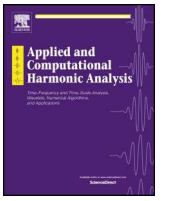
Applied and Computational Harmonic Analysis 74 (2025) 101713

Contents lists available at ScienceDirect

Applied and Computational Harmonic Analysis

journal homepage: www.elsevier.com/locate/acha



Full Length Article

Weighted variation spaces and approximation by shallow ReLU networks ☆

Ronald DeVore [a], Robert D. Nowak [b], Rahul Parhi [c,*], Jonathan W. Siegel [a]

[a] Department of Mathematics, Texas A&M University, College Station, TX 77843, United States of America
[b] Department of Electrical and Computer Engineering, University of Wisconsin–Madison, Madison, WI 53706, United States of America
[c] Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093, United States of America

ARTICLE INFO

Communicated by Hrushikesh Mhaskar

Keywords:
Neural networks
Approximation rates
Variation spaces

ABSTRACT

We investigate the approximation of functions f on a bounded domain $\Omega \subset \mathbb{R}^d$ by the outputs of single-hidden-layer ReLU neural networks of width n. This form of nonlinear n-term dictionary approximation has been intensely studied since it is the simplest case of neural network approximation (NNA). There are several celebrated approximation results for this form of NNA that introduce novel model classes of functions on Ω whose approximation rates do not grow unbounded with the input dimension. These novel classes include Barron classes, and classes based on sparsity or variation such as the Radon-domain BV classes. The present paper is concerned with the definition of these novel model classes on domains Ω. The current definition of these model classes does not depend on the domain Ω. A new and more proper definition of model classes on domains is given by introducing the concept of weighted variation spaces. These new model classes are intrinsic to the domain itself. The importance of these new model classes is that they are strictly larger than the classical (domain-independent) classes. Yet, it is shown that they maintain the same NNA rates.

1. Introduction

Neural networks (NNs) are now the numerical method of choice for the development of learning algorithms in regression and classification, especially when dealing with functions of d variables with d large. It is therefore important to understand, through mathematical theory, the reasons for this success. In learning, we are tasked with approximating an unknown function f on a domain $\Omega \subset \mathbb{R}^d$ from some finite set of data observations of f. Thus, at least part of the success in using NNs for such learning problems, must lie in their ability to effectively approximate the functions of interest. While there is no widespread agreement on exactly what are these functions of interest, i.e., which functions are encountered in applications, one can ask to describe exactly which functions are well approximated by NNs.

☆ This research was supported in part by the NSF grants DMS-2134077 and DMS-2134140 of the NSF MoDL program (RD and RN) as well as the ONR MURI grant N00014-20-1-2787 (RD, RN, and JS). RP was supported in part by the NSF Graduate Research Fellowship Program under grant DGE-1747503 while he was with the University of Wisconsin–Madison and the European Research Council under grant 101020573 while he was with the École polytechnique fédérale de Lausanne. He is now with the University of California, San Diego. JS was supported in part by the NSF grants DMS-2111387 and CCF-2205004.

* Corresponding author.
E-mail addresses: ronalddevore@tamu.edu (R. DeVore), rdnowak@wisc.edu (R.D. Nowak), rahul@ucsd.edu (R. Parhi), jwsiegel@tamu.edu (J.W. Siegel).

https://doi.org/10.1016/j.acha.2024.101713
Received 28 July 2023; Received in revised form 8 April 2024; Accepted 26 September 2024

Available online 10 October 2024

In approximation theory, questions of this second type are answered by precisely describing the set of functions which have a prescribed rate of approximation using the proposed method of approximation. In our case of NN approximation, we could for example ask for a precise characterization of the *approximation classes*

$$\mathcal{A}^\alpha((\Sigma_n)_{n\geq 0}), X) := \{f : \operatorname{dist}(f,\Sigma_n)_X = O(n^{-\alpha}),\ n\geq 1\}, \quad \alpha > 0, \tag{1.1}$$

where X is some Banach space of functions in which we measure the error of approximation (e.g. $X = L_p(\Omega)$, $1 \leq p \leq \infty$) and Σ_n is the set of functions realized by neural networks with n neurons and a prescribed activation function.

For certain methods of approximation such as polynomial or spline approximation precise characterizations of the corresponding approximation classes are known (see, e.g., [12,13]) and are described by a smoothness condition on f or equivalently a statement that a function f is in $\mathcal{A}^\alpha$ if and only if it can be represented as a sum of certain fundamental building blocks called atoms with a specified condition on the coefficients in such a representation.

In the case of NN approximation the characterization of $\mathcal{A}^\alpha$ appears to be a difficult problem, which we are not close to solving. In order to chip away at the problem of characterizing the approximation classes $\mathcal{A}^\alpha$, we can, as we do in this paper, look at the simplest case of NN approximation which corresponds to approximation by shallow neural networks (one hidden layer) with ReLU activation. Even in this simplest case, there are no characterizations of the classes $\mathcal{A}^\alpha$ save for the case $d = 1$ where the desired characterization is known and given by certain smoothness conditions known as Besov regularity (see, e.g., [12]).

The existing literature on NN approximation, briefly described below, gives sufficient conditions to be placed on f which guarantee membership in $\mathcal{A}^\alpha$. We are not aware of nontrivial necessary conditions being known. The present paper will shed some new light on the quest to characterize the approximation classes $\mathcal{A}^\alpha$ when $d \geq 2$ by showing certain deficiencies in the present theory of sufficient conditions. We treat only the case where approximation takes place in $X = L_2(\Omega)$.

We prove two important new results on shallow NN ReLU approximation in this paper. First, we show that the characterization of $\mathcal{A}^\alpha$ will depend significantly on the geometry of the domain Ω. For example, the characterization of $\mathcal{A}^\alpha$ in the case $\Omega = [0,1]^d$ will be different from that of the case when Ω is the Euclidean ball B^d of $\mathbb{R}^d$. This means that domain independent results are insufficient. Secondly, we show that if there is any hope of characterizing $\mathcal{A}^\alpha$ by requiring that f has a certain expansion in terms of the elements from the ReLU dictionary, then this will require conditions on the coefficients in such expansions which reflect the distance of the atom in the dictionary to the boundary of Ω. We believe these two new ingredients will prove to be important.

Our main result gives a sufficient condition for membership in $\mathcal{A}^\alpha$ that is much weaker than those previously known. These new sufficient conditions take the form of f having a NN dictionary expansion with conditions on the coefficients in such an expansions being weaker for atoms close to the boundary of Ω. While these new results still are not proven to characterize the approximation classes $\mathcal{A}^\alpha$ they give much weaker sufficient conditions to guarantee membership in $\mathcal{A}^\alpha$.

We turn now to a brief description of some (but not all) of the recent results on shallow NN approximation using ReLU activation functions. This brief accounting will serve to frame the new results given in the present paper.

A large number of papers have been written in recent years that give quantitative bounds on the approximation rates of various model classes of functions when using neural networks. General accountings of such results can be found in [5,11,17,34]. Two types of results have emerged. The first is to show that deep NNs with ReLU activation functions (see, e.g., [23,37,39,43]) are surprisingly effective in approximating functions from classical model classes such as finite balls in a Sobolev or Besov space when the approximation error is measured in an $L_p(\Omega)$ norm with $1 \leq p \leq \infty$. While such results are deep and interesting, they do not match the most common setting of learning in high dimensions (d large) because these model classes necessarily suffer the curse of dimensionality. Indeed, the approximation rates for such smoothness classes is of the form $O(n^{-s/d})$ with s related to the smoothness assumption on f. Here and later n always refers to the number of neurons used in the approximation. Thus, for large values of d, membership in such a model class is not a realistic assumption to make on the target function f to be learned. This negativity for classical smoothness as a model class assumption for f can be ameliorated by assuming that the input variable to f (and hence the data as well) is restricted by a probability measure μ on Ω supported on a low-dimensional submanifold.

The second type of approximation result introduces novel high-dimensional model classes for which neural network approximation (NNA) rates do not grow unbounded with the input dimension d. Thus, membership in these new model classes can be a realistic model class assumption for learning a function of many variables. The most celebrated examples of such new model classes are the Barron class $\mathcal{B}^s$, $s > 0$, introduced in [2]. The set $\mathcal{B}^s = \mathcal{B}^s(\mathbb{R}^d)$ consists of all functions f defined on $\mathbb{R}^d$ whose Fourier transform $\hat{f}$ satisfies

$$\|f\|_{\mathcal{B}^s} := \int_{\mathbb{R}^d} (1+|\omega|)^s |\hat{f}(\omega)|\, d\omega < +\infty. \tag{1.2}$$

The original result of Barron showed that on a bounded domain $\Omega \subset \mathbb{R}^d$, any function f on Ω which is the restriction of a function from $\mathcal{B}^1(\mathbb{R}^d)$ can be approximated in the $L_2(\Omega)$ norm by single-hidden-layer sigmoidal networks with n neurons to an accuracy of $C_\Omega \|f\|_{\mathcal{B}^1} n^{-1/2}$, $n \geq 1$, where the constant C_Ω only depends upon the measure of Ω. Notice that this approximation rate does not deteriorate with increasing d in contrast with classical smoothness model classes. However, one must note that the above definition of Barron spaces depend on d and indeed get more demanding as d increases.

Barron's result spurred a lot of study and generalizations over the last decades. In particular, new model classes of functions which have sparse representation of as linear combinations of neural atoms were introduced. In the case of ReLU neurons, the sparsity class is larger than the (second-order) Barron class and yet preserves the rate of approximation of n-term approximation [15]. These spaces based on sparsity are called *variation spaces* [1,22,26,40,41]. We summarize these activities for ReLU neurons in the following two

sections. For the moment, we only wish to focus on the existing theory for these model classes and their approximation rates on domains $\Omega \subset \mathbb{R}^d$. This is the typical setting in applications. The existing theory defines the corresponding model classes on $\mathbb{R}^d$ and then extends the definition to domains as the restriction of functions defined on $\mathbb{R}^d$. As such, the theory and corresponding results are in a strong sense *independent of the domain* Ω. While this leads to a simple approximation theory on domains, these results never take into consideration the nature of Ω, e.g., its geometry.

The purpose of the present paper is to show there is a more satisfactory definition of these novel model classes on domains Ω that leads to domain-dependent results that are stronger than that provided by the existing theory. We call these new model classes *weighted* variation spaces since they generalize the classical variation space for ReLU neurons by introducing a domain-dependent weighting of the ReLU atoms. These new model classes are strictly larger than the existing variation spaces while still maintaining the same rate of approximation of n-term approximation. We develop this domain-dependent theory primarily in the case when $\Omega = B^d$ is the Euclidean unit ball in $\mathbb{R}^d$. To indicate how the theory would depend on the domain Ω, we also consider the domain $Q^d := [-1,1]^d$ and contrast the difference in this case with that of B^d.

While we develop our results only for the case of ReLU neurons, we believe that the techniques developed in this paper can be applied to the case of ReLU^k neurons, $k > 1$. We leave the details to future work. So, for the remainder of this paper the activation function is

$$\sigma(t) = t_+ = \max\{0, t\}. \tag{1.3}$$

This paper is organized as follows. In the next two sections, we review some of the existing results on ReLU neural network approximation. This will serve to frame the new results proved in this paper. In §4 we introduce our new (domain-dependent) model classes. In §5, we prove our new approximation results for $\Omega = B^2$ the unit Euclidean ball in $\mathbb{R}^2$. We separate out this case since it is the simplest setting to understand. The remaining sections of this paper formulate and prove our results for $\Omega = B^d$ which is the Euclidean unit ball in $\mathbb{R}^d$. We also contrast how the results change when $\Omega = Q^d$. Finally, we discuss the possible significance of these new model classes for the problem of learning from data.

2. Approximation by shallow ReLU networks

In this paper, we concentrate on a very specific case of NNA, namely approximation by single-hidden-layer ReLU NNs, i.e., the activation function σ is given by (1.3). We study neural network approximation on a given bounded domain (the closure of an open connected set) Ω of $\mathbb{R}^d$. The most natural choices for Ω are the unit Euclidean ball B^d of $\mathbb{R}^d$ or the d-dimensional cube $Q^d := [-1,1]^d$. The case $\Omega = B^d$ will be the primary example considered in this paper. In going further, we let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^d$.

We define the ReLU atoms

$$\phi(x;\xi,t) := \sigma(\xi \cdot x - t) = (\xi \cdot x - t)_+, \quad \xi \in \mathbb{R}^d,\ \|\xi\| = 1,\ t \in \mathbb{R}. \tag{2.1}$$

Given the atom ϕ, we let

$$H_\phi := \{x \in \Omega : \xi \cdot x = t\} \tag{2.2}$$

be its hyperplane cut. H_ϕ divides Ω into two regions $H_\phi^\pm$. The function ϕ is identically zero on the region $H_\phi^- := \{x \in \Omega :\ \xi \cdot x \le t\}$ and the linear function $= \xi \cdot x - t$ on the second region $H_\phi^+ := \{x \in \Omega :\ \xi \cdot x > t\}$. Notice that for some values of t, the atom ϕ is identically zero on Ω so that $H_\phi^- = \Omega$.

For each Ω, there is a smallest interval $T = T(\Omega)$ such that for $t \notin T$, the dictionary element $\phi(\cdot;\xi,t)$ is either identically zero on Ω or a linear function on Ω. Let $\mathcal{D} = \mathcal{D}(\Omega) := \{\phi(\cdot;\xi,t)\}$ be the dictionary of all atoms ϕ for which $t \in T = T(\Omega)$. We are interested in n-term approximation from the dictionary $\mathcal{D}$. For $n = 1, 2, \dots$, let $\Sigma_n := \Sigma_n(\mathcal{D})$ be the set of functions of the form

$$S(x) = \sum_{j=1}^n a_j \phi_j(x), \quad x \in \Omega, \tag{2.3}$$

where the ϕ_j are chosen arbitrarily from $\mathcal{D}$ and $a_1, \dots, a_n$ are real numbers. When $n = 0$, we define $\Sigma_0 := \{0\}$. The functions $S \in \Sigma_n$ are precisely the functions on Ω produced by a single-hidden-layer ReLU network with n neurons, i.e., width n. The set Σ_n is thus a $(d+2)n$ dimensional parametric nonlinear manifold parameterized by the $\xi_j \in B^d$, $j = 1, \dots, n$, the $t_j \in \mathbb{R}$, $j = 1, \dots, n$, and the coefficients $a_1, \dots, a_n \in \mathbb{R}$. Note that a given $S \in \Sigma_n$ has in general many representations of the form (2.3). In other words, the dictionary $\mathcal{D}(\Omega)$ is redundant.

The above paragraph tells us that there are two ways to view shallow network approximation with ReLU activation. One view is that it is a special case of n-term approximation from a dictionary of functions. Another view is that it is a special case of manifold approximation. Therefore, a proper assessment of this form of NN approximation would be to compare it with other approximation methods of either one of these forms.

