Two-Layer Neural Networks

Instructor: Lei Wu [1]

Topics in Deep Learning Theory

Peking University, Spring 2025

[1]School of Mathematical Sciences; Center for Machine Learning Research

Models

- A **scaled** [2] two-layer neural network (2LNN) is given by

$$f_m(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(w_j^\top x + b_j) = \frac{1}{m} a^\top \sigma(Wx + b),$$

where $a_j \in \mathbb{R}, w_j \in \mathbb{R}^d, b_j \in \mathbb{R}$ and $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is the activation function.

[2] This explicit scaling does not change the expressivity but may significantly change the optimization dynamics.

Models

- A **scaled** [2] two-layer neural network (2LNN) is given by

$$f_m(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(w_j^\top x + b_j) = \frac{1}{m} a^\top \sigma(Wx + b),$$

  where $a_j \in \mathbb{R}, w_j \in \mathbb{R}^d, b_j \in \mathbb{R}$ and $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is the activation function.

- In theoretical analysis, we also consider a more general form

$$f_m(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \varphi(x, \omega_j),$$

  where $\theta = \{(a_j, \omega_j)\}_{j=1}^m$ are the learnable parameters and $\varphi : \mathcal{X} \times \Omega \mapsto \mathbb{R}$ is a generic feature function.

[2] This explicit scaling does not change the expressivity but may significantly change the optimization dynamics.

Models

- A **scaled** [2] two-layer neural network (2LNN) is given by

$$f_m(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \sigma(w_j^\top x + b_j) = \frac{1}{m} a^\top \sigma(Wx + b),$$

  where $a_j \in \mathbb{R}, w_j \in \mathbb{R}^d, b_j \in \mathbb{R}$ and $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is the activation function.

- In theoretical analysis, we also consider a more general form

$$f_m(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \varphi(x, \omega_j),$$

  where $\theta = \{(a_j, \omega_j)\}_{j=1}^m$ are the learnable parameters and $\varphi : \mathcal{X} \times \Omega \mapsto \mathbb{R}$ is a generic feature function.

- Intuitively, 2LNNs take the same form as RFMs but the features are adaptive (i.e., learned from data) instead of fixed.

[2] This explicit scaling does not change the expressivity but may significantly change the optimization dynamics.

Questions

Some motivating questions:

- What is the performance of 2LNNs in approximating classical function spaces, like $C(\Omega), W^{k,p}(\Omega)$?

Questions

Some motivating questions:

- What is the performance of 2LNNs in approximating classical function spaces, like $C(\Omega), W^{k,p}(\Omega)$?
- Why do 2LNNs perform better than RFM/kernel methods? Is the superiority of 2LNNs over RFMs limited to high-dimensional settings

Questions

Some motivating questions:

- What is the performance of 2LNNs in approximating classical function spaces, like $C(\Omega), W^{k,p}(\Omega)$?
- Why do 2LNNs perform better than RFM/kernel methods? Is the superiority of 2LNNs over RFMs limited to high-dimensional settings
- Any other questions?

Questions

Some motivating questions:

- What is the performance of 2LNNs in approximating classical function spaces, like $C(\Omega), W^{k,p}(\Omega)$?
- Why do 2LNNs perform better than RFM/kernel methods? Is the superiority of 2LNNs over RFMs limited to high-dimensional settings
- Any other questions?

Questions

Some motivating questions:

- What is the performance of 2LNNs in approximating classical function spaces, like $C(\Omega), W^{k,p}(\Omega)$?
- Why do 2LNNs perform better than RFM/kernel methods? Is the superiority of 2LNNs over RFMs limited to high-dimensional settings
- Any other questions?

We can attempt to answer these questions from the perspective of **approximation, estimation, and optimization**. In particular, we need to think about

- What roles do the weights and the adaptivity of NNs play?
- How does the choice of activation functions influence the answers.
  - In practice, it is often observed that the choice of activation function may change the performance of our neural networks.
  - The self-gated family activation functions: $\sigma(t) = t\Phi(t)$ always show superority over classical activation functions such sigmoid and ReLU.
  - When Φ is either the CDF of $\mathcal{N}(0,1)$, these activation if called Gaussian error linear unit (GELU). If Φ is the sigmoid function, the corresponding activation is called sigmoid linear unit (SiLU).

Notations

- **Input:** We use $\mathcal{X} \subset \mathbb{R}^d$ to denote the input domain.
- **Model/hypothesis class**:

$$\mathcal{F}_{\sigma,d}^m = \left\{ x \mapsto \sum_{j=1}^m a_j \sigma(w_j^\top x + b_j) : a_j \in \mathbb{R}^m, b_j \in \mathbb{R}^m, w_j \in \mathbb{R}^d \right\}$$
$$\mathcal{F}_{\sigma,d} = \cup_{m \in \mathbb{N}_+} \mathcal{F}_{\sigma,d}^m$$

  Sometimes, we shall drop the subscript d for brevity.

- In this lecture, we define the Fourier transform as follows

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

  Then, the Fourier inversion theorem is given by

$$f(x) = \int \hat{f}(\omega) e^{i\omega^\top x} \, \mathrm{d}\omega. \tag{1}$$

Notations

- We use $\|f\|_p$ denote the L^p norm of f. Consider the Sobolev spaces $W^{k,p}$

$$\|f\|_{k,p} = \begin{cases} \left(\sum_{|\alpha|\le k} \|D^\alpha f\|_p^p\right)^{1/p} \\ \max_{|\alpha|\le k} \|D^\alpha f\|_\infty. \end{cases}$$

- For a function class $\mathcal{F}$ and $\gamma \in \mathbb{R}$, let $\gamma\mathcal{F} = \{x \mapsto \gamma f(x) : f \in \mathcal{F}\}$.
- Let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$ and $\hat{x} = x/\|x\|_2$.
- For any measurable set Ω, denote by $\mathcal{P}(\Omega)$ the set of probability distribution over Ω.

Approximate Continuous Functions

In the literature, this is called the property of **universal approximation**.

Definition 1 (UAP)

Let $\mathcal{X}$ be a compact set. A function class $\mathcal{F}$ is said to be universal approximator if $\mathcal{F}$ is dense in $C(\mathcal{X})$ with respect to the uniform metric. This is equivalent to say that for any $f \in C(\mathcal{X})$ and $\varepsilon > 0$, there exists $f \in \mathcal{F}$ such that

$$\sup_{x \in \mathcal{X}} |f(x) - h(x)| \leq \varepsilon.$$

Is $\mathcal{F}_{\sigma,d}$ dense in $C(\mathcal{X})$?

