Foundations of Computational Mathematics (2024) 24:481–537
https://doi.org/10.1007/s10208-022-09595-3

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Sharp Bounds on the Approximation Rates, Metric Entropy, and *n*-Widths of Shallow Neural Networks

Jonathan W. Siegel[1] · **Jinchao Xu**[1]

Received: 8 September 2021 / Revised: 17 July 2022 / Accepted: 15 September 2022 /
Published online: 9 November 2022

Abstract

In this article, we study approximation properties of the variation spaces corresponding to shallow neural networks with a variety of activation functions. We introduce two main tools for estimating the metric entropy, approximation rates, and n-widths of these spaces. First, we introduce the notion of a smoothly parameterized dictionary and give upper bounds on the nonlinear approximation rates, metric entropy, and n-widths of their absolute convex hull. The upper bounds depend upon the order of smoothness of the parameterization. This result is applied to dictionaries of ridge functions corresponding to shallow neural networks, and they improve upon existing results in many cases. Next, we provide a method for lower bounding the metric entropy and n-widths of variation spaces which contain certain classes of ridge functions. This result gives sharp lower bounds on the L^2-approximation rates, metric entropy, and n-widths for variation spaces corresponding to neural networks with a range of important activation functions, including ReLU^k activation functions and sigmoidal activation functions with bounded variation.

Keywords Neural networks · Approximation error · n-widths · Metric entropy

Mathematics Subject Classification 62M45 · 41A46

Communicated by Albert Cohen.

✉ Jonathan W. Siegel
jus1949@psu.edu; jwsiegel@tamu.edu

Jinchao Xu
jxx1@psu.edu

[1] Department of Mathematics, Pennsylvania State University, University Park, PA 16802, USA

1 Introduction

1.1 Preliminaries

The class of shallow neural networks with activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is a popular function class used in supervised learning algorithms. This class of functions on $\mathbb{R}^d$ is given by

$$\Sigma_n^d(\sigma) = \left\{ \sum_{i=1}^n \sigma(\omega_i \cdot x + b_i) : \ \omega_i \in \mathbb{R}^d, \ b_i \in \mathbb{R} \right\}, \tag{1.1}$$

where σ is an activation function and n is the width of the network. There is a rich literature on the approximation properties and statistical inference from this class of functions [2, 4, 28, 29, 35], with a special focus on the case when σ is a sigmoidal activation function or when $\sigma = \max(0, x)^k$ is a power of the rectified linear unit.

In this work, we consider the approximation properties of shallow neural networks from the point of view of nonlinear dictionary approximation [17]. Let X be Banach space and $\mathbb{D} \subset X$ be a uniformly bounded dictionary, i.e., $\mathbb{D}$ is a subset such that $\sup_{d \in \mathbb{D}} \|d\|_X = K_{\mathbb{D}} < \infty$. Nonlinear dictionary approximation considers approximating a target function f by nonlinear n-term dictionary expansions, i.e., by an element of the set

$$\Sigma_n(\mathbb{D}) = \left\{ \sum_{j=1}^n a_j h_j : \ h_j \in \mathbb{D} \right\}. \tag{1.2}$$

The approximation is nonlinear since the elements h_j in the expansion will depend upon the target function f. It is often also important to have some control over the coefficients a_j which occur in the expansion (1.2). For this purpose, we introduce the sets

$$\begin{aligned} \Sigma_{n,M}(\mathbb{D}) &= \left\{ \sum_{j=1}^n a_j h_j : \ h_j \in \mathbb{D}, \ \sum_{i=1}^n |a_i| \le M \right\} \\ \text{and } \ \Sigma_{n,M}^\infty(\mathbb{D}) &= \left\{ \sum_{i=1}^n a_i g_i, \ g_i \in \mathbb{D}, \ \max_{i=1,\dots,n} |a_i| \le M \right\}, \end{aligned} \tag{1.3}$$

which correspond to coefficients which are bounded in ℓ^1 and ℓ^∞, respectively.

The application of this framework to shallow neural networks comes by considering the dictionary

$$\mathbb{D}_\sigma^d = \{\sigma(\omega \cdot x + b) : \ \omega \in \mathbb{R}^d, \ b \in \mathbb{R}\} \subset L^p(\Omega), \tag{1.4}$$

where σ is an appropriate activation function and $\Omega \subset \mathbb{R}^d$ is a bounded domain. For this dictionary, the set

$$\Sigma_n(\mathbb{D}_\sigma^d) = \Sigma_n^d(\sigma) = \left\{ \sum_{j=1}^n a_j \sigma(\omega_j \cdot x + b_j) : \ \omega_j \in \mathbb{R}^d, \ b_j \in \mathbb{R} \right\} \tag{1.5}$$

is exactly the set of shallow neural networks with activation function σ and width n. Note that we are suppressing the dependence on the underlying space $L^p(\Omega)$ for notational simplicity.

For activation functions σ which are bounded, the dictionary $\mathbb{D}_\sigma^d$ is uniformly bounded in $L^p(\Omega)$. This holds for the class of sigmoidal activation functions, i.e., activation functions which satisfy $\lim_{x\to-\infty} \sigma(x) = 0$ and $\lim_{x\to\infty} \sigma(x) = 1$, for example. In this case, we will consider the dictionary $\mathbb{D}_\sigma^d$ given in (1.4).

However, when the activation function σ is not bounded, the dictionary $\mathbb{D}_\sigma^d$ will not in general be uniformly bounded in $L^p(\Omega)$. This is the case for the important activation functions $\sigma_k(x) = \mathrm{ReLU}^k(x) := \max(x, 0)^k$ when $k > 0$, for instance. In this case, we modify the definition (1.4) appropriately in order to guarantee uniform boundedness of the dictionary. When $\sigma_k(x) = \mathrm{ReLU}^k(x)$ (here when $k = 0$, we interpret $\sigma_k(x)$ to be the Heaviside function), we constrain the weights ω and b and consider the dictionary

$$\mathbb{P}_k^d = \{\sigma_k(\omega \cdot x + b) : \ \omega \in S^{d-1}, \ b \in [c_1, c_2]\} \subset L^p(\Omega), \tag{1.6}$$

where $S^{d-1} = \{\omega \in \mathbb{R}^d : \ |\omega| = 1\}$ is the unit sphere and the constants c_1 and c_2 are chosen to satisfy

$$c_1 < \inf\{x \cdot \omega, \ \omega \in S^{d-1}, \ x \in \Omega\} < \sup\{x \cdot \omega, \ \omega \in S^{d-1}, \ x \in \Omega\} < c_2. \tag{1.7}$$

We remark that any choice of c_1 and c_2 satisfying the above conditions leads to a dictionary which is equivalent up to polynomials (see [56]). Due to the homogeneity of the activation function σ_k, the set of nonlinear dictionary expansions $\Sigma_n(\mathbb{P}_k^d)$ coincides with the set of shallow ReLU^k neural networks with width n.

An important model class of functions which can be efficiently approximated by nonlinear dictionary expansions is given by the variation norm of the dictionary $\mathbb{D}$ [2, 4, 5, 17, 31, 32]. Consider the set

$$B_1(\mathbb{D}) = \overline{\left\{ \sum_{j=1}^n a_j h_j : \ n \in \mathbb{N}, \ h_j \in \mathbb{D}, \ \sum_{i=1}^n |a_i| \le 1 \right\}}, \tag{1.8}$$

which is the closure of the convex, symmetric hull of $\mathbb{D}$. Using this set, we define a norm, $\|\cdot\|_{\mathscr{K}_1(\mathbb{D})}$, on X given by the gauge (see, for instance, [54]) of $B_1(\mathbb{D})$,

$$\|f\|_{\mathscr{K}_1(\mathbb{D})} = \inf\{c > 0 : \ f \in cB_1(\mathbb{D})\}. \tag{1.9}$$

This norm is defined so that $B_1(\mathbb{D})$ is the unit ball of $\|\cdot\|_{\mathscr{K}_1(\mathbb{D})}$. We also consider the subspace of X defined by the variation norm, which we denote

$$\mathscr{K}_1(\mathbb{D}) := \{f \in X : \ \|f\|_{\mathscr{K}_1(\mathbb{D})} < \infty\}. \tag{1.10}$$

As long as $\sup_{d\in\mathbb{D}} \|d\|_X = K_{\mathbb{D}} < \infty$, the space $\mathscr{K}_1(\mathbb{D})$ is a Banach space (see [56], for instance). The space $\mathscr{K}_1(\mathbb{D})$ is typically called the variation space and the norm $\|\cdot\|_{\mathscr{K}_1(\mathbb{D})}$ is typically called the variation norm or the atomic norm with respect to the dictionary $\mathbb{D}$.

As shown in [56], for the dictionary $\mathbb{P}_1^d$ corresponding to shallow neural networks with ReLU activation function, the space $\mathscr{K}_1(\mathbb{P}_1^d)$ is equivalent to the Barron space introduced and studied in [24, 25]. Further, the space $\mathscr{K}_1(\mathbb{P}_k^d)$ for general k can be characterized in terms of the Radon transform and is equivalent to the Radon BV space introduced in the context of shallow neural networks in [44–46].

The first major problem we consider in this work is how efficiently functions from the variation space $\mathscr{K}_1(\mathbb{D})$ can be approximated by nonlinear dictionary expansions from the sets $\Sigma_n(\mathbb{D})$, $\Sigma_{n,M}(\mathbb{D})$, or $\Sigma_{n,M}^\infty(\mathbb{D})$. There is a rich literature on this problem, which has significant applications to statistics and machine learning (see, for instance, [2, 4, 5, 17, 28, 29, 31, 39, 52]), and the dictionaries $\mathbb{D}_\sigma^d$ and $\mathbb{P}_k^d$ corresponding to shallow neural networks are of particular interest.

There is a close relationship between this problem and fundamental approximation theoretic quantities such as the metric entropy and n-widths of the convex hull $B_1(\mathbb{D})$ (which we recall is the unit ball of $\mathscr{K}_1(\mathbb{D})$). Let us give a brief overview of these notions, for more details, see, for instance, [33, 37, 51, 59]. We remark that the notions of entropy and n-widths can also be naturally defined for operators $T : X \to Y$ between two Banach spaces [48, 49]. Here we will only discuss these notions for subsets of a Banach space for simplicity.

The notion of metric entropy was first introduced by Kolmogorov [30]. The (dyadic) entropy numbers $\varepsilon_n(A)$ of a set $A \subset X$ are defined by

$$\varepsilon_n(A)_X = \inf\{\varepsilon > 0 : \ \text{A is covered by } 2^n \text{ balls of radius } \varepsilon\}. \tag{1.11}$$

Roughly speaking, the entropy numbers indicate how precisely we can specify elements of A given n bits of information and are closely related to approximation by stable nonlinear methods [16].

The Kolmogorov n-widths of a set $A \subset X$ measure how accurately the set A can be approximated by linear subspaces and are given by

$$d_n(A)_X = \inf_{Y_n} \sup_{x\in A} \inf_{y\in Y_n} \|x - y\|_X, \tag{1.12}$$

where the first infimum is over the collection of subspaces Y_n of dimension n. The Gelfand n-widths, which are important in compressed sensing [20], measure how accurately elements from A can be recovered from linear measurements and are given by

$$d^n(A)_X = \inf_{U_n} \sup\{\|x\|_H : x \in U_n \cap A\}, \tag{1.13}$$

where the infimum is taken over all closed subspaces U_n of codimension n. The linear n-widths are closely related to the Kolmogorov n-widths but require the approximation to be given by a linear map and are defined by

$$\delta_n(A)_X = \inf_{T_n} \sup_{x \in A} \|x - T_n(x)\|_X, \tag{1.14}$$

where the infimum is taken over all linear operators of rank n. In a Hilbert space, the linear n-widths and Kolmogorov n-widths coincide since T_n can be taken as the orthogonal projection operator. Finally, the Bernstein widths are given by

$$b_n(A)_X = \sup_{X_n} \inf_{x \in \partial(A \cap X_n)} \|x\|_X, \tag{1.15}$$

where the supremum is taken over all subspaces $X_n \subset X$ of dimension n. The Bernstein widths give a lower bound on the best possible continuous nonlinear approximation of the set A using n-parameters [18].

The second main problem we consider is the determination of the asymptotics of the metric entropy and the different types of n-widths mentioned above for the class $B_1(\mathbb{D}) \subset X$ for different dictionaries $\mathbb{D} \subset X$. This problem has been studied in functional analysis, probability theory, and approximation theory [3, 7, 10–13, 21, 22], and has important applications to statistical learning theory. For example, the asymptotics of the metric entropy can be used to determine the minimax rates of convergence for statistical estimators on a class of functions [63].

1.2 Prior Results

Let us begin with results concerning the approximation of the convex hull $B_1(\mathbb{D})$ by nonlinear dictionary expansions. A classical result of Maurey [52] (see also [4, 17, 28]) states that if X is a type-2 Banach space, which includes all Hilbert spaces and L^p for $2 \leq p < \infty$, then for functions $f \in B_1(\mathbb{D})$ we have the approximation rate

$$\inf_{f_n \in \Sigma_{n,1}(\mathbb{D})} \|f - f_n\|_X \lesssim K_{\mathbb{D}} n^{-\frac{1}{2}}, \tag{1.16}$$

where the suppressed constant depends only upon the space X. An equivalent formulation of this result, which is sometimes more convenient is that for $f \in \mathscr{K}_1(\mathbb{D})$ we have

$$\inf_{f_n \in \Sigma_{n,M}(\mathbb{D})} \|f - f_n\|_X \lesssim K_{\mathbb{D}} \|f\|_{\mathscr{K}_1(\mathbb{D})} n^{-\frac{1}{2}}, \tag{1.17}$$

where the bound M can be taken as $M = \|f\|_{\mathscr{K}_1(\mathbb{D})}$. In Sect. 2, we give the precise definition of a type-2 Banach space and prove (1.16) using Maurey's sampling argument, which is a fundamental building block of the theory.

Applying this result to the dictionaries $\mathbb{D}_\sigma^d$ and $\mathbb{P}_k^d$ immediately yields the following dimension independent approximation rates in L^p for $2 \leq p < \infty$:

$$\inf_{f_n \in \Sigma_n^d(\sigma)} \|f - f_n\|_{L^p(\Omega)} \lesssim \|f\|_{\mathscr{K}_1(\mathbb{D}_\sigma^d)} n^{-\frac{1}{2}} \tag{1.18}$$

for shallow neural networks with bounded activation function σ on the variation space $\mathscr{K}_1(\mathbb{D}_\sigma^d)$, and

$$\inf_{f_n \in \Sigma_n^d(\sigma_k)} \|f - f_n\|_{L^p(\Omega)} \lesssim \|f\|_{\mathscr{K}_1(\mathbb{P}_k^d)} n^{-\frac{1}{2}} \tag{1.19}$$

for shallow ReLU^k neural networks on the variation space $\mathscr{K}_1(\mathbb{P}_k^d)$. These results were first obtained in the L^2-norm by Jones [28] for the activation function $\sigma(x) = \cos(x)$ and by Barron [4] for sigmoidal activation functions. To be precise, Barron obtained approximation rates in terms of the spectral Barron semi-norm

$$|f|_{\mathscr{B}_s(\Omega)} := \inf_{f_e|_\Omega = f} \int_{\mathbb{R}^d} |\hat{f}(\xi)||\xi|^s \mathrm{d}\xi, \tag{1.20}$$

where the infimum is over all extensions $f_e \in L^1(\mathbb{R}^d)$ of f. Barron showed that modulo constants the semi-norm (1.20) for $s = 1$ controls the variation norm $\mathscr{K}_1(\mathbb{D}_\sigma^d)$ for any sigmoidal activation function σ, from which his seminal result

$$\inf_{f_n \in \Sigma_n^d(\sigma)} \|f - f_n\|_{L^2(\Omega)} \lesssim |f|_{\mathscr{B}_1(\Omega)} n^{-\frac{1}{2}} \tag{1.21}$$

follows [4].

It will be more convenient for us in what follows to consider the spectral Barron norm instead of the semi-norm (1.20), which is given by [55]

$$\|f\|_{\mathscr{B}_s(\Omega)} := \inf_{f_e|_\Omega = f} \int_{\mathbb{R}^d} |\hat{f}(\xi)|(1 + |\xi|)^s \mathrm{d}\xi, \tag{1.22}$$

and the corresponding spectral Barron space $\mathscr{B}_s$. Barron's results have been generalized to ReLU^k activation functions in [29, 62]. It is shown that

$$\|f\|_{\mathscr{K}_1(\mathbb{P}_k^d)} \lesssim \|f\|_{\mathscr{B}_{k+1}(\Omega)}, \tag{1.23}$$

from which it follows that the approximation rate (1.19) applies to the spectral Barron space $\mathscr{B}_{k+1}(\Omega)$. These results have also been extended to a very general class of activation functions in [55]. Interestingly, (1.23) is not an equivalence and quantifying the gap between $\mathscr{K}_1(\mathbb{P}_k^d)$ and $\mathscr{B}_{k+1}(\Omega)$ is an open problem.

An important result, first observed by Makovoz [39], is that for certain dictionaries the rate in (1.16) can be improved. In particular, for the dictionary $\mathbb{D}_\sigma^d$ corresponding

to neural networks with certain sigmoidal activation functions, Makovoz obtained the approximation rate

$$\inf_{f_n \in \Sigma_{n,M}(\mathbb{D}_\sigma^d)} \|f - f_n\|_{L^2(\Omega)} \lesssim n^{-\frac{1}{2}-\frac{1}{2d}} \tag{1.24}$$

for $f \in B_1(\mathbb{D}_\sigma^d)$. (Note that here and in what follows the implied constant is independent of n, and the ℓ^1-bound M is fixed and independent of n as well.) Furthermore, improved rates have been obtained for other dictionaries. In particular, in [29], the dictionaries $\mathbb{P}_k^d$ corresponding to neural networks with activation function $\sigma = [\max(0, x)]^k$ are studied for $k = 1, 2$ and it is shown that for $f \in B_1(\mathbb{P}_k^d)$

$$\inf_{f_n \in \Sigma_{n,M}(\mathbb{P}_k^d)} \|f - f_n\|_{L^2(\Omega)} \lesssim n^{-\frac{1}{2}-\frac{1}{d}}. \tag{1.25}$$

We remark that more generally, it is shown that these rates hold in L^∞ up to logarithmic factors. This analysis is extended to $k \geq 3$ in [62], where the same approximation rate is attained.

In addition, when $k = 1$, i.e., for the ReLU activation function, the rate (1.25) can be improved to

$$\inf_{f_n \in \Sigma_{n,M}(\mathbb{P}_1^d)} \|f - f_n\|_{L^2(\Omega)} \lesssim n^{-\frac{1}{2}-\frac{3}{2d}}, \tag{1.26}$$

and this result holds also in L^∞ [2]. This result is proved using a different set of combinatorial arguments from geometric discrepancy theory [41].

These results raise the natural question of what the optimal approximation rates for $\Sigma_{n,M}(\mathbb{P}_k^d)$ on the set $B_1(\mathbb{P}_k^d)$ are. Specifically, for each $k = 0, 1, 2, \ldots$ and dimension $d = 2, \ldots$ (the case $d = 1$ is comparatively trivial), what is the largest possible value of $\alpha := \alpha(k, d)$ such that for $f \in B_1(\mathbb{P}_k^d)$ we have

$$\inf_{f_n \in \Sigma_{n,M}(\mathbb{P}_k^d)} \|f - f_n\|_{L^2(\Omega)} \lesssim n^{-\frac{1}{2}-\alpha(k,d)}. \tag{1.27}$$

The results above imply that $\alpha(k, d) \geq \frac{1}{2d}$ for $k = 0$, $\alpha(k, d) \geq \frac{3}{2d}$ for $k = 1$, and $\alpha(k, d) \geq \frac{1}{d}$ for $k > 1$. When $d > 1$, the best available upper bounds on $\alpha(k, d)$ are $\alpha(k, d) \leq \frac{k+1}{d}$ (see [29, 39]), except in the case $k = 0, d = 2$, where Makovoz obtains the sharp bound $\alpha(0, 2) = \frac{1}{4}$ [39].

1.3 Main Results

Our results can be divided into two categories. First, we consider upper bounds on approximation rates, entropy, and n-widths. Previous approximation rates, such as those obtained in [29, 39, 62], and entropy bounds on the convex hull $B_1(\mathbb{D})$, such as

those obtained in [9, 12], rely upon covering the dictionary $\mathbb{D}$ by balls of radius ε and using a stratified sampling argument.

In order to improve upon these results, the key idea is to use the smoothness of the dictionary $\mathbb{P}_k^d$. We obtain a higher-order generalization of these results by introducing the notion of a smoothly parameterized dictionary. A dictionary $\mathbb{D}$ is smoothly parameterized to order s if there exists a parameterization map $\mathscr{P} : \mathscr{M} \to X$, where $\mathscr{M}$ is a smooth manifold, such that $\mathbb{D} \subset \mathscr{P}(\mathscr{M})$, and such that $\mathscr{P}$ is smooth to order s in an appropriate sense. We give a detailed definition in Sect. 3, where we show that if $\mathbb{D}$ is smoothly parameterized to order s by a compact d-dimensional manifold $\mathscr{M}$ and X is a type-2 Banach space, then

$$\inf_{f_n \in \Sigma_{n,M}(\mathbb{D})} \|f - f_n\|_X \lesssim n^{-\frac{1}{2}-\frac{s}{d}}, \tag{1.28}$$

for some $M < \infty$. Here the implied constant depends upon the manifold $\mathscr{M}$, the parameterization $\mathscr{P}$, and the type-2 constant of X. We apply this result to the dictionaries $\mathbb{P}_k^d \subset L^p(\Omega)$, which we show are smoothly parameterized to order $k + \frac{1}{p}$ by the compact, d-dimensional manifold $S^{d-1} \times [c_1, c_2]$. This allows us to improve upon the approximation rates for ReLU^k networks described above when $k \geq 2$, and in particular show that $\alpha(k, d) \geq \frac{2k+1}{2d}$ for all $k \geq 0$.

