical generalization bounds

Neural networks are widely used for pattern classification. In the ideal sampling scenario, the generality of Theorem 1 allows us to consider optimizations of the form

$$\min_{f \in \mathcal{F}_m} \sum_{n=1}^{N} \ell\big(y_n f(\boldsymbol{x}_n)\big) \quad \text{s.t.} \quad \|\mathrm{R}_m f\|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})} \leq B, \tag{20}$$

for some constant $B < \infty$, where $\ell(\cdot)$ is an appropriate L-Lipschitz loss function of the product $y_n f(\boldsymbol{x}_n)$. If we assume that $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ 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 $\mathrm{sgn}(f(\boldsymbol{x}_n)) \neq y_n$ (or equivalently when $y_n f(\boldsymbol{x}_n) < 0$), this is the *binary classification* setting. Given this set up, it is natural to examine if solutions to (20) predict well on new random examples $(\boldsymbol{x}, y)$ drawn independently from the same underlying distribution.

We can invoke Theorem 8 and consider optimization over neural network parameters by considering the recast optimization to (20)

$$\min_{\boldsymbol{\theta} \in \Theta} \sum_{n=1}^{N} \ell\big(y_n f_{\boldsymbol{\theta}}(\boldsymbol{x}_n)\big) \quad \text{s.t.} \quad \sum_{k=1}^{K} |v_k| \|\boldsymbol{w}_k\|_2^{m-1} \leq B, \tag{21}$$

where

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

In particular, the solution to (21) is known as an *Ivanov estimator* (Oneto et al., 2016) which is equivalent to the solution to (19) for a particular choice of λ which depends on B and the data through the data fitting term. In this section we provide a *generalization bound* for the Ivanov estimator.

Let $\bar{f}$ be a minimizer of the optimization in (21). We show that B directly controls the error probability of $\bar{f}$, i.e., $\mathbb{P}\big(y\bar{f}(\boldsymbol{x}) < 0\big)$, where $(\boldsymbol{x}, y)$ 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).

Let $\mathcal{F}$ be a hypothesis space. For every $f \in \mathcal{F}$, define its *risk* and *empirical risk*

$$R(f) \coloneqq \mathbb{E}\big[\ell\big(yf(\boldsymbol{x})\big)\big] \quad \text{and} \quad \widehat{R}_N(f) \coloneqq \frac{1}{N}\sum_{n=1}^{N} \ell\big(y_n f(\boldsymbol{x}_n)\big),$$

and assume the loss function satisfies $0 \leq \ell\big(y_n f(\boldsymbol{x}_n)\big) \leq C_0$ almost surely, for $n = 1, \ldots, N$ and some constant $C_0 < \infty$. Then, for every $f \in \mathcal{F}$, we have the following generalization bound. With probability at least $1 - \delta$,

$$R(f) \leq \widehat{R}_N(f) + L\,\mathfrak{R}(\mathcal{F}) + C_0\sqrt{\frac{\log(1/\delta)}{2N}},$$

where we use the fact that the loss ℓ is L-Lipschitz and $\mathfrak{R}(\mathcal{F})$ is the *Rademacher complexity* of the class $\mathcal{F}$ defined via

$$\mathfrak{R}(\mathcal{F}) \coloneqq 2\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{N}\sum_{n=1}^{N} \sigma_n f(\boldsymbol{x}_n)\right], \tag{22}$$

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, $(1 - yf(\boldsymbol{x}))^2$, or hinge loss, $\max\{0, 1 - yf(\boldsymbol{x})\}$), then we may use this to bound the probability of error $\mathbb{P}\big(y\bar{f}(\boldsymbol{x}) < 0\big) \leq R(\bar{f})$.

To provide a generalization bound for the minimizer $\bar{f}$ of (20), we assume that the empirical data satisfies $\|\boldsymbol{x}_n\|_2 \leq C/2$ almost surely for $n = 1, \ldots, N$ and some constant $C < \infty$ and consider the hypothesis space

$$\mathcal{F}_\Theta \coloneqq \left\{ f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta,\ \sum_{k=1}^{K} |v_k| \|\boldsymbol{w}_k\|_2^{m-1} \leq B,\ |b_k| \leq \frac{C}{2}, k = 1, \ldots, K,\ K \geq 0 \right\}.$$

The reason we impose that $|b_k| \leq C/2$, $k = 1, \ldots, K$, is because we will later show in Lemma 24 that all solutions to (20) satisfy

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

for every $k = 1, \ldots, K$. We bound the Rademacher complexity of $\mathcal{F}_\Theta$ in the following theorem.

Theorem 10 *Assume that* $\|\boldsymbol{x}_n\|_2 \leq C/2$ *almost surely for* $n = 1, \ldots, N$ *and some constant* $C < \infty$. *Then,*

$$\mathfrak{R}(\mathcal{F}_\Theta) \leq \frac{2BC^{m-1}}{\sqrt{N}(m-1)!} + \mathfrak{R}(c),$$

where $\mathfrak{R}(c)$ *denotes the Rademacher complexity of the polynomial terms* $c(\boldsymbol{x})$ *that appear in the solutions to the optimization in* (21).

Remark 11 *Theorem 10 shows that the Rademacher complexity, and hence the generalization error, is controlled by bounding the seminorm* $\|\mathrm{R}_m f\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} \leq B$. *In practice, neural networks are typically implemented without the polynomial term* $c(\boldsymbol{x})$, *in which case the same bound holds without* $\mathfrak{R}(c)$.

3. Preliminaries & Notation

In this section we will introduce the mathematical formulation and notation used in the remainder of the paper.

3.1 Spaces of functions and distributions

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

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

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

We will later see in Section 3.4 that in order to define the Radon transform of distributions, we will be interested in the *Lizorkin test functions* $\mathscr{S}_0(\mathbb{R}^d)$ of highly time-frequency localized functions over $\mathbb{R}^d$ (Holschneider, 1995). This is a closed subspace of $\mathscr{S}(\mathbb{R}^d)$ consisting of functions with all moments equal to 0, i.e., $\varphi \in \mathscr{S}_0(\mathbb{R}^d)$ when $\varphi \in \mathscr{S}(\mathbb{R}^d)$ and

$$\int_{\mathbb{R}^d} \boldsymbol{x}^{\boldsymbol{\alpha}} \varphi(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 0,$$

for every multi-index $\boldsymbol{\alpha}$. We can then define the space of *Lizorkin distributions*, $\mathscr{S}_0'(\mathbb{R}^d)$, 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 $\mathcal{P}(\mathbb{R}^d)$ is the space of polynomials on $\mathbb{R}^d$, then $\mathscr{S}_0'(\mathbb{R}^d) \cong \mathscr{S}'(\mathbb{R}^d)/\mathcal{P}(\mathbb{R}^d)$ (Yuan et al., 2010, Chapter 8). Just as above, we can define the Lizorkin

test functions on $\mathbb{S}^{d-1} \times \mathbb{R}$ as $\mathscr{S}_0(\mathbb{S}^{d-1} \times \mathbb{R}) = \mathcal{D}(\mathbb{S}^{d-1}) \mathbin{\widehat{\otimes}} \mathscr{S}_0(\mathbb{R})$ and the space of Lizorkin distributions on $\mathbb{S}^{d-1} \times \mathbb{R}$ as its continuous dual, $\mathscr{S}_0'(\mathbb{S}^{d-1} \times \mathbb{R})$.

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

$$\|u\|_{\mathcal{M}(X)} \coloneqq \sup_{\substack{\varphi \in C_0(X) \\ \|\varphi\|_\infty = 1}} \langle u, \varphi \rangle. \tag{23}$$

The norm $\|\cdot\|_{\mathcal{M}(X)}$ is exactly the *total variation* norm (in the sense of measures). As $\mathscr{S}_0(X)$ is dense in $C_0(X)$ (cf. Samko, 1995), we can associate every measure in $\mathcal{M}(X)$ with a Lizorkin distribution and view $\mathcal{M}(X) \subset \mathscr{S}_0'(X) \subset \mathscr{S}'(X)$, providing the description

$$\mathcal{M}(X) \coloneqq \Big\{ u \in \mathscr{S}_0'(X) : \ \|u\|_{\mathcal{M}(X)} < \infty \Big\},$$

and so the duality pairing $\langle \cdot, \cdot \rangle$ in (23) can be viewed, formally, as the integral

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

where u is viewed as an element of $\mathscr{S}_0'(X)$. In this paper, we will be mostly be working with $X = \mathbb{S}^{d-1} \times \mathbb{R}$, the Radon domain.

As we will later see in Section 3.4, the Radon transform of a function is necessarily even, so we will be interested in the Banach spaces of odd and even finite Radon measures on $\mathbb{S}^{d-1} \times \mathbb{R}$. Viewing $\mathcal{M}(X) \subset \mathscr{S}_0'(X)$, put

$$\begin{aligned} \mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1} \times \mathbb{R}) &\coloneqq \Big\{ u \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R}) : \ u(\boldsymbol{\gamma}, t) = u(-\boldsymbol{\gamma}, -t) \Big\} \\ \mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1} \times \mathbb{R}) &\coloneqq \Big\{ u \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R}) : \ u(\boldsymbol{\gamma}, t) = -u(-\boldsymbol{\gamma}, -t) \Big\}. \end{aligned}$$

One can then verify that the predual of $\mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1} \times \mathbb{R})$ is the subspace of odd functions in $C_0(\mathbb{S}^{d-1} \times \mathbb{R})$ and the predual of $\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$ is the subspace of even functions in $C_0(\mathbb{S}^{d-1} \times \mathbb{R})$, and so the associated norms of $\mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1} \times \mathbb{R})$ and $\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$ can be defined accordingly.

Finally, $\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$ can equivalently be viewed as $\mathcal{M}(\mathbb{P}^d)$ where $\mathbb{P}^d$ denotes the manifold of hyperplanes in $\mathbb{R}^d$. This follows from the fact that every hyperplane in $\mathbb{R}^d$ takes the form $h_{(\boldsymbol{\gamma}, t)} \coloneqq \big\{ \boldsymbol{x} \in \mathbb{R}^d : \ \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{x} = t \big\}$ for some $(\boldsymbol{\gamma}, t) \in \mathbb{S}^{d-1} \times \mathbb{R}$ and $h_{(\boldsymbol{\gamma}, t)} = h_{(-\boldsymbol{\gamma}, -t)}$. It will sometimes be convenient to work with $\mathcal{M}(\mathbb{P}^d)$ instead of $\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$ as $\mathbb{P}^d$ is a locally compact Hausdorff space.

