ETH
Swiss Federal Institute of Technology Zurich

Seminar for Statistics

Department of Mathematics

Master Thesis Fall 2022

Yongqi Wang

Analysis of the Generalization Properties and the Function Spaces Associated with Two-Layer Neural Network Model

Submission Date: March 31th 2023

Adviser: Prof. Dr. Sara van de Geer

Acknowledgements

I would like to express my sincere gratitude to my supervisor Prof. Sara van de Geer for giving me the opportunity to work on this interesting topic and the frequent discussions. I have been very hesitant and afraid over the topic selection at the beginning and I don't think I could put it through without her support. The project familiarize me with functional analysis and some metric spaces, and topology. I can really see my progress and maturity in mathematic statistics.

I would also like to thank Haiyu Chen for the inspiring discussions. I would like Dr. Andre Richter for therapy sessions, which keeps my morale high.

Finally, I am grateful for my parents, my girlfriend Peiying Cai, and my friends for their emotional support when I got stuck, which happens quite a lot during the past months. This work would not have been possible without all the support from them. Thank you.

Abstract

The use of neural network-based models for approximation has proven to be highly effective across various domains. In this thesis, we demonstrate the universal approximation property of two-layer neural networks (2NNs), which allows them to approximate any continuous function on a compact subset in $\mathbb{R}^d$ uniformly well. Additionally, we explore the class of functions that meet a smoothness constraint on their Fourier transform and can be effectively approximated by 2NNs. These functions belong to the Fourier-analytic Barron spaces, which we characterize. We then introduce the concept of infinite-width Barron spaces, where any function can be approximated well by 2NNs through an integral representation. Finally, we draw connections between these spaces and provide a comparative analysis of their properties.

Contents

Notation

The set of all natural numbers without zero is denoted by $\mathbb{N} = \{1, 2, 3, \dots\}$ and the set of all natural numbers including zero is denoted by $\mathbb{N}^0 = \mathbb{N} \setminus \{0\}$.

For an arbitrary set S, we write $|S| := \#S \in \mathbb{N}^0 \bigcup \{\infty\}$ as the number of elements of S.

For a set S in a space P, we write the closure of S in P as $\overline{S}$. The convex hull of S is denoted by $\mathrm{conv}(S)$.

Let $d \in \mathbb{N}$ and $x = (x_1, \dots, x_d)$ be a $\mathbb{R}^d$-valued vector, the p-norm of x is defined as

$$\|x\|_p = \left(\sum_{j=1}^{d} |x_j|^p \right)^{1/p}, \quad 0 < p < \infty.$$

When $p = \infty$, we define

$$\|x\|_\infty := \max_i |x_i| .$$

The standard scalar product of $x, y \in \mathbb{R}^d$ is written as $x^\intercal y = \langle x , y \rangle = \sum_{i=1}^{d} x_i y_i$. $|x|$ and $\|x\|$ is abbreviated for the L^1 norm and L^2 norm for $x \in \mathbb{R}^d$.

For a measure space $(X, \mathcal{M}, \mu)$, let f is a measurable function on X, the $\|\cdot\|_p$ is defined as

$$\|f\|_p = \left(\int |f|^p \, d\mu \right)^{1/p}, \quad 1 \le p < \infty.$$

When $p = \infty$, we define

$$\|f\|_\infty = \inf \left\{ a \ge 0 : \mu(\{x : |f(x)| > a\} = 0) \right\} = \operatorname*{ess\,sup}_{x \in X} |f(x)|.$$

We abbreviate $L^p(X, \mathcal{M}, \mu)$ by $L^p(X)$, or simply L^p when this will cause no confusion. $|f|$ and $\|f\|$ is abbreviated for the L^1 and L^2 norm.

We denote the

$$\|f - g\|_{L^p(X)} := \left(\int_X |f(x) - g(x)|^p \, d\mu \right)^{1/p}, \quad 1 \le p < \infty.$$

When $p = \infty$, we define

$$\|f - g\|_{L^\infty(X)} = \operatorname*{ess\,sup}_{x \in X} |f(x) - g(x)|.$$

We abbreviate it by $\|f - g\|_p$ when this does not cause any confusion regarding the domain of x.

We write $f(x) \lesssim g(x)$ (or simply $f \lesssim g$) if there is a constant $C > 0$, $f(x) \leq C \cdot g(x)$ for all $x \in X$. We will explicitly state the constant if C depends on f or other relevant conditions.

List of abbreviations

2NN Two-Layer Neural Network. v

ANN Artificial Neural Network. 5

CoD Curse of Dimensionality. 1

DAG Directed Acyclic Graph. 5

NN Neural Network. 1

PDE Partial Differential Equation. 1

Chapter 1

Introduction

Neural networks (NNs) have emerged as a standard method for numerical approximation and learning algorithms due to their empirical success in various fields such as computer vision and natural language processing (Shalev-Shwartz and Ben-David, 2014). Although the history of NN dates back to late 1940s, its popularity only emerged from the state-of-the-art performance in a variety of learning domains. However, despite their popularity, NN-based learning is often viewed as "black magic" due to the lack of convincing and rigorous theoretical explanations, particularly in the context of hyperparameter tuning and architecture design. The fundamental challenge of these learning tasks revolves around approximating an unknown complex function from limited observed data points. A comprehensive understanding of the approximation ability of NNs can provide a partial explanation for their success in practice.

The term *curse of dimensionality* (CoD) was coined by Bellman (1952) almost 70 years ago to describe the overwhelming complexity associated with solving a multi-stage processes through dynamic programming. In the scope of approximation theory, it amounts to the exponentially growing number of data points required to maintain the accuracy. Such dimensionality cursed problems have appeared in computational and applied mathematics. For approximations in high dimensions, the accuracy will drop exponentially as dimensionality increases but the deep learning based numerical application methods for partial differential equation (PDE) indicates satisfactory performance in E and Yu (2017); E, Han, and Jentzen (2017); Beck, E, and Jentzen (2019) but there is a absence of rigorous mathematical results to demonstrate this conjecture.

Naturally, the empirical success in numerical applications calls for a theoretical explanation. The understanding of the advantages of using NN over traditional methods as an approximation tool is crucial for explaining their performance. Classical approximation methods include polynomials, wavelets, splines, and sparse approximation from bases, and dictionaries (DeVore, 1998). Additionally, it can facilitate the integration of their use in various domains, such as numerical methods for solving PDEs.

The variety of architectures in NN presents a challenge in characterizing their approximation properties due to the wide variety of architectural choices available, such as the width, depth, activation functions, and connectivity. To address this challenge, two main directions concerning the approximation power of NN have emerged in the community. Let $\mathcal{M}^{W,L}(\sigma)$ be the collections of the outputs of NN of width W, depth L, and a activation function σ. Similar to a classical approximation scheme, one can consider the

output of 2NN, $\mathcal{M}^{W,1}(\sigma)$, where the depth is fixed at $L = 1$ and W can grow to infinity. Alternatively, one can investigate deep neural networks by fixing the width and allowing the depth to increase to infinity. We will not dig too deeply into the deep neural networks and this thesis is confined on the performance of 2NN as an approximation tool.

One of the first question in the approximation of 2NN is the question of *density*: what conditions must 2NN satisfy in order that an arbitrary target function can be approximated arbitrarily well by 2NN:

For any $f \in C(\mathbb{R}^d)$, any compact set U of $\mathbb{R}^d$, and an arbitrary accuracy $\epsilon > 0$, there is a g produced by 2NN, $g \in \mathcal{M}^{W,1}(\sigma)$ such that

$$\sup_{x \in U} |f(x) - g(x)| < \epsilon. \tag{1.1}$$

Since the late 1980s, it has been established that NNs are universal approximators (Carroll and Dickinson, 1989; Cybenko, 1989; **?**; Funahashi, 1989). Using the Hahn-Banach Theorem and the Riesz Representation Theorem, Cybenko (1989) stated the density property required is σ is sigmoidal, (defined in (2.6)). Independent from Cybenko, Funahashi's proof uses the result of Irie and Miyake (1988) on the integral representation L^1 functions, using a kernel which can be expressed as a difference of two sigmoidal functions. In paper by **?**, the activation function needs to monotone and bounded, which allows noncontinuous σ. Additionally, the summation and product of activation functions are allowed in their statement. In Jones (1992), a constructive method with a bounded sigmoidal function is sufficient to ensure the density. A sequence of papers adopted various techniques to attack this problem and yet the answer is surprisingly simple, as shown by Leshno, Lin, Pinkus, and Schocken (1993). The necessary and sufficient condition on σ for $\mathcal{M}^{W,1}(\sigma)$ to be *dense* is that σ must not be a polynomial.

While the density question is an essential step in understanding the approximation power of NNs, this fact alone does not provide a sufficient explanation for why NNs are more effective than traditional approximation methods, as methods using polynomials, splines, and wavelets are also universal approximators. In particular, we need to consider the degree of approximation as well as the possibility of developing stable algorithms to find the approximates. The effectiveness of NN-based learning in high-dimensional settings remains a mystery. To gain insights into the approximation abilities of 2NN, one can estimate the worst-case error rate of approximation by measuring the performance of the network on a target function f from a classical model class. One then tries to address the following problem 1.

Problem 1. Given a target function f in one of the classical model class such as unit balls of Lipschitz, Sobolev and Besov space, find the upper and lower bound for the approximation error with 2NN.

A different approach is to describe the class of functions that are *well approximated* by 2NN. This is generally more difficult and less straightforward. By the success in practical numerical experiments, one could assume that the model class should be quite large.

Problem 2. Describe the model classes of functions that are guaranteed to be well approximated by NN.

One of the celebrated results was introduced by Prof. Andrew Barron (1994). Given a domain $U \subset \mathbb{R}^d$, the model class K consists of all functions f in $L^2(U)$ whose Fourier

transform $\mathcal{F}(f)$ satisfies

$$\int_{\mathbb{R}^d} |\omega| \, |\mathcal{F}(f)(\omega)| \, d\omega < \infty. \tag{1.2}$$

Initially, the approximation error was obtained for any sigmoidal activation functions for L^2 norm on U. For any $n \in \mathbb{N}$, there exists a a linear combination of sigmoidal functions $f_n = \sum_{j=1}^n a_j \sigma(b_j{}^\intercal x + c_j), a_j, c_j \in \mathbb{R}, b_j \in \mathbb{R}^d$ such that

$$\|f - f_n\|_{L^2(\Omega)} \leq C n^{-1/2}, \quad f_n \in K. \tag{1.3}$$

This result clearly extend to the *rectified linear unit* (ReLU) activation as wells since $\sigma_{\mathrm{ReLU}}(x) - \sigma_{\mathrm{ReLU}}(x-1)$ is a sigmoidal function.

Inspired by Barron's result (E and Wojtowytsch, 2020; Caragea, Petersen, and Voigtlaender, 2022), many generalizations and improvements have been made. A important generalization is given by Makovoz (1996) where improved error rate is $\mathcal{O}(n^{-1/2-1/p\cdot d})$. This improvement relies on the fact that one can reduce the number of terms needed in the approximation.

As noted earlier, there is much interest in identifing new classes of functions associated with 2NN. To honor Prof. Andrew Barron's contribution in the understanding of neural nets, the term *Barron space* was coined by many but the notation of Barron spaces is not in consensus within the community, and different terms have been given to describe the same model classes or spaces. For the function spaces in which functions have finite Fourier moments, Xu (2020) calls this model classes *Barron spectral spaces* while Caragea et al. (2022) refers to *Fourier-analytic Barron space*. For function spaces in which functions admit an integral representation with a ReLU activation function, E, Ma, and Wu (2021) refer them simply as *Barron spaces* of different orders $p \in \{1, 2, 3, \cdots, \infty\}$. In some literature including Caragea et al. (2022), these spaces are named *infinite-width Barron spaces* associated with different activation functions and the term *classical Barron space* is reserved for those associated with Heaviside function[1].

To avoid confusion, we will use two definitions throughout

- Fourier-analytic Barron spaces.
- infinite-width Barron spaces.[2]

Recently, Siegel and Xu (2022a) connect the variation space introduced by Parhi and Nowak (2021, 2022). It has been shown that the spectral Barron norm in is equivalent to the variation norm of a dictionary. This allows us to properly frame these results in the context of nonlinear approximation. The compactness or smoothness condition can then be used for calculating the metric entropy of the closed convex hull of said dictionaries.

The rest of this thesis is organized as follows. Chapter 2 introduces the question setup and the 2NN model. The question of *density* with 2NN is answered for any continuous functions from $\mathbb{R}^d$ to $\mathbb{R}$. We show that the 2NN are universal approximators with a activation function that is not a polynomial when the number of nodes are allowed to grow unlimited. In Chapter 3, we consider the model classes of functions (Fourier-analytic Barron spaces) that are guaranteed to be *well approximated* by 2NN. The smoothness restriction

[1] Also called step function or unit step function

[2] We limit the model classes to those associated with ReLU only. In the case of Heaviside function, we denote the space still by the term *classical Barron space*.

is enforced by bounding the Fourier transform term of the functions. Chapter 4 describes the Infinite-width Barron spaces in which each function is also well approximated with the help of an integral representation. Furthermore, the connection between these spaces and their variation spaces is included. Each corresponding variation space is constructed using a compact dictionary of functions. Moreover, we compare and articulate the relationship between the model classes.

Chapter 2

Two-layer neural networks and nonlinear approximation

This chapter introduces the 2NN model and explores fundamental findings concerning n-term approximation using a dictionary. Section 2.1 presents an overview of the 2NN model and its basic properties. In section 2.2, we delve into the well-known results regarding n-term approximation from a dictionary. Finally, section 2.3 outlines the problem of *density* around approximation with 2NN.

The aim of this chapter is to summarize well-established results that will be utilized in subsequent sections of this thesis. For an in-depth introduction, we recommend reading DeVore (1998); Pinkus (1999).

2.1 What is a two-layer neural network

This section provides an introduction to *feed-forward neural networks* and their basic properties, with a focus on 2NN. While recurrent neural networks and other architectures are commonly used for specific applications, fully connected feed-forward neural networks offer a convenient way to balance model complexity and approximation efficiency. This is especially the case in 2NN: complexity is solely determined by the width or the number of nodes n. We will begin by defining 2NN and then discuss their elementary properties.

This section is based on DeVore, Hanin, and Petrova (2021); Shalev-Shwartz and Ben-David (2014).

2.1.1 Feed-forward neural network

The class of artificial neural networks (ANN) known as feed-forward neural networks, denoted by $\mathcal{NN}$, is characterized by a directed acyclic graph (DAG)

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}).$$

The *architecture* of $\mathcal{NN}$ is represented by the associated graph $\mathcal{G}$, which consists of a finite collection of vertices $\mathcal{V}$ and a finite set of edges $\mathcal{E}$. A computation unit associated with each $v \in \mathcal{V} \setminus \mathcal{I}$ is known as a *node* or *neuron*. The variables $\mathcal{I}$ and $\mathcal{O}$ represent the input and output layers of the neural network, respectively. The following basic properties holds for $\mathcal{NN}$:

1. For each vertex $v \in \mathcal{V} \setminus \mathcal{I}$, there is an associated activation function $\sigma : \mathbb{R} \to \mathbb{R}$.
2. For each edge $e \in \mathcal{E}$, there is an associated scalar weight $w_e \in \mathbb{R}$.

The nodes in the input layer receive scalar inputs, which are then observed by downstream nodes. Apart from the nodes in $\mathcal{I}$, each node in $\mathcal{NN}$ takes a superposition of inputs from its upstream nodes, mediated by the associated weights w_e. The edges $e = (v, v')$ between a vertex v and its upstream vertices v' correspond to these weights.

The output function of $\mathcal{NN}$ can be seen as a mapping from $\mathbb{R}^d$ to $\mathbb{R}^{d'}$, $d = |\mathcal{I}|, d' = |\mathcal{O}|$. The family of functions produced by $\mathcal{NN}$, given a fixed architecture $\mathcal{G}$, is determined by the trainable parameters, denoted by $\Theta := \{w_e, b_v\}, e \in \mathcal{E}, v \in \mathcal{V} \setminus \mathcal{I}$.

2.1.2 Fully connected neural networks

The general definition previously presented covers virtually all networks architecture encountered in practice. However, in the following chapters, we will focus specifically on a specialized class of neural networks known as *fully connected neural networks*, where the vertices are organized into layers.

In a fully connected network, each vertex is connected exclusively to all vertices in the next layer via edges and to no other layers. The input layer, $\mathcal{I}$, consists of d input vertices, where each vertex receives an external scalar signal $x_i \in \mathbb{R}, i \in \{1, \ldots, d\}$. The combined input $x := (x_1, \ldots, x_d) \in \mathbb{R}^d$ serves as the independent variable for the output function $f_{\mathcal{NN}}$. The input layer is followed by L hidden layers, with L representing the *depth* of the network. Each hidden layer contains n_j hidden vertices, where n_j is the *width* of the jth layer, $j = 1, 2, \ldots, L$.