We also give upper bounds on the metric entropy and n-widths of the convex hull $B_1(\mathbb{D})$ when $\mathbb{D}$ is a smoothly parameterized dictionary. In particular, in Sect. 3 we show that if $\mathbb{D}$ is smoothly parameterized to order s by a compact, d-dimensional manifold, then we have

$$d_n(B_1(\mathbb{D}))_X \lesssim n^{-\frac{s}{d}}. \tag{1.29}$$

If in addition the space X is a type-2 Banach space, then we have the bound

$$\varepsilon_n(B_1(\mathbb{D}))_X \lesssim n^{-\frac{1}{2}-\frac{s}{d}}. \tag{1.30}$$

Finally, if X is a Hilbert space, we obtain an analogous bound on the Gelfand numbers of the convex hull of $\mathbb{D}$. The Gelfand numbers of a convex hull are introduced and studied in [10, 12]. While closely related to the Gelfand widths $d^n(B_1(\mathbb{D}))$, they are subtly different [26]. We give the precise definition of the Gelfand numbers and provide an example illustrating this difference in Sect. 3.5.

These results generalize the results from [9–12] by taking into account the smoothness of the dictionary in addition to the compactness. The application to $\mathbb{P}_k^d$ gives the bound

$$d_n(B_1(\mathbb{P}_k^d))_{L^p(\Omega)} \lesssim n^{-\frac{pk+1}{pd}} \tag{1.31}$$

on the Kolmogorov n-widths for $1 < p < \infty$, and the bound

$$\varepsilon_n(B_1(\mathbb{P}_k^d))_{L^p(\Omega)} \lesssim n^{-\frac{1}{2}-\frac{pk+1}{pd}} \tag{1.32}$$

on the metric entropy for $2 \leq p < \infty$. Based on the L^∞ approximation rates which can be derived using geometric discrepancy theory [2, 41] in the case $k = 1$, we only expect the entropy upper bound to be sharp for $p = 2$, however.

Next, we consider lower bounds on approximation rates, metric entropy, and n-widths of the spaces $B_1(\mathbb{D}_\sigma^d)$ and $B_1(\mathbb{P}_k^d)$. The key here is a generalization of a construction of Makovoz [39], which enables us to find a large collection of nearly orthogonal functions in $B_1(\mathbb{D})$ when the dictionary $\mathbb{D}$ contains a certain class of ridge functions. As a consequence of this construction, we obtain the following lower bounds on the entropy and n-widths in L^2

$$d_n\left(B_1(\mathbb{P}_k^d)\right)_{L^2(\Omega)} \gtrsim n^{-\frac{2k+1}{2d}},\ \varepsilon_n\left(B_1(\mathbb{P}_k^d)\right)_{L^2(\Omega)} \gtrsim n^{-\frac{1}{2}-\frac{2k+1}{2d}},\ b_n(B_1(\mathbb{P}_k^d))_{L^2(\Omega)} \gtrsim n^{-\frac{1}{2}-\frac{2k+1}{2d}}. \tag{1.33}$$

This method is quite general, and one can also obtain lower bounds for $B_1(\mathbb{D}_\sigma^d)$ for a variety of other activation functions σ, but we do not pursue this further here. Note that this lower bound combined with the upper bound (1.32) for $p = 2$ gives the sharp rate of decay for the metric entropy of $B_1(\mathbb{P}_k^d)$

$$\varepsilon_n(B_1(\mathbb{P}_k^d))_{L^2(\Omega)} \eqsim n^{-\frac{1}{2}-\frac{2k+1}{2d}}. \tag{1.34}$$

In addition, the lower bounds on the entropy enable us to obtain the sharp upper bound $\alpha(k, d) \leq \frac{2k+1}{2d}$ on the exponent of approximation by shallow ReLU^k networks. Further, we use these lower bounds to show that the exponent in the approximation rates (1.27) cannot be improved even when relaxing the ℓ^1-norm constraint on the weights to an ℓ^∞-norm constraint, i.e., by using the set $\Sigma_{n,M}^\infty(\mathbb{P}_k^d)$ instead of $\Sigma_{n,M}(\mathbb{P}_k^d)$.

Making use of a technical lemma proved in Sect. 4.1, these lower bounds also carry over to sigmoidal activation functions σ which have bounded variation. This generalizes previous results, which require more stringent assumptions on the activation function σ in order to obtain lower bounds [38].

The entropy lower bounds also quantify the gap between the spaces $\mathscr{K}_1(\mathbb{P}_k^d)$ and $\mathscr{B}_{k+1}(\Omega)$. Specifically, the unit ball in the spectral Barron space $\mathscr{B}_{k+1}(\Omega)$ satisfies

$$\varepsilon_{n\log n}(\{f : \|f\|_{\mathscr{B}_{k+1}(\Omega)} \leq 1\})_{L^2(\Omega)} \lesssim n^{-\frac{1}{2}-\frac{k+1}{d}}. \tag{1.35}$$

This follows from the approximation results obtained in [58] combined with Theorem 10. Conversely, the metric entropy of $B_1(\mathbb{P}_k^d)$ decays at the rate (1.34), which is slower by a factor of $n^{1/2d}$.

Finally, the lower bounds on the metric entropy, Kolmogorov, and Bernstein n-widths show that linear methods are dramatically worse that shallow neural networks, and even nonlinear methods cannot improve upon shallow neural networks on the class $B_1(\mathbb{P}_k^d)$ if either stability [16] or continuity [18] of the approximation method is required.

The paper is organized as follows. In Sect. 3, we study smoothly parameterized dictionaries in detail and derive approximation rates for the set $B_1(\mathbb{D})$ when $\mathbb{D}$ is

smoothly parameterized by a compact manifold. Here we also extend existing methods to obtain upper bounds on the entropy and n-widths of $B_1(\mathbb{D})$ for such dictionaries $\mathbb{D}$. We apply these results to obtain an upper bound on the approximation rates, entropy numbers and n-widths of $B_1(\mathbb{P}_k^d)$. Next, in Section, Finally, we give some concluding remarks and further research directions.

2 Type-2 Spaces and Maurey's Sampling Argument

In this section, we give the definition and basic properties of type-2 Banach spaces and prove the approximation rate (1.16) using a sampling argument due to Maurey [52]. This sampling argument was first used in the context of neural network approximation by Barron [4]. We refer to [34], Chapter 9 and [1], Chapter 6 for a detailed theory of type-2 spaces (and the more general type-p spaces which we will not discuss further here).

Definition 1 A Banach space X is a type-2 Banach space if there exists a constant $C_{2,X} < \infty$ such that for any $n \geq 1$ and $f_1, ..., f_n \in X$ we have

$$\left(\mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i f_i \right\|_X^2 \right)^{1/2} \leq C_{2,X} \left(\sum_{i=1}^n \| f_i \|_X^2 \right)^{1/2}, \tag{2.1}$$

where the expectation is taken over independent Rademacher random variables $\varepsilon_1, ..., \varepsilon_n$, i.e.,

$$\mathbb{P}(\varepsilon_i = 1) = \mathbb{P}(\varepsilon_i = -1) = \frac{1}{2}.$$

The constant $C_{2,X}$ is called the type-2 constant of the space X.

It is clear that a Hilbert space X is a type-2 space with type-2 constant $C_{2,X} = 1$, since by expanding the inner product we have

$$\mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i f_i \right\|_X^2 = \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}(\varepsilon_i \varepsilon_j) \langle f_i, f_j \rangle_X = \sum_{i=1}^n \| f_i \|_X^2. \tag{2.2}$$

Here independence of the Rademacher variables implies that $\mathbb{E}(\varepsilon_i \varepsilon_j) = \delta_{ij}$.

Let $(X, \mathscr{A}, \mu)$ be a measure space. We also have that $L^p(\mu)$ for $2 \leq p < \infty$ is a type-2 Banach space. This follows from Khintchine's inequality, which states that for $0 < p < \infty$ there exists a constant $C_p < \infty$ such that for $n \geq 1$ and real numbers $x_1, ..., x_n$ we have

$$\left(\mathbb{E} \left| \sum_{i=1}^n \varepsilon_i x_i \right|^p \right)^{\frac{1}{p}} \leq C_p \left(\sum_{i=1}^n |x_i|^2 \right)^{\frac{1}{2}}. \tag{2.3}$$

Letting $f_1, ..., f_n \in L^p(\mu)$ and interchanging the expectation and integration, we calculate

$$\begin{aligned}
&\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i f_i\right\|_{L^p(\mu)}^p \\
&= \int_X \mathbb{E}\left|\sum_{i=1}^n \varepsilon_i f_i(x)\right|^p \mathrm{d}\mu(x) \le C_p^p \int_X \left(\sum_{i=1}^n f_i(x)^2\right)^{p/2} \mathrm{d}\mu(x) \\
&= C_p^p \left\|\sum_{i=1}^n f_i^2\right\|_{L^{p/2}(\mu)}^{p/2}. \qquad (2.4)
\end{aligned}$$

Taking both sides to the power $2/p$ and using the triangle inequality (valid since $p \ge 2$), we get

$$\mathbb{E}\left\|\sum_{i=1}^n \varepsilon_i f_i\right\|_{L^p(\mu)}^2 \le C_p^2 \left\|\sum_{i=1}^n f_i^2\right\|_{L^{p/2}(\mu)} \le C_p^2 \sum_{i=1}^n \|f_i^2\|_{L^{p/2}(\mu)} = C_p^2 \left\|\sum_{i=1}^n f_i\right\|_{L^p(\mu)}^2, \qquad (2.5)$$

which proves that $L^p(\mathrm{d}\mu)$ is a type-2 Banach space with type-2 constant C_p. We remark that the optimal constant C_p in Khintchine's inequality is known [27] and scales like $\sqrt{p}$.

Finally, we prove the approximation rate (1.16) in type-2 Banach spaces, which is originally due to Maurey [52].

Theorem 1 *Suppose that X is a type-2 Banach space and $\mathbb{D} \subset X$ is a dictionary with $K_{\mathbb{D}} := \sup_{d\in\mathbb{D}} \|d\|_X < \infty$. Then for $f \in B_1(\mathbb{D})$, we have*

$$\inf_{f_n \in \Sigma_{n,1}(\mathbb{D})} \|f - f_n\|_X \le 4C_{2,X} K_{\mathbb{D}} n^{-\frac{1}{2}}. \qquad (2.6)$$

Proof Let $\delta > 0$ be arbitrary. Since $f \in B_1(\mathbb{D})$, for some integer $N = N(\delta)$, there exists an f_δ of the form

$$f_\delta = \sum_{i=1}^N a_i d_i \qquad (2.7)$$

with $d_i \in \mathbb{D}$ and $\sum_{i=1}^N |a_i| = 1$ such that $\|f - f_\delta\|_X < \delta$. We will show that there exists an $f_n \in \Sigma_{n,1}$ with

$$\|f_\delta - f_n\|_X \le 4C_{2,X} K_{\mathbb{D}} n^{-1/2}$$

which completes the proof since $\delta > 0$ was arbitrary. To this end, define a random variable Y with values in X by (recall that $\sum_{i=1}^{N} |a_i| = 1$)

$$\mathbb{P}(Y = \operatorname{sgn}(a_i) d_i) = |a_i|. \tag{2.8}$$

Note that by construction we have $\mathbb{E}(Y) = f_\delta$. Let $Y_1, ..., Y_n$ be independent copies of Y and consider the empirical average

$$Z_n = \frac{1}{n} \sum_{i=1}^{n} Y_i. \tag{2.9}$$

We will show that

$$\mathbb{E}\|Z_n - f_\delta\|_X^2 = \mathbb{E} \left\| \frac{1}{n} \sum_{i=1}^{n} (Y_n - f_\delta) \right\|_X^2 \leq 16 C_{2,X}^2 K_{\mathbb{D}}^2 n^{-1}. \tag{2.10}$$

This implies that there must be a realization of the random variable Z_n, i.e., a value $Z_n(\omega)$ for some $\omega \in \Omega$ in the underlying probability space, such that $\|Z_n(\omega) - f_\delta\|_X \leq 2C_{2,X} K_{\mathbb{D}} n^{-1/2}$. Since the values of the random variable Z_n lie in $\Sigma_{n,1}(\mathbb{D})$ by construction, this completes the proof.

To prove (2.10), we will show that if a sequence of i.i.d. random variables $R_1, ..., R_n$ with values in X satisfies $\mathbb{E}(R_i) = 0$ and $\|R_i\|_X \leq M$ almost surely, then

$$\mathbb{E} \left\| \sum_{i=1}^{n} R_i \right\|_X^2 \leq 4n C_{2,X}^2 M^2. \tag{2.11}$$

Applying this to the sequence $R_i = n^{-1}(Y_n - f_\delta)$ completes the proof since $\|Y_n - f_\delta\|_X \leq \|Y_n\| + \|f_\delta\|_X \leq 2K_{\mathbb{D}}$ almost surely.

Finally, we prove (2.11) using a symmetrization argument and the type-2 property of X. Let $R_1', ..., R_n'$ denote a new set of independent copies of $R_1, ..., R_n$ and let $\mathbb{E}'$ denote the expectation over $R_1', ..., R_n'$ and $\mathbb{E}$ denote the expectation over the original $R_1, ..., R_n$. Using the zero mean property of the R_i and Jensen's inequality, we get

$$\mathbb{E} \left\| \sum_{i=1}^{n} R_i \right\|_X^2 \leq \mathbb{E}\mathbb{E}' \left\| \sum_{i=1}^{n} R_i - R_i' \right\|_X^2. \tag{2.12}$$

For any fixed choice of signs $\varepsilon_1, ..., \varepsilon_n$, the distribution of $\sum_{i=1}^{n} \varepsilon_i (R_i - R_i')$ is the same. This is due the fact that R_i and R_i' are i.i.d. and switching the sign ε_i is the same as swapping R_i and R_i'. This means that

$$\mathbb{E}\mathbb{E}' \left\| \sum_{i=1}^{n} R_i - R_i' \right\|_X^2 = \mathbb{E}_\varepsilon \mathbb{E}\mathbb{E}' \left\| \sum_{i=1}^{n} \varepsilon_i (R_i - R_i') \right\|_X^2, \tag{2.13}$$

where $\mathbb{E}_\varepsilon$ denotes an average over the Rademacher random variables $\varepsilon_1, \ldots, \varepsilon_n$. Switching the order of the expectation and using the type-2 property of X, we get

$$\mathbb{E}\mathbb{E}'\mathbb{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i (R_i - R_i') \right\|_X^2 \leq C_{2,X}^2 \mathbb{E}\mathbb{E}' \sum_{i=1}^n \|R_i - R_i'\|_X^2. \tag{2.14}$$

Finally, since $\|R_i\|_X \leq M$ almost surely, we get that $\|R_i - R_i'\|_X \leq 2M$ almost surely so that

$$\mathbb{E} \left\| \sum_{i=1}^n R_i \right\|_X^2 \leq 4nC_{2,X}^2 M^2, \tag{2.15}$$

which completes the proof. □

3 Smoothly Parameterized Dictionaries

Let X be a Banach space and consider a dictionary $\mathbb{D} \subset X$ which is parameterized by a smooth manifold $\mathscr{M}$, i.e., for which there exists a surjection

$$\mathscr{P} : \mathscr{M} \to \mathbb{D}. \tag{3.1}$$

In this section, we consider dictionaries $\mathbb{D}$ which are parameterized by a smooth compact manifold $\mathscr{M}$ and study how the approximation properties of the convex hull $B_1(\mathbb{D})$ depend upon the smoothness of the parameterization map $\mathscr{P}$. Specifically, we give upper bounds on the metric entropy and n-widths of $B_1(\mathbb{D})$, and upper bounds on the approximation rates for $B_1(\mathbb{D})$ from sparse convex combinations $\Sigma_{n,M}(\mathbb{D})$, as a function of the degree of smoothness of the parameterization map $\mathscr{P}$.

We being by introducing the relevant notion of smoothness. Throughout this section, $\mathscr{M}$ will denote a smooth manifold with a smooth boundary. For a given $s > 0$ and domain $\Omega \subset \mathbb{R}^d$, we consider the Lipschitz space $\mathrm{Lip}(s, L^\infty(\Omega))$ (see [37] Chapter 2, for instance) with semi-norm given by

$$|f|_{\mathrm{Lip}(s,L^\infty(\Omega))} = \sup_{x,y\in\Omega} \frac{|D^k f(x) - D^k f(y)|}{|x-y|^\alpha}. \tag{3.2}$$

Here $s = k + \alpha$, k is an integer, D^k represents the k-th derivative (tensor), and $0 < \alpha \leq 1$. If s is not an integer, it is well known that the space $\mathrm{Lip}(s, L^\infty(\Omega))$ is equivalent to the Besov space $B^s_{\infty,\infty}(\Omega)$, but when s is an integer, the Lipschitz space is slightly smaller [37]. The first step is to extend this definition to Banach space valued functions f.

Definition 2 Let X be a Banach space, $U \subset \mathbb{R}^d$ an open set and $s > 0$. A function $\mathscr{F} : U \to X$ is of smoothness class s, which we write $\mathscr{F} \in \mathrm{Lip}_\infty(s, X)$, if for every

$\xi \in X^*$, the function $f_\xi : U \to \mathbb{R}$ defined by

$$f_\xi(x) = \langle \xi, \mathscr{F}(x) \rangle \tag{3.3}$$

satisfies

$$|f_\xi|_{\mathrm{Lip}(s, L^\infty(\Omega))} \leq K \|\xi\|_{X^*} \tag{3.4}$$

for a constant $K < \infty$. The smallest constant K above is the semi-norm $|\mathscr{F}|_{\mathrm{Lip}_\infty(s,X)}$.

In order to apply our method to more general dictionaries, we generalize this definition to allow the domain to be a smooth manifold.

Definition 3 Let X be a Banach space, $\mathscr{M}$ a smooth d-dimensional manifold, and $s > 0$. A map $\mathscr{P} : \mathscr{M} \to X$ is of smoothness class s, which we write $\mathscr{P} \in \mathrm{Lip}_\infty^{\mathscr{M}}(s, X)$, if for each coordinate chart (U, ϕ) we have $\mathscr{P} \circ \phi \in \mathrm{Lip}_\infty(s, X)$.

To illustrate this definition, we consider the dictionary $\mathbb{P}_k^d$ with respect to $X = L^p(\Omega)$.

Lemma 1 *Let $\mathscr{M} = S^{d-1} \times [c_1, c_2]$, $1 \leq p < \infty$, and consider the parameterization map $\mathscr{P}_k^d : \mathscr{M} \to L^p(\Omega)$ given by*

$$\mathscr{P}_k^d(\omega, b) = \sigma_k(\omega \cdot x + b) \in L^p(\Omega). \tag{3.5}$$

Then $\mathscr{P}_k^d \in \mathrm{Lip}_\infty^{\mathscr{M}}\left(k + \frac{1}{p}, L^p(\Omega)\right)$.

Proof Let $\xi(x) \in L^q(\Omega)$ with $\frac{1}{p} + \frac{1}{q} = 1$. Then we have

$$f_\xi(\omega, b) = \int_\Omega \sigma_k(\omega \cdot x + b)\xi(x)\mathrm{d}x. \tag{3.6}$$

We view $\mathscr{M}$ as embedded into $\mathbb{R}^{d+1}$ and calculate the derivatives

$$\begin{aligned} \frac{\partial^k}{\partial\omega_{i_1} \cdots \partial\omega_{i_j} \partial b^{k-j}} f_\xi(\omega, b) &= \int_\Omega \frac{\partial^k}{\partial\omega_{i_1} \cdots \partial\omega_{i_j} \partial b^{k-j}} \sigma_k(\omega \cdot x + b)\xi(x)\mathrm{d}x \\ &= k! \int_\Omega \left(\prod_{l=1}^{j} x_{i_l} \right) \sigma_0(\omega \cdot x + b)\xi(x)\mathrm{d}x, \end{aligned} \tag{3.7}$$

since $\sigma_k^{(k)} = k!\sigma_0$. Because Ω is a bounded domain, $\left(\prod_{l=1}^{j} x_{i_l}\right) \in L^\infty(\Omega)$ and so there exists a constant $C(k, \Omega)$ such that for all indices $i_1, .., i_j$ we have

$$\left\| k! \left(\prod_{l=1}^{j} x_{i_l} \right) \xi(x) \right\|_{L^q(\Omega, \mathrm{d}x)} \leq C(k, \Omega) \|\xi\|_{L^q(\Omega)}. \tag{3.8}$$

Then we have

$$\begin{aligned}&\left|\frac{\partial^k}{\partial\omega_{i_1}\cdots\partial\omega_{i_j}\partial b^{k-j}}f_\xi(\omega,b)-\frac{\partial^k}{\partial\omega_{i_1}\cdots\partial\omega_{i_j}\partial b^{k-j}}f_\xi(\omega',b')\right|\\&\leq k!\int_\Omega\left(\prod_{l=1}^{j}x_{i_l}\right)|\sigma_0(\omega\cdot x+b)-\sigma_0(\omega'\cdot x+b')|\xi(x)\mathrm{d}x\\&\leq C(k,\Omega)\|\xi\|_{L^q(\Omega)}\|\sigma_0(\omega\cdot x+b)-\sigma_0(\omega'\cdot x+b')\|_{L^p(\Omega)},\end{aligned}\tag{3.9}$$

which means that

$$\begin{aligned}&\left|D^k f_\xi(\omega,b)-D^k f_\xi(\omega',b')\right|\lesssim_{k,\Omega}\|\xi\|_{L^q(\Omega)}\|\sigma_0(\omega\cdot x+b)\\&\quad-\sigma_0(\omega'\cdot x+b')\|_{L^p(\Omega)}.\end{aligned}\tag{3.10}$$

The proof will be complete if we can show that

$$\|\sigma_0(\omega\cdot x+b)-\sigma_0(\omega'\cdot x+b')\|_{L^p(\Omega)}\lesssim_{p,\Omega}(|\omega-\omega'|+|b-b'|)^{\frac{1}{p}}.\tag{3.11}$$

This follows since the function $\sigma_0(\omega\cdot x+b)-\sigma_0(\omega'\cdot x+b')$ is zero except on the set

$$\begin{aligned}S=\{x\in\Omega:\omega\cdot x+b<0\leq\omega'\cdot x+b'\}\cup\{x\in\Omega:\omega'\cdot x\\+b'<0\leq\omega\cdot x+b\},\end{aligned}\tag{3.12}$$

where it is ±1. The set S is a wedge or strip within Ω of width proportional to $|\omega-\omega'|+|b-b'|$ (see also [39]), and thus, we have

$$\|\sigma_0(\omega\cdot x+b)-\sigma_0(\omega'\cdot x+b')\|_{L^p(\Omega)}^p=|S|\lesssim_\Omega|\omega-\omega'|+|b-b'|,\tag{3.13}$$

which completes the proof. □

In the following analysis, it will be convenient to use the following technical lemma which allows us to reduce to the case where $\mathscr{M}$ is a d-dimensional cube.

