Journal of Machine Learning **ISSN**: 2790-2048(e), 2790-203X(p)

Embedding Inequalities for Barron-Type Spaces

Lei Wu * [1,2]

[1]School of Mathematical Sciences, Peking University, Beijing, China.
[2]Center for Machine Learning, Peking University, Beijing, China.

Abstract. An important problem in machine learning theory is to understand the approximation and generalization properties of two-layer neural networks in high dimensions. To this end, researchers have introduced the Barron space $\mathcal{B}_s(\Omega)$ and the spectral Barron space $\mathcal{F}_s(\Omega)$, where the index $s \in [0,\infty)$ indicates the smoothness of functions within these spaces and $\Omega \subset \mathbb{R}^d$ denotes the input domain. However, the precise relationship between the two types of Barron spaces remains unclear. In this paper, we establish a continuous embedding between them as implied by the following inequality: For any $\delta \in (0,1), s \in \mathbb{N}^+$ and $f : \Omega \mapsto \mathbb{R}$, it holds that

$$\delta \|f\|_{\mathcal{F}_{s-\delta}(\Omega)} \lesssim_s \|f\|_{\mathcal{B}_s(\Omega)} \lesssim_s \|f\|_{\mathcal{F}_{s+1}(\Omega)}.$$

Importantly, the constants do not depend on the input dimension d, suggesting that the embedding is effective in high dimensions. Moreover, we also show that the lower and upper bound are both tight.

Keywords:
Barron space,
Two-layer neural network,
High-dimensional approximation,
Embedding theorem,
Fourier transform.

Article Info.:
Volume: 2
Number: 4
Pages: 259 - 270
Date: December/2023
doi.org/10.4208/jml.230530

Article History:
Received: 30/05/2023
Accepted: 17/12/2023

Communicated by:
Weinan E

1 Introduction

A (scaled) two-layer neural network is given by

$$f_m(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \sigma\big(w_j^T x + b_j\big), \tag{1.1}$$

where $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is a nonlinear activation function; $a_j, b_j \in \mathbb{R}, w_j \in \mathbb{R}^d, \theta = \{(a_j, w_j, b_j)\}_{j=1}^m$, m and d denote the network width and the input dimension, respectively. The extra scale factor in (1.1) is introduced to facilitate our subsequent analysis and it does not change the network's approximation power. Additionally, throughout this paper, we assume the input domain $\Omega \subset \mathbb{R}^d$ to be compact and focus on the case of activation function ReLU^s with $s \geq 0$

$$\sigma(z) = \max(0,z)^s.$$

The cases of $s = 0$ and $s = 1$ correspond to the Heaviside step function and vanilla ReLU function, respectively. The case of $s \geq 2$ has also found applications in solving PDEs [11,13,26] and natural language processing [25].

*leiwu@math.pku.edu.cn

Cybenko [5] showed that functions in $C(\Omega)$ can be approximated arbitrarily well by two-layer neural networks with respect to the uniform metric. However, the approximation can be arbitrarily slow. Pinkus [21] expanded on this by showing that for functions belonging in $C^k(\Omega)$, the approximation by two-layer neural networks can achieve a rate of $\mathcal{O}(m^{-k/d})$. This rate, unfortunately, is subject to the curse of dimensionality since it diminishes as d increases. These suggest that mere continuity and smoothness are not sufficient to ensure an efficient approximation in high dimensions. Then it is natural to ask: What kind of regularity can ensure the efficient approximation by two-layer neural networks? Before proceeding to review previous studies attempting to answer this question. We need a dual norm for handling the compactness of input domain.

Definition 1.1 ([1])**.** *Given a compact set Ω, we define $\|v\|_\Omega = \sup_{x\in\Omega} |v^T x|$.*

We begin by considering the spectral Barron spaces [3, 22, 24, 26], which are defined as follows.

Definition 1.2. *Let $\Omega \subset \mathbb{R}^d$ be a compact domain. For $f : \Omega \mapsto \mathbb{R}$ and $s \geq 0$, define*

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

where the infimum is taken over all extensions of f. Let

$$\mathcal{F}_s(\Omega) := \{f : \Omega \mapsto \mathbb{R} \, : \, \|f\|_{\mathcal{F}_s(\Omega)} < \infty\}.$$

Then, the spectral Barron space is defined as $\mathcal{F}_s(\Omega)$ equipped with the $\|\cdot\|_{\mathcal{F}_s(\Omega)}$ norm.

In the above definition, we consider measure-valued Fourier transform as done in [1]. It is worth noting that Definition 1.2 bears resemblance to the Fourier-based characterization of Sobolev spaces, denoted as

$$\|f\|_{H_s}^2 = \int_{\mathbb{R}^d} (1 + \|\xi\|)^s |\hat{f}(\xi)|^2 \, \mathrm{d}\xi.$$

The major distinction lies in the fact that the moment in Definition 1.2 is calculated with respect to $|\hat{f}(\xi)|$ instead of $|\hat{f}(\xi)|^2$.