A Reduction Result

Theorem 2 (Pinkus (1999))

Suppose the activation function is chosen such that $\mathcal{F}_{\sigma,1}$ is dense in $C([0,1])$. Then, $\mathcal{F}_{\sigma,d}$ is dense in $C([0,1]^d)$.

A Reduction Result

Theorem 2 (Pinkus (1999))

Suppose the activation function is chosen such that $\mathcal{F}_{\sigma,1}$ is dense in $C([0,1])$. Then, $\mathcal{F}_{\sigma,d}$ is dense in $C([0,1]^d)$.

Proof. If $\sigma \in C^\infty(\mathbb{R})$. Then, for any $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$,

$$\frac{\partial}{\partial w_i}\sigma(w^\top x + b) = \lim_{\epsilon \to 0} \frac{\sigma(w^\top x + \epsilon e_i^\top x + b) - \sigma(w^\top x + b)}{\epsilon} \in \overline{\mathcal{F}}_{\sigma,d}$$

for $i = 1, \ldots, d$. Similarly, for any $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{N}^d$,

$$\frac{\partial}{\partial w^\alpha}\sigma(w^\top x + b) = x^\alpha \sigma^{|\alpha|}(w^\top x + b) \in \overline{\mathcal{F}}_{\sigma,d}.$$

- $\mathcal{F}_{\sigma,1}$ is dense in $C([0,1])$ implies that we can take $w = 0$ and $b \in \mathbb{R}$ such that $\sigma^k(b) \neq 0$ for any $k \in \mathbb{N}$. Therefore, $x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ are in $\bar{\mathcal{F}}_{\sigma,d}$.
- $\bar{\mathcal{F}}_{\sigma,d}$ contains all the polynomials.By Weierstrass-Stone theorem, $\bar{\mathcal{F}}_{\sigma,d}$ is dense in $C(\mathcal{X})$.
- For non-smooth σ, since $\mathcal{F}_{\sigma,1}$ is dense in $C([0,1])$, we can use a 2LNN to approximate a smooth one.

One-dimensional ReLU Nets

Lemma 3

Let $\mathrm{PL}([0,1])$ *denote the space of picewise linear functions in the domain* $[0,1]$. *Show that* $\mathrm{PL}([0,1]) = \mathcal{F}_{\mathrm{ReLU},1}$.

The proof is omitted. A direct consequence is that $\mathcal{F}_{\mathrm{ReLU},1}$ is dense in $C([0,1])$.

Reproduce the Classical Result: Cybenko (1989)

Math. Control Signals Systems (1989) 2: 303–314

Mathematics of Control, Signals, and Systems

Approximation by Superpositions of a Sigmoidal Function*

G. Cybenko†

Abstract. In this paper we demonstrate that finite linear combinations of compositions of a fixed, univariate function and a set of affine functionals can uniformly approximate any continuous function of n real variables with support in the unit hypercube; only mild conditions are imposed on the univariate function. Our results settle an open question about representability in the class of single hidden layer neural networks. In particular, we show that arbitrary decision regions can be arbitrarily well approximated by continuous feedforward neural networks with

Definition. We say that σ is *sigmoidal* if

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

Theorem 1. *Let σ be any continuous discriminatory function. Then finite sums of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j) \tag{2}$$

are dense in $C(I_n)$. In other words, given any $f \in C(I_n)$ and $\varepsilon > 0$, there is a sum, $G(x)$, of the above form, for which

$$|G(x) - f(x)| < \varepsilon \qquad \text{for all} \quad x \in I_n.$$

Reproduce the Classical Result: Cybenko (1989)

Math. Control Signals Systems (1989) 2: 303–314

Mathematics of Control, Signals, and Systems

Approximation by Superpositions of a Sigmoidal Function*

G. Cybenko†

Abstract. In this paper we demonstrate that finite linear combinations of compositions of a fixed, univariate function and a set of affine functionals can uniformly approximate any continuous function of n real variables with support in the unit hypercube; only mild conditions are imposed on the univariate function. Our results settle an open question about representability in the class of single hidden layer neural networks. In particular, we show that arbitrary decision regions can be arbitrarily well approximated by continuous feedforward neural networks with

Definition. We say that σ is *sigmoidal* if

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

Theorem 1. *Let σ be any continuous discriminatory function. Then finite sums of the form*

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(y_j^T x + \theta_j) \tag{2}$$

are dense in $C(I_n)$. In other words, given any $f \in C(I_n)$ and $\varepsilon > 0$, there is a sum, $G(x)$, of the above form, for which

$$|G(x) - f(x)| < \varepsilon \qquad \text{for all} \quad x \in I_n.$$

- Let $H(t) = 1_{[0,\infty]}(t)$ be the Heaviside step function. Then, $\mathcal{F}_{H,1}$ contains all piecewise constant functions. Therefore, $C([0,1]) \subset \overline{\mathcal{F}}_{H,1}$.
- Let σ be a generic sigmoidal function. Then, for any $t \in \mathbb{R}$, $\sigma(\beta t) \mapsto H(t)$ as $\beta \to \infty$. Therefore, $\mathcal{F}_{H,1} \subset \overline{\mathcal{F}}_{\sigma,1}$.
- Combining them gives $C([0,1]) \subset \mathcal{F}_{\sigma,1}$

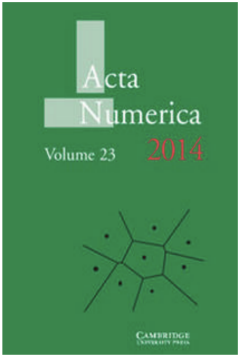
Approximate Sobolev Spaces

Acta Numerica (1999), pp. 143–195

Approximation theory of the MLP model in neural networks

Allan Pinkus
Department of Mathematics,
Technion – Israel Institute of Technology,
Haifa 32000, Israel
E-mail: pinkus@tx.technion.ac.il

In this survey we discuss various approximation-theoretic problems that arise in the multilayer feedforward perceptron (MLP) model in neural networks. The MLP model is one of the more popular and practical of the many neural network models. Mathematically it is also one of the simpler models. Nonetheless the mathematics of this model is not well understood, and many of these problems are approximation-theoretic in character. Most of the research we will discuss is of very recent vintage. We will report on what has been done and on various unanswered questions. We will not be presenting practical (algorithmic) methods. We will, however, be exploring the capabilities and limitations of this model.



Pinkus, (1999) is a **VERY** nice paper.