Lemma 2 *Suppose that $\mathscr{M}$ is a d-dimensional compact smooth manifold (potentially with boundary) and we are given a parameterization $\mathscr{P}\in\mathrm{Lip}_\infty(s,X)$. Then there exist finitely many maps $\mathscr{P}_j:[-1,1]^d\to X$, $j=1,\ldots,T$, such that $\mathscr{P}_j\in\mathrm{Lip}_\infty(s,X)$ for each j and*

$$\mathscr{P}(\mathscr{M})=\bigcup_{j=1}^{T}\mathscr{P}_j([-1,1]^d).\tag{3.14}$$

Proof Let $x \in \mathscr{M}$ be an interior point. Take a chart $\phi_x : U_x \to \mathscr{M}$ containing x. Here U_x may be homeomorphic to the upper half-plane with boundary if the chart contains boundary points. Let $T_x : C \to U_x$ be a smooth injective map from the cube $C = [-1, 1]^d$ to U_x such that $\phi_x^{-1}(x)$ is in the interior of $T_x(C)$. Such a map can clearly always be found since $\phi_x^{-1}(x) \in \text{int}(U_x)$ as $x \in \text{int}(\mathscr{M})$. Define $V_x = \phi_x \circ T_x(C^o)$, where C^o is the interior of C.

For boundary points $x \in \mathscr{M}$, we take a chart $U_x \to \mathscr{M}$ containing x. In this case, U_x is homeomorphic to the upper half-plane with boundary and $\phi_x^{-1}(x)$ is a boundary point of U_x. Since the boundary is smooth, we may in this case find a smooth injective map $T_x : C \to U_x$ such that $\phi_x^{-1}(x) \in \{-1\} \times (-1, 1)^{d-1} \subset C$. In this case, we define $V_x = \phi_x \circ T_x(C^o \cup \{-1\} \times (-1, 1)^{d-1})$.

In either case, we have $x \in V_x$ and that V_x is relatively open in $\mathscr{M}$. Since $\mathscr{M}$ is compact, it can be covered by V_{x_j} for finitely many $x_1, ..., x_T$. The maps $\mathscr{P}_j = \mathscr{P} \circ \phi_{x_j} \circ T_{x_j}$ satisfy the conclusion of the lemma. Indeed, since T_{x_i} is smooth and by definition $\mathscr{P} \circ \phi_i \in \text{Lip}_\infty(s, X)$, it is easy to see that $\mathscr{P}_j \in \text{Lip}_\infty(s, X)$. In addition, by construction the images $\mathscr{P}_j(C)$ cover the image $\mathscr{P}(\mathscr{M})$. □

3.1 Polynomial Interpolation Error Bounds

The significance of the smoothness Definition 2 lies in its relationship with approximation by polynomial interpolation. Let use briefly review some basic facts concerning polynomial interpolation.

To set the stage, let $U \subset \mathbb{R}^d$ be an open domain, let Π_k^d denote the space of polynomials of degree at most k in d variables, set $M = \dim \Pi_k^d = \binom{k+d}{d}$, and let $x_1, ..., x_M \subset U$ be a set of points which is unisolvent for the space Π_k^d, i.e., such that no polynomial in Π_k^d vanishes at all of the x_i. In one dimension, it is well known that any k distinct points are unisolvent for the space Π_k^1. It is also possible to explicitly give sets of $\binom{d+k}{k}$ points which are unisolvent for the space Π_k^d for $d > 1$ [14, 15, 43].

In this setting, we can find Lagrange polynomials $l_1, ..., l_M \in \Pi_k^d$ such that $l_i(x_j) = \delta_{ij}$. Given values $y_1, ..., y_M \in \mathbb{R}$ at the points $x_1, ..., x_M$ the unique polynomial in Π_k^d interpolating these values is given by

$$\sum_{i=1}^{M} y_i l_i \in \Pi_k^d. \tag{3.15}$$

The norm of the interpolation map as a map from ℓ_M^∞ to $C(U)$, called the Lesbesgue constant, is given by

$$\Lambda_k^d(U, \{x_1, ..., x_M\}) = \sup_{x \in U} \sum_{i=1}^{M} |l_i(x)|. \tag{3.16}$$

An elementary, yet critical fact about the Lesbesgue constant is its invariance under invertible affine transformations, which is immediate since the space of polynomials Π_k^d is invariant under such transformations.

Lemma 3 *Let A be an invertible linear map on $\mathbb{R}^d$ and $b \in \mathbb{R}^d$ a fixed vector. Then we have*

$$\Lambda_k^d(U, \{x_1, ..., x_M\}) = \Lambda_k^d(AU + b, \{Ax_1 + b, ..., Ax_M + b\}). \tag{3.17}$$

Here $AU + b = \{Ax + b,\ x \in U\}$.

Next, we recall the classical Bramble–Hilbert lemma [6] (see also [60, 61]), which bounds the polynomial interpolation error for functions $f \in \mathrm{Lip}(s, L^\infty(U))$.

Lemma 4 *Let $s = k + \alpha$ with $k \in \mathbb{Z}$ and $\alpha \in (0, 1]$ and suppose that $f \in \mathrm{Lip}(s, L^\infty(U))$ for a convex domain U. Let*

$$f_I(x) = \sum_{i=1}^M f(x_i) l_i(x) \in \Pi_k^d \tag{3.18}$$

denote the polynomial which interpolates the values of f at the points $x_1, ..., x_M$. Then for any $y \in U$ we have

$$|f(y) - f_I(y)| \leq C|f|_{\mathrm{Lip}(s, L^\infty(\Omega))}[\mathrm{diam}(U)]^s \Lambda_k^d(U, \{x_1, ..., x_M\}), \tag{3.19}$$

where the constant C only depends upon k and d.

Proof For each $i = 1, ..., M$ consider the function $r_i(t) = f(y + t(x_i - y))$. Taylor expanding r_i about $t = 0$, we obtain

$$r_i(1) - r_i(0) = \sum_{j=1}^k \frac{1}{j!} r_i^{(j)}(0) + \frac{1}{(k-1)!} \int_0^1 [r_i^{(k)}(t) - r_i^{(k)}(0)] t^{k-1} \mathrm{d}t. \tag{3.20}$$

Next, note that by construction $r_i^{(j)}(0) = D^j f(y) \cdot (x_i - y)^{\otimes j}$ and $r_i^{(k)}(t) = D^k f(y + t(x_i - y)) \cdot (x_i - y)^{\otimes k}$. Plugging this into equation (3.20), we get

$$\begin{aligned} f(x_i) - f(y) &= \sum_{j=1}^k \frac{1}{j!} D^j f(y) \cdot (x_i - y)^{\otimes j} \\ &+ \left(\frac{1}{(k-1)!} \int_0^1 [D^k f(y + t(x_i - y)) - D^k f(y)] t^{k-1} \mathrm{d}t \right) \cdot (x_i - y)^{\otimes k}. \end{aligned} \tag{3.21}$$

We now multiply this equation by $l_i(y)$ and sum over y to obtain

$$\begin{aligned} f_I(y) - f(y) = \sum_{i=1}^M l_i(y) \left(\int_0^1 [D^k f(y + t(x_i - y)) - D^k f(y)] t^{k-1} \mathrm{d}t \right) \\ \cdot (x_i - y)^{\otimes k}. \end{aligned} \tag{3.22}$$

Here we have used the identities

$$\sum_{i=1}^{M} l_i(y) = 1, \quad \sum_{i=1}^{M} l_i(y)(x_i - y)^{\otimes j} = 0, \tag{3.23}$$

for $j = 1, ..., k$. The first of these identities holds since left-hand side is the interpolation of the constant function 1 evaluated at y. The second holds since the left-hand side is the interpolation of the degree j polynomial $p(x) = (x - y)^{\otimes j}$ evaluated at y. Since $j \leq k$, this polynomial is reproduced exactly and so we get $p(y) = 0$.

Since $f \in \mathrm{Lip}(s, L^\infty(U))$ and U is convex so that $y + t(x_i - y) \in U$ for $t \in [0, 1]$, we get

$$|D^k f(y + t(x_i - y)) - D^k f(y)| \leq |f|_{\mathrm{Lip}(s, L^\infty(\Omega))}(t|x_i - y|)^\alpha. \tag{3.24}$$

Since also $|(x_i - y)^{\otimes k}| \leq C(k, d)|x_i - y|^k$, we get

$$\begin{aligned} &\left(\int_0^1 [D^k f(y + t(x_i - y)) - D^k f(y)] t^{k-1} \mathrm{d}t\right) \cdot (x_i - y)^{\otimes k} \\ &\quad \leq C(k, d)|f|_{\mathrm{Lip}(s, L^\infty(\Omega))}|x_i - y|^{k+\alpha} \\ &\quad \leq C(k, d)|f|_{\mathrm{Lip}(s, L^\infty(\Omega))}[\mathrm{diam}(U)]^s. \end{aligned} \tag{3.25}$$

for each $i = 1, ...M$. Plugging this into (3.22), finally get

$$\begin{aligned} |f_I(y) - f(y)| &\leq C(k, d)|f|_{\mathrm{Lip}(s, L^\infty(\Omega))}[\mathrm{diam}(U)]^s \sum_{i=1}^{M} |l_i(y)| \\ &\leq C(k, d)|f|_{\mathrm{Lip}(s, L^\infty(\Omega))}[\mathrm{diam}(U)]^s \Lambda_k^d(U, \{x_1, ..., x_M\}), \end{aligned} \tag{3.26}$$

as desired. □

Given Banach space values $y_1, ..., y_M \in X$ at the points $x_1, ..., x_M$, we can analogously form the interpolating polynomial $P_k : U \to X$ by

$$P_k(x) = \sum_{i=1}^{M} y_i l_i(x). \tag{3.27}$$

The next lemma shows that if the $y_i = \mathscr{F}(x_i)$ are the values of a map $\mathscr{F} \in \mathrm{Lip}_\infty(s, X)$, then the Bramble–Hilbert lemma holds in the Banach space setting.

Lemma 5 *Let $U \subset \mathbb{R}^d$ be a convex domain and suppose that a map $\mathscr{F} : U \to X$ satisfies $\mathscr{F} \in \mathrm{Lip}_\infty(s, X)$ where $s = k + \alpha$ with k an integer and $\alpha \in (0, 1]$. Let*

$$\mathscr{F}_I(x) = \sum_{i=1}^{M} \mathscr{F}(x_i) l_i(x) \tag{3.28}$$

denote the polynomial which interpolates the values of $\mathscr{F}$ *at the points* $x_1, ..., x_M$. *Then for any* $y \in U$ *we have*

$$\|\mathscr{F}(y) - \mathscr{F}_I\|_X \leq C|\mathscr{F}|_{\mathrm{Lip}(s,X)}[\mathrm{diam}(U)]^s \Lambda_k^d(U, \{x_1, ..., x_M\}), \tag{3.29}$$

where the constant C *only depends upon* k *and* d.

Proof The proof is by duality. Let $\xi \in X^*$ and note that by definition the function $f_\xi = \langle \xi, \mathscr{F}(x)\rangle$ satisfies

$$|f_\xi|_{\mathrm{Lip}(s,L^\infty(\Omega))} \leq |\mathscr{F}|_{\mathrm{Lip}(s,X)}\|\xi\|_{X^*}. \tag{3.30}$$

We also have that

$$f_{\xi,I} := \langle \xi, \mathscr{F}_I\rangle = \sum_{i=1}^M \langle \xi, \mathscr{F}(x_i)\rangle l_i(x) = \sum_{i=1}^M f_\xi(x_i) l_i(x) \tag{3.31}$$

is the interpolation of f_ξ at the points $x_1, ..., x_M$. Applying the Bramble–Hilbert lemma 4 to the function f_ξ, we get

$$\begin{aligned} |\langle \xi, \mathscr{F}_I(y) - \mathscr{F}(y)\rangle| &= |f_{\xi,I}(y) \\ &- f_\xi(y)| \leq C|\mathscr{F}|_{\mathrm{Lip}(s,X)}\|\xi\|_{X^*}[\mathrm{diam}(U)]^s \Lambda_k^d(U, \{x_1, ..., x_M\}) \end{aligned} \tag{3.32}$$

where the constant C only depends upon k and d. Since this is true for all $\xi \in X^*$, the result follows. □

3.2 Approximation Rates for Smoothly Parameterized Dictionaries

Next, we give upper bounds on the approximation rates of $B_1(\mathbb{D})$ from sparse convex combinations $\Sigma_{n,M}(\mathbb{D})$ for smoothly parameterized dictionaries. We have the following theorem.

Theorem 2 *Let* $s > 0$ *and* X *be a type-2 Banach space. Suppose that* $\mathscr{M}$ *is a compact* d*-dimensional smooth manifold,* $\mathscr{P} \in \mathrm{Lip}_\infty^{\mathscr{M}}(s, X)$, *and the dictionary* $\mathbb{D} \subset \mathscr{P}(\mathscr{M})$. *Then there exists an* $M > 0$ *such that for* $f \in B_1(\mathbb{D})$ *we have*

$$\inf_{f_n \in \Sigma_{n,M}(\mathbb{D})} \|f - f_n\|_X \lesssim n^{-\frac{1}{2} - \frac{s}{d}}. \tag{3.33}$$

Here both M *and the implied constant depend only upon* s, d, *the parameterization map* $\mathscr{P}$, *and the type-2 constant of the space* X.

We note that although the implied constants here are independent of n, they may be quite large and accurately estimating them would require careful consideration of the structure of the manifold $\mathscr{M}$ and the parameterization map $\mathscr{P}$. The proof of this theorem is a higher-order generalization of the stratified sampling argument [29, 39],

which corresponds to the $k = 0$ case. Before proving this theorem, we note a corollary obtained when applying it to the dictionary $\mathbb{P}_k^d$.

Theorem 3 *Let $k \geq 0$ and $2 \leq p < \infty$. Then there exists an $M = M(p, k, d) > 0$ such that for all $f \in B_1(\mathbb{P}_k^d)$ we have*

$$\inf_{f_n \in \Sigma_{n,M}(\mathbb{P}_k^d)} \|f - f_n\|_{L^p(\Omega)} \lesssim n^{-\frac{1}{2} - \frac{pk+1}{pd}}. \tag{3.34}$$

In the case $p = 2$, we obtain in particular that $\alpha(k, d) \geq \frac{2k+1}{2d}$ in (1.27). We will show in Sect. 4 that this rate is sharp when $p = 2$. However, we expect that this rate can be improved when $2 < p < \infty$ using the techniques of geometric discrepancy theory [2, 41, 42].

Proof This follows immediately from Theorem 2 given the smoothness condition of the map $\mathscr{P}_k^d$ proven in Lemma 1 and the fact that $S^{d-1} \times [c_1, c_2]$ is a compact d-dimensional manifold. □

Proof of Theorem 2 We apply lemma 2 to $\mathscr{P}$ and $\mathscr{M}$ to obtain a collection of maps $\mathscr{P}_j : C := [-1, 1]^d \to X$ such that $\mathbb{D} = \cup_{j=1}^T \mathscr{P}_j(C)$ and $\mathscr{P}_j \in \mathrm{Lip}_\infty(s, X)$. We remark that using cubes here is not strictly necessary, but we do this for convenience since cubes can be easily subdivided in a straightforward manner.

It suffices to prove the result for $\mathbb{D} = \mathbb{D}_j := \mathscr{P}_j(C)$. This follows since $B_1(\mathbb{D}) = \mathrm{conv}(\cup_{j=1}^T B_1(\mathbb{D}_j))$ and given a convex combination $f = \alpha_1 f_1 + \cdots + \alpha_T f_T$ with $f_j \in B_1(\mathbb{D}_j)$ and $\sum_{i=1}^T \alpha_j = 1$, we get

$$\inf_{f_n \in \Sigma_{Tn,M}(\mathbb{D})} \|f - f_n\|_H \leq \sum_{j=1}^T \alpha_j \inf_{f_{n,j} \in \Sigma_{n,M}(\mathbb{D}_j)} \|f_j - f_{n,j}\|_X, \tag{3.35}$$

which easily follows by setting $f_n = \sum_{j=1}^T \alpha_j f_{n,j}$ and noting that $\mathbb{D} = \cup_{j=1}^T \mathbb{D}_j$.

So in what follows we consider $\mathbb{D} = \mathbb{D}_j$, $\mathscr{P} = \mathscr{P}_j$ and $\mathscr{M} = C$. In other words, we assume without loss of generality that $T = 1$ (at the cost of introducing a constant which depends upon T and thus upon $\mathscr{P}$ and $\mathscr{M}$).

Now let $f \in B_1(\mathbb{D})$ and $\delta > 0$. Then there exists a convex combination (with potentially very large $N := N_\delta$)

$$f_\delta = \sum_{i=1}^N a_i d_i, \tag{3.36}$$

with $d_i \in \mathbb{D}$, $\sum |a_i| \leq 1$, and $\|f - f_\delta\|_X < \delta$. Since $\mathbb{D} = \mathscr{P}(C)$, each $d_i = \mathscr{P}(z_i)$ for some $z_i \in C$, so we get

$$f_\delta = \sum_{i=1}^N a_i \mathscr{P}(z_i). \tag{3.37}$$

We remark that in what follows all implied constants will be independent of n and δ.

Let $n \geq 1$ be given and subdivide the cube C into n sub-cubes $C_1, ..., C_n$ such that each C_r has diameter $O(n^{-\frac{1}{d}})$. This can easily be done by considering a uniform subdivision in each direction. (This is also why we chose to use cubes in this construction.)

We proceed to approximate the map $\mathscr{P}$ by a piecewise polynomial on the sub-cubes $C_1, ..., C_n$. To this end, let $M = \binom{d+k}{k}$ and $x_1, ..., x_M \in C$ a set of points which is unisolvent for the space of polynomials of degree at most k in d variables and let $l_1, ..., l_M$ be the associated Lagrange interpolation polynomials, as discussed in Sect. 3.1. Here the integer k is determined by $s = k + \alpha$ with $\alpha \in (0, 1]$.

For each cube C_r, denote by $x_1^r, ..., x_M^r$ and $l_1^r, ..., l_M^r$ the image of the interpolation points $x_1, ..., x_M$ on the cube C_r and their associated Lagrange polynomials. We rewrite f_δ as

$$f_\delta = \sum_{r=1}^{n} \sum_{z_i \in C_r} \mathscr{P}(z_i) = \sum_{r=1}^{n} \sum_{z_i \in C_r} a_i \mathscr{P}_{r,I}(z_i) + \sum_{l=1}^{n} \sum_{z_i \in C_r} a_i \mathscr{E}_r(z_i), \tag{3.38}$$

where the polynomial interpolation in the cube C_r is given by

$$\mathscr{P}_{r,I}(z) = \sum_{i=1}^{M} \mathscr{P}(x_i^r) l_i^r(z), \tag{3.39}$$

and the error in the approximation is given by

$$\mathscr{E}_r(z) = \mathscr{P}(z) - \mathscr{P}_{r,I}(z). \tag{3.40}$$

We use the Banach space Bramble–Hilbert Lemma 5 and Maurey's sampling argument (Theorem 1) to bound the second term in (3.38). We apply Theorem 1 with the dictionary $\mathbb{D}_{\mathscr{E}} = \{\mathscr{E}_r(z_i),\ z_i \in C_r\}$ to the term

$$\sum_{l=1}^{n} \sum_{z_i \in C_r} a_i \mathscr{E}_r(z_i) \in B_1(\mathbb{D}_{\mathscr{E}}). \tag{3.41}$$

This yields the existence of an n-term convex combination

$$f_n' = \frac{1}{n} \sum_{s=1}^{n} \mathscr{E}_{r_s}(z_{i_s}) \tag{3.42}$$

satisfying

$$\left\| f_n' - \sum_{l=1}^{n} \sum_{z_i \in C_r} a_i \mathscr{E}_r(z_i) \right\|_X \leq C K_{\mathbb{D}_{\mathscr{E}}} n^{-\frac{1}{2}}, \tag{3.43}$$

where C only depends upon the space X. Lemma 5 implies that

$$\begin{aligned} K_{\mathbb{D}_{\mathscr{E}}} &\leq \sup_{z \in C_r} \|\mathscr{E}_r(z)\|_X \\ &\leq C(k,d)|\mathscr{P}|_{\mathrm{Lip}(s,X)}[\mathrm{diam}(C_r)]^s \Lambda_k^d(C_r, x_1^r, ..., x_M^r) = Cn^{-\frac{s}{d}}, \end{aligned} \tag{3.44}$$

where C is independent of n. This holds since by Lemma 3 the Lebesgue constant satisfies

$$\Lambda_k^d(C_r, x_1^r, ..., x_M^r) = \Lambda_k^d(C, x_1, ..., x_M)$$

and is thus independent of n. Setting

$$f_n = \sum_{r=1}^{n} \sum_{z_i \in C_r} a_i \mathscr{P}_{r,I}(z_i) + f_n', \tag{3.45}$$

we thus obtain

$$\|f_n - f_\delta\|_X = \left\| f_n' - \sum_{l=1}^{n} \sum_{z_i \in C_r} a_i \mathscr{E}_r(z_i) \right\|_X \lesssim n^{-\frac{1}{2}-\frac{s}{d}}, \tag{3.46}$$

where the implied constant is independent of n. Finally, we observe that

$$\begin{aligned} f_n = \sum_{r=1}^{n} \sum_{z_i \in C_r} a_i \mathscr{P}_{r,I}(z_i) + \frac{1}{n} \sum_{s=1}^{n} (\mathscr{P}(z_{i_s}) \\ - \mathscr{P}_{r_s,I}(z_{i_s})) \in \Sigma_{(M+1)n,2K+1}(\mathbb{D}), \end{aligned} \tag{3.47}$$

for $K := \sup_{x \in C_r} \sum_{i=1}^{M} |l_i^r(x)| = \Lambda_k^d(C_r, x_1^r, ..., x_M^r) = \Lambda_k^d(C, x_1, ..., x_M)$. This holds since the interpolating polynomial $\mathscr{P}_{r,I}(z_i)$ only involves evaluations of the map $\mathscr{P}$ at the fixed interpolation points $x_1^r, ..., x_M^r$ in the cube C_r and the coefficients of those evaluations are bounded in ℓ^1 by the Lebesgue constant $\Lambda_k^d(C, x_1, ..., x_M)$. Since there are n cubes $C_1, ..., C_n$ which each contain M interpolation points, there are Mn interpolation points in total, so we have

$$\sum_{r=1}^{n} \sum_{z_i \in C_r} a_i \mathscr{P}_{r,I}(z_i) - \frac{1}{n} \sum_{s=1}^{n} \mathscr{P}_{r_s,I}(z_{i_s}) \in \Sigma_{Mn,2K}(\mathbb{D}), \tag{3.48}$$

from which (3.47) follows. Since $\delta > 0$ was arbitrary, this completes the proof. □

3.3 Metric Entropy Bounds for Smoothly Parameterized Dictionaries

Next, we bound the metric entropy of $B_1(\mathbb{D})$ for smoothly parameterized dictionaries $\mathbb{D}$. We first observe that the approximation rate proven in Theorem 2 implies a bound

on the metric entropy via Theorem 10 (see also the proofs of Theorem 4 in [39] and in [29]). Specifically, under the assumptions of Theorem 2 we get

$$\varepsilon_{n\log n}(B_1(\mathbb{D})) \lesssim n^{-\frac{1}{2}-\frac{s}{d}}. \tag{3.49}$$

The main result in this section is that the logarithmic factor in (3.49) can be removed.