It was proved in [26] that if $\|f\|_{\mathcal{F}_s(\Omega)} < \infty$, then functions in $\mathcal{F}_s(\Omega)$ can be approximated by two-layer ReLU^{s-1} networks without suffering the curse of dimensionality. Specifically, the approximation error obeys the Monte-Carlo error rate $\mathcal{O}(m^{-1/2})$, where m denotes the network width. The special case of $s = 1$ was first considered in the pioneer work of Barron [1]. Subsequently, the case of $s = 2$ was studied in [2, 12]. More recently, the extension to general positive integer s was provided in [3, 22, 26].

The Fourier-based characterization, while explicit, is not necessarily tight as it may exclude functions that can be effectively approximated by two-layer neural networks. [19,20] considered similar characterizations based on Radon transform instead of Fourier transform, which can yield a tight characterization for the case of $d = 1$. Moreover, [7,8] offered a probabilistic generalization of Barron's analysis [1]. In these studies, functions satisfying the following expectation representation are taken into consideration:

$$f_\rho(x) = \mathbb{E}_{(a,w,b)\sim\rho}\left[a\sigma(w^T x + b)\right], \quad \forall\, x \in \Omega, \tag{1.2}$$

where $\rho \in \mathcal{P}(\mathbb{R} \times \mathbb{R}^d \times \mathbb{R})$. This can be obtained from (1.1) by taking $m \to \infty$ and applying the law of large numbers. One can view f_ρ as an infinitely-wide two-layer neural network. It is important to note that the expectation representation in (1.2) only needs to hold in Ω instead of the entire space $\mathbb{R}^d$. Accordingly, the (probabilistic) Barron spaces are defined as follows.

Definition 1.3. *Given $s \geq 0$ and $f : \Omega \mapsto \mathbb{R}$, let*

$$A_f := \left\{\rho \in \mathcal{P}(\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}) \,:\, f_\rho|_\Omega = f\right\}.$$

Then, the Barron norm of f and the associated Barron space is defined by

$$\|f\|_{\mathcal{B}_s(\Omega)} := \inf_{\rho\in A_f} \mathbb{E}_{(a,w,b)\sim\rho}\left[|a|(\|w\|_\Omega + |b|)^s\right].$$

Let

$$\mathcal{B}_s(\Omega) = \{f : \Omega \mapsto \mathbb{R} \,:\, \|f\|_{\mathcal{B}_s(\Omega)} < \infty\}.$$

Then the Barron space is defined as $\mathcal{B}_s(\Omega)$ equipped with the $\|\cdot\|_{\mathcal{B}_s(\Omega)}$ norm.

The above definition is a slight generalization of the one originally proposed in [8], where only the case of $s = 1$ is considered. Following the proofs in [7, 8], one can easily show that approximating and estimation error for learning functions in $\mathcal{B}_s$ with two-layer ReLU^s networks follow the Monte-Carlo rates $\mathcal{O}(m^{-1/2})$ and $\mathcal{O}(n^{-1/2})$, respectively. Here n denotes the number of training samples. Recently, [23] established a sharper approximation rate of $\mathcal{O}(m^{-1/2-(s+1/2)/d})$. However, it is important to note that this rate improvement is less significant in high dimensions and additionally, the hidden constants in [23] may have an exponential dependence on d. Compared with the Fourier-based characterization in Definition 1.2, the above expectation-based characterization is more natural and complete. Specifically, [8] provided an inverse approximation theorem, showing that if f can be approximated by two-layer ReLU networks with bounded path norm [18], it must lie in $\mathcal{B}_1(\Omega)$.

1.1 Our contribution

Recently, Barron-type spaces defined above have been adopted to explore various high-dimensional problems. For instance, [4, 9, 15, 16] established some regularity theories of high-dimensional PDEs with Barron-type spaces. Hence, it is natural to ask: What is the relationship between them? [1,2,12,26] already showed that $\mathcal{F}_{s+1}(\Omega) \subseteq \mathcal{B}_s(\Omega)$. Moreover, [10] provided a specific example showing that $\mathcal{F}_2(\Omega) \subsetneq \mathcal{B}_1(\Omega)$, implying that $\mathcal{B}_1(\Omega)$ is strictly larger than $\mathcal{F}_2(\Omega)$. Along this line of work, our major contribution is the following precise embedding result.

Theorem 1.1. *Let $\Omega \subset \mathbb{R}^d$ be a compact set. For any $s \in \mathbb{N}^+$, $f \in \mathcal{B}_s(\Omega)$, $\delta \in (0, 1)$, we have*

$$\delta\|f\|_{\mathcal{F}_{s-\delta}(\Omega)} \lesssim_s \|f\|_{\mathcal{B}_s(\Omega)} \lesssim_s \|f\|_{\mathcal{F}_{s+1}(\Omega)},$$

where s in the upper bound can take the value of 0.