Main Result

Theorem 4 (Theorem 6.8 in Pinkus (1999))

Assume $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is such that $\sigma \in C^{\infty}(\Theta)$ on some open interval Θ, and σ is not a polynomial on Θ. Then, for any $p \in [1, +\infty]$ and $m \geq 1$ and $d \geq 2$,

$$\sup_{\|f^*\|_{W^{s,p}} \leq 1} \inf_{f \in \mathcal{F}^m_{\sigma,d}} \|f - f^*\|_p \leq C m^{-\frac{s}{d}}$$

- Does this imply that smooth activations can enable the adaptivity of NNs to any order of smoothness? This may explain the popularity of smooth activation functions like GELU!
- **Can we establish similar results for the sample complexity?**

Approximate Sobolev Spaces (Cont'd)

Theorem 5 (Corollary 6.10 in Pinkus (1999))

Let $\sigma(z) = \max(0, z^k)$. *Then,*

$$\sup_{\|f^*\|_{W^{2,s}} \leq 1} \inf_{f \in \mathcal{F}_{m,\sigma}} \|f - f^*\|_2 \lesssim m^{-s/d}$$

for $s = 1, 2, \ldots, k + 1 + \frac{d-1}{2}$.

- The rate looks strange [3] as it implies that ReLU-activated 2LNNs can adapt to the smoothness of target function beyond $s = 2$.
- However, the rate suffers from the **curse of dimensionality (CoD)**.

[3] I will check the proof and get back to you!

Avoid CoD via Monte-Carlo Approximation?

- Take a Monte-Carlo discretization viewpoint of 2LNN approximations:

$$f_m(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \varphi(x, w_j) \to \int_{\mathbb{R}\times\Omega} a\varphi(x,w)\,\mathrm{d}\rho(a,w) = f_\rho(x),$$

  where $(a, w_j) \overset{iid}{\sim} \rho$ and $\rho \in \mathcal{P}(\mathbb{R}\times\Omega)$.

Avoid CoD via Monte-Carlo Approximation?

- Take a Monte-Carlo discretization viewpoint of 2LNN approximations:

$$f_m(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \varphi(x, w_j) \to \int_{\mathbb{R}\times\Omega} a\varphi(x,w)\,\mathrm{d}\rho(a,w) = f_\rho(x),$$

  where $(a, w_j) \overset{iid}{\sim} \rho$ and $\rho \in \mathcal{P}(\mathbb{R}\times\Omega)$.
- One can view f_ρ as an infinitely-wide 2LNN.

Avoid CoD via Monte-Carlo Approximation?

- Take a Monte-Carlo discretization viewpoint of 2LNN approximations:

$$f_m(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \varphi(x, w_j) \to \int_{\mathbb{R}\times\Omega} a\varphi(x,w)\,\mathrm{d}\rho(a,w) = f_\rho(x),$$

  where $(a, w_j) \overset{iid}{\sim} \rho$ and $\rho \in \mathcal{P}(\mathbb{R}\times\Omega)$.
- One can view f_ρ as an infinitely-wide 2LNN.
- Then the approximation error is given by standard Monte-Carlo estimate:

$$\mathbb{E}_\theta[|f_m(\cdot;\theta) - f_\rho|^2] \le \frac{\mathbb{E}_{x\sim\mathbb{P}_x,(a,w)\sim\rho}[a^2\varphi^2(x,w)]}{m}$$

Avoid CoD via Monte-Carlo Approximation?

- Take a Monte-Carlo discretization viewpoint of 2LNN approximations:

$$f_m(x;\theta) = \frac{1}{m}\sum_{j=1}^{m} a_j \varphi(x, w_j) \to \int_{\mathbb{R}\times\Omega} a\varphi(x,w)\,\mathrm{d}\rho(a,w) = f_\rho(x),$$

  where $(a, w_j) \overset{iid}{\sim} \rho$ and $\rho \in \mathcal{P}(\mathbb{R}\times\Omega)$.
- One can view f_ρ as an infinitely-wide 2LNN.
- Then the approximation error is given by standard Monte-Carlo estimate:

$$\mathbb{E}_\theta[|f_m(\cdot;\theta) - f_\rho|^2] \le \frac{\mathbb{E}_{x\sim\mathbb{P}_x,(a,w)\sim\rho}[a^2\varphi^2(x,w)]}{m}$$

- Assume $\sup_{x,w}|\varphi(x,w)| \le 1$. Then,

$$\mathbb{E}_\theta[|f_m(\cdot;\theta) - f_\rho|^2] \le \frac{\mathbb{E}[a^2]}{m}.$$

Barron Spaces

The previous derivation motivates the following characterization of function complexity:

Definition 6 (Barron Spaces)

Given a function $f \in \mathbb{R}^{\mathcal{X}}$, let $Q_f = \{\rho \in \mathcal{P}(\mathbb{R} \times \Omega) : f(x) = f_\rho(x) \text{ for } x \in \mathcal{X}\}$. Then, define the Barron norm

$$\|f\|_{\mathcal{B}_p} = \inf_{\rho \in Q_f} \sqrt[p]{\mathbb{E}[|a|^p]},$$

and accordingly $\mathcal{B}_p = \{f : \|f\|_{\mathcal{B}_p} < \infty\}$.

Barron Spaces

The previous derivation motivates the following characterization of function complexity:

Definition 6 (Barron Spaces)

Given a function $f \in \mathbb{R}^{\mathcal{X}}$, let $Q_f = \{\rho \in \mathcal{P}(\mathbb{R} \times \Omega) : f(x) = f_\rho(x) \text{ for } x \in \mathcal{X}\}$. Then, define the Barron norm

$$\|f\|_{\mathcal{B}_p} = \inf_{\rho \in Q_f} \sqrt[p]{\mathbb{E}[|a|^p]},$$

and accordingly $\mathcal{B}_p = \{f : \|f\|_{\mathcal{B}_p} < \infty\}$.

- The step "inf" is crucial as the representations of a function Q_f are non-unique and NNs can adapt the "optimal" one. For instance, for $f(x) := \sigma(x_1) = \sigma(x_1) - \sigma(x_2) + \sigma(x_2)$, $\|f\|_{\mathcal{B}} = 1$ instead of 3.

Barron Spaces

The previous derivation motivates the following characterization of function complexity:

Definition 6 (Barron Spaces)

Given a function $f \in \mathbb{R}^{\mathcal{X}}$, let $Q_f = \{\rho \in \mathcal{P}(\mathbb{R} \times \Omega) : f(x) = f_\rho(x) \text{ for } x \in \mathcal{X}\}$. Then, define the Barron norm

$$\|f\|_{\mathcal{B}_p} = \inf_{\rho \in Q_f} \sqrt[p]{\mathbb{E}[|a|^p]},$$

and accordingly $\mathcal{B}_p = \{f : \|f\|_{\mathcal{B}_p} < \infty\}$.

