Constructive Approximation (2023) 57:1109–1132
https://doi.org/10.1007/s00365-023-09626-4

CONSTRUCTIVE APPROXIMATION

Characterization of the Variation Spaces Corresponding to Shallow Neural Networks

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

Received: 28 June 2021 / Revised: 9 April 2022 / Accepted: 2 June 2022 /
Published online: 22 February 2023

Abstract
We study the variation space corresponding to a dictionary of functions in $L^2(\Omega)$ for a bounded domain $\Omega \subset \mathbb{R}^d$. Specifically, we compare the variation space, which is defined in terms of a convex hull with related notions based on integral representations. This allows us to show that three important notions relating to the approximation theory of shallow neural networks, the Barron space, the spectral Barron space, and the Radon BV space, are actually variation spaces with respect to certain natural dictionaries.

Keywords Function space · Neural networks · Approximation

Mathematics Subject Classification 68T05 · 46B99

1 Introduction

In this work we consider the variation space with respect to a dictionary $\mathbb{D} \subset H$ in a separable Hilbert space H. This notion arises in the study of non-linear approximation by an expansion of dictionary elements [2, 3]. Suppose that $\sup_{d\in\mathbb{D}} \|d\|_H = K_{\mathbb{D}} < \infty$, and consider the variation norm [16, 17] of $\mathbb{D}$ defined by

$$\|f\|_{\mathbb{D}} = \inf\left\{c > 0 : \ f/c \in \overline{\operatorname{conv}(\pm\mathbb{D})}\right\}. \tag{1.1}$$

Communicated by Zuowei Shen.

✉ Jonathan W. Siegel
jus1949@psu.edu

Jinchao Xu
jxx1@psu.edu

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

This is the gauge, or Minkowski functional of the closed symmetric convex hull of $\mathbb{D}$,

$$\overline{\text{conv}(\pm\mathbb{D})} = \left\{ \sum_{j=1}^{n} a_j h_j : \ n \in \mathbb{N}, \ h_j \in \mathbb{D}, \ \sum_{i=1}^{n} |a_i| \leq 1 \right\}.$$

The variation space $\mathscr{K}(\mathbb{D})$ is then given by

$$\mathscr{K}(\mathbb{D}) := \{f \in H : \ \|f\|_{\mathbb{D}} < \infty\}.$$

The varation norm and variation space have been introduced in different forms in the literature and play an important role in the approximation theory of neural networks [1, 14, 22, 23, 35], the convergence theory of greedy algorithms [2, 4, 21, 36–38] and in non-linear approximation [3, 16, 17].

In this work, we begin by developing the basic properties of the variation space and variation norm. Specifically, we show that the set $\mathscr{K}(\mathbb{D})$ is a Banach space with the $\|\cdot\|_{\mathbb{D}}$-norm. Next, we study the variation space $\mathscr{K}(\mathbb{D})$ for the following two dictionaries which arise in the study of shallow neural networks and compare them with related notions in the approximation theory of shallow neural networks.

The first type of dictionary arises when studying neural networks with ReLU^k activation function [35]

$$\sigma_k(x) = \text{ReLU}^k(x) := [\max(0, x)]^k.$$

Here when $k = 0$, we interpret $\sigma_k(x)$ to be the Heaviside function. Let $\Omega \subset \mathbb{R}^d$ be a compact domain and consider the dictionary

$$\mathbb{P}_k = \{\sigma_k(\omega \cdot x + b) : \ \omega \in S^{d-1}, \ b \in [c_1, c_2]\} \subset L^2(\Omega), \tag{1.2}$$

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

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

The important point is that $\sigma_k(\omega \cdot x + b)$ for all planes $\omega \cdot x + b = 0$ which intersect Ω must be strictly contained in $\mathbb{P}_k$. Note that we are suppressing the dependence on the domain Ω and dimension d for notational convenience. We explain in more detail where this definition comes from in Sect. 3.

Recently, neural networks with ReLU activation function have shown remarkable empirical success on problems in computer vision and natural language processing [18]. A shallow neural network of width n with ReLU activation function is a function of the form

$$f_n(x) = \sum_{i=1}^{n} a_i \sigma_1(\omega_i \cdot x + b_i),$$

for some parameters $a_i, b_i \in \mathbb{R}$ and $\omega_i \in \mathbb{R}^d$, where $\sigma_1(x) = \max(0, x)$ is the rectified linear unit [25]. A natural measure of complexity on the parameters a_i, b_i, ω_i is the squared ℓ^2-norm

$$C(f_n) := C(\{a_i, b_i, \omega_i\}_{i=1}^n) := \sum_{i=1}^n a_i^2 + \|\omega_i\|_2^2,$$

which corresponds to the regularizer induced by the common practice of weight decay [15]. In [26], a semi-norm is defined by taking the complexity required to uniformly approximate f on compact subsets as the width $n \to \infty$. Specifically, they define the semi-norm

$$\bar{R}(f) = \liminf_{\epsilon \to 0} \left\{ C(f_n) \text{ s.t. } |f_n(x) - f(x)| \le \epsilon, \text{ for } |x| \le \epsilon^{-1} \right\}.$$

Functions for which $\bar{R}(f)$ is finite can be approximated arbitrarily closely by shallow ReLU neural networks with bounded complexity. It is shown in [26] that this semi-norm is given by

$$\bar{R}(f) = \min \left\{ \|\alpha\|_1, \text{ s.t. } f(x) = \int_{S^{d-1} \times \mathbb{R}} [\sigma_1(\omega \cdot x + b) - \sigma_1(b)] \mathrm{d}\alpha(\omega, b) + c \right\},$$

where the infemum is taken over all signed Borel measures α and constants c, and $\|\alpha\|_1$ denotes the total variation norm of α. Further, they provide a characterization of the semi-norm $\bar{R}$ in terms of the Radon transform, showing that (roughly speaking, see [26], Theorem 2)

$$\bar{R}(f) = \gamma_d \|\mathscr{R}(\Delta^{\frac{d+1}{2}} f)\|_1 + |\nabla f(\infty)|, \tag{1.3}$$

where $\mathscr{R}$ is the Radon transform, γ_d is a dimension dependent constant, and the gradient at ∞ is defined by $\nabla f(\infty) = \lim_{r \to \infty} \frac{1}{r^{d-1}|S^{d-1}|} \int_{|x|=r} \nabla f(x) dx$. Note that one part of this equality, namely that the left hand side is less than or equal to the right, was also proved in [29]. When d is even, the fractional power of the Laplacian appearing in (1.3) must be defined in terms of the ramp filter in the Radon domain (see [26, 27] for details).

This notion is extended the higher powers of the ReLU, i.e. to $\sigma_k = [\max(0, x)]^k$ for $k \ge 2$ in [27, 28] (the case $k = 0$ was treated in [13]). They propose a family of seminorms, called the Radon BV semi-norms, denoted by $|f|_{(k)}$, indexed by k and defined via the Radon transform $\mathscr{R}$ (again, roughly speaking, see [27] for details) as

$$|f|_{(m)} = \gamma_d \|\mathscr{R}(\Delta^{\frac{d+2k-1}{2}} f)\|_1.$$

Further, they prove a representer theorem for this semi-norm, i.e. they show that minimizers to the regularized problem

$$\arg\min_f \sum_{i=1}^{N} \ell(f(x_i), y_i) + \gamma |f|_{(m)},$$

where $(x_1, y_1), ..., (x_N, y_N)$ is a finite data sample are shallow neural networks with ReLU^k activation function and finite width (depending on N).

A closely related notion introduced recently concerning approximation by shallow ReLU neural networks is the Barron norm [7, 40], defined by

$$\|f\|_{\mathscr{B}} = \min\left\{\mathbb{E}_\rho(|a|(|\omega|_1 + |b|)) : \ f(x) = \int_{\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}} a\sigma_1(\omega\cdot x + b)\rho(\mathrm{d}a, \mathrm{d}\omega, \mathrm{d}b)\right\},$$

where the infemum is over all probability measures ρ on $\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}$. The properties of this norm have also been studied in [8], for instance.

Our contribution is to show that on a bounded domain Ω, the notion of Barron norm coincides with the variation norm of the dictionary $\mathbb{P}_1$. Specifically, we show the equivalence

$$\|f\|_{\mathbb{P}_1} \eqsim \|f\|_{\mathscr{B},\Omega} := \inf_{f_e|\Omega=f} \|f_e\|_{\mathscr{B}}.$$

Here the constant in the above bound sclaes with the square root of the dimension and is due to the fact that the Barron norm measures the norm of ω in ℓ^1 while we measure it in ℓ^2.

In addition, we show that up to a polynomial kernel, the Radon BV semi-norm coincides with the variation norm of $\mathbb{P}_k$ on a bounded domain Ω. Specifically, we have for $k \geq 0$

$$\inf_{p\in\mathscr{P}_k} \|f + p\|_{\mathbb{P}_k} \eqsim |f|_{(k+1),\Omega} := \inf_{f_e|\Omega=f} |f_e|_{(k+1)}, \tag{1.4}$$

where $\mathscr{P}_k$ is the space of polynomials of degree at most k and the infimum is taken over all extensions f_e of f to the whole of $\mathbb{R}^d$.