Theorem 4 *Let $s > 0$ and X be a type-2 Banach space. Suppose that $\mathscr{M}$ is a compact d-dimensional smooth manifold, $\mathscr{P} \in \mathrm{Lip}^{\mathscr{M}}_{\infty}(s, X)$, and the dictionary $\mathbb{D} \subset \mathscr{P}(\mathscr{M})$. Then*

$$\varepsilon_n(B_1(\mathbb{D}))_X \lesssim n^{-\frac{1}{2}-\frac{s}{d}}. \tag{3.50}$$

Here the implied constant depends only upon s, d, the parameterization map $\mathscr{P}$, and the type-2 constant of the space X.

We note that combined with the bounds in Lemma 1, Theorem 4 implies the following bound for the variation space corresponding to shallow ReLU^k networks

$$\varepsilon_n(B_1(\mathbb{P}_k^d))_{L^2(\Omega)} \lesssim n^{-\frac{1}{2}-\frac{2k+1}{2d}}. \tag{3.51}$$

In Sect. 4, we will show that this rate is sharp up to a constant factor.

To prove Theorem 4 we will use the following two lemmas. The first is a triangle inequality for the entropy numbers.

Lemma 6 (see [37] Section 15.7) *Let $A, B \subset X$. Then for any $0 \le m \le n$*

$$\varepsilon_n(A + B) \le \varepsilon_m(A) + \varepsilon_{n-m}(B). \tag{3.52}$$

Proof If balls of radius $\varepsilon_m(S)$ around $s_1, \ldots, s_{2^m} \in X$ cover A and balls of radius $\varepsilon_{n-m}(T)$ around $t_1, \ldots, t_{2^{n-m}} \in X$ cover B, then balls of radius $\varepsilon_m(S) + \varepsilon_{n-m}(T)$ around the 2^n points $s_i + t_j$ cover $A + B$. If the infimum in the entropy is not achieved, then a simple limiting argument can be used to complete the proof. □

The second, due to Carl (Proposition 1 in [8]), gives a bound on the metric entropy of the convex hull of a finite dictionary $\mathbb{D} \subset X$.

Lemma 7 *Let X be a type-2 Banach space and $\mathbb{D} \subset X$ a dictionary with n elements, i.e., $\mathbb{D} = \{d_1, \ldots, d_n\}$. Set $K_{\mathbb{D}} = \max_{i=1,\ldots,n} \|d_i\|_X$. Then*

$$\varepsilon_m(B_1(\mathbb{D})) \lesssim \begin{cases} K_{\mathbb{D}} & m = 0 \\ \sqrt{1 + \log \frac{n}{m}}\, m^{-\frac{1}{2}} K_{\mathbb{D}} & 1 \le m \le n \\ 2^{-\frac{m}{n}} n^{-\frac{1}{2}} K_{\mathbb{D}} & m > n, \end{cases} \tag{3.53}$$

where the implied constant only depends upon the type-2 constant of X.

Proof of Theorem 4 As in the proof of Theorem 2, we begin by reducing to the case where $\mathscr{M} = C := [-1, 1]^d$ is the cube. Using Lemma 2, we see that there exists an integer T and a collection of maps $\mathscr{P}_j : C \to X$ such that $\mathbb{D} \subset \cup_{j=1}^T \mathscr{P}_j(C)$ and $\mathscr{P}_j \in \mathrm{Lip}_\infty(s, X)$. Again, using the cube here is not strictly necessary, but simplifies the argument somewhat.

Now $B_1(\mathbb{D}) \subset \sum_{j=1}^T B_1(\mathscr{P}_j(C))$ and applying Lemma 6 implies that

$$\varepsilon_{Tn}(B_1(\mathbb{D})) \leq \sum_{j=1}^T \varepsilon_n(B_1(\mathscr{P}_j(C))). \tag{3.54}$$

Thus, at the cost of a constant factor it suffices to prove Theorem 3.82 for $\mathbb{D} = \mathscr{P}_j(C)$ for each j. So we set $\mathscr{P} = \mathscr{P}_j$ and consider the case where $\mathscr{M} = C$ and $\mathbb{D} = \mathscr{P}(C)$.

Let $s = k+\alpha$ with k and integer and $\alpha \in (0, 1]$. Set $M = \binom{d+k}{k}$ and choose M points $x_1, ..., x_M \in C$ which are unisolvent for the space of polynomials of degree at most k in d variables and let $l_1, ..., l_M$ be the associated Lagrange interpolation polynomials. For each integer $i \geq 0$, we subdivide the cube into 2^{di} sub-cubes $C_1, ..., C_{2^{di}}$ of side length 2^{-i}. We let $\mathscr{P}_i$ denote the piecewise degree k interpolation of the map $\mathscr{P}$ on the sub-cubes C_r. Specifically, for $z \in C$, we denote by $r_i(z) \in \{1, ..., 2^{di}\}$ the index such that $z \in C_{r_i(z)}$ (for points on the boundary of a sub-cube where this index may not be unique we simply choose one) and define

$$\mathscr{P}_i(z) = \sum_{j=1}^M l_j^{r_i(z)}(z)\mathscr{P}(x_j^{r_i(z)}) \tag{3.55}$$

where $l_j^{r_i(z)}$ and $x_j^{r_i(z)}$ are the images of the Lagrange polynomials and interpolation points on the sub-cube $C_{r_i(z)}$ containing z at discretization level i.

Next, we define dictionaries $\mathbb{D}_i$ for $i \geq 0$ by

$$\mathbb{D}_i = \{\mathscr{P}_i(z) - \mathscr{P}_{i-1}(z),\ z \in C\}. \tag{3.56}$$

Here we set $\mathscr{P}_{-1}(z) = 0$. Note that

$$B_1(\mathbb{D}) \subset \overline{\sum_{i=1}^\infty B_1(\mathbb{D}_i)}. \tag{3.57}$$

Indeed, by definition $B_1(\mathbb{D})$ is the closure of elements of the form

$$\sum_{l=1}^N a_l \mathscr{P}(z_l) = \sum_{i=1}^\infty \sum_{l=1}^N a_l(\mathscr{P}_i(z_l) - \mathscr{P}_{i-1}(z_l)) \in \sum_{i=1}^\infty B_1(\mathbb{D}_i), \tag{3.58}$$

where $\sum_{i=1}^{N} |a_i| \leq 1$. Using Lemma 6 inductively, this implies that for any sequence of integers $n_1, n_2, \ldots \geq 0$ such that $\sum_{i=1}^{\infty} n_i = n$ we have the bound

$$\varepsilon_n(B_1(\mathbb{D})) \leq \sum_{i=1}^{\infty} \varepsilon_{n_i}(B_1(\mathbb{D}_i)). \tag{3.59}$$

Note that since the entropy is decreasing this also holds if $\sum_{i=1}^{\infty} n_i \leq n$.

The next step is to bound $\varepsilon_{n_i}(B_1(\mathbb{D}_i))$. For this, we note the following composition property of interpolation

$$\mathscr{P}_{i-1}(z) = \sum_{j=1}^{M} l_j^{r_i(z)}(z)\mathscr{P}_{i-1}(x_j^{r_i(z)}), \tag{3.60}$$

which follows since the function $\mathscr{P}_{i-1}$ equals its interpolation on the finer grid at level i. Thus, the dictionary $\mathbb{D}_i$ can be rewritten as

$$\mathbb{D}_i = \left\{ \sum_{j=1}^{M} l_j^{r_i(z)}(z)[x_j^{r_i(z)} - \mathscr{P}_{i-1}(x_j^{r_i(z)})],\ z \in C \right\}, \tag{3.61}$$

from which it follows that

$$B_1(\mathbb{D}_i) \subset \left(\max_{z \in C} \sum_{j=1}^{M} |l_j^{r_i(z)}(z)| \right) B_1(\overline{\mathbb{D}}_i) = \Lambda_k^d(C, \{x_1, \ldots, x_M\}) B_1(\overline{\mathbb{D}}_i), \tag{3.62}$$

where $\overline{\mathbb{D}}_i$ is the finite dictionary given by

$$\overline{\mathbb{D}}_i = \{x_j^r - \mathscr{P}_{i-1}(x_j^r),\ j = 1, \ldots, M,\ r = 1, \ldots, 2^{di}\}.$$

We note that the number of elements in $\overline{\mathbb{D}}_i$ is $M2^{di}$ and the Banach space Bramble–Hilbert Lemma 5 implies that $K_{\overline{\mathbb{D}}_i} \lesssim 2^{-si}$. We now use Lemma 7 to get

$$\begin{aligned} \varepsilon_m(B_1(\mathbb{D}_i)) &\leq \Lambda_k^d(C, \{x_1, \ldots, x_M\}) \varepsilon_m(B_1(\overline{\mathbb{D}}_i)) \\ &\lesssim \begin{cases} 2^{-si} & m = 0 \\ 2^{-si} m^{-\frac{1}{2}} \sqrt{1 - \log m + di + \log M} & 1 \leq m \leq M2^{di} \\ 2^{-\left(s + \frac{\mathrm{d}}{2}\right) i} 2^{-\frac{m}{M2^{di}}} & m > M2^{di}, \end{cases} \end{aligned} \tag{3.63}$$

where the implied constant is independent of i and m (specifically it will only depend upon k, d, the parameterization map $\mathscr{P}$, and the type-2 constant of X).

Finally, the proof is completed by substituting (3.63) into (3.59) and optimizing over the choice of n_i. This is a somewhat standard, but involved calculation (see [3, 9, 12, 40], for instance).

It suffices to prove Theorem 4 for n of the form $n = K2^{rd}$ for a fixed integer K which will be determined later. This follows since the entropy is a decreasing function and our bound is polynomial in n, so extending to all values of n will only increase the implied constant. For such a value of n, we wish to show that

$$\varepsilon_n(B_1(\mathbb{D})) \lesssim 2^{-\left(s+\frac{d}{2}\right)r}, \tag{3.64}$$

where the implied constant is independent of r. Let $\delta > 0$ and choose the n_i as

$$n_i = \begin{cases} M\left(s+\frac{d}{2}+\delta\right)(r-i+1)2^{di} & 0 \le i < r \\ M2^{rd-\delta(i-r)} & r \le i < \left(1+\frac{d}{2s}\right)r \\ 0 & i \ge \left(1+\frac{d}{2s}\right)r. \end{cases} \tag{3.65}$$

For simplicity, we allow the n_i to not necessarily be integers for now. We must show two things. First, that

$$\sum_{i=0}^{\infty} n_i \lesssim 2^{rd}, \tag{3.66}$$

which will ensure that $\sum_{i=1}^{\infty} n_i \le n = K2^{rd}$ for sufficiently large K. Second, we must use the bound (3.63) to show that

$$\sum_{i=0}^{\infty} \varepsilon_{n_i}(B_1(\mathbb{D}_i)) \lesssim 2^{-\left(s+\frac{d}{2}\right)r}. \tag{3.67}$$

First, we calculate

$$\begin{aligned} \sum_{i=1}^{\infty} n_i &= M\left(\left(s+\frac{d}{2}+\delta\right)\sum_{i=1}^{r-1}(r-i+1)2^{di} + 2^{rd}\sum_{i=r}^{\left(1+\frac{d}{2s}\right)r} 2^{-\delta(i-r)}\right) \\ &\le M2^{rd}\left(\left(s+\frac{d}{2}+\delta\right)\sum_{i=1}^{\infty}(i+1)2^{-di} + \sum_{i=0}^{\infty}2^{-\delta i}\right) \\ &\lesssim 2^{rd}. \end{aligned} \tag{3.68}$$

Next, we note that if $i < r$, then $(s+\frac{d}{2}+\delta)(r-i+1) > d \ge 1$ and so for the first r indices $0 \le i < r$, the last branch of (3.63) is taken. This gives

$$\begin{aligned} \sum_{i=0}^{r-1} \varepsilon_{n_i}(B_1(\mathbb{D}_i)) &\lesssim 2^{\left(s+\frac{d}{2}\right)(r+1)}\sum_{i=0}^{r-1} 2^{-\delta(r-i+1)} \le 2^{\left(s+\frac{d}{2}\right)(r+1)}\sum_{i=1}^{\infty}2^{-\delta i} \\ &\lesssim 2^{\left(s+\frac{d}{2}\right)r}. \end{aligned} \tag{3.69}$$

When $r \le i < \left(1+\frac{d}{2s}\right) r$, the middle branch in (3.63) is taken, which gives

$$\begin{aligned}\sum_{i=r}^{\left(1+\frac{d}{2s}\right)r} \varepsilon_{n_i}(B_1(\mathbb{D}_i)) &\lesssim 2^{-\left(s+\frac{d}{2}\right)r} \sum_{i=r}^{\left(1+\frac{d}{2s}\right)r} 2^{-(s-\delta)(i-r)}\sqrt{1+(d-\delta)(i-r)} \\ &\lesssim 2^{-\left(s+\frac{d}{2}\right)r} \sum_{i=0}^{\infty} 2^{-(s-\delta)i}\sqrt{1+i} \lesssim 2^{-\left(s+\frac{d}{2}\right)r},\end{aligned} \tag{3.70}$$

as long as δ is chosen to be less than s. If $i \ge \left(1+\frac{d}{2s}\right) r$, then the first branch of (3.63) is taken and we calculate

$$\sum_{i=\left(1+\frac{d}{2s}\right)r}^{\infty} \varepsilon_{n_i}(B_1(\mathbb{D}_i)) \lesssim 2^{-s\left(1+\frac{d}{2s}\right)r} \sum_{i=0}^{\infty} 2^{-si} \lesssim 2^{-\left(s+\frac{d}{2}\right)r}. \tag{3.71}$$

Finally, since the n_i must be chosen to be integers, we replace n_i by $\lceil n_i \rceil$. Since the right-hand side of (3.63) is a decreasing function of m, this can only reduce our bound on $\sum_{i=0}^{\infty} \varepsilon_{n_i}(B_1(\mathbb{D}_i))$. Further, since at most $\left(1+\frac{d}{2}\right) r$ of the n_i's are nonzero, the sum $\sum_{i=0}^{\infty} n_i$ can increase by at most $\left(1+\frac{d}{2}\right) r \le 2^{rd}$. Thus, after making this change conditions (3.66) and (3.67) will still be satisfied, which completes the proof. □

3.4 Kolmogorov *n*-Width Bounds for Smoothly Parameterized Dictionaries

Next, we bound the Kolmogorov n-widths of $B_1(\mathbb{D})$ for smoothly parameterized dictionaries $\mathbb{D}$. We have the following theorem.

Theorem 5 *For $s > 0$ and X a Banach space, suppose that $\mathscr{M}$ is a compact d-dimensional manifold, $\mathscr{P} : \mathscr{M} \to X$ is of smoothness class s, i.e., $\mathscr{P} \in \mathrm{Lip}_{\infty}^{\mathscr{M}}(s, X)$, and $\mathbb{D} \subset \mathscr{P}(\mathscr{M})$. Then we have the bound*

$$d_n(B_1(\mathbb{D}))_X \lesssim n^{-\frac{s}{d}}. \tag{3.72}$$

Here the implied constant depends only upon s, d, and the parameterization map $\mathscr{P}$.

As a corollary, we obtain an upper bound on the Kolmogorov widths of $B_1(\mathbb{P}_k^d)$ in $L^p(\Omega)$ of $O(n^{-\frac{pk+1}{pd}})$ for $1 \le p < \infty$.

Proof For any subspace $V \in X$, the distance map $d(x, V) = \inf_{y \in V} \|x - y\|_X$ is a convex function of x. This means that the Kolmogorov n-widths are invariant under taking convex hulls, i.e.,

$$d_n(B_1(\mathbb{D}))_X = d_n(\mathbb{D})_X. \tag{3.73}$$

Thus, it suffices to bound the n-widths of the dictionary $\mathbb{D}$.

We use Lemma 2 to obtain a collection of maps $\mathscr{P}_1, ..., \mathscr{P}_T : C \to X$, where $C = [0, 1]^d$ is the unit cube, such that $\mathbb{D} = \cup_{i=1}^T \mathscr{P}_i(C)$. Set $\mathbb{D}_i = \mathscr{P}_i(C)$.

Now let $n \geq 1$ be a fixed integer. We proceed to subdivide the cube C into n sub-cubes $C_1, ..., C_n$ of diameter $O(n^{-\frac{1}{d}})$. Further, let $s = k + \alpha$ with k an integer and $\alpha \in (0, 1]$. Let $M = \binom{d+k}{k}$ and choose interpolation points $x_1, ..., x_M$ which are unisolvent for the space of polynomials of degree at most k. Denote by $x_1^r, ..., x_M^r$ the images of these interpolation points in the cube C_r and by $l_1^r, ..., l_M^r$ the corresponding Lagrange polynomials. Consider the space

$$V_n = \text{span}\{\mathscr{P}_i(x_j^r),\ i = 1, ..., T,\ j = 1, ..., M,\ r = 1, ..., n\}, \tag{3.74}$$

which satisfies $\dim(V_n) \leq TMn$. Given any $d \in \mathbb{D}$, by definition $d = \mathscr{P}_i(z)$ for some $1 \leq i \leq T$ and $z \in C_r$ for some $1 \leq r \leq n$. Consider the interpolated value

$$\mathscr{P}_{i,r,I}(z) = \sum_{i=1}^{M} \mathscr{P}_i(x_i^r) l_i^r(z) \in V_n. \tag{3.75}$$

The Banach space Bramble–Hilbert Lemma 5 implies that

$$\begin{aligned} \|d - \mathscr{P}_{i,r,I}(z)\|_X &\leq C(k, d) |\mathscr{P}|_{\text{Lip}(s,X)} [\text{diam}(C_r)]^s \Lambda_k^d(C_r, x_1^r, ..., x_M^r) \\ &\lesssim n^{-\frac{s}{d}}, \end{aligned} \tag{3.76}$$

where the implied constant is independent of n since by Lemma 3 the Lebesgue constant $\Lambda_k^d(C_r, x_1^r, ..., x_M^r) = \Lambda_k^d(C, x_1, ..., x_M)$ is independent of n. This means that

$$d_{TMn}(\mathbb{D})_X \leq \sup_{d \in \mathbb{D})} \inf_{y \in V_n} \|d - y\|_X \lesssim n^{-\frac{s}{d}}, \tag{3.77}$$

which completes the proof since T and M are fixed constants independent of n. □

3.5 Gelfand Numbers of Smoothly Parameterized Dictionaries

Finally, we consider the Gelfand numbers of smoothly parameterized dictionaries $\mathbb{D}$. Denote by $\ell^1(\mathbb{D})$ the Banach space of absolutely summable functions on the dictionary $\mathbb{D}$, i.e.,

$$\ell^1(\mathbb{D}) = \{f : \mathbb{D} \to \mathbb{R},\ \|f\|_{\ell^1(\mathbb{D})} < \infty\}, \tag{3.78}$$

where the norm $\|f\|_{\ell^1(\mathbb{D})}$ is given by

$$\|f\|_{\ell^1(\mathbb{D})} = \sup_{\mathbb{D}_n \subset \mathbb{D}} \sum_{d \in \mathbb{D}_n} |f(d)|. \tag{3.79}$$

Here the supremum $\mathbb{D}_n$ is over all finite subsets of the dictionary $\mathbb{D}$. Define the evaluation map $\mathscr{T}_{\mathbb{D}} : \ell^1(\mathbb{D}) \to X$ by

$$\mathscr{T}_{\mathbb{D}} = \sum_{d \in \mathbb{D}} f(d)d. \tag{3.80}$$

It is easy to see that if $\|f\|_{\ell^1(\mathbb{D})} < \infty$ and the dictionary $\mathbb{D}$ is uniformly bounded, then f is nonzero for at most countably many dictionary elements d and the sum in (3.80) converges absolutely.

For an operator $T : X \to Y$ between two Banach spaces X and Y, we define the Gelfand numbers of the operator T by

$$c_n(T) = \inf_{U_n \subset X} \|T|_{U_n}\|, \tag{3.81}$$

where the infimum is taken over all closed subspaces of codimension n (see [47], Section 11.5). The Gelfand numbers of the convex hull of a dictionary $\mathbb{D}$ are defined to be $c_n(\mathscr{T}_{\mathbb{D}})$ [10, 12, 13].

We have the following result, which generalizes the results from [10, 12] to the case to smoothly parameterized dictionaries.

Theorem 6 *Let $s > 0$ and X a Hilbert space. Suppose that $\mathscr{M}$ is a compact d-dimensional smooth manifold, $\mathscr{P} \in \mathrm{Lip}^{\mathscr{M}}_{\infty}(s, X)$, and the dictionary $\mathbb{D} \subset \mathscr{P}(\mathscr{M})$. Then*

$$c_n(\mathscr{T}_{\mathbb{D}}) \lesssim n^{-\frac{1}{2}-\frac{s}{d}}, \tag{3.82}$$

where the implied constants are independent of n.

The proof of Theorem 6 is analogous to the proof of the entropy bound Theorem 4, and for the sake of brevity, we leave these details to the reader. The main difference is that Lemmas 6 and 7 are replaced by the following three results concerning Gelfand numbers of operators. The last result, Theorem 7 is a rather deep theorem of Carl and Pajor [13].