- The step "inf" is crucial as the representations of a function Q_f are non-unique and NNs can adapt the "optimal" one. For instance, for $f(x) := \sigma(x_1) = \sigma(x_1) - \sigma(x_2) + \sigma(x_2)$, $\|f\|_{\mathcal{B}} = 1$ instead of 3.
- The case of $p = 2$ corresponds to "variance" complexity of a function. Moreover, we can show $\mathcal{B}_p = \mathcal{B}_q$ for all $p, q \in [1, \infty]$ (see the next slide), we shall use a unifying notation $\mathcal{B}$ to denote the Barron spaces.

Barron Spaces

The previous derivation motivates the following characterization of function complexity:

Definition 6 (Barron Spaces)

Given a function $f \in \mathbb{R}^{\mathcal{X}}$, let $Q_f = \{\rho \in \mathcal{P}(\mathbb{R} \times \Omega) : f(x) = f_\rho(x) \text{ for } x \in \mathcal{X}\}$. Then, define the Barron norm

$$\|f\|_{\mathcal{B}_p} = \inf_{\rho \in Q_f} \sqrt[p]{\mathbb{E}[|a|^p]},$$

and accordingly $\mathcal{B}_p = \{f : \|f\|_{\mathcal{B}_p} < \infty\}$.

- The step "inf" is crucial as the representations of a function Q_f are non-unique and NNs can adapt the "optimal" one. For instance, for $f(x) := \sigma(x_1) = \sigma(x_1) - \sigma(x_2) + \sigma(x_2)$, $\|f\|_{\mathcal{B}} = 1$ instead of 3.
- The case of $p = 2$ corresponds to "variance" complexity of a function. Moreover, we can show $\mathcal{B}_p = \mathcal{B}_q$ for all $p, q \in [1, \infty]$ (see the next slide), we shall use a unifying notation $\mathcal{B}$ to denote the Barron spaces.
- One can prove that $\mathcal{B}$ is a Banach space, i.e., $\|\cdot\|_{\mathcal{B}}$ is well-defined norm and $\mathcal{B}$ is complete wrt this norm.

Barron Spaces (Cont'd)

Lemma 7

$\mathcal{B}_p = \mathcal{B}_q$ for all $p, q \in [1, \infty]$.

Proof:

- By Jensen's inequality, it is obvious that $\mathcal{B}_\infty \subset \mathcal{B}_p \subset \mathcal{B}_q \subset \mathcal{B}_1$ for $1 \leq q \leq p \leq \infty$.
- Next, we only need to prove $\mathcal{B}_1 \subset \mathcal{B}_\infty$.
- If $f \in \mathcal{B}_1$, then there exists ρ such that $f = f_\rho$ and $\mathbb{E}[|a|] = \|f\|_{\mathcal{B}_1}$. Moreover, due to

$$\int a\varphi(x, w)\, \mathrm{d}\rho(a, w) = \int a_+\varphi(x, w)\, \mathrm{d}\rho(a, w) - \int a_-\varphi(x, w)\, \mathrm{d}\rho(a, w),$$

  WLOG, we can assume $a > 0$ almost surely.

- Then, we have

$$\int a\varphi(x,w)\,\mathrm{d}\rho(a,w) = \|f\|_{\mathcal{B}_1}\int \varphi(x,w)\frac{a}{\|f\|_{\mathcal{B}_1}}\rho(a)\rho(w|a)\,\mathrm{d}a\,\mathrm{d}w$$
$$= \|f\|_{\mathcal{B}_1}\int \varphi(x,w)\tilde{\rho}(a,w)\,\mathrm{d}a\,\mathrm{d}w,$$

  where

$$\tilde{\rho}(a,w) = \frac{a}{\|f\|_{\mathcal{B}_1}}\rho(a)\rho(w|a).$$

Remark: The underlying reason stems from the **adaptivity** of NNs. For any $f \in \mathcal{B}$, there are many different representations ρ's such that $f = f_\rho$ and NNs can adapt to the best one.

Approximation Theorem

Theorem 8 (Controllable Approximation)

There exist $\tilde{\theta}$ such that

$$\|f(\cdot;\tilde{\theta}) - f^*\|_2 \leq \frac{\|f^*\|_{\mathcal{B}}}{\sqrt{m}}, \quad \frac{1}{m}\sum_{j=1}^{m}|a_j| \leq \|f^*\|_{\mathcal{B}}.$$

Remark: The norm control of the approximator is important. For instance, we will use this property to establish an upper bound of the estimation error.

Approximation Theorem

Theorem 8 (Controllable Approximation)

There exist $\tilde{\theta}$ such that

$$\|f(\cdot;\tilde{\theta}) - f^*\|_2 \leq \frac{\|f^*\|_{\mathcal{B}}}{\sqrt{m}}, \quad \frac{1}{m}\sum_{j=1}^{m}|a_j| \leq \|f^*\|_{\mathcal{B}}.$$

Remark: The norm control of the approximator is important. For instance, we will use this property to establish an upper bound of the estimation error.

Proof: There exist ρ such that $f = f_\rho$ and $\|a\|_\infty \leq \|f^*\|_{\mathcal{B}}$. Let $(a_j, w_j) \overset{iid}{\sim} \rho$. Then,

$$\mathbb{E}_\theta \|f(\cdot;\theta) - f^*\|_2^2 \leq \frac{\mathbb{E}[a^2]}{m} \leq \frac{\|f^*\|_{\mathcal{B}}^2}{m} \tag{2}$$

Thus, there must exist $\tilde{\theta}$ such the property to be hold.

Capacity Control: Setup

- For controlling the capacity, we shall focus on the ridge-type feature $\varphi(x, w) = \sigma(w^\top x)$ [4].

[4] The bias is omitted for brevity!

Capacity Control: Setup

- For controlling the capacity, we shall focus on the ridge-type feature $\varphi(x, w) = \sigma(w^\top x)$ [4].
  - For a generic $\varphi(\cdot, w)$, it is impossible to establish the capacity control as $\{\varphi(\cdot, w)\}_{w \in \Omega}$ can be arbitrarily complex.

[4] The bias is omitted for brevity!