Note that the hidden embedding constants depend solely on the value of s. This suggests that the embedding revealed in Theorem 1.1 is effective in high dimensions. Additionally, as per our current proof, the smoothness index s is required to be a positive integer, though $\mathcal{B}_s(\Omega)$ is defined for any s in the range of $[0,\infty)$. However, we conjecture that analogous results would apply for any $s \in (0,+\infty)$ as discussed in Remark 2.1, which we leave for future work.

Additionally, we would like to clarify that the upper bound in Theorem 1.1 has already been implicitly established in previous works. Specifically, the case of $s = 0$ was proven in the pioneering work of Barron [1] albeit presented in a different form. Subsequently, the analysis was extended to the case of $s = 1$ in [2, 12], and further generalized to arbitrary non-negative integer values of s in [22, 26]. Our major contribution is the lower bound, which is critical for establishing the embedding between the two types of Barron spaces and the proof is presented in Section 2.1. In Theorem 1.1, the upper bound is stated for the sake of completeness.

We mention that [17] establishes the embedding among spectral Barron spaces and some classical spaces such as the Sobolev space, Besov space, and Bessel potential space. In contrast, we focus on the embedding between the Barron spaces and spectral Barron spaces.

Tightness. For the upper bound, [3, Proposition 7.4] shows that when Ω has nonempty interior, if $\mathcal{F}_s(\Omega) \subset \mathcal{B}_1(\Omega)$, then we must have $s \geq 2$. This implies that the upper bound is tight. The following proposition shows that the lower bound in Theorem 1.1 is also tight in the sense that the value of δ cannot be taken to zero.

Proposition 1.1. *Let* $\Omega = [-1,1]$ *and* $f(x) = \max(1-|x|,0)$ *for* $x \in \Omega$. *Then,*

$$\|f\|_{\mathcal{B}_1(\Omega)} \leq 3, \quad \|f\|_{\mathcal{F}_1(\Omega)} = +\infty.$$

Let $t(x) := \max(1-|x|,0)$ for any $x \in \mathbb{R}$ be the triangular hat function (see Fig. 1.1).

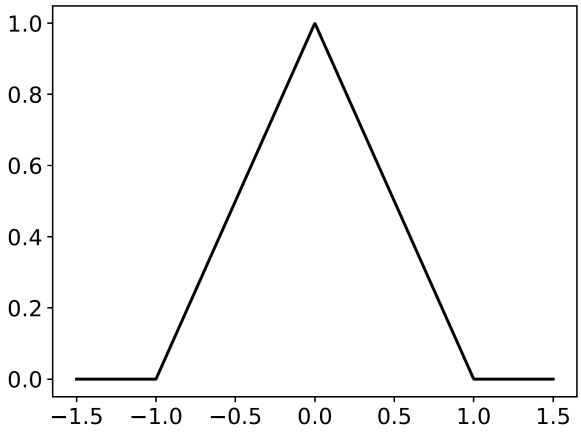

Figure 1.1: The triangular function $t(x) := \max(1-|x|,0)$.

We have

$$\hat{t}(\xi) = \frac{1-\cos(\xi)}{\pi\xi^2}.$$

Note that $t(\cdot)$ is a zero extension of f and

$$\int_{\mathbb{R}} (1+|\xi|)\left|\hat{t}(\xi)\right| \mathrm{d}\xi = +\infty.$$

However, this does not directly imply $\|f\|_{\mathcal{F}_1(\Omega)} = \infty$, since the spectral Barron norm is defined by taking the infimum over all possible extensions. We refer to Section 2.2 for a rigorous proof.

2 Proofs

Notation. We use $X \lesssim_\alpha Y$ to denote $X \leq C_\alpha Y$ where C_α is a positive constant that depends only on α. For a vector v, let $\|v\|_p = (\sum_j v_j^p)^{1/p}$. Let

$$\mathbb{S}^{d-1} = \left\{x \in \mathbb{R}^d \,:\, \|x\|_2 = 1\right\},$$
$$\mathbb{S}_\Omega^{d-1} = \left\{x \in \mathbb{R}^d \,:\, \|x\|_\Omega = 1\right\}.$$

Denote by 1_S the indicator function of the set S, satisfying $1_S(x) = 1$ for $x \in S$, and 0 otherwise. For a metric space X, denote by $\mathcal{P}(X)$ the set of probability measures over X.

Throughout this paper, we define Fourier transform as follows:

$$\hat{f}(\xi) = \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i\xi^T x} f(x)\, \mathrm{d}x,$$

and the inverse Fourier transform is given by

$$f(x) = \int_{\mathbb{R}^d} e^{i\xi^T x} \hat{f}(\xi)\, \mathrm{d}\xi.$$

Note that in these definitions, the terms $f(x)\,\mathrm{d}x$ and $\hat{f}(\xi)\,\mathrm{d}\xi$ should be interpreted as a finite measure in a broad sense. Moreover, we will use the identity: for $d = 1$,

$$\delta(\xi) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-i\xi x}\, \mathrm{d}x.$$

Before proceeding to the proof, we first clarify several important issues that might be ignored. Both types of Barron functions are defined on a compact domain Ω instead of the whole space $\mathbb{R}^d$ and thus, Barron norms depend on the underlying domain Ω. When estimating Barron norms, one need to be careful with the choice of extensions. A naive extension may yield a significantly loose bound of the $\mathcal{F}_s(\Omega)$ norm [6] and $\mathcal{B}_s(\Omega)$ norm.

2.1 Proof of Theorem 1.1

We start by considering the case of single neurons. For any $w \in \mathbb{S}_\Omega^{d-1}, b \in \mathbb{R}$, the single neuron $\sigma_{w,b} : \Omega \mapsto \mathbb{R}$ is given by $\sigma_{w,b}(x) = \sigma(w^T x + b)$. Note that the domain of $\sigma_{w,b}$ is Ω

instead of $\mathbb{R}^d$. In particular, when $d = 1$ and $w = 1$, we write $\sigma_b = \sigma_{w,b}$ for simplicity. The following lemma characterizes the Fourier transform of a single neuron.

Lemma 2.1. *Let $\sigma_{w,b} : \Omega \mapsto \mathbb{R}$ with $\|w\|_\Omega = 1$ be a single neuron and $g : \mathbb{R} \mapsto \mathbb{R}$ be any extension of $1_{[-1,1]}$. Then, $G_{w,b}(x) := \sigma_{w,b}(x)g(w^T x)$ is an extension of $\sigma_{w,b}$, satisfying*

$$\int_{\mathbb{R}^d} (1 + \|\xi\|_\Omega)^s \left|\widehat{G}_{w,b}(\xi)\right| \mathrm{d}\xi = \int_{\mathbb{R}} (1 + |v|)^s \left|\hat{h}_{\sigma,b}(v)\right| \mathrm{d}v, \tag{2.1}$$

where $h_{\sigma,b}(z) = \sigma(z + b)g(z)$ is an extension of $\sigma_b : [-1, 1] \mapsto \mathbb{R}$.

Proof. Let $Q = (w, w_2, \ldots, w_d)^T \in \mathbb{R}^{d\times d}$ with $w_2, \ldots, w_d$ being orthonormal and $w_i^T w = 0$ for $i = 2, \ldots, d$. Then, by letting $\bar{\xi} = (Q^{-1})^T \xi$, we have

$$\begin{aligned}
\widehat{G}_{w,b}(\xi) &= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \sigma\big(w^T x + b\big) g\big(w^T x\big) e^{-i\xi^T x} \,\mathrm{d}x \\
&= \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} \sigma(y_1 + b) g(y_1) e^{-i\xi^T Q^{-1} y} \frac{1}{|\det Q|} \,\mathrm{d}y \quad (y = Qx) \\
&= \frac{1}{|\det Q|} \left(\frac{1}{2\pi} \int_{\mathbb{R}^d} \sigma(y_1 + b) g(y_1) e^{-i\bar{\xi}_1 y_1} \,\mathrm{d}y_1 \right) \prod_{j=2}^{d} \delta(\bar{\xi}_j) \\
&= \frac{1}{|\det Q|} \hat{h}_{\sigma,b}(\bar{\xi}_1) \prod_{j=2}^{d} \delta(\bar{\xi}_j).
\end{aligned} \tag{2.2}$$

Now, we have

$$\begin{aligned}
&\int_{\mathbb{R}^d} (1 + \|\xi\|_\Omega)^s \left|\widehat{G}_{w,b}(\xi)\right| \mathrm{d}\xi \\
&= \int_{\mathbb{R}^d} \left(1 + \|Q^T \bar{\xi}\|_\Omega\right)^s \frac{1}{|\det Q|} |\hat{h}_{\sigma,b}(\bar{\xi}_1)| \prod_{j=2}^{d} \delta(\bar{\xi}_j) |\det Q| \,\mathrm{d}\bar{\xi} \\
&= \int_{\mathbb{R}^d} \left(1 + \left\| \bar{\xi}_1 w + \sum_{j=2}^{d} \bar{\xi}_j w_j \right\|_\Omega\right)^s \left|\hat{h}_{\sigma,b}(\bar{\xi}_1)\right| \prod_{j=2}^{d} \delta(\bar{\xi}_j) \,\mathrm{d}\bar{\xi} \\
&= \int_{\mathbb{R}} \left(1 + \|w\|_\Omega |\bar{\xi}_1|\right)^s \left|\hat{h}_{\sigma,b}(\bar{\xi}_1)\right| \mathrm{d}\bar{\xi}_1 \\
&= \int_{\mathbb{R}} \left(1 + |\bar{\xi}_1|\right)^s \left|\hat{h}_{\sigma,b}(\bar{\xi}_1)\right| \mathrm{d}\bar{\xi}_1,
\end{aligned} \tag{2.3}$$

where the first step uses $\xi = Q^T \bar{\xi}$ and the last step is due to $\|w\|_\Omega = 1$. □

The above lemma provides a way to estimating spectral Barron norms of single neurons. What remains is to determine an extension g such that the right-hand side of the Eq. (2.1) to be as small as possible. To this end, we first consider the one-dimensional case.