3.2 The Fourier transform

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

$$\mathscr{F}\{f\}(\boldsymbol{\xi}) \coloneqq \int_{\mathbb{R}^d} f(\boldsymbol{x}) e^{-\mathrm{i}\boldsymbol{x}^\mathsf{T}\boldsymbol{\xi}} \,\mathrm{d}\boldsymbol{x}, \quad \boldsymbol{\xi} \in \mathbb{R}^d$$
$$\mathscr{F}^{-1}\{F\}(\boldsymbol{x}) \coloneqq \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{\mathrm{i}\boldsymbol{x}^\mathsf{T}\boldsymbol{\xi}} F(\boldsymbol{\xi}) \,\mathrm{d}\boldsymbol{\xi}, \quad \boldsymbol{x} \in \mathbb{R}^d,$$

where $\mathrm{i}^2 = -1$. We will usually write $\widehat{\cdot}$ for $\mathscr{F}\{\cdot\}$. The Fourier transform and its inverse can be applied to functions in $\mathscr{S}(\mathbb{R}^d)$, resulting in functions in $\mathscr{S}(\mathbb{R}^d)$. These transforms can be extended to act on $\mathscr{S}'(\mathbb{R}^d)$ by duality.

3.3 The Hilbert transform

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

$$\mathscr{H}\{f\}(x) \coloneqq \frac{\mathrm{i}}{\pi} \mathrm{p.v.} \int_{-\infty}^{\infty} \frac{f(y)}{x - y} \,\mathrm{d}y, \quad x \in \mathbb{R},$$

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

$$\widehat{\mathscr{H}\{f\}}(\omega) = (\operatorname{sgn}\omega)\widehat{f}(\omega) \quad \text{and} \quad \mathscr{H}\,\mathscr{H}\,f = f.$$

The Hilbert transform can be applied to functions in $\mathscr{S}(\mathbb{R}^d)$, resulting in functions in $\mathscr{S}(\mathbb{R}^d)$. This transform can be extended to act on $\mathscr{S}'(\mathbb{R}^d)$ by duality.

3.4 The Radon transform

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

$$\mathscr{R}\{f\}(\boldsymbol{\gamma}, t) \coloneqq \int_{\{\boldsymbol{x}:\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x}=t\}} f(\boldsymbol{x}) \,\mathrm{d}s(\boldsymbol{x}), \quad (\boldsymbol{\gamma}, t) \in \mathbb{S}^{d-1} \times \mathbb{R} \tag{24}$$
$$\mathscr{R}^*\{\Phi\}(\boldsymbol{x}) \coloneqq \int_{\mathbb{S}^{d-1}} \Phi(\boldsymbol{\gamma}, \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x}) \,\mathrm{d}\sigma(\boldsymbol{\gamma}), \quad \boldsymbol{x} \in \mathbb{R}^d,$$

where s denotes the surface measure on the plane $\{\boldsymbol{x} : \boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x} = t\}$, and σ denotes the surface measure on $\mathbb{S}^{d-1}$. We will sometimes write $\boldsymbol{z} = (\boldsymbol{\gamma}, t)$ as the variable in the Radon domain. We discuss the spaces which we can apply the Radon transform and its dual in the sequel. We also remark that the Radon transform of a function is always even, i.e., $\mathscr{R}\{f\}(\boldsymbol{\gamma}, t) = \mathscr{R}\{f\}(-\boldsymbol{\gamma}, -t)$.

Another way to view the Radon transform and its dual is to consider, formally, the integrals

$$\mathscr{R}\{f\}(\boldsymbol{\gamma}, t) = \int_{\mathbb{R}^d} f(\boldsymbol{x}) \delta_{\mathbb{R}}(\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x} - t) \,\mathrm{d}\boldsymbol{x}, \quad (\boldsymbol{\gamma}, t) \in \mathbb{S}^{d-1} \times \mathbb{R} \tag{25}$$
$$\mathscr{R}^*\{\Phi\}(\boldsymbol{x}) = \int_{\mathbb{S}^{d-1}\times\mathbb{R}} \delta_{\mathbb{R}}(\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x} - t)\Phi(\boldsymbol{\gamma}, t) \,\mathrm{d}(\sigma \times \lambda)(\boldsymbol{\gamma}, t), \quad \boldsymbol{x} \in \mathbb{R}^d, \tag{26}$$

where λ denotes the univariate Lebesgue measure.

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

$$2(2\pi)^{d-1} f = \mathscr{R}^* \, \Lambda^{d-1} \, \mathscr{R} \, f, \tag{27}$$

where the *ramp filter* Λ^d of a function $\Phi(\boldsymbol{\gamma}, t)$ is given by

$$\Lambda^d\{\Phi\}(\boldsymbol{\gamma}, t) \coloneqq \begin{cases} \partial_t^d \Phi(\boldsymbol{\gamma}, t), & d \text{ even} \\ \mathscr{H}_t \, \partial_t^d \Phi(\boldsymbol{\gamma}, t), & d \text{ odd}, \end{cases} \tag{28}$$

where $\mathscr{H}_t$ is the Hilbert transform (in the variable t) and ∂_t is the partial derivative with respect to t. It is easier to see that Λ^d is indeed a ramp filter by looking at its frequency response with respect to the t variable. We have

$$\widehat{\Lambda^d \Phi}(\boldsymbol{\gamma}, \omega) = \mathrm{i}^d |\omega|^d \widehat{\Phi}(\boldsymbol{\gamma}, \omega).$$

Some care has to be taken to understand the Radon transforms of distributions. Just as the Fourier and Hilbert transforms can be extended to distributions via duality, we do the same with the Radon transform, though some care has to be taken. In particular, the choice of test functions must be carefully chosen and cannot be the space of Schwartz functions. It is easy to verify that if $\varphi \in \mathscr{S}(\mathbb{R}^d)$, then $\mathscr{R}\{\varphi\} \in \mathscr{S}(\mathbb{S}^{d-1} \times \mathbb{R})$. This is not true about the dual transform. Indeed, if $\psi \in \mathscr{S}(\mathbb{S}^{d-1} \times \mathbb{R})$, then it may not be true that $\mathscr{R}^*\{\psi\} \in \mathscr{S}(\mathbb{R}^d)$.

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.

Proposition 12 (Kostadinova et al. (2014, Corollary 6.1)) *The transforms*

$$\mathscr{R} : \mathscr{S}_0(\mathbb{R}^d) \to \mathscr{S}_0(\mathbb{P}^d)$$
$$\mathscr{R}^* : \mathscr{S}_0(\mathbb{P}^d) \to \mathscr{S}_0(\mathbb{R}^d)$$

are continuous bijections, where $\mathscr{S}_0(\mathbb{P}^d) \subset \mathscr{S}_0(\mathbb{S}^{d-1} \times \mathbb{R})$ denotes the subspace of even Lizorkin test functions.

Proposition 12 allows us to define the Radon transform and dual Radon transform of distributions by duality by choosing our test functions to be Lizorkin test functions, i.e., the action of the Radon transform of $f \in \mathscr{S}_0'(\mathbb{R}^d)$ on $\psi \in \mathscr{S}_0(\mathbb{P}^d)$ is defined to be $\langle \mathscr{R} \, f, \psi \rangle \coloneqq \langle f, \mathscr{R}^* \, \psi \rangle$, and the action of the dual Radon transform of $\Phi \in \mathscr{S}_0'(\mathbb{P}^d)$ on $\varphi \in \mathscr{S}_0(\mathbb{R}^d)$ is defined to be $\langle \mathscr{R}^* \, \Phi, \varphi \rangle \coloneqq \langle \Phi, \mathscr{R} \, \varphi \rangle$. This means we have the following corollary to Proposition 12.

Corollary 13 *The transforms*

$$\mathscr{R} : \mathscr{S}_0'(\mathbb{R}^d) \to \mathscr{S}_0'(\mathbb{P}^d)$$
$$\mathscr{R}^* : \mathscr{S}_0'(\mathbb{P}^d) \to \mathscr{S}_0'(\mathbb{R}^d)$$

are continuous bijections.

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

$$2(2\pi)^{d-1}\Phi = \Lambda^{d-1}\,\mathscr{R}\,\mathscr{R}^*\,\Phi. \tag{29}$$

The inversion formulas in (27) and (29) 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

$$(-\Delta)^{\frac{d-1}{2}}\,\mathscr{R}^* = \mathscr{R}^*\,\Lambda^{d-1} \qquad \text{and} \qquad \mathscr{R}(-\Delta)^{\frac{d-1}{2}} = \Lambda^{d-1}\,\mathscr{R}\,. \tag{30}$$

As the constant $2(2\pi)^{d-1}$ arises often when working with the Radon transform, we put

$$c_d \coloneqq \frac{1}{2(2\pi)^{d-1}}. \tag{31}$$

Remark 14 *We warn the reader that 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 different notation for different pairings to reduce clutter. The exact pairings should be clear from context.*

With these definitions in hand, we see that the seminorms (6) studied in this paper are well-defined. Recall from (6) that

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

where $f \in \mathcal{F}_m$. We will later show in Lemma 19 that the null space $\mathcal{N}_m$ of R_m is the space of polynomials of degree strictly less than m. Thus, to understand that the seminorms studied in this paper are well-defined, we can view $f \in \mathscr{S}_0'(\mathbb{R})$. From Corollary 13, it follows that $\mathscr{R}\,f \in \mathscr{S}_0'(\mathbb{P}^d) \subset \mathscr{S}_0'(\mathbb{S}^{d-1}\times\mathbb{R}) \subset \mathscr{S}'(\mathbb{S}^{d-1}\times\mathbb{R})$. Since $\partial_t^m\Lambda^{d-1}$ is a Fourier multiplier and from the definition of $\mathcal{F}_m$, we see that $\partial_t^m\Lambda^{d-1}\,\mathscr{R}\,f \in \mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})$, and so the seminorms are well-defined.