Approximation by Σ_n is one of the simplest examples of neural network approximation (NNA). It is therefore a fundamental problem to completely understand the approximation properties of Σ_n, $n \ge 1$, i.e., what are the properties of a function f that determine how well f is approximated by the elements of Σ_n. In the case $d = 1$ and Ω is an interval, the set Σ_n is the space of

piecewise linear function with n breakpoints. In this special case, approximation by Σ_n is well understood (see, e.g., [11,12]). So, we restrict ourselves to the case $d \geq 2$ in going further in this paper.

For a function f in $L_p(\Omega)$, $1 \leq p \leq \infty$, we define

$$E_n(f)_p := E_n(f)_{L_p(\Omega)} := \inf_{S \in \Sigma_n} \|f - S\|_{L_p(\Omega)}. \tag{2.4}$$

This is a form of nonlinear approximation since the set Σ_n is not a linear space but rather a nonlinear manifold. Rightfully, we often put this form of approximation in competition with other examples of manifold approximation (see, e.g., [10,11]).

From the viewpoint of approximation theory, an understanding of the approximation properties of Σ_n would seek to precisely characterize the approximation classes for Σ_n approximation. An approximation class is the collection of all functions whose approximation error decays at a prescribed decay rate. For example, for a given $\alpha > 0$, we seek a characterization of the set

$$\mathcal{A}^\alpha := \mathcal{A}^\alpha((\Sigma_n)_{n\geq 0}, L_p(\Omega)) \tag{2.5}$$

of functions $f \in L_p(\Omega)$ for which

$$E_n(f)_p \leq M(n+1)^{-\alpha}, \quad n = 0, 1, 2, \dots. \tag{2.6}$$

Note that by definition, $\Sigma_0 = \{0\}$ and hence $E_0(f)_p = \|f\|_{L_p(\Omega)}$ The smallest value of M for which (2.6) holds is defined as $\|f\|_{\mathcal{A}^\alpha}$. Notice that $\mathcal{A}^\alpha$ is a quasi-normed linear space. While for most classical methods of linear and nonlinear approximation, e.g. polynomials, splines, n-term wavelets, there is a characterization of the spaces $\mathcal{A}^\alpha$ (at least for a certain range of α), the case for neural network approximation is much different. There is at present no known characterization of $\mathcal{A}^\alpha$ for any value of $\alpha > 0$. There are however many sufficient conditions that guarantee membership in $\mathcal{A}^\alpha$ (see [11]).

Another (less ambitious) viewpoint of approximation by Σ_n is to propose model classes K, i.e., compact subsets $K \subset L_p(\Omega)$, and study how well the elements of K can be approximated by the elements of Σ_n. This leads to the study of

$$E_n(K)_p := \sup_{f \in K} E_n(f)_p, \quad n \geq 0. \tag{2.7}$$

If one comes up with a set K for which $E_n(K) \leq Cn^{-\alpha}$, $n \geq 1$, then clearly $K \subset \mathcal{A}^\alpha$ and we gain some information about $\mathcal{A}^\alpha$. Many interesting approximation results have been proven for various classical model classes K such as Sobolev and Besov balls, however, the best approximation rates are not known in all cases [11]. These results show no gain in approximation efficiency when compared with more classical methods of approximation such as those that use splines or wavelets. Moreover, these classical model classes all suffer the curse of dimensionality: smoothness of order s gives rate decay $E_n(K)_p \geq Cn^{-s/d}$, $n \geq 0$.

One of the celebrated accomplishments in the study of NNA was the introduction of new model classes K whose NNA rates do not grow unbounded with the input dimension d. They also give us information on $\mathcal{A}^\alpha$. We discuss these model classes in the next two sections. In going further in this paper we only treat the case of approximation in $L_2(\Omega)$. However, the case of $L_p(\Omega)$ approximation has also been well studied (see [40]).

3. Novel (non-classical) model classes

While the classical model classes based on smoothness all suffer the curse of dimensionality, certain novel model classes K have been introduced whose rates do not grow unbounded with the input dimension. The discovery of these novel model classes begin with the celebrated work of Barron [2]. We have already defined the Barron spaces $\mathcal{B}^s(\mathbb{R}^d)$ in the introduction.

Barron's original results on NNA were for sigmoidal activation and the Barron class $\mathcal{B}^1(\mathbb{R}^d)$ where he showed that functions in this class, when restricted to a domain $\Omega \subset \mathbb{R}^d$, had an $L_2(\Omega)$ approximation rate $n^{-1/2}$, $n \geq 1$. It was rather straightforward to extend his approach to proving that functions in $\mathcal{B}^2$ had the same approximation rate when using ReLU activation. Several follow up papers significantly improved on these original results as we now describe.

Notice that the Barron classes are formulated for functions which are defined on all of $\mathbb{R}^d$. Given a bounded domain Ω, it is not obvious how these classes should be defined on Ω. The definition employed in the literature is that the space $\mathcal{B}^s(\Omega)$ is the set of function f defined on Ω which are the restriction of a function $F \in \mathcal{B}^s(\mathbb{R}^d)$ with norm given by

$$\|f\|_{\mathcal{B}^s(\Omega)} := \inf_{F|_\Omega = f} \|F\|_{\mathcal{B}^s(\mathbb{R}^d)}, \quad s > 0. \tag{3.1}$$

With this definition, we have

$$E_n(U(\mathcal{B}^2(\Omega))_{L_2(\Omega)} \leq Cn^{-1/2}, \quad n \geq 1, \tag{3.2}$$

where C depends only on the diameter and measure of Ω. Here and later we use the notation $U(Y)$ to denote the unit ball of a normed space Y. This approximation rate was improved over the years starting with Makovoz [24] and continuing on with the results of [1,19,40]. The current best known approximation rate for n-term ReLU NNA is

$$E_n(U(\mathcal{B}^2(\Omega))_{L_2(\Omega)} \leq Cn^{-\frac{1}{2}-\frac{3}{2d}}, \quad n \geq 1, \tag{3.3}$$

where again C depends only on d. We refer the reader to [40] for a more detailed discussion of these approximation results. It is still not known if this rate can be improved for the Barron class $\mathcal{B}^2$.

We turn next to a second family of novel model classes for NNA referred to as variation spaces. Let $\mathcal{D} = \mathcal{D}(\Omega)$ be the dictionary of ReLU atoms whose hyperplane cut intersects Ω. Consider any function $S = \sum_{j=1}^n a_j \phi_j$, i.e., $S \in \Sigma_n$. Recall that this representation is not unique. We define

$$V(S) := \inf \left\{ \sum_{j=1}^n |a_j| : S = \sum_{j=1}^n a_j \phi_j \right\}, \tag{3.4}$$

which is called the variation of S with respect to the dictionary $\mathcal{D}$.

With this notation in hand, we can define a new space $\mathcal{V} := \mathcal{V}(\Omega) = \mathcal{V}(\Omega, \mathcal{D})$ as the set of all f in $L_2(\Omega)$ for which there is a sequence $S_n \in \Sigma_n$, $n \geq 1$, such that $\|f - S_n\|_{L_2(\Omega)} \to 0$, $n \to \infty$, and $V(S_n) \leq M$, $n \geq 1$. Throughout the paper, we will use $\mathcal{V}$ when the domain Ω and dictionary $\mathcal{D}$ are clear from the context, and use $\mathcal{V}(\Omega)$, $\mathcal{V}(\mathcal{D})$, or $\mathcal{V}(\Omega, \mathcal{D})$ when we want to call attention to the domain and/or dictionary. The smallest M for which this is true is defined as $\|f\|_{\mathcal{V}(\Omega)}$. This space is called the *variation space* of the dictionary $\mathcal{D}$. The space $\mathcal{V}(\Omega)$ is a Banach space with respect to this norm (see [41] for properties of variation spaces). A fundamental relation between the Barron and variation space is the embedding

$$\|f\|_{\mathcal{V}(\Omega)} \leq C_\Omega \|f\|_{\mathcal{B}^2(\Omega)}, \quad f \in \mathcal{B}^2(\Omega), \tag{3.5}$$

with C_Ω the embedding constant (which depends only on the diameter of Ω). The space $\mathcal{V}(\Omega)$ is strictly larger than $\mathcal{B}^2(\Omega)$. We remark that the variation space $\mathcal{V}(\Omega)$ has also been introduced under other names such as the $\mathcal{F}_1$ space [1] and the Barron space [15].

The variation space $\mathcal{V}(\Omega)$ has been carefully studied and in particular it has been proven that (see [40])

$$E_n(U(\mathcal{V}(\Omega)))_{L_2(\Omega)} \leq C n^{-\frac{1}{2} - \frac{3}{2d}}, \quad n \geq 1, \tag{3.6}$$

where C depends only on Ω and d. This approximation rate also matches the decay rate of the metric entropy of $U(\mathcal{V}(\Omega))$ [40]. Notice that this gives the bound (3.3) and is in fact how approximation rates for the Barron class are proved. The important thing to note here is that $\mathcal{V}$ is a larger space than $\mathcal{B}^2$ but the current best known approximation rates (with shallow ReLU NNs) for both of these classes is the same, namely $O(n^{-\frac{1}{2} - \frac{3}{2d}})$, $n \geq 1$.

We remark that the rate (3.6) has also been obtained in the L_∞-norm (on the sphere) in [1] using deep results from geometric discrepancy theory [6,7,25], although a gap exists in dimensions $d = 2, 3$, which was apparently overlooked by the author of [1]. Recently, this gap has been completely closed and these results have been generalized to ReLU^k networks for all $k \geq 0$ in [38]. Similar uniform approximation rates have also been obtained using an entirely different method for a smaller class of functions in [27]. Similar results have also been investigated for ReLU networks whose inputs and outputs take values in Banach spaces [20]. In that work, it is shown that the n-term approximation rate is bounded by $O(n^{-\frac{1}{2}})$.

A major breakthrough in the understanding of $\mathcal{V}(\Omega)$ was made by characterizing membership of a function f in $\mathcal{V}(\Omega)$ through the smoothness of its Radon transform. Namely, it was originally proved in [29] that a function f is in $\mathcal{V}(\Omega)$ if and only if f has an extension F to all of $\mathbb{R}^d$ such that the Radon transform $\mathcal{R}(F; \xi, t)$ is in a certain smoothness space. Properties and generalizations of this notion of smoothness were extensively studied in [30,31,33], giving rise to a new family of Banach spaces, now referred to as the Radon-domain BV spaces. These spaces are denoted by $\mathcal{R}\text{BV}^k$, $k \in \mathbb{N}$.

The key result of [30] is the following *representer theorem* for these spaces. Let $x_i \in \mathbb{R}^d$, $i = 1, \dots, m$, and $y_i \in \mathbb{R}$, $i = 1, \dots, m$. Then, there always exists a solution to the data-fitting problem

$$\min_{f \in \mathcal{R}\text{BV}^k} \sum_{i=1}^m \mathcal{L}(y_i, f(x_i)) + \lambda |f|_{\mathcal{R}\text{BV}^k} \tag{3.7}$$

that takes the form of a function S which is the output of a single-hidden-layer neural network with $\leq m$ neurons and ReLU^{k-1} activation functions. Here, $\mathcal{L}$ is any loss function which is lower-semi-continuous in its second argument and $|f|_{\mathcal{R}\text{BV}^k}$ is the semi-norm which defines the $\mathcal{R}\text{BV}^k$ spaces, which measures smoothness in the Radon domain. The Radon BV spaces are defined on domains $\Omega \subset \mathbb{R}^d$ via restrictions. For the case $k = 2$ (which corresponds to shallow ReLU NNs) it has been shown in [33, Theorem 6] (see also [41, Theorem 2 and Corollary 1]) that

$$\mathcal{R}\text{BV}^2(\Omega) = \mathcal{V}(\Omega), \tag{3.8}$$

with equivalent norms. It has also been shown that there exists a solution S to (3.7) which is in $\Sigma_m(\mathcal{D})$ on any bounded domain $\Omega \subset \mathbb{R}^d$ [33, Theorem 5].

4. Weighted variation model classes

One of the main points of the present paper is that one can derive improved results on approximation by shallow ReLU networks if one considers new model classes that generalize the standard variation space by including weights on the atoms. In this section, we introduce these new model classes for the case when we want the error of approximation to be taken in the $L_2(\Omega)$ norm with Ω a bounded domain in $\mathbb{R}^d$. We begin with the general principle of *weighted variation spaces*.

Let $\mathcal{D}$ be the dictionary of ReLU atoms. Let S^{d-1} be the boundary of the unit Euclidean ball B^d of $\mathbb{R}^d$. That is, $S^{d-1} := \{\xi \in \mathbb{R}^d : \|\xi\| = 1\}$. Any atom ϕ in $\mathcal{D}$ is of the form $\phi(x) = (\xi \cdot x - t)_+$ where $t \in \mathbb{R}$. We are interested in the atoms ϕ whose hyperplane cut intersects Ω (since otherwise the atom is identically an affine function). Accordingly, we define

$$Z(\Omega) := \{(\xi, t) : \xi \in S^{d-1},\ t \in \mathbb{R} \text{ such that } H_{\phi(\cdot;\xi,t)} \cap \Omega \neq \emptyset\} \tag{4.1}$$

and $\bar{Z}(\Omega)$ its closure in the Euclidean norm. Note that whenever $\phi(x) = \phi(x;\xi,t)$, $(\xi,t) \in Z(\Omega)$, is positive for some $x \in \Omega$, it is positive in a neighborhood of x and hence $\|\phi\|_{L_2(\Omega)} > 0$. Given the domain Ω we define the dictionary

$$\mathcal{D}(\Omega) := \{\phi(\cdot,\xi,t) :\ (\xi,t) \in \bar{Z}(\Omega)\}. \tag{4.2}$$

The set $\bar{Z}(\Omega)$ is a compact subset of $S^{d-1} \times \mathbb{R}$. If we equip $\bar{Z}(\Omega)$ with the Euclidean norm topology then the mapping $(\xi,t) \mapsto \phi(\cdot;\xi,t)$ is a continuous mapping from $\bar{Z}(\Omega)$ into $L_2(\Omega)$.

Here is an important observation about the atoms in this dictionary which underlies the improved approximation results of this paper. While each atom $\phi \in \mathcal{D}(\Omega)$ is in $L_2(\Omega)$ whenever Ω is a bounded domain, the $L_2(\Omega)$ norm of ϕ will depend heavily on ϕ and Ω. Namely, if the support of ϕ lies near the boundary of Ω then this norm will be small and we expect that ϕ has a less important role in approximating a given target function $f \in L_2(\Omega)$.

As an example, consider the case when $\Omega = B^d$ is the d-dimensional Euclidean ball. It is easy to see that the atom $\phi(x) = (\xi \cdot x - t)_+$, has $L_2(\Omega)$-norm satisfying

$$\|\phi\|_{L_2(\Omega)} \approx (1-t)^{\frac{3}{2}+\frac{d-1}{4}}, \quad -1 \le t \le 1, \tag{4.3}$$

with constants of equivalence depending only on d. Indeed, the $L_\infty(\Omega)$ norm of ϕ is $1-t$ and the measure of its support $\approx (1-t)[\sqrt{1-t}]^{d-1}$. It follows that the norms of atoms get smaller as t approaches one.