Lemma 8 (Theorem 11.8.2 in [47]) *Let $S, T : X \to Y$. Then for any $0 \le m \le n$ we have*

$$c_n(S + T) \le c_m(S) + c_{n-m}(T). \tag{3.83}$$

Proof Let $U_m, U_{n-m} \subset X$ be codimension m and $n - m$ subspaces of X, respectively, such that

$$\|S|_{U_m}\| \le c_m(S), \ \ \|T|_{U_{n-m}}\| \le c_{n-m}(T). \tag{3.84}$$

If the infimum in the definition of the Gelfand numbers is not achieved, a standard limiting argument can be used here. Set $U_n = U_m \cap U_{n-m}$, which is a subspace of

codimension at most n. Then we have

$$\begin{aligned}\|(S+T)|_{U_n}\| &\leq \|S|_{U_n}\| + \|T|_{U_n}\| \leq \|S|_{U_m}\| \\ &+ \|T|_{U_{n-m}}\| = c_m(S) + c_{n-m}(T). \end{aligned} \tag{3.85}$$

□

Lemma 9 *Let $S : X \to Y$ and $T : Y \to Z$. Then we have*

$$c_n(ST) = \|S\| c_n(T). \tag{3.86}$$

Proof Let $U_n \subset Y$ be a subspace of codimension n. Then $V_n := S^{-1}(U_n) \subset X$ is a subspace of codimension at most n. Then $\|ST|_{V_n}\| \leq \|S\|\|T_{U_n}\|$ and the result follows. □

Theorem 7 (Theorem 2.2 in [13]) *Let $T : \ell_1^n \to H$ be a bounded linear operator where H is a Hilbert space. Then we have*

$$c_m(T) \lesssim \begin{cases} \|T\| & m = 0 \\ \sqrt{1 + \log \frac{n}{m}} m^{-\frac{1}{2}} \|T\| & 1 \leq m < n \\ 0 & m \geq n. \end{cases} \tag{3.87}$$

Note the implied constant here is absolute.

Given the bound in Theorem 6, a natural question is how the Gelfand numbers of the dictionary $\mathbb{D}$ are related to the Gelfand widths $d^n(B_1(\mathbb{D}))$ of the convex hull of $\mathbb{D}$, which measure how efficiently functions from $B_1(\mathbb{D})$ can be recovered from linear measurements. By definition,

$$B_1(\mathbb{D}) = \overline{\mathscr{T}_{\mathbb{D}}(B_{\ell^1(\mathbb{D})})}, \tag{3.88}$$

where $B_{\ell^1(\mathbb{D})}$ denote the unit ball in $\ell^1(\mathbb{D})$. Thus, this question is a special case of the general question of how the Gelfand numbers $c_n(T)$ are related to the Gelfand widths $d^n(T(B_X))$ for a general operator $T : X \to Y$ between two Banach spaces X and Y, where B_X is the unit ball of X.

Expanding out the definitions, we have

$$c_n(T) = \inf_{\xi_1,\ldots,\xi_n \in X^*} \sup\{\|T(x)\|_Y : x \in B_X,\ \xi_i(x) = 0,\ i = 1, \ldots, n\}, \tag{3.89}$$

and on the other hand,

$$\begin{aligned} d^n(T(B_X)) = \inf_{\xi_1,\ldots,\xi_n \in Y^*} &\sup\{\|T(x)\|_y : x \in B_X, \\ &\xi_i(T(x)) = 0,\ i = 1, \ldots, n\}. \end{aligned} \tag{3.90}$$

Since $\xi_i(T(x)) = (T^*\xi_i)(x)$ and $T^*\xi_i \in X^*$, we see that the infimum in (3.89) is over a larger set, so that $c_n(T) \leq d^n(B_X)$. However, if T is not injective, the map T^* will not be surjective and the infimum in (3.89) is over a strictly larger set. In such a situation it is possible that strict inequality holds [26, 50], i.e., that $c_n(T) < d^n(B_X)$.

Equality of the Gelfand numbers and widths has been established under certain conditions on the operator T, see [26], for instance. However, these conditions all suppose the injectivity of the operator T. In fact, when T is not injective the typical situation is that $c_n(T) < d^n(B_X)$. To illustrate this, we give the following example of a small finite dictionary $\mathbb{D}$ for which $c_n(\mathscr{T}_\mathbb{D}) < d^n(B_1(\mathbb{D}))$. This shows that in general there is not much hope to bound the Gelfand widths $d^n(\mathbb{D})$ using Theorem 6 and leaves open the problem of developing techniques for bounding $d^n(B_1(\mathbb{D}))$ in the case where the evaluation map $\mathscr{T}_\mathbb{D}$ is not injective.

Proposition 1 *Consider the following dictionary* $\mathbb{D} \subset \mathbb{R}^3$

$$\mathbb{D} = \left\{ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1/2 \\ \sqrt{3}/2 \\ 0 \end{pmatrix}, \begin{pmatrix} -1/2 \\ \sqrt{3}/2 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ \sqrt{3} \end{pmatrix} \right\}. \tag{3.91}$$

Then $c_1(\mathscr{T}_\mathbb{D}) < d^1(B_1(\mathbb{D}))$.

Proof Note that for this dictionary the map $\mathscr{T}_\mathbb{D} : \ell_1^4 \to \ell_2^3$ is given by the following matrix

$$\begin{pmatrix} 1 & 1/2 & -1/2 & 0 \\ 0 & \sqrt{3}/2 & \sqrt{3}/2 & 0 \\ 0 & 0 & 0 & \sqrt{3} \end{pmatrix}. \tag{3.92}$$

Consider the subspace $U_1 \subset \ell_1^4$ defined by $\xi \cdot x = 0$ where

$$\xi = \begin{pmatrix} 1 \\ -1 \\ 1 \\ 3 \end{pmatrix}. \tag{3.93}$$

In order to calculate $\|\mathscr{T}_\mathbb{D}|_{U_1}\|$, we determine the extreme points of the intersection

$$U_1 \cap B_{\ell_1^4} = \{x \in \mathbb{R}^4,\ \xi \cdot x = 0,\ |\xi|_1 \leq 1\}. \tag{3.94}$$

The unit ball of ℓ_1^4 is the convex hull of $\{\pm e_1, ..., \pm e_4\}$ and the extreme points of $U_1 \cap B_{\ell_1^4}$ must be a linear combinations of at most two of these vectors. Using the form of ξ, these extreme points are

$$E := \left\{ \frac{1}{2}(\pm e_1 \pm e_2), \frac{1}{2}(\pm e_1 \mp e_3), \pm\frac{3}{4}e_1 \mp \frac{1}{4}e_4, \frac{1}{2}(\pm e_2 \pm e_3), \right.$$

$$\pm\frac{3}{4}e_2 \pm \frac{1}{4}e_4, \pm\frac{3}{4}e_3 \mp \frac{1}{4}e_4\Bigg\}. \tag{3.95}$$

The norm $\|\mathscr{T}_{\mathbb{D}}|_{U_1}\|$ is equal to the maximum value of $\|\mathscr{T}_{\mathbb{D}}(x)\|_2$ for $e \in E$ and a straightforward calculation yields

$$\|\mathscr{T}_{\mathbb{D}}|_{U_1}\| = \frac{\sqrt{3}}{2}. \tag{3.96}$$

Thus, we get $c_1(\mathscr{T}_{\mathbb{D}}) \leq \sqrt{3}/2$.

Next, we will show that $d^1(B_1(\mathbb{D})) > \sqrt{3}/2$. Note that the shape of $B_1(\mathbb{D})$ is a hexagonal bipyramid. Suppose that there exists a plane $U \subset \mathbb{R}^3$ such that $U \cap B_1(\mathbb{D})$ is contained in a ball of radius $\sqrt{3}/2$. Consider $U \cap \operatorname{span}(e_1, e_2)$. This intersection cannot contain points of the hexagonal base of $B_1(\mathbb{D})$ which are longer than $\sqrt{3}/2$. Thus, $U \cap \operatorname{span}(e_1, e_2)$ must be a line connecting the midpoints of two opposite side of this hexagon. So we can assume without loss of generality that

$$U \cap \operatorname{span}(e_1, e_2) = \operatorname{span}\left\{\begin{pmatrix}\sqrt{3}/2\\ -1/2\\ 0\end{pmatrix}\right\}. \tag{3.97}$$

This in turn implies that U must intersect the line segments connecting

$$\begin{pmatrix}1\\0\\0\end{pmatrix}, \begin{pmatrix}1/2\\ \sqrt{3}/2\\ 0\end{pmatrix}, \begin{pmatrix}-1/2\\ \sqrt{3}/2\\ 0\end{pmatrix} \text{ to either } \begin{pmatrix}0\\0\\ \sqrt{3}\end{pmatrix} \text{ or } \begin{pmatrix}0\\0\\ -\sqrt{3}\end{pmatrix}. \tag{3.98}$$

By reflecting, we may assume without loss of generality that the former occurs. However, each of these line segments contains a unique point with length at most $\sqrt{3}/2$, which are given by

$$\begin{pmatrix}3/4\\0\\ \sqrt{4}/4\end{pmatrix}, \begin{pmatrix}3/8\\ 3\sqrt{3}/8\\ \sqrt{4}/4\end{pmatrix}, \begin{pmatrix}-3/8\\ 3\sqrt{3}/8\\ \sqrt{4}/4\end{pmatrix}, \tag{3.99}$$

respectively. Since $B_1(\mathbb{D})$ contains each of the line segments in (3.98) and $U \cap B_1(\mathbb{D})$ is contained in a ball of radius $\sqrt{3}/2$, this implies that U must contain each of the points in (3.99). Finally, we note that the points in (3.99) are linearly independent and thus cannot all be contained in the two-dimensional subspace U. This contradiction shows that $d^1(B_1(\mathbb{D})) > \sqrt{3}/2$ and completes the proof. □

4 Lower Bounds for Dictionaries of Ridge Functions

In this section, we consider lower bounds on the metric entropy, Kolmogorov, and Bernstein n-widths of convex subsets A of $L^2(\Omega)$. We show that if A contains a

certain class of ridge functions, then these quantities must be bounded below. We will apply this result to lower bound the entropy and n-widths of variation spaces corresponding to shallow neural networks.

Our method works by constructing a large collection of nearly orthogonal vectors in A and then obtaining lower bounds by noting that A must contain the convex hull of these vectors. We begin with some Lemmas lower bounding the entropy, Kolmogorov, and Bernstein n-widths of such a convex hull. This idea has been used to lower bound the entropy in [29, 39], yet these authors did not find as large a collection of nearly orthogonal vectors and obtained suboptimal bounds as a result.

Lemma 10 *Let H be a Hilbert space and $A \subset H$ a convex and symmetric set. Suppose that $g_1, ..., g_n \subset A$. Then*

$$\varepsilon_n(A) \geq \frac{1}{2}\sqrt{\frac{\lambda_{min}}{n}}, \ b_n(A) \geq \sqrt{\frac{\lambda_{min}}{n}} \tag{4.1}$$

where λ_{min} is the smallest eigenvalue of the Gram matrix G defined by $G_{ij} = \langle g_i, g_j \rangle_H$.

Proof Consider a maximal set of points $x_1, ..., x_N \in b_1^n(0,1) := \{x \in \mathbb{R}^n : |x|_1 \leq 1\}$ in the ℓ^1-unit ball satisfying $|x_i - x_j| \geq \frac{1}{2}$ for each $i \neq j$. We claim that $N \geq 2^n$. Indeed, if the set $\{x_i\}_{i=1}^N$ is maximal, then the balls

$$b_1^n(x_i, 1/2) = \left\{x \in \mathbb{R}^n : |x - x_i|_1 \leq \frac{1}{2}\right\}$$

must cover the ball $b_1^n(0,1)$. This implies that

$$\sum_{i=1}^N |b_1^n(x_i, 1/2)| \geq |b_1^n(0,1)|. \tag{4.2}$$

Since we obviously have $|b_1^n(x_i, 1/2)| = (1/2)^n |b_1^n(0,1)|$ for each i, it follows that $N \geq 2^n$.

Consider the collection of elements $f_1, ..., f_N \in H$ defined by

$$f_i = \sum_{k=1}^n x_i^k g_k. \tag{4.3}$$

Since A is symmetric and convex, we have $f_i \in A$ for each $i = 1, ..., N$. Moreover, if $i \neq j$, then

$$\|f_i - f_j\|_H^2 = v_{ij}^T G v_{ij}, \tag{4.4}$$

where $v_{ij} = x_i - x_j$. Since $|x_i - x_j|_1 \geq \frac{1}{2}$, it follows from Hölder's inequality that $|v_{ij}|_2^2 \geq \frac{1}{4n}$. From the eigenvalues of G, we then see that $\|f_i - f_j\|_H^2 \geq \frac{\lambda_{min}}{4n}$ for all $i \neq j$. This gives the entropy lower bound (4.1).

To lower bound the Bernstein widths, we note that if $g_1, ..., g_n$ are linearly dependent, then $\lambda_{min} = 0$ and there is noting to prove. On the other hand, consider the linear subspace V_n spanned by the g_i. Then $V_n \cap A$ contains the convex hull of $g_1, ...g_n$ and so for every $x \in \partial(A \cap V_n)$ we have

$$\|x\|_H^2 \geq \inf_{\|a\|_1=1} \left\| \sum_{i=1}^{n} a_i g_i \right\|_H^2 \geq \lambda_{min} n^{-1}, \tag{4.5}$$

since $\|a\|_1 = 1$ implies that $\|a\|_2^2 \geq n^{-1}$. This completes the bound on the Bernstein widths. □

This lemma can be applied to sequences of almost orthogonal vectors to obtain Lemma 3 from [39], which we state here as a corollary for completeness.

Corollary 1 *Let H be a Hilbert space and $A \subset H$ a convex and symmetric set. Suppose that $g_1, ..., g_n \subset A$ and the g_i are almost orthogonal in the sense that for all $i = 1, ..., n$,*

$$\sum_{j \neq i} |\langle g_i, g_j \rangle_H| \leq \frac{1}{2} \|g_i\|_H^2. \tag{4.6}$$

Then

$$\varepsilon_n(A) \geq \frac{\min_i \|g_i\|_H}{\sqrt{8n}}, \; b_n(A) \geq \frac{\min_i \|g_i\|_H}{\sqrt{2n}} \tag{4.7}$$

Proof This follows from Lemma 10 if we can show that the Gram matrix G satisfies

$$\lambda_{min}(G) \geq \frac{1}{2} \min_i \|g_i\|_H^2. \tag{4.8}$$

This follows immediately from the diagonal dominance condition 4.6 and the Gerschgorin circle theorem (see the proof in [39] for details). □

In order to lower bound the Kolmogorov n-widths, we will need the following Lemma, which generalizes Lemma 6 in [4] to almost orthogonal sets, which satisfy a stronger notion of almost orthogonality than that in Corollary 1.

Lemma 11 *Let H be a Hilbert space and $A \subset H$ a convex and symmetric set. Suppose that $g_1, ..., g_{2n} \subset A$ and the g_i are almost orthogonal in the sense that for all $i = 1, ..., 2n$,*

$$\sum_{j \neq i} |\langle g_i, g_j \rangle_H| \leq \frac{1}{2} \min_j \|g_j\|_H^2. \tag{4.9}$$

Then

$$d_n(A) \geq \frac{1}{2} \min_j \|g_j\|_H. \tag{4.10}$$

Proof By scaling down the g_j if necessary, we may assume that $\|g_j\|_H = \min_i \|g_i\|_H$ for all j. This follows since the rescaled vectors will clearly be in A (due to symmetry) and condition (4.9) will still be satisfied (since the left-hand side can only decrease while the right-hand side is unchanged). We can further assume without loss of generality that $\|g_j\|_H = 1$ for all $j = 1, ..., 2n$.

Let V_n be an n-dimensional subspace of H with orthonormal basis $e_1, ..., e_n$. For each index $i = 1, ..., n$, we will have

$$\sum_{j=1}^{2n} |\langle e_i, g_j \rangle|^2 \leq \lambda_{max}(G), \tag{4.11}$$

where G is the Gram matrix of the g_j. Using the Gerschgorin circle theorem and condition (4.9), we get $\lambda_{max}(G) \leq \frac{3}{2}$. Summing over i and switching the order of summation, we get

$$\sum_{j=1}^{2n} \sum_{i=1}^{n} |\langle e_i, g_j \rangle|^2 = \sum_{i=1}^{n} \sum_{j=1}^{2n} |\langle e_i, g_j \rangle|^2 \leq \frac{3}{2} n. \tag{4.12}$$

From this, we see that there must exist an index j, such that

$$\sum_{i=1}^{n} |\langle e_i, g_j \rangle|^2 \leq \frac{3}{4}. \tag{4.13}$$

But this means that the projection of g_j onto V_n has norm at most $\frac{3}{4}$, so that $d(g_j, V_n) \geq \frac{1}{2}$. Since this bound holds for some j for any subspace V_n of dimension n, we get the desired lower bound. □

Using the relationship between the $\mathscr{K}_1(\mathbb{P}_k^d)$-norm and the spectral Barron norm (1.23), we obtain that

$$\|f_\xi\|_{\mathscr{K}_1(\mathbb{P}_k^d)} \lesssim 1, \tag{4.14}$$

where $f_\xi(x) = (1 + |\xi|)^{-(k+1)} e^{2\pi i \xi \cdot x}$. In other words, the space $\mathscr{K}_1(\mathbb{P}_k^d)$ contains appropriately rescaled frequencies.

By considering the collection of functions f_ξ for $\xi \in \mathbb{Z}^d$ with $|\xi|_\infty \leq R$, we can make the f_ξ orthogonal on $[0, 1]^d$. Applying Lemmas 1 and 4.9 with this set of functions yields the bounds

$$\begin{aligned} &\varepsilon_n(B_1(\mathbb{P}_k^d))_{L^2([0,1]^d)} \gtrsim_{k,d} n^{-\frac{1}{2}-\frac{k+1}{d}}, \ b_n(B_1(\mathbb{P}_k^d))_{L^2([0,1]^d)} \\ &\gtrsim_{k,d} n^{-\frac{1}{2}-\frac{k+1}{d}}, \ d_n(B_1(\mathbb{P}_k^d))_{L^2([0,1]^d)} \gtrsim_{k,d} n^{-\frac{k+1}{d}}. \end{aligned} \tag{4.15}$$

This argument, which is essentially using the fact that the spectral Barron norm bounds the $\mathscr{K}_1(\mathbb{P}_k^d)$-norm, was used in [29, 39] to obtain a lower bound on the metric entropy $\varepsilon_n(B_1(\mathbb{P}_k^d))_{L^2([0,1]^d)}$.

However, it is known that $\mathscr{B}_{k+1} \subsetneq \mathscr{K}_1(\mathbb{P}_k^d)$, in other words that the spectral Barron space is strictly smaller than the variation space $\mathscr{K}_1(\mathbb{P}_k^d)$ [23, 24]. Consequently it should be possible to obtain a better lower bound on $\varepsilon_n(B_1(\mathbb{P}_k^d))_{L^2}$, $b_n(B_1(\mathbb{P}_k^d))_{L^2}$, and $d_n(B_1(\mathbb{P}_k^d))_{L^2}$, which would precisely quantify the gap between the spectral Barron space and $\mathscr{K}_1(\mathbb{P}_k^d)$.

The first such improved lower bound on $\varepsilon_n(B_1(\mathbb{P}_k^d))_{L^2}$ was obtained by Makovoz [39] in the case $k = 0$, $d = 2$, and it was conjectured that an improved lower bound holds more generally. We settle this conjecture by deriving an improved lower bound for all $k \geq 0$ and $d \geq 2$ and removing a logarithm from Makovoz's original bound. In addition, we also derive an improved lower bound on the Kolmogorov n-widths $d_n(B_1(\mathbb{P}_k^d))_{L^2}$ and a bound on the Bernstein widths $b_n(B_1(\mathbb{P}_k^d))_{L^2}$.

Theorem 8 *Let $d \geq 2$, $k \geq 0$, and denote the unit ball in $\mathbb{R}^d$ by*

$$B_1^d = \{x \in \mathbb{R}^d,\ |x|_2 \leq 1\}. \tag{4.16}$$

Let $A \subset L^2(B_1^d)$ be a convex and symmetric set. Suppose that for every profile $\phi \in C_c^\infty([-2,2])$ such that $\|\phi^{(k+1)}\|_{L^1(\mathbb{R})} \leq 1$, and any direction $\omega \in S^{d-1}$, the ridge function $\phi(\omega \cdot x) \in L^2(B_1^d)$ satisfies

$$\phi(\omega \cdot x) \in A. \tag{4.17}$$

Then

$$\begin{aligned} \varepsilon_n(A)_{L^2(B_1^d)} &\gtrsim_{k,d} n^{-\frac{1}{2}-\frac{2k+1}{2d}},\ b_n(A)_{L^2(B_1^d)} \\ &\gtrsim_{k,d} n^{-\frac{1}{2}-\frac{2k+1}{2d}},\ d_n(A)_{L^2(B_1^d)} \gtrsim_{k,d} n^{-\frac{2k+1}{2d}}. \end{aligned} \tag{4.18}$$

The argument we give here adapts the argument in the proof of Theorem 4 in [39]. A careful analysis allows us extend the result to higher dimensions and remove a logarithmic factor. The key is to consider profiles ϕ whose higher-order moments vanish in combination with a weighted L^2-norm with a Bochner–Riesz type weight.

Before we give the proof, we observe that the Peano kernel formula

$$\begin{aligned} \phi(x) &= \frac{1}{k!}\int_{-2}^{2} \phi^{(k+1)}(t)[\max(0, x-t)]^k \mathrm{d}t \\ &= \frac{1}{k!}\int_{-2}^{2} \phi^{(k+1)}(t)\sigma_k(0, x-t)\mathrm{d}t, \end{aligned} \tag{4.19}$$

which holds for all $\phi \in C_c^\infty([-2,2])$, implies that for a constant $C = C(k,d)$, the unit ball $C\,B_1(\mathbb{P}_k^d)$ satisfies the conditions of Theorem 8. Combining this with the fact that any bounded domain Ω is contained in a large enough ball yields the result given in the introduction:

Theorem 9 *Let $d \geq 2$ and $\Omega \subset \mathbb{R}^d$ a bounded domain. Then*

$$\varepsilon_n(B_1(\mathbb{P}_k^d))_{L^2(\Omega)} \gtrsim_{k,d} n^{-\frac{1}{2}-\frac{2k+1}{2d}}, \ b_n(B_1(\mathbb{P}_k^d))_{L^2(\Omega)} \gtrsim_{k,d} n^{-\frac{1}{2}-\frac{2k+1}{2d}}, \ d_n(B_1(\mathbb{P}_k^d))_{L^2(\Omega)} \gtrsim_{k,d} n^{-\frac{2k+1}{2d}}. \tag{4.20}$$

Note that the lower bound for $k = 0$ also applies to the variation spaces for networks with more general sigmoidal activation functions as well. This follows by a standard argument which scales the sigmoidal function to approximate a Heaviside activation function [4]. In addition, Theorem 8 can be applied to more general activation functions as well, for instance the B-spline activation functions $\sigma_{k,B}$, but we do not give the details here.