Each hidden node is associated with an activation function σ, while nodes at the output layer $\mathcal{O}$ take the identity function to ensure that $f_{\mathcal{NN}}(x)$ is a linear combination of nodes at the Lth layer plus a bias term. Consequently, for a fixed architecture, the *weight matrices* and *bias vectors* alone suffice to describe the output function $f_{\mathcal{NN}}(x)$,

$$W^{(l)} \in \mathbb{R}^l \times \mathbb{R}^{l-1}, b^{(l)} \in \mathbb{R}^l \quad l = 1, 2, \ldots, L+1. \tag{2.1.1}$$

Let $X^{(l)} \in \mathbb{R}^l$ be the outputs of layer l, the output function is

$$f_{\mathcal{NN}}(x) := W^{L+1} X^L + b^{L+1} \tag{2.1.2}$$

and each layer is calculated recursively

$$X^{(0)} = x \in \mathbb{R}^d, \quad d = |\mathcal{I}| \tag{2.1.3}$$

$$X^{(l)} = \sigma(W^{(l)} X^{(l-1)} + b^{(l)}), \quad l = 1, 2, \ldots, L+1, \tag{2.1.4}$$

Here the activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is assumed to be coordinate-wise or element-wise independent:

$$\sigma(x) = (\sigma(x_1), \sigma(x_2), \ldots, \sigma(x_d)),$$

for $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d, d \in \mathbb{N}$.

2.1.3 The family of functions from neural networks

Throughout the following discussion, we will assume that each layer of the fully connected neural network has an identical width, that is,

$$n_1 = n_2 = \cdots = n_L = W.$$

A NN of different width can be embedded into a wider NN of fixed width

$$W := \max_{j=1,\ldots,L} n_j.$$

This can be achieved simply by inserting zero bias terms in each bias vector $b^{(l)}$ in each lth layer and adding zero-weighted edges between layers. To account for the newly added vertices, new edges with a weight of 0 are added, connecting to the forthcoming $(j+1)$th layer. The fully connected neural networks we discuss will be of a fixed width.

Consider a fully connected neural network with width W and depth L, and let σ be the activation function. If d and d' are the input and output dimension, respectively, we denote the set of all possible functions $f : \mathbb{R}^d \to \mathbb{R}^{d'}$ produced by a fixed architecture with trainable parameters as

$$\mathcal{M}^{W,L}(\sigma). \tag{2.1.5}$$

The number of trainable parameters is given by

$$n(W, L) := (d+1)W + W(W+1)(L-1) + d'(W+1). \tag{2.1.6}$$

Remark. It is worth noting that $\mathcal{M}^{W,L}(\sigma)$ is not a linear space, even when $L = 1$, as it is not closed under addition of functions. For example, there exist functions $f_1, f_2 \in \mathcal{M}^{W,L}(\sigma)$ such that their sum $f_1 + f_2$ does not belong to $\mathcal{M}^{W,L}(\sigma)$.

In what follows, we only consider the specialized case of 2NN ($L = 1$), and the set of functions by 2NN is denoted by

$$\mathcal{M}^{W,1}(\sigma) = \mathcal{M}(\sigma) = \left\{ b_0 + \sum_{j=1}^{W} a_j \sigma({b_j}^{\intercal} x + c_j), \quad a_j, c_j \in \mathbb{R}, b_j \in \mathbb{R}^d \right\}. \tag{2.1.7}$$

Here, W refers to the width of the 2NN model, i.e. the number of nodes. When discussing complexity in approximation, n is usually used to represent the number of nodes required for a certain level of accuracy.

2.2 n-term approximation

This section introduces some fundamental findings regarding n-term approximations derived from a dictionary. Nonlinear approximation problems are typically presented as follows: given a target function $f \in \mathcal{H}$, select a basis and an n-term approximation to f from that basis. The class of bases is known as *library*, as seen in wavelets or splines approximation. Our focus is on a closely related form of approximation: n-term approximation from a dictionary $\mathbb{D} \subset \mathcal{H}$. The dictionary $\mathbb{D}$ can be an arbitrary subset but its choice is subject to practical computational limitations.

This section is mainly based on (DeVore, 1998, Section 8) and van der Vaart and Wellner (1996), unless stated otherwise.

Let $\mathbb{D} = \{d_1, d_2, \dots\}$ be a uniformly bounded domain in a Hilbert space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$

$$\sup_{d_j \in \mathbb{D}} \|d_j\|_{\mathcal{H}} < \infty \quad \forall d_j \in \mathbb{D}. \tag{2.2.1}$$

For every $n \in \mathbb{N}$, the collection of all functions in $\mathcal{H}$ which can be expressed as a linear combination of at most n elements of $\mathbb{D}$ is denoted by

$$\Sigma_n(\mathbb{D}) = \left\{ \sum_{j=1}^{n} a_j d_j, \quad d_j \in \mathbb{D}, a_j \in \mathbb{R} \right\}. \tag{2.2.2}$$

Any function $f_n \in \Sigma_n(\mathbb{D})$ with $n \in \mathbb{N}$ is represented as

$$f_n = \sum_{j=1}^{n} a_j d_j, \quad d_j \in \mathbb{D}. \tag{2.2.3}$$

To include the control over the coefficients, also called *outer parameters*, α_j in the above expansion, we introduce

$$\Sigma_n^t(\mathbb{D}, M) := \left\{ f = \sum_{j=1}^{n} \alpha_j d_j : d_j \in \mathbb{D} \text{ and } \sum_{j=1}^{n} |\alpha_j|^t \leq M^t, \quad n \in \mathbb{N}, \alpha_j \in \mathbb{R} \right\} \tag{2.2.4}$$

for any $t \in \mathbb{N} \bigcup \{\infty\}$ and any $M > 0$.

Let $K_n^t(\mathbb{D}, M)$ be the closure of $\Sigma_n^t(\mathbb{D}, M)$ in $\mathcal{H}$

$$K_n^t(\mathbb{D}, M) := \overline{\Sigma_n^t(\mathbb{D}, M)}. \tag{2.2.5}$$

It should be noted that $t = 1$, $K_n^1(\mathbb{D}, M)$ is the class of functions that are a signed *convex* combination of elements in $\mathbb{D}$. When $t = \infty$, $K_n^\infty(\mathbb{D}, M)$ corresponds to the sets whose coefficients are bounded in L^∞.

Next, we define the union for all $M > 0$

$$K_n^t(\mathbb{D}) = \bigcup_{M>0} K_n^t(\mathbb{D}, M). \tag{2.2.6}$$

We define a seminorm for functions $f \in K^t(\mathbb{D})$

$$|f|_{K^t(\mathbb{D})} = \inf\{M > 0 : f \in K_n^t(\mathbb{D}, M)\}. \tag{2.2.7}$$

For $f \in \mathcal{H}$, the approximation error is defined as

$$e_n(f, \mathbb{D}, \mathcal{H}) := \sup_{g \in \Sigma_n(\mathbb{D})} \|f - g\|_{\mathcal{H}}. \tag{2.2.8}$$

where $\Sigma_n(\mathbb{D})$ is the collection of all functions in $\mathcal{H}$ that can be expressed as a linear combination of at most n elements of $\mathbb{D}$ (2.2.2).

Assuming that approximation holds with a dictionary $\mathbb{D}$ in $\mathcal{H}$ such that for any function $f \in \mathcal{H}$, the approximation error is

$$e_n(f, \mathbb{D}, H) \leq n^{-\alpha} C_{\mathbb{D}}. \tag{2.2.9}$$

We are concerned with the coefficients in the exponents α and what $C_{\mathbb{D}}$ is dependent on.

2.2.1 Maurey's Theorem

Approximation with a finite linear combination of elements from a bounded dictionary in a Hilbert space $\mathcal{H}$ achieves an error rate of $\mathcal{O}(n^{-1/2})$ by Pisier (1980) and this holds for any bounded $\mathbb{D} \subset \mathcal{H}$.

Definition 2.1 (Covering numbers and entropy)**.** *Let $(\mathcal{F}, \|\cdot\|_{\mathcal{F}})$ be a subset of a normed space. The covering number $N(\epsilon, \mathcal{F}, \|\cdot\|_{\mathcal{F}})$ is the minimal number of balls $\{g : \|f - g\|_{\mathcal{F}} < \epsilon\}$ of radius ϵ required to cover the set $\mathcal{F}$.*

Definition 2.2 (n-th dyadic entropy number)**.** *The entropy is the logarithm of the covering number. For $n \in \mathbb{N}$, the n-th (dyadic) entropy number is given by*

$$\varepsilon_n(\mathcal{F}) = \inf\{\epsilon > 0 : \mathcal{F} \text{ is covered by } 2^n \text{ balls of radius } \epsilon\}. \tag{2.2.10}$$

Theorem 2.3 (Maurey's Theorem)**.** Let $\mathbb{D} = \{d_1, d_2, \dots\}$ be a subset of a Hilbert space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$. $\mathbb{D}$ is uniformly bounded

$$\sup_{d \in \mathbb{D}} \|d\|_{\mathcal{H}} < \infty. \tag{2.2.11}$$

For every $f \in \mathcal{H}$ of the form

$$f = \sum_{j=1}^{m} c_j d_j, \quad \sum_{j=1}^{m} |c_i| < \infty, \quad c_j \in \mathbb{R}, m > 0 \tag{2.2.12}$$

and every $n \in \mathbb{N}$, there exists a $g_n = \sum_{j=1}^{n} a_j d_j$ with at most n non-zero coefficients with

$$\sum_{j=1}^{n} |a_j| \leq \sum_{j=1}^{n} |c_j|, \quad a_j \in \mathbb{R} \tag{2.2.13}$$

such that

$$\|f - g_n\|_{\mathcal{H}} \leq 2\varepsilon_n(\mathbb{D}) \cdot n^{-\frac{1}{2}} \cdot \sum_{j=1}^{n} |c_j|. \tag{2.2.14}$$

Proof. Without loss of generality, we can assume $m < \infty$ and the coefficients c_j in (2.2.12) are non-negative and the sum $\sum_{j=1}^{m} |c_j| = 1$.

We partition the index set $I = \{1, 2, \cdots, m\}$ into n nonempty subsets I_p, $p = 1, 2, \cdots, n$ such that each subset $\mathbb{D}_p$ of $\mathbb{D}$ has a diameter smaller than ϵ.

For each partition $p = 1, 2, \cdots, n$, $n_p := |I_p|$. Let $S_p = \sum_{i \in I_p} c_i$ and

$$\hat{f}_p := \frac{S_p}{n_p}(\hat{d}_1^{(p)} + \cdots + \hat{d}_{n_p}^{(p)}) \tag{2.2.15}$$

$$\hat{f} := \sum_{p=1}^{n} \hat{f}_p \tag{2.2.16}$$

where each $\hat{d}_k^{(p)}$ is identically distributed and pairwise independent from $\mathbb{D}_p$ and the probability of equalling to $d_i \in \mathbb{D}_p$ is $\frac{c_i}{S_p}$.

$$\mathrm{P}\left[\hat{d}_k^{(p)} = d_i\right] = \frac{c_i}{S_p}, \quad i \in I_p, p = 1, 2, \cdots, n. \tag{2.2.17}$$

Next, we can show that

$$\mathbf{E}\left[\hat{f}_p\right] = \frac{S_p}{n_p}\sum_{k}^{n_p}\mathbf{E}\left[\hat{d}_k^{(p)}\right] = \frac{S_p}{n_p}\sum_{i\in I_p}\frac{c_i}{S_p}d_i = f_p \tag{2.2.18}$$

$$\mathbf{E}\left[\left\|f-\hat{f}\right\|^2\right] = \sum_{p}^{n}\mathrm{Var}\left(f_p-\hat{f}_p\right) = \sum_{p}^{n}\mathrm{Var}\left(\hat{f}_p\right) \tag{2.2.19}$$

As each partition I_p, the $\mathbb{D}_p$ has a diameter of smaller than ϵ which implies variance of $\hat{d}_p^{(p)}$ is controlled by ϵ^2, we have

$$\mathbf{E}\left[\left\|f-\hat{f}\right\|^2\right] = \sum_{p}^{n}\mathrm{Var}\left(\hat{f}_p\right) \le \sum_{p}^{n}\frac{\epsilon^2}{n}S_p = \epsilon^2/n. \tag{2.2.20}$$

In conclusion, there must exist some realization $\hat{f}$ such that $\left\|f-\hat{f}\right\|^2$ is smaller than ϵ^2/n and ϵ can be chosen close to the entropy. Such realization is a linear combination of at most $2n$ elements from $\mathbb{D}$. □

Remark. This is refinement of the result in Pisier (1980) in a Hilbert space. The approximation rate by Maurey is not always sharp and one can use the compactness of the dictionary $\mathbb{D}$ to improve it.

Corollary 2.4. If f is in the closure of the convex hull of a set $\mathbb{D}$ in a Hilbert space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ and every function $d \in \mathbb{D}$ is bounded, then for every $n \in \mathbb{N}$, there is an f_n in the convex hull of n elements in $\mathbb{D}$ such that

$$\|f-f_n\|_{\mathcal{H}} \le Cn^{-\frac{1}{2}} \tag{2.2.21}$$

where $f_n = \sum_{j=1}^{n}\alpha_j\cdot d_j, \quad d_j\in\mathbb{D}, \quad \sum_{j=1}^{n}\alpha_j = 1$, and $C^2 \ge \sup_{d\in\mathbb{D}}\|g\|_{\mathcal{H}}^2 - \|f\|_{\mathcal{H}}^2$.

2.2.2 Iterative approximation

Jones (1992) has also shown the same approximation error $\mathcal{O}(n^{-1/2})$ using an iterative argument. One can find a proof in (Jones, 1992, p. 611) and a refinement of Jone's result in (Barron, 1993, Theorem 5).

Let $\mathbb{D}$ be a nonempty bounded subset of a Hilbert space $\mathcal{H}$. Suppose a sequence of function f_n were used to approximate a element f in $\mathcal{H}$ with the following algorithm

$$f_n = \alpha_n + f_{n-1} + (1-\alpha_n)d_n, \quad d_n\in\mathbb{D}, \alpha_n\in[0,1], n\ge 1. \tag{2.2.22}$$

The iteration starts with $f_1 = d_1 \in \mathbb{D}$ and $\alpha_1 = 0$ and the sequence $\{f_n\}$ defined as above is in the convex hull of the points $\{d_1,\ldots,d_n\}$.

Theorem 2.5 (Jone's Theorem)**.** Let $\mathbb{D}$ be a nonempty bounded domain of a Hilbert space $\mathcal{H}$. Let $C_{\mathbb{D}} = \sup_{d\in\mathbb{D}}\|d\|^2 < \infty$ and set $C_f = C_{\mathbb{D}} - \|f\|^2$. If f is in the closure of the convex hull of $\mathbb{D}$, i.e. $f\in\mathrm{conv}(\mathbb{D})$. Then for every $n\in\mathbb{N}$, f_n is chosen iteratively as

$$\|f_n - f\|^2 \le \inf_{\alpha\in[0,1]}\inf_{d\in\mathbb{D}}\|\alpha f_{n-1} + (1-\alpha)d - f\|^2 + e_n \tag{2.2.23}$$

where $e_n \le \frac{C}{n(n+C/C_f-1)}$ for some $C > C_f$. Then approximation error for each $n\in\mathbb{N}$ is $\mathcal{O}(n^{-1/2})$.

Proof. Set $f_\delta = \sum_{j=1}^m a_j d_j$ be a convex combination of m elements from $\mathbb{D}$ for some $\delta > 0$ such that f and f_δ is close

$$\|f - f_\delta\| \leq \delta. \tag{2.2.24}$$

We have

$$\|\alpha(f_{n-1} - f) + (1-\alpha)(d - f_\delta)\|^2 = \alpha^2 \|f_{n-1} - f\|^2 + (1-\alpha)^2 \|d - f_\delta\|^2 \tag{2.2.25}$$
$$+ 2\alpha(1-\alpha) \langle f_{n-1} - f \, , d - f_\delta \rangle \tag{2.2.26}$$

We can then calculate the average of terms containing d

$$\sum_{j=1}^m a_j (1-\alpha)^2 \|d_j - f_\delta\|^2 + 2a_j\alpha(1-\alpha) \langle f_{n-1} - f \, , d_j - f_\delta \rangle \tag{2.2.27}$$
$$= (1-\alpha)^2 \sum_{j=1}^m a_j \|d_j - f_\delta\|^2 + 2\alpha(1-\alpha) \left\langle f_{n-1} - f \, , \sum_{j=1}^m a_j d_j - f_\delta \right\rangle \tag{2.2.28}$$
$$= (1-\alpha)^2 \sum_{j=1}^m a_j \|d_j - f_\delta\|^2 + 0 \tag{2.2.29}$$
$$= (1-\alpha)^2 \sum_{j=1}^m a_j \|d_j\|^2 - \|f_\delta\|^2 \tag{2.2.30}$$
$$\leq (1-\alpha)^2 (C_{\mathbb{D}} - \|f_\delta\|^2). \tag{2.2.31}$$

This implies that there must exist a $d \in \mathbb{D}$ such that

$$\|\alpha(f_{n-1} - f) + (1-\alpha)(d - f_\delta)\|^2 \leq \alpha^2 \|f_{n-1} - f\|^2 + (1-\alpha)^2 \|(d - f_\delta)\|^2 + \delta. \tag{2.2.32}$$

By triangle inequality and setting $\delta \to 0$, we have

$$\inf_{g \in \mathbb{D}} \|\alpha f_{n-1} + (1-\alpha)d - f\|^2 = \inf_{g \in \mathbb{D}} \|\alpha(f_{n-1} - f) + (1-\alpha)(d - f)\|^2 \tag{2.2.33}$$
$$\leq \alpha^2 \|f_{n-1} - f\|^2 + (1-\alpha)^2 C_f. \tag{2.2.34}$$

Then proof is complete when we set $\alpha = \frac{C_f}{C_f + \|f_{n-1} - f\|^2}$. □

2.3 Universal approximation theorem

Neural networks, including those with a single hidden layer, such as 2NN, have the ability to approximate any continuous function f on a compact set up to an arbitrary precision. This is known as universal approximation, which can be achieved with mild restrictions on the activation functions σ. For functions of d-variables, a linear combination of functions from the dictionary

$$\mathbb{D} = \{\sigma(b^\intercal x + c) : b \in \mathbb{R}^d, c \in \mathbb{R}\} \tag{2.3.1}$$

can be used to obtain a good approximation to f. These functions in $\mathbb{D}$, also known as ridge functions or planar waves, are subject to the requirements on σ.