By equivalence of the $\bar{R}$ semi-norm and Radon BV semi-norm noted in [27] this also implies that

$$\inf_{p\in\mathscr{P}_k} \|f + p\|_{\mathbb{P}_1} \eqsim \inf_{f_e|\Omega=f} \bar{R}(f_e). \tag{1.5}$$

The constants implicit in the equivalences (1.4) and (1.5) do not depend upon the dimension. We also prove the equivalences

$$\|f\|_{\mathbb{P}_k} \eqsim |f|_{(k+1),\Omega} + \|f\|_{L^2(\Omega)}, \ \|f\|_{\mathbb{P}_1} \eqsim \inf_{f_e|\Omega=f} \bar{R}(f) + \|f\|_{L^2(\Omega)}.$$

However, for these the implied constant does depend upon the dimension.

Finally, we also give a complete characterization of the variation space corresponding to $\mathbb{P}_k$ in one dimension. In particular, we prove that

$$\|f\|_{\mathbb{P}_k} \eqsim \sum_{j=0}^{k-1} |f^{(j)}(-1)| + \|f^{(k)}\|_{BV([-1,1])}.$$

The second type of dictionary related to shallow neural networks which we consider is the spectral dictionary of order $s \geq 0$, given by

$$\mathbb{F}_s = \{(1+|\omega|)^{-s} e^{2\pi i \omega \cdot x} : \omega \in \mathbb{R}^d\} \subset L^2(\Omega).$$

Our contribution is to show that the variation space of $\mathbb{F}_s$ can be completely characterized in terms of the Fourier transform. In particular, in Sect. 5 we prove that

$$\|f\|_{\mathbb{F}_s} = \inf_{f_e|_\Omega = f} \int_{\mathbb{R}^d} (1+|\xi|)^s |\hat{f}_e(\xi)| d\xi, \tag{1.6}$$

where the infimum is taken over all extensions $f_e \in L^1(\mathbb{R}^d)$. (Note that we have here equality, not just equivalence.)

The norm in (1.6) was first introduced by Barron [1], who showed that functions in the space $\mathscr{K}(\mathbb{F}_1)$ could be approximated with rate $O(n^{-\frac{1}{2}})$ using shallow networks with sigmoidal activation function. These results have been extended to networks with ReLU^k activation functions in [14, 41]. The spectral Barron norm (1.6) has also been important in understanding the approximation properties of shallow neural networks with more general activation functions [11, 33]. Our contribution is to show that the spectral Barron norm (1.6) is equivalent to the variation norm with respect to a suitable dictionary of decaying Fourier modes.

The paper is organized as follows. In Sect. 2 we discuss the basic properties of the variation spaces. In particular, show that they are Banach spaces. In Sect. 3, we analyze the spaces $\mathscr{K}(\mathbb{P}_k)$. We show that when $k = 1$, the space is equivalent to the Barron space studied in [7] and also compare them with the Radon BV spaces. In Sect. 4, we give a characterization of $\mathscr{K}(\mathbb{P}_k)$ when $d = 1$ in terms of the space of bounded variation. Then, in Sect. 5, we give a characterization of $\mathscr{K}(\mathbb{F}_s^d)$ in terms of the Fourier transform, showing that it is equivalent to the spectral Barron norm. Finally, we give some concluding remarks and further research directions.

2 Basic Properties of $\mathscr{K}_1(\mathbb{D})$

Let us first develop the elementary properties of the variation space $\mathscr{K}(\mathbb{D})$. The key result is that $\mathscr{K}(\mathbb{D})$ is a Banach space with the variation norm $\|\cdot\|_{\mathbb{D}}$.

Lemma 1 *Suppose that* $\sup_{d\in\mathbb{D}} \|d\|_H = K_{\mathbb{D}} < \infty$. *Then the* $\|\cdot\|_{\mathbb{D}}$ *norm satisfies the following properties.*

- $\overline{conv(\pm\mathbb{D})} = \{f \in H : \|f\|_{\mathbb{D}} \leq 1\}$
- $\|f\|_H \leq K_{\mathbb{D}}\|f\|_{\mathbb{D}}$
- $\mathscr{K}(\mathbb{D}) := \{f \in H : \|f\|_{\mathbb{D}} < \infty\}$ *is a Banach space with the* $\|\cdot\|_{\mathbb{D}}$ *norm*

Proof The first two statements are well-known and can be found for instance in [16]. For the third statement we must show that the set $\mathscr{K}(\mathbb{D})$ is complete with respect to the $\|\cdot\|_{\mathbb{D}}$ norm.

Let $\{f_n\}_{n=1}^{\infty}$ be a Cauchy sequence with respect to the $\|\cdot\|_{\mathbb{D}}$ norm. From the second statement, we have $\|f_n - f_m\|_H \leq K_{\mathbb{D}}\|f_n - f_m\|_{\mathbb{D}}$, so that the sequence is Cauchy with respect the the H-norm as well. Thus, there exists an $f \in H$ such that $f_n \to f$ in H, i.e. such that $\|f_n - f\|_H \to 0$.

We will show that also $\|f_n - f\|_{\mathbb{D}} \to 0$, i.e. that we have convergence in the variation norm as well (note that this automatically implies that $\|f\|_{\mathbb{D}} < \infty$).

To this end, let $\epsilon > 0$ and choose N such that $\|f_n - f_m\|_{\mathbb{D}} < \epsilon/2$ for $n, m \geq N$ ($\{f_n\}$ is Cauchy, so this is possible). In particular, this means that $\|f_N - f_m\|_{\mathbb{D}} \leq \epsilon/2$ for all $m > N$. Now the first statement implies that $f_m - f_N \in (\epsilon/2)\overline{\text{conv}(\pm\mathbb{D})}$, or in other words that $f_m \in f_N + (\epsilon/2)\overline{\text{conv}(\pm\mathbb{D})}$. Since $f_m \to f$ in H, and $\overline{\text{conv}(\pm\mathbb{D})}$ is closed in H by definition, we get $f \in f_N + (\epsilon/2)\overline{\text{conv}(\pm\mathbb{D})}$. Hence $\|f - f_N\|_{\mathbb{D}} \leq \epsilon/2$ and the triangle inequality finally implies that $\|f - f_m\|_{\mathbb{D}} \leq \epsilon$ for all $m \geq N$. Thus $f_n \to f$ in the variation norm and $\mathscr{K}(\mathbb{D})$ is complete. □

Let us remark that for some dictionaries $\mathbb{D}$ the $\mathscr{K}(\mathbb{D})$ space can be substantially smaller than H. In fact, if the dictionary $\mathbb{D}$ is contained in a closed subspace of H, then we have the following elementary result.

Lemma 2 *Let $K \subset H$ be a closed subspace of H. Then $\mathbb{D} \subset K$ iff $\mathscr{K}_1(\mathbb{D}) \subset K$.*

Proof We have $\mathbb{D} \subset \mathscr{K}(\mathbb{D})$ so that the reverse implication is trivial. For the forward implication, since $\mathbb{D} \subset K$ and K is closed, it follows that $\overline{\text{conv}(\pm\mathbb{D})} \subset K$. Then, from the definition (1.1), it follows that

$$\mathscr{K}(\mathbb{D}) = \bigcup_{r>0} r \cdot \overline{\text{conv}(\pm\mathbb{D})} \subset K.$$

□

A simple example when this occurs is when considering a shallow neural network with activation function σ which is a polynomial of degree k. In this case the space $\mathscr{K}(\mathbb{D})$ is contained in the finite-dimensional space of polynomials of degree k, and the $\|\cdot\|_{\mathbb{D}}$ norm is infinite on non-polynomial functions. This is related to the well-known result that neural network functions are dense iff the activation function is not a polynomial [19].

Proposition 1 *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain and $\mathbb{D} = \{\sigma(\omega \cdot x + b) : (\omega, b) \in \mathbb{R}^d \times \mathbb{R}\} \subset L^2(\Omega)$, where the activation function $\sigma \in L^{\infty}_{loc}(\mathbb{R})$, i.e. $\|\sigma\|_{L^{\infty}(K)} < \infty$ for any compact set $K \subset \mathbb{R}$. Suppose further that the set of discontinuities of σ has Lebesgue measure 0. Then $\mathscr{K}(\mathbb{D})$ is finite dimensional iff σ is a polynomial (a.e.).*

Proof If σ is a polynomial, $\mathbb{D}$ is contained in the space of polynomials of degree at most $\deg(\sigma)$, which is finite dimensional. This implies the result by Lemma 2. For the reverse implication, we use Theorem 1 of [19], which states that if σ is not a polynomial, then

$$C(\Omega) \subset \overline{\left\{ \sum_{i=1}^{n} a_i \sigma(\omega_i \cdot x + b_i) \right\}},$$

where the closure is taken in $L^\infty(\Omega)$ (note that this cumbersome statement is necessary since σ may not be continuous). This immediately implies that $\mathscr{K}(\mathbb{D})$ is dense in $L^2(\Omega)$ (since $C(\Omega)$ is dense in $L^2(\Omega)$), and thus obviously not finite dimensional. □

While in this example the variation norm is finite dimensional, this is typically not the case for most dictionaries of interest. Specifically for the dictionaries $\mathbb{P}_k$ and $\mathbb{F}_s$ which we study, this space is infinite dimensional but still much smaller than H. The size of the variation space has been precisely quantified for $\mathbb{P}_k$ in terms of the metric entropy in [35] and these spaces have been used as trial spaces for solving PDEs in [10]. However, it remains an interesting open question what the practical utility of $\mathscr{K}(\mathbb{P}_k)$ and $\mathscr{K}(\mathbb{F}_s)$ really are.

The significance of the variation norm is that functions $f \in \mathscr{K}(\mathbb{D})$ can be efficiently approximated by convex combinations of small numbers of dictionary elements. In particular, we have the following result of Maurey [1, 12, 30]. Denote

$$\Sigma_{n,M}(\mathbb{D}) = \left\{ \sum_{j=1}^{n} a_j h_j : h_j \in \mathbb{D}, \ \sum_{i=1}^{n} |a_i| \le M \right\}.$$

Then we have the following result.

Theorem 1 *(Lemma 1 in [1]) Suppose that $f \in \mathscr{K}(\mathbb{D})$. Then for $M = \|f\|_{\mathbb{D}}$ we have*

$$\inf_{f_n \in \Sigma_{n,M}(\mathbb{D})} \|f - f_n\|_H \le K_{\mathbb{D}} \|f\|_{\mathscr{K}(\mathbb{D})} n^{-\frac{1}{2}},$$

We note the following simple converse to Maurey's approximation rate. In particular, if a function can be approximated by elements from $\Sigma_{n,M}(\mathbb{D})$ with fixed M, then it must be in the space $\mathscr{K}(\mathbb{D})$.

Proposition 2 *Let H be a Hilbert space and $f \in H$. Suppose that $f_n \to f$ in H with $f_n \in \Sigma_{n,M}(\mathbb{D})$ for a fixed $M < \infty$. Then $f \in \mathscr{K}(\mathbb{D})$ and*

$$\|f\|_{\mathbb{D}} \le M.$$

Proof It is clear that we must only prove this for $M = 1$. From the definitions we have $\Sigma_{n,1}(\mathbb{D}) \subset \overline{\text{conv}(\pm\mathbb{D})}$ for every n. Thus $f_n \in \overline{\text{conv}(\pm\mathbb{D})}$ and since $\overline{\text{conv}(\pm\mathbb{D})}$ is closed, we get $f \in \overline{\text{conv}(\pm\mathbb{D})}$, so that $\|f\|_{\mathbb{D}} \le 1$, as desired. □

Next, we wish to connect the space $\mathscr{K}(\mathbb{D})$ defined via the closed symmetric convex hull of $\mathbb{D}$ to integral representations, which have recently become a popular concept in the approximation theory of shallow neural networks [8, 26, 27, 39]. An integral representation of a function f over the dictionary $\mathbb{D}$ is given by

$$f = \int_{\mathbb{D}} i_{\mathbb{D}\to H} \mathrm{d}\mu. \tag{2.1}$$

Here the dictionary $\mathbb{D}$ inherits the subspace topology from the Hilbert space H, $d\mu$ is a (signed) Borel measure with finite variation on $\mathbb{D}$, i.e.

$$\|\mu\| = \sup_{\substack{g\colon \mathbb{D}\to[-1,1]\\ g \text{ measurable}}} \int_{\mathbb{D}} g \mathrm{d}\mu < \infty,$$

and the integral is the Bochner integral of the inclusion map $i_{\mathbb{D}\to H} : \mathbb{D} \to H$. Note that since H is separable, the inclusion map is μ-measurable by the Pettis measurability theorem. Further, if $\mathbb{D}$ is bounded, i.e. if $|\mathbb{D}| = \sup_{d\in\mathbb{D}} \|d\|_H < \infty$, then since μ has finite variation the inclusion map $i_{\mathbb{D}\to H}$ is absolutely integrable and so the Bochner integral exists (see [5], Chapter 4).

We prove that if the dictionary $\mathbb{D}$ is compact, then membership in $\mathscr{K}(\mathbb{D})$ is equivalent to the existence of an integral representation.

Lemma 3 *Suppose that $\mathbb{D} \subset H$ is compact. Then $f \in \mathscr{K}(\mathbb{D})$ iff there exists a Borel measure μ on $\mathbb{D}$*

$$f = \int_{\mathbb{D}} i_{\mathbb{D}\to H} \mathrm{d}\mu.$$

Moreover,

$$\|f\|_{\mathbb{D}} = \inf\left\{\|\mu\| : \ f = \int_{\mathbb{D}} i_{\mathbb{D}\to H} \mathrm{d}\mu\right\}.$$

Proof From the definition of the variation norm (1.1) we must show that

$$\overline{\mathrm{conv}(\pm\mathbb{D})} = M(\mathbb{D}) := \left\{\int_{\mathbb{D}} i_{\mathbb{D}\to H} \mathrm{d}\mu : \ \|\mu\| \le 1\right\}.$$

We first show that $M(\mathbb{D}) \subset \overline{\mathrm{conv}(\pm\mathbb{D})}$. The idea of the proof is to approximate the inclusion map $i_{\mathbb{D}\to H}$ by a simple function. The only technical issue is that we must be able to restrict the range of this simple function to lie in $\mathbb{D}$. We proceed as in the proof of Bochner's theorem (see [5], Chapter IV) with minor modification.

Let μ be a Borel measure on $\mathbb{D}$ with variation $\|\mu\| \le 1$. Since H is separable, the Pettis measurability theorem implies that the inclusion $i_{\mathbb{D}\to H}$ is μ-measurable.

So for each n we can choose a countably valued μ-measurable function f_n such that $\|f_n - i_{\mathbb{D}\to H}\|_H \leq 1/2n$ μ-almost everywhere. Thus we can write

$$f_n = \sum_{k=1}^{\infty} a_{n,k}\chi_{E_{n,k}}$$

for elements $a_{n,k} \in H$ and μ-measurable sets $E_{n,k}$ which satisfy $E_{n,i} \cap E_{n,j} = \emptyset$ when $i \neq j$. The condition $\|f_n - i_{\mathbb{D}\to H}\|_H \leq 1/2n$ means that for every $d \in E_{n,k} \subset \mathbb{D}$ we have $\|a_{n,k} - d\|_H \leq 1/2n$. Using the triangle inequality this means that for any $d, d' \in E_{n,k}$, we have $\|d - d'\|_H \leq 1/n$. Now for each $E_{n,k}$ we choose $d_{n,k} \in E_{n,k}$ and set

$$\tilde{f}_n = \sum_{k=1}^{\infty} d_{n,k}\chi_{E_{n,k}}.$$

Then we have $\|\tilde{f}_n - i_{\mathbb{D}\to H}\|_H \leq 1/n$ μ-almost everywhere and the range of $\tilde{f}_n$ lies in H. Finally, for each n we choose p_n such that

$$\int_{\left(\cup_{k=p_n+1}^{\infty} E_{n,k}\right)} \|\tilde{f}_n\|_H \mathrm{d}\mu \leq \frac{1}{n}.$$

Since $\mathbb{D}$ is compact and thus bounded and μ satisfies $\|\mu\| \leq 1$, the function $\|\tilde{f}_n\|_H$ is in $L^1(d\mu)$ so that such a p_n can always be chosen. We now set

$$g_n = \sum_{k=1}^{p_n} d_{n,k}\chi_{E_{n,k}}.$$

Then g_n is a simple function satisfying $\int_{\mathbb{D}} \|i_{\mathbb{D}\to H} - g_n\|_H d\mu \leq \frac{2}{n}$. This means that

$$\left| \int_{\mathbb{D}} i_{\mathbb{D}\to H} d\mu - \sum_{k=1}^{p_n} \mathrm{d}_{n,k}\mu(E_{n,k}) \right| \leq \frac{2}{n}.$$

By design, $d_{n,k} \in \mathbb{D}$ and since $\|\mu\| \leq 1$, we get $\sum_{k=1}^{p_n} |\mu(E_{n,k})| \leq 1$. Thus

$$\sum_{k=1}^{p_n} \mathrm{d}_{n,k}\mu(E_{n,k}) \in \overline{\mathrm{conv}(\pm\mathbb{D})}$$

for every n. Letting $n \to \infty$, we see that

$$\int_{\mathbb{D}} i_{\mathbb{D}\to H} \mathrm{d}\mu \in \overline{\mathrm{conv}(\pm\mathbb{D})}.$$

Since μ was an arbitrary measure we get $M(\mathbb{D}) \subset \overline{\mathrm{conv}(\pm\mathbb{D})}$.