The compactness of $\bar{Z}(\Omega)$ implies that the dictionary $\mathcal{D}(\Omega)$ is a compact subset of $L_2(\Omega)$. Thus, there is another useful characterization of the functions in $\mathcal{V}(\Omega)$. Consider the space $\mathcal{M} := \mathcal{M}(\bar{Z}(\Omega))$ of all finite (signed) Radon measures on $\bar{Z}(\Omega)$, equipped with the variation norm $\|\mu\|_{\mathcal{M}} := \int\limits_{\bar{Z}(\Omega)} d|\mu|$. For $\mu \in \mathcal{M}$, we introduce the function

$$f_\mu := \int\limits_{\bar{Z}(\Omega)} \phi(\cdot;\xi,t)\, d\mu(\xi,t), \tag{4.4}$$

where the integral in (4.4) can be understood as a Bochner integral (see [41, Lemma 3] for more details). Then, any $f \in \mathcal{V}(\Omega)$ has a representation

$$f = f_\mu, \quad \text{for some } \mu \in \mathcal{M}. \tag{4.5}$$

This representation is not unique in the sense that different measures μ can give rise to the same f. It then follows (see [41]) that the $\mathcal{V}$-norm can be alternatively specified by

$$\|f\|_{\mathcal{V}} = \inf\{\|\mu\|_{\mathcal{M}} :\ f = f_\mu,\ \mu \in \mathcal{M}\}. \tag{4.6}$$

In order to simplify the geometry, in going further in this section, we assume that Ω is a convex subset of $\mathbb{R}^d$ and $\mathcal{D} := \mathcal{D}(\Omega)$. We say that

$$w(\xi,t), \quad (\xi,t) \in \bar{Z}(\Omega), \tag{4.7}$$

is a weight function if w is a non-negative continuous function on $\bar{Z}(\Omega)$. Given an atom $\phi(\cdot;\xi,t)$ we will abuse notation and also write $w(\phi)$ or $w(\phi(\cdot;\xi,t))$ for $w(\xi,t)$.

Admissible Weights: *Given a weight function w defined on $\bar{Z}(\Omega)$, we define*

$$\tilde{\phi}(\cdot;\xi,t) := \frac{\phi(\cdot;\xi,t)}{w(\xi,t)}, \quad (\xi,t) \in \bar{Z}(\Omega), \tag{4.8}$$

where $\tilde{\phi}(\cdot;\xi,t)$ is defined to be the zero function whenever $w(\xi,t) = 0$. We say that the weight function w is admissible for Ω, if the mapping $(\xi,t) \to \tilde{\phi}(\cdot;\xi,t)$ is continuous as a mapping from $\bar{Z}(\Omega)$ into $L_2(\Omega)$. It follows that

$$\|\tilde{\phi}(\cdot;\xi,t)\|_{L_2(\Omega)} \le C_w, \tag{4.9}$$

with C_w an absolute constant. Notice that if a weight function w is admissible, then any larger weight function $\tilde{w}$ is also admissible.

When given an admissible weight w, the set of functions

$$\mathcal{D}_w := \mathcal{D}_w(\Omega) = \{\tilde{\phi}(\cdot;\xi,t) :\ (\xi,t) \in \bar{Z}(\Omega)\}, \tag{4.10}$$

is a new dictionary contained in $L_2(\Omega)$. Furthermore, this dictionary is compact in $L_2(\Omega)$. We define the weighted variation space $\mathcal{V}_w := \mathcal{V}_w(\Omega)$ to be variation space of this new dictionary $\mathcal{D}_w$. Since the admissibility conditions ensure that the dictionary $\mathcal{D}_w$ is

compact in $L_2(\Omega)$, we have, from the discussion above, that, for every $f \in \mathcal{V}_w(\Omega)$, there exists a signed Radon measure $\mu = \mu_f$ on $\bar{Z}(\Omega)$ such that

$$f = \tilde{f}_\mu := \int_{\bar{Z}(\Omega)} \tilde{\phi}(\cdot;\xi,t)\, d\mu(\xi,t) \quad \text{with} \quad \|f\|_{\mathcal{V}_w(\Omega)} = \|\tilde{f}_\mu\|_{\mathcal{V}_w(\Omega)} = \|\mu_f\|_{\mathcal{M}}. \tag{4.11}$$

We also clearly have

$$\|f\|_{L_2(\Omega)} \le C_w \|f\|_{\mathcal{V}_w(\Omega)}, \quad f \in \mathcal{V}_w(\Omega), \tag{4.12}$$

where C_w is the constant in (4.9). We also have that, if $\tilde{w} \ge w$, then $\mathcal{V}_{\tilde{w}}(\Omega) \subset \mathcal{V}_w(\Omega)$ and $\|f\|_{\mathcal{V}_{\tilde{w}}} \le \|f\|_{\mathcal{V}_w(\Omega)}$ which implies that

$$E_n(U(\mathcal{V}_{\tilde{w}}(\Omega))) \le E_n(U(\mathcal{V}_w(\Omega))), \quad n \ge 0, \tag{4.13}$$

where we note that $\Sigma_n(\mathcal{D}) = \Sigma_n(\mathcal{D}_w)$ for any admissible w.

While $\mathcal{V}_w(\Omega)$ is defined for any nonnegative weight w which is admissible, there is a particular choice of w which we will consider in this paper. Specifically, we show that the approximation rates derived for shallow ReLU neural networks on the unweighted space $\mathcal{V}(\Omega)$ actually hold on the larger space $\mathcal{V}_w(\Omega)$ for a certain collection of admissible weights w. As we will later see, the smallest admissible weight with this property will depend upon the domain Ω. This domain-dependent smallest weight is related to the measure of the intersection of the hyperplane of ϕ restricted to Ω. To describe this particular weight w and our new approximation results, we start with the case $d = 2$ where the proofs of approximation rates are simplest to understand. We consider the two domains $\Omega = B^2$ and $\Omega = Q^2$. Later, we treat the general cases $\Omega = B^d$, $d \ge 2$. We then explain how the same theory carries over to $\Omega = Q^d$ (see Remark (6.8)).

Variation spaces $\mathcal{V}(\mathcal{D}_0)$ are defined as above for any dictionary $\mathcal{D}_0$ in any Hilbert space H provided that the dictionary elements $\psi \in \mathcal{D}_0$ satisfy $\|\psi\|_H \le \delta$ for a fixed value of $\delta > 0$. Given such a dictionary $\mathcal{D}_0$, we define $\Sigma_n := \Sigma_n(\mathcal{D}_0)$ as the set of all functions $S \in H$ that are a linear combination of at most n terms of $\mathcal{D}_0$. For any $f \in H$, we define the error of n term approximation to be

$$E(f,\Sigma_n)_H := \inf_{S \in \Sigma_n} \|f - S\|_H. \tag{4.14}$$

This n-term approximation error from a dictionary is well studied. A fundamental result for such n-term approximation is the theorem of Maurey [35] (see also [2,18]). Maurey's theorem says that for each $n \ge 0$ and $f \in \mathcal{V}(\mathcal{D}_0)$ we have

$$\inf_{S_n \in \Sigma_n} \|f - S_n\|_H \le \|f\|_{\mathcal{V}(\mathcal{D}_0)} \delta n^{-1/2}, \quad n \ge 1. \tag{4.15}$$

In fact, Maurey's theorem can be generalized beyond the setting of a Hilbert space to the class of type-2 Banach spaces (see [40] for the application to non-linear dictionary approximation). This introduces an extra constant factor which depends upon the type-2 constant of the space. We shall use this theorem going forward, but restrict ourselves to the Hilbert space setting.

5. Approximation in $\Omega = B^2$

In this section, we develop our results in the case $\Omega = B^2$ where B^2 is the unit Euclidean ball in $\mathbb{R}^2$. Here, $\bar{Z}(\Omega) = S^1 \times [-1,1]$. This will illustrate, in their simplest form, all of the principles needed to treat the more general case $\Omega = B^d$, $d \ge 2$. The treatment of B^d is given in §6 but with a significant increase in the level of technicality.

In this section, we let $\mathcal{D} = \mathcal{D}(\Omega)$ be the ReLU dictionary of atoms $\phi = \phi(\cdot;\xi,t)$, $\xi \in S^1$ and $t \in [-1,1]$. Note that since $d = 2$, the hyperplane H_ϕ associated to the atom ϕ is a line and $L_\phi := H_\phi \cap \Omega$ is a line segment whose length is $|L_\phi| = (1-t^2)^{1/2}$. We define the weight of this atom by

$$w(\phi) = w(\phi(\cdot;\xi,t)) := 1 - t, \quad t \in [-1,1]. \tag{5.1}$$

It is easy to check that this weight is admissible since $\|\phi(\cdot;\xi,t)\|_{L_2(\Omega)} \approx (1-t)^{7/4}$ (see (4.3)). We discuss where this weight comes from in §6 in the sequel.

We first want to prove results on the linear approximation of the atoms ϕ. Namely, for each $n = 1,2,\dots$, we want to construct an n dimensional linear space X_n which is good at approximating all of the atoms $\phi \in \mathcal{D}(\Omega)$. The linear space X_n will be the span of n well chosen atoms ϕ_j, $j = 1,\dots,n$, from $\mathcal{D}(\Omega)$. The construction we give for X_n is a modification of ideas from [40]. Our analysis of the approximation error in approximating ϕ by the elements of X_n is new in that it gives an improved error estimate when the support of ϕ is near the boundary of Ω.

To define the space $X_n := \operatorname{span}\{\phi_1,\dots,\phi_n\}$, we want to choose the atoms ϕ_j, $j = 1,\dots,n$, to have as a special discrete distribution from $\mathcal{D}$. In the case $d = 2$, these atoms are rather easy to describe geometrically as is given in the next paragraph. When $d > 2$, we will need more sophisticated arguments (see §6).

We fix $m \ge 4$ and let $P = P_m$ be the set of points

$$\mu_j = \mu_j(m) := (\cos\theta_j, \sin\theta_j), \quad \theta_j = \theta_{j,m} := \frac{2\pi j}{m}, \ j \in \mathbb{Z}. \tag{5.2}$$

There are m distinct points and $\mu_j = \mu_{j'}$ if j and j' are congruent modulo m, i.e., if $j \equiv j'$. These points are equally spaced on the circle.

Let X_n be the linear space spanned by the dictionary elements ϕ whose line segment L_ϕ has end points μ_i and μ_j, $1 \le i < j \le m$. Notice that for each pair i, j there are two such atoms. Hence, the dimension of X_n is $n := m(m-1)$. We also note that X_n contains all linear functions on Ω.

Given $i, j \in \mathbb{Z}$, we define the distance between i and j by

$$d(i,j) := \min\{|i' - j'| : \ i \equiv i', \ j \equiv j'\},$$

i.e. to be the periodic distance between the indices i and j.

Let $\mathcal{L}_{i,j} = \mathcal{L}_{i,j}(m)$ be the set of all line segments L whose end points a, b are the points $(\cos\theta, \sin\theta)$ where $\theta \in [\theta_i, \theta_{i+1}]$ in the case of a and $\theta \in [\theta_j, \theta_{j+1}]$ in the case of b. We denote by $S_{i,j} = S_{i,j}(m)$ the union of all the line segments L_ϕ in $\mathcal{L}_{i,j}$.

Note that the length $L_{i,j}$ and width $W_{i,j}$ of $S_{i,j}$ satisfy

$$|L_{ij}| \approx \frac{d(i,j)+1}{m}, \quad |W_{ij}| \approx \frac{d(i,j)+1}{m^2}, \quad 1 \le i \le j \le m. \tag{5.3}$$

Here and later in this section, all constants of equivalence are absolute. It follows that the measure of $S_{i,j}$ satisfies

$$|S_{i,j}| \lesssim \frac{(d(i,j)+1)^2}{m^3}, \quad 1 \le i \le j \le m. \tag{5.4}$$

Lemma 5.1. *Suppose that $m \ge 4$ is an even integer, $n = m(m-1)$, and $\phi = \sigma(\cdot;\xi,t)$ is any dictionary element whose line segment L_ϕ is in $\mathcal{L}_{i,j} = \mathcal{L}_{i,j}(m)$ with $\mu_i \ne \mu_j$. Then there is a function $g \in X_n$ such that*
(i) $\phi(x) = g(x), \ x \notin S_{i,j}$,
(ii) $\|\phi - g\|_{L_\infty(\Omega)} \le C\frac{d(i,j)}{m^2}$, *with C an absolute constant.*
(iii) $\|\phi - g\|_{L_2(\Omega)} \le C w(\phi) n^{-3/4}$, *with C an absolute constant.*
If $\phi \in \mathcal{L}_{i,i}$ for some i, then there is a $g \in X_n$ such that statement (iii) *holds.*

Proof. We first assume that $0 \le i < j \le m$. Also, by reversing the roles of i and j if necessary, we can also assume that $j < m/2$. Because of rotational symmetry we can assume that $i = 0$, $i+1 = 1$, $0 < j < m/2$. Consider the linear function $\ell(x) := \xi \cdot x - t$. Let the line segment $L_\phi = H_\phi \cap \Omega$ associated with ϕ be in $\mathcal{L}_{i,j}$. Let $\mu_i = \mu_i(m)$, $i \in \mathbb{Z}$. We use the following three functions ϕ_1, ϕ_2, ϕ_3 in X_n each of whose line segments L_{ϕ_i} are contained in $\mathcal{L}_{i,j}$. Here, L_{ϕ_1} has endpoints μ_i, μ_{j+1}, the second segment L_{ϕ_2} has end points μ_i, μ_j, and the third function ϕ_3 has line segment L_{ϕ_3} with endpoints μ_{i+1}, μ_{j+1}. The orientation of these three atoms matches that of ϕ. By this we mean that whenever $x \in \Omega$ is strictly outside $S_{i,j}$ and $\phi(x) > 0$ then each of the functions ϕ_i, $i = 1,2,3$, will likewise be positive. Similarly, if x is strictly outside this strip and $\phi(x) = 0$ the three functions ϕ_i, $i = 1,2,3$, will likewise vanish.

Consider the three linear functions ℓ_j, $j = 1,2,3$, corresponding to these line segments. That is, we have $\ell_i(x) = \xi_i' \cdot x - t_i'$ and $\phi_i(x) = \ell_i(x)_+$ with $\xi_i' \in S^1$ and $t_i' \in [-1,1]$. Since these three linear functions are linearly independent, we can write

$$\ell = c_1\ell_1 + c_2\ell_2 + c_3\ell_3. \tag{5.5}$$

Specifically, let ζ be the point where $\ell_2(\zeta) = \ell_3(\zeta) = 0$. Then,

$$c_1 = \frac{\ell(\zeta)}{\ell_1(\zeta)}, \quad c_2 = \frac{\ell(\mu_{j+1})}{\ell_2(\mu_{j+1})}, \quad c_3 = \frac{\ell(\mu_i)}{\ell_3(\mu_i)}. \tag{5.6}$$

This follows by noting that with this choice (5.5) holds at the affinely independent set of points ζ, μ_{j+1} and μ_i.