However, the problem of complexity in approximation remains a challenge, as it is unclear how many nodes n in the hidden layer are required to achieve a desired degree of approximation. While n is allowed to grow uncontrolled, one is unable to obtain quantitative information about the error rate. Despite this limitation, this section will present the univariate approximation theorem based on Cybenko (1989), unless otherwise specified.

Definition 2.6 (sigmoidal function)**.** *A function σ is* ***sigmoidal*** *if*

$$\sigma(t) = \begin{cases} 1 & as \quad t \to +\infty \\ 0 & as \quad t \to -\infty. \end{cases} \tag{2.3.2}$$

Notations and setup

Let $I_d = [0,1]^d$ denote the d-dimensional unit cube, and the space of continuous functions over I_d is denoted by $C(I_d)$. We denote the supremum norm of a function $f \in C(I_d)$ by

$$\|f\|_\infty = \sup_{x \in I_d} |f(x)| . \tag{2.3.3}$$

We use $M(I_d)$ to denote the space of finite, signed regular Borel measures, i.e. $\mu(A) \in \mathbb{R}$ for all Borel sets $A \in I_d$ and $\mu(\emptyset) = 0$. We refer readers to Rudin (1991, 1987) for a detailed presentation of functional construction used and we include some basic materials in Appendix A.

Let $h(x)$ be a linear combinations of elements from the dictionary $\mathbb{D}$ (2.3.1)

$$h(x) \in \Sigma_n(\mathbb{D}_\sigma) = \Sigma_n^\sigma := \left\{ \sum_{j=1}^n a_j \sigma(b_j{}^\intercal x + c_j), \quad a_j, c_j \in \mathbb{R}, b_j \in \mathbb{R}^d \right\} \tag{2.3.4}$$

where $n < \infty$, and σ is a univariate function from $\mathbb{R}$ to $\mathbb{R}$.

The main contribution of approximation theorem is the statement on the conditions of σ such that the above finite linear combination $h(x)$ is dense in $C(I_d)$ with respect to the supremum norm. It should also be stressed that there is no restriction for the number of combinations. Here the set $\Sigma_n(\mathbb{D}_\sigma)$ is the set $\mathcal{M}(\sigma)$ in (2.1.7) where $n = W$ and b_0 is omitted.

Theorem 2.7 (Universal approximation theorem)**.** If σ is sigmoidal as defined in Definition 2.6, then any function $f \in C(I_d)$ be approximated uniformly well by a finite linear combination of ridge functions of the form (2.3.4).

In other words, for any function $f : I_d \to \mathbb{R}$ and any $\epsilon > 0$, there exists a $f_n \in \Sigma_n^\sigma$ such that

$$\|f - f_n\|_{L^\infty(I_d)} < \epsilon, \quad n \in \mathbb{N}. \tag{2.3.5}$$

Theorem 2.8. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous function. We consider the set

$$\Sigma_n^\sigma = \left\{ \sum_{j=1}^n \alpha_j \sigma(b^\intercal x + c), \quad b_j \in \mathbb{R}^d, \alpha_j, c_j \in \mathbb{R}, n \in \mathbb{N} \right\}. \tag{2.3.6}$$

For any function $f \in C(\mathbb{R}^d)$, and any compact set U of $\mathbb{R}^d$, and any $\epsilon > 0$, there exists a $f_n \in \Sigma_n^\sigma$ such that

$$\|f - f_n\|_{L^\infty(U)} < \epsilon \tag{2.3.7}$$

if and only if σ is not a polynomial. In other words, Σ_n^σ is dense in the space $C(\mathbb{R}^d)$ in the topology of uniform convergence on compact sets.

Remark. The requirements for the set Σ_n^σ constructed using a function σ is surprisingly simple and it is proven in Theorem 1 (Leshno et al., 1993, p. 10) that σ simply need not to be a polynomial.

The main structure of the proof for Theorem 2.7 is as follows:

1. Any finite sums of the form (2.3.4) with a discriminatory function σ are dense in $C(I_d)$ with respective to the supremum norm;
2. Any bounded sigmoidal function is discriminatory.

Definition 2.9 (Discriminatory function)**.** *A function σ is **discriminatory** if for every measure $\mu \in M(I_d)$,*

$$\int_{I_d} \sigma(b^\intercal x + c)\, d\mu(x) = 0 \quad \forall\, b \in \mathbb{R}^d \text{ and } c \in \mathbb{R} \tag{2.3.8}$$

implies $\mu = 0$.

We continue to show that the linear span of any continuous discriminatory functions are dense in the space of $(C(I_d), \|\cdot\|_\infty)$.

Theorem 2.10. Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a continuous discriminatory function, the finite sums of the form (2.3.4) are dense in the space $(C(I_d), \|\cdot\|_\infty)$. In other words, for any $\epsilon > 0$ and any $f \in C(I_d)$, there exists a $n \in \mathbb{N}$ and a sum h of the above form (2.3.4), where

$$\|f - h\|_{L^\infty(I_d)} < \epsilon. \tag{2.3.9}$$

Proof. Let $G := \operatorname{span}(\{\sigma(b^\intercal x + c) : b \in \mathbb{R}^d, c \in \mathbb{R}\})$ be the linear span for every $b \in \mathbb{R}^d, c \in \mathbb{R}$. G clearly is a linear subspace of $C(I_d)$. We claim that the closure of G, $\overline{G}$, is all of $C(I_d)$.

We continue the proof by contradiction. Assuming $\overline{G}$ is not $C(I_d)$, then there is a bounded linear functional L on $C(I_d)$ such that $L \not\equiv 0$ on $C(I_d)$ and $L(G) = L(\overline{G}) = 0$ by the Hahn-Banach Theorem A.8.

By the Riesz Representation Theorem, there is a unique $\mu \in M(I_d)$ for this L such that

$$L(f) = \int_{I_d} f(x)\, d\mu(x) \quad \forall f \in C(I_d) \tag{2.3.10}$$

Since L is identically zero on G, we must have for all b and c that

$$\int_{I_d} \sigma(b^\intercal x + c)\, d\mu(x) = 0 \tag{2.3.11}$$

However, the condition that σ is discriminatory implies $\mu = 0$ and consequently $L = 0$. By (A.9), subspace G must be dense in $C(I_d)$. □

Now it remains to show that sigmoidal functions are discriminatory with the help of Lemma 2.11.

Lemma 2.11. Rudin (1991) if μ is a signed finite Borel measure on $\mathbb{R}^d$ such that the Fourier transform of μ

$$\mathcal{F}(\mu)(u) = \int_{\mathbb{R}^d} e^{-iu^\intercal x}\, \mu(x) = 0, \tag{2.3.12}$$

for all $x \in \mathbb{R}^d$, then $\mu = 0$ for all measurable sets of $\mathbb{R}^d$.

Lemma 2.12. Any bounded, measurable sigmoidal function is discriminatory.

Proof. **Step 0** *(Assume discriminatory and construct pointwise convergence function)*: Let $\sigma : \mathbb{R} \to \mathbb{R}$ be a sigmoidal function. Assume that the σ is discriminatory with a measure $\mu \in M(I_d)$ as in Definition 2.9 and the goal is to show that $\mu = 0$.

Fix an arbitrary $b_0 \in \mathbb{R}^d \setminus \{0\}$ and define $\sigma_\lambda(x) := \sigma(\lambda(b_0{}^\intercal x + c) + \varphi)$. For any $c, \lambda, \varphi \in \mathbb{R}$, one can write

$$\sigma_\lambda(x) = \begin{cases} \to 1, & \text{for } b_0{}^\intercal x + c > 0 \quad \text{as } \lambda \to +\infty \\ \to 0, & \text{for } b_0{}^\intercal x + c < 0 \quad \text{as } \lambda \to -\infty \\ \sigma(\varphi), & \text{for } b_0{}^\intercal x + c = 0, \quad \forall \lambda \in \mathbb{R} \end{cases}. \tag{2.3.13}$$

Therefore, the function σ_λ converges pointwise to a function $\gamma(x) : I_d \to \mathbb{R}$ as $\lambda \to +\infty$.

$$\gamma(x) = \begin{cases} 1, & \text{for } b_0{}^\intercal x + c > 0 \\ 0, & \text{for } b_0{}^\intercal x + c < 0 \\ \sigma(\varphi), & \text{for } b_0{}^\intercal x + c = 0 \end{cases} \tag{2.3.14}$$

Let $\Pi_{b_0,c}$ denote the hyperplane and $H_{b_0,c}$ denote the half-space as:

$$\Pi_{b_0,c} = \{x \in \mathbb{R}^d \mid b_0{}^\intercal x + c = 0\} \tag{2.3.15}$$

$$H_{b_0,c} = \{x \in \mathbb{R}^d \mid b_0{}^\intercal x + c > 0\} \tag{2.3.16}$$

for all $c \in \mathbb{R}$.

By the Lebesgue Convergence Theorem, we have

$$\begin{aligned} \sigma(\varphi)\mu(\Pi_{b_0,c}) + \mu(H_{b_0,c}) &= \int_{I_d} \gamma(x)\, d\mu(x) \\ &= \int_{I_d} \lim_{\lambda\to\infty} \sigma_\lambda(x)\, d\mu(x) \\ &= \lim_{\lambda\to\infty} \int_{I_d} \sigma_\lambda(x)\, d\mu(x) = 0 \end{aligned}$$

for all $\lambda, \varphi \in \mathbb{R}$.

Thanks to the function σ being sigmoidal, we have $\lim_{\varphi\to+\infty} \sigma(\varphi) = 1$ and $\lim_{\varphi\to-\infty} \sigma(\varphi) = 0$ and consequently it is easy to see

$$\mu(\Pi_{b_0,c}) = 0 \quad \text{and} \quad \mu(H_{b_0,c}) = 0 \quad \forall c \in \mathbb{R}. \tag{2.3.17}$$

This in turn implies $\mu(I_d) = 0$ as c can be chosen arbitrarily large.

We would like to show that the measure of all half-planes being zero implies that the measure μ must be zero. If μ is a positive Borel measure, this would be trivial by (2.3.17) but μ here is a signed measure.

Step 1 *(Construct a signed measure)*: Let ϕ be a finite signed, Borel measure on $\mathbb{R}$

$$\phi(A) = \mu(\{x \in I_d : b_o{}^\intercal x \in A\}) \quad \forall A \subseteq \mathbb{R}. \tag{2.3.18}$$

By construction, we have

$$\begin{aligned}
\forall a < b \in \mathbb{R}, \quad \phi((a,b)) &= \phi((a,\infty)) - \phi([b,\infty)) && (2.3.19)\\
&= \mu(\{x \in I_d : {b_o}^\intercal x > a\}) && (2.3.20)\\
&\quad - \Big(\mu(\{x \in I_d : {b_o}^\intercal x > b\}) + \mu(\{x \in I_d : {b_o}^\intercal x = b\})\Big) && (2.3.21)\\
&= \mu(H_{b_0,-a}) - (\mu(H_{b_0,-b}) + \mu(\Pi_{b_0,-b})) && (2.3.22)\\
&= 0 - 0 = 0. && (2.3.23)
\end{aligned}$$

Therefore $\phi(A) = 0$ for all Borel sets $A \subseteq \mathbb{R}$.

Step 2 *(Define a linear functional L)*: Let $L^\infty(\mathbb{R})$ denote the space of all measurable bounded functions $f : \mathbb{R} \to \mathbb{R}$. For a function $h \in L^\infty(\mathbb{R})$, we define a functional $L : L^\infty(\mathbb{R}) \to \mathbb{R}$:

$$L(h) = \int_{I_d} h({b}^\intercal x)\, d\mu(x). \tag{2.3.24}$$

L is linear because for all $\alpha, \beta \in \mathbb{R},$ and $g, h \in L^\infty(\mathbb{R})$

$$\begin{aligned}
L(\alpha g + \beta h) &= \int_{I_d} (\alpha g + \beta h)({b_0}^\intercal x)\, d\mu(x)\\
&= \int_{I_d} (\alpha g({b_0}^\intercal x) + \beta h({b_0}^\intercal x))\, d\mu(x)\\
&= \alpha \int_{I_d} g({b_0}^\intercal x)\, d\mu(x) + \beta \int_{I_d} h({b_0}^\intercal x)\, d\mu(x)\\
&= \alpha F(g) + \beta L(h)
\end{aligned}$$

Now we look at the indicator function for all Borel sets of $\mathbb{R}$

$$\mathbb{1}_A(x) = \begin{cases} 1, & \text{if } x \in A,\\ 0, & \text{if } x \notin A. \end{cases} \tag{2.3.25}$$

Combining $\mathbb{1}_A \in L^\infty(\mathbb{R})$ and (2.3.18), we have

$$L(\mathbb{1}_A) = \int_{I_d} \mathbb{1}_A({b_0}^\intercal x)\, d\mu(x) = \mu(\{x \in I_d : {b_0}^\intercal x \in A\}) = \phi(A) = 0 \tag{2.3.26}$$

for all Borel sets $A \subseteq \mathbb{R}$.

Since L is a linear functional, the finite sum of functions of the form (simple functions):

$$s_n(x) = \sum_{j=1}^{n} a_j \mathbb{1}_{A_j}(x) \tag{2.3.27}$$

is zero for all $n \in \mathbb{N}$ where $a_j \in \mathbb{R}$, and $A_j \subseteq \mathbb{R}$ are measurable and pairwise disjoint sets of $\mathbb{R}$. Since the simple functions (2.3.27) are dense in $L^\infty(\mathbb{R})$ (Folland, 1999, Proposition 6.7) [1], then for a function $h \in L^\infty(\mathbb{R})$, there s exists a sequence of functions s_n converge

[1] A detailed proof can be found in the Theorem 7.8 in the lecture note of Measure Theory by Prof. Dr. John K. Hunter (2011)

pointwise to h for all $n \in \mathbb{N}$ and $x \in \mathbb{R}$

$$L(h) = \int_{I_d} h(b_0{}^\intercal x)\, d\mu(x) = \int_{I_d} \lim_{n\to\infty} s_n(b_0{}^\intercal x)\, d\mu(x) \tag{2.3.28}$$

$$= \lim_{n\to\infty} \int_{I_d} s_n(b_0{}^\intercal x)\, d\mu(x) = \lim_{n\to\infty} L(s_n) = 0 \tag{2.3.29}$$

where the limit in the integral is moved outside the integral by the Lebesgue Convergence Theorem.

Step 3 *(Tidy up)*: sine and cosine functions are in $L^\infty(\mathbb{R})$, which implies $L(\cos) = L(\sin) = 0$ for all b_0. For a bounded measurable function $h'(x) = e^{ib_0{}^\intercal x} = \cos(b_0{}^\intercal x) + i\sin(b_0{}^\intercal x)$, we have

$$L(h') = \int_{I_d} h'(b_0{}^\intercal x)\, d\mu(x) = \int_{I_d} \cos(b_0{}^\intercal x) + i\sin(b_0{}^\intercal x)\, d\mu(x) = 0. \tag{2.3.30}$$

It is easy to verify that $\mu = 0$ on $\mathbb{R}^d \setminus \{0\}$ as the Fourier transform of μ is zero from Lemma 2.11. Combining with the fact that $\mu(I_d) = 0$, (2.3.30) establishes that σ is discriminatory.

□

2.3.1 Application to the classification problem

This section will explain the implications of Theorem 2.7 for classification problems. It should be noted that the decision function defined below is not continuous on I_d. We would like to check whether such classification problem can also be well understood as the regression problem.

Definition 2.13 (Decision function)**.** *Let $\{P_1, \ldots, P_k\}$ be a partition of I_d where each partition P_j are pairwise disjoint nonempty Borel sets, i.e. $P_j \neq \emptyset$ and $\bigcup_{j=1}^k P_j = I_d$. f is a decision function for I_d of the form*

$$f : I_d \to \{1, \ldots, k\} \tag{2.3.31}$$

where $f(x) = j$ for $x \in P_j$.

Theorem 2.14. Let σ be a continuous sigmoidal function and f be the decision function for any finite, measurable partition of I_d. Let ϕ be a Borel measure on I_d and $\mu(I_d) = 1$. Then for any $\epsilon > 0$, there exists a finite sum of the form

$$h(x) = \sum_{j=1}^{n} a_j \sigma(b_j{}^\intercal x + c_j) \tag{2.3.32}$$

where $a_j, c_j \in \mathbb{R}, b_j \in \mathbb{R}^d$ and a set $D \subset I_d$, such that the measure of the set D, $\mu(D) \geq 1-\epsilon$ and

$$\sup_{x\in D} |f(x) - h(x)| < \epsilon \tag{2.3.33}$$

Theorem 2.14 is a anoloy of the UAT and the proof is straightforward using Lusin's Theorem (A.12).