When $d = 1$, for any $b \in \mathbb{R}$, let $\sigma_b = \sigma(\cdot + b) : [-1, 1] \mapsto \mathbb{R}$. When it is clear from the context, we also use σ_b denote the single neuron define on the entire space. Let $\chi : \mathbb{R} \mapsto \mathbb{R}$

DOI https://doi.org/10.4208/jml.230530 | Generated on 2024-12-19 02:38:29

be a smooth cutoff function, satisfying $\chi \in C_c^\infty(\mathbb{R})$ and $\chi(z) = 1$ for any $z \in [-1,1]$ and $\operatorname{supp}\chi = [-2,2]$. Given any $b \in \mathbb{R}$, we shall consider the following extension of a single neuron:

$$h_{\sigma,b}(z) = \chi(z)\sigma_b(z) : \mathbb{R} \ \mapsto \ \mathbb{R}.$$

Lemma 2.2. *Let $\sigma(z) = \max(0,z)^s$ with $s \in \mathbb{N}$. Then,*

$$|\hat{h}_{\sigma,b}(\xi)| \lesssim_s \frac{(1+|b|)^s}{(1+|\xi|)^{s+1}}.$$

Remark 2.1. The proof uses explicitly the condition of $s \in \mathbb{N}$. However, according to the relationship between the smoothness of a function and the decay of the Fourier transform, we anticipate that the same result holds for any $s \in [0,\infty)$.

Proof. Using the product rule, we have for any $k \in \mathbb{N}^+$,

$$h_{\sigma,b}^{(k)}(z) = \sum_{i=0}^{k} \binom{k}{i} \sigma_b^{(i)}(z)\chi^{(k-i)}(z),$$

and prove the theorem for the following two cases separately. Without lose of generality, we consider here only the case of $b \geq 0$. When b is negative, the proof is similar.

Case 1: $b \geq 2$. In this case, $h_{\sigma,b}(\cdot) = \sigma_b(\cdot)\chi(\cdot) \in C^\infty(\mathbb{R})$. Without loss of generality, we consider the case of $b \geq 1$, for which

$$h_{\sigma,b}(z) = \begin{cases} 0, & \text{if} \quad z < -2, \\ (z+b)^s\chi(z), & \text{if} \quad z \in [-2,2], \\ 0, & \text{if} \quad z > 2. \end{cases}$$

- When $z \in [-2,2]$, we have $\sigma_b^{(k)}(z) = 0$ for $k > s$ and $|\sigma_b^{(k)}(z)| \lesssim_s (1+|b|)^s$ for $k \leq s$. Hence, for any $k \in \mathbb{N}$, we have

$$\begin{aligned} |h_b^{(k)}(z)| &= \sum_{i=0}^{\min\{k,s\}} \binom{k}{i} \sigma_b^{(i)}(z)\chi^{(k-i)}(z) \\ &\lesssim_s \sum_{i=0}^{\min\{k,s\}} |\sigma_b^{(i)}(z)||\chi^{(k-i)}(z)| \\ &\lesssim_s \sum_{k=0}^{\min\{k,s\}} |\sigma_b^{(k)}(z)| \lesssim_s (1+|b|)^s. \end{aligned}$$

- When $|z| > 2$, $|h_{\sigma,b}^{(k)}(z)| = 0$ for any $k \in \mathbb{N}$.

Combining two cases leads to for any $k \in \mathbb{N}$,

$$\|h_{\sigma,b}^{(k)}\|_{L^1(\mathbb{R})} = \int_{-2}^{2} |h_{\sigma,b}^{(k)}(z)| \, \mathrm{d}z \lesssim_s (1+|b|)^s.$$

This implies

$$\left|\hat{h}_b(\xi)\right| = \left|\frac{1}{2\pi(-i\xi)^{s+1}}\int_{\mathbb{R}} h_{\sigma,b}^{(s+1)}(x)e^{-i\xi x}\,\mathrm{d}x\right| \lesssim_s \frac{(1+|b|)^s}{|\xi|^{s+1}}. \tag{2.4}$$

Case 2: $0 \le b < 2$. In this case, $h_{\sigma,b}$ is piecewise smooth, given by

$$h_{\sigma,b}(z) = \begin{cases} 0, & \text{if} \quad z \le -b, \\ (z+b)^s\chi(z), & \text{if} \quad z > -b. \end{cases}$$

Consequently, $h_{\sigma,b}^{(s)}(\cdot)$ is bounded and has only one discontinuity point at $z = -b$. By adopting the product rule in a way similar as the above, it is not hard to show that for all $k \in \mathbb{N}$,

$$\begin{aligned} &h_{\sigma,b}^{(k)}(z) = 0, \qquad \forall\, z \in (-\infty, -b) \cup [2, +\infty), \\ &\left|h_{\sigma,b}^{(k)}(z)\right| \lesssim_s 1, \quad \forall\, z \in (-b, 2], \end{aligned} \tag{2.5}$$

and $\lim_{z\to(-b)^+} h_{\sigma,b}^{(s)}(z)$ exists with $|\lim_{z\to(-b)^+} h_{\sigma,b}^{(s)}(z)| \lesssim_s 1$.