4. The Representer Theorem

In this section we will prove Theorem 1, our representer theorem. The general strategy will be to reduce the problem in (9) 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

$$\min_{u\in\mathcal{M}(\Omega)} \|u\|_{\mathcal{M}(\Omega)} \quad \text{s.t.} \quad \mathcal{V}u = \boldsymbol{y}, \tag{32}$$

where $\mathcal{V} : \mathcal{M}(\Omega) \to \mathbb{R}^N$ is a continuous linear operator and $\boldsymbol{y} \in \mathbb{R}^N$. The first "representer theorem" for (32) is from Zuhovickiĭ (1948). This representer theorem essentially states that there exists a sparse solution to (32) of the form

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

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), and (Bredies and Carioni, 2020, Theorem 4.2). For our problem, we reduce (9) to one of the following Radon measure recovery problems in Propositions 15 and 16, depending on if m is even or odd. Reducing our problem to one of these Radon measure recovery problems requires several steps. The first step is to understand what functions are sparsified by R_m. The next step is the construction of a stable right-inverse of R_m. The final step is to understand the topological structure of the native space $\mathcal{F}_m$. These are outlined in the remainder of this section before finally proving Theorem 1 at the end of this section.

Proposition 15 (Based on Bredies and Carioni (2020, Theorem 4.2)) *Assume the following:*

1. *The function $G : \mathbb{R}^N \to \mathbb{R}$ is strictly convex, coercive, and lower semi-continuous.*
2. *The operator $\mathcal{V} : \mathcal{M}(\mathbb{P}^d) \to \mathbb{R}^N$ is continuous, linear, and surjective, where we recall that $\mathbb{P}^d$ is the manifold of hyperplanes in $\mathbb{R}^d$.*

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

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

of the form

$$\bar{u} = \sum_{k=1}^{K} v_k \, \delta_{\mathbb{P}^d}(\cdot - \boldsymbol{z}_k) = \sum_{k=1}^{K} v_k \left[\frac{\delta_{\mathbb{S}^{d-1} \times \mathbb{R}}(\cdot - \boldsymbol{z}_k) + \delta_{\mathbb{S}^{d-1} \times \mathbb{R}}(\cdot + \boldsymbol{z}_k)}{2} \right]$$

with $K \leq N$, $v_k \in \mathbb{R} \setminus \{0\}$, and $\boldsymbol{z}_k = (\boldsymbol{w}_k, b_k) \in \mathbb{S}^{d-1} \times \mathbb{R}$, $k = 1, \ldots, K$.

Although Bredies and Carioni (2020, Theorem 4.2) considers an open, bounded subset of $\mathbb{R}^d$ rather than $\mathbb{P}^d$, their proof is general enough to apply to any locally compact Hausdorff space, and, in particular, for $\mathbb{P}^d$. Analogously, we immediately have the following result for the Radon measure recovery problem posed over $\mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1} \times \mathbb{R})$.

Proposition 16 *Assume the following:*

1. *The function $G : \mathbb{R}^N \to \mathbb{R}$ is a strictly convex function that is coercive and lower semi-continuous.*
2. *The operator $\mathcal{V} : \mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1} \times \mathbb{R}) \to \mathbb{R}^N$ is continuous, linear, and surjective.*

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

$$\min_{u \in \mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1} \times \mathbb{R})} G(\mathcal{V}u) + \|u\|_{\mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1} \times \mathbb{R})}$$

of the form

$$\bar{u} = \sum_{k=1}^{K} v_k \left[\frac{\delta_{\mathbb{S}^{d-1} \times \mathbb{R}}(\cdot - \boldsymbol{z}_k) - \delta_{\mathbb{S}^{d-1} \times \mathbb{R}}(\cdot + \boldsymbol{z}_k)}{2} \right]$$

with $K \leq N$, $v_k \in \mathbb{R} \setminus \{0\}$, and $\boldsymbol{z}_k = (\boldsymbol{w}_k, b_k) \in \mathbb{S}^{d-1} \times \mathbb{R}$, $k = 1, \ldots, K$.

In order to reduce (9) to one of the problems in Propositions 15 and 16, we need to understand which functions are sparsified by R_m. As claimed in Remark 7, these are exactly the neurons in (10).

Lemma 17 *The atoms of the single-hidden layer neural network as in* (1) *are "sparsified" by* R_m *in the sense that*

$$\mathrm{R}_m\, r^{(m)}_{(\boldsymbol{w},b)} = \frac{\delta_{\mathbb{S}^{d-1}\times\mathbb{R}}(\cdot - \boldsymbol{z}) + (-1)^m \delta_{\mathbb{S}^{d-1}\times\mathbb{R}}(\cdot + \boldsymbol{z})}{2},$$

where $\boldsymbol{z} = (\boldsymbol{w}, b) \in \mathbb{S}^{d-1} \times \mathbb{R}$ *and* $r^{(m)}_{(\boldsymbol{w},b)}(\boldsymbol{x}) \coloneqq \rho_m(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)$.

Before proving Lemma 17, we first prove the following intermediary result which may be of independent interest.

Lemma 18 *Let* $r^{(m)}_{(\boldsymbol{\gamma},t)}(\boldsymbol{x}) \coloneqq \rho_m(\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x} - t)$. *For all* $\varphi \in \mathscr{S}_0(\mathbb{R}^d)$ *and any* $m \in \mathbb{N}$ *we have*

$$\left\langle r^{(m)}_{(\boldsymbol{\gamma},t)}, \varphi \right\rangle = (-1)^m (\partial_t^{-m})\, \mathscr{R}\{\varphi\}(\boldsymbol{\gamma}, t),$$

where ∂_t^{-m} *is a stable left-inverse of* ∂_t^m, *i.e., an* m*-fold integrator in the* t *variable of the Radon domain. In particular, by the intertwining relations of the Radon transform with the Laplacian operator in* (30), *this says for even* m

$$\left\langle \Delta^{m/2} r^{(m)}_{(\boldsymbol{\gamma},t)}, \varphi \right\rangle = \mathscr{R}\{\varphi\}(\boldsymbol{\gamma}, t),$$

for all $\varphi \in \mathscr{S}_0(\mathbb{R}^d)$.

Proof The proof is a direct computation. For all $\varphi \in \mathscr{S}_0(\mathbb{R}^d)$ and any $m \in \mathbb{N}$ we have

$$\begin{aligned}
\left\langle r^{(m)}_{(\boldsymbol{\gamma},t)}, \varphi \right\rangle &= \int_{\mathbb{R}^d} \rho_m(\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x} - t)\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\
&= \int_{\mathbb{R}^d} (-1)^m (\partial_t^{-m}) \delta_{\mathbb{R}}(\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x} - t)\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\
&= (-1)^m (\partial_t^{-m}) \int_{\mathbb{R}^d} \delta_{\mathbb{R}}(\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{x} - t)\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \\
&= (-1)^m (\partial_t^{-m})\, \mathscr{R}\{\varphi\}(\boldsymbol{\gamma}, t).
\end{aligned}$$

In particular, the stability of ∂_t^{-m} ensures that the third line is well-defined[5]. The last line follows from (25). ∎

Proof [Proof of Lemma 17] One may verify that when m is even, $\mathrm{R}_m f$ is even and when m is odd, $\mathrm{R}_m f$ is odd. Next, as discussed in Remark 6, if we consider the subspace of even or odd Lizorkin distributions, then the point evaluation functionals (i.e., Dirac impulses) are

$$\frac{\delta_{\mathbb{S}^{d-1}\times\mathbb{R}}(\cdot - \boldsymbol{z}_k) + \delta_{\mathbb{S}^{d-1}\times\mathbb{R}}(\cdot + \boldsymbol{z}_k)}{2}$$

5. One choice of ∂_t^{-m} appears in Unser et al. (2017, pg. 781), where the (Schwartz) kernel of ∂_t^{-m} is compactly supported and bounded, which justifies changing the order of the operator and the integral.

for the even subspace and

$$\frac{\delta_{\mathbb{S}^{d-1}\times\mathbb{R}}(\cdot - \boldsymbol{z}_k) - \delta_{\mathbb{S}^{d-1}\times\mathbb{R}}(\cdot + \boldsymbol{z}_k)}{2}$$

for the odd subspace. Thus, it suffices to check that

$$c_d\Big\langle \partial_t^m \Lambda^{d-1}\, \mathscr{R}\, r^{(m)}_{(\boldsymbol{w},b)}, \psi \Big\rangle = \psi(\boldsymbol{w}, b) \tag{33}$$

for either all even Lizorkin test functions $\psi \in \mathscr{S}_0(\mathbb{S}^{d-1}\times\mathbb{R})$ if m is even or for all odd Lizorkin test functions $\psi \in \mathscr{S}_0(\mathbb{S}^{d-1}\times\mathbb{R})$ is m is odd. We remark that for the even (respectively odd) subspace, the pairing $\langle\cdot,\cdot\rangle$ in the above display is of even (respectively odd) Lizorkin test functions and even (respectively odd) Lizorkin distributions.

We now verify (33). Suppose that m is even. For all even $\psi \in \mathscr{S}_0(\mathbb{S}^{d-1}\times\mathbb{R})$ we have

$$\begin{aligned}
c_d\Big\langle \partial_t^m \Lambda^{d-1}\, \mathscr{R}\, r^{(m)}_{(\boldsymbol{w},b)}, \psi \Big\rangle &= c_d\Big\langle r^{(m)}_{(\boldsymbol{w},b)}, \mathscr{R}^*\, \Lambda^{d-1}\{(-1)^m \partial_t^m \psi\} \Big\rangle \\
&= \Big[(-1)^m \partial_t^{-m} c_d\, \mathscr{R}\, \mathscr{R}^*\, \Lambda^{d-1}\{(-1)^m \partial_t^m \psi\}\Big](\boldsymbol{w}, b) \\
&= \big[(-1)^m \partial_t^{-m} (-1)^m \partial_t^m \psi\big](\boldsymbol{w}, b) \\
&= \psi(\boldsymbol{w}, b),
\end{aligned}$$

where the first line holds since the formal adjoint of ∂_t^m is $(-1)^m\partial_t^m$, the second line holds via Lemma 18, the third line holds by the dual Radon transform inversion formula (29) and the intertwining relations (30) combined with the fact that since ψ is even and $\psi \in \mathscr{S}_0(\mathbb{S}^{d-1}\times\mathbb{R})$ we have that $\partial_t^m\psi$ is also even and $\partial_t^m\psi \in \mathscr{S}_0(\mathbb{S}^{d-1}\times\mathbb{R})$, and the fourth line holds by the left-inverse property for our choice of construction for ∂_t^{-m}. The case when m is odd is analogous, with the key fact that if ψ is odd and $\psi \in \mathscr{S}_0(\mathbb{S}^{d-1}\times\mathbb{R})$ we have that $\partial_t^m\psi$ is *even* and $\partial_t^m\psi \in \mathscr{S}_0(\mathbb{S}^{d-1}\times\mathbb{R})$, which justifies the use of the dual Radon transform inversion formula. ∎

Since R_m sparsifies functions of the form $r^{(m)}_{(\boldsymbol{w},b)} = \rho_m(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)$, we choose to define the growth restriction previously discussed in Section 2.1 for the null space and native space of R_m via the algebraic growth rate

$$n_0 \coloneqq \inf\Big\{ n \in \mathbb{N} :\ r^{(m)}_{(\boldsymbol{w},b)} \in L^{\infty,n}(\mathbb{R}^d) \Big\} = m - 1. \tag{34}$$

In (9), we are optimizing over the native space $\mathcal{F}_m$. Although $\mathcal{F}_m$ is defined by the seminorm $\|\mathrm{R}_m f\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})}$, we show in Theorem 22 that, 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 two intermediary results.

Lemma 19 *The null space $\mathcal{N}_m$ of R_m defined in (7) is finite-dimensional. In particular, it is the space of polynomials of degree strictly less than m.*

Lemma 19 says, in particular, that we can find a *biorthogonal system* for $\mathcal{N}_m$.

Definition 20 *Consider a finite-dimensional space $\mathcal{N}$ with $N_0 \coloneqq \dim \mathcal{N}$. The pair $(\boldsymbol{\phi}, \boldsymbol{p}) = \{(\phi_n, p_n)\}_{n=1}^{N_0}$ is called a* biorthogonal system *for $\mathcal{N}$ if $\{p_n\}_{n=1}^{N_0}$ is a basis of $\mathcal{N}$ and the vector of "boundary" functionals $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_{N_0})$ 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 $N_0 \coloneqq \dim \mathcal{N}_m$ and let $(\boldsymbol{\phi}, \boldsymbol{p})$ be a biorthogonal system for $\mathcal{N}_m$. Definition 20 says that any $q \in \mathcal{N}_m$ has the *unique representation*

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

We will sometimes write $\boldsymbol{\phi}(f)$ to denote the vector $(\langle \phi_1, f \rangle, \ldots, \langle \phi_{N_0}, f \rangle) \in \mathbb{R}^{N_0}$.

Lemma 21 *Let $(\boldsymbol{\phi}, \boldsymbol{p})$ be a biorthogonal system for $\mathcal{N}_m \subset \mathcal{F}_m \subset L^{\infty, m-1}(\mathbb{R}^d)$. Then, there exists a* unique operator

$$\mathrm{R}_{m,\boldsymbol{\phi}}^{-1} : \psi \mapsto \mathrm{R}_{m,\boldsymbol{\phi}}^{-1} \psi = \int_{\mathbb{S}^{d-1} \times \mathbb{R}} g_{m,\boldsymbol{\phi}}(\cdot, \boldsymbol{z}) \psi(\boldsymbol{z}) \, \mathrm{d}(\sigma \times \lambda)(\boldsymbol{z}),$$

where we recall that σ is the surface measure on $\mathbb{S}^{d-1}$ and λ is the univariate Lebesgue measure. Then, for all even $\psi \in \mathscr{S}_0(\mathbb{S}^{d-1} \times \mathbb{R})$ if m is even and all odd $\psi \in \mathscr{S}_0(\mathbb{S}^{d-1} \times \mathbb{R})$ if m is odd, the operator $\mathrm{R}_{m,\boldsymbol{\phi}}^{-1}$ satisfies

$$\begin{aligned} \mathrm{R}_m \, \mathrm{R}_{m,\boldsymbol{\phi}}^{-1} \psi &= \psi \qquad \textit{(right-inverse property)} \\ \boldsymbol{\phi}(\mathrm{R}_{m,\boldsymbol{\phi}}^{-1} \psi) &= \mathbf{0} \qquad \textit{(boundary conditions)} \end{aligned} \tag{35}$$

The kernel of this operator is

$$g_{m,\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{z}) = r_{\boldsymbol{z}}^{(m)}(\boldsymbol{x}) - \sum_{n=1}^{N_0} p_n(\boldsymbol{x}) q_n(\boldsymbol{z}),$$

where $r_{\boldsymbol{z}}^{(m)} = r_{(\boldsymbol{w},b)}^{(m)} = \rho_m(\boldsymbol{w}^\mathsf{T}(\cdot) - b)$ and $q_n(\boldsymbol{z}) \coloneqq \langle \phi_n, r_{\boldsymbol{z}} \rangle$.

If $\mathrm{R}_{m,\boldsymbol{\phi}}^{-1}$ is bounded, then it admits a continuous extension from $\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$ or $\mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1} \times \mathbb{R})$ to $L^{\infty, m-1}(\mathbb{R}^d)$ when m is even or odd, with (35) holding for all $\psi \in \mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$ or $\psi \in \mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1} \times \mathbb{R})$.

With these two results, we can now establish the Banach space structure of $\mathcal{F}_m$.

Theorem 22 *Let $(\boldsymbol{\phi}, \boldsymbol{p})$ be a biorthogonal system for the null space $\mathcal{N}_m$ of R_m as defined in (7) and let $\mathcal{F}_m$ be the native space of R_m as defined in (8). Then, the following hold:*

1. *The right-inverse operator $\mathrm{R}_{m,\boldsymbol{\phi}}^{-1}$ specified by Lemma 21 isometrically maps $\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$ (respectively $\mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1} \times \mathbb{R})$) to $\mathcal{F}_m$ when m is even (respectively odd). Moreover, this map is necessarily bounded.*

2. Every $f \in \mathcal{F}_m$ admits a unique *representation*

$$f = \mathrm{R}_{m,\boldsymbol{\phi}}^{-1}\, u + q, \tag{36}$$

where $u = \mathrm{R}_m f \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})$[6] and $q = \sum_{n=1}^{N_0} \langle \phi_n, f \rangle p_n \in \mathcal{N}_m$. In particular, this specifies the structural property $\mathcal{F}_m = \mathcal{F}_{m,\boldsymbol{\phi}} \oplus \mathcal{N}_m$, where

$$\mathcal{F}_{m,\boldsymbol{\phi}} \coloneqq \{f \in \mathcal{F} : \ \boldsymbol{\phi}(f) = \mathbf{0}\}.$$

3. $\mathcal{F}_m$ is a Banach space when equipped with the norm

$$\|f\|_{\mathcal{F}_m} \coloneqq \|\mathrm{R}_m f\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} + \|\boldsymbol{\phi}(f)\|_2.$$

Remark 23 *Item 2. in Theorem 22 says that every $f \in \mathcal{F}_m$ admits an* integral representation *via* (36)*, that can be viewed as an infinite-width (continuum-width) neural network. Since u in* (36) *depends on f, it follows that* (36) *is a kind of* Calderón-type reproducing formula *(Calderón, 1964). Integral representations have been studied in several recent works (Bach, 2017; Ongie et al., 2020; Mhaskar, 2020). We also remark that* (36) *shares many similarities with the* dual Ridgelet transform *(Murata, 1996; Rubin, 1998; Candès, 1998, 1999).*

The proofs of Lemmas 19 and 21 and Theorem 22 appear in Appendix A. We will now prove Theorem 1.

4.1 Proof of Theorem 1

Proof We first recast the problem in (9) as one with interpolation constraints. To do this use a technique from Unser (2020). We use the fact that G is a strictly convex function. In particular, for any two solutions $\bar{f}$, $\widetilde{f}$ of (9), we must have $\mathcal{V}\bar{f} = \mathcal{V}\widetilde{f}$ (since otherwise, it would contradict the strict convexity of G). Hence, there exists $\boldsymbol{z} \in \mathbb{R}^N$ such that $\boldsymbol{z} = \mathcal{V}\bar{f} = \mathcal{V}\widetilde{f}$. Although $\boldsymbol{z} \in \mathbb{R}^N$ is not usually known before hand, this property provides us with the parametric characterization of the solution set to (9) as

$$S_{\boldsymbol{z}} \coloneqq \operatorname*{arg\,min}_{f \in \mathcal{F}_m} \|\mathrm{R}_m f\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} \quad \text{s.t.} \quad \mathcal{V}f = \boldsymbol{z}, \tag{37}$$

for some $\boldsymbol{z} \in \mathbb{R}^N$. Hence, it suffices to show that there exists a solution to (37) of the form in (10). We will now show this for a fixed $\boldsymbol{z} \in \mathbb{R}^N$.

Consider the $N \times N_0$ matrix

$$\mathbf{A} \coloneqq \begin{bmatrix} \mathcal{V}p_1 & \cdots & \mathcal{V}p_{N_0} \end{bmatrix}, \tag{38}$$

6. More specifically, we have that $u \in \mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$ when m is even and $u \in \mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1} \times \mathbb{R})$ when m is odd.

where $\{p_n\}_{n=1}^{N_0}$ is a basis for $\mathcal{N}_m$. Since every $q \in \mathcal{N}_m$ has a *unique* expansion $q = \sum_{n=1}^{N_0} c_n p_n$, we have that the linear system $\mathbf{A}\boldsymbol{c} = \mathcal{V}q$, where $\boldsymbol{c} = (c_1, \ldots, c_{N_0})$, has a unique solution. From Item 3. in Theorem 1, we see that the system in (38) satisfies $N \geq N_0$. Hence, it is solvable if and only if $\mathbf{A}^\mathsf{T}\mathbf{A}$ is invertible and the solution is given by the least-squares solution

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

Thus, we know $\mathbf{A}^\mathsf{T}\mathbf{A}$ must be invertible. Invertibility of $\mathbf{A}^\mathsf{T}\mathbf{A}$ says, in particular, that $\operatorname{span}\{\boldsymbol{a}_n\}_{n=1}^{N} = \mathbb{R}^{N_0}$, where $\boldsymbol{a}_n^\mathsf{T}$ is the nth row of $\mathbf{A}$. Therefore, there exists a subset of N_0 rows of $\mathbf{A}$ that span $\mathbb{R}^{N_0}$. Without loss of generality, suppose this subset is $\{\boldsymbol{a}_n\}_{n=1}^{N_0}$. Then, the submatrix $\mathbf{A}_0$ of $\mathbf{A}$ defined by

$$\mathbf{A}_0 \coloneqq \begin{bmatrix} \boldsymbol{a}_1^\mathsf{T} \\ \vdots \\ \boldsymbol{a}_{N_0}^\mathsf{T} \end{bmatrix}$$

is invertible. Consider the components $(\nu_1, \ldots, \nu_N)$ of $\mathcal{V}$ via $\mathcal{V} : f \mapsto (\langle \nu_1, f\rangle, \ldots, \langle \nu_N, f\rangle)$. We can write $\boldsymbol{a}_n$ 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 $\mathcal{N}_m$. 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 nth column of $\mathbf{A}_0$. Let $\boldsymbol{e}_n$ denote the nth canonical basis vector. Then, we have the equality $\mathcal{V}_0 p_n = \mathbf{A}_0 \boldsymbol{e}_n$. Using the invertibility of $\mathbf{A}_0$, we have

$$\mathbf{A}_0^{-1}(\mathcal{V}_0 p_n) = \boldsymbol{e}_n.$$

If we put $\boldsymbol{\phi}_0 \coloneqq \mathbf{A}_0^{-1} \circ \mathcal{V}_0$, the above display is exactly the biorthogonality property and hence $(\boldsymbol{\phi}_0, \boldsymbol{p})$ form a biorthogonal system for $\mathcal{N}_m$.