Proof. By Lusin's Theorem (A.12), there exists a continuous function $g \in C(I_d)$ such that $\mu(\{x \in I_d \mid f(x) \neq g(x)\}) < \epsilon$. Now we have a continuous g and we are able to a find a sum of the form above satisfying $|h(x) - g(x)| < \epsilon$ by Theorem (2.7) for all $x \in I_d$. Let set $D = \{x \in I_d \mid f(x) = g(x)\}$ and we have $\mu(D) \geq 1 - \epsilon$. Then for $x \in D$, we have

$$\sup_{x \in D} |h(x) - f(x)| = \sup_{x \in D} |h(x) - g(x)| < \epsilon \quad \forall x \in D. \tag{2.3.34}$$

□

The above result shows that the total measure of the misclassified points can be made arbitrarily small when n is allowed to grow.

Chapter 3

Fourier-analytic Barron spaces

In this chapter, we study the model classes of functions that are *well approximated* by 2NN. Barron (1993) has demonstrated that under some mild assumptions for the target functions, a dimension-independent approximation rate can be achieved by 2NN. Specifically, this assumption is formulated as a spectral condition on the Fourier transform of the target function. When approximating functions with finite Fourier moment by 2NN with n nodes, an approximation rate of $\mathcal{O}(n^{-1/2})$ is obtained in L^2. In section 3.4, we connect this result with n-term dictionary approximation, where we identify the dictionary of choice for the spectral condition.

The chapter is structured as follows. In section 3.1, we introduce the class of functions that satisfies certain smoothness constraints. Section 3.2 characterizes the function spaces constructed based on the previous smoothness restriction. In section 3.3, we prove an approximation rate of $\mathcal{O}(n^{-\frac{1}{2}})$ with respect to the supremum norm, where n is the number of nodes. Section 3.4 presents the dictionary corresponding to the smoothness condition. Finally, we provide some high order error rates for $\mathrm{ReLU}^{\mathrm{k}}$ networks.

3.1 Barron class and spectral condition

In this section, we would identify the class of functions originally proposed by Barron (1993). We define the *spectral condition* on which the smoothness constrain of a function is imposed. A modified spectral condition proposed by Siegel and Xu (2022a) is also introduced.

Definition 3.1 (Barron class)**.** *Let U be a nonempty bounded domain in $\mathbb{R}^d$. A function $f : U \to \mathbb{R}$ is said to be in Barron class with a constant $C > 0$, if there is a x_0 in U, $c \in [-C, C]$, and a measurable function $f : \mathbb{R}^d \to \mathbb{C}$ satisfying:*

$$\int_{\mathbb{R}^d} |\omega|_{U,x_0} \cdot |\mathcal{F}(f)(\omega)| \, d\omega < C \tag{3.1.1}$$

$$f(x) = c + \int_{\mathbb{R}^d} (e^{i\omega^\intercal x} - e^{i\omega^\intercal x_0}) \cdot \mathcal{F}(f)(\omega) \, d\omega \tag{3.1.2}$$

where $|\omega|_{U,x_0} := \sup_{x \in U} \|\langle \omega \, , x - x_0 \rangle\|$ and it is denoted by $|\omega|_U$ when $x_0 = 0$ for simplicity. We refer the class of all functions satisfying the above condition as $\Gamma_C(U, x_0)$[1]*.*

[1] $\Gamma_C(U, x_0)$ is the Fourier-analytic Barron space $\mathcal{B}_{\mathcal{F},1}(U)$ discussed before. The integral condition in (3.1.1) is the integral condition $v_{f,1} < \infty$ defined in (3.1.6).

Remark. In the above definition, the domain U bounded using $|\omega|_U$ is interpreted as bounding the trigonometric component $e^{i\omega^\intercal x}$ where ω is the frequency in the Fourier representation of the functions restricted. It it easier to see that the constant C

$$C \leq r \int_{\mathbb{R}^d} |\omega| \, |\mathcal{F}(f)(\omega)| \, d\omega. \tag{3.1.3}$$

if U is in a ball of radius r since $|\omega|_U \leq r \cdot |\omega|$ by the Cauchy-Schwarz inequality.

More generally, if U is a l_p ball of radius r, we can rewrite the condition as

$$r \int_{\mathbb{R}^d} \|\omega\|_p \, |\mathcal{F}(f)(\omega)| \, d\omega < \infty, \quad 1 \leq p \leq \infty. \tag{3.1.4}$$

Definition 3.2 (Spectral condition)**.** *Let U be a nonempty bounded domain in $\mathbb{R}^d$. Suppose that a function $f : U \to \mathbb{R}$ admits a Fourier representation*

$$f(x) = \int_{\mathbb{R}^d} e^{i\omega^\intercal x} \mathcal{F}(f)(\omega) \, d\omega \tag{3.1.5}$$

where $\mathcal{F}(f) : \mathbb{R}^d \to \mathbb{C}$ is the Fourier transform of f.

For any $s \in \mathbb{N}$, the spectral condition of f is defined as

$$v_{f,s} = \int_{\mathbb{R}^d} |\omega|^s \, |\mathcal{F}(f)(\omega)| \, d\omega. \tag{3.1.6}$$

Definition 3.3 (Modified spectral condition)**.** *Under the setup of Definition 3.2, we define a modified spectral condition*

$$v'_{f,s} = \int_{\mathbb{R}^d} (1 + |\omega|)^s \, |\mathcal{F}(f)(\omega)| \, d\omega. \tag{3.1.7}$$

3.2 Construction of Fourier-analytic Barron spaces

Firstly, we define the spectral condition and the seminorm and norm[2] by which the smoothness of a function is controlled.

Definition 3.4 (Spectral seminorm)**.** *Let U be a nonempty bounded domain in $\mathbb{R}^d$. Suppose that a function $f : U \to \mathbb{R}$ admits a Fourier representation*

$$f(x) = \int_{\mathbb{R}^d} e^{i\omega^\intercal x} \mathcal{F}(f)(\omega) \, d\omega \tag{3.2.1}$$

where $\mathcal{F}(f) : \mathbb{R}^d \to \mathbb{C}$ is the Fourier transform of f.

For any $s \in \mathbb{N}$, the spectral seminorm of f is defined as

$$|f|_{\mathcal{F},s} = \inf_{f_e|_U = f} v_{f,s} \tag{3.2.2}$$

where the infimum is taken over all extensions f_e of f in $L^1(U)$.

[2] Often papers include "Barron" as readers can deduce from the context. In the following chapters, we would like to call them *spectral semi/norm* as another definition of norm in infinite-width Barron spaces is named *Barron norm*.

The notion of spectral norm was first introduced in Siegel and Xu (2021a) since it is more convenient compared to the seminorm.

Definition 3.5 (Spectral norm)**.** *Under the setup of Definition 3.4, we define a spectra norm*

$$\|f\|_{\mathcal{F},s} = \inf_{f_e|_U=f} v'_{f,s}. \tag{3.2.3}$$

Definition 3.6 (Fourier-analytic Barron spaces)**.** *Let U be a nonempty bounded domain in $\mathbb{R}^d$. The Fourier-analytic Barron spaces are*

$$\mathcal{B}_{\mathcal{F},s}(U) := \left\{ f : U \to \mathbb{R} : v'_{f,s} < \infty \text{ and } \forall x \in U, f(x) = \int_{\mathbb{R}^d} e^{i\omega^\intercal x} \mathcal{F}(f)(\omega)\, d\omega \right\} \tag{3.2.4}$$

equipped with a norm $\|f\|_{\mathcal{F},s}$ for all $s \in \mathbb{N}$.

The smoothness index $s \in \mathbb{N}$ refers to the degree of smoothness of functions in $\mathcal{B}_{\mathcal{F},s}$. It is reasonable to expect that as s increases, the functions in $\mathcal{B}_{\mathcal{F},s}$ become smoother. Consequently, the size of the function spaces of higher smoothness index is expected to shrink, which may suggest better approximation error rates. In the following sections, we will provide a precise statement of the improved approximation error rates, along with a detailed explanation of the complexities inherent within and between these spaces.

In general, the activation function associated with the infinite-width Barron spaces is ReLU and we will explicitly state when other functions (e.g. squared ReLU, ReLU^k, Heaviside) are used.

3.3 Approximation rate in Fourier-analytic Barron spaces

Barron (1993) has demonstrated that functions of d-variables with finite Fourier moments can be approximated using the superpositions of sigmoidal functions at a rate that is independent of the dimensionality d. This result has also been extended to the ReLU activation function. In other words, any function in this class can be approximated using a 2NN with an error rate of $\mathcal{O}(n^{-1} \cdot C)$, where n is the number of nodes in the single hidden layer, and C is a constant that depends only on the smoothness of the target function. Although the convergence rate is independent of the dimensionality d of the input vector $x \in \mathbb{R}^d$, the constant C could be dimension-dependent, as the Fourier transform is used in this approach.

This section is based on Barron (1993).

Theorem 3.7. Let U be a nonempty bounded domain in $\mathbb{R}^d$, $x_0 \in U$, and $C > 0$ a constant. For every function in the Barron class $\Gamma_C(U, x_0)$, every sigmoidal function σ, and every $n \in \mathbb{N}$, there exists a linear combination of sigmoidal function $f_n(x) = \sum_{j=1}^n a_j \sigma(b_j{}^\intercal x + c_j), a_j, c_j \in \mathbb{R}, b_j \in \mathbb{R}^d$ such that

$$\|f - f_n\|_{L^2(U)} \leq n^{-\frac{1}{2}} \cdot 2C \tag{3.3.1}$$

Proof. The main idea behind the the proof is to show functions with finite Fourier moment are in the closure of the convex hull of the set of half planes.

Step 0 (*Fix x_0 to 0*): Let x_0, x_1 be two arbitrarily selected points in U, and $f \in \Gamma_C(U, x_0)$. For any $\omega \in \mathbb{R}^d$, given x_0, x_1, we have

$$|\omega|_{U,x_0} = \sup_{x \in U} \|\langle \omega\,, x - x_0 \rangle\| \leq \sup_{x \in U} \|\langle \omega\,, x - x_1 \rangle\| + \|\langle \omega\,, x_0 - x_1 \rangle\| \leq 2\,|\omega|_{U,x_1} \tag{3.3.2}$$

Therefore, we have $\int_{\mathbb{R}^d} |\omega|_{U,x_1} |\mathcal{F}(f)(\omega)| \, d\omega \leq 2C$. If we have $\tilde{c} = c + \int_{\mathbb{R}^d} (e^{\omega^\intercal x_0} - e^{\omega^\intercal x_1}) d\omega$, then $f(x) = \tilde{c} + \int_{\mathbb{R}^d} (e^{i\omega^\intercal x} - e^{i\omega^\intercal x_1})$ with $\tilde{c} \leq 2C$.

This shows that changing x_0 would only affect the constant in the RHS of Theorem 3.7 by a factor of at most two, i.e. $\Gamma_C(U, x_0) \subset \Gamma_{2C}(U, x_1)$. Therefore, we continue the proof assuming $x_0 = 0$.

Step 1 (*Represent f via Inverse Fourier Transform*): With the polar decomposition, we have $\mathcal{F}(f)(\omega) = e^{i\theta(\omega)} \cdot |\mathcal{F}(f)(\omega)|$ where $\theta(\omega) \in \mathbb{R}$ denote the magnitude decomposition. From the assumption, and the fact that f is real-valued ($f : U \to \mathbb{R}$), the real-valued part of $f(x) - f(0)$ can be written as:

$$f(x) - f(0) = \Re \int (e^{i\omega^\intercal x} - e^{i\omega^\intercal 0}) e^{i\theta(\omega)} \cdot |\mathcal{F}(f)(\omega)| \, d\omega \tag{3.3.3}$$

$$= \int_\Omega \Big(\cos(\omega^\intercal x + \theta(\omega)) - \cos(\theta(\omega)) \Big) |\mathcal{F}(f)(\omega)| \, d\omega \tag{3.3.4}$$

$$= \int_\Omega \frac{C_{f,U}}{|\omega|_{U,0}} \Big(\cos(\omega^\intercal x + \theta(\omega)) - \cos(\theta(\omega)) \Big) \, d\mu_g \tag{3.3.5}$$

$$= \int_\Omega g(x, \omega) \, d\mu_g. \tag{3.3.6}$$

where we denote $\int_{\mathbb{R}^d} |\omega|_U \cdot |\mathcal{F}(f)(\omega)| \, d\omega \leq C$ by $C_{f,U}$.

μ_g is a probability distribution $d\mu_g = |\omega|_U / C_{f,U} \, |\mathcal{F}(f)(\omega)| \, d\omega$, the integral is evaluated on $\Omega = \{\omega \in \mathbb{R}^d : \omega \neq 0\}$ and

$$g(x, \omega) = \frac{C_{f,U}}{|\omega|_U} \Big(\cos(\omega^\intercal x + \theta(\omega)) - \cos(\theta(\omega)) \Big). \tag{3.3.7}$$

Step 2 ($f(x) - f(0)$ *is in the closure of the convex hull of* $G_{\cos}$): The integral form in (3.3.3) shows that $f(x) - f(0)$ can be represented as an infinite convex combination of functions in the class

$$G_{\cos} = \left\{ \frac{|\gamma|}{|\omega|_U} \Big(\cos(\omega^\intercal x + b) - \cos(b) \Big) : \omega \neq 0, |\gamma| \leq C, b \in \mathbb{R} \right\} \tag{3.3.8}$$

Suppose we have drawn n samples ($\{\omega_i, i = 1, \ldots, n\}$) from μ_g, the expected norm in $L^2(U, \mu_g)$ converges to zero as $n \to \infty$ by L^2 law of large numbers. Therefore, there exist a convex combination of elements in $G_{\cos}$ that converges to $f(x) - f(0)$ in L^2.

Step 3 ($G_{\cos}$ *is in the closure of the convex hull of* G_{step}): It is sufficient to check $g(z), z = \alpha x, \alpha = \omega / |\omega|_U$ on $[-1, 1]$ for some $\omega \neq 0$. As $g(z)$ is a uniformly continuous sinusoidal function on $[-1, 1]$, it can be uniformly approximately by piecewise constant step function.

Restricting $g(z)$ on $[0, 1]$, for a partition $0 \leq p_1 \leq p_2 \leq \cdots \leq p_k = 1$, define

$$g_{k,+}(z) = \sum_{i=1}^{k-1} \Big(g(p_i) - g(p_{i-1}) \Big) \cdot \mathbb{1}_{\{z \geq p_i\}}(z) \tag{3.3.9}$$

Similarly, we can construct $g_{k,-}(z) = \sum_{i=1}^{k-1} (g(-p_i) - g(-p_{i-1})) \cdot \mathbb{1}_{\{z \leq -p_i\}}(z)$, resulting in a sequence of piecewise step function on $[-1, 1]$ uniformly close to $g(z)$. We have

$g(z) = g_{k,+} + g_{k,-}$, a linear combination of step function (or heaviside function) and the sum of the coefficients is bounded by 2C (The sum of coefficients of $g_{k,+}$ is bounded by C as a result of the derivative of g bounded by C, so does $g_{k,-}$ and hence $2C$).

We can see that functions $g(z)$ are in the closure of the convex hull of the step functions (by Lemma 1 in Barron (1993))

By substituting $z = \frac{\omega}{|\omega|_U} x$, we have $G_{\cos} \subset G_{\text{step}}$,

$$G_{\text{step}} = \left\{ \gamma \mathbb{1}_{\{\alpha x - t\}}(x) : |\gamma| \leq 2C, |t| \leq 1, |\alpha|_U = 1 \right\}. \tag{3.3.10}$$

Step 4 (*Closure of G_ϕ*): There exists a sequence of sigmoidal functions $\phi(|c|(\alpha x - t))$, as $|c| \to \infty$, they converge to step functions pointwise (except at points where $\alpha x - t = 0$). If we introduce a measure μ that has zero measure at those points, previous statement on $G_{\cos} \subset G^{\mu}_{\text{step}}$ still holds on $\{|t| \leq 1 : \alpha x - t \neq 0\}$ given a particular α. We subsequently have convergence in $L_2(U, \mu)$ by the Dominated Convergence Theorem, which implies that $G^{\mu}_{\text{step}} \subset G_\phi$.

Finally, we arrive at the following relationship since the closure of a convex set is also convex (A.3)

$$\Gamma_{U,x_0} \subset \overline{G_{\cos}} \subset \overline{G_{\text{step}}} \subset \overline{G_\phi}.$$

It has been shown above that function $f(x) - f(0)$ is in the closure of the convex hull of G_ϕ where $\|g\| \leq (2C)^2$ for every $g \in G_\phi$. Hence the L_2 norm of the approximation error is bounded for any choice of $C' > (2C)^2 - \|f(x) - f(0)\|^2$ by Corollary 2.4.

□

3.4 Connection with variation space

As a expansion of nonlinear approximation of dictionary, the variation space is defined in terms of a convex hull based on integral representations (Parhi and Nowak, 2021, 2022). Suppose that the dictionary is uniformly bounded in a separable Hilbert space $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$.

$$\sup_{d \in \mathbb{D}} \|d\|_{\mathcal{H}} < \infty. \tag{3.4.1}$$

If G is a subset of $\mathcal{H}$ and $c \in \mathbb{R}$, then we define the set

$$cG = \{cg : g \in G\}. \tag{3.4.2}$$

Definition 3.8 (Varitaion norm)**.** *The variation norm of $\|f\|_{\mathcal{K}(\mathbb{D})}$ of a subset $\mathbb{D}$ of a linear space X is defined for all $f \in X$ as*

$$\|f\|_{\mathcal{K}(\mathbb{D})} := \inf\{c > 0 : f/c \in \overline{\text{conv}(\mathbb{D} \bigcup -\mathbb{D})}\}. \tag{3.4.3}$$

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

$$\overline{\text{conv}(\mathbb{D} \bigcup -\mathbb{D})} := \left\{ \sum_{j=1}^{n} a_j d_j : n \in \mathbb{N}, d_j \in \mathbb{D}, \sum_{j=1}^{n} |a_j| \leq 1 \right\}. \tag{3.4.4}$$

Definition 3.9 (Varitaion space)**.** *The variation space $\mathcal{K}(\mathbb{D})$ is given by*

$$\mathcal{K}(\mathbb{D}) := \{f \in \mathcal{H} : \|f\|_{\mathcal{K}(\mathbb{D})} < \infty\}. \tag{3.4.5}$$

By the definitions, the following elementary properties of the variation space hold.