We claim that

$$|c_i| \le 1, \quad i = 1,2,3. \tag{5.7}$$

Indeed, since the ξ, ξ_i lie on the sphere it is clear that $|\ell_i(x)| = d(x, L_{\phi_i})$ and $|\ell(x)| = d(x, L_\phi)$ for any $x \in \mathbb{R}^2$ (here $d(x,L)$ denotes the distance from the point x to the line L). We will show that

$$d(\mu_i, L_\phi) \le d(\mu_i, L_{\phi_3}), \tag{5.8}$$

which implies $|c_3| \le 1$. A completely analogous argument shows that $|c_2| \le 1$.

For the proof of (5.8), we assume that $j > i+1$. If $j = i+1$, a similar argument applies (which we leave to the reader). Consider the trapezoid whose vertices are $\mu_i, \mu_{i+1}, \mu_j, \mu_{j+1}$ and let T denote its interior. Let $\bar{\mu}_i$ denote the orthogonal projection of μ_i onto the line L_{ϕ_3}. The angle formed by the vertices $\mu_{j+1}, \mu_{i+1}, \mu_i$ is larger than or equal to $\pi/2$. This means that $\bar{\mu}_i$ lies either on or outside of the circle. By the defining property of L_ϕ this line must intersect the segment $[\mu_i, \bar{\mu}_i]$. Therefore, $d(\mu_i, L_\phi) \le d(\mu_i, L_{\phi_3})$ which proves (5.8) as desired.

Next, we consider bounding $|c_1|$. The line segments $[\mu_i, \mu_{j+1}]$ and $[\mu_{i+1}, \mu_j]$ are parallel, and the intersection point ζ lies on the perpendicular line L_p connecting the midpoints of these two line segments. Moreover, the lengths of these segments satisfy

$l([\mu_i,\mu_{j+1}]) > l([\mu_{i+1},\mu_j])$. This means that the distance from ζ to L_{ϕ_1} is greater than the distance to the parallel line segment $[\mu_{i+1},\mu_j]$. Finally, since $L_\phi \in \mathcal{L}_{i,j}$, L_ϕ must intersect L_p, which implies that $d(\zeta,L_\phi) \le d(\zeta,L_{\phi_1})$. This means that $|c_1| \le 1$ as claimed.

Now, consider the function

$$g := c_1\phi_1 + c_2\phi_2 + c_3\phi_3, \tag{5.9}$$

which is in the linear space X_n. This function agrees with ϕ outside $S_{i,j}$ so that (i) is satisfied. Each of the functions ϕ and g have $L_\infty(S_{i,j})$ norm not exceeding the width $W_{ij} \le C\frac{d(i,j)}{m^2}$ (they are 1-Lipschitz and vanish on one edge) and so the upper bound in (ii) follows. The function $\phi - g$ is supported on $S_{i,j}$ and we have

$$\|\phi - g\|_{L_2(\Omega)} \le \|\phi - g\|_{L_\infty(\Omega)}|S_{i,j}|^{1/2} \le Cd(i,j)^2m^{-2-3/2} \le C|L_{ij}|^2m^{-1-1/2}. \tag{5.10}$$

Note that in this calculation we have use that $d(i,j) \approx (d(i,j)+1)$. Since $d(i,j) > 1$, we easily see that $|L_{ij}| \approx |L_\phi| = w(\phi)$, which verifies (iii).

Finally, if L_ϕ is in $\mathcal{L}_{i,j}$ with $d(i,j) \le 1$ then the conclusion follows in the same way we proved (5.10) by taking either $v = 0$ or $v = w\cdot x + b$ to be linear function which matches the linear part of ϕ. □

5.1. The approximation theorem

Throughout this section $E_n(f) := E_n(f)_{L_2(\Omega)}$, $n \ge 1$ for any $f \in L_2(\Omega)$. We can now state the main theorem to be proved in this section.

Theorem 5.2. *Let $\Omega = B^2$ and $w(\phi)$, $\phi \in D(\Omega)$, be defined by (5.1). Then for any $f \in \mathcal{V}_w$, we have*

$$E_n(f) \le C\|f\|_{\mathcal{V}_w(\Omega)}n^{-\frac{5}{4}}, \quad n \ge 1, \tag{5.11}$$

where C is an absolute constant.

Proof. Since $\Sigma_n \subset \Sigma_{n+1}$, $n \ge 0$, it is enough to prove the theorem for any $n = m(m-1)$ with $m \ge 4$ an even integer. This means that we can apply Lemma 5.1. It is enough to prove the theorem for any function f from $U(\mathcal{V}_w(\Omega))$. According to the definition of $\mathcal{V}_w(\Omega)$, for N sufficiently large, there is an $S \in \Sigma_N$ with $S = \sum_{j=1}^N a_j\phi_j$ such that

$$\|f - S\|_{L_2(\Omega)} \le n^{-5/4} \quad \text{and} \quad \sum_{j=1}^N w(\phi_j)|a_j| \le 1. \tag{5.12}$$

For each j, let $g_j \in X_n$ approximate the function ϕ_j appearing in the representation of S according to (iii) of Lemma 5.1. That is, we have

$$\|\phi_j - g_j\|_{L_2(\Omega)} \le C_0w(\phi_j)n^{-3/4}, \tag{5.13}$$

with C_0 an absolute constant. The function $g := \sum_{j=1}^N a_jg_j$ is in X_n and hence in Σ_n. We write

$$f = f - S + h + g, \quad h := S - g. \tag{5.14}$$

Therefore,

$$E_{3n}(f) \le n^{-5/4} + E_{2n}(h). \tag{5.15}$$

We want to bound $E_{2n}(h)$. We have $h = \sum_{j=1}^N a_j[\phi_j - v_j]$. We consider the dictionary $\mathcal{D}' = \{\psi_j\}_{j=1}^N$ with $\psi_j := w(\phi_j)^{-1}(\phi_j - g_j)$. According to (5.12) and (5.13), each ψ_j has $L_2(\Omega)$ norm at most $C_0n^{-3/4}$ and $h = \sum_{j=1}^N c_j'\psi_j$ with $\sum_{j=1}^N |c_j'| \le 1$. It follows from Maurey's theorem (see (4.15)) that h can be approximated by a sum T of n terms from the dictionary $\mathcal{D}'$ with error

$$\|h - T\|_{L_2(\Omega)} \le Cn^{-3/4}n^{-1/2} = Cn^{-5/4}, \tag{5.16}$$

with C an absolute constant. The function T is a sum of at most $2n$ terms from the original dictionary $\mathcal{D}$. Hence,

$$E_{2n}(h) \le Cn^{-5/4}. \tag{5.17}$$

If we place this inequality back into (5.15), we obtain

$$E_{3n}(f) \le [1 + C]n^{-5/4} \tag{5.18}$$

and the theorem follows. □

We close this section with two remarks that clarify Theorem 5.2.

Remark 5.3. We emphasize that Theorem 5.2 is an improvement on the known theorem that any $f \in \mathcal{V}(D)$ satisfies $E_n(f) \le Cn^{-5/4}$ because the weighted variation space $\mathcal{V}_w$ is strictly larger than the standard variation space $\mathcal{V}$.

Remark 5.4. While Theorem 5.2 only applies to the approximation rate $O(n^{-\alpha})$, when $\alpha = 5/4$, there is a standard technique to obtain results for more general rates $O(n^{-\beta})$, for any $\beta \le 5/4$, by considering the interpolation spaces between $L_2(\Omega)$ and $\mathcal{V}_w(\Omega)$ as is explained in [14] and [3]. This approach gives new sufficient conditions for membership in $\mathcal{A}^\beta$. This remark also applies to later results in this paper. We do not elaborate further on this point.

5.2. Weighted variation spaces for $\Omega = Q^2$

Although we do not formulate a general result, it will be clear that the techniques of this paper can be generalized to any convex domain Ω. In this section, we want to point out what such a result is for $Q^2 := [-1,1]^2$ since this will allow us to see the effect of the geometry of Ω. So, in going further in this section, we take $\Omega = Q^2$.

If ϕ is a ReLU atom, then the line segment L_ϕ relative to Ω is $H_\phi \cap \Omega$. The length $|L_\phi|$ can now be large even if L_ϕ is close to the boundary of Ω, for example when L_ϕ is parallel to one of the sides of Ω. In other words, many fewer atoms ϕ will have small $|L_\phi|$.

Let us sketch how the results and analysis for approximating general atoms ϕ given in §5.1 for B^2, changes in this case. We now take a set of $m \sim \sqrt{n}$ equally spaced points on the boundary of Q^2. We can associate each ϕ to a $\mathcal{L}_{i,j}$ similar to the case of B^2 and create a linear space X_n of dimension $m^2 \sim n$ as before. Now the analogue of Lemma 5.1 says that any dictionary element ϕ can be approximated by an element of $g \in X_n$ to an accuracy (corresponding to (iii) in that lemma)

$$\|\phi - g\|_{L_2(\Omega)} \le C_0 m^{-1} [|L_\phi| m^{-1}]^{1/2} = C_0 m^{-3/2} |L_\phi|^{1/2}. \tag{5.19}$$

Here the factor m^{-1} reflects the L_∞ error and the bracketed factor is the measure of the support where ϕ and g differ.

Given the above calculations, we define $w(\phi) := |L_\phi|^{1/2}$ as the weight of the atom ϕ and use this weight to define $\mathcal{V}_w(Q^2)$. The proof of Theorem 5.2 now gives

Theorem 5.5. *Let $d = 2$ and $\Omega = Q^2$ and define $w(\phi) := |L_\phi|^{1/2}$. Then for any $f \in \mathcal{V}_w$, we have*

$$E_n(f) \le C \|f\|_{\mathcal{V}_w(Q^2)} n^{-\frac{5}{4}}, \quad n \ge 1, \tag{5.20}$$

where C is an absolute constant.

6. Approximation in $L_2(B^d)$

We turn now to the case of approximation on the domain $\Omega = B^d$, $d > 2$, i.e., Ω is the unit Euclidean ball of $\mathbb{R}^d$. Recall that each atom $\phi = \phi(\cdot; \xi, t)$, satisfies $(\xi, t) \in \bar{Z}(\Omega) = S^{d-1} \times [-1,1]$. To each atom, we assign the special weight

$$w(\xi, t) := w^*(\xi, t) := (1-t)^{\frac{1}{2} + \frac{d}{4}}. \tag{6.1}$$

From (4.3), we see that this weight is admissible for Ω. Thus, w is taken to be given by (6.1) throughout this section. One can ask where the particular form of the weight comes from. It arises due to the norm of the atoms in L_2 and the smoothness of the parameterization of the atoms. The effects of these two ingredients will become clear in the details of the proof.

We recall the variation space $\mathcal{V}_w$ introduced and studied in §4. The main result of this paper is the following theorem

Theorem 6.1. *If $f \in \mathcal{V}_w = \mathcal{V}_w(\Omega)$ then*

$$E_n(f) := E_n(f)_{L_2(\Omega)} \le C \|f\|_{\mathcal{V}_w} n^{-\frac{1}{2} - \frac{3}{2d}}, \quad n \ge 1, \tag{6.2}$$

where C depends only on d.

Notice that this theorem gives a stronger result than the previously known results on approximation by shallow neural networks with ReLU activation. Indeed, although the approximation rate $O(n^{-\frac{1}{2} - \frac{3}{2d}})$ is the same as known whenever $f \in \mathcal{V}$, the assumption of membership in $\mathcal{V}_w$ is a strictly weaker assumption than the membership in the traditional variation space $\mathcal{V}$.

The remainder of this section is devoted to proving Theorem 6.1. The proof is similar, in spirit, to the case $d = 2$ which was given in Theorem 5.2, but it is quite a bit more technical. Our first goal is to construct certain linear spaces X_n of dimension at most n, which can be used to effectively approximate general ReLU atoms. The space X_n will be the span of at most n well chosen atoms from $\mathcal{D}(\Omega)$. The choice of the atoms used to define X_n would intuitively be gotten by discretizing the unit Euclidean sphere S^{d-1} with m^{d-1} uniformly spaced vectors and then and to discretize the offsets in $T = [-1,1]$ with m points. Here m is chosen so that $n \approx m^d$. The discretization of T will not be uniform but instead will be done in such a way that atoms whose support is small, i.e., atoms whose associated hyperplane lies near the boundary S^{d-1} of B^d will be very well approximated.

Since there is no natural discretization of S^{d-1}, when $d > 2$, we proceed as follows. Let $Q := Q^d := [-1,1]^d$ and F be a face of Q. Each face F is gotten by setting one of the coordinates, say, coordinate i, equal to either $+1$ or -1. Given one of these faces F,

we shall use dyadic partitions of F into $d-1$ dimensional cubes of side length 2^{-k}. We let $V_k(F)$ be the set of all vertices of this partition. Thus, the cardinality of $V_k(F)$ is $(2^k+1)^{d-1}$. We define V_k to be the union of all of the sets $V_k(F)$ as F runs through the $2d$ faces of Q. This gives a discrete set of points on the boundary of Q with ℓ_∞ spacing 2^{-k}. To obtain our discrete set of points on S^{d-1}, we simply renormalize. Namely,

$$W_k := \left\{ \xi = \frac{\bar{\xi}}{\|\bar{\xi}\|} \,:\, \bar{\xi} \in V_k \right\} \tag{6.3}$$

gives a set of points on the boundary of B^d that are quasi-uniformly spaced in the sense that

$$c_0 2^{-k} \le \operatorname{dist}(\xi_i, W_k \setminus \{\xi_i\}) \le C_0 2^{-k}, \tag{6.4}$$

where the constants[1] depend only on d. After adjusting for redundancy, we see that the cardinality of W_k is $2d(2^k)^{d-1}$. It is important to note that

$$V_k \subset V_{k+1} \quad \text{and} \quad W_k \subset W_{k+1}, \quad k \ge 1. \tag{6.5}$$

We also want to discretize the offsets t. For this, we take

$$T_m := \{-1 < t_1 < \cdots < t_{2m} = 1\}. \tag{6.6}$$

We take the first m of these to be equally spaced in $[-1,0]$, i.e., $t_j := -1 + j/m$, $j = 1, \dots, m$. For the remainder of these points, we take

$$t_{j+m} := \cos \frac{\pi(m-j)}{2m} =: \cos \theta_{j,m}, \quad j = 1, \dots, m. \tag{6.7}$$

Notice that the points in T_m have a finer spacing near one. Concerning this spacing, in going further we will use the fact that for each $m < j < 2m-1$ and $t \in [t_j, t_{j+1}]$ we have

$$\frac{\pi\sqrt{1-t_{j+1}^2}}{2m} \le |t_{j+1} - t_j| \le \frac{\pi\sqrt{1-t_j^2}}{2m} \quad \text{and} \quad \sqrt{1-t_j^2} \le 2\sqrt{1-t_{j+1}^2} \le 2\sqrt{1-t^2}. \tag{6.8}$$

To prove this, we note that for fixed $j = i+m$, $1 \le i < m$, we have

$$|t_{j+1} - t_j| = \frac{\pi}{2m} \sin\zeta = \frac{\pi}{2m}\sqrt{1-\cos^2\zeta}$$

where $\zeta \in [\theta_{i+1,m}, \theta_{i,m}]$. This gives the first inequalities in (6.8). The second inequalities are proved similarly. The inequalities in (6.8) show that for any given $t \in [t_j, t_{j+1}]$, $j \le 2m-2$, we have $|t_{j+1} - t_j| \approx \sqrt{1-t^2}/m$ with absolute constants in this comparison. We shall use this fact repeatedly.