One can verify that

$$\|\mathrm{R}_m f\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} = \begin{cases} \|\mathrm{R}_m f\|_{\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1}\times\mathbb{R})}, & m \text{ even} \\ \|\mathrm{R}_m f\|_{\mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1}\times\mathbb{R})}, & m \text{ odd}. \end{cases} \tag{39}$$

Suppose that m is even. By Theorem 22, every $f \in \mathcal{F}_m$ admits a unique representation $f = \mathrm{R}_{m,\boldsymbol{\phi}_0}^{-1} u + q_0$, where $u = \mathrm{R}_m f \in \mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1} \times \mathbb{R})$ and $q_0 \in \mathcal{N}_m$. We can use the above display to rewrite the problem in (37) as

$$\begin{aligned} \min_{u \in \mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1}\times\mathbb{R})} \quad & \|u\|_{\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1}\times\mathbb{R})} \\ \text{s.t.} \quad & \mathcal{V}f = \boldsymbol{z}, \\ & f = \mathrm{R}_{m,\boldsymbol{\phi}}^{-1} u + q_0. \end{aligned} \tag{40}$$

Write $\boldsymbol{z}_0 = (z_1, \ldots, z_{N_0})$ and $\boldsymbol{z}_1 = (z_{N_0+1}, \ldots, z_N)$. Then, the constraint $\mathcal{V}f = \boldsymbol{z}$ can be written as the two constraints $\mathcal{V}_0 f = \boldsymbol{z}_0$ and $\mathcal{V}_1 f = \boldsymbol{z}_1$. By the boundary conditions in (35) we have that $\boldsymbol{\phi}_0(f) = \boldsymbol{\phi}_0(\mathrm{R}_{m,\boldsymbol{\phi}_0}^{-1} u + q_0) = \boldsymbol{\phi}_0(\mathrm{R}_{m,\boldsymbol{\phi}_0}^{-1} u) + \boldsymbol{\phi}_0(q_0) = \boldsymbol{\phi}_0(q_0)$. Thus, by

definition of $\boldsymbol{\phi}_0$, we have that $\boldsymbol{z}_0 = \mathcal{V}_0 f = \mathcal{V}_0 q_0$. Hence, (40) can be rewritten as

$$\begin{aligned} \min_{u \in \mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1}\times\mathbb{R})} \quad & \|u\|_{\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1}\times\mathbb{R})} \\ \text{s.t.} \quad & \mathcal{V}_1 f = \boldsymbol{z}_1, \\ & f = \mathrm{R}^{-1}_{m,\boldsymbol{\phi}}\, u + q_0 \end{aligned} \quad = \quad \begin{aligned} \min_{u \in \mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1}\times\mathbb{R})} \quad & \|u\|_{\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1}\times\mathbb{R})} \\ \text{s.t.} \quad & \mathcal{V}_1(\mathrm{R}^{-1}_{m,\boldsymbol{\phi}_0}\, u) = \boldsymbol{z}_1 - \mathcal{V}_1 q_0, \\ & f = \mathrm{R}^{-1}_{m,\boldsymbol{\phi}}\, u + q_0. \end{aligned}$$

This says there exists a solution to the above display of the form $\bar{f} = \mathrm{R}^{-1}_{m,\boldsymbol{\phi}_0}\, \bar{u} + q_0$, where

$$\bar{u} \in \operatorname*{arg\,min}_{u \in \mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1}\times\mathbb{R})} \|u\|_{\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1}\times\mathbb{R})} \quad \text{s.t.} \quad \mathcal{V}_1(\mathrm{R}^{-1}_{m,\boldsymbol{\phi}_0}\, u) = \boldsymbol{z}_1 - \mathcal{V}_1 q_0.$$

By Proposition 15, there exists a sparse minimizer to the above display with $N - N_0$ terms. The result then follows by invoking Lemma 21 and Lemma 17, which says $\bar{f} = \mathrm{R}^{-1}_{m,\boldsymbol{\phi}_0}\, \bar{u} + q_0$ takes the form in (10). An analogous argument holds when m is odd by invoking Proposition 16 instead of Proposition 15. ∎

5. Ridge Splines and Polynomial Splines

In this section we establish connections between ridge splines and classical polynomial splines in both the univariate ($d = 1$) and multivariate ($d > 1$) cases.

5.1 Univariate ridge splines are univariate polynomial splines

Univariate ridge splines and classical univariate splines are in fact the same object. To see this, it suffices to verify that

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

when $d = 1$ and then simply invoke the result of Unser et al. (2017), where D^m is the univariate mth derivative operator. Certainly this is true. Indeed, when $d = 1$, we have that $c_d = 1/2$ and from (25) that the univariate Radon transform is simply

$$\mathscr{R}\{f\}(\gamma, t) = \int_{\mathbb{R}} f(x)\delta_{\mathbb{R}}(\gamma x - t)\,\mathrm{d}x = \int_{\mathbb{R}} \frac{f(x)}{|\gamma|}\delta_{\mathbb{R}}\left(x - \frac{t}{\gamma}\right)\mathrm{d}x = \int_{\mathbb{R}} f(x)\delta_{\mathbb{R}}\left(x - \frac{t}{\gamma}\right)\mathrm{d}x = f\left(\frac{t}{\gamma}\right),$$

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,

$$\begin{aligned} c_d \left\| \partial_t^m \Lambda^{d-1}\, \mathscr{R} f \right\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})}\Big|_{d=1} &= \frac{1}{2}\|\partial_t^m\, \mathscr{R} f\|_{\mathcal{M}(\{-1,+1\}\times\mathbb{R})} \\ &= \frac{1}{2}\sum_{\gamma\in\{-1,+1\}} \left\| \mathrm{D}^m f\left(\frac{\cdot}{\gamma}\right)\right\|_{\mathcal{M}(\mathbb{R})} \\ &= \|\mathrm{D}^m f\|_{\mathcal{M}(\mathbb{R})}, \end{aligned}$$

where the last equality holds since $f(\cdot/\gamma)$ is either f or its reflection, both of which will have the same $\|\mathrm{D}^m\{\cdot\}\|_{\mathcal{M}(\mathbb{R})}$ value. Thus, 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 being fit to data have been recently explored in Savarese et al. (2019); Parhi and Nowak (2020). This says, by Theorem 8, 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 an optimal polynomial spline fit 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 $\boldsymbol{\gamma} \in \mathbb{S}^{d-1}$. Suppose $\mathcal{V}$ corresponds to the ideal sampling setting where the sampling locations are located at $\{\boldsymbol{x}_n\}_{n=1}^N \subset \mathbb{R}^d$. Then, using the same technique we did in the proof of Theorem 1, we can recast the continuous-domain inverse problem in (9) as one with interpolation constraints:

$$\min_{f \in \mathcal{F}_m} \|\mathrm{R}_m f\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} \quad \text{s.t.} \quad f(\boldsymbol{x}_n) = z_n,\ n = 1, \dots, N, \tag{41}$$

for some $\boldsymbol{z} \in \mathbb{R}^N$. By (27), the Radon inversion formula, we can always write $f = c_d \mathscr{R}^* \Lambda^{d-1} \mathscr{R} f$ for any $f \in \mathcal{F}_m$, where the operators are understood in the distributional sense via Corollary 13. Thus, we see that the above optimization can be rewritten as

$$\min_{f \in \mathcal{F}_m} c_d \left\| \partial_t^m \Lambda^{d-1} \mathscr{R} f \right\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} \quad \text{s.t.} \quad (c_d \mathscr{R}^* \Lambda^{d-1} \mathscr{R} f)(\boldsymbol{x}_n) = z_n,\ n = 1, \dots, N.$$

If we put $\Phi \coloneqq c_d \Lambda^{d-1} \mathscr{R} f$, then the above optimization is

$$\min_{\Phi \in \mathfrak{F}_m} \|\partial_t^m \Phi\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} \quad \text{s.t.} \quad \mathscr{R}^*\{\Phi\}(\boldsymbol{x}_n) = \int_{\mathbb{S}^{d-1}} \Phi(\boldsymbol{\gamma}, \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{x}_n)\, \mathrm{d}\sigma(\boldsymbol{\gamma}) = z_n,\ n = 1, \dots, N, \tag{42}$$

where $\mathfrak{F}_m$ is the image of $c_d \Lambda^{d-1} \mathscr{R}$ applied to $\mathcal{F}_m$. This essentially says for a fixed direction $\boldsymbol{\gamma} \in \mathbb{S}^{d-1}$, the function $\bar{\Phi}(\boldsymbol{\gamma}, \cdot) : \mathbb{R} \to \mathbb{R}$ is an mth-order polynomial spline. This follows by considering an optimization for each $\boldsymbol{\gamma} \in \mathbb{S}^{d-1}$:

$$\min_{\Phi(\boldsymbol{\gamma},\cdot)} \|\partial_t^m \Phi(\boldsymbol{\gamma}, \cdot)\|_{\mathcal{M}(\mathbb{R})} \quad \text{s.t.} \quad \Phi(\boldsymbol{\gamma}, \boldsymbol{\gamma}^\mathsf{T} \boldsymbol{x}_n) = z_n(\boldsymbol{\gamma}),\ n = 1, \dots, N, \tag{43}$$

where

$$\int_{\mathbb{S}^{d-1}} z_n(\boldsymbol{\gamma})\, \mathrm{d}\sigma(\boldsymbol{\gamma}) = z_n,\ n = 1, \dots, N$$

and noting that by finding a solution for each fixed $\boldsymbol{\gamma} \in \mathbb{S}^{d-1}$, we can find a $\bar{\Phi}$ that attains a lower bound for (42)[7], but this $\bar{\Phi}$ is clearly feasible for (42) and is hence a solution to (42). It then follows that $\bar{\Phi}(\boldsymbol{\gamma}, \cdot)$ is a polynomial spline of order m. In particular, due