Proof of Theorem 8 We introduce the weight

$$\mathrm{d}\mu = (1 - |x|^2)_+^{\frac{d}{2}} \mathrm{d}x$$

of Bochner–Riesz type on B_1^d and consider the space $H = L^2(B_1^d, \mathrm{d}\mu)$. Since $1 - |x|^2 \leq 1$, it follows that $\|f\|_H \leq \|f\|_{L^2(\Omega)}$, and so it suffices to lower bound the entropy and n-widths of A with respect to the weighted space H.

Choose $0 \not\equiv \psi \in C_c^\infty([-1, 1])$ such that $2d - 1$ of its moments vanish, i.e., such that

$$\int_{-1}^{1} x^r \psi(x) \mathrm{d}x = 0, \tag{4.21}$$

for $r = 0, ..., 2d - 2$. Such a function ψ can easily be obtained by convolving an arbitrary compactly supported function whose moments vanish (such as a Legendre polynomial) with a C^∞ bump function.

Our assumptions on the set A imply that by scaling ψ appropriately, we can ensure that for $0 < \delta < 1$

$$\delta^k \psi(\delta^{-1}\omega \cdot x + b) \in A, \tag{4.22}$$

for any $\omega \in S^{d-1}$ and $b \in [-\delta^{-1}, \delta^{-1}]$. Note that ψ, which will be fixed in what follows, depends upon both d and k.

Let $N \geq 1$ be an integer and fix $n = N^{d-1}$ directions $\omega_1, ..., \omega_n \in S^{d-1}$ with $\min(|\omega_i - \omega_j|_2, |\omega_i + \omega_j|_2) \gtrsim_d N^{-1}$. This can certainly be done since projective space $P^{d-1} = S^{d-1}/\{\pm\}$ has dimension $d - 1$. In particular, if $\omega_1, ..., \omega_n$ is a maximal set satisfying $\min(|\omega_i - \omega_j|_2, |\omega_i + \omega_j|_2) \geq cN^{-1}$, then balls of radius cN^{-1} centered at the ω_i must cover P^{d-1}. So we must have $n = \Omega(N^{d-1})$, and by choosing c appropriately, we can arrange $n = N^{d-1}$.

Further, let $a \leq \frac{1}{4}$ be a sufficiently small constant to be specified later and consider for $\delta = aN^{-1}$ the collection of functions

$$g_{p,l}(x) = \delta^k \psi(\delta^{-1}\omega_p \cdot x + 2l) \in A, \tag{4.23}$$

for $p = 1, ..., n$ and $l = -\frac{N}{2}, ..., \frac{N}{2}$.

The intuition here is that $g_{p.l}$ is a ridge function which varies in the direction ω_p and has the compactly supported profile ψ dilated to have width δ (and scaled appropriately to remain in A). The different values of l give different non-overlapping shifts of these functions. The proof proceeds by checking that the $g_{p,l}$ can be made 'nearly orthogonal' by choosing a sufficiently small.

Indeed, we claim that if a is chosen small enough, then the $g_{p,l}$ satisfy the conditions of Lemma 11, i.e., for each (p, l)

$$\sum_{(p',l')\neq(p,l)} |\langle g_{p,l}, g_{p',l'}\rangle_H| \leq \frac{1}{2} \min_{(p',l')} \|g_{p',l'}\|_H^2. \tag{4.24}$$

This will of course also imply that the weaker conditions of Corollary 1 will be satisfied.

Before giving the detailed calculation, we describe the key ideas.

If we consider two different directions ω_p and $\omega_{p'}$, functions $g_{p,l}$ and $g_{p',l}$ will be constant along the $(d-2)$-dimensional subspace orthogonal to both ω_p and $\omega_{p'}$. Thus, the inner product $\langle g_{p,l}, g_{p',l'}\rangle_H$ corresponds to an integral over a circle in the plane spanned by ω_p and $\omega_{p'}$. The integrand is given by a product of the profile ψ supported in two intersecting strips multiplied by the integral of the Bochner–Riesz weight $\mathrm{d}\mu$ along the $(d-2)$-dimensional subspace orthogonal to ω_p and $\omega_{p'}$. The weight $\mathrm{d}\mu$ has been chosen so that when we integrate out this $(d-2)$-dimensional subspace, we will obtain a polynomial which vanishes to a high degree at the boundary of the circle (and is zero outside). This, combined with the high-order vanishing of the profile ψ, results in the functions $g_{p,l}$ and $g_{p',l}$ satisfying the required 'near orthogonality' bounds. We give the detailed calculations in the following.

We begin by estimating $\|g_{p,l}\|_H^2$, as follows

$$\|g_{p,l}\|_H^2 = \delta^{2k} \int_{B_1^d} |\psi(\delta^{-1}\omega_p \cdot x + 2l)|^2 (1-|x|^2)^{\frac{d}{2}} \mathrm{d}x. \tag{4.25}$$

We proceed to complete ω_p to an orthonormal basis of $\mathbb{R}^d$, $b_1 = \omega_p, b_2, ..., b_d$ and denote the coordinates of x with respect to this basis by $y_i = x \cdot b_i$. Rewriting the above integral in this new orthonormal basis, we get

$$\begin{aligned} \|g_{p,l}\|_H^2 &= \delta^{2k} \int_{B_1^d} |\psi(\delta^{-1}y_1 + 2l)|^2 \left(1 - \sum_{i=1}^d y_i^2\right)^{\frac{d}{2}} dy_1 \cdots dy_d \\ &= \delta^{2k} \int_{-1}^1 |\psi(\delta^{-1}y_1 + 2l)|^2 \rho_d(y_1) dy_1, \end{aligned} \tag{4.26}$$

where

$$\begin{aligned}\rho_d(y) &= \int_0^{\sqrt{1-y^2}} (1-y^2-r^2)^{\frac{d}{2}} r^{d-2}\mathrm{d}r \\ &= (1-y^2)^{d-\frac{1}{2}} \int_0^1 (1-r^2)^{\frac{d}{2}} r^{d-2}\mathrm{d}r = K_d(1-y^2)^{d-\frac{1}{2}},\end{aligned} \tag{4.27}$$

for a dimension dependent constant K_d.

Further, we change variables, setting $y = \delta^{-1}y_1 + 2l$ and use the fact that ψ is supported in $[-1, 1]$, to get

$$\|g_{p,l}\|_H^2 = K_d\delta^{2k+1}\int_{-1}^1 |\psi(y)|^2(1-[\delta(y-2l)]^2)^{d-\frac{1}{2}}\mathrm{d}y. \tag{4.28}$$

Since $|y| \le 1$ and $|2\,l| \le N$, as long as $\delta(N+1) \le 1/2$, which is guaranteed by $a \le \frac{1}{4}$, the coordinate $y_1 = \delta(y-2l)$ will satisfy $|y_1| \le 1/2$. This means that

$$(1-[\delta(y-2l)]^2)^{d-\frac{1}{2}} = (1-y_1^2)^{d-\frac{1}{2}} \ge (3/4)^{d-\frac{1}{2}}$$

uniformly in p, l, N and δ, and thus,

$$\|g_{p,l}\|_H^2 \ge K_d(3/4)^{d-\frac{1}{2}}\delta^{2k+1}\int_{-1}^1 |\psi(y)|^2\mathrm{d}y \gtrsim_{k,d} \delta^{2k+1}. \tag{4.29}$$

Next consider $|\langle g_{p,l}, g_{p',l'}\rangle_H|$ for $(p,l) \ne (p',l')$.

If $p = p'$, then $\omega_p = \omega_{p'}$, but $l \ne l'$. In this case, we easily see that the supports of $g_{p,l}$ and $g_{p,l'}$ are disjoint and so the inner product $\langle g_{p,l}, g_{p',l'}\rangle_H = 0$.

On the other hand, if $p \ne p'$ we get

$$\begin{aligned}\langle g_{p,l}, g_{p',l'}\rangle_H = \delta^{2k}\int_{B_1^d} &\psi(\delta^{-1}\omega_p\cdot x \\ &+2l)\psi(\delta^{-1}\omega_{p'}\cdot x + 2l')(1-|x|^2)^{\frac{d}{2}}\mathrm{d}x.\end{aligned} \tag{4.30}$$

Since $p \ne p'$, the vectors ω_p and $\omega_{p'}$ are linearly independent and we complete them to a basis $b_1 = \omega_p, b_2 = \omega_{p'}, b_3, ..., b_d$, where $b_3, ..., b_d$ is an orthonormal basis for the subspace orthogonal to ω_p and $\omega_{p'}$.

Letting $b_1', b_2', b_3' = b_3, ..., b_d' = b_d$ be a dual basis (i.e., satisfying $b_i'\cdot b_j = \delta_{ij}$) and making the change of variables $x = y_1b_1' + \cdots + y_db_d'$ in the above integral, we get

$$\begin{aligned}\langle g_{p,l}, g_{p',l'}\rangle_H = \delta^{2k}\det(D_{p,p'})^{-\frac{1}{2}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\psi(\delta^{-1}y_1+2l)\psi \\ (\delta^{-1}y_2+2l')\gamma_d(|y_1b_1'+y_2b_2'|)\mathrm{d}y_1\mathrm{d}y_2,\end{aligned} \tag{4.31}$$

where $D_{p,p'}$ is the Graham matrix of ω_1 and ω_2 (notice that then $D^{-1}_{p,p'}$ is the Graham matrix of b'_1 and b'_2) and

$$\begin{aligned}\gamma_d(y) &= \int_0^{\sqrt{1-y^2}} (1-y^2-r^2)_+^{\frac{d}{2}} r^{d-3}\mathrm{d}r \\ &= (1-y^2)_+^{d-1}\int_0^1 (1-r^2)_+^{\frac{d}{2}} r^{d-3}\mathrm{d}r = K'_d(1-y^2)_+^{d-1},\end{aligned} \tag{4.32}$$

for a second dimension dependent constant K'_d. (Note that if $d = 2$, then the above calculation is not correct, but we still have $\gamma_d(y) = (1-y^2)_+^{\frac{d}{2}} = (1-y^2)_+^{d-1}$.) We remark that the choice of Bochner–Riesz weight $\mathrm{d}\mu = (1-|x|^2)_+^{\frac{d}{2}}$ was made precisely so that γ_d is a piecewise polynomial with continuous derivatives of order $d-2$, which will be important in what follows.

Next, we fix y_1 and analyze, as a function of z,

$$\tau_{p,p'}(y_1, z) = \gamma_d(|y_1 b'_1 + z b'_2|) = K'_d(1 - q_{p,p'}(y_1, z))_+^{d-1},$$

where $q_{p,p'}$ is the quadratic

$$q_{p,p'}(y_1, z) = (b'_1 \cdot b'_1)y_1^2 - 2(b'_1 \cdot b'_2)y_1 z - (b'_2 \cdot b'_2)z^2, \tag{4.33}$$

We observe that, depending upon the value of y_1, $\tau_{p,p'}(y_1, z)$ is either identically 0 or is a piecewise polynomial function of degree $2d-2$ with exactly two break points at the roots z_1, z_2 of $q_{p,p'}(y_1, z) = 1$. Furthermore, utilizing Faà di Bruno's formula [19] and the fact that $q_{p,p'}(y_1, \cdot)$ is quadratic, we see that

$$\begin{aligned}\frac{\mathrm{d}^k}{\mathrm{d}z^k}\tau_{p,p'}(y_1, z)\Big|_{z_i} &= \sum_{m_1+2m_2=k} \frac{k!}{m_1! m_2! 2^{m_2}} f_d^{(m_1+m_2)}(1)\left[\frac{\mathrm{d}}{\mathrm{d}z}q_{p,p'}(y_1, z)|_{z_i}\right]^{m_1} \\ &\left[\frac{\mathrm{d}^2}{\mathrm{d}z^2}q_{p,p'}(y_1, z)|_{z_i}\right]^{m_2},\end{aligned} \tag{4.34}$$

where $f_d(x) = (1-x)^{d-1}$.

Since $f_d^{(m)}(1) = 0$ for all $m \le d-2$, we see that the derivative in (4.34) is equal to 0 for $0 \le k \le d-2$. Thus, the function $\tau_{p,p}(y_1, \cdot)$ has continuous derivatives up to order $d-2$ at the break points z_1 and z_2. Moreover, if we consider the derivative of order $k = d-1$, then only the term with $m_2 = 0$ in (4.34) survives and we get

$$\begin{aligned}\frac{\mathrm{d}^{d-1}}{\mathrm{d}z^{d-1}}\tau_{p,p'}(y_1, z)\Big|_{z_i} &= f_d^{(d-1)}(1)\left[\frac{\mathrm{d}}{\mathrm{d}z}q_{p,p'}(y_1, z)|_{z_i}\right]^{d-1} \\ &= (-1)^{d-1}(d-1)!\left[\frac{\mathrm{d}}{\mathrm{d}z}q_{p,p'}(y_1, z)|_{z_i}\right]^{d-1}.\end{aligned} \tag{4.35}$$

Utilizing the fact that the derivative of a quadratic $q(x) = ax^2 + bx + c$ at its roots is given by $\pm\sqrt{b^2 - 4ac}$ combined with the formula for $q_{p,p'}$ (4.33), we get

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}z} q_{p,p'}(y_1, z)|_{z_i} &= \pm 2\sqrt{(b_1' \cdot b_1')(b_1' \cdot b_2')^2 - (b_2' \cdot b_2')(b_1' \cdot b_1')} \\ &= \pm 2 \det(D_{p,p'})^{-\frac{1}{2}}. \end{aligned} \tag{4.36}$$

Taken together, this shows that the jump in the $d-1$-st derivative of $\tau_{p,p'}(y_1, z)$ at the break points z_1 and z_2 has magnitude

$$\left| \frac{\mathrm{d}^{d-1}}{\mathrm{d}z^{d-1}} \tau_{p,p'}(y_1, z) \Big|_{z_i} \right| \lesssim_d \det(D_{p,p'})^{-\frac{d-1}{2}}. \tag{4.37}$$

Going back to equation (4.31), we see that due to the compact support of ψ, the integral in (4.31) is supported on a square with side length 2δ in y_1 and y_2. To clarify this, we make the change of variables $s = \delta^{-1} y_1 + 2l$, $t = \delta^{-1} y_2 + 2l'$, and use that ψ is supported on $[-1, 1]$, to get (for notational convenience we let $y(s, l) = \delta(s - 2l)$)

$$\begin{aligned}&\langle g_{p,l}, g_{p',l'} \rangle_H \\ &\quad = \delta^{2k+2} \det(D_{p,p'})^{-\frac{1}{2}} \int_{-1}^{1} \int_{-1}^{1} \psi(s)\psi(t)\tau_{p,p'}(y(s, l), y(t, l'))\mathrm{d}s\mathrm{d}t. \end{aligned} \tag{4.38}$$

We now estimate the sum over l' as

$$\begin{aligned}&\sum_{l'=-\frac{N}{2}}^{\frac{N}{2}} |\langle g_{p,l}, g_{p',l'} \rangle_H| \\ &\quad = \delta^{2k+2} \det(D_{p,p'})^{-\frac{1}{2}} \sum_{l'=-\frac{N}{2}}^{\frac{N}{2}} \left| \int_{-1}^{1} \int_{-1}^{1} \psi(s)\psi(t)\tau_{p,p'}(y(s, l), y(t, l'))\mathrm{d}s\mathrm{d}t \right| \\ &\quad \leq \delta^{2k+2} \det(D_{p,p'})^{-\frac{1}{2}} \sum_{l'=-\frac{N}{2}}^{\frac{N}{2}} \int_{-1}^{1} \left| \int_{-1}^{1} \psi(s)\psi(t)\tau_{p,p'}(y(s, l), y(t, l'))\mathrm{d}t \right| \mathrm{d}s \\ &\quad = \delta^{2k+2} \det(D_{p,p'})^{-\frac{1}{2}} \int_{-1}^{1} |\psi(s)| \sum_{l'=-\frac{N}{2}}^{\frac{N}{2}} \left| \int_{-1}^{1} \psi(t)\tau_{p,p'}(y(s, l), y(t, l'))\mathrm{d}t \right| \mathrm{d}s. \end{aligned} \tag{4.39}$$

For fixed s and l, consider the inner sum

$$\sum_{l'=-\frac{N}{2}}^{\frac{N}{2}} \left| \int_{-1}^{1} \psi(t)\tau_{p,p'}(y(s, l), y(t, l'))\mathrm{d}t \right|$$

$$= \sum_{l'=-\frac{N}{2}}^{\frac{N}{2}} \left| \int_{-1}^{1} \psi(t) \tau_{p,p'}(y(s,l), \delta(t-2l')) \mathrm{d}t \right|. \tag{4.40}$$

In the integrals appearing in this sum, the variable $z = \delta(t - 2l')$ runs over the line segment $[\delta(2l' - 1), \delta(2l' + 1)]$. These segments are disjoint for distinct l' and are each of length 2δ.

Further, recall that for fixed $y_1 = y(s,l)$, the function $\tau_{p,p'}(y_1, z)$ is a piecewise polynomial of degree $2d-2$ with at most two break points z_1 and z_2. Combined with the fact that $2d-1$ moments of ψ vanish, this implies that at most two terms in the above sum are nonzero, namely those where the corresponding integral contains a break point.

Furthermore, the bound on the jump in the $d-1$st-order derivatives at the break points (4.37) implies that in the intervals (of length 2δ) which contain a break point, there exists a polynomial q_i of degree $d-2$ for which

$$|\tau_{p,p'}(y_1, z) - q_i(z)| \leq \frac{(2\delta)^{d-1}}{(d-1)!} M_d \det(D_{p,p'})^{-\frac{d-1}{2}} \lesssim_d \delta^{d-1} \det(D_{p,p'})^{-\frac{d-1}{2}} \tag{4.41}$$

on the given interval. Using again the vanishing moments of ψ, we see that the nonzero integrals in the sum (4.40) (of which there are at most 2) satisfy

$$\left| \int_{-1}^{1} \psi(t) \tau_{p,p'}(y(s,l), \delta(t-2l')) \mathrm{d}t \right| \lesssim_{k,d} \delta^{d-1} \det(D_{p,p'})^{-\frac{d-1}{2}}.$$

So for each fixed s and l, we get the bound

$$\sum_{l'=-\frac{N}{2}}^{\frac{N}{2}} \left| \int_{-1}^{1} \psi(t) \tau_{p,p'}(y(s,l), y(t,l')) \mathrm{d}t \right| \lesssim_{k,d} \delta^{d-1} \det(D_{p,p'})^{-\frac{d-1}{2}}. \tag{4.42}$$

Plugging this into equation (4.39), we get

$$\begin{aligned} &\sum_{l'=-\frac{N}{2}}^{\frac{N}{2}} |\langle g_{p,l}, g_{p',l'} \rangle_H| \\ &\lesssim_{k,d} \delta^{2k+d+1} \det(D_{p,p'})^{-\frac{d}{2}} \int_{-1}^{1} |\psi(s)| \mathrm{d}s \lesssim_{k,d} \delta^{2k+d+1} \det(D_{p,p'})^{-\frac{d}{2}}. \end{aligned} \tag{4.43}$$

We analyze the $\det(D_{p,p'})^{-\frac{d}{2}}$ term using that ω_p and $\omega_{p'}$ are on the sphere to get

$$\det(D_{p,p'})^{-\frac{d}{2}} = (1 - \langle \omega_p, \omega_{p'} \rangle^2)^{-\frac{d}{2}} = \frac{1}{\sin(\theta_{p,p'})^d}, \tag{4.44}$$

where $\theta_{p,p'}$ represents the angle between ω_p and $\omega_{p'}$.

Summing over $p' \neq p$, we get

$$\sum_{(p',l') \neq (p,l)} |\langle g_{p,l}, g_{p',l'} \rangle_H| \lesssim_{k,d} \delta^{2k+d+1} \sum_{p' \neq p} \frac{1}{\sin(\theta_{p,p'})^d}. \tag{4.45}$$

The final step is to bound the above sum. This is done in a relatively straightforward manner by noting that this sum is comparable to the following integral

$$\sum_{p' \neq p} \frac{1}{\sin(\theta_{p,p'})^d} \eqsim_d N^{d-1} \int_{P^{d-1} - B(p,r)} |x - p|^{-d} \mathrm{d}x, \tag{4.46}$$

where we are integrating over projective space minus a ball of radius $r \gtrsim_d N^{-1}$ around p. Integrating around this pole of order d in the $d-1$-dimensional P^{d-1}, this gives

$$\sum_{p' \neq p} \frac{1}{\sin(\theta_{p,p'})^d} \eqsim_d N^d. \tag{4.47}$$

To be more precise, we present the detailed estimates in what follows.

We bound the sum over one hemisphere

$$\sum_{0 < \theta_{p,p'} \leq \frac{\pi}{2}} \frac{1}{\sin(\theta_{p,p'})^d}, \tag{4.48}$$

and note that the sum over the other hemisphere can be handled in an analogous manner. To this end, we decompose this sum as

$$\sum_{0 < \theta_{p,p'} \leq \frac{\pi}{2}} \frac{1}{\sin(\theta_{p,p'})^d} = \sum_{0 < \theta_{p,p'} \leq \frac{\pi}{4}} \frac{1}{\sin(\theta_{p,p'})^d} + \sum_{\frac{\pi}{4} < \theta_{p,p'} \leq \frac{\pi}{2}} \frac{1}{\sin(\theta_{p,p'})^d}. \tag{4.49}$$

For the second sum, we note that $\sin(\theta_{p,p'}) \geq \frac{1}{\sqrt{2}}$, and the number of terms is at most $n = N^{d-1}$, so that the second sum is $\lesssim N^{d-1}$.