Proposition 3.10. Let $\mathbb{D}$ be a uniformly bounded subset of a Hilbert space $\mathcal{H}$

1. $\overline{\text{conv}(\mathbb{D} \bigcup -\mathbb{D})} = \{f \in \mathcal{H} : \|f\|_{\mathcal{K}(\mathbb{D})} \leq 1\}$
2. $\|f\|_{\mathcal{H}} \leq \|f\|_{\mathcal{K}(\mathbb{D})} \cdot \sup_{d\in\mathbb{D}} \|d\|_{\mathcal{H}}$
3. $\mathcal{K}(\mathbb{D})$ is a Banach space equipped with norm $\|f\|_{\mathcal{K}(\mathbb{D})}$

Proof. (1) and (2) is clear from the previous definitions. In order for $\mathcal{K}(\mathbb{D})$ to be a Banach space, we need to show $\mathcal{K}(\mathbb{D})$ is complete with norm $\|\cdot\|_{\mathcal{K}(\mathbb{D})}$.

Let $\{f_n\}$ be a Cauchy sequence w.r.t. $\|\cdot\|_{\mathcal{K}(\mathbb{D})}$. By (2), this automatically implies that $f_n \to f$ in $\mathcal{H}$ so the sequence is Cauchy w.r.t. $\|\cdot\|_{\mathcal{H}}$. □

Proposition 3.11. Let $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ be a separable Hilbert space and f be a function in $\mathcal{H}$. Suppose that a sequence $\{f_n\}$ in $\mathcal{H}$ where $f_n \in \Sigma_n(\mathbb{D}, M)$

$$\Sigma_n(\mathbb{D}, M) := \left\{ f = \sum_{j=1}^{n} \alpha_j d_j : d_j \in \mathbb{D} \text{ and } \sum_{j=1}^{n} |\alpha_j| \leq M, \quad n \in \mathbb{N}, \alpha_j \in \mathbb{R} \right\}. \tag{3.4.6}$$

$f_n \to f$ in $\mathcal{H}$ for some fixed $M < \infty$. Then f is in the variation space $\mathcal{K}(\mathbb{D})$ and its variation norm is bounded by M

$$f \in \mathcal{K}(\mathbb{D}), \quad \|f\|_{\mathcal{K}(\mathbb{D})} \leq M. \tag{3.4.7}$$

Proof. Without loss of generality, we only need to prove for $M = 1$.

From the definition, $\Sigma_n(\mathbb{D}, 1) \subset \overline{\text{conv}(\mathbb{D} \bigcup -\mathbb{D})}$ for every $n \in \mathbb{N}$. Therefore, $f_n \in \overline{\text{conv}(\mathbb{D} \bigcup -\mathbb{D})}$. Noted that the set is closed, we immediately have $f \in \overline{\text{conv}(\mathbb{D} \bigcup -\mathbb{D})}$ and hence the $\|f\|_{\mathcal{K}(\mathbb{D})} \leq 1$. □

Here we would like to connect $\mathbb{D}$ with the integral representation. An integral representation of a function f using dictionary $\mathbb{D}$ is

$$f = \int_{\mathbb{D}} i_{\mathbb{D}\to\mathcal{H}} \, d\mu. \tag{3.4.8}$$

Here, μ is a signed Borel measure on $\mathbb{D}$ with finite variation on $\mathbb{D}$:

$$\|\mu\|_{\mathbb{D}} := \sup_{\substack{g:\, \mathbb{D}\to[-1,1] \\ g \text{ is measurable}}} \int_{\mathbb{D}} g \, d\mu < \infty. \tag{3.4.9}$$

Remark. The existence of μ by is detailed in (Siegel and Xu, 2022a, Proposition 2) and (Diestel, 1984, Chapter 4)

Definition 3.12 (Inclusion map)**.** *Let $A \subseteq B$. The injection $i_{A\to B} : A \to B$ is an inclusion map if*

$$i_{A\to B}(a) = a, \quad \forall a \in A. \tag{3.4.10}$$

Lemma 3.13. Let $\mathbb{D}$ be a compact subset on $\mathcal{H}$. Then $f \in \mathcal{K}(\mathbb{D})$ if and only if there exists a Borel measure μ on $\mathbb{D}$ such that

$$f = \int_{\mathbb{D}} i_{\mathbb{D}\to\mathcal{H}}\, d\mu \tag{3.4.11}$$

where $i_{\mathbb{D}\to\mathcal{H}}$ is an inclusion map from $\mathbb{D}$ to $\mathcal{H}$. Furthermore, the variation norm of f is given by the infimum of measure's variation over all the signed Borel measure on $\mathbb{D}$:

$$\|f\|_{\mathcal{K}(\mathbb{D})} = \inf_{\mu}\Big\{ \|\mu\|_{\mathbb{D}} : f = \int_{\mathbb{D}} i_{\mathbb{D}\to\mathcal{H}}\, d\mu\Big\}. \tag{3.4.12}$$

Definition 3.14 (Spectral dictionary)**.** *The spectral dictionary of order $s \in \mathbb{N}^0$ is given by*

$$\mathbb{F}_s := \left\{ (1+|\omega|)^{-s} \cdot e^{2\pi i \omega^\intercal x} : \omega \in \mathbb{R}^d \right\} \tag{3.4.13}$$

Theorem 3.15 (Equal norm)**.** The Fourier-analytic Barron space is equivalent to the variation space constructed with $\mathbb{F}_s$

$$\mathcal{B}_{\mathcal{F},s} = \mathcal{K}(\mathbb{F}_s). \tag{3.4.14}$$

Furthermore, the spectral norm is equivalent to the variation norm:

$$\|f\|_{\mathcal{K}(\mathbb{F}_s)} = \inf_{f_{e|U}=f} \int_{\mathbb{R}^d} (1+|\omega|)^s \, |\mathcal{F}(f)|\, d\omega. \tag{3.4.15}$$

Proof. One can find a proof in the Section 5 of Siegel and Xu (2022a). □

3.5 Improved rate via Heaviside function

In the following sections, we will discuss how the approximation error rates can be improved by imposing stricter conditions on the smoothness of the dictionary set or by ensuring its compactness. The key idea is to derive the n-th dyadic entropy number corresponding to the dictionary $\mathbb{D}$. Maurey's Theorem 2.3 provides a bound for any compact dictionary in a Hilbert space.

The first instance of improvement in this area was reported in Makovoz (1996), where the compactness of the dictionary of Heaviside functions was used and showed that the error can be improved to $\mathcal{O}(n^{\frac{1}{2}-\frac{1}{p\cdot d}})$ in $L^p(\Omega)$, $p < \infty$. More recently, Klusowski and Barron (2018a) demonstrated that ReLU and ReLU^2 networks can also be made compact by controlling the L^1 and L^∞ norms of their parameters. In the latter case, the inner parameters $b \in \mathbb{R}^d$ in (3.6.1) are suitably constrained to ensure compactness.

Our discussion will primarily be based on the works of Ma, Siegel, and Xu (2022); Siegel and Xu (2022b); Klusowski and Barron (2018a), unless we explicitly cite other sources.

Theorem 3.16. Let U be bounded set on $\mathbb{R}^d$. Let $f : U \to \mathbb{R}$ be a function and V be the closure of all functions $f : U \to \mathbb{R}$ of the following form

$$V = \{f(x) = \sum_{j=1}^{n} a_j H(b_j{}^\intercal x + c_j) \quad \sum a_j \le 1, |b_j| = 1, c_j \in \mathbb{R}\} \tag{3.5.1}$$

where H is the Heaviside function.

Then there exists a finite linear combination f_n

$$f_n = \sum_{j=1}^{n} a_j \sigma(b_j{}^\intercal x + c_j), b_j \in R^d, a_j, c_j \in \mathbb{R} \tag{3.5.2}$$

where $\sigma(x)$ is a sigmoidal function from $\mathbb{R}$ to $\mathbb{R}$.

For any $f \in V$ and any $n \in \mathbb{N}$ such that

$$\|f - f_n\|_{L^p(U)} \leq C n^{-\frac{1}{2}-\frac{1}{p \cdot d}} \quad 1 \leq p < \infty \tag{3.5.3}$$

where C is a constant dependent only U, p.

Note that this theorem does not cover the case $p = \infty$. However, Barron (1993) has already showed that $f \in V$ are dense w.r.t. supremum norm ($\|f - f_n\|_{L^\infty(U)} = \mathcal{O}(n^{-1/2})$), which implies $\|f - f_n\|_{L^p(U)}$ for all $L^p, 1 \leq p < \infty$.

As the dimensionality d increases, this rate approaches to the original rate and therefore this theorem is insignificant for high dimensionality. It is sufficient to prove the case σ is the Heaviside function. It is easy to check $\sigma(\lambda t) \to H(t)$ as $\lambda \to \infty$ and $\sigma(\lambda t) \to H(t)$ as $\lambda \to -\infty$. On a closed interval $[-u, u]$ on $\mathbb{R}$, note that the difference $H(t) - \sigma(\lambda t)$ is still bounded everywhere. Therefore, we can see that distance between H and σ can be made arbitrarily small on the space $L^d(\mathbb{R})$ with a sufficiently large b.

Proof. We denote the set of sigmoidal functions

$$A = \{\sigma(b, c) : \sigma(b, c) = \sigma(b^\intercal x + c), \quad b \in \mathbb{R}^d, c \in \mathbb{R}\} \tag{3.5.4}$$

As we already knew from Barron (1993) that V is the closure of the convex, symmetric hull of A in $L^p(U)$

$$V = \overline{\text{conv}(A \bigcup -A)}. \tag{3.5.5}$$

Now it remains to finding an estimate for the entropy of A. Without loss of generality, only the case σ with $|b| = 1$ is considered. We can assume U is inside a ball with a suitable radius r since U is bounded subset in $\mathbb{R}^d$. Then this implies $|c| \leq r$ as $\sigma(b^\intercal x + c)$ would be ones or zeros over all U. Suppose b_0, b_1, c_0, c_1 and $|b_0 - b_1| \leq \epsilon$ and $|c_0 - c_1| \leq \epsilon$ for some $\epsilon < 0$. It is easy to check that

$$\sup_{x \in U} \|\sigma(b_0, c_0) - \sigma(b_1, c_1)\|_2 \lesssim \sqrt{\epsilon} \tag{3.5.6}$$

in L^2 with some constant independent of d.

By the volume ratio argument (van der Vaart and Wellner, 1996), one can obtain a $\mathcal{O}(\sqrt{\epsilon})$-net for A in $L^2(U)$ if we are able to find a ϵ-net for the set $P := \{(b, c) \in \mathbb{R}^{d+1} : |b| = 1, |c| \leq r\}$. To build a ϵ-net for the sphere $|b| = 1$, $\mathcal{O}(\epsilon^{1-d})$ elements is needed for the sphere $\{v \in \mathbb{R}^d, |v| = 1\}$. An interval $[-r, r]$ requires $\mathcal{O}(\epsilon^{-d})$ elements. Therefore, a ϵ-net of $\mathcal{O}(\epsilon^{-2d})$ elements can be constructed for A and hence the covering number for A is of the order $\mathcal{O}(\epsilon^{-\frac{1}{2d}})$.

The statement then follows the Corollary on (Makovoz, 1996, p. 104).

□

3.6 Improved rate with higher smoothness index

In section 3.3, an error rate of $\mathcal{O}(n^{-1/2})$ is obtained for functions in $\mathcal{B}_{\mathcal{F},1}$ using n elements from the dictionary (2.2.2)

$$\mathbb{D} = \{\sigma(b^\intercal x + c), \quad b \in \mathbb{R}^d, c \in \mathbb{R}\}. \tag{3.6.1}$$

where σ is a sigmoidal function.

The smoothness of a function f is expressed through its *first* Fourier representation and controlled via the spectral condition $v_{f,s}$ (3.1.6). In particular, how "oscillating" or "fluctuating" of a function f is measured by the mean of the norm of the frequency vector weighted by the Fourier magnitude distribution. Naturally, we would like to extend the findings with tighter restriction on the smoothness and ideally decrease the error rate $\mathcal{O}(n^{-1/2})$. Tighter rates of approximation is made possible with stricter conditions on the Barron spectral norm while bounding the inner parameter $b \in \mathbb{R}^d$ in (3.6.1).

It it shown in section 3.4, the Fourier analytic Barron space $\mathcal{B}_{\mathcal{F},s}$ is equivalent to the variation space $\mathcal{K}(\mathcal{F}_s)$ of the dictionary:

$$\mathbb{F}_s := \{(1 + |\omega|)^{-s} e^{i\omega^\intercal x} : \omega \in \mathbb{R}^d\}. \tag{3.6.2}$$

This implies that we can apply Maurey's argument regarding n-term approximation using dictionaries. Improvements in this direction Siegel and Xu (2022a, 2021b, 2022b); Klusowski and Barron (2018b) is concerned about calculating the metric entropy of the convex hull.

We define the collection of functions which can be expressed as linear combinations of elements from $\mathbb{F}_s$

$$\Sigma_n := \left\{ \sum_{j=1}^n \alpha_j d_j : d_j \in \mathbb{F}_s \text{ and } \sum_{j=1}^n |\alpha_j| \leq M, \quad n \in \mathbb{N}, \alpha_j \in \mathbb{R} \right\}. \tag{3.6.3}$$

Theorem 3.17 (Approximation in L^∞ with bounded coefficients)**.** Let $\Omega = [0,1]^d$ and $s > 0$. If $f \in \mathcal{K}(\mathbb{F}_s)$, then for $n \in \mathbb{N}$, there exists a $f_n \in \Sigma_n$ such that

$$\|f - f_n\|_{L^\infty(\Omega)} \lesssim n^{-\frac{1}{2}-\frac{s}{d}} \sqrt{\log n}\, \|f\|_{\mathcal{B}_{\mathcal{F},s}}. \tag{3.6.4}$$

Proof. One can find the proof in Klusowski and Barron (2018a); Siegel and Xu (2022b). □

3.7 Approximation with ReLU^k activation function

This section focuses on the problem of approximating functions $f \in \mathcal{K}(\mathbb{F}_s)$ uniformly using 2NN with the ReLU^k activation function. In the previous section, we examined the approximation rates for ReLU and Heaviside networks in the space $\mathcal{K}(\mathbb{F}_s)$. In this section, we extend that analysis to include the approximation by ReLU^k networks.

This part builds on the research by Klusowski and Barron (2018a). Siegel and Xu (2021b) obtained the error rate in L^2 and the result is extended to L^∞ in Ma et al. (2022).

We define

$$\Sigma_n^k := \left\{ \sum_{j=1}^n a_j \sigma_k(b_j{}^\intercal x + c_j) : b_j \in \mathbb{R}^d, a_j, c_j \in \mathbb{R} \right\}. \tag{3.7.1}$$

Theorem 3.18. Let $\Omega = [0,1]^d$, $k \geq 1$, and $f \in \mathcal{K}(\mathbb{F}_s)$. The smoothness index $1 < s \leq (d+1)k+1$. Then for a large $n \in \mathbb{N}$, there exists a finite linear combination of elements $f_n \in \Sigma_n^k$ such that

$$\|f - f_n\|_{L^\infty(\Omega)} \lesssim n^{-\frac{1}{2}-\frac{s-1}{d+1}} (\log n)^{1+\frac{(s-1)t}{d+1}} \|f\|_{\mathcal{B}_{\mathcal{F},s}} \tag{3.7.2}$$

where $t = 0$ when $s < (d+1)k+1$ and $t = 1$ when $s = (d+1)k+1$.

An error rate of $\mathcal{O}(n^{-\frac{1}{2}-\frac{2k+1}{2(d+1)}})$ is obtained for L^p norm in Theorem 3 (Siegel and Xu, 2021b).

In the case where the target functions are highly smooth (large s), we obtained the error rate below.

Theorem 3.19. Let $\Omega = [0,1]^d$, $k \geq 1$, and $f \in \mathcal{K}(\mathbb{F}_s)$. The smoothness index $s \geq (d+1)(k+1)$. Then for a large $n \in \mathbb{N}$, there exists a finite linear combination of elements $f_n \in \Sigma_n^k$ such that

$$\|f - f_n\|_{L^\infty(\Omega)} \lesssim n^{-(k+1)} (\log n)^t \|f\|_{\mathcal{B}_{\mathcal{F},s}} \tag{3.7.3}$$

where $t = 0$ when $s < (d+1)(k+1)$ and $t = 1$ when $s = (d+1)(k+1)$.

Remark. The implied constants denoted by $\lesssim$ can be seen to be only depend on s, k, p, d but not on the target function f or the number nodes n. Furthermore, the constant might depend on the dimension d. This dependence could be exponential, i.e. C^d for some C.