7. Since the integral of a min is less than or equal to the min of an integral.

to the structure of the interpolation constraints in (43), we see that $\bar{\Phi}(\boldsymbol{\gamma}, \cdot)$ is interpolates data with sampling locations at $\left\{\boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{x}_n\right\}_{n=1}^{N} \subset \mathbb{R}$. This viewpoint allows us to understand additional structural information about the sparse (i.e., single-hidden layer neural network) solutions to (41). In particular, its classically known[8] that the univariate spline $\bar{\Phi}(\boldsymbol{\gamma}, \cdot)$ has some set of adaptive knot locations $\{t_\ell(\boldsymbol{\gamma})\}_{\ell=1}^{K_{\boldsymbol{\gamma}}} \subset \mathbb{R}$ with $K_{\boldsymbol{\gamma}} \leq N - m$ and there are no knots outside the sampling locations[9], i.e.,

$$|t_\ell(\boldsymbol{\gamma})| \leq \max_{n=1,\ldots,N} \left|\boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{x}_n\right|, \quad \ell = 1, \ldots, K_{\boldsymbol{\gamma}}.$$

It is then clear that for $\bar{\Phi}$ to be a sparse minimizer of (9), it must satisfy Definition 5. This implies that the *biases* in a ridge spline solution to (9) exactly correspond to these knot locations. Thus, we can see that for a ridge spline solution to (9) as in (10) with K neurons, we have the additional information about a bound on the bias terms, which we summarize in the following lemma.

Lemma 24 *In the ideal sampling scenario, the biases in the sparse solution* (10) *of the variational problem in* (9) *satisfy*

$$|b_k| \leq \max_{n=1,\ldots,N} \|\boldsymbol{x}_n\|_2,$$

for all $k = 1, \ldots, K$.

Proof The proof follows from the discussion above. In particular,

$$|b_k| \leq \sup_{\boldsymbol{\gamma} \in \mathbb{S}^{d-1}} \max_{\ell=1,\ldots,K_{\boldsymbol{\gamma}}} |t_\ell(\boldsymbol{\gamma})| \leq \sup_{\boldsymbol{\gamma} \in \mathbb{S}^{d-1}} \max_{n=1,\ldots,N} \left|\boldsymbol{\gamma}^{\mathsf{T}}\boldsymbol{x}_n\right| \leq \max_{n=1,\ldots,N} \|\boldsymbol{x}_n\|_2.$$

for all $k = 1, \ldots, K$. ■

6. Applications to Neural Networks

In this section we will apply Theorem 1 to neural network training, regularization, and generalization.

6.1 Finite-dimensional neural network training problems

In this section we will prove Theorem 8.

Lemma 25 *Consider the single-hidden layer neural network*

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

where $\boldsymbol{\theta} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K, v_1, \ldots, v_K, b_1, \ldots, b_K, c)$ *contains the neural network parameters such that* $v_k \in \mathbb{R}$, $\boldsymbol{w}_k \in \mathbb{R}^d$, *and* $b_k \in \mathbb{R}$ *for* $k = 1, \ldots, K$, *and where* c *is a polynomial of*

8. Since we can always explicitly construct spline solutions with the spline knots bounded by the data.
9. Notice that the number of knots and the knot locations depend on the direction $\boldsymbol{\gamma} \in \mathbb{S}^{d-1}$.

degree strictly less than m. Also assume without loss of generality that the weight-bias pairs $(\boldsymbol{w}_k, b_k)$ are unique[10]*. Then,*

$$\|\mathrm{R}_m f_{\boldsymbol{\theta}}\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} = \sum_{k=1}^{K} |v_k| \|\boldsymbol{w}_k\|_2^{m-1}.$$

Proof This proof is a direct calculation. Write

$$\begin{aligned}
\mathrm{R}_m f_{\boldsymbol{\theta}} &= \sum_{k=1}^{K} v_k \, \mathrm{R}_m \, \rho_m(\boldsymbol{w}_k^{\mathsf{T}}(\cdot) - b_k) \\
&= \sum_{k=1}^{K} v_k \|\boldsymbol{w}_k\|_2^{m-1} \, \mathrm{R}_m \, \rho_m(\widetilde{\boldsymbol{w}}_k^{\mathsf{T}}(\cdot) - \widetilde{b}_k) \\
&= \sum_{k=1}^{K} v_k \|\boldsymbol{w}_k\|_2^{m-1} \left[\frac{\delta_{\mathbb{S}^{d-1}\times\mathbb{R}}(\cdot - (\widetilde{\boldsymbol{w}}_k, \widetilde{b}_k)) + (-1)^m \delta_{\mathbb{S}^{d-1}\times\mathbb{R}}(\cdot + (\widetilde{\boldsymbol{w}}_k, \widetilde{b}_k))}{2} \right],
\end{aligned}$$

where the second line follows from the substitution $\widetilde{\boldsymbol{w}}_k \coloneqq \boldsymbol{w}_k / \|\boldsymbol{w}_k\|_2 \in \mathbb{S}^{d-1}$ and $\widetilde{b}_k \coloneqq b_k / \|\boldsymbol{w}_k\|_2 \in \mathbb{R}$ combined with the homogenity of degree $m-1$ of ρ_m and the third line follows from Lemma 17. Taking the $\mathcal{M}$-norm proves the lemma. ∎

6.1.1 Proof of Theorem 8

Proof Recasting the problem in (9) as (14) follows from Theorem 1. Equivalence of the problem in (14) and (15) follows from Lemma 25. Thus, we just need to show that the solutions to the problem in (16) are also solutions to problem in (15). To see this, let $\boldsymbol{\theta}$ be a solution to (16) with network weights $\{(v_k, \boldsymbol{w}_k)\}_{k=1}^K$. Consider the regularizer from (16):

$$\frac{1}{2} \sum_{k=1}^{K} \Big(|v_k|^2 + \|\boldsymbol{w}_k\|_2^{2m-2} \Big).$$

Since ρ_m is homogeneous of degree $m-1$, the weights may be rescaled so that $|v_k| = \|\boldsymbol{w}_k\|_2^{m-1}$, $k = 1, \ldots, K$, without altering the function of the network and its fit to the data. Note that each term of the regularizer is a sum of squares $|v_k|^2 + \big(\|\boldsymbol{w}_k\|_2^{m-1}\big)^2$, and thus each term is minimized when $|v_k| = \|\boldsymbol{w}_k\|_2^{m-1}$. Thus, at the minimizer we have

$$\frac{1}{2} \sum_{k=1}^{K} \Big(|v_k|^2 + \|\boldsymbol{w}_k\|_2^{2m-2} \Big) = \sum_{k=1}^{K} |v_k| \|\boldsymbol{w}_k\|_2^{m-1},$$

which is exactly the regularizer of (15). ∎

10. In the sense that $(\boldsymbol{w}_k, b_k) \neq (\boldsymbol{w}_n, b_n)$ for $k \neq n$.

6.2 Generalization bounds for binary classification

In this section we will prove Theorem 10.

6.2.1 Proof of Theorem 10

Proof Using the rescaling technique discussed in Remark 3, without loss of generality, we may assume that $\|\boldsymbol{w}_k\|_2 = 1$ (since we can absorb the norm of $\boldsymbol{w}_k$ into the magnitude of v_k). In this case,

$$\|\mathrm{R}_m f_{\boldsymbol{\theta}}\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} = \sum_{k=1}^{K} |v_k| \leq B.$$

To begin we bound the *empirical Rademacher complexity* of $\mathcal{F}_\Theta$. The empirical Rademacher complexity, denoted by $\widehat{\mathfrak{R}}(\mathcal{F}_\Theta)$, is computed by taking the conditional expectation, conditioning on $\{\boldsymbol{x}_n\}_{n=1}^N$ in place of the total expectation in (22). In other words, the only random variables are $\{\sigma_n\}_{n=1}^N$. The Rademacher complexity is then

$$\mathfrak{R}(\mathcal{F}_\Theta) = \mathbb{E}\Big[\widehat{\mathfrak{R}}(\mathcal{F}_\Theta)\Big].$$

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

$$\widehat{\mathfrak{R}}\Big(\rho_m(\boldsymbol{w}^\mathsf{T}(\cdot) - b)\Big) \coloneqq 2\,\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\boldsymbol{w}:\|\boldsymbol{w}\|_2=1\\ b:|b|\leq C/2}} \frac{1}{N}\sum_{n=1}^{N} \sigma_n \rho_m(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n - b)\right].$$

First notice that when m is odd

$$\rho_m(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b) = \frac{(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)^{m-1} + \big|\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b\big|^{m-2}(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)}{2(m-1)!}, \tag{44}$$

and when m is even

$$\rho_m(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b) = \frac{(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b)^{m-1} + \big|\boldsymbol{w}^\mathsf{T}\boldsymbol{x} - b\big|^{m-1}}{2(m-1)!}. \tag{45}$$

It is well-known that for two function spaces $\mathcal{F}$ and $\mathcal{G}$, the empirical Rademacher complexity satisfies $\widehat{\mathfrak{R}}(\mathcal{F}\oplus\mathcal{G}) = \widehat{\mathfrak{R}}(\mathcal{F}) + \widehat{\mathfrak{R}}(\mathcal{G})$, where $\oplus$ is the direct-sum. With this property, we see from (44) and (45) that the empirical Rademacher complexity of a single neuron is

$$\begin{aligned}
&\widehat{\mathfrak{R}}\Big(\rho_m(\boldsymbol{w}^\mathsf{T}(\cdot) - b)\Big)\\
&\quad= \frac{1}{2(m-1)!}\Bigg(\underbrace{\widehat{\mathfrak{R}}\Big((\boldsymbol{w}^\mathsf{T}(\cdot) - b)^{m-1}\Big)}_{(*)} + \underbrace{\begin{cases}\widehat{\mathfrak{R}}\Big(\big|\boldsymbol{w}^\mathsf{T}(\cdot) - b\big|^{m-2}(\boldsymbol{w}^\mathsf{T}(\cdot) - b)\Big), & m \text{ is odd}\\ \widehat{\mathfrak{R}}\Big(\big|\boldsymbol{w}^\mathsf{T}(\cdot) - b\big|^{m-1}\Big), & m \text{ is even}\end{cases}}_{(\S)}\Bigg).
\end{aligned}$$