Noting that

$$\begin{aligned} \hat{h}_{\sigma,b}(\xi) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} h_{\sigma,b}(z)e^{-i\xi z}\,\mathrm{d}z \\ &= \frac{1}{2\pi(-i\xi)^s}\int_{-\infty}^{\infty} h_{\sigma,b}^{(s)}(z)e^{-i\xi z}\,\mathrm{d}z \\ &= \frac{1}{2\pi(-i\xi)^s}\int_{-b}^{2} h_{\sigma,b}^{(s)}(z)e^{-i\xi z}\,\mathrm{d}z \\ &= \frac{1}{2\pi(-i\xi)^s}\left(\frac{e^{-i\xi z}}{-i\xi}h_{\sigma,b}^{(s)}(z)\Big|_{-b}^{2} + \int_{-b}^{2} h_{\sigma,b}^{(s+1)}(z)\frac{e^{-i\xi z}}{i\xi}\,\mathrm{d}z\right), \end{aligned}$$

and applying (2.5), we have

$$|\hat{h}_{\sigma,b}(\xi)| \lesssim_s \frac{1}{|\xi|^{s+1}}\left(1 + \int_{-b}^{2}\mathrm{d}z\right) \lesssim \frac{1}{|\xi|^{s+1}} \lesssim_s \frac{(1+|b|)^s}{|\xi|^{s+1}}, \tag{2.6}$$

where the last step uses the assumption of $|b| \le 2$.

On the other hand, when $|\xi| \le 1$, we have for any $b \in \mathbb{R}$ that

$$\left|\hat{h}_{\sigma,b}(\xi)\right| \le \frac{1}{2\pi}\int_{\mathbb{R}} \left|h_{\sigma,b}(z)e^{-i\xi z}\,\mathrm{d}z\right| \le \frac{1}{2\pi}\int_{-2}^{2} |h_{\sigma,b}(z)|\,\mathrm{d}z \lesssim_s (1+|b|)^s. \tag{2.7}$$

Then, combining (2.7) with (2.4) and (2.6) yields

$$\left|\hat{h}_{\sigma,b}(\xi)\right| \lesssim_s \frac{(1+|b|)^s}{(1+|\xi|)^{s+1}}.$$

The proof is complete. □

DOI https://doi.org/10.4208/jml.230530 | Generated on 2024-12-19 02:38:29

Lemma 2.3. *Given any $w \in \mathbb{S}_{\Omega}^{d-1}$, $b \in \mathbb{R}$, consider the extension*

$$H_{w,b}(x) := \sigma\big(w^T x + b\big)\chi\big(w^T x\big).$$

Then, for any $\delta \in (0,1)$, we have

$$\int_{\mathbb{R}} (1 + \|\xi\|_{\Omega})^{s-\delta} \big|\hat{H}_{w,b}(\xi)\big| \, \mathrm{d}\xi \lesssim_s \delta^{-1}(1+|b|)^s. \tag{2.8}$$

Proof. By Lemmas 2.1 and 2.2, we have

$$\begin{aligned}
&\int_{\mathbb{R}} (1 + \|\xi\|_{\Omega})^{s-\delta} \big|\hat{H}_{w,b}(\xi)\big| \\
&= \int_{\mathbb{R}} (1+|v|)^{s-\delta} \big|\hat{h}_{\sigma,b}(v)\big| \, \mathrm{d}v \\
&\lesssim_s \int_{\mathbb{R}} (1+|v|)^{s-\delta} \frac{(1+|b|)^s}{(1+|v|)^{s+1}} \, \mathrm{d}v \\
&\lesssim_s (1+|b|)^s \int_{\mathbb{R}} \frac{1}{(1+|v|)^{1+\delta}} \, \mathrm{d}v \lesssim_s \frac{(1+|b|)^s}{\delta}.
\end{aligned}$$

The proof is complete. □

The proof of Theorem 1.1. We are now ready to prove the main theorem. For any $f \in \mathcal{B}_s(\Omega)$ and $\varepsilon > 0$, there exists $\rho_\epsilon \in \mathcal{P}(\mathbb{R} \times \mathbb{S}_{\Omega}^{d-1} \times \mathbb{R})$ such that

$$\begin{aligned}
f(x) &= \int a\sigma(w^T x + b) \, \mathrm{d}\rho_\epsilon(a,w,b), \quad \forall\, x \in \Omega, \\
\int |a|(1+|b|)^s \, \mathrm{d}\rho_\epsilon(a,w,b) &\le \|f\|_{\mathcal{B}_s(\Omega)} + \epsilon,
\end{aligned}$$