Next we prove the reverse inclusion. Given any convex combination

$$f = \sum_{i=1}^{N} a_i d_i,$$

with $d_i \in \mathbb{D}$ and $\sum_{i=1}^{N} |a_i| \leq 1$, we can choose $\mu = \sum_{i=1}^{\infty} a_i \delta_{d_i}$ to be a linear combination of Dirac deltas to see that $f \in M(\mathbb{D})$. To complete the proof we must show that $M(\mathbb{D})$ is closed. We will prove this using Prokhorov's theorem [31] (see also [6], Theorem 11.5.4, for instance). Let $f_n \to f$ with $f_n \in M(\mathbb{D})$ and let μ_n be the corresponding sequence of Borel measures on $\mathbb{D}$ such that

$$f_n = \int_{\mathbb{D}} i_{\mathbb{D}\to H} \mathrm{d}\mu_n$$

and $\|\mu_n\| \leq 1$. By the compactness of $\mathbb{D}$ and Prokhorov's theorem, by taking a subsequence if necessary we may assume that the $\mu_n \to \mu$ weakly, i.e. that the integrals against continuous functions on $\mathbb{D}$ converges. Set $\tilde{f} = \int_{\mathbb{D}} i_{\mathbb{D}\to H} d\mu$, which is Bochner integrable by the comments prior to the lemma. Choose a countable dense sequence $\{\lambda_i\}_{i=1}^{\infty} \in H$. The weak convergence implies that

$$\lim_{n\to\infty} \langle \lambda_i, f_n \rangle_H = \left\langle \lambda_i, \tilde{f} \right\rangle_H.$$

for every i. The strong convergence $f_n \to f$ implies the same with f replacing $\tilde{f}$. Thus $\langle \lambda_i, f \rangle_H = \langle \lambda_i, \tilde{f} \rangle_H$ for all i. Hence $f = \tilde{f} \in M(\mathbb{D})$, as desired. □

Finally, we note that the compactness in the preceding theorem was necessary. Indeed, we have the following simple example.

Proposition 3 *Suppose that $\Omega \subset \mathbb{R}^d$ is bounded and σ is a smooth sigmoidal function. Let $H = L^2(\Omega)$ and $\mathbb{D}_\sigma$ defined by*

$$\mathbb{D}_\sigma = \{\sigma(\omega \cdot x + b),\ \omega \in \mathbb{R}^d,\ b \in \mathbb{R}\}.$$

Then $\overline{conv(\pm\mathbb{D}_\sigma)} \supsetneq M(\mathbb{D}_\sigma)$, where $M(\mathbb{D}_\sigma)$ is defined as in the proof of the previous lemma.

Proof Let σ_0 be the Heaviside activation function. Then we have

$$\lim_{r\to\infty} \|\sigma_0(x_1) - \sigma(r x_1)\|_H = 0,$$

since σ is sogmoidal. Thus $\sigma_0(x_1) \in \overline{\mathrm{conv}(\pm\mathbb{D}_\sigma)}$. However, since σ is smooth, the discontinuous function $\sigma_0(x_1)$ cannot have an integral representation of the form (2.1), so that $\sigma_0(x_1) \notin M(\mathbb{D}_\sigma)$. □

3 Properties of $\mathscr{K}(\mathbb{P}_k^d)$ and Relationship with the Barron and Badon BV Spaces

In this section we study the space $\mathscr{K}(\mathbb{P}_k)$ in more detail. We begin by explaining the precise definition (1.2), i.e. how we define an appropriate dictionary corresponding to the ReLU^k activation function. The problem with letting $\sigma_k(x) = [\max(0, x)]^k$ and setting

$$\mathbb{D} = \{\sigma_k(\omega \cdot x + b) : \ \omega \in \mathbb{R}^d, \ b \in \mathbb{R}\},$$

is that unless $k = 0$ the dictionary elements are not bounded in $L^2(\Omega)$, since σ_k is not bounded and we can shift b arbitrarily. This manifests itself in the fact that $\|\cdot\|_{\mathscr{K}_1(\mathbb{D})}$ is a semi-norm which contains the set of polynomials of degree at most $k-1$ in its kernel (this occurs since the arbirtrarily large elements in $\mathbb{D}$ are polynomials on the domain Ω).

We rectify this issue by considering the dictionary

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

where c_1 and c_2 are chosen to satisfy

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

This has the effect of ensuring that the dictionary $\mathbb{P}_k$ is bounded and the constants c_1 and c_2 are chosen so that $\sigma_k(\omega \cdot x + b) \in \mathbb{P}_k$ whenever the hyperplane $\{\omega \cdot x + b = 0\}$ intersects Ω. Further, when $c_1 < b < \inf\{x \cdot \omega : x \in \Omega, \omega \in S^{d-1}\}$ or $\sup\{x \cdot \omega : x \in \Omega, \omega \in S^{d-1}\} < b < c_2$, we recover all polynomials of degree at most k on Ω as well.

Next, we consider the relationship between $\mathscr{K}(\mathbb{P}_1)$ and the Barron norm introduced in [7], which is given by

$$\|f\|_{\mathscr{B}} = \inf\left\{\mathbb{E}_\rho(|a|(|\omega|_1 + |b|)) : \ f(x) = \int_{\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}} a\sigma_1(\omega \cdot x + b)\rho(\mathrm{d}a, \mathrm{d}\omega, \mathrm{d}b)\right\},$$

where we recall that σ_1 is the rectified linear unit and the infimum is taken over all integral representations of f. Here ρ is a probability distribution on $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$, and the expectation is taken with respect to ρ. We show that the $\mathscr{K}(\mathbb{P}_1)$ space is equivalent to the Barron space when restricted to bounded domains Ω.

Proposition 4 *For any bounded domain Ω, we have*

$$\sqrt{d}\|f\|_{\mathbb{P}_1} \eqsim \inf_{f_e|_\Omega = f} \|f_e\|_{\mathscr{B}},$$

where the infemum is taken over all extensions of f to the whole of $\mathbb{R}^d$. Here the implied constant depends only upon the constants c_1 and c_2 taken in the definition of $\mathbb{P}_1$.

Proof Consider the dictionary

$$\mathbb{B} = \{(|\omega|_1 + |b|)^{-1}\sigma_1(\omega \cdot x + b) : \ \omega \in \mathbb{R}^d, \ b \in \mathbb{R}\} \subset L^2(\Omega).$$

From Lemma 3, it follows that $\|f\|_{\mathbb{B}} = \|f\|_{\mathscr{B}}$. Indeed, by making the change of variables $\mu = |a|(|\omega|_1 + |b|)\rho$, we get

$$\|f\|_{\mathscr{B}} = \inf\left\{\|\mu\| : \ f = \int_{\mathbb{B}} i_{\mathbb{B}\to L^2(\Omega)} d\mu\right\}.$$

Thus, it suffices to show that $\mathbb{P}_1^d \subset C\sqrt{d} \cdot \overline{\text{conv}(\pm\mathbb{B})}$ and $\mathbb{B} \subset C \cdot \overline{\text{conv}(\pm\mathbb{P}_1)}$ for a constant $C(c_1, c_2)$.

So let $g \in \mathbb{P}_1^d$. This means that $g(x) = \sigma_1(\omega \cdot x + b)$ for some $\omega \in S^{d-1}$ and $b \in [-c_1, c_2]$. Thus

$$(|\omega|_1 + |b|) \leq (\sqrt{d} + \max(c_1, c_2)) \leq C(c_1, c_2)\sqrt{d}$$

and since $(|\omega|_1 + |b|)^{-1}\sigma_1(\omega \cdot x + b) \in \mathbb{B}$, we see that $g \in C\sqrt{d} \cdot \overline{\text{conv}(\pm\mathbb{B})}$.

Now, let $g \in \mathbb{B}$. Then $g(x) = (|\omega|_1 + |b|)^{-1}\sigma_1(\omega \cdot x + b)$ for some $\omega \in \mathbb{R}^d$ and $b \in \mathbb{R}$.

Consider first the case when $\omega \neq 0$. Note that by the positive homogeneity of σ_1 we can assume that $|\omega| = 1$, i.e. that $\omega \in S^{d-1}$. Further, we have that $(|\omega|_1 + |b|)^{-1} \leq (1 + |b|)^{-1}$. Thus, we must show that

$$\tilde{g}(x) := (1 + |b|)^{-1}\sigma_1(\omega \cdot x + b) \in C \cdot \overline{\text{conv}(\pm\mathbb{P}_1)}$$

for $\omega \in S^{d-1}$ and $b \in \mathbb{R}$. For $b \in [c_1, c_2]$ this clearly holds with $C = 1$ since $(1 + |b|)^{-1} \leq 1$ and for such values of b, we have $\sigma_1(\omega \cdot x + b) \in \mathbb{P}_1^d$. If $b < c_1$, then $\tilde{g}(x) = 0$, so we trivially have $\tilde{g} \in \overline{\text{conv}(\pm\mathbb{P}_1)}$. Finally, if $b > c_2$, then $\omega \cdot x + b$ is positive on Ω, so that

$$\tilde{g}(x) = (1 + |b|)^{-1}(\omega \cdot x + b) = (1 + |b|)^{-1}\omega \cdot x + b(1 + |b|)^{-1}.$$

Now $\omega \cdot x \in 2 \cdot \overline{\text{conv}(\pm\mathbb{P}_1)}$ and $1 = [\sigma_1(\omega \cdot x + 2) - \sigma_1(\omega \cdot x + 1)] \in 2 \cdot \overline{\text{conv}(\pm\mathbb{P}_1)}$. Combined with the above and the fact that $(1 + |b|)^{-1}, |b|(1 + |b|)^{-1} \leq 1$, we get $\tilde{g} \in 4 \cdot \overline{\text{conv}(\pm\mathbb{P}_1)}$.