Since the functions in $(*)$ and $(\S)$ are the same up to a sign, it follows that the Rademacher complexities are the same due to the symmetry of the Rademacher random variables. Thus,

$$\widehat{\mathfrak{R}}\Big(\rho_m(\boldsymbol{w}^\mathsf{T}(\cdot)-b)\Big)=\frac{\widehat{\mathfrak{R}}\big((\boldsymbol{w}^\mathsf{T}(\cdot)-b)^{m-1}\big)}{(m-1)!}=\frac{2}{N(m-1)!}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\boldsymbol{w}:\|\boldsymbol{w}\|_2=1\\ b:|b|\leq C/2}}\sum_{n=1}^{N}\sigma_n(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n-b)^{m-1}\right].$$

Next, by the binomial theorem

$$\begin{aligned}
\widehat{\mathfrak{R}}\Big(\rho_m(\boldsymbol{w}^\mathsf{T}(\cdot)-b)\Big) &\leq \frac{2}{N(m-1)!}\sum_{k=0}^{m-1}\binom{m-1}{k}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\substack{\boldsymbol{w}:\|\boldsymbol{w}\|_2=1\\ b:|b|\leq C/2}}\sum_{n=1}^{N}\sigma_n(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n)^k(-b)^{m-1-k}\right]\\
&\leq \frac{2}{N(m-1)!}\sum_{k=0}^{m-1}\binom{m-1}{k}\left(\frac{C}{2}\right)^{m-1-k}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2=1}\sum_{n=1}^{N}\sigma_n(\boldsymbol{w}^\mathsf{T}\boldsymbol{x}_n)^k\right]\\
&= \frac{2}{N(m-1)!}\sum_{k=0}^{m-1}\binom{m-1}{k}\left(\frac{C}{2}\right)^{m-1-k}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\boldsymbol{w}:\|\boldsymbol{w}\|_2=1}\left(\sum_{n=1}^{N}\sigma_n\boldsymbol{x}_n^{\otimes k}\right)^\mathsf{T}\boldsymbol{w}^{\otimes k}\right]\\
&\leq \frac{2}{N(m-1)!}\sum_{k=0}^{m-1}\binom{m-1}{k}\left(\frac{C}{2}\right)^{m-1-k}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{n=1}^{N}\sigma_n\boldsymbol{x}_n^{\otimes k}\right\|_2\right],
\end{aligned}$$

where $(\cdot)^{\otimes k}$ denotes the kth order Kronecker product. By Jensen's inequality we have

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{n=1}^{N}\sigma_n\boldsymbol{x}_n^{\otimes k}\right\|_2\right]\leq\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{n=1}^{N}\sigma_n\boldsymbol{x}_n^{\otimes k}\right\|_2^2\right]^{1/2}=\left(\sum_{n=1}^{N}\left\|\boldsymbol{x}_n^{\otimes k}\right\|_2^2\right)^{1/2}\leq\sqrt{N}\left(\frac{C}{2}\right)^k,$$

and so

$$\widehat{\mathfrak{R}}\Big(\rho_m(\boldsymbol{w}^\mathsf{T}(\cdot)-b)\Big)\leq\frac{2C^{m-1}}{\sqrt{N}(m-1)!}.$$

Therefore, the empirical Rademacher complexity of $\mathcal{F}_\Theta$ is bounded as follows

$$\widehat{\mathfrak{R}}(\mathcal{F}_\Theta)=\sum_{k=1}^{K}|v_k|\widehat{\mathfrak{R}}\Big(\rho_m(\boldsymbol{w}_k^\mathsf{T}(\cdot)-b_k)\Big)+\widehat{\mathfrak{R}}(c)\leq\frac{2BC^{m-1}}{\sqrt{N}(m-1)!}+\widehat{\mathfrak{R}}(c).$$

Taking the expectation of both sides proves the theorem. ∎

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 polynomial splines. There are a number of followup research questions that may be asked.

7.1 Computational issues

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., 2019b,a), though full theoretical justifications of why these algorithms work currently do not exist. Thus, it remains an open question about designing neural network training algorithms that provably find global minimizers. Such algorithms could then be used to find the sparse solutions the continuous-domain inverse problems studied in this paper.

7.2 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 on data? Answering this question would require posing a continuous-domain inverse problem and deriving a representer theorem showing that deep networks are solutions. We believe answering this question will be challenging, due to the compositions of ridge functions that arise in deep networks.

Acknowledgments

This work is partially supported by AFOSR/AFRL grant FA9550-18-1-0166, the NSF Research Traineeship Program under grant 1545481, and the NSF Graduate Research Fellowship Program under grant DGE-1747503. The authors thank Greg Ongie for helpful feedback and discussions related to the initial draft of this paper. The authors also thank the anonymous reviewers for their constructive feedback.

Appendix A. Auxiliary Proofs

A.1 Proof of Lemma 9

Proof It suffices to prove that for a fixed $\boldsymbol{x}_0 \in \mathbb{R}^d$, the operator $f \mapsto \langle \delta_{\mathbb{R}^d}(\cdot - \boldsymbol{x}_0), f \rangle$ is bounded. Let $(\boldsymbol{\phi}, \boldsymbol{p})$ be a biorthogonal system for the null space $\mathcal{N}_m$ of R_m. By Theorem 22, we know every $f \in \mathcal{F}_m$ admits the unique representation $f = \mathrm{R}_{m,\boldsymbol{\phi}}^{-1} u + q$, where $u = \mathrm{R}_m f$. Moreover, from the proof of Theorem 22, we know that $\mathrm{R}_{m,\boldsymbol{\phi}}^{-1}$ is bounded and hence its kernel $g_{m,\boldsymbol{\phi}}(\boldsymbol{x}_0, \cdot)$ is bounded for fixed $\boldsymbol{x}_0 \in \mathbb{R}^d$. We also have by construction that $g_{m,\boldsymbol{\phi}}(\boldsymbol{x}_0, \cdot)$

is continuous. Next, suppose $|g_{m,\boldsymbol{\phi}}(\boldsymbol{x}_0, \cdot)|$ is bounded by $0 < M_{\boldsymbol{x}_0} < \infty$. Then, for any $f \in \mathcal{F}_m$

$$\begin{aligned}
|\langle \delta_{\mathbb{R}^d}(\cdot - \boldsymbol{x}_0), f\rangle| &= |f(\boldsymbol{x}_0)| \\
&= \left|\mathrm{R}_{m,\boldsymbol{\phi}}^{-1}\{u\}(\boldsymbol{x}_0) + q(\boldsymbol{x}_0)\right| \\
&\leq |\langle u, g_{m,\boldsymbol{\phi}}(\boldsymbol{x}_0, \cdot)\rangle| + |q(\boldsymbol{x}_0)| \\
&\leq M_{\boldsymbol{x}_0}\|u\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} + |q(\boldsymbol{x}_0)| \\
&= M_{\boldsymbol{x}_0}\|\mathrm{R}_m\, u\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} + |q(\boldsymbol{x}_0)|,
\end{aligned}$$

where the fourth line follows since we can extend $u \in \mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})$ to act continuously on $C_{\mathrm{b}}(\mathbb{S}^{d-1} \times \mathbb{R})$, the space of bounded continuous functions on $\mathbb{S}^{d-1} \times \mathbb{R}$. From Lemma 19, we know $\mathcal{N}_m$ is the space of polynomials of degree strictly less than m. Thus, it's clear that point evaluations are continuous on $\mathcal{N}_m$. From the proof of Theorem 22, we know that the norm $q \mapsto \|\boldsymbol{\phi}(q)\|_2$ equips $\mathcal{N}_m$ with a Banach space structure. Therefore, there exists a constant $0 < \widetilde{M}_{\boldsymbol{x}_0} < \infty$ such that $|q(\boldsymbol{x}_0)| \leq \widetilde{M}_{\boldsymbol{x}_0}\|\boldsymbol{\phi}(q)\|_2 = \widetilde{M}_{\boldsymbol{x}_0}\|\boldsymbol{\phi}(f)\|_2$. Thus, there exists a constant $C_{\boldsymbol{x}_0} \coloneqq M_{\boldsymbol{x}_0} + \widetilde{M}_{\boldsymbol{x}_0}$ such that

$$|\langle \delta_{\mathbb{R}^d}(\cdot - \boldsymbol{x}_0), f\rangle| \leq C_{\boldsymbol{x}_0}\Big(\|\mathrm{R}_m\, f\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} + \|\boldsymbol{\phi}(f)\|_2\Big) = C_{\boldsymbol{x}_0}\|f\|_{\mathcal{F}_m},$$

where the last inequality is from Item 3. in Theorem 22. ■

A.2 Proof of Lemma 19

We can prove that the null space $\mathcal{N}_m$ of R_m is finite-dimensional via the *Fourier slice theorem* (Helgason, 2014, pg. 4), which says for any $f \in \mathscr{S}'(\mathbb{R}^d)$,

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

where the Fourier transform on the left-hand side is a one-dimensional Fourier transform with respect to $t \to \omega$ and the Fourier transform on the right-hand side is the multivariate Fourier transform of f.