where we have used the positive homogeneity of ReLU^s and set $w \in \mathbb{S}_{\Omega}^{d-1}$. Let

$$f_e(x) = \int a\sigma(w^T x + b)\chi(w^T x) \, \mathrm{d}\rho_\epsilon(a,w,b) = \int a H_{w,b}(x) \, \mathrm{d}\rho_\epsilon(a,w,b), \quad \forall\, x \in \mathbb{R}^d,$$

where $H_{w,b} : \mathbb{R}^d \mapsto \mathbb{R}$ is the extension of $\sigma_{w,b}$ defined in Lemma 2.3. Then, f_e is an extension of f and satisfies

$$\hat{f}_e(\xi) = \int a\hat{H}_{w,b}(\xi) \, \mathrm{d}\rho_\epsilon(a,w,b).$$

According to Lemma 2.3, we have

$$\begin{aligned}
&\int_{\mathbb{R}^d} (1 + \|\xi\|_{\Omega})^{s-\delta} \big|\hat{f}_e(\xi)\big| \, \mathrm{d}\xi \\
&\le \int |a| \left(\int_{\mathbb{R}^d} (1 + \|\xi\|_{\Omega})^{s-\delta} \big|\hat{H}_{w,b}(\xi)\big| \, \mathrm{d}\xi \right) \mathrm{d}\rho_\epsilon(a,w,b) \\
&\lesssim_s \int |a| \frac{(1+|b|)^s}{\delta} \, \mathrm{d}\rho_\epsilon(a,w,b) \le \frac{1}{\delta}(\|f\|_{\mathcal{B}_s(\Omega)} + \epsilon).
\end{aligned}$$

By the definition of spectral Barron norm, it follows that

$$\|f\|_{\mathcal{F}_{s-\delta}(\Omega)} \lesssim_s \delta^{-1}(\|f\|_{\mathcal{B}_s(\Omega)} + \epsilon).$$

Taking $\epsilon \to 0$ completes the proof. The converse direction follows from [12,24,26]. □

2.2 Proof of Proposition 1.1

Proof. Notice that $f(\cdot)$ can be exactly represented as a two-layer neural network for $x \in [-1,1]$:

$$f(x) = \sigma(1) - \sigma(x) - \sigma(-x).$$

Hence, f is a Barron function and obviously, $\|f\|_{\mathcal{B}_1(\Omega)} \leq 3$.

What remains is to show that

$$\int (1+|\xi|)\big|\hat{f}_e(\xi)\big|\,\mathrm{d}\xi = \infty$$

holds for any extension f_e. Suppose, to the contrary, that there exists an extension f_e such that

$$\int (1+|\xi|)\big|\widehat{f_e}(\xi)\big|\,\mathrm{d}\xi < \infty.$$

Then $\hat{f}_e\,\mathrm{d}\xi$ represents a finite measure over $\mathbb{R}^d$ and f_e is continuous in Ω. By the Fourier inverse theorem, we have

$$f(x) = \int e^{i\xi x}\hat{f}_e(\xi)\,\mathrm{d}x, \quad \forall\, x \in \Omega.$$

For any $x \in (-1/2,0)\cup(0,1/2)$ and sufficiently small δ,

$$\frac{f(x+\delta)-f(x)}{\delta} = \int e^{i\xi x}\frac{e^{i\xi\delta}-1}{\delta}\hat{f}_e(\xi)\,\mathrm{d}\xi. \tag{2.9}$$

The integrand on the right side of (2.9) is bounded by $|\xi||\hat{f}_e(\xi)|$, which is integrable by the assumption. Consequently, by the dominated convergence theorem, for $x \in (-1/2,0)\cup(0,1/2)$, we have

$$f'(x) = \lim_{\delta\to 0}\frac{f(x+\delta)-f(x)}{\delta} = \int e^{i\xi x}\lim_{\delta\to 0}\frac{e^{i\xi\delta}-1}{\delta}\hat{f}_e(\xi)\,\mathrm{d}\xi = \int i\xi e^{i\xi x}\hat{f}_e(\xi)\,\mathrm{d}\xi.$$

Again, by dominated convergence theorem and taking $x \to 0$,

$$\lim_{x\to 0} f'(x) = \lim_{x\to 0}\int i\xi e^{i\xi x}\hat{f}_e(\xi)\,\mathrm{d}\xi = \int \lim_{x\to 0} i\xi e^{i\xi x}\hat{f}_e(\xi)\,\mathrm{d}\xi = i\int \xi\hat{f}_e(\xi)\,\mathrm{d}\xi.$$

This contradicts the fact that $\lim_{x\to 0} f'(x)$ does not exist. □

3 Concluding remark

In this paper, we establish a continuous embedding for Barron-type spaces over compact domains. Crucially, the embedding constants do not depend on the input dimension,

implying that the embedding is effective in high dimensions. We thus establish a more unifying perspective for understanding the high-dimensional approximation of two-layer neural networks. This embedding result has potential implications for the analysis of approximating solutions of high-dimensional PDEs with two-layer neural networks [4, 9, 16].

