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

2025 Feb 07

Table of Contents I

NN excels in function approximation

- 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 $(512 \times 512) = 3 \times 512 \times 512 = 786,432$

ChatGPT and DeepSeek

"Yes, you can say that both DeepSeek and ChatGPT are results of function approximation in a broad sense, as they are built using machine learning techniques that involve approximating complex functions."

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.

Decomposition of error in ML models

- f_n: best approximation to f_{True} in $\mathcal{H}_n$, n is the width in 2NN case
- $\tilde{f}_{n,M}$: best approximation to f_{True} in $\mathcal{H}_n$ given M samples
- $\hat{f}$: output of ML model

We can decompose the error between f_{True} and $\hat{f}$

$$f_{\text{True}} - \hat{f} = \underbrace{f_{\text{True}} - f_n}_{\text{app. err.}} + \underbrace{f_n - \tilde{f}_{n,M}}_{\text{est. err.}} + \underbrace{\tilde{f}_{n,M} - \hat{f}}_{\text{opt. err.}}$$

- $f_{\text{True}} - f_n$: only due to the hypothesis space chosen
- $f_n - \tilde{f}_{n,M}$: limited by training data S_n
- $\tilde{f}_{n,M} - \hat{f}$: by training algorithm

Problem with approximation: Curse of dimensionality

Definition (CoD)

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 parameters by a factor of 10^d.
Holds for all classical algorithms, e.g. approximating functions using polynomials, trigonometric polynomials or wavelets.

2NN: a special class of functions

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

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

Common activation functions:

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

Two main problems in approximation by 2NN

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

Density: Cybenko's Universal Approximation Theorem

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"

Later: The activation must not be a polynomials (Leshno, Lin, Pinkus, and Schocken, 1993)

Finding the correct function spaces associated with 2NN

Find the functions in $\mathbb{R}^d$ that are **well approximated** by 2NN

By **well approximated**, we mean the approximation error rate is of the order $n^{-1/2}$, not depend on d

Fourier-analytic Barron spaces: Approximation error rate

Theorem (Barron 1993)

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

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

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

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

What is the norm?

A norm can be defined as:

Definition

$$|f|_{\mathcal{F},s} := \inf_{f_{e|U}=f} v'_{f,s} \tag{2}$$

Here the infimum is taken over all f is taken over all extensions f_e of f in $L_1(U)$.

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

Inverse and **Direct** approximation theorem.

Infinite-width Barron spaces: Approximation error rate

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

What is the norm?

Definition (Barron norm)

For a function f that admits the integral representation

$$\|f\|_{\mathcal{B}_p} := \inf_{\rho} \left(\mathbf{E}_{\rho} \left[|a|^p (|b| + |c|)^p \right] \right)^{1/p}, \quad 1 \leq p \leq \infty. \tag{3}$$

The infimum is taken over all ρ which the integral representation holds.

E et al. showed that For any $f \in \mathcal{B}_1$, f also $\in \mathcal{B}_\infty$ and the spaces $\mathcal{B}_\infty = \cdots = \mathcal{B}_2 = \mathcal{B}_1$. Space $\mathcal{B}$ is a Banach space.

Relationship between Fourier-analytic and Infinite-width Barron spaces

- ReLU 2NN, $\mathcal{B}(U)$ is "sandwiched" between $\mathcal{B}_{\mathcal{F},1}(U)$ and $\mathcal{B}_{\mathcal{F},2}(U)$

Given a nonempty bounded domain U in $\mathbb{R}^d$, the following holds:

$$\begin{aligned}
\mathcal{B}_{\mathcal{F},s}(U) &\subset \mathcal{B}(U) \quad \forall s \geq 2 \\
\mathcal{B}_{\mathcal{F},1}(U) &\not\subset \mathcal{B}(U) \\
\mathcal{B}(U) &\underbrace{\subset}_{?} \mathcal{B}_{\mathcal{F},1}(U)
\end{aligned}$$

Improved rates in later research

Improved rate with higher s in $\mathcal{B}_{\mathcal{F},s}$ Barron and Klusowski (2018)

- 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 $\mathcal{B}$ Ma, Siegel, and Xu (2022)

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

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

Scholars

- Andrew Barron
- Weinan E
- Jason M. Klusowski
- Jonathan W. Siegel
- Robert D Nowak
- Rahul Parhi
- Josef Teichmann
- Felix Voigtlaender, Philipp Petersen
- Frank Gao
- Maurey's Theorem

Bibliography I

Barron, A. R. and J. M. Klusowski (2018, September). Approximation and Estimation for High-Dimensional Deep Learning Networks. *arXiv:1809.03090 [cs, stat]*.

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.

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.