Capacity Control: Setup

- For controlling the capacity, we shall focus on the ridge-type feature $\varphi(x, w) = \sigma(w^\top x)$ [4].
  - For a generic $\varphi(\cdot, w)$, it is impossible to establish the capacity control as $\{\varphi(\cdot, w)\}_{w \in \Omega}$ can be arbitrarily complex.
  - Question: 1) Why is this condition not necessary for approximation? 2) Why is this condition not necessary for establishing the capacity control for RFMs?

[4] The bias is omitted for brevity!

Capacity Control: Setup

- For controlling the capacity, we shall focus on the ridge-type feature $\varphi(x, w) = \sigma(w^\top x)$ [4].
  - For a generic $\varphi(\cdot, w)$, it is impossible to establish the capacity control as $\{\varphi(\cdot, w)\}_{w\in\Omega}$ can be arbitrarily complex.
  - Question: 1) Why is this condition not necessary for approximation? 2) Why is this condition not necessary for establishing the capacity control for RFMs?
- Specifically, we shall focus on ReLU-activated 2LNNs, i.e., $\sigma(z) = \max(0, z)$. Analogous results can be extended to general Lipschitz activation functions.

[4] The bias is omitted for brevity!

Capacity Control: Setup

- For controlling the capacity, we shall focus on the ridge-type feature $\varphi(x, w) = \sigma(w^\top x)$ [4].
  - For a generic $\varphi(\cdot, w)$, it is impossible to establish the capacity control as $\{\varphi(\cdot, w)\}_{w\in\Omega}$ can be arbitrarily complex.
  - Question: 1) Why is this condition not necessary for approximation? 2) Why is this condition not necessary for establishing the capacity control for RFMs?
- Specifically, we shall focus on ReLU-activated 2LNNs, i.e., $\sigma(z) = \max(0, z)$. Analogous results can be extended to general Lipschitz activation functions.
- Assume $\mathcal{X} = \mathbb{S}^{d-1}$, $\Omega = \mathbb{S}^{d-1}$. Then, $|\varphi(x, w)| \leq 1$.

[4] The bias is omitted for brevity!

Rademacher Complexity Bound

Proposition 9

Let $\mathcal{F}_Q = \{f \in \mathcal{B} : \|f\|_{\mathcal{B}} \leq Q\}$. Then, under the aforementioned condition, we have

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}_Q) \lesssim \frac{Q}{\sqrt{n}}$$

Rademacher Complexity Bound

Proposition 9

Let $\mathcal{F}_Q = \{f \in \mathcal{B} : \|f\|_{\mathcal{B}} \leq Q\}$. Then, under the aforementioned condition, we have

$$\widehat{\mathrm{Rad}}_n(\mathcal{F}_Q) \lesssim \frac{Q}{\sqrt{n}}$$

- For ReLU networks, $f_m(x;\theta) = \frac{1}{m}\sum_{j=1}^m a_j\sigma(w_j^\top x)$, we define the path norm

$$\|\theta\|_{\mathcal{P}} = \frac{1}{m}\sum_{j=1}^m |a_j|\|w_j\|.$$

Let $\mathcal{F}_Q^m = \{f_m(\cdot;\theta) : \|\theta\|_{\mathcal{P}} \leq Q\}$. Noticing

$$\frac{1}{m}\sum_{j=1} a_j\sigma(w_j^\top x) = \frac{1}{m}\sum_{j=1} a_j\|w_j\|_1\sigma(\hat{w}_j^\top x) = \int a\sigma(w^\top x)\,\mathrm{d}\hat{\rho}_m(a,w),$$

where $\hat{\rho}_m(a,w) = \frac{1}{m}\sum_{j=1}^m \delta(a - a_j\|w_j\|)\delta(w - \hat{w}_j)$. We have $\mathcal{F}_Q^m \subset \mathcal{F}_Q$. This is due to

- Therefore, we have the capacity control for the model class: $\widehat{\mathrm{Rad}}_n(\mathcal{F}_Q^m) \lesssim \frac{Q}{\sqrt{n}}$.

Proof of Proposition 9

$$
\begin{aligned}
n\widehat{\mathrm{Rad}}_n(\mathcal{F}_Q) &= \mathbb{E}_\xi[\sup_{f\in\mathcal{F}_Q}\sum_{i=1}^n \xi_i\,\mathbb{E}_\rho[a\sigma(w^\top x_i)]] = \mathbb{E}_{\boldsymbol{\xi}}[\sup_{f\in\mathcal{F}_Q}\mathbb{E}_\rho[|a|\|w\|_1\sum_{i=1}^n \xi_i\sigma(\hat{w}^\top x_i)]] \\
&\le \mathbb{E}_\xi[\sup_{f\in\mathcal{F}_Q}\mathbb{E}_\rho[|a|\|w\|_1 \sup_{\|w\|_1\le 1}|\sum_{i=1}^n \xi_i\sigma(w^\top x_i)|] \\
&\le Q\,\mathbb{E}_\xi[\sup_{\|w\|_1\le 1}|\sum_{i=1}^n \xi_i\sigma(w^\top x_i)|] \\
&\le Q\,\mathbb{E}_\xi[\sup_{\|w\|_1\le 1}\sum_{i=1}^n \xi_i\sigma(w^\top x_i)] + Q\,\mathbb{E}_\xi[\sup_{\|w\|_1\le 1}-\sum_{i=1}^n \xi_i\sigma(w^\top x_i)] \\
&= 2Q\,\mathbb{E}_\xi[\sup_{\|w\|_1\le 1}\sum_{i=1}^n \xi_i\sigma(w^\top x_i)] \le 2Q\,\mathbb{E}_\xi[\sup_{\|w\|_1\le 1}\sum_{i=1}^n \xi_i w^\top x_i]
\end{aligned}
$$

Hence, the problem is reduced to bound the Rademacher complexity of linear class.

Generalization Bound

Consider the path norm-regularized estimator:

$$\hat{\theta}_n = \operatorname*{argmin}_{\theta} \frac{1}{2n} \sum_{i=1}^{n} (f(x_i) - f^*(x_i))^2 + \frac{\lambda}{\sqrt{n}} \|\theta\|_{\mathcal{P}}. \tag{3}$$

For technical simplicity, assume $\sup_{x \in X} |f^*(x)| \leq 1$ and use the truncated network:

$$\tilde{f}_m(x; \theta) = \min(\max(f_m(x; \theta), -1), 1).$$

Theorem 10

Assume $\lambda \geq C$, where C is an absolute constant. For any $\delta \in (0, 1)$, with probability $1 - \delta$ over the choice of training samples, we have

$$\|f(\cdot; \hat{\theta}_n) - f^*\|_2^2 \lesssim \frac{\|f^*\|_{\mathcal{B}}^2}{m} + \frac{\|f^*\|_{\mathcal{B}}}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}}.$$

Remark: No CoD for both approximation and estimation errors. The proof is left as homework.