Chapter 4

Infinite-width Barron spaces

The objective of this chapter is to establish and characterize the function spaces related to 2NN utilizing various activation functions, including Heaviside function and ReLU function family. As Caragea et al. (2022) state, functions belong to *infinite-width Barron spaces*if they possess an integral representation and can be adequately approximated by 2NN. The approximation error rate is of the order $\mathcal{O}(n^{-1/2})$ when n represents the number of nodes. Unlike the Fourier-analytic Barron spaces discussed in Chapter 3, these spaces rely more heavily on the choice of activation functions and do not cover all possible functions with integral representation.

Section 4.1 defines the norm and integral condition for the functions in infinite-width Barron spaces and proceeds to construct these function spaces. The elementary properties of these spaces are also included for completeness. Section 4.2 examines the approximation error rate for functions in these spaces associated with L^2 and L^∞. In section 4.3, the concept of the variation space and variation norm is connected to infinite-width Barron space through the construction of compact dictionaries. In section 4.5, the connections between infinite-width Barron spaces and Fourier-analytic Barron space are explored.

The goal of this chapter is to provide a concise summary of well-established findings related to the approximation properties of 2NN utilizing the ReLU activation function family. For a comprehensive review, we recommend the works by E, Ma, Wojtowytsch, and Wu (2020); Berner, Grohs, Kutyniok, and Petersen (2022) for a comprehensive review.

4.1 Construction of infinite-width Barron spaces

This section introduces the infinite-width Barron space and its elementary properties, which is mostly based on E et al. (2021) unless stated otherwise.

Let U be a nonempty and bounded domain in $\mathbb{R}^d$. For functions $f : U \to \mathbb{R}$, we consider those that admit the following integral representation:

$$f(x) = \int_\Omega a\sigma(b^\intercal x + c)\mu(da, db, dc), \quad x \in U, a, c \in \mathbb{R}, b \in \mathbb{R}^d. \tag{4.1.1}$$

Let $\Omega = \mathbb{R}^1 \times \mathbb{R}^d \times \mathbb{R}^1$ and Σ_Ω be the σ-algebra on Ω, and $\sigma(\cdot)$ is the ReLU activation function. We define an integral condition for f w.r.t. a a probability distribution μ on

(Ω, Σ_Ω)

$$r(f,\mu,p) = \Big(\int_{\mathbb{R}^d} |a|^p \, (|b|+|c|)^p \, d\mu(a,b,c) \Big)^{1/p} \tag{4.1.2}$$

$$= \Big(\mathbf{E}_\mu \left[|a|^p \, (|b|+|c|)^p \right] \Big)^{1/p}, \quad 1 \le p \le +\infty. \tag{4.1.3}$$

We consider special case where the ReLU function is replaced by the Heaviside function.

Definition 4.1 (Heaviside function)**.**

$$H(x) = \begin{cases} 1 & x > 0, \\ 0 & x \le 0 \end{cases} \tag{4.1.4}$$

Similarly, we consider functions f that admit the representation below

$$f(x) = \int_\Omega a H(b^\intercal x + c) \mu(da, db, dc), \quad x \in U. \tag{4.1.5}$$

Accordingly, we define an integral condition for that particular μ where (4.1.5) holds

$$r(f,\mu,H) = \int_\Omega |a| \; d\mu(a,b,c) = \mathbf{E}_\mu \left[|a| \right] \tag{4.1.6}$$

The integral representation above can be seen as a continuum analogy of the 2NN with n hidden nodes:

$$f_n(x,\Theta) := \frac{1}{n} \sum_{j=1}^{n} a_j \sigma(b^\intercal x + c_j), \quad \Theta = \{(a_j, b_j, c_j), \; j = 1, \dots, n\}. \tag{4.1.7}$$

Definition 4.2 (Barron norm)**.** *For a function f that admits the integral representation in (4.1.1), its Barron norm is defined as*

$$\|f\|_{\mathcal{B}_p} := \inf_\rho \Big(\mathbf{E}_\rho \left[|a|^p \, (|b|+|c|)^p \right] \Big)^{1/p}, \quad 1 \le p \le \infty. \tag{4.1.8}$$

The infimum is taken over all probability distribution where (4.1.1) holds for all $x \in U$. When $p = \infty$, the Barron norm reads

$$\|f\|_{\mathcal{B}_\infty} := \inf_\rho \max_{a,b,c \in \mathrm{supp}(\rho)} |a| \, (|b|+|c|). \tag{4.1.9}$$

Similarly, we can define a norm associated with Heaviside function where the infimum is taken for all measure ρ where (4.1.5) holds

$$\|f\|_{\mathcal{B}_H} = \inf_\rho \mathbf{E}_\rho \left[|a| \right], \quad 1 \le p \le +\infty. \tag{4.1.10}$$

Definition 4.3 (Infinite-width Barron space)**.** *Let U be a nonempty bounded domain in $\mathbb{R}^d$. For functions that admit the integral representation in (4.1.1), the infinite-width Barron space with an order of $1 \le p \le \infty$ is*

$$\mathcal{B}_p(U) = \Big\{ f : U \to \mathbb{R} : \exists \, \mu \in \Sigma_\Omega : r(f,\mu,p) < \infty \text{ and}$$

$$\forall \, x \in U, f(x) = \int_\Omega a \sigma(b^\intercal x + c) \mu(da, db, dc) \Big\}. \tag{4.1.11}$$

A normed space can be defined for those associated with Heaviside function.

Definition 4.4 (Classical Barron space)**.** *Let U be a nonempty unbounded domain in $\mathbb{R}^d$. For functions that admit the integral representation in (4.1.5), the infinite-width Barron space associated with Heaviside function is*

$$\mathcal{B}_H(U) = \Big\{ f : U \to \mathbb{R} : \exists\, \mu \in \Sigma_\Omega : r(f,\mu,H) < \infty \text{ and } \forall\, x \in U, f(x) = \int_\Omega aH(b^\intercal x + c)\mu(da,db,dc) \Big\}. \quad (4.1.12)$$

Barron spaces are denoted by $\mathcal{B}_p$ [1], consist of all the functions whose $r(f,\mu,p)$ is finite for a measure $\mu \in \Sigma_\Omega$.

Proposition 4.5. By the definition of Barron norm, it is easy to see that

$$\mathcal{B}_\infty \subset \cdots \subset \mathcal{B}_2 \subset \mathcal{B}_1. \quad (4.1.13)$$

Proof. The idea is similar to the inclusion of L_p, L_q space.

Applying Hölder's inequality, for any $1 \le p \le q < \infty$

$$\begin{aligned}\int |a|^p (|b|+|c|)^p \, d\rho &= \int |a|^p (|b|+|c|)^p \cdot 1 \, d\rho \\ &\le \Big(\int |a|^{p\cdot q/p} (|b|+|c|)^{p\cdot q/p} \, d\rho \Big)^{p/q} \Big(\int d\rho \Big)^{1-p/q} \\ &= \Big(\int |a|^q (|b|+|c|)^q \, d\rho \Big)^{p/q} \Big(\int d\rho \Big)^{1-p/q}\end{aligned}$$

Therefore we have the inclusion $\mathcal{B}_q \subset \mathcal{B}_p$ for $1 \le p \le q \le \infty$. □

As the reverse also holds in the class of ReLU functions, we have $\mathcal{B}_\infty = \mathcal{B}_p$, $\|\cdot\|_{\mathcal{B}_\infty} = \|\cdot\|_{\mathcal{B}_p}$ for all $1 \le p \le \infty$.

Proposition 4.6. For any $f \in \mathcal{B}_1$, f also $\in \mathcal{B}_\infty$ and $\|f\|_{\mathcal{B}_\infty} = \|f\|_{\mathcal{B}_p}$ and hence $\mathcal{B}_\infty = \cdots = \mathcal{B}_2 = \mathcal{B}_1$ when $\sigma(\cdot)$ is the ReLU function.

Proposition 4.7. $\mathcal{B}$ is a Banach space with norm $\|\cdot\|_{\mathcal{B}}$.

Proof. See Theorem 2.3 in (E and Wojtowytsch, 2020, p. 7). □

[1] Going forward, we will simplify $\mathcal{B}_{\mathcal{F},s}(U)$, $\mathcal{B}_p(U)$ and $\mathcal{B}_H(U)$ as $\mathcal{B}_{\mathcal{F},S}$, $\mathcal{B}_p$ and $\mathcal{B}_H$ to avoid cluttering the notations when U is a bounded domain in $\mathbb{R}^d$.

4.2 Approximation rate in infinite-width Barron spaces

Theorem 4.8 (Approximation in L^2)**.** Let U be a nonempty unbounded domain in $[0,1]^d$ and $f : U \to \mathbb{R}$. For any function $f \in \mathcal{B}(U)$ and any integer $n \in \mathbb{N}$, there exists a 2NN $f_n = f(x,\Theta) = \frac{1}{n}\sum_{j=1}^n a_j\sigma(b_j{}^\intercal x + c_j)$ such that

$$\|f - f_n\|_{L^2(U)} \lesssim n^{-\frac{1}{2}} \|f\|_{\mathcal{B}}. \tag{4.2.1}$$

where Θ is the set of parameters $\Theta = \{(a_j, b_j, c_j), j = 1, \dots, m\}$.

Proof. Let $\epsilon > 0$ and μ be a probability measure on Ω such that

$$f(x) = \mathbf{E}_\mu\left[a\sigma(b^\intercal x + c)\right] \tag{4.2.2}$$

$$\mathbf{E}_\mu\left[a^2(|b| + |c|)^2\right] \le (1+\epsilon)\|f\|_{\mathcal{B}}^2. \tag{4.2.3}$$

The second inequality means that the probability measure μ is chosen such that this integral representation is ***not*** the minimum over all possible probability measures where the integral representation exists.

Let $\hat{f}_n(x)$ be the average drawn from the measure μ, namely

$$\hat{f}_n(x) := \frac{1}{n}\sum_{j=1}^n a_j\sigma(b_j{}^\intercal x + c_j), \quad a_j, b_j, c_j \sim \mu. \tag{4.2.4}$$

Next we would like to evaluate the approximation error between $f(x)$ and $\hat{f}_n(x)$ on $x \in [0,1]^d$.

Let $e(\mu) = \mathbf{E}_{\mathrm{x}}\left[\left|\hat{f}_n(x) - f(x)\right|^2\right]$ and we evaluate the expectation of approximation error

$$\mathbf{E}_\mu\left[e(\mu)\right] = \mathbf{E}_\mu\left[\mathbf{E}_{\mathrm{x}}\left[\left|\hat{f}_n(x) - f(x)\right|^2\right]\right] \tag{4.2.5}$$

$$= \mathbf{E}_{\mathrm{x}}\left[\mathbf{E}_\mu\left[\left|\hat{f}_n(x) - f(x)\right|^2\right]\right] \tag{4.2.6}$$

$$\le \frac{1}{n^2}\sum_{j=1}^n \mathbf{E}_{\mathrm{x}}\left[\mathbf{E}_{(\mathrm{a_j},\mathrm{b_j},\mathrm{c_j})\sim\mu}\left[\left(a_j\sigma(b_j{}^\intercal x + c_j) - f(x)\right)^2\right]\right] \tag{4.2.7}$$

$$\le \frac{1}{n}\mathbf{E}_{\mathrm{x}}\left[\mathbf{E}_{(\mathrm{a,b,c})\sim\mu}\left[\left(a\sigma(b^\intercal x + c)\right)^2\right]\right] \tag{4.2.8}$$

$$= \frac{1}{n}\mathbf{E}_\mu\left[a^2(|b| + |c|)^2\right] \le \frac{(1+\epsilon)\|f\|_{\mathcal{B}}}{n} \tag{4.2.9}$$

Define the event $E = \{e(\mu) \le C\|f\|_{\mathcal{B}}/n\}$, it is easy to check

$$\mathrm{P}[E] = 1 - \mathrm{P}[E^c] \ge 1 - \frac{\mathbf{E}_\mu[e(\mu)]}{C\|f\|_{\mathcal{B}}/n} = 1 - \frac{1+\epsilon}{C} = \frac{C-1-\epsilon}{C} \tag{4.2.10}$$

For any $C > 0$, we can find a corresponding ϵ, and choose a 2NN with parameter Θ satisfies the condition in the theorem.

□

Theorem 4.9 (Approximation in L^∞)**.** Given the conditions outlined in Theorem 4.8, it can be established that for any function $f \in \mathcal{B}(U)$ and any integer $n \in \mathbb{N}$, there exists a 2NN such that

$$\|f - f_n\|_{L^\infty(U)} \lesssim \|f\|_{\mathcal{B}} \sqrt{\frac{d+1}{n}}. \tag{4.2.11}$$

Proof. For simplicity, we only prove the case where $U = I_d = [0,1]^d$. Let μ be a probability measure on Σ_Ω where (4.1.1) holds. Without loss of generality, we can assume $\|f\|_{\mathcal{B}} = |a|$ almost everywhere on μ thanks to the homogeneity of ReLU function.

A bound can found using Rademacher variables thanks to Lemma 26.2 in (Shalev-Shwartz and Ben-David, 2014, p. 376)

Let ξ be Rademacher variables, i.e. $\mathrm{P}[\xi = 1] = \mathrm{P}[\xi = -1] = \frac{1}{2}$, we have

$$\mathbf{E}_\mu \left[\sup_{x \in I_d} \frac{1}{n} \sum_{j=1}^n (a_j \sigma(b_j{}^\intercal x + c_j) - f(x)) \right] \leq$$
$$2\mathbf{E}_\mu \left[\frac{1}{n} \mathbf{E}_\xi \left[\sum_{j=1}^n \sup \xi_j a_j \sigma(b_j{}^\intercal x + c_j) \right] \right]$$

where $(a, b, c) \sim \mu$, and ξ_j are Rademacher variables.

We continue bounding the RHS

$$\mathbf{E}_\xi \left[\sup_{x \in I_d} \frac{1}{n} \sum_{j=1}^n \xi_j a_j \sigma(b_j{}^\intercal x + c_j) \right] \tag{4.2.12}$$
$$= \mathbf{E}_\xi \left[\sup_{x \in I_d} \frac{1}{n} \sum_{j=1}^n \xi_j |a_j| \sigma(b_j{}^\intercal x + c_j) \right] \quad (\text{ symmetry of } \xi) \tag{4.2.13}$$
$$\leq \mathbf{E}_\xi \left[\sup_{x \in I_d} \frac{1}{n} \sum_{j=1}^n \xi_j |a_j| (b_j{}^\intercal x + c_j) \right] \quad (\text{ReLU}) \tag{4.2.14}$$
$$= \mathbf{E}_\xi \left[\sup_{x \in I_d} x'{}^\intercal \frac{1}{n} \sum_{j=1}^n \xi_j |a_j| b'_j \right] \quad (b^\intercal x + c = b'{}^\intercal x') \tag{4.2.15}$$
$$= \mathbf{E}_\xi \left[\left\| \frac{1}{n} \sum_{j=1}^n \xi_j |a_j| b'_j \right\|_1 \right] \tag{4.2.16}$$

Let $u = |a|(|b| + |c|)$,

$$\|u\|_1 \leq \|f\|_{\mathcal{B}} \tag{4.2.17}$$

Then we can rewrite (4.2.12) from the expectation w.r.t. all probability measure μ where

$f(x) = \mathbf{E}_\mu [a\sigma(b^\intercal x + c)]$ holds to the supremum as below:

$$\mathbf{E}_\mu \left[\sup_{x \in I_d} \frac{1}{n} \sum_{j=1}^{n} (a_j \sigma({b_j}^\intercal x + c_j) - f(x)) \right] \tag{4.2.18}$$

$$\leq 2\mathbf{E}_\mu \left[\mathbf{E}_\xi \left[\left\| \frac{1}{n} \sum_{j=1}^{n} \xi_j \|u\|_1 \right\|_1 \right] \right] \tag{4.2.19}$$

$$= 2 \sup_{\|u\|_1 \leq \|f\|_\mathcal{B}} \mathbf{E}_\xi \left[\left\| \frac{1}{n} \sum_{j=1}^{n} \xi_j \|u\|_1 \right\|_1 \right] \tag{4.2.20}$$

$$= 2 \|f\|_\mathcal{B} \sup_{\|u\|_1 \leq 1} \mathbf{E}_\xi \left[\left\| \frac{1}{n} \sum_{j=1}^{n} \xi_j \|u\|_1 \right\|_1 \right] \tag{4.2.21}$$

$$= 2 \|f\|_\mathcal{B} \sqrt{d+1} \sup_{\|u\|_2 \leq 1} \mathbf{E}_\xi \left[\left\| \frac{1}{n} \sum_{j=1}^{n} \xi_j \|u\|_2 \right\|_2 \right] \tag{4.2.22}$$

$$\leq 2 \|f\|_\mathcal{B} \sqrt{\frac{d+1}{n}} \tag{4.2.23}$$

The last inequality follows the results of Rademacher variables in a bounded unit ball in a Hilbert space ($|a|\,|b'| \leq 1$). A proof can be found in Lemma 26.10 in (Shalev-Shwartz and Ben-David, 2014, p. 383) □

4.3 Connection with variation space

In this section, we consider the n-term approximation using ReLU^k functions

$$\sigma_k(x) = [\max(0, x)]^k, \quad k \in \mathbb{N}^0 \tag{4.3.1}$$

when $k = 0$, σ_0 is the Heaviside function in Definition 4.1.5.

Similarly, an approximation upper bound using n elements from the dictionary

$$\mathbb{D} = \{\sigma_k(b^\intercal x + b) : b \in \mathbb{R}^d, c \in \mathbb{R}\} \quad k \in \mathbb{N}^0. \tag{4.3.2}$$

It should be noted that the dictionary with ReLU^0 function is compact. However, the compactness of $\mathbb{D}^k$ when $k > 0$ is not guaranteed so the Maurey's argument does not apply automatically.