We now want to define the linear space X_n. Consider the set of atoms given by

$$\Phi_{k,m} := \{\phi(\cdot;\xi,t),\ \xi \in W_k,\ t \in T_m\}. \tag{6.9}$$

This is a set of at most $4dm(2^k)^{d-1}$ ReLU atoms. We choose k as the largest integer such that $4d2^k(2^k)^{d-1} \le n$ and then take $m = 2^k$. Then, $\Phi_n := \Phi_{k,m}$ is a set of at most n atoms. We define X_n as the linear space

$$X_n := \operatorname{span}(\Phi_n). \tag{6.10}$$

Then, X_n is a linear space of dimension at most n.

We caution the reader that for the remainder of this paper, the integer n is always taken of the form $n = 4dm^d$, where $m = 2^k$. It is enough to prove our approximation results for these n.

We now proceed to show that any atom $\phi := \phi(\cdot;\xi,t)$ from $\mathcal{D}(\Omega)$ can be well approximated by an element of the linear space X_n. We fix ξ, t and n. The approximation result we prove is given in the following theorem.

Theorem 6.2. *For any $(\xi,t) \in \bar{Z}(\Omega) = S^{d-1} \times [-1,1]$, there is an element $g = g_\phi \in X_n$ such that*

$$\|\phi - g\|_{L_2(\Omega)} \le C w(\phi) n^{-\frac{3}{2d}}, \tag{6.11}$$

with the constant C depending only on d.

The proof of this theorem is a bit technical and given in the next subsection. After proving this theorem we prove Theorem 6.1. In the constructions given below there are two important constants A and L which depend only on d. It will be useful to the reader

[1] In this paper, all constants depend only on d and may change from line to line. We use c for small constants and C for large constants, sometimes with subscripts.

if we explain their role and their definition. To prove Theorem 6.2, we are presented with an atom $\phi(x) = (\xi \cdot x - t)_+$ and need to construct an element $g \in X_n$ that approximates ϕ to the given accuracy. From the definition of X_n, the function g will take the form

$$g = \sum_{j=1}^{n} a_j \phi_j, \tag{6.12}$$

where the ϕ_j are the atoms $\phi_j(x) = (\xi_j \cdot x - \tau_j)_+$ used to define X_n, where $\tau_j = t_{i(j)}$. The function g that we construct to provide the approximation will agree with ϕ except on a certain set of small measure. The only atoms active in the definition of g, i.e., which have nonzero coefficients, will satisfy $\|\xi - \xi_j\| \le \frac{A}{m}$, with $A \ge 2$ a fixed integer constant depending only on d. The size of A is determined by the proof of Lemma 6.6 which is formulated later in this section and then proved in the Appendix. Hence, in going forward, we can consider d arbitrary but fixed and A depending only on d to be fixed as well.

The constant L is an integer which is chosen in the proof of Lemma 6.5. We will only have to consider values of t such that $t \le t_{2m-L}$. This restriction can be applied on t because of the following remark.

Remark 6.3. Let us note and record the following:
(i) If $t \ge t_{2m}$ or even $t \ge t_{m'}$ with $m' = 2m - L$ with L a fixed integer, then for any atom $\phi(\cdot;\xi,t)$ the statement (6.11) holds by simply taking $g = 0$ and using the estimate (4.3) for the norm of the atom.
(ii) If $t \le C < 1$ with C fixed then the weight $w(\xi,t) \ge c$. In this case, the existence of a space X_n spanned by n atoms that provides the estimate (6.11) was given in [40]. While our space X_n is defined differently (we use a different discretization of the offsets t), the proof in this case is simpler and we exclude this case going forward.
(iii) If ξ is one of the discrete vectors from W_k, then the proof of the existence of a g for which (6.11) is quite simple. Indeed, one can take $g = a\phi(\cdot;\xi,t_i) + (1-a)\phi(\cdot;\xi,t_{i+1})$ where t_i is the closest discrete offset to t and a is chosen so that $at_i + (1-a)t_{i+1} = t$.

In the proof of (6.11) we only need to provide a proof in the case that none of the special cases (i-iii) holds.

6.1. The proof of Theorem 6.2

Obviously, we only need to prove the theorem for m sufficiently large, say $m \ge m^*$ where m^* depends on d. The integer m^* will be specified as we go along. Because of Remark 6.3 we only need to prove the theorem in the case $1/2 < t \le t_{m-L}$, where L is a fixed integer depending only on d. Again, we shall specify L as we proceed in the proof. Similarly, we can assume ξ is not in W_k. We fix such an $\xi \in S^{d-1}$ and such a t throughout this subsection.

We define

$$H^+ := \{x : \ \xi \cdot x \ge t\} \quad \text{and} \quad H^- := \{x : \ \xi \cdot x < t\}. \tag{6.13}$$

So ϕ is identically zero on H^- and the linear function $\xi \cdot x - t$ on H^+. For any one of the vectors ξ_i appearing in the set W_k and any given a $t_j \in T_m$, we similarly define

$$H_j^+ := H_j(\xi_i) := \{x \in \Omega : \xi_i \cdot x \ge t_j\}, \quad H_j^- := H_j^-(\xi_i) := \{x \in \Omega : \xi_i \cdot x < t_j\}. \tag{6.14}$$

Given ξ_i, we want to choose a t_j with $j \le 2m-1$ (depending on i) that is close to t and so that H_j^+ is a subset of H^+. This is always possible whenever $\|\xi - \xi_i\| \le A/m$ and $t \le t_{m-L}$ and L is sufficiently large (depending only on A). One such choice for t_j is to take

$$t_i^+ := t^+(\xi_i) := \min\{t_j \in T_m : \ H_j^+ \subset H^+\}. \tag{6.15}$$

If $t_i^+ = t_j$, we let

$$\tilde{t}_i := t_{j+1}. \tag{6.16}$$

Then, we will also have $H_{j+1}^+ \subset H^+$.

We now proceed to proving Theorem 6.2. We begin by recalling the following fact.

Lemma 6.4. *If $\xi, \xi' \in S^{d-1}$ with $\|\xi - \xi'\| = \delta$, then we have*

$$\xi \cdot \xi' = 1 - \delta^2/2. \tag{6.17}$$

Proof. By rotation, we can assume $\xi = e_1 = (1,0,\dots,0)$ and $\xi' = \alpha e_1 + \eta$ where η is orthogonal to e_1 and $\|\eta\|^2 = 1 - \alpha^2$. Therefore,

$$\delta^2 = (1-\alpha)^2 + \|\eta\|^2 = 2 - 2\alpha = 2 - 2\xi \cdot \xi'$$

and so (6.17) follows. □

The last lemma allows us to compare t_i^+ with t.

Lemma 6.5. *Given the integer A, we define*

$$L := (A+1)^2 = A^2 + 2A + 1. \tag{6.18}$$

If m^* is sufficiently large, depending only on d and $m \geq m^*$, then whenever $t \in [1/2, t_{2m-L}]$ and $\|\xi - \xi_i\| \leq \frac{A}{m}$, then t_i^+ and $\tilde{t}_i$ are well defined, and we have

$$t \leq t_i^+ \leq \tilde{t}_i \leq t + \frac{C_1}{m}\sqrt{1-t^2}, \tag{6.19}$$

where C_1 depends only on d.

Proof. Consider first the existence of $t_i^+, \tilde{t}_i$. It is enough to show that if $t \leq t_{2m-L}$, and ξ_i satisfies $\|\xi - \xi_i\| \leq \frac{A}{m}$, then there is a $j \leq 2m-2$ such that $H_j^+ \subset H^+$. Suppose that j is such that $t_j \geq t$ but H_j^+ is not contained in H^+. Then, there is an $x = t_j\xi_i + \eta$ with η orthogonal to ξ_i and $\|\eta\| \leq \sqrt{1-t_j^2}$ and

$$x \cdot \xi = t_j \xi \cdot \xi_i + \eta \cdot (\xi - \xi_i) < t.$$

From Lemma 6.4, we know that $\xi \cdot \xi_j \geq 1 - \frac{A^2}{m^2}$ and so we must have

$$\left(1 - \frac{A^2}{m^2}\right) t_j \leq t + \|\xi - \xi_i\|\sqrt{1-t_j^2} \leq t + \frac{A}{m}\sqrt{1-t_j^2}. \tag{6.20}$$

That is, we must have

$$t_j \leq t + \frac{A}{m}\sqrt{1-t_j^2} + \frac{A^2}{m^2} \leq t_{2m-L} + \frac{A}{m}\sqrt{1-t_j^2} + \frac{A^2}{m^2}. \tag{6.21}$$

If we write $j = 2m-k$, and use the definition of the t_j (see (6.7)) we can rewrite (6.21) as

$$\cos\frac{\pi k}{2m} - \cos\frac{\pi L}{2m} \leq \frac{A}{m}\sin\frac{\pi k}{2m} + \frac{A^2}{m^2} \leq \frac{\pi A k}{2m^2} + \frac{A^2}{m^2} \leq \frac{A^2 + 2Ak}{m^2}. \tag{6.22}$$

The left side of (6.22) is larger than $\frac{k(L-k)}{m^2}$ and so we see with the above definition of L, (6.22) is violated when $k = 2$. This proves that t_i^+ and $\tilde{t}_i$ are well defined.

We turn now to proving (6.19). First note that if there is $j \in \{m+1, \dots, 2m\}$ such that $H_j^+ \subset H^+$, then we must have $t_j\xi_i \cdot \xi \geq t$ which gives the left inequality in (6.19). To prove the right inequality in (6.19), we let $t_i^+ = t_j$, $\tilde{t}_i = t_{j+1}$. It follows from the minimality in the definition of t_i^+ that we must have H_{j-1}^+ is not contained in H^+. Thus, the inequality (6.20) holds with j replaced by $j-1$. This gives

$$\left(1 - \frac{A^2}{m^2}\right) t_{j-1} \leq t + \|\xi - \xi_i\|\sqrt{1-t_{j-1}^2} \leq t + \frac{2A}{m}\sqrt{1-t_i^{+2}}, \tag{6.23}$$

where the last inequality uses (6.8). From (6.8), we also have

$$t_j \leq t_{j-1} + (t_j - t_{j-1}) \leq t_{j-1} + \frac{\pi}{m}\sqrt{1-t_j^2}.$$

If we multiply both sides of this last inequality by $(1 - \frac{A^2}{m^2})$ and use (6.23) we obtain

$$\left(1 - \frac{A^2}{m^2}\right) t_i^+ \leq t + \frac{2(A+2)}{m}\sqrt{1-t^2},$$

where we used that $t_i^+ \geq t$. We also have $\tilde{t}_i \leq t_i^+ + \frac{\pi}{2m}\sqrt{1-t^2}$ because of (6.8). When these facts are used in the last inequality we obtain the right inequality in (6.19). □

Now consider any $\xi \in S^{d-1}$ and define

$$W_k(\xi) := \left\{\xi_i \in W_k : \ \|\xi - \xi_i\| \leq \frac{A}{m}\right\} \quad \text{and} \quad t^+ := t^+(\xi) := \max_{\xi_i \in W_k(\xi)} t_i^+. \tag{6.24}$$

We can write $t_i^+ = t_j$ for some $j \leq 2m-1$ and define $\tilde{t} := \tilde{t}(\xi) := t_{j+1}$. From the previous lemma, we know that

$$t \leq t^+ \leq \tilde{t} \leq t + \frac{C_1}{m}\sqrt{1-t^2}, \tag{6.25}$$

where C_1 depends only on d.

For the construction of g, we use the following lemma.

Lemma 6.6. *There is a constant m^* depending only on d such that the following holds. Given any $m \geq m^*$ and any $\xi \in S^{d-1}$ and any $1/2 \leq t \leq t_{2m-L}$, there exists (a_i^+), $(\tilde{a}_i)$ such that*
(i) $\xi = \sum_{\xi_i \in W_k(\xi)} a_i^+ \xi_i + \sum_{\xi_i \in W_k(\xi)} \tilde{a}_i \xi_i$,
(ii) $t^+ \sum_{\xi_i \in W_k(\xi)} a_i^+ + \tilde{t} \sum_{\xi_i \in W_k(\xi)} \tilde{a}_i = t$,
(iii) $\sum_{\xi_i \in W_k(\xi)} |a_i^+| + \sum_{\xi_i \in W_k(\xi)} |\tilde{a}_i| \leq C_1$, *where C_1 depends only on d.*

The proof of this lemma is technical and so we place it in the Appendix so as not to interrupt the flow of the proof of Theorem 6.2. We define the following function g which will be used to approximate $\phi = \phi(\cdot; \xi, t)$ in the case $1/2 \leq t \leq t_{2m-L}$

$$g(x) := \sum_{\xi_i \in W_k(\xi)} [a_i^+ (\xi_i \cdot x - t^+)_+ + \tilde{a}_i (\xi_i \cdot x - \tilde{t})_+] \tag{6.26}$$

where the coefficients come from Lemma 6.6. The functions appearing in the representation of g are all in X_n and therefore g is also in X_n. From Lemma 6.6, we obtain

$$\phi(x) - g(x) = \left[\sum_{\xi_i \in W_k(\xi)} a_i^+ (\xi_i \cdot x - t^+) + \tilde{a}_i (\xi_i \cdot x - \tilde{t}) \right]_+ - \sum_{\xi_i \in W_k(\xi)} [a_i^+ (\xi_i \cdot x - t^+)_+ + \tilde{a}_i (\xi_i \cdot x - \tilde{t})_+] \tag{6.27}$$

Before bounding the error in approximating ϕ by g, let us make some remarks to motivate the idea of how to estimate this error. Notice that if $x \in \Omega$ is such that $\xi_i \cdot x \geq \tilde{t}$ for all $\xi_i \in W_k(\xi)$, then x is also in H^+ and so $g(x) = \phi(x)$. Similarly, if $x \in H^-$ then $\phi(x) = g(x) = 0$. This means that the only points $x \in \Omega$ where $\phi(x) \neq g(x)$ must be in one of the sets

$$\tilde{\Omega}_i := \{x \in \Omega : \xi \cdot x > t,\ \xi_i \cdot x \leq \tilde{t}\}. \tag{6.28}$$

We will now proceed to bound the measure of each of these sets and also bound the error $|\phi(x) - g(x)|$ on each of these sets.