Proof [Proof of Lemma 19] Recall that

$$\mathrm{R}_m = c_d\, \partial_t^m \Lambda^{d-1}\, \mathscr{R}\,.$$

Next, let $f \in L^{\infty,m-1}(\mathbb{R}^d)$. Then,

$$\widehat{\mathrm{R}_m\{f\}}(\boldsymbol{\gamma}, \omega) = c_d\, (\mathrm{i}\omega)^m\, \mathrm{i}^d|\omega|^d\, \widehat{\mathscr{R}\{f\}}(\boldsymbol{\gamma}, \omega) = c_d\, (\mathrm{i}\omega)^m\, \mathrm{i}^d|\omega|^d\, \widehat{f}(\omega\boldsymbol{\gamma}),$$

where the Fourier transform is with respect to $t \to \omega$ and the second equality holds via the Fourier slice theorem. Next, we notice that for $f \in \mathcal{N}_m$ we require that the above display is 0 for all $(\boldsymbol{\gamma}, \omega) \in \mathbb{S}^{d-1} \times \mathbb{R}$. Since the above display is zero at $\omega = 0$ we see that it must be that $\widehat{f}$ is supported only at $\mathbf{0}$. This says that f must be a polynomial. Finally, since $f \in L^{\infty,m-1}(\mathbb{R}^d)$, we see that f must be a polynomial of degree strictly less than m. ■

A.3 Proof of Lemma 21

Proof Using Lemma 17, which says that $r_{\boldsymbol{z}}$ is a translated Green's function of R_m, and that $p_n \in \mathcal{N}_m$, $n = 1, \ldots, N_0$, a direct calculation results in

$$\mathrm{R}_m \, \mathrm{R}_{m,\boldsymbol{\phi}}^{-1} \, \psi = \psi$$

for all even (respectively odd) $\psi \in \mathscr{S}_0(\mathbb{S}^{d-1} \times \mathbb{R})$ when m is even (respectively odd).

Suppose m is even (since the case of m being odd is analogous). To check the boundary conditions we check for all even $\psi \in \mathscr{S}_0(\mathbb{S}^{d-1} \times \mathbb{R})$ that $\left\langle \phi_k, \mathrm{R}_{m,\boldsymbol{\phi}}^{-1} \, \psi \right\rangle = 0$, $k = 1, \ldots, N_0$. Write

$$\left\langle \phi_k, \mathrm{R}_{m,\boldsymbol{\phi}}^{-1} \, \psi \right\rangle = \left\langle \phi_k, \int_{\mathbb{S}^{d-1} \times \mathbb{R}} r_{\boldsymbol{z}}(\cdot) \psi(\boldsymbol{z}) \, \mathrm{d}(\sigma \times \lambda)(\boldsymbol{z}) \right\rangle - \underbrace{\sum_{n=1}^{N_0} \langle \phi_k, p_n \rangle \langle q_n, \psi \rangle}_{= \langle q_k, \psi \rangle} .$$

Next,

$$\begin{aligned} \langle q_k, \psi \rangle = \int_{\mathbb{S}^{d-1} \times \mathbb{R}} q_k(\boldsymbol{z}) \psi(\boldsymbol{z}) \, \mathrm{d}(\sigma \times \lambda)(\boldsymbol{z}) &= \int_{\mathbb{S}^{d-1} \times \mathbb{R}} \langle \phi_k, r_{\boldsymbol{z}} \rangle \psi(\boldsymbol{z}) \, \mathrm{d}(\sigma \times \lambda)(\boldsymbol{z}) \\ &= \left\langle \phi_k, \int_{\mathbb{S}^{d-1} \times \mathbb{R}} r_{\boldsymbol{z}}(\cdot) \psi(\boldsymbol{z}) \, \mathrm{d}(\sigma \times \lambda)(\boldsymbol{z}) \right\rangle. \end{aligned}$$

Thus, the previous two displays imply $\left\langle \phi_k, \mathrm{R}_{m,\boldsymbol{\phi}}^{-1} \, \psi \right\rangle = 0$, $k = 1, \ldots, N_0$. Uniqueness of $\mathrm{R}_{m,\boldsymbol{\phi}}^{-1}$ follows from the fact that the biorthogonal system provides a unique representation of elements of $\mathcal{N}_m$.

Again suppose that m is even (since the case of m being odd is analagous). Assume that $\mathrm{R}_{m,\boldsymbol{\phi}}^{-1}$ is bounded, in other words, that this inverse is *stable*. Since $\mathscr{S}_0(\mathbb{P}^d) \subset \mathscr{S}(\mathbb{P}^d)$ and $\mathscr{S}_0(\mathbb{P}^d)$ and $\mathscr{S}(\mathbb{P}^d)$ are both dense in $C_0(\mathbb{P}^d)$ it follows that $\mathscr{S}_0(\mathbb{P}^d)$ is dense in $\mathscr{S}(\mathbb{P}^d)$. Next, it is well known that the space of Schwartz functions is dense in the space of tempered distributions (Schwartz, 1966). Thus, since we are assuming $\mathrm{R}_{m,\boldsymbol{\phi}}^{-1}$ is bounded, we can continuously extend it to act on elements of $\mathcal{M}(\mathbb{P}^d) \subset \mathscr{S}'(\mathbb{P}^d)$ with the properties in (35) holding for every $\psi \in \mathcal{M}(\mathbb{P}^d)$. We also remark that we show that $\mathrm{R}_{m,\boldsymbol{\phi}}^{-1}$ is necessarily bounded when acting on elements of $\mathcal{M}(\mathbb{P}^d)$ in Theorem 22. ∎

A.4 Proof of Theorem 22

Proof Before proving the individual items in the theorem, first recall from the theorem statement the definition

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

Clearly this is a vector space. Since $f \mapsto \| \mathrm{R}_m \, f \|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})}$ is a seminorm, it is a a norm except for lacking the property that $\| \mathrm{R}_m \, f \|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})} = 0$ if and only if $f \equiv 0$ (since every $q \in \mathcal{N}_m$ has $\| \mathrm{R}_m \, q \|_{\mathcal{M}(\mathbb{S}^{d-1} \times \mathbb{R})} = 0$). By imposing the boundary conditions $\boldsymbol{\phi}(f) = \mathbf{0}$ in the definition of $\mathcal{F}_{m,\boldsymbol{\phi}}$, we enforce that every $f \in \mathcal{F}_{m,\boldsymbol{\phi}}$ has no null space component. Thus,

$\mathcal{F}_{m,\boldsymbol{\phi}}$ is a bona fide Banach space when equipped with the norm $f \mapsto \|\mathrm{R}_m f\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})}$, more specifically this shows that $\mathcal{F}_{m,\boldsymbol{\phi}}$ is isometrically isomorphic to $\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1}\times\mathbb{R})$ (respectively $\mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1}\times\mathbb{R})$) when m is even (respectively odd)[11]. In particular, this says we have the topological isomorphism $\mathcal{F}_{m,\boldsymbol{\phi}} \cong \mathcal{F}_m/\mathcal{N}_m$.

1. The above discussion along with (39) has shown that R_m is a bijective isometry from $\mathcal{F}_{m,\boldsymbol{\phi}}$ to $\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1}\times\mathbb{R})$ or $\mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1}\times\mathbb{R})$ when m is even or odd. Since R_m is a bijective isometry, it is bounded, allowing us to invoke the bounded inverse theorem (Folland, 1999, Chapter 5). This says there exists a bounded inverse R_m^{-1} (in this case, an isometry) of R_m mapping $\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1}\times\mathbb{R})$ or $\mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1}\times\mathbb{R})$ to $\mathcal{F}_{m,\boldsymbol{\phi}}$ when m is even or odd. This inverse is necessarily the *unique operator* $\mathrm{R}_{m,\boldsymbol{\phi}}^{-1}$ constructed in Lemma 21 as it imposes the boundary conditions in (35). This result indirectly shows that the operator $\mathrm{R}_{m,\boldsymbol{\phi}}^{-1}$ when acting on $\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1}\times\mathbb{R})$ or $\mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1}\times\mathbb{R})$ is bounded when m is even or odd.

2. Given the biorthogonal system $(\boldsymbol{\phi}, \boldsymbol{p})$ of $\mathcal{N}_m$, consider the projection operator

$$\mathrm{proj}_{\mathcal{N}_m} : f \mapsto \sum_{n=1}^{N_0} \langle \phi_n, f \rangle p_n.$$

   Then, for every $f \in \mathcal{F}_m$ we can write $f = \widetilde{f} + q$, where $q \coloneqq \mathrm{proj}_{\mathcal{N}_m}$ and so $\widetilde{f} = f - q$. By this construction, $\boldsymbol{\phi}(\widetilde{f}) = 0$, so $\widetilde{f} \in \mathcal{F}_{m,\boldsymbol{\phi}}$. Clearly, $\widetilde{f} = \mathrm{R}_{m,\boldsymbol{\phi}}^{-1} u$, where $u \coloneqq \mathrm{R}_m \widetilde{f} = \mathrm{R}_m f$. Indeed, this is true since $\mathrm{R}_{m,\boldsymbol{\phi}}^{-1}$ is a bona fide inverse when acting on $\mathcal{M}_{\mathsf{even}}(\mathbb{S}^{d-1}\times\mathbb{R})$ or $\mathcal{M}_{\mathsf{odd}}(\mathbb{S}^{d-1}\times\mathbb{R})$ when m is even or odd. Thus, (36) holds. Since $\mathcal{F}_{m,\boldsymbol{\phi}} \cong \mathcal{F}_m/\mathcal{N}_m$, we have $\mathcal{F}_{m,\boldsymbol{\phi}} \cap \mathcal{N}_m = \{0\}$. Hence, we have the direct-sum decomposition $\mathcal{F}_m = \mathcal{F}_{m,\boldsymbol{\phi}} \oplus \mathcal{N}_m$.

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

$$\|f\|_{\mathcal{F}_m} \coloneqq \|\mathrm{R}_m f\|_{\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})} + \|\boldsymbol{\phi}(f)\|_2$$

   to provide $\mathcal{F}_m$ a Banach space structure.

■

11. Where we are using the fact that the range of R_m is even or odd elements of $\mathcal{M}(\mathbb{S}^{d-1}\times\mathbb{R})$ when m is even or odd combined with (39).

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.

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.

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:1607.07819v1*, 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.

L. Oneto, S. Ridella, and D. Anguita. Tikhonov, ivanov and morozov regularization for support vector machine learning. *Machine Learning*, 103(1):103–136, 2016.

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. The role of neural network activation functions. *IEEE Signal Processing Letters*, 27:1779–1783, 2020. doi: 10.1109/LSP.2020.3027517.

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. ℓ_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 Φ_V of Lizorkin type in the mixed $L^{\overline{p}}(\mathbb{R}^n)$-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.

L. Schwartz. *Théorie des distributions*, volume 2. Hermann Paris, 1966.

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, 2019.

M. Unser. A unifying representer theorem for inverse problems and machine learning. *Foundations of Computational Mathematics*, pages 1–20, 2020.

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. Zuhovickiĭ. Remarks on problems in approximation theory. *Mat. Zbirnik KDU*, pages 169–183, 1948.