Proposition 4.10. When $k > 0$, $\mathbb{D}^k$ is not compact in $L^p(\Omega)$, $1 \leq p < \infty$.

To ensure the compactness of the dictionary, we limit the b, c in the following dictionary for $k > 0$

$$\mathbb{D}_k = \left\{ \sigma_k(b^\intercal x + c), \quad b \in S^{d-1}, c \in [c_1, c_2] \right\} \tag{4.3.3}$$

where σ_k is ReLU^k function described above, $S^{d-1} := \{x \in \mathbb{R}^d : \|x\| = 1\}$ is the unit sphere in $\mathbb{R}^d$. c_1, c_2 is chosen such that

$$c_1 < \inf_{x \in \Omega} b^\intercal x < \sup_{x \in \Omega} b^\intercal x < c_2, \quad b \in S^{d-1}. \tag{4.3.4}$$

The parameters b and c are restricted such that the set $\mathbb{D}_k$ is bounded in $L^p(\Omega)$.

Proposition 4.11. Let U be a nonempty bounded domain in $\mathbb{R}^d$ and $\mathbb{D}_1$ be the dictionary constructed in (4.3.3). The variation space constructed using $\mathbb{D}_1$ is equivalent to the infinite-width Barron space $\mathcal{B}(U)$:

$$\mathcal{B}(U) = \mathcal{K}(\mathbb{D}_1). \tag{4.3.5}$$

Furthermore, the variation norm is equivalent to the Barron norm defined in (4.1.8) up to some constant C which is only dependent on the choice of c_1, c_2

$$\|f\|_{\mathcal{K}(\mathbb{D}_1)} = \|f\|_{\mathcal{B}} \cdot C, \quad C = C(c_1, c_2). \tag{4.3.6}$$

4.4 Improved rate

This section details the improved approximation error rate.

Theorem 4.12. (Makovoz, 1998, Theorem 2, p. 218)

Let $U = [0,1]^d$ be the unit ball in R^d and $f : U \to \mathbb{R}^d$ a function of the form

$$f(x) = \int_{S^d \times [-1,1]} c(b,c) H(b^\intercal x + b)\, d\mu \tag{4.4.1}$$

where c is bounded

$$\sup_{(b,c) \subset S^{d-1} \times [-1,1]} |c(b,c)| \leq 1. \tag{4.4.2}$$

H is the Heaviside function. Then for any $n \in \mathbb{N}$, there exists a linear combination in the form:

$$f_n = \sum_{j=1}^{n} \frac{1}{n} a_j H(b_j{}^\intercal x + c_j), \quad a_j, c_j \in \mathbb{R}, b_j \in \mathbb{R}^d \tag{4.4.3}$$

such that

$$\|f - f_n\|_\infty \lesssim n^{-\frac{1}{2} - \frac{1}{2d}} \sqrt{\log n} \tag{4.4.4}$$

Remark. It should be highlighted that the condition above define a strict subset of the $\overline{\text{conv}(\mathbb{D}_0 \bigcup -\mathbb{D}_0)}$. This stricter condition gives a better approximation error rate.

4.4.1 Improved rate in classical Barron space with less strict conditions

Compared to the condition above, the error rate given below is for all functions f in the closed convex hull of $\mathbb{D}_0$ and the uniform approximation is over all of $\mathbb{R}^d$.

Theorem 4.13. (Ma et al., 2022, Theorem 4, p. 45)

Let $\mathbb{D}_0 = \{\sigma_0(b^\intercal x + c), b \in \mathbb{R}^d, c \in \mathbb{R}\}$. Let f be any function from the closed convex hull of $\mathbb{D}_0$. Then for any $n \in \mathbb{N}$, there exists a linear combination in the form (4.4.3) such that

$$\|f - f_n\|_\infty \lesssim n^{-\frac{1}{2} - \frac{1}{2d}}. \tag{4.4.5}$$

Proof. Let $N = 2^l$ for some integer l. We consider a f that can be represented as $f = \frac{1}{N} \sum_{j=1}^{N} H(b_j{}^\intercal x + c_j)$, which is clearly positive.

Let $X = \{1, 2, \cdots, N\}$ and R_x be the index set of hyperplanes in $\mathbb{R}^d$

$$R_x = \{j : b_j{}^\intercal x + c_j \geq 0\} \tag{4.4.6}$$

and $\mathcal{R} = \{R_x : x \in \mathbb{R}^d\}$.

From Theorem A.16, we can find a coloring $\chi : X \to \{-1, +1\}$ of the set system $(X, \mathcal{R})$

$$\left|\sum_{j=1}^{N} \chi(j) H(b_j{}^\intercal x + c_j)\right| = \left|\sum_{j \in R_x} \chi(j)\right| \lesssim N^{\frac{1}{2}-\frac{1}{2d}}. \tag{4.4.7}$$

Let $R_{+x} = \{i : b_j{}^\intercal x + c_j > 0\}$ and $R_{-x} = \{i : b_j{}^\intercal x + c_j < 0\}$. The measure of $b_j{}^\intercal x + c_j = 0$ is zero for each b_j, c_j. It is easy to check that on $R_{+x} \bigcup R_{-x}$

$$\left|\sum_{j=1}^{N} \chi(j)\right| \leq \left|\sum_{j \in R_{+x}} \chi(j)\right| + \left|\sum_{j \in \mathbb{R}_{-x}} \chi(j)\right| \tag{4.4.8}$$

$$\lesssim 2N^{\frac{1}{2}-\frac{1}{2d}}. \tag{4.4.9}$$

From that we can conclude that we can find a balanced χ such that (4.4.7) holds up to a constant 2. A balanced coloring means that $\sum_{j=1}^{N} \chi(j) = 0$.

Let S be the set $\{i : \chi(i) = -1\}$ fulfilling (4.4.7) and χ is a balanced coloring such that $|S| = N/2$. Consequently, we can find

$$\left|f - \frac{1}{|S|}\sum_{j \in S} H(b_j{}^\intercal x + c_j)\right| = \frac{1}{N}\left|\sum_{j=1}^{N} \chi(j) H(b_j{}^\intercal x + c_j)\right| \tag{4.4.10}$$

$$= \frac{1}{N}\left|\sum_{j \in R_x} \chi(j)\right| \lesssim N^{-\frac{1}{2}-\frac{1}{2d}}. \tag{4.4.11}$$

It it clear that $\frac{1}{|S|}\sum_{j \in S} H(b_j{}^\intercal x + c_j) \in \Sigma_{n.1}(\mathbb{D}_0)$. Setting $|S| \in [\frac{n}{2}, n]$ yields the RHS of the theorem.

The above results holds for general $f \in \overline{\text{conv}(\mathbb{D}_0 \bigcup -\mathbb{D}_0)}$ since one can decompose f into a convex combinations of negative and positive parts easily. □

4.5 Difference and connection between different Barron spaces

This section will clarify the relationships between Barron spaces, namely the *Fourier-analytic Barron spaces* and the *infinite-width Barron spaces*. Although some relationships between these spaces has been examined and understood partially in E et al. (2021, 2020), we hope to clarify this problem in this section inspirerd by the work from Caragea et al. (2022).

Firstly, Let Σ_Ω denote the set of all Borel probability measures on $\Omega = \mathbb{R}^1 \times \mathbb{R}^d \times \mathbb{R}^1$ and we write functions that admits the integral form

$$f(x) = \int_\omega a\sigma(b^\intercal x + c)\, d\mu(a, b, c), \quad \forall x \in \mathbb{R}^d. \tag{4.5.1}$$

Given a nonempty, bounded domain U in $\mathbb{R}^d$, we have already defined various Barron spaces:

- $\mathcal{B}_{\mathcal{F},s}$, $s \in \{1,2\}$ Fourier-analytic Barron spaces (3.6) with $\|\cdot\|_{\mathcal{F},s}$ (3.5)
- $\mathcal{B}$ infinite-width Barron spaces (4.3) with $\|\cdot\|_{\mathcal{B}}$ (4.1.8)
  - $\mathcal{B}_H$ classical Barron space (4.4) with $\|\cdot\|_{\mathcal{B}_H}$ (4.1.10)

We *could* include the classical Barron space within the infinite-width Barron spaces when $\mathcal{B}_p(U), p = 0$ with a ReLU^0 activation function which is essentially the Heaviside function but we decide against it to emphasize on $\mathcal{B}_H(U)$ as it is frequently visited. We denote infinite-width Barron spaces with $\mathcal{B}(U)$ after the equivalence of $\mathcal{B}_p(U), 1 \leq p \leq \infty$ has been shown in Proposition 4.6.

Proposition 4.14. Given the constructions of spaces above and a nonempty bounded domain U in $\mathbb{R}^d$, the following relationships holds:

1) $\mathcal{B}(U) \subset \mathcal{B}_H(U)$

2) $\mathcal{B}_{\mathcal{F},1}(U) \subset \mathcal{B}_H(U)$

3) $\mathcal{B}_{\mathcal{F},2}(U) \subset \mathcal{B}(U)$

Proof. **1):** One can begin with the connection between the ReLU and the Heaviside function. As U is nonempty and bounded in $\mathbb{R}^d$, there is a open ball for some $x \in \mathbb{R}^d$, $B_r(\cdot)$, with a radius $\delta > 0$ whose closure contains U such that

$$U \subset \overline{B_\delta(x)}. \tag{4.5.2}$$

For $x = 0$ and a suitable δ, $U \subset \overline{B_\delta(0)}$ then we have:

$$\sigma(x) = \int_0^{1+\delta} H(x-t)\,dt \quad \forall x \in \mathbb{R} \text{ and } |x| < \delta. \tag{4.5.3}$$

Let $\beta_{b,c}$ be $|b| + |c|$ for any $b \in \mathbb{R}^d, c \in \mathbb{R}$. It is easy to see that $|b^\intercal x + c| \leq (1+\delta)\beta_{b,c}$. Thanks to the positive homogeneity of ReLU function σ, i.e. $\sigma(\lambda x) = \lambda\sigma(x)$ for $x \in \mathbb{R}$, we observe that any function $f : U \to \mathbb{R}$ that admits such an integral representation with a measure $\mu \in \Sigma_\Omega$ can be rewritten as

$$\begin{aligned}
f(x) &= \int_\Omega a\sigma(b^\intercal x + c)\,d\mu(a,b,c) \\
&= \int_\Omega \beta_{b,c}\sigma\Big(\frac{b^\intercal x}{\beta_{b,c}} + \frac{c}{\beta_{b,c}}\Big)\,d\mu(a,b,c) \\
&= \int_\Omega \int_0^{1+\delta} a\beta_{b,c} H\Big(\frac{b^\intercal x}{\beta_{b,c}} + \frac{c}{\beta_{b,c}} - t\Big)\,dt\,d\mu(a,b,c) \quad \text{(Fubini's Theorem)} \\
&= \int_\Omega a' H(b^\intercal x + c')\,d\nu(a',b,c'), \quad \forall x \in U
\end{aligned}$$

where $a', c' \in \mathbb{R}$ for some $v \in \Sigma_\Omega$.

The inclusion is immediate if one can find a measure v and the integral condition $r(f,v,H)$ is also finite.

With a mapping

$$T : \Omega \times [0, 1+\delta] \to \Omega, \quad ((a,b,c),t) \mapsto \Big(a\beta_{b,c}, \frac{b}{\beta_{b,c}}, \frac{c}{\beta_{b,c}} - t\Big) \tag{4.5.4}$$

we can construct the measure v via the pushforward of the product measure $\mu \otimes \lambda$, given λ is the Lebesgue measure on the interval $[0, 1+\delta]$,

$$v := T^{-1}(\mu \otimes \lambda). \tag{4.5.5}$$

Furthermore, we can evaluate the $r(f, v, H)$

$$r(f, v, H) = \int_\Omega |a| \; dv(a, v, c) = \int_\Omega \int_0^{1+\delta} |a\beta_{b,c}| \; dt \, d\mu(a, b, c) \tag{4.5.6}$$

$$= (1+\delta) \, |a| \, (|b| + |c|) \, d\mu(a, b, c) = (1+\delta) r(f, \mu) < \infty. \tag{4.5.7}$$

Therefore, it shows that for any function $f \in \mathcal{B}(U)$

$$\|f\|_{\mathcal{B}_H} \lesssim \|f\|_{\mathcal{B}} < \infty \tag{4.5.8}$$

hence the inclusion holds.

2): This is a direct consequence of (Barron, 1992, Theorem 2).

3): As U is nonempty and bounded in $\mathbb{R}^d$, we can fix a point $x_0 \in \mathbb{R}^d$ and a radius $\delta > 0$ such that $U \subset x_0 + [0, \delta]^d$. Without loss of generality, it is safe to assume that f is a function in $\mathcal{B}_{\mathcal{F},2}$ with the spectral condition $v_{f,2} \leq 1$ (3.4). This implies spetral norm $\|f\|_{\mathcal{B}_{\mathcal{F},2}} \leq 2$ by direct calculation.

One can prove the inclusion if f can be represented as in (4.5.1) with a measure in Σ_Ω.

We define two mapping $G, H : \mathbb{R}^d \to \mathbb{C}$:

$$G(\omega) = \frac{1}{2}(\mathcal{F}(f)(\omega) + \overline{\mathcal{F}(f)(-\omega)})$$

$$H(\omega) = \frac{1}{\delta^d} \cdot e^{\frac{i\omega^\intercal x_0}{\delta}} \cdot G(\omega/\delta)$$

where $\mathcal{F}(f)$ is the Fourier transform of f and $\overline{\mathcal{F}(f)}$ is the complex conjugate of $\mathcal{F}(f)$.

We calculate their respective spectral norm in $\mathcal{B}_{\mathcal{F},2}$:

$$\|G\|_{\mathcal{F},2} \leq 2 \tag{4.5.9}$$

$$\|H\|_{\mathcal{F},2} \leq 2\delta^2 \tag{4.5.10}$$

We then define two functions from U to $\mathbb{R}$ with G, H as their Fourier transform, respectively.

$$g(x) := \int_{\mathbb{R}^d} e^{i\omega^\intercal x} G(\omega) \, d\omega$$

$$h(x) := \int_{\mathbb{R}^d} e^{i\omega^\intercal x} H(\omega) \, d\omega$$

It is easy to check that for all $x \in U$, $f(x) = g(x) = h(\frac{x - x_0}{\delta})$.

By construction, the spectral condition of h, $v_{h,2}$, is finite. We have $\|h\|_{\mathcal{B}}$ is finite with some constant C_h thanks to Theorem 9 in E et al. (2020). Therefore, $h(y)$ can be represented as

$$h(y) = \int_{[0,1]^d} a\sigma(b^\intercal x + c) \, d\mu(a, b, c) \quad \forall y \in [0, 1]^d \tag{4.5.11}$$

where $\mu \in \Sigma_\Omega$ and $\|f\|_{\mathcal{B}} < \infty$ w.r.t. some constant only dependent on δ and d.

Since $y = \frac{x-x_0}{\delta}$ for all $x \in U$, the results above implies that

$$f(x) = h(\frac{x-x_0}{\delta}) \int_\Omega a\sigma(\frac{b^\intercal x}{\delta} + c - \frac{b^\intercal x_0}{\delta})\, dv(a,b,c) \tag{4.5.12}$$

for some measure $v \in \Sigma_\Omega$.

We continue the construction of measure via the pushforward of $v = T(\mu)$. Let T be a mapping:

$$T : \Omega \to \Omega, \quad (a,b,c) \mapsto (a, \frac{b}{\delta}, c - \frac{b^\intercal x_0}{\delta}) \tag{4.5.13}$$

By calculation, $r(f,v) \leq (1+|x_0|)r(h,\mu)$ is finite w.r.t some constant C dependent only on d, δ, x_0. Hence the inclusion is shown.

□

Proposition 4.15. Let $U \subset \mathbb{R}^d$ be bounded and have nonempty interior. For $s > 0$, if $\mathcal{B}_{\mathcal{F},s}(U) \subset \mathcal{B}_1(U)$, then $s \geq 2$. In particular $\mathcal{B}_{\mathcal{F},1}(U) \not\subset \mathcal{B}_1(U)$.

Remark. It has been argued in Theorem 3.1 (E and Wojtowytsch, 2020, p. 12) that $\mathcal{B}_{\mathcal{F},1}$ embeds into the $\mathcal{B}$ but Caragea et al. (2022) proven that this embedding wrongly interpreted the results of Barron (1993, 1992). In other words, the original model class (i.e. $\mathcal{B}_{\mathcal{F},1}$) proposed by Barron (1992) is not *contained* or *embedded* in the novel Barron space $\mathcal{B}$ introduced recently by E et al. (2021).

One can observe that the class of functions in which f admits a Fourier representation with finite *second* moment, $\mathcal{B}_{\mathcal{F},2}$, is well contained inside the *infinite-width Barron space*. However, $\mathcal{B}_{\mathcal{F},1}$ still encapsulates a boarder class of functions. In other words, the infinite-width Barron space $\mathcal{B}$ is *sandwiched* between $\mathcal{B}_{\mathcal{F},1}$ and $\mathcal{B}_{\mathcal{F},2}$.

Chapter 5

Summary and future work

In this thesis, we investigate the approximation properties of two-layer neural networks with various activation functions. Chapter 2 summarizes the results from n-term approximation theory and addresses the question of density regarding 2NN. We state Maurey's theorem for approximation with an n-term dictionary and introduce Jones' iterative approach for obtaining the same approximation error rate.