Finally, if $\omega = 0$, then $g(x) = 1$ and by the above paragraph we clearly also have $g \in 2 \cdot \overline{\text{conv}(\pm\mathbb{P}_1)}$. This completes the proof. □

Note that it follows from this result that the Barron space $\mathscr{B}$ is a Banach space, which was first proven in [8].

Next, we compare the spaces $\mathscr{K}(\mathbb{P}_k)$ and their variation norms to the Radon BV semi-norms introduced in [26, 27]. Specifically, these norms coincide with the $\mathbb{P}_k$-variation norms on a bounded domain Ω up to a kernel consisting of polynomials.

Theorem 2 *For any bounded domain Ω, we have*

$$\inf_{p\in\mathscr{P}_k} \|f+p\|_{\mathbb{P}_k} = \frac{1}{k!}\inf_{f_e|\Omega=f} |f_e|_{(k+1)}$$

where $|\cdot|_{(k+1)}$ is the Radon BV seminorm introduced in [27], f_e is an extension of f to the whole of $\mathbb{R}^d$, and $\mathscr{P}_k$ is the space of polynomial of degree at most k.

Note that by the remarks in [27], when $k=1$ this theorem also applies to the semi-norm introduced in [26], which is equivalent to the Radon BV semi-norm.

Proof Theorem 22 in [27] implies that $|f|_{(k+1)} \le 1$ is equivalent to an integral representation of the form

$$f(x) = \frac{1}{k!}\int_{S^{d-1}\times\mathbb{R}} [\sigma_k(\omega\cdot x+b) - (\omega\cdot x+b)^k]\mathrm{d}\mu(\omega,b) + p(x), \tag{3.1}$$

where μ is a Borel measure on $S^{d-1}\times\mathbb{R}$, $p(x)$ is a polynomial of degree at most k and μ satisfies $\|\mu\| = 1$.

Further, Lemma 3 means that $\|f\|_{\mathbb{P}_k} \le 1$ is equivalent to the existence of an integral representation

$$f(x) = \int_{S^{d-1}\times[c_1,c_2]} \sigma_k(\omega\cdot x+b)\mathrm{d}\mu(\omega,b) \tag{3.2}$$

on Ω, where μ is a Borel measure on $S^{d-1}\times[c_1,c_2]$ and $\|\mu\| \le 1$. This follows since $\mathbb{P}_k$ is a compact subset of $L^2(\Omega)$ (it is continuously parameterized by the compact set $S^{d-1}\times[c_1,c_2]$).

So if $\|f\|_{\mathbb{P}_k} \le 1$ we use the integral representation and set (3.2) and set

$$f_e(x) = k!\left[\frac{1}{k!}\int_{S^{d-1}\times[c_1,c_2]} [\sigma_k(\omega\cdot x+b) - (\omega\cdot x+b)^k]\mathrm{d}\mu(\omega,b) + p(x)\right],$$

where

$$p(x) = \frac{1}{k!}\int_{S^{d-1}\times[c_1,c_2]} (\omega\cdot x+b)^k\mathrm{d}\mu(\omega,b).$$

Since $S^{d-1}\times[c_1,c_2] \subset S^{d-1}\times\mathbb{R}$ we see that $|f_e|_{k+1} \le k!$. This implies that $\inf_{f_e|\Omega=f}|f_e|_{(k+1)} \le k!\|f\|_{\mathbb{P}_k}$. Since $\mathscr{P}_k$ is the kernel of the $|\cdot|_{(k+1)}$ (see Lemma 19 in [27]), we can take an infemum over $\mathscr{P}_k$ to get

$$\inf_{f_e|\Omega=f}|f_e|_{(k+1)} \le k!\inf_{p\in\mathscr{P}_k}\|f+p\|_{\mathbb{P}_k}.$$

For the converse, suppose that f satisfies $\inf_{f_e|\Omega=f}|f_e|_{(k+1)} \le 1$ and let f_e be an extension of f such that

$$|f_e|_{(k+1)} \le 1+\epsilon.$$

We now apply integral representation (3.1) and note that if $b \notin [c_1, c_2]$, then $\sigma_k(\omega \cdot x + b) - (\omega \cdot x + b)^k$ is a polynomial on the domain Ω. So we can write

$$f_e(x) = f(x) = f'(x) + q(x)$$

for $x \in \Omega$, where $\|f\|_{\mathbb{P}_k} \leq 1/k!$ and $q(x)$ is a polynomial of degree at most k. This implies that

$$\inf_{p \in \mathscr{P}_k} \|f + p\|_{\mathbb{P}_k} \leq \|f'\|_{\mathbb{P}_k} = 1.$$

Hence $\inf_{p \in \mathscr{P}_k} \|f + p\|_{\mathbb{P}_k} \leq (1/k!) \inf_{f_e|\Omega = f} |f_e|_{(k+1)}$ as desired. □

As a corollary of this result, we have the following equivalence when we strengthen the Radon BV semi-norm to a norm.

Corollary 1 *For any bounded domain Ω, we have*

$$\|f\|_{\mathbb{P}_k} \eqsim \inf_{f_e|\Omega = f} |f_e|_{(k+1)} + \|f\|_{L^2(\Omega)}.$$

Note that by the remarks in [27], when $k = 1$ this corollary also applies to the semi-norm introduced in [26].

Proof By Theorem 2 we have

$$\inf_{f_e|\Omega = f} |f_e|_{(k+1)} = k! \inf_{p \in \mathscr{P}_k} \|f + p\|_{\mathbb{P}_k} \leq k! \|f\|_{\mathbb{P}_k}.$$

Further, by Lemma 1 we have $\|f\|_{L^2(\Omega)} \lesssim \|f\|_{\mathbb{P}_k}$ since the dictionary $\mathbb{P}_k$ is bounded in $L^2(\Omega)$. Putting these together, we get

$$\inf_{f_e|\Omega = f} |f_e|_{(k+1)} + \|f\|_{L^2(\Omega)} \lesssim \|f\|_{\mathbb{P}_k}.$$

To prove the other direction, let $p^* \in \mathscr{P}_k$ be such that

$$\|f - p^*\|_{\mathbb{P}_k} = \inf_{p \in \mathscr{P}_k} \|f + p\|_{\mathbb{P}_k}.$$

Such a p^* can always be found since $\mathscr{P}_k$ is a finite dimensional space and the function $\|f + p\|_{\mathbb{P}_k} \to \infty$ as $p \to \infty$. Then, using Theorem 2, we have

$$\|f\|_{\mathbb{P}_k} \leq \|f - p^*\|_{\mathbb{P}_k} + \|p^*\|_{\mathbb{P}_k} = \frac{1}{k!} \inf_{f_e|\Omega = f} |f_e|_{(k+1)} + \|p^*\|_{\mathbb{P}_k}.$$

Since $\mathscr{K}(\mathbb{P}_k)$ contains all polynomials of degree at most k on Ω, $\|\cdot\|_{\mathbb{P}_k}$ is a finite norm on $\mathbb{P}_k$. As all norms on the finite dimensional space $\mathscr{P}_k$ are equivalent we get

$$\|p^*\|_{\mathbb{P}_k} \lesssim \|p^*\|_{L^2(\Omega)}.$$

Next, we notice that

$$\|p^*\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} + \|f - p^*\|_{L^2(\Omega)}.$$

Further, $\|f - p^*\|_{L^2(\Omega)} \lesssim \|f\|_{\mathbb{P}_k} \lesssim \inf_{f_e|\Omega=f} |f_e|_{(k+1)}$ for a constant C. This follows by Lemma 1, Theorem 2 and the fact that $|p|_{(k+1)} = 0$ for any $p \in \mathscr{P}_k$ (Lemma 19 in [27]). It follows that

$$\|p^*\|_{\mathbb{P}_k} \lesssim \|p^*\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} + \|f - p^*\|_{L^2(\Omega)} \lesssim \inf_{f_e|\Omega=f} |f_e|_{(k+1)} + \|f\|_{L^2(\Omega)},$$

so we finally get

$$\|f\|_{\mathbb{P}_k} \leq \frac{1}{k!} \inf_{f_e|\Omega=f} |f_e|_{(k+1)} + \|p^*\|_{\mathbb{P}_k} \lesssim \inf_{f_e|\Omega=f} |f_e|_{(k+1)} + \|f\|_{L^2(\Omega)}.$$

□

4 Characterization of $\mathscr{K}(\mathbb{P}_k)$ in One Dimension

In this section, we prove a characterization of $\mathscr{K}(\mathbb{P}_k)$ in one dimension. In this case, the space $\mathscr{K}(\mathbb{P}_k)$ has a relatively simple characterization in terms of the space of bounded variation. In the case where $k = 1$ an analogous characterization can be found in [8], sect. 4. Earlier results characterizing the Barron space in one dimension on the while of $\mathbb{R}$ were obtained in [20, 32]. Note that by the results of the previous section, a higher dimensional characterization in terms of the Radon transform is given in [26, 27].