What kind of functions lie in the Barron Space?

We shall focus on the ReLU features.

- Finite-width neural nets: $f_m(x) = \sum_{j=1}^m a_j \sigma(w_j^\top x)$. Obviously,

$$\|f_m\|_{\mathcal{B}} \leq \sum_{j=1}^m |a_j| \|w_j\|.$$

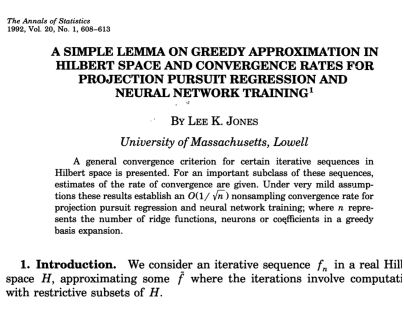
  This implies that a very wide 2LNN can be expressed as long as the norm is controlled!!!

- General functions with a linear low-dimensional structure: $f(x) = g(W^\top x)$ with $g : \mathbb{R}^k \mapsto \mathbb{R}$. Obviously,

$$\|f\|_{\mathcal{B}} \leq \|W\|_2 \|g\|_{\mathcal{B}}.$$

  This implies that $\|f\|_{\mathcal{B}}$ only depends on the intrinsic dimension k rather than the ambient dimension d.

Fourier Analysis of Barron Spaces



The Annals of Statistics
1992, Vol. 20, No. 1, 608–613

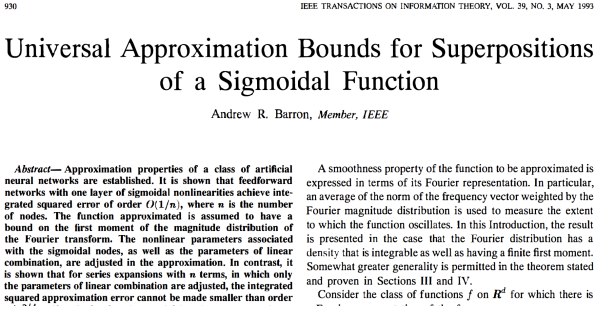
A SIMPLE LEMMA ON GREEDY APPROXIMATION IN HILBERT SPACE AND CONVERGENCE RATES FOR PROJECTION PURSUIT REGRESSION AND NEURAL NETWORK TRAINING[1]

BY LEE K. JONES

University of Massachusetts, Lowell

A general convergence criterion for certain iterative sequences in Hilbert space is presented. For an important subclass of these sequences, estimates of the rate of convergence are given. Under very mild assumptions these results establish an $O(1/\sqrt{n})$ nonsampling convergence rate for projection pursuit regression and neural network training; where n represents the number of ridge functions, neurons or coefficients in a greedy basis expansion.

1. Introduction. We consider an iterative sequence f_n in a real Hill space H, approximating some $\hat{f}$ where the iterations involve computati with restrictive subsets of H.



930 IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 39, NO. 3, MAY 1993

Universal Approximation Bounds for Superpositions of a Sigmoidal Function

Andrew R. Barron, *Member, IEEE*

***Abstract*— Approximation properties of a class of artificial neural networks are established. It is shown that feedforward networks with one layer of sigmoidal nonlinearities achieve integrated squared error of order $O(1/n)$, where n is the number of nodes. The function approximated is assumed to have a bound on the first moment of the magnitude distribution of the Fourier transform. The nonlinear parameters associated with the sigmoidal nodes, as well as the parameters of linear combination, are adjusted in the approximation. In contrast, it is shown that for series expansions with n terms, in which only the parameters of linear combination are adjusted, the integrated squared approximation error cannot be made smaller than order**

A smoothness property of the function to be approximated is expressed in terms of its Fourier representation. In particular, an average of the norm of the frequency vector weighted by the Fourier magnitude distribution is used to measure the extent to which the function oscillates. In this Introduction, the result is presented in the case that the Fourier distribution has a density that is integrable as well as having a finite first moment. Somewhat greater generality is permitted in the theorem stated and proven in Sections III and IV.

Consider the class of functions f on R^d for which there is

Figure 1: **Left:** 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, pages 608–613. **Right:** Barron, A. R. (1993). *Universal approximation bounds for superpositions of a sigmoidal function.* IEEE Transactions on Information theory, 39(3):930–945.

The Jones's trick

- Recall the Fourier inversion theorem:

$$f(x) = \int \hat{f}(\omega) e^{i\omega x} \, \mathrm{d}\omega. \tag{4}$$

The Jones's trick

- Recall the Fourier inversion theorem:

$$f(x) = \int \hat{f}(\omega) e^{i\omega x} \, \mathrm{d}\omega. \tag{4}$$

- Let $\hat{f}(\omega) = |\hat{f}(\omega)| e^{ib(\omega)}$. Then, we can rewrite (4) as follows

$$f(x) = \int |\hat{f}(\omega)| e^{i(b(\omega) + \omega^\top x)} \, \mathrm{d}\omega = \int |\hat{f}(\omega)| \cos(b(\omega) + \omega^\top x) \, \mathrm{d}\omega. \tag{5}$$

Assume $C_f = \int |\hat{f}(\omega)| \, \mathrm{d}\omega$ and let $\mathrm{d}\pi(\omega) = \frac{|\hat{f}(\omega)|}{C_f} \, \mathrm{d}\omega$. Then,

$$f(x) = C_f \, \mathbb{E}_{\omega \sim \pi}[\cos(\omega^\top x + b(\omega))]. \tag{6}$$

The Jones's trick

- Recall the Fourier inversion theorem:

$$f(x) = \int \hat{f}(\omega) e^{i\omega x} \, \mathrm{d}\omega. \tag{4}$$

- Let $\hat{f}(\omega) = |\hat{f}(\omega)| e^{ib(\omega)}$. Then, we can rewrite (4) as follows

$$f(x) = \int |\hat{f}(\omega)| e^{i(b(\omega) + \omega^\top x)} \, \mathrm{d}\omega = \int |\hat{f}(\omega)| \cos(b(\omega) + \omega^\top x) \, \mathrm{d}\omega. \tag{5}$$

  Assume $C_f = \int |\hat{f}(\omega)| \, \mathrm{d}\omega$ and let $\mathrm{d}\pi(\omega) = \frac{|\hat{f}(\omega)|}{C_f} \, \mathrm{d}\omega$. Then,

$$f(x) = C_f \, \mathbb{E}_{\omega \sim \pi}[\cos(\omega^\top x + b(\omega))]. \tag{6}$$

- This corresponds to two-layer neural networks with $\cos$ activation functions.