For future work, it is promising to extend our embedding result to the case of $s \in (0, \infty)$ as discussed in Remark 2.1. Additionally, our proof heavily relies on the positive homogeneity property of the ReLU^s activation function. It would be interesting to extend our analysis to Barron spaces associated with general activation functions [14].

Acknowledgments

We would like to thank Professor Weinan E and Dr. Jihong Long for helpful discussions and to anonymous reviewers for detailed and constructive feedback.

References

[1] A. R. Barron, Universal approximation bounds for superpositions of a sigmoidal function, *IEEE Trans. Inf. Theory*, **39**(3):930–945, 1993.

[2] L. Breiman, Hinging hyperplanes for regression, classification, and function approximation, *IEEE Trans. Inf. Theory*, **39**(3):999–1013, 1993.

[3] A. Caragea, P. Petersen, and F. Voigtlaender, Neural network approximation and estimation of classifiers with classification boundary in a Barron class, *arXiv:2011.09363*, 2020.

[4] Z. Chen, J. Lu, and Y. Lu, On the representation of solutions to elliptic PDEs in Barron spaces, in: *Advances in Neural Information Processing Systems*, **8**:6454–6465, 2021.

[5] G. Cybenko, Approximation by superpositions of a sigmoidal function, *Math. Control Signals Systems*, **2**(4):303–314, 1989.

[6] C. Domingo-Enrich and Y. Mroueh, Tighter sparse approximation bounds for ReLU neural networks, in: *International Conference on Learning Representations*, 2022.

[7] W. E, C. Ma, and L. Wu, A priori estimates of the population risk for two-layer neural networks, *Commun. Math. Sci.*, **17**(5):1407–1425, 2019.

[8] W. E, C. Ma, and L. Wu, The Barron space and the flow-induced function spaces for neural network models, *Constr. Approx.*, **55**:369–406, 2022.

[9] W. E and S. Wojtowytsch, Some observations on high-dimensional partial differential equations with Barron data, in: *Proceedings of Machine Learning Research*, **107**:253–269, 2021.

[10] W. E and S. Wojtowytsch, Representation formulas and pointwise properties for Barron functions, *Calc. Var. Partial Differential Equations*, **61**(2):46, 2022.

[11] W. E and B. Yu, The deep Ritz method: A deep learning-based numerical algorithm for solving variational problems, *Commun. Math. Stat.*, **6**(1):1–12, 2018.

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

[13] B. Li, S. Tang, and H. Yu, Better approximations of high dimensional smooth functions by deep neural networks with rectified power units, *Commun. Comput. Phys.*, **27**(2):379–411, 2019.

[14] Z. Li, C. Ma, and L. Wu, Complexity measures for neural networks with general activation functions using path-based norms, *arXiv:2009.06132*, 2020.

[15] J. Lu and Y. Lu, A priori generalization error analysis of two-layer neural networks for solving high dimensional Schrödinger eigenvalue problems, *Comm. Amer. Math. Soc.*, **2**(1):1–21, 2022.

[16] Y. Lu, J. Lu, and M. Wang, A priori generalization analysis of the deep Ritz method for solving high dimensional elliptic partial differential equations, in: *Proceedings of Machine Learning Research*, **134**:1–46, 2021.
[17] Y. Meng and P. Ming, A new function space from Barron class and application to neural network approximation, *Commun. Comput. Phys.*, **32**(5):1361–1400, 2022.
[18] B. Neyshabur, R. Tomioka, and N. Srebro, Norm-based capacity control in neural networks, in: *JMLR: Workshop and Conference Proceedings*, **40**:1–26, 2015.
[19] G. Ongie, R. Willett, D. Soudry, and N. Srebro, A function space view of bounded norm infinite width ReLU nets: The multivariate case, in: *International Conference on Learning Representations*, 2019.
[20] R. Parhi and R. D. Nowak, Banach space representer theorems for neural networks and ridge splines, *J. Mach. Learn. Res.*, **22**(1):1960–1999, 2021.
[21] A. Pinkus, Approximation theory of the MLP model in neural networks, *Acta Numer.*, **8**:143–195, 1999.
[22] J. W. Siegel and J. Xu, Approximation rates for neural networks with general activation functions, *Neural Netw.*, **128**:313–321, 2020.
[23] J. W. Siegel and J. Xu, Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks, *Found. Comput. Math.*, 1–57, 2022.
[24] J. W. Siegel and J. Xu, Characterization of the variation spaces corresponding to shallow neural networks, *Constr. Approx.*, **57**(3):1109–1132, 2023.
[25] D. So, W. Mańke, H. Liu, Z. Dai, N. Shazeer, and Q. V. Le, Searching for efficient transformers for language modeling, in: *Advances in Neural Information Processing Systems*, **34**:6010–6022, 2021.
[26] J. Xu, Finite neuron method and convergence analysis, *Commun. Comput. Phys.*, **28**(5):1707–1745, 2020.