Theorem 3 *Let $\Omega = [-1, 1]$. We have*

$$\mathscr{K}(\mathbb{P}_k) = \{f \in L^2([-1, 1]) : \ f \text{ is } k\text{-times differentiable a.e. and } f^{(k)} \in BV([-1, 1])\}.$$

In particular, it holds that

$$\|f\|_{\mathbb{P}_k} \eqsim \sum_{j=0}^{k-1} |f^{(j)}(-1)| + \|f^{(k)}\|_{BV([-1,1])}.$$

Proof We first prove that

$$\|f\|_{\mathbb{P}_k} \lesssim \sum_{j=0}^{k-1} |f^{(j)}(-1)| + \|f^{(k)}\|_{BV([-1,1])}. \tag{4.1}$$

Note that the right hand side is uniformly bounded for all $f = \sigma_k(\pm x + b) \in \mathbb{P}_k$, since $\sigma_k^{(k)}$ is a multiple of the Heaviside function and b is bounded by $\max(|c_1|, |c_2|)$.

By taking convex combinations, this means that for some constant C, we have

$$\left\{ \sum_{j=1}^{n} a_j h_j : \ h_j \in \mathbb{P}_k^1, \ \sum_{i=1}^{n} |a_i| \leq 1 \right\} \subset C B_{BV,k}^1,$$

where

$$B_{BV,k}^1 := \left\{ f \in L^2([-1,1]) : \ \sum_{j=0}^{k-1} |f^{(j)}(-1)| + \|f^{(k)}\|_{BV([-1,1])} \leq 1 \right\}.$$

It is well-known that $B_{BV,k}^1$ is compact in $L^1([-1,1])$ (see, for instance Theorem 4 of Chapter 5 in [9]). This implies that $B_{BV,k}^1$ is closed in $L^2([-1,1])$, since if $f_n \to_{L^2} f$ with $f_n \in B_{BV,k}^1$, then there must exist a subsequence $f_{k_n} \to_{L^1} \tilde{f} \in B_{BV,k}^1$. Clearly $f = \tilde{f}$ and so $B_{BV,k}^1$ is closed in $L^2([-1,1])$. From this it follows that $\overline{\text{conv}(\pm\mathbb{P}_k)} \subset C B_{BV,k}^1$ and we obtain (4.1).

Next, we prove the reverse inequality. So let $f \in B_{BV,k}^1$. By Theorem 2 in Chapter 5 of [9], there exist $f_n \in C^\infty \cap B_{BV,k}^1$ such that $f_n \to f$ in $L^1([-1,1])$. Further, since $f_n, f \in B_{BV,k}^1$, we have that $\|f - f_n\|_{L^\infty([-1,1])}$ is uniformly bounded. Thus

$$\|f - f_n\|_{L^2([-1,1])}^2 \leq \|f - f_n\|_{L^1([-1,1])} \|f - f_n\|_{L^\infty([-1,1])} \to 0$$

and so $f_n \to f$ in $L^2([-1,1])$ as well.

Using the Peano kernel formula, we see that

$$f_n(x) = \sum_{j=0}^{k} \frac{f_n^{(j)}(-1)}{j!}(x+1)^j + \int_{-1}^{1} \frac{f_n^{(k+1)}(b)}{k!} \sigma_k(x-b) db.$$

From the definition of the BV-norm and the fact that $f_n \in B_{BV,k}^1$, we see that

$$\sum_{j=0}^{k} \frac{|f_n^{(j)}(-1)|}{j!} + \int_{-1}^{1} \frac{|f_n^{(k+1)}(b)|}{k!} db \leq C_1$$

for a fixed constant C_1. Choose $k+1$ distinct $b_1, ..., b_{k+1} \in [1, c_2]$ (note that we need $c_2 > 1$). Then by construction $\sigma_k(x + b_i) = (x + b_i)^k$ is a polynomial on $[-1,1]$. Moreover, it is well-known that the polynomials $(x + b_i)^k$ span the space of polynomials of degree at most k (using for instance the determinant of Vandermonde matrix). Combined with the coefficient bound

$$\sum_{j=0}^{k} \frac{|f_n^{(j)}(-1)|}{j!} \leq C_1,$$

we see that

$$\sum_{j=0}^{k} \frac{f_n^{(j)}(-1)}{j!}(x-a)^j \in C_2 \cdot \overline{\mathrm{conv}(\pm\mathbb{P}_k)}$$

for a fixed constant C_2 (independent of f_n). Furthermore, since also

$$\int_{-1}^{1} \frac{|f_n^{(k+1)}(b)|}{k!} db \leq C_1,$$

we obtain

$$\int_{-1}^{1} \frac{f_n^{(k+1)}(b)}{k!} \sigma_k(x-b) db \in C_1 \cdot \overline{\mathrm{conv}(\pm\mathbb{P}_k)}.$$

This implies that $f_n \in C \cdot \overline{\mathrm{conv}(\pm\mathbb{P}_k)}$ for $C = C_1 + C_2$ and since $f_n \to f$ and $\overline{\mathrm{conv}(\pm\mathbb{P}_k)}$ is closed in $L^2([-1,1])$, we get $f \in C \cdot \overline{\mathrm{conv}(\pm\mathbb{P}_k)}$, which completes the proof. □

5 Characterization of $\mathscr{K}(\mathbb{F}_s^d)$

In this section we characterize the space $\mathscr{K}(\mathbb{F}_s)$ and the variation norm corresponding to the dictionary $\mathbb{F}_s$. In particular, we have that this variation norm is equivalent to the spectral Barron norm which has been widely used in the approximation theory of shallow neural networks [1, 14, 33–35].

Theorem 4 *We have*

$$\|f\|_{\mathbb{F}_s^d} = \inf_{f_e|_\Omega = f} \int_{\mathbb{R}^d} (1+|\xi|)^s |\hat{f}_e(\xi)| d\xi, \tag{5.1}$$

where the infimum is taken over all extensions $f_e \in L^1(\mathbb{R}^d)$.

Note that we have equality in the above theorem, not just equivalence of the norms. We remark that throughout this section, we use the following convention for the Fourier transform

$$\hat{f}(\xi) = \int_{\mathbb{R}^d} f(x) e^{-2\pi i \xi \cdot x} dx,$$

for which the inverse transform is given by

$$f(x) = \int_{\mathbb{R}^d} \hat{f}(\xi) e^{2\pi i \xi \cdot x} d\xi.$$

To prove Theorem 4 we will need the following technical lemma concerning cutoff functions.

Lemma 4 *Suppose that $\Omega \subset \mathbb{R}^d$ is bounded. Let $\epsilon > 0$ and $s \geq 0$. Then there exists a function $\phi \in L^1(\mathbb{R}^d)$, such that $\phi(x) = 1$ for $x \in \Omega$ and*

$$\int_{\mathbb{R}^d} (1 + |\xi|)^s |\hat{\phi}(\xi)| d\xi \leq 1 + \epsilon.$$

Proof Since Ω is bounded, it suffices to consider the case where $\Omega = [-L, L]^d$ for a sufficiently large L. We consider separable $\phi = \phi_1(x_1) \cdots \phi_d(x_d)$, and note that

$$\int_{\mathbb{R}^d} (1 + |\xi|)^s |\hat{\phi}(\xi)| d\xi \leq \int_{\mathbb{R}^d} \prod_{i=1}^{d} (1 + |\xi_i|)^s |\hat{\phi}_i(\xi_i)| d\xi \leq \prod_{i=1}^{d} \int_{\mathbb{R}} (1 + |\xi|)^s |\hat{\phi}_i(\xi)| d\xi,$$

and this reduces us to the one-dimensional case where $\Omega = [-L, L]$.

For the one-dimensional case, consider a Gaussian $g_R(x) = e^{-\frac{x^2}{2R}}$. A simple calculation shows that the Fourier transform of the Gaussian is $\hat{g}_R(\xi) = \sqrt{\frac{R}{2\pi}} e^{-\frac{R\xi^2}{2}}$. This implies that

$$\lim_{R \to \infty} \int_{\mathbb{R}} (1 + |\xi|)^s |\hat{g}_R(\xi)| d\xi = 1,$$

and thus by choosing R large enough, we can make this arbitrarily close to 1.

Now consider $\tau_R \in C^k(\mathbb{R})$ for $k > s + 2$ such that $\tau_R(x) = 1 - g_R(x)$ for $x \in [-L, L]$. Then we have

$$\|\tau_R\|_{L^\infty([-L,L])}, \|\tau_R'\|_{L^\infty([-L,L])}, \cdots, \|\tau_R^{(k)}\|_{L^\infty([-L,L])} \to 0$$

as $R \to \infty$. Consequently, it is possible to extend τ_R to $\mathbb{R}$ so that

$$\|\tau_R\|_{L^1(\mathbb{R})}, \|\tau_R^{(k)}\|_{L^1(\mathbb{R})} \to 0.$$

as $R \to \infty$. For instance, for $x > L$ we can take τ_R to be a polynomial which matches the first k derivatives at L times a fixed smooth cutoff function which is identically 1 in some neighborhood of L (and similarly at $-L$).