The Barron's trick

- Can we represent $t \to \cos(\|\omega\| t + b(\omega))$ as an "expectation form" using traditional activation functions like ReLU?

The Barron's trick

- Can we represent $t \to \cos(\|\omega\| t + b(\omega))$ as an "expectation form" using traditional activation functions like ReLU?
- Consider $f \in C(\mathcal{X})$ and let f_e be an extension of f. Since, $f(0) = \int \hat{f}_e(\omega)\, \mathrm{d}\omega$, we can express f as follows

$$\begin{aligned} f(x) - f(0) &= \int (e^{i\omega^\top x} - 1)\hat{f}_e(\omega)\, \mathrm{d}\omega \\ &= \int \frac{e^{i\omega^\top x} - 1}{\|\omega\|} \|\omega\| \hat{f}_e(\omega)\, \mathrm{d}\omega \\ &= \int \frac{\cos(\omega^\top x + b(\omega)) - \cos(b(\omega))}{\|\omega\|} \|\omega\| |\hat{f}_e(\omega)|\, \mathrm{d}\omega \\ &= \int g(\omega, x) \|\omega\| |\hat{f}_e(\omega)|\, \mathrm{d}\omega, \end{aligned}$$

where

$$g(x, w) = \frac{\cos(\omega^\top x + b(\omega)) - \cos(b(\omega))}{\|\omega\|}.$$

The Barron's trick (Cont'd)

- Assume that $C_1(f) := \int \|\omega\| |\hat{f}(\omega)| \, \mathrm{d}\omega < \infty$. Then,

$$f(x) - f(0) = C_1(f) \, \mathbb{E}_{\omega \sim \pi}[g(x, \omega)] = C_1(f) \, \mathbb{E}_{\omega \sim \pi}[h(\hat{w}^\top x, \omega)], \tag{7}$$

  where

$$h(t, \omega) = (\cos(\|\omega\| t + b(\omega)) - \cos(b(\omega)))/\|\omega\|$$

  is Lipschitz with respect to t.

- If we can further express a Lipschitz function in an "expectation form", then we complete the proof. This is indeed doable as explained later!!!

Lipschitz Functions in Expectation Forms.

Lemma 11

Suppose $h \in C^1([-1,1])$. Then, we have

$$h(t) = h(0) + \int_0^1 h'(s)H(t-s)\,\mathrm{d}s + \int_0^{-1} h'(s)H(-t+s)\,\mathrm{d}s.$$

Proof: When $t \geq 0$, we have

$$h(t) = h(0) + \int_0^t h'(s)\,\mathrm{d}s = h(0) + \int_0^1 h'(s)H(t-s)\,\mathrm{d}s.$$

If $t < 0$, the proof is similar.

Remark:

- Plugging this lemma into (7), we obtain

$$f(x) - f(0) = C_1(f)\,\mathbb{E}_{\omega\sim\pi, s\sim\mu_0}[-\sin(\|\omega\|s + b(\omega))H(\hat{\omega}^\top x - s)] + I_2,$$

where $\mu_0 = \mathrm{Unif}([0,1])$ and I_2 accounts for the negative part .

An Alternative Approach

- By Lemma 11, $e^{ict} = 0 + ic\int_0^1 e^{is}H(t-s)\,\mathrm{d}s + ic\int_0^{-1} e^{is}H(s-t)\,\mathrm{d}s$.

An Alternative Approach

- By Lemma 11, $e^{ict} = 0 + ic\int_0^1 e^{is}H(t-s)\,\mathrm{d}s + ic\int_0^{-1} e^{is}H(s-t)\,\mathrm{d}s$.
-

$$f(x) = \int e^{i\omega^\top x}\hat{f}_e(\omega)\,\mathrm{d}\omega = \int e^{i\|\omega\|\hat{\omega}^\top x}\hat{f}_e(\omega)\,\mathrm{d}\omega$$
$$= \int \hat{f}_e(\omega)\,\mathrm{d}\omega + \int \left(i\|\omega\|\int_0^1 e^{i\|\omega\|s}H(\hat{\omega}^\top x - s)\,\mathrm{d}s\right)\hat{f}_e(\omega)\,\mathrm{d}\omega + I_2,$$

where I_2 accounts for the negative part.

An Alternative Approach

- By Lemma 11, $e^{ict} = 0 + ic\int_0^1 e^{is}H(t-s)\,\mathrm{d}s + ic\int_0^{-1} e^{is}H(s-t)\,\mathrm{d}s$.
- $$\begin{aligned} f(x) &= \int e^{i\omega^\top x}\hat{f}_e(\omega)\,\mathrm{d}\omega = \int e^{i\|\omega\|\hat{\omega}^\top x}\hat{f}_e(\omega)\,\mathrm{d}\omega \\ &= \int \hat{f}_e(\omega)\,\mathrm{d}\omega + \int \left(i\|\omega\| \int_0^1 e^{i\|\omega\|s} H(\hat{\omega}^\top x - s)\,\mathrm{d}s \right) \hat{f}_e(\omega)\,\mathrm{d}\omega + I_2, \end{aligned}$$

  where I_2 accounts for the negative part.
- $$\begin{aligned} f(x) - f(0) &= i\int_{\mathbb{R}^d}\int_0^1 e^{i\|\omega\|s} H(\omega^\top x - s)\,\mathrm{d}s\hat{f}_e(\omega)\,\mathrm{d}\omega + I_2 \\ &= i\int_{\mathbb{R}}\int_0^1 e^{i\|\omega\|t+b(\omega)} H(\hat{\omega}^\top x - t)\|\omega\||\hat{f}_e(\omega)|\,\mathrm{d}t\,\mathrm{d}\omega + I_2 \\ &= \underbrace{-\int_{\mathbb{R}}\int_0^1 \sin(\|\omega\|t + b(\omega)) H(\hat{\omega}^\top x - t)\|\omega\||\hat{f}_e(\omega)|\,\mathrm{d}t\,\mathrm{d}\omega}_{I_1} + I_2. \end{aligned}$$