To bound the first sum in (4.49), we rotate the sphere so that $\omega_p = (0, ..., 0, 1)$ is the north pole. We then take the $\omega_{p'}$ for which $\theta_{p,p'} \leq \frac{\pi}{4}$ and project them onto the tangent plane at ω_p. Specifically, this corresponds to the map $\omega_{p'} = (x_1, ...x_{d-1}, x_d) \to x_{p'} = (x_1, ...x_{d-1})$, which removes the last coordinate.

It is now elementary to check that this maps distorts distances by at most a constant (since the $\omega_{p'}$ are all contained in a spherical cap of radius $\frac{\pi}{4}$), i.e., that for $p_1' \neq p_2'$, we have

$$|x_{p_1'} - x_{p_2'}| \leq |\omega_{p_1'} - \omega_{p_2'}| \lesssim |x_{p_1'} - x_{p_2'}|, \tag{4.50}$$

and also that $\sin(\theta_{p,p'}) = |x_{p'}|$.

This allows us to write the first sum in (4.49) as

$$\sum_{0<\theta_{p,p'}\leq\frac{\pi}{4}} \frac{1}{\sin(\theta_{p,p'})^d} = \sum_{0<|x_{p'}|\leq\frac{1}{\sqrt{2}}} \frac{1}{|x_{p'}|^d}, \tag{4.51}$$

where by construction we have $|\omega_{p_1'} - \omega_{p_2'}| \gtrsim_d N^{-1}$ for $p_1' \neq p_2'$, and thus, $|x_{p_1'} - x_{p_2'}| \gtrsim_d N^{-1}$ as well. In addition, $|\omega_p - \omega_{p'}| \gtrsim_d N^{-1}$, and thus, also $|x_{p'}| \gtrsim_d N^{-1}$.

Now let $r \gtrsim_d N^{-1}$ be such that the balls of radius r around each of the $x_{p'}$, and around 0, are disjoint. Notice that since $|x|^{-d}$ is a subharmonic function on $\mathbb{R}^{d-1}/\{0\}$, we have

$$\frac{1}{|x_{p'}|^d} \leq \frac{1}{|B(x_{p'}, r)|} \int_{B(x_{p'},r)} |y|^{-d} \mathrm{d}y \lesssim_d N^{d-1} \int_{B(x_{p'},r)} |y|^{-d} \mathrm{d}y. \tag{4.52}$$

Since all of the balls $B(x_{p'}, r)$ are disjoint and are disjoint from $B(0, r)$, we get (note that these integrals are in $\mathbb{R}^{d-1}$)

$$\begin{aligned} \sum_{0<|x_{p'}|\leq\frac{1}{\sqrt{2}}} \frac{1}{|x_{p'}|^d} &\lesssim_d N^{d-1} \int_{r\leq|y|\leq\frac{\pi}{2}+r} |y|^{-d} \mathrm{d}y \leq N^{d-1} \int_{r\leq|y|} |y|^{-d} \mathrm{d}y \\ &\lesssim_d N^{d-1} r^{-1} \lesssim_d N^d. \end{aligned} \tag{4.53}$$

Plugging this into equation (4.49) and bounding the sum over the other hemisphere in a similar manner, we get

$$\sum_{p'\neq p} \frac{1}{\sin(\theta_{p,p'})^d} \lesssim_d N^d. \tag{4.54}$$

Using equation (4.45), we finally obtain

$$\sum_{(p',l')\neq(p,l)} |\langle g_{p,l}, g_{p',l'}\rangle_H| \lesssim_{k,d} \delta^{2k+d+1} N^d. \tag{4.55}$$

Combined with the lower bound (4.29), which gives $\|g_{p,l}\|_H^2 \gtrsim_{k,d} \delta^{2k+1}$ for all (p, l), we see that by choosing the factor a in $\delta = aN^{-1}$ small enough (independently of N, of course), we can guarantee that the conditions of Lemma 11 (and thus also Corollary 1) are satisfied.

Applying Corollary 1, we see that

$$\varepsilon_n(A) \geq \frac{\min_{(p,l)} \|g_{p,l}\|_H}{\sqrt{8n}} \gtrsim_{k,d} n^{-\frac{1}{2}} \delta^{\frac{2k+1}{2}} \gtrsim_{k,d,a} n^{-\frac{1}{2}} N^{-\frac{2k+1}{2}}, \tag{4.56}$$

where $n = N^d$ is the total number of functions $g_{p,l}$. We obtain a completely analogous result for the Bernstein widths as well. This finally gives (since a is fixed depending only upon k and d)

$$\varepsilon_n(A) \gtrsim_{k,d} n^{-\frac{1}{2}-\frac{2k+1}{2d}}, \ b_n(A) \gtrsim_{k,d} n^{-\frac{1}{2}-\frac{2k+1}{2d}}. \tag{4.57}$$

Applying Lemma 11, we get

$$d_n(A) \geq \frac{1}{2} \min_{(p,l)} \|g_{p,l}\|_H \gtrsim_{k,d} \delta^{\frac{2k+1}{2}} \gtrsim_{k,d,a} N^{-\frac{2k+1}{2}}. \tag{4.58}$$

Since $n = N^d$ is the total number of functions $g_{p,l}$, we get as before

$$d_n(A) \gtrsim_{k,d} n^{-\frac{2k+1}{2d}}. \tag{4.59}$$

The monotonicity of the entropy and n-widths extends this bound to all n. This completes the proof. □

We remark that in the case of ReLU^k activation functions on the sphere, the high degree of symmetry allows the Kolmogorov n-widths to be determined exactly in terms of the spectrum of a kernel operator [2, 36], which we briefly describe in an abstract form here.

Specifically, the abstract situation here consists of the convex hull of a dictionary $\mathbb{D} = \{g \cdot f_e : g \in G\} \subset H$, where H is a Hilbert space, G is a compact Hausdorff topological group of isometries on the space H, and $f_e \in H$ is a fixed element. One simple example of this framework is the case where $H = \mathbb{R}^d$, $f_e = e_1$ is the first unit basis vector, and $G = \mathbb{Z}_n$ is the cyclic group on d elements. The action of G on $\mathbb{R}^n$ is given by cyclically shifting the indices. For this example, $B_1(\mathbb{D})$ is the unit ball of the ℓ^1-norm in $\mathbb{R}^n$ and this approach can be used to calculate its Kolmogorov n-widths with respect to ℓ^2 (see [37], chapter 14, for instance).

Another example, which we are primarily interested in here, is where $H = L^2(S^{d-1})$, $f_e(x) = \sigma(x_1) \in L^2(S^{d-1})$ for an activation function σ, and the group $G = O_d$ is the group of orthogonal transformations on $\mathbb{R}^d$. The action of $g \in G$, $(g \cdot f)(x) = f(g^{-1}x)$ is given by rotating the function f. In this case, the dictionary $\mathbb{D}$ is given by $\{\sigma(\omega \cdot x) : \omega \in S^{d-1}\} \subset L^2(S^{d-1})$. This is the situation which has been studied in [2, 36].

In this situation, we can lower bound the Kolmogorov n-widths by averaging over the group G. Let $x_1, ..., x_n$ be an orthonormal basis of a subspace X_n, and let $\mathrm{d}\mu$

denote the normalized Haar measure on G. Consider the average distance to X_n

$$\mathbb{E}_{g\sim \mathrm{d}\mu} d(f_g, X_n)^2 = \mathbb{E}_{g\sim \mathrm{d}\mu}\left(\|f_g\|_H^2 - \sum_{i=1}^n \langle f_g, x_i\rangle_H^2 \right), \tag{4.60}$$

where to simplify notation we have written f_g for $g \cdot f_e$. Using the assumption that G consists of isometries, we get

$$\mathbb{E}_{g\sim \mathrm{d}\mu} d(f_g, X_n)^2 = \|f_e\|_H^2 - \sum_{i=1}^n \mathbb{E}_{g\sim \mathrm{d}\mu} \langle f_g, x_i\rangle_H^2 = \|f_e\|_H^2 - \sum_{i=1}^n \langle x_i, T_G(x_i)\rangle_H, \tag{4.61}$$

where the operator $T_G : H \to H$ is given by the average of rank 1 operators:

$$T_G(x) = \mathbb{E}_{g\sim \mathrm{d}\mu} \langle f_g, x\rangle_H f_g. \tag{4.62}$$

From this formula, it is clear that T_G is a self-adjoint, compact, G-invariant operator on H with trace $\mathrm{Tr}(T_G) = \|f_e\|_H^2$. If we let $\lambda_1 \geq \lambda_2 \geq \cdots$ denote the eigenvalues of the operator T_G, we get from (4.61) and the minimax characterization of the eigenvalues that for any n-dimensional subspace $X_n \subset H$

$$\mathbb{E}_{g\sim \mathrm{d}\mu} d(f_g, X_n)^2 \geq \|f_e\|_H^2 - \sum_{i=1}^n \lambda_i = \sum_{i=n+1}^\infty \lambda_i, \tag{4.63}$$

with equality if X_n is the space spanned by $\phi_1, \ldots, \phi_n$, the eigenfunctions corresponding to the n largest eigenvectors. Since a maximum bounds an average, this gives the following lower bound on the Kolmogorov n-widths

$$d_n(B_1(\mathbb{D})) \geq \sqrt{\sum_{i=n+1}^\infty \lambda_i}. \tag{4.64}$$

Furthermore, suppose that $\lambda_n > \lambda_{n+1}$ and X_n is taken to be the space spanned by $\phi_1, \ldots, \phi_n$. Since $\lambda_n > \lambda_{n+1}$ and T_G is G-invariant, we have that X_n must be a G-invariant subspace. This means that $d(f_g, X_n)$ does not depend upon g and so the average and maximum coincide. Thus, if $\lambda_n > \lambda_{n+1}$, we actually have equality above and so

$$d_n(B_1(\mathbb{D})) = \sqrt{\sum_{i=n+1}^\infty \lambda_i}. \tag{4.65}$$

In the case of shallow neural networks on the sphere considered in [2, 36], the operator T_G is given by integration against an appropriate kernel and the eigenvalues λ_i can be explicitly calculated in the case where $\sigma = \mathrm{ReLU}^k$. (This is done in [2] and the result used to bound the Kolmogorov widths in [36].) This method allows an accurate determination of the constants in the n-width rates as well.

Finally, we will prove the following general result, from which a bound on the Kolmogorov n-widths leads to a lower bound on the metric entropy (see also [57] for a version of this argument, which we call the skewed simplex argument since we find a skewed image of the ℓ^1-unit ball in our space).

Proposition 2 *Let H be a Hilbert space and $A \subset H$ a symmetric, convex set. Then*

$$\varepsilon_n(A)_H \geq C d_n(A)_H n^{-\frac{1}{2}}, \tag{4.66}$$

for an absolute constant C.

Proof Let $\delta > 0$ and define a collection of elements $g_1, ..., g_n \in A$ recursively as follows:

$$\|g_i - P_{i-1} g_i\|_H \geq (1-\delta) \sup_{g \in A} \|g - P_{i-1} g\|_H, \tag{4.67}$$

where P_{i-1} is the orthogonal projection onto the span of $g_1, ..., g_{i-1}$. By definition of the n-widths, we have

$$\|g_i - P_{i-1} g_i\|_H \geq (1-\delta) d_i(A)_H \geq (1-\delta) d_n(A)_H. \tag{4.68}$$

Let $\tilde{g}_1, ..., \tilde{g_n}$ be the Gram–Schmidt orthogonalization of $g_1, ..., g_n$. Since the change of basis between $g_1, ..., g_n$ and $\tilde{g}_1, ..., \tilde{g_n}$ is upper triangular with ones on the diagonal, the volume (viewed in the n-dimensional Euclidean space spanned by $g_1, ..., g_n$) of the convex hull of $g_1, ..., g_n$ and $\tilde{g}_1, ..., \tilde{g_n}$ is the same. Since $\tilde{g}_1, ..., \tilde{g_n}$ are orthogonal with length at least $(1-\delta) d_n(A)_H$, we get (from the volume of the ℓ^1-unit ball)

$$|\mathrm{co}(g_1, ..., g_n)| \geq ((1-\delta) d_n(A)_H)^n \frac{2^n}{n!}. \tag{4.69}$$

Using the covering definition of the entropy and comparing this with the volume of 2^n balls of radius $\varepsilon := \varepsilon_n(A)_H$, we get

$$((1-\delta) d_n(A)_H)^n \frac{2^n}{n!} \leq |\mathrm{co}(g_1, ..., g_n)| \leq (2\varepsilon)^n \frac{\pi^{n/2}}{\Gamma\left(\frac{n}{2}+1\right)}. \tag{4.70}$$

Utilizing Sterling's formula, we get

$$\varepsilon \geq C(1-\delta) d_n(A)_H n^{-\frac{1}{2}}, \tag{4.71}$$

for an absolute constant C. Letting $\delta \to 0$ completes the proof. □

Although the preceding method is simpler and allows a more precise estimate of the constants in the n-width and entropy rates for $B_1(\mathbb{D})$, we note that Theorem 8 is more general. Specifically, it does not require the high degree of symmetry that the preceding argument does and thus applies to more general domains Ω and dictionaries $\mathbb{D}$. In addition, Theorem 8 finds a collection of nearly orthogonal vectors as opposed to a (potentially highly) skewed image of the simplex within the set $B_1(\mathbb{D})$. This stronger condition enables us to obtain a lower bound on the Bernstein widths as well.

4.1 Lower Bounds on Approximation Rates for Shallow Neural Networks

In this section, we use Theorem 8 to obtain lower bounds on the approximation rates of shallow neural networks. The key is the following relationship between metric entropy and nonlinear approximation rates, which can be viewed as an analogue of Carl's inequality [7].

Theorem 10 *Let X be a Banach space and $\mathbb{D} \subset X$ a dictionary with $K_{\mathbb{D}} := \sup_{h\in\mathbb{D}} \|h\|_X < \infty$. Suppose that for some constants $0 < l < \infty$, $C < \infty$, the dictionary $\mathbb{D}$ can be covered by $C\varepsilon^{-l}$ sets of diameter ε for any $\varepsilon > 0$. If there exists an M, $K < \infty$ and $\alpha > 0$ such that for all $f \in B_1(\mathbb{D})$*

$$\inf_{f_n\in\Sigma^\infty_{n,M}(\mathbb{D})} \|f - f_n\|_X \le Kn^{-\alpha}, \tag{4.72}$$

then the entropy numbers of $B_1(\mathbb{D})$ are bounded by

$$\varepsilon_{n\log n}(B_1(\mathbb{D}))_X \lesssim n^{-\alpha}, \tag{4.73}$$

where the implied constant is independent of n.

Thus, a given approximation rate from the set $\Sigma^\infty_{n,M}(\mathbb{D})$ implies a corresponding bound on the metric entropy.

Note that we are considering approximation by the set $\Sigma^\infty_{n,M}(\mathbb{D})$, defined in (1.3), which corresponds to expansions with coefficients bounded in ℓ^∞. This is in contrast to previous results [29, 39] which obtained lower bounds when the coefficients were bounded in ℓ^1.

For the dictionaries $\mathbb{P}^d_k$, i.e., for ReLU^k networks, the set $\Sigma^\infty_{n,M}(\mathbb{P}^d_k)$ corresponds to shallow neural networks with n neurons, inner coefficients bounded, and outer coefficients bounded in ℓ^∞. For the dictionary $\mathbb{D}_\sigma := \{\sigma(\omega\cdot x + b)\ \omega\in\mathbb{R}^d,\ b\in\mathbb{R}\}$ where σ is a sigmoidal activation function, the set $\Sigma^\infty_{n,M}(\mathbb{D}_\sigma)$ corresponds to shallow neural networks with n neurons and outer coefficients bounded in ℓ^∞, with no bound on the inner coefficients.

Proof In what follows, all implied constants will be independent of n.

Let $n \ge 1$ be an integer. We use our assumption on $\mathbb{D}$ and set $\varepsilon = n^{-\alpha-1}$. Then that there is a subset $\mathscr{D}_n \subset \mathbb{D}$ such that $|\mathscr{D}_n| \le Cn^{(\alpha+1)l}$ and

$$\sup_{d\in\mathbb{D}} \inf_{s\in\mathscr{D}_n} \|d - s\|_H \le \varepsilon = n^{-\alpha-1}. \tag{4.74}$$

The next step, in which the argument differs from that in [29, 39], is to cover the unit ball in ℓ_∞^n by unit balls in ℓ_1^n of radius δ. Indeed, denoting a ball of radius R in a Banach space Y by $B_R(Y) = \{x \in Y : |x|_Y \leq R\}$, we see that

$$B_1(\ell_\infty^n) \subset B_n(\ell_1^n). \tag{4.75}$$

Furthermore, we can cover the unit ball in a space Y by $(1+\frac{2}{\delta})^n$ balls in Y of radius δ (see [53], page 63). Applying this to $Y = \ell_1^n$ and scaling the unit ball appropriately, we see that we can cover

$$B_M(\ell_\infty^n) \subset B_{Mn}(\ell_1^n) \tag{4.76}$$

by $(1+\frac{2Mn}{\delta})^n$ ℓ_1^n-balls of radius δ. Now we set $\delta = 2Mn^{-\alpha}$, so the number of balls will be at most

$$(1+n^{\alpha+1})^n = n^{(\alpha+1)n}(1+n^{-(\alpha+1)})^n \lesssim n^{(\alpha+1)n},$$

where the last inequality is due to $\alpha > 0$. Denote by $\mathscr{L}_n$ the centers of these balls.

Denote by $\mathscr{S}_n$ the set of all linear combinations of n elements of $\mathscr{D}_n$ with coefficients in $\mathscr{L}_n$. Then clearly

$$|\mathscr{S}_n| \leq |\mathscr{D}_n|^n |\mathscr{L}_n| \lesssim C^n n^{(\alpha+1)ln} n^{(\alpha+1)n} = C^n n^{(\alpha+1)(l+1)n}. \tag{4.77}$$

By (4.72), we have for every $f \in B_1(\mathbb{D})$ an $f_n \in \Sigma_{n,M}(\mathbb{D})$ such that

$$f_n = \sum_{j=1}^n a_j h_j \tag{4.78}$$

and $\|f - f_n\|_X \lesssim n^{-\alpha}$, $h_j \in \mathbb{D}$ and $|a_j| \leq M$ for each j.

We now replace the h_j by their closest elements in $\mathscr{D}_n$ and the coefficients a_j by their closest point in $\mathscr{L}_n$. Since $\|h_j\|_H \leq K_\mathbb{D}$ and $|a_j| \leq M$ for each j, this results in a point $\tilde{f}_n \in \mathscr{S}_n$ with

$$\|f_n - \tilde{f}_n\|_H \leq Mn\varepsilon + K_\mathbb{D}\delta = Mn^{-\alpha} + 2K_\mathbb{D}Mn^{-\alpha} \lesssim n^{-\alpha}.$$

Thus, $\|f - \tilde{f}_n\|_H \lesssim n^{-\alpha}$ and so

$$\varepsilon_{\log|\mathscr{S}_n|} \lesssim n^{-\alpha}. \tag{4.79}$$

By equation (4.77), we see that $\log|\mathscr{S}_n| \lesssim n\log n$, which completes the proof. □

Using Theorem 10 and Theorem 8, we can immediately conclude the following lower bound on the approximation rates by neural networks with ReLU^k activation function.

Corollary 2 *Let $k \geq 0$ and $M < \infty$ be fixed and suppose that $\alpha > \frac{1}{2} + \frac{2k+1}{2d}$. Then*

$$\sup_{n \geq 1} n^{\alpha} \left[\sup_{f \in B_1(\mathbb{P}_k^d)} \inf_{f_n \in \Sigma_{n,M}^{\infty}(\mathbb{P}_k^d)} \|f - f_n\|_{L^2(\Omega)} \right] = \infty. \tag{4.80}$$

This corollary shows that the exponent in the approximation rates for shallow ReLU^k neural networks with respect to the variation norm cannot be improved beyond $-\frac{1}{2} - \frac{2k+1}{2d}$, even if the ℓ^1 bound on the outer coefficients is relaxed to an ℓ^∞ bound.

Proof From the theory developed in Sect. 3, it is clear that the dictionaries $\mathbb{P}_k^d$ satisfy the assumptions of Theorem 10 since they are smoothly parameterized by compact manifolds. If the supremum in (4.80) were finite, then by Theorem 10 we would have $\varepsilon_{n \log n}(B_1(\mathbb{P}_k^d)) \lesssim n^{-\alpha}$. This contradicts the lower bound from Theorem 8 since $\alpha > \frac{1}{2} + \frac{2k+1}{2d}$. □

Next, we extend this result to sigmoidal activation functions with bounded variation. For this, we need the following technical lemma.

Lemma 12 *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain and suppose that σ is a sigmoidal function with bounded variation. Then there exist $C, l < \infty$ such that the dictionary*

$$\mathbb{D}_\sigma := \left\{ \sigma(\omega \cdot x + b),\ \omega \in \mathbb{R}^d,\ b \in \mathbb{R} \right\} \tag{4.81}$$

can be covered by $C\varepsilon^{-l}$ balls of radius ε in $L^2(\Omega)$. In particular, $\mathbb{D}_\sigma$ satisfies the assumptions of Theorem 10.

This result generalizes Lemma 2 in [39] by relaxing the assumption on σ. Instead of requiring a Lipschitz condition and the assumption that σ approaches the Heaviside σ_0 at a polynomial rate, we only require the activation function σ to have bounded variation.

Proof Consider the Jordan decomposition of the function $\sigma = \sigma^+ - \sigma^-$, where σ^+ and σ^- are non-decreasing functions and

$$\|\sigma\|_{BV} = \lim_{x \to \infty} (\sigma^+(x) + \sigma^-(x)) - \lim_{x \to -\infty} (\sigma^+(x) + \sigma^-(x)) < \infty. \tag{4.82}$$

Denote by $a^+ := \lim_{x \to -\infty} \sigma^+(x)$ and $b^+ := \lim_{x \to \infty} \sigma^+(x)$ and likewise for σ^-. By (4.82) a^+, b^+, a^- and b^- are all finite. Further, $[a^+, b^+]$ is the closure of the range of σ^+ and $[a^-, b^-]$ is the closure of the range of σ^-.