This implies that $\|\hat{\tau}_R(\xi)\|_{L^\infty(\mathbb{R})}, \|\xi^k \hat{\tau}_R(\xi)\|_{L^\infty(\mathbb{R})} \to 0$ as $R \to \infty$. Together, these imply that

$$\lim_{R \to \infty} \int_{\mathbb{R}} (1 + |\xi|)^s |\hat{\tau}_R(\xi)| d\xi \to 0,$$

since $k - 2 > s$.

Finally, set $\phi_R = g_R(x) + \tau_R(x)$. Then clearly $\phi_R = 1$ on $[-L, L]$ and also

$$\lim_{R \to \infty} \int_{\mathbb{R}} (1 + |\xi|)^s |\hat{\phi}_R(\xi)| d\xi \leq \lim_{R \to \infty} \int_{\mathbb{R}} (1 + |\xi|)^s |\hat{\tau}_R(\xi)| d\xi +$$

$$\lim_{R\to\infty}\int_{\mathbb{R}}(1+|\xi|)^s|\hat{g}_R(\xi)|d\xi = 1.$$

Choosing R large enough, we obtain the desired result. □

Using this lemma, we now show that integral representations of the form (2.1) over the dictionary $\mathbb{F}_s$ are equivalent to the right hand side of (5.1).

Proposition 5 *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain and $s \geq 0$. Then*

$$\inf\left\{\|\mu\| : f = \int_{\mathbb{F}_s} i_{\mathbb{F}_s\to L^2(\Omega)}\mathrm{d}\mu\right\} = \inf_{f_e|_\Omega = f}\int_{\mathbb{R}^d}(1+|\xi|)^s|\hat{f}_e(\xi)|d\xi. \quad (5.2)$$

Proof We first prove the inequality

$$\inf\left\{\|\mu\| : f = \int_{\mathbb{F}_s} i_{\mathbb{F}_s\to L^2(\Omega)}\mathrm{d}\mu\right\} \leq \inf_{f_e|_\Omega = f}\int_{\mathbb{R}^d}(1+|\xi|)^s|\hat{f}_e(\xi)|d\xi.$$

If the right hand side is infinite, there is nothing to prove. So let $f_e \in L^1(\mathbb{R}^d)$ be an extension such that

$$\int_{\mathbb{R}^d}(1+|\xi|)^s|\hat{f}_e(\xi)|d\xi < \infty.$$

In particular, this means that $\hat{f} \in L^1(\mathbb{R}^d)$ as well and Fourier inversion holds almost everywhere. So we get

$$f(x) = \int_{\mathbb{R}^d}(1+|\xi|)^{-s}e^{2\pi \mathrm{i}\xi\cdot x}(1+|\xi|)^s\hat{f}(\xi)d\xi$$

for almost every $x \in \Omega$. Thus, by choosing $\mu = (1+|\xi|)^s\hat{f}(\xi)d\xi$ we get

$$f = \int_{\mathbb{F}_s} i_{\mathbb{F}_s\to L^2(\Omega)}\mathrm{d}\mu,$$

where the Bochner integral in is justified since $\mathbb{F}_s$ is uniformly bounded in $L^2(\Omega)$ and $\|\mu\| < \infty$. The right hand side is then a function in $L^2(\Omega)$ which agrees with f almost everywhere (hence we have equality). Thus we get

$$\inf\left\{\|\mu\| : f = \int_{\mathbb{F}_s} i_{\mathbb{F}_s\to L^2(\Omega)}\mathrm{d}\mu\right\} \leq \inf_{f_e|_\Omega = f}\int_{\mathbb{R}^d}(1+|\xi|)^s|\hat{f}_e(\xi)|d\xi.$$

Now let us prove the reverse inequality. Let λ be a regular Borel measure such that the integral on the right hand side of (5.2) is finite (note this must mean that λ has finite mass) and

$$f(x) = \int_{\mathbb{F}_s} i_{\mathbb{F}_s\to L^2(\Omega)}d\lambda = \int_{\mathbb{R}^d}(1+|\xi|)^{-s}e^{2\pi \mathrm{i}\xi\cdot x}d\lambda(\xi)$$

for $x \in \Omega$. Let $\mu = (1+|\xi|)^{-s}\lambda$, so that we have

$$f(x) = \int_{\mathbb{R}^d} e^{2\pi \mathrm{i}\xi\cdot x} \mathrm{d}\mu(\xi)$$

and

$$\int_{\mathbb{R}^d} (1+|\nu|)^s d|\mu|(\nu) = \|\lambda\|.$$

Choose $\epsilon > 0$. By Lemma 4 we can find a $\phi \in L^1(\mathbb{R}^d)$ such $\phi|_\Omega = 1$ and

$$\int_{\mathbb{R}^d} (1+|\xi|)^s |\hat{\phi}(\xi)| d\xi \leq 1+\epsilon.$$

We now set

$$f_e(x) = \phi(x)\left[\int_{\mathbb{R}^d} e^{2\pi \mathrm{i}\xi\cdot x} \mathrm{d}\mu(\xi)\right] \in L^1(\mathbb{R}^d),$$

since $\phi \in L^1(\mathbb{R}^d)$ and μ has finite mass, so the second factor must be bounded.

Then we have that for $x \in \Omega$,

$$f(x) = f(x)\phi(x) = f_e(x),$$

and $\hat{f}_e = \hat{\phi} * \mu$, where the function $\hat{\phi} * \mu$ is given by

$$(\hat{\phi} * \mu)(\xi) = \int_{\mathbb{R}^d} \hat{\phi}(\xi - \nu) \mathrm{d}\mu(\nu).$$

We now calculate

$$\int_{\mathbb{R}^d} (1+|\xi|)^s |(\hat{\phi} * \mu)(\xi)| d\xi \leq \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} (1+|\xi|)^s |\hat{\phi}(\xi - \nu)| d|\mu|(\nu) d\xi.$$

Finally, we use the simple inequality $(1+|\xi|)^s \leq (1+|\nu|)^s (1+|\xi-\nu|)^s$ combined with a change of variables, to get

$$\begin{aligned}\int_{\mathbb{R}^d} (1+|\xi|)^s |(\hat{\phi} * \mu)(\xi)| d\xi &\leq \left(\int_{\mathbb{R}^d} (1+|\xi|)^s |\hat{\phi}(\xi)| d\xi\right)\left(\int_{\mathbb{R}^d} (1+|\nu|)^s d|\mu|(\nu)\right) \\ &\leq (1+\epsilon)\left(\int_{\mathbb{R}^d} (1+|\nu|)^s d|\mu|(\nu)\right) = (1+\epsilon)\|\lambda\|.\end{aligned}$$

This shows that

$$\inf_{f_e|_\Omega = f} \int_{\mathbb{R}^d} (1+|\xi|)^s |\hat{f}_e(\xi)| d\xi \leq (1+\epsilon) \inf\left\{\|\mu\| : \ f = \int_{\mathbb{F}_s} i_{\mathbb{F}_s \to L^2(\Omega)} \mathrm{d}\mu\right\}.$$

Since $\epsilon > 0$ was arbitrary, we get the desired result. □

This completes the proof of Theorem 4 if $s > 0$ since then $\mathbb{F}_s$ is compact in $L^2(\Omega)$ and we can invoke Lemma 3 to obtain the equality

$$\|f\|_{\mathbb{F}_s} = \int_{\mathbb{F}_s} i_{\mathbb{F}_s \to L^2(\Omega)} \mathrm{d}\mu$$

The final step is thus to prove the left equality in 5.1 when $s = 0$. For this, we use the following.

Proposition 6 *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. Then*

$$B_e(\Omega) = \left\{ f : \Omega \to \mathbb{R} : \inf_{f_e|_\Omega = f} \int_{\mathbb{R}^d} |\hat{f_e}(\xi)| d\xi \leq 1 \right\} \tag{5.3}$$

is closed in $L^2(\Omega)$.

Proof Let $f_n \to f$ in $L^2(\Omega)$ with $f_n \in B_e(\Omega)$. Choose $\epsilon > 0$ and consider the corresponding sequence of $h_n = \hat{f}_{n,e}$ in (5.3) which satisfy

$$\int_{\mathbb{R}^d} |h_n(\xi)| d\xi \leq 1 + \epsilon, \;\; f_n(x) = \hat{h}_n(x) = \int_{\mathbb{R}^d} h_n(\xi) e^{2\pi \mathrm{i} \xi \cdot x} d\xi. \tag{5.4}$$

By assumption $f_n \to f$ in $L^2(\Omega)$ so that for any $g \in L^2(\Omega)$, we have

$$\langle f_n, g \rangle_{L^2(\Omega)} \to \langle f, g \rangle_{L^2(\Omega)}. \tag{5.5}$$

Choose g to be any element in the dense subset $C_c^\infty(\Omega) \subset L^2(\Omega)$ and note that in this case we have by Plancherel's theorem

$$\langle f_n, g \rangle_{L^2(\Omega)} = \langle f_n, g \rangle_{L^2(\mathbb{R}^d)} = \langle h_n, \hat{g} \rangle_{L^2(\mathbb{R}^d)}.$$

Note that $\hat{g}$ is a Schwartz function and so is in $C_0(\mathbb{R}^d)$, the space of continuous, decaying functions

$$C_0(\mathbb{R}^d) = \{\phi \in C(\mathbb{R}) : \lim_{\xi \to \infty} |\phi(\xi)| = 0\},$$

with the supremum norm.