Lemma 6.7. *There are constants C and m^* depending only on d such that the following holds for $m \geq m^*$ and any $1/2 \leq t \leq t_{2m-L}$:*
(i) *If $x \in \tilde{\Omega} := \bigcup_{\xi_i \in W_k(\xi)} \tilde{\Omega}_i$, then*

$$|\phi(x) - g(x)| \leq C \sqrt{1 - t^2}/m. \tag{6.29}$$

(ii) *The measure of $\tilde{\Omega}$ is bounded by*

$$|\tilde{\Omega}|_d \leq C(1 - t^2)^{d/2}/m. \tag{6.30}$$

Proof. From Lemma 6.5 we have that

$$t \leq t^+ \leq \tilde{t} \leq t + C_1 \frac{\sqrt{1 - t^2}}{m}, \quad i = 1, \dots, M \tag{6.31}$$

where C_1 depends only on d. Now, for any fixed i consider any $x \in \tilde{\Omega}_i$. Our goal is to estimate the distance from x to the hyperplane $H := \{z : \ z \cdot \xi = t\}$. From the definition of $W_k(\xi)$, we have $\|\xi - \xi_i\| \leq A/m$. Since $\alpha := x \cdot \xi > t$, we can write

$$x = \alpha \xi + \eta, \tag{6.32}$$

where η is orthogonal to ξ and

$$\|\eta\| \leq \sqrt{1 - \alpha^2} \leq \sqrt{1 - t^2}. \tag{6.33}$$

We want to show that α cannot be too large. Since $x \in \tilde{\Omega}_i$, we know that

$$(x \cdot \xi_i) = \alpha \xi \cdot \xi_i + \eta \cdot \xi_i \leq \tilde{t}, \tag{6.34}$$

and

$$|\eta \cdot \xi_i| = |\eta \cdot (\xi_i - \xi)| \leq A \frac{\sqrt{1 - t^2}}{m}. \tag{6.35}$$

Using the inequality $\xi \cdot \xi_i \geq 1 - A^2/m^2$ (see Lemma 6.4) and (6.35) back in (6.34), we obtain

$$(1 - A^2 m^{-2})\alpha \leq \tilde{t} + A \frac{\sqrt{1 - t^2}}{m} \leq t + C_1 \frac{\sqrt{1 - t^2}}{m} + A \frac{\sqrt{1 - t^2}}{m}, \tag{6.36}$$

where the last inequality used (6.31). Going further, we note that since by assumption $t \leq t_{2m-L}$, we have

$$(1 - A^2 m^{-2})^{-1} \leq 1 + 2Am^{-2} \leq 1 + 2A \frac{\sqrt{1 - t^2}}{m}.$$

Using this estimate back in (6.36), we arrive at

$$\alpha \le t + C_2 \frac{\sqrt{1-t^2}}{m}, \tag{6.37}$$

with C_2 depending only on d. It is important to note that this bound is independent of i. Therefore any point x in $\tilde{\Omega}$ can be written as $x = \alpha\xi + \eta$ where α satisfies (6.37) and $\|\eta\| \le \sqrt{1-t^2}$. The measure of the set of such η is $\le C_3[\sqrt{1-t^2}]^{d-1}$. Hence, we have proven (ii).

The inequality (6.37) also shows that

$$\phi(x) \le C_2 \frac{\sqrt{1-t^2}}{m}, \quad x \in \tilde{\Omega}. \tag{6.38}$$

Since the functions appearing in the sum for g are all smaller than ϕ from (iii) of Lemma 6.6 we conclude that

$$|g(x)| \le C_3 \frac{\sqrt{1-t^2}}{m}, \quad x \in \tilde{\Omega}. \tag{6.39}$$

This proves (i) and concludes the proof of the lemma. □

Proof of Theorem 6.2. According to Remark 6.3, we only need to consider the case $\phi = \phi(\cdot;\xi,t)$ when $1/2 \le t < t_{2m-L}$. We return to our representation (6.27). As already mentioned $\phi(x) = g(x)$ outside the set $\tilde{\Omega}$. From Lemma 6.7 it follows that

$$\|\phi - g\|_{L_2(\Omega)} \le C|\tilde{\Omega}|_d^{1/2} \frac{\sqrt{1-t^2}}{m} \le Cw(\phi)m^{-3/2}. \tag{6.40}$$

Since $m^d \ge c_d n$, we have completed the proof of Theorem 6.2. □

6.2. *Proof of Theorem 6.1*

We can now prove Theorem 6.1 in the same way we proved the case $d = 2$. Let f be any fixed function from $\mathcal{V}_w(\Omega)$. According to the definition of $\mathcal{V}_w(\Omega)$, for N sufficiently large, there is an $S \in \Sigma_N$ with $S = \sum_{j=1}^N a_j\phi_j$ such that

$$\|f - S\|_{L_2(\Omega)} \le Mn^{-\frac{1}{2}-\frac{3}{2d}} \quad \text{and} \quad \sum_{j=1}^N w(\phi_j)|a_j| \le \|f\|_{\mathcal{V}_w} =: M. \tag{6.41}$$

For each j, let $g_j \in X_n$ approximate the function ϕ_j appearing in the representation of S according to Theorem 6.2. That is, we have

$$\|\phi_j - g_j\|_{L_2(\Omega)} \le Cw(\phi_j)n^{-\frac{3}{2d}}, \tag{6.42}$$

with C depending only on d. The function $g := \sum_{j=1}^N a_j g_j$ is in X_n and hence in Σ_n. We write

$$f = f - S + h + g, \quad h := S - g = \sum_{j=1}^N a_j[\phi_j - g_j]. \tag{6.43}$$

Therefore,

$$E_{3n}(f) \le Mn^{-\frac{1}{2}-\frac{3}{2d}} + E_{2n}(h). \tag{6.44}$$

To bound $E_{2n}(h)$, we consider the dictionary $\mathcal{D}' = \{\psi_j\}_{j=1}^N$ with $\psi_j := w(\phi_j)^{-1}(\phi_j - g_j)$. According to Theorem 6.2 each ψ_j has $L_2(\Omega)$ norm at most $Cn^{-\frac{3}{3d}}$ and $h = \sum_{j=1}^N c_j\psi_j$ with $\sum_{j=1}^N |c_j| \le M$. It follows from Maurey's Theorem (see (4.15)) that h can be approximated by a sum T of n terms from the dictionary $\mathcal{D}'$ with error

$$\|h - T\|_{L_2(\Omega)} \le CMn^{-\frac{3}{2d}}n^{-1/2} = CMn^{\frac{1}{2}-\frac{3}{2d}}. \tag{6.45}$$

The function T is a sum of at most $2n$ terms from the original dictionary $\mathcal{D}$. Hence,

$$E_{2n}(h) \le CMn^{-\frac{1}{2}-\frac{3}{2d}}. \tag{6.46}$$

If we place this inequality back into (6.44), we obtain

$$E_{3n}(f) \le CMn^{-\frac{1}{2}-\frac{3}{2d}}, \tag{6.47}$$

and the theorem follows. □

Remark 6.8. If we consider $\Omega = Q^d$ in place of B^d then for the weight $w(\phi(\cdot;\xi,t)) = |L_\phi|^{1/2}$ where $|L_\phi|$ is the $(d-1)$-dimensional measure of the intersection $Q^d \cap L_\phi$, we obtain

$$E_n(f)_{L_2(Q^d)} \le C n^{-\frac{1}{2}-\frac{3}{2d}} \|f\|_{\mathcal{V}_w}, \quad f \in \mathcal{V}_w. \tag{6.48}$$

We do not give the proof which follows along the lines of the case $d = 2$ given in §5.2.

7. Concluding remarks

The main theme of this paper was to introduce, for a bounded domain $\Omega \subset \mathbb{R}^d$, new model classes $\mathcal{V}_w := \mathcal{V}_w(\mathcal{D}(\Omega))$, called *weighted variation spaces*, and to prove bounds on how well functions in these classes can be approximated by a linear combination of n terms of the ReLU dictionary $\mathcal{D}(\Omega)$. That is, we provided bounds on the error $E_n(f)_{L_2(\Omega)}$ in approximating $f \in \mathcal{V}_w(\Omega)$ in the $L_2(\Omega)$ norm by the elements of $\Sigma_n := \Sigma_n(\mathcal{D}(\Omega))$ where Σ_n is the nonlinear manifold of functions g that are a linear combination of n elements of the ReLU dictionary. We showed that for certain choices of the weight w (dependent on Ω) the functions in these new model classes have the same approximation rate as those in the classical variation spaces $\mathcal{V}(\Omega)$. Since $\mathcal{V}_w$ is strictly larger than the classical variation spaces $\mathcal{V} := \mathcal{V}(\Omega)$, this gives stronger results on n-term ReLU approximation than those in the literature. Thus, these new model classes $\mathcal{V}_w$ are important in trying to understand which functions are well approximated by Σ_n.

A natural follow-up question would be to then consider the problem of learning from data generated from the samples of a function from $\mathcal{V}_w$, in both the noiseless and noisy settings. In the literature, the former is referred to as *optimal recovery* and the latter is referred to as *minimax estimation*. For the classical variation spaces, the minimax estimation rates have been determined [33]. On the other hand, the optimal recovery rates are currently unknown. Once the data sites are fixed, it is well-known that the procedure for optimal recovery takes the form of solving a regularized least-squares problem over the model class [4]. Theorem 7.1 below motivates a numerical method (posed as a neural network training problem) to investigate the problem of optimal recovery (as well as for minimax estimation for $\mathcal{V}_w$).

Assume that $d \ge 2$ and $\Omega = B^d$ in the sequel. Then, $\bar{Z}(\Omega) = S^{d-1} \times [-1,1]$. Let w be any admissible weight function in the sense of (4.9) and let $\mathcal{D}_w$ be the weighted dictionary defined in (4.10). We use the results and notation of §4 in going forward. In particular, the functions in $\mathcal{V}_w := \mathcal{V}_w(\Omega)$ all take the form (see (4.11))

$$\bar{f}_\mu := \int_{\bar{Z}(\Omega)} \bar{\phi}(\cdot;\xi,t)\, d\mu \quad \text{and} \quad \|\bar{f}_\mu\|_{\mathcal{V}_w(\Omega)} = \|\mu\|_{\mathcal{M}}. \tag{7.1}$$

Consider the following data-fitting problem. Suppose that x_i, $i = 1,\dots,m$, are points from the interior of Ω and y_i, $i = 1,\dots,m$, are real numbers. The data-fitting problem

$$\inf_{f \in \mathcal{V}_w(\Omega)} \sum_{i=1}^m |y_i - f(x_i)|^2 + \lambda \|f\|_{\mathcal{V}_w}, \tag{7.2}$$

with $\lambda > 0$ is equivalent to the data-fitting problem

$$\inf_{\mu \in \mathcal{M}(\bar{Z}(\Omega))} \sum_{i=1}^m |y_i - f_\mu(x_i)|^2 + \lambda \|\mu\|_{\mathcal{M}}, \tag{7.3}$$

in the sense that their infimal values are the same and if $\mu^\star$ is a minimizer of (7.3), then f_{μ^*} is a minimizer of (7.2). Note that the minimization problem (7.2) does not depend on the ambient space $L_2(\Omega)$ in which we measure error of performance for the approximation problem. An important property of the weighted variation spaces is that solutions to data-fitting problems over this model class admit finite-parameter representations as neural networks. This is summarized in the next theorem.

Theorem 7.1. *Suppose that w is an admissible weight function and the $\{x_i\}_{i=1}^m$ lie in the interior of Ω. Then, there exists a solution f^* to (7.2) that takes the form of a shallow ReLU network*

$$f^\star(x) = \sum_{j=1}^n a_j \phi(x;\xi_j,t_j) = \sum_{j=1}^n a_j (\xi_j \cdot x - t_j)_+, \tag{7.4}$$

where the number of atoms satisfies $n \le m$, $\{a_j\}_{j=1}^n \subset \mathbb{R} \setminus \{0\}$, and $\{(\xi_j,t_j)\}_{j=1}^n \subset \bar{Z}(\Omega)$ are data-dependent and not known a priori. *Furthermore, the regularization cost is $\|f^\star\|_{\mathcal{V}_w} = \sum_{j=1}^n w(\xi_j,t_j)|a_j|$.*

Proof. Let $C(\bar{Z}(\Omega))$ denote the space of real valued functions on $\bar{Z}(\Omega)$. This is a Banach space when equipped with the L_∞-norm. By the Riesz–Markov–Kakutani representation theorem [16, Chapter 7], the dual of $C(\bar{Z}(\Omega))$ can be identified with the space of signed Radon measures $\mathcal{M} := \mathcal{M}(\bar{Z}(\Omega))$. It is well-known that the extreme points of the unit ball

$$\{\mu \in \mathcal{M} : \|\mu\|_{\mathcal{M}} \le 1\} \tag{7.5}$$

are the Dirac measures $\pm\delta_{(\xi,t)}$, $(\xi,t) \in \bar{Z}(\Omega)$ (see, e.g., [9, Proposition 4.1]).

Next, for $i = 1, \dots, m$, we introduce the functions

$$h_i(\xi, t) := \tilde{\phi}(x_i; \xi, t) = \frac{\phi(x_i; \xi, t)}{w(\xi, t)}, \quad (\xi, t) \in \bar{Z}(\Omega). \tag{7.6}$$

We can rewrite (7.3) as

$$\inf_{\mu \in \mathcal{M}} \sum_{i=1}^{m} |y_i - \langle \mu, h_i \rangle|^2 + \lambda \|\mu\|_{\mathcal{M}}, \tag{7.7}$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $C(\bar{Z}(\Omega))$ and $\mathcal{M}$. Since the functions h_i, $i = 1, \dots, m$, are in $C(\bar{Z}(\Omega))$, the mappings $\mu \mapsto \langle \mu, h_i \rangle$ are weak* continuous [36, Theorem IV.20, p. 114]. This shows that the hypothesis of the abstract representer theorem [8, 9,42] are satisfied. That theorem shows that there exists a solution to (7.7) that takes the form of a linear combination of the extreme points of the unit regularization ball. Thus, there exists a solution that takes the form

$$\mu^\star = \sum_{j=1}^{n} c_j \delta_{(\xi_j, t_j)}, \tag{7.8}$$

where the number of atoms satisfies $n \leq m$, $\{c_j\}_{j=1}^n \subset \mathbb{R} \setminus \{0\}$, and $\{(\xi_j, t_j)\}_{j=1}^n \subset \bar{Z}(\Omega)$ are distinct, data dependent, and not known *a priori*. Clearly $\|\mu^\star\|_{\mathcal{M}} = \sum_{j=1}^n |c_j|$.