In Chapter 3, we focus on the smooth functions identified by Barron in the early 1990s. We show that a model class of functions (Fourier-analytic Barron space $\mathcal{B}_{\mathcal{F},s}$) bounded in their first Fourier moment is the closed convex hull of a dictionary of sigmoidal functions. This result implies an error rate of $\mathcal{O}(n^{-1/2})$ when n is the number of nodes, which connects the *complexity* of 2NN models with an upper bound for the approximation error rate. A lot of improvements on the error rate follows, by either placing more restrictions on the activation function (such as the Heaviside function) or requiring the target functions to be smoother by bounding their Fourier moment.

Finally, we consider functions that can be represented in integral form and well approximated by 2NNs. The function space, infinite-width Barron space, is a Banach space. We first construct the infinite-width Barron space via integral representation and show how it connects to the variation space. We describe the relationship between Fourier-analytic Barron spaces and infinite-width Barron spaces, where we find that $\mathcal{B}$ is sandwiched between $\mathcal{B}_{\mathcal{F},1}$ and $\mathcal{B}_{\mathcal{F},2}$.

There remains some interesting problems to investigate in the future:

- First, despite the exponent term of n in the approximation error not containing d, the constant implied in the error rate by $\lesssim$ may be exponential in d.
- The connection between the variation space and the Barron spaces helps unify the techniques and notation for obtaining the error rate, but there has not been much progress other than the representation theorem by Parhi and Nowak (2021).
- A detailed investigation of the model class from the closed convex hull of $\mathrm{ReLU}^{\mathrm{k}}$ functions still remains open in L^p for the $\mathrm{ReLU}^{\mathrm{k}}$ family $k > 1$.

Bibliography

Barron, A. R. (1992). Neural Net Approximation. In *Proc. 7th Yale Workshop on Adaptive and Learning Systems*, Volume 1, pp. 69–72. 38, 39

Barron, A. R. (1993, May). Universal Approximation Bounds for Superpositions of a Sigmoidal Function. *IEEE Transactions on Information Theory 39*(3), 930–945. 10, 19, 21, 23, 26, 39

Barron, A. R. (1994, January). Approximation and Estimation Bounds for Artificial Neural Networks. *Machine Learning 14*(1), 115–133. 2

Beck, C., W. E, and A. Jentzen (2019, August). Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations. *Journal of Nonlinear Science 29*(4), 1563–1619. 1

Bellman, R. (1952, August). On the Theory of Dynamic Programming. *Proceedings of the National Academy of Sciences 38*(8), 716–719. 1

Berner, J., P. Grohs, G. Kutyniok, and P. Petersen (2022). The Modern Mathematics of Deep Learning. In P. Grohs and G. Kutyniok (Eds.), *Mathematical Aspects of Deep Learning*, pp. 1–111. Cambridge: Cambridge University Press. 29

Caragea, A., P. Petersen, and F. Voigtlaender (2022, March). Neural network approximation and estimation of classifiers with classification boundary in a Barron class. 3, 29, 36, 39

Carroll and Dickinson (1989). Construction of Neural Nets Using the Radon Transform. In *International 1989 Joint Conference on Neural Networks*, pp. 607–611 vol.1. 2

Cybenko, G. (1989, December). Approximation by Superpositions of a Sigmoidal Function. *Mathematics of Control, Signals and Systems 2*(4), 303–314. 2, 11

DeVore, R., B. Hanin, and G. Petrova (2021, May). Neural network approximation. *Acta Numerica 30*, 327–444. 5

DeVore, R. A. (1998). Nonlinear approximation. *Acta Numerica 7*, 51–150. 1, 5, 7

Diestel, J. (1984). *Sequences and Series in Banach Spaces*, Volume 92 of *Graduate Texts in Mathematics*. New York, NY: Springer New York. 24

E, W., J. Han, and A. Jentzen (2017, December). Deep learning-based numerical methods for high-dimensional parabolic partial differential equations and backward stochastic differential equations. *Communications in Mathematics and Statistics 5*(4), 349–380. 1

E, W., C. Ma, S. Wojtowytsch, and L. Wu (2020, December). Towards a Mathematical Understanding of Neural Network-Based Machine Learning: What we know and what we don't. *arXiv:2009.10713 [cs, math, stat]*. 29, 36, 38

E, W., C. Ma, and L. Wu (2021, March). The Barron Space and the Flow-induced Function Spaces for Neural Network Models. *arXiv:1906.08039 [cs, math, stat]*. 3, 29, 36, 39

E, W. and S. Wojtowytsch (2020, June). Representation formulas and pointwise properties for Barron functions. *arXiv:2006.05982 [cs, math, stat]*. 3, 31, 39

E, W. and B. Yu (2017, September). The Deep Ritz method: A deep learning-based numerical algorithm for solving variational problems. 1

Folland, G. B. (1999). *Real Analysis : Modern Techniques and Their Applications* (Second edition. ed.). Pure and Applied Mathematics. New York: John Wiley & Sons. 15

Funahashi, K.-I. (1989, January). On the Approximate Realization of Continuous Mappings by Neural Networks. *Neural Networks 2*(3), 183–192. 2

Hunter, J. K. (2011). L^p spaces. https://www.math.ucdavis.edu/~hunter/measure_theory/measure_notes_ch7.pdf. 15

Irie and Miyake (1988, July). Capabilities of Three-layered Perceptrons. In *IEEE 1988 International Conference on Neural Networks*, pp. 641–648 vol.1. 2

Jones, L. K. (1992). A Simple Lemma on Greedy Approximation in Hilbert Space and Convergence Rates for Projection Pursuit Regression and Neural Network Training. *The Annals of Statistics 20*(1), 608–613. 2, 10

Klusowski, J. M. and A. R. Barron (2018a, December). Approximation by Combinations of ReLU and Squared ReLU Ridge Functions With ℓ^1 and ℓ^0 Controls. *IEEE Transactions on Information Theory 64*(12), 7649–7656. 25, 27

Klusowski, J. M. and A. R. Barron (2018b, October). Risk Bounds for High-dimensional Ridge Function Combinations Including Neural Networks. *arXiv:1607.01434 [math, stat]*. 27

Leshno, M., V. Y. Lin, A. Pinkus, and S. Schocken (1993, January). Multilayer Feedforward Networks with Non-Polynomial Activation Function Can Approximate Any Function. *Neural Networks 6*(6), 861–867. 2, 13

Ma, L., J. W. Siegel, and J. Xu (2022, September). Uniform approximation rates and metric entropy of shallow neural networks. *Research in the Mathematical Sciences 9*(3), 46. 25, 27, 35

Makovoz, Y. (1996, April). Random Approximants and Neural Networks. *Journal of Approximation Theory 85*(1), 98–109. 3, 25, 26

Makovoz, Y. (1998, November). Uniform Approximation by Neural Networks. *Journal of Approximation Theory 95*(2), 215–228. 35

Matoušek, J., E. Welzl, and L. Wernisch (1993, December). Discrepancy and approximations for bounded VC-dimension. *Combinatorica 13*(4), 455–466. 49

Parhi, R. and R. D. Nowak (2021, February). Banach Space Representer Theorems for Neural Networks and Ridge Splines. 3, 23, 41

Parhi, R. and R. D. Nowak (2022, June). What Kinds of Functions do Deep Neural Networks Learn? Insights from Variational Spline Theory. *SIAM Journal on Mathematics of Data Science 4*(2), 464–489. 3, 23

Pinkus, A. (1999, January). Approximation theory of the MLP model in neural networks. *Acta Numerica 8*, 143–195. 5

Pisier, G. (1980). Remarques sur un résultat non publié de B. Maurey. *Séminaire d'Analyse fonctionnelle (dit "Maurey-Schwartz")*, 1–12. 9, 10

Rudin, W. (1987). *Real and Complex Analysis* (3rd ed ed.). New York: McGraw-Hill. 12, 47, 48

Rudin, W. (1991). *Functional Analysis* (2nd ed ed.). International Series in Pure and Applied Mathematics. New York: McGraw-Hill. 12, 13, 47

Shalev-Shwartz, S. and S. Ben-David (2014). *Understanding Machine Learning: From Theory to Algorithms*. New York, NY, USA: Cambridge University Press. 1, 5, 33, 34

Siegel, J. W. and J. Xu (2021a, January). Approximation Rates for Neural Networks with General Activation Functions. 21

Siegel, J. W. and J. Xu (2021b, December). High-Order Approximation Rates for Shallow Neural Networks with Cosine and ReLU$\hat{\text{k}}$ Activation Functions. 27, 28

Siegel, J. W. and J. Xu (2022a, April). Characterization of the Variation Spaces Corresponding to Shallow Neural Networks. 3, 19, 24, 25, 27

Siegel, J. W. and J. Xu (2022b, July). Sharp Bounds on the Approximation Rates, Metric Entropy, and n-widths of Shallow Neural Networks. 25, 27

van der Vaart, A. W. and J. A. Wellner (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics. New York, NY: Springer New York. 7, 26

Xu, J. (2020, June). The Finite Neuron Method and Convergence Analysis. *Communications in Computational Physics 28*(5), 1707–1745. 3

Appendix A

Functional Analysis

We will briefly review some of the constructions and theorems of functional analysis and measure theory used throughout this paper. Further background material can be found in Rudin (1991, 1987). Familiarity with the basic notions of topology is assumed (inner product spaces, normed spaces, Banach and Hilbert spaces).

Let X be any vector space, and $K \subset X$ a subset.

Definition A.1 (Convex hull)**.** *The convex hull of E is the set:*

$$\text{conv}(E) = \{x = a_1x_1 + \cdots + a_nx_n \in X \mid x_1, \cdots, x_n \in E \quad t_1 + \cdots t_n = 1, t_i \geq 0\} \quad \text{(A.1)}$$

and an element x in $\text{conv}(E)$ is called a convex combination of $x_1, \cdots, x_n$.

From the definition, it is easy to see that

- $E \subset \text{conv}(E)$, $\text{conv}(E)$ is convex.
- If a set $A \subset X$ is convex and $E \subset A$, then $\text{conv}(E) \subset A$.

Since any intersection of convex sets is still convex, we can get the following equivalent definition of convex hull.

Proposition A.2. The convex hull of a set E is

$$\text{conv}(E) = \bigcap_{E \subset A, A \text{ is convex}} A. \quad \text{(A.2)}$$

Definition A.3 (Closed convex hull)**.** *The closure of the convex hull of E is called closed convex hull of E, denoted by $\overline{\text{conv}(E)}$.*

It is easy to see that $\overline{\text{conv}(E)}$ is a closed convex set, and it is the smallest closed convex set containing E. Another way to view it is that $\overline{\text{conv}(E))}$ is the intersection of all closed convex sets that contains E:

$$\overline{\text{conv}(E)} = \bigcap_{E \subset A, A \text{ is convex and closed}} A \quad \text{(A.3)}$$

We have the following for $K \subseteq E$ in X, $\overline{\text{conv}(K)} \subseteq \overline{\text{conv}(E)}$.

Definition A.4 (Linear opertor)**.** *Let X and Y be normed spaces over $\mathbb{R}$*[1]*. $T : X \to Y$ is a linear operator if T is linear, i.e.*

$$T(x+y) = T(x) + T(y) \quad \forall x \in X, y \in Y$$
$$T(\lambda x) = \lambda T(x) \quad \forall x \in X, \lambda \in \mathbb{N}$$

Definition A.5 (Bounded linear operator)**.** *Let X and Y be normed spaces. A linear operator $T : X \to Y$ is bounded if there exist a $M > 0$ such that for all $x \in X$,*

$$\|T(x)\|_Y \leq M \|x\|_X$$

A linear operator between normed spaces is bounded if and only if it is continuous. We use $B(X, Y)$ to denote the space of Bounded operator between X and Y.

Definition A.6 (Linear functional)**.** *Let X be a normed space over $\mathbb{R}$ and $T : X \to \mathbb{R}$ is a linear operator, then T is a linear functional on X.*

A linear functional T is *bounded* if and only if there exists a $\lambda > 0$ such that $T(x) \leq \lambda \|x\|_X$ for all $x \in X$.

Definition A.7 (Dual space)**.** *The dual space of normed space X is the vector space X^* whose elements are the continuous linear functionals on X. In other words, $X^* = B(X, \mathbb{R})$.*

Theorem A.8 (Consequence II of Hahn-Banach Theorem)**.** Let $(X, \|\cdot\|_X)$ be a subspace of $(Y, \|\cdot\|_Y)$ and $x_0 \in X$. x_0 is in the closure $\overline{X}$ of X if and only if there is no bounded linear functional $T : Y \to \mathbb{R}$ on $(Y, \|\cdot\|_Y)$ such that $T(x) = 0$ for all $x \in X$ while $F(x_0) \neq 0$

Corollary A.9 (Consequence II of Hahn-Banach Theorem)**.** Let X be subspace of a normed linear space Y. Assume $\overline{X}$, the closure of X, is not Y. Then there exists a bounded linear functional L on Y such that $L \not\equiv 0$ and $L(x) = 0$ for all $x \in X$. Suppose every bounded functional L on Y is identically zero on Y. Then X is dense in Y.

Theorem A.10 (Riesz Representation Theorem)**.** (Rudin, 1987, Theorem 6.19, p. 130)

If X is a locally compact Hausdorff space, then every bounded linear functional L on the the space of all continuous functions on X, $C(X)$, is represented by a unique regular complex Borel measure μ

$$L(f) = \int_X f(x)\, d\mu(x), \quad \forall f \in C(X).$$

Moreover, the norm of L is the total variation of μ:

$$\|L\| = |\mu|(X) \tag{A.4}$$

Theorem A.11 (Lesbegue Bounded Convergence Theorem)**.** Let f_n be a sequence of uniformly bounded functions for all $n \in \mathbb{N}$ that satisfy

$$\lim_{n\to\infty} f_n(x) = f(x), \quad \text{pointwise} \tag{A.5}$$

Then,

$$\lim_{n\to\infty} \int f_n\, d\mu = \int \lim_{n\to\infty} f_n\, d\mu = \int f\, d\mu \tag{A.6}$$

[1] It can be extended to $\mathbb{C}$

Theorem A.12 (Lusin's Theorem)**.** Let X be a locally compact Hausdorff space, let μ be a Radon measure on X and f is a measurable function $f : X \to \mathbb{R}$. Suppose that there is a set $A \subseteq X$ with finite measure such that $f(x) = 0$ if $x \notin A$. Then for every $\epsilon > 0$, there exists a compactly supported continuous function $g : X \to \mathbb{R}$ with $\|g\|_{infty} \leq \|f\|_{\infty}$ such that

$$\mu(\{x \in X \mid f(x) \neq g(x)\}) < \epsilon \tag{A.7}$$

Theorem A.13 (Discrepancy of a set)**.** Let $(X, \mathcal{R})$ be a set system and $\mathcal{R} = P(X)$ is the power set of X. Let a mapping $\chi : X \to \{-1, +1\}$ and we name it a *coloring* of X

$$\chi(A) = \sum_{x \in A} \chi(x), A \subseteq X. \tag{A.8}$$

The *discrepancy* of χ on $\mathcal{R}$ is given by

$$\operatorname{disc}(\mathcal{R}, \chi) = \max_{R \in \mathcal{R}} |\chi(R)| . \tag{A.9}$$

The discrepancy of $\mathcal{R}$ is defined as

$$\operatorname{disc}(\mathcal{R}) = \min_{\chi : X \to \{-1,+1\}} \Big\{ \operatorname{disc}(\mathcal{R}, \chi) \Big\} = \min_{\chi : X \to \{-1,+1\}} \max_{R \in \mathcal{R}} |\chi(R)| \tag{A.10}$$

$$= \min_{\chi : X \to \{-1,+1\}} \max_{R \in \mathcal{R}} \left| \sum_{x \in R} \chi(x) \right| . \tag{A.11}$$

Definition A.14 (VC-class)**.** *VC-dimension of the set system $(X, \mathcal{R})$ is defined as the maximum size of a shattered subset of X. A subset $S \subset X$ is called a ϵ-net provided that $S \bigcap R \neq \emptyset$ for every set $R \subset \mathcal{R}$.*

Proposition A.15. Let $X = \{1, \ldots, N\}$ and a subset $R_x \subset X$ is

$$R_x = \{i : a_i{}^{\intercal} x + b_i \geq 0, \quad a_i \in \mathbb{R}^d, b_i \in \mathbb{R}\}. \tag{A.12}$$

Let $\mathcal{R} = \{R_x : x \in \mathbb{R}^d\}$. Then VC-dimension of the set system of $(X, \mathcal{R})$ is at most d.

Theorem A.16. (Matoušek, Welzl, and Wernisch, 1993, Theorem 1.2, p. 595) Let d be the VC-dimension of a set system $(X, \mathcal{R})$. Then for every $N > 1$, there is a coloring of the set X, $\chi : X \to \{-1, 1\}$ such that

$$\max_{R \in \mathcal{R}} \left| \sum_{i \in R} \chi(i) \right| \lesssim N^{1/2 - 1/2d}. \tag{A.13}$$

Declaration of Originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor .

Title of work (in block letters):

Authored by (in block letters):
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
----------	----------
----------	----------
----------	----------
----------	----------
----------	----------

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the Citation etiquette information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work .
- I am aware that the work may be screened electronically for plagiarism.
- I have understood and followed the guidelines in the document *Scientific Works in Mathematics*.

Place, date:	**Signature(s):**
----------	----------

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.