This implies that the map

$$h : \phi \to \lim_{n \to \infty} \langle h_n, \phi \rangle_{L^2(\mathbb{R}^d)}$$

defines a bounded linear functional on the subspace of $C_0(\mathbb{R}^d)$ which is spanned by $\{\hat{g} : g \in C_c^\infty(\Omega)\}$. The limit above exists by (5.6) and the assumption that $f_n \to f$. Further, the bound has norm $\leq 1 + \epsilon$ by equation (5.4).

By the Hahn-Banach theorem, we can extend h to an element $\mu \in C_0^*(\mathbb{R}^d)$, such that $\|\mu\|_{C_0^*(\mathbb{R}^d)} \leq 1+\epsilon$. By the Riesz-Markov theorem (Theorem 22 in [24]), the dual space $C_0^*(\mathbb{R}^d)$ is exactly the space of Borel measures with the total variation norm. Thus we get

$$\|\mu\|_{C_0^*(\mathbb{R}^d)} = \int_{\mathbb{R}^d} d|\mu|(\xi) \leq 1+\epsilon.$$

But we also have that for every $g \in C_c^\infty(\Omega)$, $\langle \mu, \hat{g} \rangle = \langle f, g \rangle$. Taking the Fourier transform, we see that the function

$$f_\mu = \int_{\mathbb{R}^d} e^{2\pi \mathrm{i} \xi \cdot x} \mathrm{d}\mu(\xi)$$

satisfies $\langle f_\mu, g \rangle = \langle f, g \rangle$ for all $g \in C_c^\infty(\Omega)$. Thus $f = f_\mu$ in $L^2(\Omega)$ and so by (5.2), we have

$$\inf_{f_e|_\Omega = f} \int_{\mathbb{R}^d} |\hat{f}_e(\xi)| d\xi \leq \int_{\mathbb{R}^d} |\hat{f}_\mu(\xi)| d\xi \leq 1+\epsilon.$$

Since ϵ was arbitrary, this completes the proof. □

To complete the proof in the case of $s = 0$, we simply note that by (5.2), $B_e(\Omega)$ contains all of the complex exponentials $e^{2\pi i \omega \cdot x}$. Since it is clearly convex and is closed by Proposition 5, it must be equal to $\overline{\mathrm{conv}(\pm \mathbb{F}_0)}$. This completes the proof of Theorem 4.

6 Conclusion

We have provided some foundational analysis of the variation spaces with respect to dictionaries arising in the study of shallow neural networks. The precise analysis of approximation theoretic properties such as the metric entropy and n-widths of these spaces is a major research direction which we propose. In addition, it must be investigated whether these variation spaces are useful for any particular practical applications, for instance solving PDEs.

Acknowledgements We would like to thank Professors Russel Caflisch, Ronald DeVore, Weinan E, Albert Cohen, Stephan Wojtowytsch and Jason Klusowski for helpful discussions. We would also like to thank the anonymous reviewers for their helpful comments. 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. Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Trans. Inform. Theory **39**(3), 930–945 (1993)
2. Barron, A.R., Cohen, A., Dahmen, W., DeVore, R.A.: Approximation and learning by greedy algorithms. Annal Stat. **36**(1), 64–94 (2008)
3. DeVore, R.A.: Nonlinear approximation. Acta Numer. **7**, 51–150 (1998)
4. DeVore, R.A., Temlyakov, V.N.: Some remarks on greedy algorithms. Adv. Comput. Math. **5**(1), 173–187 (1996)
5. Diestel, J.: Sequences and series in Banach spaces, vol. 92. Springer Science & Business Media (2012)
6. Dudley, R.M.: Real analysis and probability. CRC Press (2018)
7. E, W., Ma, C., Wu, L.: Barron spaces and the compositional function spaces for neural network models. arXiv preprint arXiv:1906.08039 (2019)
8. Weinan, E., Wojtowytsch, S.: Representation formulas and pointwise properties for barron functions. CoRR (2020)
9. Evans, L.C., Gariepy, R.F.: Measure theory and fine properties of functions. CRC press (2015)
10. Hao, W., Jin, X., Siegel, J.W., Xu, J.: An efficient greedy training algorithm for neural networks and applications in pdes. arXiv preprint arXiv:2107.04466 (2021)
11. Hornik, K., Stinchcombe, M., White, H., Auer, P.: Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. Neural Comput. **6**(6), 1262–1275 (1994)
12. Jones, L.K.: A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training. Ann. Stat. **20**(1), 608–613 (1992)
13. Kainen, P.C., Krková, V., Vogt, A.: Integral combinations of heavisides. Math. Nachr. **283**(6), 854–878 (2010)
14. Klusowski, J.M., Barron, A.R.: Approximation by combinations of relu and squared relu ridge functions with ℓ^1 and ℓ^0 controls. IEEE Trans. Inform. Theory **64**(12), 7649–7656 (2018)
15. Krogh, A., Hertz, J.: A simple weight decay can improve generalization. Adv. Neural Inform. Process. Syst.**4** (1991)
16. Kurková, V., Sanguineti, M.: Bounds on rates of variable-basis and neural-network approximation. IEEE Trans. Inform. Theory **47**(6), 2659–2665 (2001)
17. Kurková, V., Sanguineti, M.: Comparison of worst case errors in linear and neural network approximation. IEEE Trans. Inform. Theory **48**(1), 264–275 (2002)
18. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
19. Leshno, M., Lin, V.Y., Pinkus, A., Schocken, S.: Multilayer feedforward networks with a non-polynomial activation function can approximate any function. Neural Networks **6**(6), 861–867 (1993)
20. Li, Z., Ma, C., Wu, L.: Complexity measures for neural networks with general activation functions using path-based norms. arXiv preprint arXiv:2009.06132 (2020)
21. Livshits, E.D.: Lower bounds for the rate of convergence of greedy algorithms. Izvestiya: Mathematics **73**(6), 1197 (2009)
22. Makovoz, Y.: Random approximants and neural networks. J. Approx. Theory **85**(1), 98–109 (1996)
23. Makovoz, Y.: Uniform approximation by neural networks. J. Approx. Theory **95**(2), 215–228 (1998)
24. Markoff, A.: On mean values and exterior densities. Rec. Math. **4**(1), 165–191 (1938)
25. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML (2010)
26. Ongie, G., Willett, R., Soudry, D., Srebro, N.: A function space view of bounded norm infinite width relu nets: The multivariate case. In: International Conference on Learning Representations (ICLR 2020) (2019)
27. Parhi, R., Nowak, R.D.: Banach space representer theorems for neural networks and ridge splines. arXiv preprint arXiv:2006.05626 (2020)
28. Parhi, R., Nowak, R.D.: What kinds of functions do deep neural networks learn? insights from variational spline theory. arXiv preprint arXiv:2105.03361 (2021)
29. Petrosyan, A., Dereventsov, A., Webster, C.G.: Neural network integral representations with the relu activation function. In: Mathematical and Scientific Machine Learning, pp. 128–143. PMLR (2020)
30. Pisier, G.: Remarques sur un résultat non publié de b. maurey. Séminaire Analyse fonctionnelle (dit "Maurey-Schwartz") pp. 1–12 (1981)

31. Prokhorov, Y.V.: Convergence of random processes and limit theorems in probability theory. Theory Probab. Appl. **1**(2), 157–214 (1956)
32. Savarese, P., Evron, I., Soudry, D., Srebro, N.: How do infinite width bounded norm networks look in function space? In: Conference on Learning Theory, pp. 2667–2690. PMLR (2019)
33. Siegel, J.W., Xu, J.: Approximation rates for neural networks with general activation functions. Neural Networks **128**, 313–321 (2020)
34. Siegel, J.W., Xu, J.: High-order approximation rates for neural networks with ReLU^k activation functions. arXiv preprint arXiv:2012.07205 (2020)
35. Siegel, J.W., Xu, J.: Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks. arXiv preprint arXiv:2101.12365 (2021)
36. Sil'nichenko, A.: Rate of convergence of greedy algorithms. Math. Notes **76**(3), 582–586 (2004)
37. Temlyakov, V.: Greedy approximation, vol. 20. Cambridge University Press (2011)
38. Temlyakov, V.N.: Greedy approximation. Acta Numerica **17**(235), 409 (2008)
39. Weinan, E., Ma, C., Wu, L.: Barron spaces and the compositional function spaces for neural network models. arXiv preprint arXiv:1906.08039 (2019)
40. Weinan, E., Ma, C., Wu, L.: The barron space and the flow-induced function spaces for neural network models. Constructive Approximation pp. 1–38 (2021)
41. Xu, J.: Finite neuron method and convergence analysis. Commun. Comput. Phys. **28**(5), 1707–1745 (2020). https://doi.org/10.4208/cicp.OA-2020-0191