We proceed to divide the intervals $[a^+, b^+]$ and $[a^-, b^-]$ into intervals of length at most $\frac{\varepsilon}{2}$. Denote these intervals by $[x_{i-1}, x_i)$ and $[y_{i-1}, y_i)$ where we have

$$a^+ = x_0 < \cdots < x_{n_1} = b^+ \tag{4.83}$$

and

$$a^- = y_0 < \cdots < y_{n_2} = b^-. \tag{4.84}$$

This partitions the domain $\mathbb{R}$ into two sets of disjoint intervals: $\sigma^{-1}([x_{i-1}, x_i))$ for $i = 1, ..., n_1$ and $\sigma^{-1}([y_{i-1}, y_i))$ for $i = 1, ..., n_2$. Take the common refinement of these intervals, i.e., consider non-empty all intervals of the form $\sigma^{-1}([x_{i-1}, x_i)) \cap \sigma^{-1}([y_{j-1}, y_j))$ and define a piecewise constant function σ_ε by

$$\sigma_\varepsilon(x) = x_{i-1} + y_{j-1} \text{ if } x \in \sigma^{-1}([x_{i-1}, x_i)) \cap \sigma^{-1}([y_{j-1}, y_j)). \tag{4.85}$$

By construction, we have for any x that

$$|\sigma_\varepsilon(x) - \sigma(x)| \leq |x_{i-1} - \sigma^+(x)| + |y_{j-1} - \sigma^-(x)| \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon, \tag{4.86}$$

since $\sigma^+(x) \in [x_{i-1}, x_i)$ and $\sigma^-(x) \in [y_{j-1}, y_j)$. Thus, we have

$$\|\sigma_\varepsilon(\omega \cdot x + b) - \sigma(\omega \cdot x + b)\|_{L^2(\Omega, \mathrm{d}x)} \leq |\Omega|^{\frac{1}{2}} \varepsilon \tag{4.87}$$

uniformly in $\omega, b \in \mathbb{R}^d \times \mathbb{R}$.

In addition, it is easy to see that there are points $z_0 < z_1 < \cdots < z_n$ with $n = n_1 + n_2 \lesssim \varepsilon^{-1}$ such that σ_ε is constant on (z_i, z_{i+1}) and on $(-\infty, z_0)$ and (z_n, ∞).

Next, choose an ε^3-net for the slightly enlarged domain

$$\Omega_\varepsilon = \{x : \mathrm{dist}(x, \Omega) \leq \varepsilon^3\}, \tag{4.88}$$

which will contain at most $N \lesssim \varepsilon^{-3d}$ points $x_1, ..., x_N \in \Omega_\varepsilon$. For each x_i and each z_j, consider the hyperplane in the parameter space $\mathbb{R}^d \times \mathbb{R}$ given by

$$H_{ij} = \{(\omega, b) \in \mathbb{R}^d \times \mathbb{R} : \ \omega \cdot x_i + b = z_j\}. \tag{4.89}$$

It is well known that K hyperplanes in $\mathbb{R}^{d+1}$ cut the space $\mathbb{R}^{d+1}$ into at most

$$\sum_{i=0}^{d+1} \binom{K}{i} \leq K^{d+1} \tag{4.90}$$

regions. Thus, the hyperplanes H_{ij} cut the parameter space $\mathbb{R}^d \times \mathbb{R}$ into at most

$$M = (nN)^{d+1} \lesssim \varepsilon^{-(3d+1)(d+2)} \tag{4.91}$$

regions $R_1, ..., R_M$.

We claim that for each $i = 1, ..., M$, the set

$$S_i := \{\sigma(\omega \cdot x + b) : \ (\omega, b) \in R_i\} \tag{4.92}$$

is contained in a ball of radius $r \lesssim \varepsilon$ in $L^2(\Omega, \mathrm{d}x)$. Setting $l = (3d + 1)(d + 2)$ and choosing C appropriately large, we obtain the desired result.

Fix $(\omega, b), (\omega', b') \in R_i$. From the triangle inequality and equation (4.87), we see that

$$\begin{aligned}
\|\sigma(\omega \cdot x + b) - \sigma(\omega' \cdot x + b')\|_{L^2(\Omega,\mathrm{d}x)} &\leq \|\sigma(\omega \cdot x + b) \\
&- \sigma_\varepsilon(\omega \cdot x + b)\|_{L^2(\Omega,\mathrm{d}x)} \\
&+ \|\sigma_\varepsilon(\omega \cdot x + b) - \sigma_\varepsilon(\omega' \cdot x + b')\|_{L^2(\Omega,\mathrm{d}x)} \\
&+ \|\sigma_\varepsilon(\omega' \cdot x + b') - \sigma(\omega' \cdot x + b')\|_{L^2(\Omega,\mathrm{d}x)} \\
&\leq 2|\Omega|^{\frac{1}{2}}\varepsilon + \|\sigma_\varepsilon(\omega \cdot x + b) - \sigma_\varepsilon(\omega' \cdot x + b')\|_{L^2(\Omega,\mathrm{d}x)}.
\end{aligned} \tag{4.93}$$

To conclude the proof, we bound the difference

$$\begin{aligned}
&\|\sigma_\varepsilon(\omega \cdot x + b) - \sigma_\varepsilon(\omega' \cdot x + b')\|^2_{L^2(\Omega,\mathrm{d}x)} \\
&\quad = \int_\Omega (\sigma_\varepsilon(\omega \cdot x + b) - \sigma_\varepsilon(\omega' \cdot x + b'))^2 \mathrm{d}x.
\end{aligned} \tag{4.94}$$

For this, we consider the set

$$D = \{x \in \Omega : \ \sigma_\varepsilon(\omega \cdot x + b) \neq \sigma_\varepsilon(\omega' \cdot x + b')\}. \tag{4.95}$$

From the definition of σ_ε, we see that $x \in D$ only if there exists a z_j such that

$$\omega \cdot x + b \leq z_j \leq \omega' \cdot x + b', \tag{4.96}$$

or vice versa (i.e., with the order reversed). Thus, we have

$$D \subset \bigcup_{j=0}^{n} D_j^+ \cup \bigcup_{j=0}^{n} D_j^-, \tag{4.97}$$

where

$$D_j^+ = \{\omega \cdot x + b \leq z_j\} \cap \{z_j \leq \omega' \cdot x + b'\} \tag{4.98}$$

and

$$D_j^- = \{\omega \cdot x + b \geq z_j\} \cap \{z_j \geq \omega' \cdot x + b'\}. \tag{4.99}$$

By construction, none of the sets $D_j^\pm$ contain any of the points $x_1, ..., x_N$ since (ω, b) and (ω', b') are both in the same region R_i. Since $x_1, ..., x_N$ forms an ε^3-net for Ω_ε,

this implies that none of the $D_j^{\pm} \cap \Omega_\varepsilon$ can contain a ball of radius ε^3. Consider the sets

$$\begin{aligned}\Sigma_j &:= \{x \in \Omega : \operatorname{dist}(x, \{y : \omega \cdot y + b = z_j\}) \leq \varepsilon^3\} \\ \text{and } \Sigma_j' &:= \{x \in \Omega : \operatorname{dist}(x, \{y : \omega' \cdot y + b' = z_j\}) \leq \varepsilon^3\},\end{aligned} \tag{4.100}$$

which are strips of width ε^3 around the hyperplanes defined by $\omega \cdot y + b = z_j$ and $\omega' \cdot y + b' = z_j$ intersected with Ω, respectively. We claim that

$$D_j^{\pm} \cap \Omega \subset \Sigma_j \cup \Sigma_j', \tag{4.101}$$

for each j and choice of sign $\pm$. Suppose to the contrary that for some j there exists an $x \in D_j^+ \cap \Omega$ (the case of negative sign is exactly the same) such that

$$\begin{aligned}\operatorname{dist}(x, \{y : \omega \cdot y + b = z_j\}) &> \varepsilon^3 \text{ and} \\ \operatorname{dist}(x, \{y : \omega' \cdot y + b' = z_j\}) &> \varepsilon^3.\end{aligned} \tag{4.102}$$

These two conditions imply that the ball of radius ε^3 about x is contained in D_j^+. Further, since $x \in \Omega$, this ball is also contained in Ω_ε. But $D_j^+ \cap \Omega_\varepsilon$ cannot contain a ball of radius ε^3. This contradiction shows that (4.101) holds. From this, we deduce that

$$|D_j^+ \cap \Omega| \leq |\Sigma_j| + |\Sigma_j'| \lesssim \varepsilon^3, \tag{4.103}$$

since Σ_j and Σ_j' are strips of width ε^3 and Ω is a bounded domain. Using (4.97) and a union bound, we obtain

$$|D| \lesssim n\varepsilon^3 \lesssim \varepsilon^2. \tag{4.104}$$

Finally, the difference $\sigma_\varepsilon(\omega \cdot x + b) - \sigma_\varepsilon(\omega' \cdot x + b')$ is equal to 0 outside of D and on D it is bounded by $\sup_x \sigma(x) - \inf_x \sigma(x) \leq \|\sigma\|_{BV} \lesssim 1$. This implies that

$$\int_\Omega (\sigma_\varepsilon(\omega \cdot x + b) - \sigma_\varepsilon(\omega' \cdot x + b'))^2 \mathrm{d}x \lesssim \varepsilon^2, \tag{4.105}$$

and finally that

$$\|\sigma_\varepsilon(\omega \cdot x + b) - \sigma_\varepsilon(\omega' \cdot x + b')\|_{L^2(\Omega, \mathrm{d}x)} \lesssim \varepsilon. \tag{4.106}$$

Using (4.93) and that $(\omega, b), (\omega', b') \in S_i$ were arbitrary, we see that the diameter of the sets S_i is $\lesssim \varepsilon$, which completes the proof. □

Using Lemma 12, we show that the lower bound on the approximation rates holds even for a sigmoidal activation function with bounded variation.

Corollary 3 *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain and σ be a sigmoidal activation function with bounded variation. Consider the dictionary $\mathbb{D}_\sigma^d \subset L^2(B_1^d)$ defined in (1.4). Then for any $M < \infty$ and $\alpha > \frac{1}{2} + \frac{1}{2d}$ we have*

$$\sup_{n\geq 1} n^\alpha \left[\sup_{f\in B_1(\mathbb{D}_\sigma)} \inf_{f_n \in \Sigma_{n,M}^\infty(\mathbb{D}_\sigma)} \|f - f_n\|_{L^2(B_1^d)} \right] = \infty. \tag{4.107}$$

This shows that the exponent in the approximation rate derived by Makovoz [39] is optimal, even if the outer coefficients of the network are only bounded in ℓ^∞ and the activation function is a general sigmoidal function with bounded variation.

Proof We observe that since σ is a sigmoidal activation function and Ω is a bounded domain, we have

$$\lim_{a\to\infty} \|\sigma(a(\omega \cdot x + b)) - \sigma_0(\omega \cdot x + b)\|_{L^2(\Omega)} = 0, \tag{4.108}$$

where we recall that σ_0 is the Heaviside activation function. Since

$$\sigma(a(\omega \cdot x + b)) \in \mathbb{D}_\sigma \tag{4.109}$$

for every $a \in \mathbb{R}$, this implies that $B_1(\mathbb{D}_\sigma) \supset B_1(\mathbb{P}_0^d)$. By Lemma 12 and Theorem 10, if the supremum in (4.107) were finite, then the metric entropy would satisfy

$$\varepsilon_{n\log n}(B_1(\mathbb{P}_k^d)) \leq \varepsilon_{n\log n}(B_1(\mathbb{D}_\sigma)) \lesssim n^{-\alpha}.$$

This contradicts the lower bound from Theorem 8 since $\alpha > \frac{1}{2} + \frac{1}{2d}$. □

5 Conclusion

We have introduced the notion of a smoothly parameterized dictionary and have bounded both approximation rates and fundamental quantities such as the metric entropy and n-widths for convex hulls of such dictionaries. Further, we have developed a method for lower bounding n-widths and metric entropy of convex hulls of certain classes of ridge functions. Applying these results to shallow neural networks, we obtain sharp approximation rates for neural networks with ReLU^k activation functions, improving upon several results in the literature. In addition, this allows us to compare ReLU^k networks with other methods and to show that they are optimal on their corresponding variation space.

There are a few further questions we would like to propose. First, it is unclear how to compute entropy or n-width bounds on $B_1(\mathbb{D})$, and specifically $B_1(\mathbb{P}_k^d)$, in L^p for $p \neq 2$. For this problem, partial results appear in [2, 29, 38], but a complete solution seems to require significant new ideas. Second, we have been primarily interested in the rates for fixed dimension in this work and have not taken care to precisely determine the implied constants. As such, our work it mainly interesting for problems in a fix

dimension which is not too large. Obtaining tighter bounds on the constants will be important in quantifying the curse of dimensionality. Finally, we would like to extend this theory to approximation by deeper neural networks.

Acknowledgements We would like to thank Professors Russel Caflisch, Ronald DeVore, Weinan E, Albert Cohen, Stephan Wojtowytsch, Jason Klusowski, and Lei Wu for helpful discussions. This work was supported by the Verne M. Willaman Chair Fund at the Pennsylvania State University and the National Science Foundation (Grant No. DMS-1819157 and DMS-2111387).

References

1. Fernando Albiac and Nigel John Kalton, Topics in banach space theory, vol. 233, Springer, 2006.
2. Francis Bach, Breaking the curse of dimensionality with convex neural networks, The Journal of Machine Learning Research **18** (2017), no. 1, 629–681.
3. Keith Ball and Alain Pajor, The entropy of convex bodies with "few" extreme points, Proceedings of the 1989 Conference in Banach Spaces at Strob. Austria. Cambridge Univ. Press, 1990.
4. Andrew R Barron, Universal approximation bounds for superpositions of a sigmoidal function, IEEE Transactions on Information theory **39** (1993), no. 3, 930–945.
5. Andrew R Barron, Albert Cohen, Wolfgang Dahmen, and Ronald A DeVore, Approximation and learning by greedy algorithms, **36** (2008), no. 1, 64–94.
6. James H Bramble and SR Hilbert, Estimation of linear functionals on sobolev spaces with application to fourier transforms and spline interpolation, SIAM Journal on Numerical Analysis **7** (1970), no. 1, 112–124.
7. Bernd Carl, Entropy numbers, s-numbers, and eigenvalue problems, Journal of Functional Analysis **41** (1981), no. 3, 290–306.
8. Bernd Carl,Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in banach spaces, Annales de l'institut Fourier, vol. 35, 1985, pp. 79–118.
9. Bernd Carl, Metric entropy of convex hulls in hilbert spaces, Bulletin of the London Mathematical Society **29** (1997), no. 4, 452–458.
10. Bernd Carl, Aicke Hinrichs, and Alain Pajor, Gelfand numbers and metric entropy of convex hulls in hilbert spaces, Positivity **17** (2013), no. 1, 171–203.
11. Bernd Carl, Aicke Hinrichs, and Philipp Rudolph, Entropy numbers of convex hulls in banach spaces and applications, Journal of Complexity **30** (2014), no. 5, 555–587.
12. Bernd Carl, Ioanna Kyrezi, and Alain Pajor, Metric entropy of convex hulls in banach spaces, Journal of the London Mathematical Society **60** (1999), no. 3, 871–896.
13. Bernd Carl and Alain Pajor, Gelfand numbers of operators with values in a hilbert space, Inventiones mathematicae **94** (1988), no. 3, 479–504.
14. Kwok Chiu Chung and Te Hai Yao, On lattices admitting unique lagrange interpolations, SIAM Journal on Numerical Analysis **14** (1977), no. 4, 735–743.
15. Philippe G Ciarlet and Pierre-Arnaud Raviart, General lagrange and hermite interpolation in rn with applications to finite element methods, Archive for Rational Mechanics and Analysis **46** (1972), no. 3, 177–199.
16. Albert Cohen, Ronald Devore, Guergana Petrova, and Przemyslaw Wojtaszczyk, Optimal stable nonlinear approximation, Foundations of Computational Mathematics (2021), 1–42.
17. Ronald A DeVore, Nonlinear approximation, Acta numerica **7** (1998), 51–150.
18. Ronald A DeVore, Ralph Howard, and Charles Micchelli, Optimal nonlinear approximation, Manuscripta mathematica **63** (1989), no. 4, 469–478.
19. F Faà Di Bruno,Note sur une nouvelle formule de calcul différentiel, Quarterly J. Pure Appl. Math **1** (1857), no. 359-360, .
20. David L Donoho, Compressed sensing, IEEE Transactions on information theory **52** (2006), no. 4, 1289–1306.
21. Richard M Dudley, The sizes of compact subsets of hilbert space and continuity of gaussian processes, Journal of Functional Analysis **1** (1967), no. 3, 290–330.
22. RM Dudley, Universal donsker classes and metric entropy, Ann. Probab. **15** (1987), no. 4, 1306–1326.

23. W. E, Chao Ma, and Lei Wu, Barron spaces and the compositional function spaces for neural network models, arXiv preprint arXiv:1906.08039 (2019).
24. W. E and Stephan Wojtowytsch, Representation formulas and pointwise properties for barron functions., CoRR (2020).
25. Weinan E, Chao Ma, and Lei Wu, Barron spaces and the compositional function spaces for neural network models, arXiv preprint arXiv:1906.08039 (2019).
26. David E Edmunds and Jan Lang, Gelfand numbers and widths, Journal of Approximation Theory **166** (2013), 78–84.
27. Uffe Haagerup, The best constants in the khintchine inequality, Studia Mathematica **70** (1981), 231–283.
28. Lee K Jones, A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training, The annals of Statistics **20** (1992), no. 1, 608–613.
29. Jason M Klusowski and Andrew R Barron, Approximation by combinations of relu and squared relu ridge functions with ℓ^1 and ℓ^0 controls, IEEE Transactions on Information Theory **64** (2018), no. 12, 7649–7656.
30. Andrei Nikolaevich Kolmogorov, On linear dimensionality of topological vector spaces, Doklady Akademii Nauk, vol. 120, Russian Academy of Sciences, 1958, pp. 239–241.
31. Vera Kurková and Marcello Sanguineti, Bounds on rates of variable-basis and neural-network approximation, IEEE Transactions on Information Theory **47** (2001), no. 6, 2659–2665.
32. Vera Kurková and Marcello Sanguineti,Comparison of worst case errors in linear and neural network approximation, IEEE Transactions on Information Theory **48** (2002), no. 1, 264–275.
33. Jan Lang and David Edmunds, Eigenvalues, embeddings and generalised trigonometric functions, vol. 2016, Springer Science & Business Media, 2011.
34. Michel Ledoux and Michel Talagrand, Probability in banach spaces: isoperimetry and processes, Springer Science & Business Media, 2013.
35. Wee Sun Lee, Peter L Bartlett, and Robert C Williamson, Efficient agnostic learning of neural networks with bounded fan-in, IEEE Transactions on Information Theory **42** (1996), no. 6, 2118–2132.
36. Jihao Long and Lei Wu, Linear approximability of two-layer neural networks: A comprehensive analysis based on spectral decay, arXiv preprint arXiv:2108.04964 (2021).
37. George G Lorentz, Manfred v Golitschek, and Yuly Makovoz, Constructive approximation: advanced problems, vol. 304, Springer, 1996.
38. Y Makovoz, Uniform approximation by neural networks, Journal of Approximation Theory **95** (1998), no. 2, 215–228.
39. Yuly Makovoz, Random approximants and neural networks, Journal of Approximation Theory **85** (1996), no. 1, 98–109.
40. Jiří Matoušek, Tight upper bounds for the discrepancy of half-spaces, Discrete & Computational Geometry **13** (1995), no. 3, 593–601.
41. Jiří Matoušek,Improved upper bounds for approximation by zonotopes, Acta Mathematica **177** (1996), no. 1, 55–73.
42. Jiri Matousek, Geometric discrepancy: An illustrated guide, vol. 18, Springer Science & Business Media, 1999.
43. RA Nicolaides, On a class of finite elements generated by lagrange interpolation, SIAM Journal on Numerical Analysis **9** (1972), no. 3, 435–445.
44. Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro, A function space view of bounded norm infinite width relu nets: The multivariate case, International Conference on Learning Representations (ICLR 2020), 2019.
45. Rahul Parhi and Robert D Nowak, Banach space representer theorems for neural networks and ridge splines, arXiv preprint arXiv:2006.05626 (2020).
46. Rahul Parhi and Robert D Nowak, What kinds of functions do deep neural networks learn? insights from variational spline theory, arXiv preprint arXiv:2105.03361 (2021).
47. A. Pietsch, Ideals (Algebra), and North-Holland Publishing Company, Operator ideals, Mathematical Studies, North-Holland Publishing Company, 1980.
48. Albrecht Pietsch, s-numbers of operators in banach spaces, Studia Mathematica **51** (1974), 201–223.
49. Albrecht Pietsch,Operator ideals, vol. 16, Deutscher Verl d Wiss, 1978.
50. Albrecht Pietsch,History of banach spaces and linear operators, Springer Science & Business Media, 2007.

51. Allan Pinkus, N-widths in approximation theory, vol. 7, Springer Science & Business Media, 2012.
52. Gilles Pisier, Remarques sur un résultat non publié de b. maurey, Séminaire Analyse fonctionnelle (dit "Maurey-Schwartz") (1981), 1–12.
53. Gilles Pisier,The volume of convex bodies and banach space geometry, vol. 94, Cambridge University Press, 1999.
54. R Tyrrell Rockafellar, Convex analysis, no. 28, Princeton university press, 1970.
55. Jonathan W Siegel and Jinchao Xu, Approximation rates for neural networks with general activation functions, Neural Networks **128** (2020), 313–321.
56. Jonathan W Siegel and Jinchao Xu,Characterization of the variation spaces corresponding to shallow neural networks, arXiv preprint arXiv:2106.15002 (2021).
57. Jonathan W Siegel and Jinchao Xu,Improved convergence rates for the orthogonal greedy algorithm, arXiv preprint arXiv:2106.15000 (2021).
58. Jonathan W Siegel and Jinchao Xu,High-order approximation rates for shallow neural networks with cosine and reluk activation functions, Applied and Computational Harmonic Analysis **58** (2022), 1–26.
59. Hans Triebel, Interpolation theory, function spaces, Differential Operators (1995).
60. Jinchao Xu, Error estimates of the finite element method for the 2nd order elliptic equations with discontinuous coefficients, J. Xiangtan University **1** (1982), 1–5.
61. Jinchao Xu, Estimate of the convergence rate of finite element solutions to elliptic equations of second order with discontinuous coefficients, arXiv preprint arXiv:1311.4178 (2013).
62. Jinchao Xu, Finite neuron method and convergence analysis, Communications in Computational Physics **28** (2020), no. 5, 1707–1745.
63. Yuhong Yang and Andrew Barron, Information-theoretic determination of minimax rates of convergence, Annals of Statistics (1999), 1564–1599.