From the equivalence between (7.2) and (7.7), we see that the function

$$f_{\mu^\star} = \int_{\bar{Z}(\Omega)} \tilde{\phi}(\cdot; \xi, t) d\mu^\star(\xi, t) = \sum_{j=1}^{n} \frac{c_j}{w(\xi_j, t_j)} \phi(x; \xi_j, t_j) \tag{7.9}$$

is a minimizer of (7.2). The theorem follows by the substitution $a_j := c_j / w(\xi_j, t_j)$. □

We have not indicated the fact that the solution (7.9) to the data-fitting problem depends on λ. If we let λ tend to zero then the solutions converge to a minimum-norm interpolant $f^\#$ of the data

$$f^\# \in \operatorname{argmin}\{\|f\|_{\mathcal{V}_w} : f(x_i) = y_i, \ i = 1, \dots, m\}. \tag{7.10}$$

In which case, there always exists an $f^\#$ that has a representation

$$f^\# = \sum_{j=1}^{n} a_j^\# \phi(x; \xi_j^\#, t_j^\#), \tag{7.11}$$

with $n \leq m$.

The theorem statement also holds when the first term in the objective in (7.2) is replaced by any loss function $\mathcal{L}(\cdot, \cdot)$ which is lower semi-continuous (see [31, Proof of Theorem 3.2]). In neural network parlance, the ξ_j are referred to as the *input weights*, the a_j are referred to as the *output weights* and the t_j are referred to as the *biases*. Observe that the norm of a single neuron $\phi(\cdot; \xi, t)$, where $\xi \in \mathbb{R}^d$ and $t \in \mathbb{R}$, takes the form

$$\|\phi(\cdot; \xi, t)\|_{\mathcal{V}_w} = \|\xi\| \, w\left(\frac{\xi}{\|\xi\|}, \frac{t}{\|\xi\|}\right), \tag{7.12}$$

where we took advantage of the fact that the ReLU is positively homogeneous of degree 1. In this form, the input weights are not restricted to be unit norm. Theorem 7.1 then implies that a solution to the variational problem in (7.2) can be found by training a sufficiently wide (fixed width $n \geq m$) neural network to a global minimizer with an appropriate regularization term. This follows, in particular, by finding a solution to the neural network training problem

$$\min_{\theta} \sum_{i=1}^{m} |y_i - f_\theta(x_i)|^2 + \lambda \sum_{j=1}^{n} |a_j| \|\xi_j\| w\left(\frac{\xi_j}{\|\xi_j\|}, \frac{t_j}{\|\xi_j\|}\right), \tag{7.13}$$

where

$$f_\theta(x) = \sum_{j=1}^{n} a_j \phi(x; \xi_j, t_j) = \sum_{j=1}^{n} a_j (\xi_j \cdot x - t_j)_+ \tag{7.14}$$

is a shallow ReLU neural network and $\theta = (a_j, \xi_j, t_j)_{j=1}^n$ denotes the neural network parameters and $n \geq m$. When λ is chosen to be sufficiently small, the estimator $f_{\tilde{\theta}}$ achieves the optimal recovery rate for the model class $\mathcal{V}_w$, where $\tilde{\theta}$ is any minimizer of (7.13) [4].

When w is the weight specified in (6.1) (which satisfies the hypotheses of Theorem 7.1), the resulting regularizer takes the form

$$\sum_{j=1}^{n} |a_j| \|\xi_j\| \left(1 - \frac{t_j}{\|\xi_j\|}\right)^{\frac{1}{2} + \frac{d}{4}} \tag{7.15}$$

This is a new regularizer for training neural networks which directly penalizes the biases. If we assume the data sites $\{x_i\}_{i=1}^m$ are drawn i.i.d. uniformly on B^d, then this penalization reflects the volume of the subset of B^d where the neuron is "active" (nonzero output). This suggests a new, data-adaptive regularization scheme in which the penalty on a given neuron is proportional to the number of data in its support. This regularizer should be contrasted with the unweighted case in which the regularizer takes the form

$$\sum_{j=1}^n |a_j| \|\xi_j\|, \tag{7.16}$$

which is sometimes referred to as the *path-norm* [28] of the neural network. Remarkably, path-norm regularization is equivalent to the common procedure of training a neural network with *weight decay* [21] which corresponds to a regularizer of the form

$$\frac{1}{2}\sum_{j=1}^n |a_j|^2 + \|\xi_j\|^2. \tag{7.17}$$

We refer the reader to [32] for more details about this equivalence. The new regularizer in (7.15) requires further study in both theory and practice.

7.1. Open problems

The results presented in this paper open the door to several new research directions.

(i) We have shown that the classical variation space $\mathcal{V}(\Omega)$ is not the approximation space $\mathcal{A}^\alpha = \mathcal{A}^\alpha(L_2(\Omega))$, $\alpha = \frac{1}{2} + \frac{3}{2d}$, since the (strictly larger) weighted variation space $\mathcal{V}_w(\Omega)$ admits the same n-term approximation rate with shallow ReLU networks. Thus, the results of this paper bring us one step closer to characterizing the approximation space $\mathcal{A}^\alpha$, $\alpha = \frac{1}{2} + \frac{3}{2d}$. Future work will be devoted to finding a characterization of this approximation space.

(ii) The results of this paper only consider L_2-approximation. We conjecture that the same rates hold for weighted variation spaces for all L_p, $1 \le p \le \infty$, where now the admissibility condition on the weights will depend on p. That is to say, for each $1 \le p \le \infty$, there exists a weight function w_p^* such that the optimal rate $n^{-\frac{1}{2}-\frac{3}{2d}}$ is achieved.

(iii) The determination of the optimal recovery rates and minimax estimation rates for $\mathcal{V}_w$ is a natural follow-up research direction. Theorem 7.1 and (7.13) provide a numerical method (posed as a neural network training problem) whose solutions are known to achieve the optimal recovery rate. A characterization of this rate is has not been determined, even in the unweighted scenario.

(iv) The weighted variation spaces motivates a new form of data-adaptive regularization for neural networks. Theoretical and experimental comparisons of this new form of regularization compared with more conventional regularization techniques is a direction of future work. Furthermore, extensions of this regularizer to deep neural networks is also a direction of future work.

Appendix A

In this appendix, we prove Lemma 6.6. We let ξ be arbitrary but fixed throughout this section. We begin by recalling some well known results on the representation of points x in a cube $R \subset \mathbb{R}^d$ in terms of the vertices of R. Given any cube $R \subset \mathbb{R}^d$, we let $V(R)$ denote its set of vertices. Let us first consider the case $R = U$ where $U := U^d := [0,1]^d$. We denote the vertices in $V(U)$ by e. So e is a vector with d components $e = (e_1, \dots, e_d)$ with each $e_j \in \{0,1\}$. There are 2^d such e.

Let

$$\ell_0(s) := (1-s) \quad \ell_1(s) = s, \quad s \in \mathbb{R}. \tag{A.1}$$

For each $e \in V$, we define

$$\ell_e(x) := \prod_{j=1}^d \ell_{e_j}(x_j), \quad x = (x_1, \dots, x_d) \in U. \tag{A.2}$$

Then, $\ell_e(e') = 0$, $e' \neq e$ and $\ell_e(e) = 1$. Any $x \in U$ is represented as

$$x = \sum_{e \in V} \ell_e(x) e. \tag{A.3}$$

This is a convex representation in that the coefficients $\ell_e(x) \in [0,1]$, $e \in V(U)$, and they sum to one.

Now consider an arbitrary cube $R \subset \mathbb{R}^d$. We can write $R = v + \alpha[0,1]^d = v + \alpha U$ with $\alpha > 0$. This cube has vertices $v + \alpha e$, $e \in V$. Any point $x = v + \alpha y$, $y \in U$, from this cube, has the representation

$$x = v + \alpha \sum_{e \in V} \ell_e(y) e = \sum_{e \in V} \ell_e(y)[v + \alpha e], \tag{A.4}$$

because $\sum_{e \in V} \ell_e(y) = 1$. Again this is the representation of x as a convex combination of the vertices $V(R)$ of R. Let us note that here we are taking v as the smallest vertex of R. We can derive a similar decomposition by starting with any other vertex of R.

We use the above to find a variety of representations of any x on the boundary of the cube $Q^d := [-1,1]^d$. Later, we shall apply these representations to $x = \bar{\xi}$ and subsequently to $\xi \in S^{d-1}$. Let x be in the face F of Q^d. We assume that the $d-1$ dimensional face F of Q^d corresponds to $x_1 = 1$. We will derive representations for points $x \in F$. Similar representations hold for any of the other faces of Q^d.

Any $x \in F$ takes the form $x = (1, \tilde{x})$ with $\tilde{x} \in [-1,1]^{d-1}$. Suppose now that R is any $d-1$ dimensional cube on F, i.e., R consist of points $(1, \tilde{x})$ where $\tilde{x}$ is in a $d-1$ dimensional cube $\tilde{R}$. From the above, we can write $\tilde{x} = \tilde{v} + \alpha y$, $y \in U^{d-1}$, where $\tilde{v}$ is the smallest vertex of R. Therefore, we have

$$\tilde{x} = \tilde{v} + \alpha y = \tilde{v} + \alpha \sum_{e \in V(U^{d-1})} \ell_e(y) e = \sum_{e \in V(U^{d-1})} \ell_e(y)[\tilde{v} + \alpha e] = \sum_{v \in V(\tilde{R})} \gamma_v v, \tag{A.5}$$

where the v are the vertices of $\tilde{R}$ and

$$\gamma_v = \ell_e(y), \quad \text{when } v = \tilde{v} + \alpha e. \tag{A.6}$$

This is a representation of $\tilde{x}$ as a convex combination of the vertices $V(\tilde{R})$.

We turn now to representations of $\xi \in S^{d-1}$. We write $\xi = \frac{\bar{\xi}}{\|\bar{\xi}\|}$ where $\bar{\xi}$ lies on the boundary of $Q^d = [-1,1]^d$. We assume $\bar{\xi}$ lies on the face F corresponding to first coordinate equal to one. All other cases are handled similarly. We write $\bar{\xi} = (1, \tilde{x})$ with $\tilde{x} \in [-1,1]^{d-1}$ and use the representations of $\tilde{x}$ given above. Recall the discrete set of points F_k, with $m = 2^k$. If $k' < k$ then $F_{k'} \subset F_k$. We fix such a k' to be chosen in a moment.

We let $A \geq 1$ be an integer whose value will be chosen below. We place ourselves in the following situation where $\tilde{x} \in \tilde{v} + \delta U^{d-1} =: \tilde{R} \subset \tilde{R}' := \tilde{v} + \delta' U^{d-1}$ where $\tilde{v} \in F_k$, $\delta = 2^{-k} = 1/m$ and $\delta' = 2^{-k'} = A\delta$ with $A = 2^{k-k'}$. The assumption that $\tilde{x} \in \tilde{R}$ for which there is such a $\tilde{R}$ and $\tilde{R}'$ is a restriction on the position of $\tilde{x}$ in $[-1,1]^{d-1}$. When this is not the case, the argument below needs to be adjusted by changing the choice of the initial vertex and the direction for the representation. Since the adjustment is purely notational, we leave it to the reader.

We will give two representations of $\tilde{x}$, respectively $\bar{\xi}$; the one in terms of the vertices of $\tilde{R}$ and the second in terms of the vertices of $\tilde{R}'$. For the first representation, we use (A.5) with $\alpha = \delta$ to write

$$\bar{\xi} = \sum_{v \in V(\tilde{R})} \gamma_v (1, v) = \sum_{v \in V(\tilde{R})} \gamma_v \sqrt{(1 + \|v\|^2)} \xi_v, \quad \xi_v = \frac{(1, v)}{\sqrt{1 + \|v\|^2}}, \tag{A.7}$$

with the coefficients γ_v given by (A.6). Notice that the ξ_v are all in W_k. This gives the representation

$$\xi = \sum_{v \in V(\tilde{R})} a_v \xi_v, \quad a_v := \frac{\gamma_v \sqrt{(1 + \|v\|^2)}}{\|\bar{\xi}\|}. \tag{A.8}$$

We obtain a second representation as follows. We again write $\tilde{x} = \tilde{v} + \delta y$ with $y \in U^{d-1}$. Then,

$$\tilde{x} = \tilde{v} + \delta \sum_{e \in V(U^{d-1})} \ell_e(y) e = \tilde{v} + \sum_{e \in V(U^{d-1})} \frac{\ell_e(y)}{A} A\delta e = \left(1 - \frac{1}{A}\right) \tilde{v} + \sum_{e \in V(U^{d-1})} \frac{\ell_e(y)}{A} [\tilde{v} + A\delta e]. \tag{A.9}$$

This gives the representation

$$\tilde{x} = \sum_{v \in V(\tilde{R}')} \gamma'_v v, \tag{A.10}$$

where

$$\gamma'_v = \frac{\ell_e(y)}{A}, \quad \text{when } v = \tilde{v} + A\delta e \text{ with } e \neq 0 \text{ and } \gamma'_0 = 1 - \frac{1}{A} + \frac{\ell_0(y)}{A}. \tag{A.11}$$

Notice that this representation of $\tilde{x}$ is again a convex combination of the vertices of $\tilde{R}'$. It follows that

$$\xi = \sum_{v \in V(\tilde{R}')} a'_v \xi'_v, \quad a'_v := \frac{\gamma'_v \sqrt{(1 + \|v\|^2)}}{\|\bar{\xi}\|}. \quad \xi'_v = \frac{(1, v)}{\sqrt{1 + \|v\|^2}} \tag{A.12}$$

We now want to estimate the sums

$$S := \sum_{v \in V(\tilde{R})} a_v, \quad S' = \sum_{v \in V(\tilde{R}')} a'_v, \tag{A.13}$$

Lemma A.1. *There is an $m^* = m^*(d)$, depending only on d, such that whenever $m \geq m^*$ and A is sufficiently large (depending only on d), the following hold. Whenever ξ is not a vertex in W_k, i.e., $\bar{\xi} = (1, \tilde{x})$ where $\tilde{x} = \tilde{v} + y$ where $y \neq 0$, we have*

$$S = 1 + \epsilon \quad \text{and} \quad S' = 1 + \epsilon', \quad \text{where} \quad 0 < 2|\epsilon| < |\epsilon'| \leq \sigma\delta^2, \tag{A.14}$$

and

$$\sigma := \sigma(y) = \sum_{e\neq 0} \ell_e(y) > 0, \tag{A.15}$$

where the strict inequality holds because $y \neq 0$.

Proof. Let $B^2 := 1 + \|\tilde{v}\|^2$ and recall that $\delta := 1/m$. Observe that when $v = \tilde{v} + a\delta e$, $e \in V(U^{d-1})$, we have

$$1 + \|v\|^2 = B^2 + 2a\delta\langle \tilde{v}, e\rangle + a^2\delta^2\|e\|^2 = B^2 + s_v(a), \tag{A.16}$$

where

$$s_v(a) := 2a\delta\langle \tilde{v}, e\rangle + a^2\delta^2\|e\|^2. \tag{A.17}$$

We are interested in the cases, $a = 1, A$. Notice that $s_v(a) = 0$ when $e = 0$, i.e., $v = \tilde{v}$, and also $|s_v(a)| \leq 1/2$ for these two values of a provided m^* is large enough. These facts will be used without further mention in what follows.

