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: