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

Yongqi Wang, Master Thesis

Adviser: Prof. Dr. Sara van de Geer

D-BSSE, Swiss Federal Institute of Technology Zurich

2023 Apr 11

Table of Contents I

1. Introduction
2. Question of density in approximation
3. Fourier-analytic Barron spaces
4. Infinite-width Barron spaces

Why neural networks (NN) excel across domains

- Image and video processing, segmentation
- Time series methods, NLP
- Generative models

Even simplest one are very capable! Two-layer neural network (2NN)

These are not new problems in computational mathematics

- Image classification: approximating function
- Generative models: approximating and sampling distribution with finite samples
- Go game: solving differential and difference equations

The major difference is **dimensionality** d!

d for a RGB image $(512 \times 512) = 3 \times 512 \times 512 = 786,432$

These are not new problems in computational mathematics

Given observed data x, y, often with noise in practical cases.
Find the target function $f_{\text{True}} : x \to y$

- y: labels in classification task
- y: response in most prediction tasks.

Curse of dimensionality

Definition

For a specified accuracy $\epsilon > 0$, the number of parameters to satisfy is growing exponentially.

To reduce the error by a factor of 10, we need to increase m by a factor of 10^d.
Holds for all classical algorithms, e.g. approximating functions using polynomials, trigonometric polynomials or wavelets.

$$f(x) = \sum_{j=1}^{n} a_j \sigma(b_j \cdot x + c_j)$$

where $a_j, c_j \in \mathbb{R}$, $b_j \in \mathbb{R}^d$ and σ is the activation function.

Common activation functions:

- sigmoid: $\sigma(z) = \frac{1}{1+e^{-z}}$
- ReLU, $\sigma(z) = \max\{z, 0\}$
- ...

Two main problems in approximation by NN (2NN)

- **density**: the conditions where f_{target} can be approximated arbitrarily well
- **complexity**: how "large" are necessary to give a prescribed degree of approximation ϵ

Cybenko's Theorem: density

Any continuous functions on $\mathbb{R}^d$ can be approximated uniformly well with 2NN.

Theorem

If σ is sigmoidal as $\sigma(t) = 1$ as $t \to \infty$ and $\sigma(t) = 0$ as $t \to -\infty$, then any continuous functions over $[0,1]^d$ be approximated uniformly well by 2NN.

Necessary and sufficient condition condition for "density"
The activation must not be a polynomials (Leshno, Lin, Pinkus, and Schocken, 1993)

Finding the correct function spaces associated with 2NN

Find the functions that are **well approximated** by 2NN

Fourier-analytic Barron spaces: construction

Let U be a nonempty bounded set on $\mathbb{R}^d$, functions $f : U \to \mathbb{R}$ is said to be in

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

where $v'_{f,s} = \int_{\mathbb{R}^d} (1 + |\omega|)^s \, |\mathcal{F}(f)(\omega)| \; d\omega$.
$f \in \mathcal{B}_{\mathcal{F},1}(U)$: functions with finite Fourier first moment.

Theorem

For any $f \in \mathcal{B}_{\mathcal{F},s}(U)$, there exists a $n > 0$ such that

$$\|f - f_n\|_2 \lesssim n^{-1/2} \tag{1}$$

and the implied constant does depend upon the dimension.

- $\mathcal{B}_{\mathcal{F},1}(U)$ in $L^2(U)$: $\|f - f_n\|_2 \lesssim n^{-1/2}$
- $\mathcal{B}_{\mathcal{F},1}(U)$ in $L^\infty(U)$: $\|f - f_n\|_\infty \lesssim n^{-1/2}$

Infinite-width Barron spaces: construction

Let U be a nonempty bounded set on $\mathbb{R}^d$

$$\mathcal{B}(U) := \Big\{ f : U \to \mathbb{R} : r(f, \mu, p) < \infty \text{ and } \\ \forall x \in U, f(x) = \int_{\Omega} a\sigma(b^\intercal x + c)\mu(da, db, dc) \Big\}$$

where $r(f, \mu, p) = \mathbf{E}_\mu \left[|a| \left(|b| + |c| \right) \right]$

Infinite-width Barron spaces: Approximation error rate

- in $L^2(U)$: $\|f - f_n\|_2 \lesssim n^{-1/2}$
- in $L^\infty(U)$: $\|f - f_n\|_\infty \lesssim n^{-1/2}$

relationship between Fourier-analytic and Infinite-width Barron spaces

- $\mathcal{B}(U)$ depends on the choice of activation function σ
- ReLU 2NN, $\mathcal{B}(U)$ is sandwiched between $\mathcal{B}_{\mathcal{F},1}(U)$ and $\mathcal{B}_{\mathcal{F},2}(U)$

Varitaion space and variation norm

Definition (Varitaion norm)

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

$$\|f\|_{\mathcal{K}(\mathbb{D})} := \inf\{c > 0 : f/c \in \text{closed convex hull of } \mathbb{D}\}$$

Definition (Varitaion space)

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

Connection to variation space

One can find a dictionary $\mathbb{D}$ and the variation space $\mathcal{K}(\mathbb{D})$ for both $\mathcal{B}_{\mathcal{F},s}(U)$ $\mathcal{B}(U)$

- $\mathbb{F}_s := \left\{ (1 + |\omega|)^{-s} \cdot e^{2\pi i \omega^\intercal x} : \omega \in \mathbb{R}^d \right\}$
- $\mathbb{D}_k = \left\{ \sigma_k(b^\intercal x + c), \quad b \in S^{d-1}, c \in [c_1, c_2] \right\}$, S^{d-1} is the unit sphere, c is chosen to ensure $\mathbb{D}_k$'s compactness

Improved rates

Improved rate with higher s

- in $L^\infty(U)$: $\|f - f_n\|_\infty \lesssim n^{-1/2-s/d}\sqrt{\log n}$

If σ is Heaviside function, improved rate

- in $L^\infty(U)$: $\|f - f_n\|_\infty \lesssim n^{-1/2-1/2d}$

Bibliography I

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.