We will use the Taylor expansion of the function $F(s) := \sqrt{B^2 + s}$. We have

$$F(s) = B + \frac{1}{2}B^{-1}s - \frac{1}{4}B^{-3}s^2 + O(s^3), \quad |s| < 1. \tag{A.18}$$

This gives that

$$\sqrt{1 + \|v\|^2} = F(s_v(a)) = B + \frac{1}{2}B^{-1}s_v(a) - \frac{1}{4}B^{-3}s_v(a)^2 + O(s_v(a)^3) \tag{A.19}$$

From the above observations, we can write

$$\begin{aligned}
\|\bar{\xi}\|S' &= \sum_{v\in V(\tilde{R}')} \gamma'_v F(s_v(A)) \\
&= (1 - 1/A)B + \sum_{e\in V(U^{d-1})} \frac{\ell_e(y)}{A}\left(B + \frac{1}{2}B^{-1}s_v(A) - \frac{1}{4}B^{-3}s_v(A)^2 + O(s_v(A)^3)\right) \\
&= B + \frac{B^{-1}}{2}\sum_{e\in V(U^{d-1})} \frac{\ell_e(y)s_v(A)}{A} - \frac{B^{-3}}{4}\sum_{e\in V(U^{d-1})} \frac{\ell_e(y)s_v(A)^2}{A} + O(A^2\sigma\delta^3).
\end{aligned} \tag{A.20}$$

Let us analyze the first sum Σ_1 in (A.20). Using the definition of the s_v, we see that this sum equals

$$\Sigma_1 = C_1\delta + C_2A\delta^2, \quad \text{where} \quad C_1 = B^{-1}\sum_{e\neq 0} \ell_e(y)\langle \tilde{v}, e\rangle \text{ and } C_2 = \frac{B^{-1}}{2}\sum_{e\neq 0} \ell_e(y)\|e\|^2. \tag{A.21}$$

A similar analysis of the second sum Σ_2 gives

$$\Sigma_2 = C_3A\delta^2 + C_4A^2\delta^3 + C_5A^3\delta^4, \quad \text{where} \quad C_3 = B^{-3}\sum_{e\neq 0} \ell_e(y)\langle \tilde{v}, e\rangle^2, \quad \text{and } |C_4|, |C_5| \leq C_0\sigma, \tag{A.22}$$

where C_0 depends at most on d. In total, this gives

$$\|\bar{\xi}\|S' = B + C_1\delta + \tilde{C}\sigma A\delta^2 + O(\sigma A^2\delta^3), \tag{A.23}$$

where

$$\sigma\tilde{C} = C_2 + C_3, \tag{A.24}$$

and where the constants in the "O" term depend only on d. It is important to notice that

$$\tilde{C} \geq \sigma^{-1}C_2 \geq 1/4. \tag{A.25}$$

Replacing A by one, we get

$$\|\bar{\xi}\|S = B + C_1\delta + \tilde{C}\sigma\delta^2 + O(\sigma\delta^3). \tag{A.26}$$

Notice that these constants are the same as those in (A.23) and again the constants in the "O" term depend only on d.

Next, we want to compute $\|\bar{\xi}\|$ and compare this number with $B + C_1\delta$. We have $\bar{\xi} = (1, \tilde{x})$ where

$$\tilde{x} = \tilde{v} + \delta y = \tilde{v} + \delta\sum_{e\in V(U^{d-1})} \ell_e(y)e.$$

Therefore,

$$\|\tilde{\xi}\|^2 = 1 + \|\tilde{v}\|^2 + 2\delta \sum_{e\in V(U^{d-1})} \ell_e(y)\langle \tilde{v}, e\rangle + \delta^2\|y\|^2 = B^2 + s. \tag{A.27}$$

If we now use (A.18), we obtain

$$\|\bar{\xi}\| = F(s) = B + \frac{B^{-1}}{2}s - \frac{B^{-3}}{4}s^2 + O(s^3) = B + C_1\delta + \tilde{C}'\sigma\delta^2 + O(\sigma\delta^3),$$

where

$$\sigma\tilde{C}' = \frac{B^{-1}}{2}\delta^2\|y\|^2 + B^{-3}\left[\sum_{e\neq 0} \ell_e(y)\langle \tilde{v}, e\rangle\right]^2 \delta^2. \tag{A.28}$$

Here, we have also used the fact that $\|y\| \le \sigma$. If we use this expression for $\|\bar{\xi}\|$ in (A.26), we obtain

$$S = 1 + C^*\sigma\delta^2 + O(\sigma\delta^3) =: 1 + \varepsilon, \quad C^* = \tilde{C} - \tilde{C}'. \tag{A.29}$$

Similarly

$$S' = 1 + C^{**}\sigma\delta^2 + O(\sigma A^2\delta^3) =: 1 + \varepsilon', \quad C^{**} = \tilde{C}A^2 - \tilde{C}'. \tag{A.30}$$

If we choose m sufficiently large ($m \ge m^*$ with m^* depending only on d and A as a sufficiently large integer depending only on d we will have $0 < 2\varepsilon < \varepsilon'$ (see (A.25)). This completes the proof of the Lemma. □

Note that the constant A of this lemma serves to define A for this paper and then $L = (A+1)^2$ is defined as in Lemma 6.5.

Proof of Lemma 6.6. Case $\xi \in W_k$: Let $\xi = \xi_i \in W_k$. Given $t \in [1/2, t_{2m-L}]$, we have $t_i^+ = t_j$ and we take $\tilde{t}_i := t_{j+1}$. We define α by the requirement

$$\alpha t_i^+ + (1-\alpha)\tilde{t}_i = t, \quad \text{i.e.} \quad \alpha = \frac{t - \tilde{t}_i}{t_i^+ - \tilde{t}_i}. \tag{A.31}$$

Then, $\xi = \alpha\xi + (1-\alpha)\xi$, which is the decomposition for ξ required in Lemma 6.6. Indeed, $|\alpha| \le C$ with C depending only on d because of Lemma 6.5 and (6.8).

Case ξ is not in W_k: We will use the constructions given above. We take A to be an integer as given in Lemma A.1. We have given two ways of representing ξ as given in (A.8) and (A.12). The ξ_ν and ξ'_ν appearing in these representations are all from W_k. We take $W_k(\xi)$ as the collection of all these points. Property (ii) of Lemma 6.6 is satisfied since $\|\xi - \xi_i\| \le A/m$ for each i. We define α by the requirement

$$\alpha\epsilon + (1-\alpha)\epsilon' = 0, \quad \text{i.e.} \quad \alpha = \frac{\epsilon'}{\epsilon' - \epsilon}. \tag{A.32}$$

It follows that

$$\xi = \alpha \sum_{\nu\in V(R)} a_\nu\xi_\nu + (1-\alpha) \sum_{\nu\in V(R')} a'_\nu\xi'_\nu, = \sum_{j=1}^M b_j\xi_j, \quad \sum_{j=1}^M b_j = 1, \tag{A.33}$$

where all of the ξ_j are in $W(\xi)$. The key here is that the coefficients in this representation sum to one.

Now, given $t \in [1/2, t_{m-L}]$, we define

$$t^+ := \max\{t_i^+ : \ \xi_i \in W_k(\xi)\} = t_j, \quad \tilde{t} := t_{j+1}. \tag{A.34}$$

Similar to the above, we define β by requiring that

$$\beta t^+ + (1-\beta)\tilde{t} = t, \quad \text{i.e.} \quad \beta = \frac{t - t^+}{t^+ - \tilde{t}}. \tag{A.35}$$

It follows that

$$\xi\cdot x - t = \sum_{j=1}^M \beta b_j(\xi_j\cdot x - t^+) + \sum_{j=1}^M (1-\beta)b_j(\xi_j\cdot x - \tilde{t}). \tag{A.36}$$

This is the decomposition promised in Lemma 6.6 and thereby completes the proof of the lemma. □

Data availability

No data was used for the research described in the article.

References

[1] Francis Bach, Breaking the curse of dimensionality with convex neural networks, J. Mach. Learn. Res. 18 (1) (2017) 629–681.
[2] Andrew R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, IEEE Trans. Inf. Theory 39 (3) (1993) 930–945.
[3] Andrew R. Barron, Albert Cohen, Wolfgang Dahmen, Ronald A. DeVore, Approximation and learning by greedy algorithms, Ann. Stat. 36 (1) (2008) 64–94.
[4] Peter Binev, Andrea Bonito, Ronald DeVore, Guergana Petrova, Optimal learning, Calcolo 61 (1) (2024) 15.
[5] Helmut Bolcskei, Philipp Grohs, Gitta Kutyniok, Philipp Petersen, Optimal approximation with sparsely connected deep neural networks, SIAM J. Math. Data Sci. 1 (1) (2019) 8–45.
[6] Jean Bourgain, Joram Lindenstrauss, Distribution of points on spheres and approximation by zonotopes, Isr. J. Math. 64 (1988) 25–31.
[7] Jean Bourgain, Joram Lindenstrauss, Vitali Milman, Approximation of zonoids by zonotopes, Acta Math. 162 (1) (1989) 73–141.
[8] Claire Boyer, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric de Gournay, Pierre Weiss, On representer theorems and convex regularization, SIAM J. Optim. 29 (2) (2019) 1260–1281.
[9] Kristian Bredies, Marcello Carioni, Sparsity of solutions for variational inverse problems with finite-dimensional data, Calc. Var. Partial Differ. Equ. 59 (1) (2020) 14.
[10] Albert Cohen, Ronald DeVore, Guergana Petrova, Przemyslaw Wojtaszczyk, Optimal stable nonlinear approximation, Found. Comput. Math. 22 (3) (2022) 607–648.
[11] Ronald DeVore, Boris Hanin, Guergana Petrova, Neural network approximation, Acta Numer. 30 (2021) 327–444.
[12] Ronald A. DeVore, Nonlinear approximation, Acta Numer. 7 (1998) 51–150.
[13] Ronald A. DeVore, George G. Lorentz, Constructive Approximation, Grundlehren der Mathematischen Wissenschaften, Springer Berlin Heidelberg, 1993.
[14] Ronald A. DeVore, Vasil A. Popov, Interpolation spaces and non-linear approximation, in: Function Spaces and Applications: Proceedings of the US-Swedish Seminar Held in Lund, Sweden, June 15–21, 1986, Springer, 2006, pp. 191–205.
[15] Weinan E, Chao Ma, Lei Wu, The Barron space and the flow-induced function spaces for neural network models, Constr. Approx. 55 (1) (2022) 369–406.
[16] Gerald B. Folland, Real Analysis: Modern Techniques and Their Applications, second edition, Pure and Applied Mathematics (New York), John Wiley & Sons, Inc., New York, 1999.
[17] Rémi Gribonval, Gitta Kutyniok, Morten Nielsen, Felix Voigtlaender, Approximation spaces of deep neural networks, Constr. Approx. 55 (1) (2022) 259–367.
[18] Lee K. Jones, A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, Ann. Stat. (1992) 608–613.
[19] Jason M. Klusowski, Andrew R. Barron, Approximation by combinations of ReLU and squared ReLU ridge functions with ℓ^1 and ℓ^0 controls, IEEE Trans. Inf. Theory 64 (12) (2018) 7649–7656.
[20] Yury Korolev, Two-layer neural networks with values in a Banach space, SIAM J. Math. Anal. 54 (6) (2022) 6358–6389.
[21] Anders Krogh, John Hertz, A simple weight decay can improve generalization, Adv. Neural Inf. Process. Syst. 4 (1991).
[22] Věra Kůrková, Marcello Sanguineti, Bounds on rates of variable-basis and neural-network approximation, IEEE Trans. Inf. Theory 47 (6) (2001) 2659–2665.
[23] Jianfeng Lu, Zuowei Shen, Haizhao Yang, Shijun Zhang, Deep network approximation for smooth functions, SIAM J. Math. Anal. 53 (5) (2021) 5465–5506.
[24] Yuly Makovoz, Uniform approximation by neural networks, J. Approx. Theory 95 (2) (1998) 215–228.
[25] Jiří Matoušek, Improved upper bounds for approximation by zonotopes, Acta Math. 177 (1) (1996) 55–73.
[26] Hrushikesh N. Mhaskar, On the tractability of multivariate integration and approximation by neural networks, J. Complex. 20 (4) (2004) 561–590.
[27] Hrushikesh N. Mhaskar, Dimension independent bounds for general shallow networks, Neural Netw. 123 (2020) 142–152.
[28] Behnam Neyshabur, Russ R. Salakhutdinov, Nati Srebro, Path-SGD: path-normalized optimization in deep neural networks, Adv. Neural Inf. Process. Syst. 28 (2015).
[29] Greg Ongie, Rebecca Willett, Daniel Soudry, Nathan Srebro, A function space view of bounded norm infinite width ReLU nets: the multivariate case, in: International Conference on Learning Representations, 2020.
[30] Rahul Parhi, Robert D. Nowak, Banach space representer theorems for neural networks and ridge splines, J. Mach. Learn. Res. 22 (43) (2021) 1–40.
[31] Rahul Parhi, Robert D. Nowak, What kinds of functions do deep neural networks learn? Insights from variational spline theory, SIAM J. Math. Data Sci. 4 (2) (2022) 464–489.
[32] Rahul Parhi, Robert D. Nowak, Deep learning meets sparse regularization: a signal processing perspective, IEEE Signal Process. Mag. 40 (6) (2023) 63–74.
[33] Rahul Parhi, Robert D. Nowak, Near-minimax optimal estimation with shallow ReLU neural networks, IEEE Trans. Inf. Theory 69 (2) (2023) 1125–1140.
[34] Allan Pinkus, Approximation theory of the MLP model in neural networks, Acta Numer. 8 (1999) 143–195.
[35] Gilles Pisier, Remarques sur un résultat non publié de B. Maurey, in: Séminaire d'Analyse Fonctionnelle (Dit "Maurey-Schwartz"), April 1981, pp. 1–12.
[36] Michael Reed, Barry Simon, Methods of Modern Mathematical Physics I: Functional Analysis, Academic Press, 1972.
[37] Zuowei Shen, Haizhao Yang, Shijun Zhang, Optimal approximation rate of ReLU networks in terms of width and depth, J. Math. Pures Appl. 157 (2022) 101–135.
[38] Jonathan W. Siegel, Optimal approximation of zonoids and uniform approximation by shallow neural networks, arXiv preprint, arXiv:2307.15285, 2023.
[39] Jonathan W. Siegel, Optimal approximation rates for deep ReLU neural networks on Sobolev and Besov spaces, J. Mach. Learn. Res. 24 (357) (2023) 1–52.
[40] Jonathan W. Siegel, Jinchao Xu, Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks, Found. Comput. Math. (2022) 1–57.
[41] Jonathan W. Siegel, Jinchao Xu, Characterization of the variation spaces corresponding to shallow neural networks, Constr. Approx. (2023) 1–24.
[42] Michael Unser, A unifying representer theorem for inverse problems and machine learning, Found. Comput. Math. 21 (4) (2021) 941–960.
[43] Dmitry Yarotsky, Error bounds for approximations with deep ReLU networks, Neural Netw. 94 (2017) 103–114.