An Alternative Approach

- By Lemma 11, $e^{ict} = 0 + ic\int_0^1 e^{is}H(t-s)\,\mathrm{d}s + ic\int_0^{-1} e^{is}H(s-t)\,\mathrm{d}s$.
- $$\begin{aligned} f(x) &= \int e^{i\omega^\top x}\hat{f}_e(\omega)\,\mathrm{d}\omega = \int e^{i\|\omega\|\hat{\omega}^\top x}\hat{f}_e(\omega)\,\mathrm{d}\omega \\ &= \int \hat{f}_e(\omega)\,\mathrm{d}\omega + \int \left(i\|\omega\| \int_0^1 e^{i\|\omega\|s} H(\hat{\omega}^\top x - s)\,\mathrm{d}s \right) \hat{f}_e(\omega)\,\mathrm{d}\omega + I_2, \end{aligned}$$

  where I_2 accounts for the negative part.
- $$\begin{aligned} f(x) - f(0) &= i\int_{\mathbb{R}^d}\int_0^1 e^{i\|\omega\|s} H(\omega^\top x - s)\,\mathrm{d}s\hat{f}_e(\omega)\,\mathrm{d}\omega + I_2 \\ &= i\int_{\mathbb{R}}\int_0^1 e^{i\|\omega\|t + b(\omega)} H(\hat{\omega}^\top x - t)\|\omega\|\,|\hat{f}_e(\omega)|\,\mathrm{d}t\,\mathrm{d}\omega + I_2 \\ &= \underbrace{-\int_{\mathbb{R}}\int_0^1 \sin(\|\omega\|t + b(\omega)) H(\hat{\omega}^\top x - t)\|\omega\|\,|\hat{f}_e(\omega)|\,\mathrm{d}t\,\mathrm{d}\omega}_{I_1} + I_2. \end{aligned}$$
- Lastly, we can apply the Jones's trick to convert the above integration into expectation.

ReLU Nets: One-dimensional Case

Lemma 12

Suppose $h \in C^2([-1,1])$. *Then, we have*

$$h(t) = h(0) + h'(0)t + \int_0^1 h''(s)\sigma_{\mathrm{ReLU}}(t-s)\,\mathrm{d}s + \int_0^{-1} h''(s)\sigma_{\mathrm{ReLU}}(-t+s)\,\mathrm{d}s.$$

ReLU Nets: One-dimensional Case

Lemma 12

Suppose $h \in C^2([-1,1])$. Then, we have

$$h(t) = h(0) + h'(0)t + \int_0^1 h''(s)\sigma_{\mathrm{ReLU}}(t-s)\,\mathrm{d}s + \int_0^{-1} h''(s)\sigma_{\mathrm{ReLU}}(-t+s)\,\mathrm{d}s.$$

Proof: When $t \geq 0$ (the proof for $t < 0$ is similar), we have

$$\begin{aligned}
h(t) &= h(0) + \int_0^t h'(\tau)\,\mathrm{d}\tau = h(0) + \int_0^t \left(h'(0) + \int_0^s h''(s)\,\mathrm{d}s\right)\mathrm{d}\tau \\
&= h(0) + h'(0)t + \int_0^t \int_0^s h''(s)\,\mathrm{d}s\,\mathrm{d}\tau = h(0) + h'(0)t + \int_0^t \int_s^t h''(s)\,\mathrm{d}s\,\mathrm{d}\tau \\
&= h(0) + h'(0)t + \int_0^t h''(s)(t-s)\,\mathrm{d}s \\
&= h(0) + h'(0)t + \int_0^1 h''(s)(t-s)H(t-s)\,\mathrm{d}s \\
&= h(0) + h'(0)t + \int_0^1 h''(s)\sigma_{\mathrm{ReLU}}(t-s)\,\mathrm{d}s.
\end{aligned}$$

ReLU Nets: The High-dimensional Case

- Applying the preceding lemma to e^{ict} gives

$$e^{ict} - ict - 1 = -c^2 \int_0^1 e^{ics}\sigma(t-s)\,\mathrm{d}s - c^2 \int_0^{-1} e^{ics}\sigma(-t+s)\,\mathrm{d}s. \tag{8}$$

- Then, by the Fourier inverse theorem, we have

$$\begin{aligned} f(x) - \nabla f(0)^\top x - f(0) &= \int_{\mathbb{R}^d} (e^{i\omega^\top x} - i\omega^\top x - 1)\hat{f}_e(\omega)\,\mathrm{d}\omega \\ &= -\int_{\mathbb{R}^d}\int_0^1 \|\omega\|^2 \sigma(\hat{\omega}^\top x - s) e^{i\|\omega\|s}\,\mathrm{d}s\hat{f}_e(\omega)\,\mathrm{d}\omega + I_2 \\ &= \underbrace{-\int_{\mathbb{R}^d}\int_0^1 \cos(\|\omega\|t + b(\omega))\sigma(\hat{\omega}^\top x - t)\|\omega\|^2 |\hat{f}(\omega)|\,\mathrm{d}t\,\mathrm{d}\omega}_{I_1} + I_2, \end{aligned} \tag{9}$$

  where the I_2 is similar to I_1, accounting for the case $\hat{\omega}^\top x \le 0$. The explicit form of I_2 is omitted for notation simplicity.
- If $\int \|\omega\|^2 |\hat{f}(\omega)|\,\mathrm{d}\omega < \infty$, by using the Jones' trick, we can write (9) in an expectation form.
- The linear part can be expressed with two ReLU neurons: $\nabla f(0)^\top x = \mathrm{ReLU}(w^\top x) - \mathrm{ReLU}(-w^\top x)$ with $w = \nabla f(0)$.

The ReLU^k Activation Functions

The previous idea can be extended to the general ReLU^k activation function:

$$\mathrm{ReLU}^k(z) = \max(0, z^k).$$

Definition 13 (Spectral Barron norm)

For any $f \in C(\mathcal{X})$, define

$$\|f\|_{\mathbb{F}_s} := \inf_{f_e|_\Omega = f} \int (1 + \|\omega\|)^s |\hat{f}_e(\omega)| \, \mathrm{d}\omega,$$

where the infimum is taken over all the extension of f.

Theorem 14

If $\|f\|_{\mathbb{F}_{k+1}} < \infty$, then there exits a two-layer neural net f_m activated by ReLU^k such that

$$\|f_m - f\|^2_{L^2(\rho)} \lesssim \frac{\|f\|^2_{\mathbb{F}_{k+1}}}{m}.$$

For Fourier-based analysis of two-layer nets for more general activation functions, we refer to (Siegel and Xu, 2020).

Why do 2LNNs perform better than RFM/kernel methods?

The Perspective of Adaptive Kernel Methods

- Let $\pi(w) = \int \rho(a, w)\,\mathrm{d}a$ and $a(w) = \int a\rho(a|w)\,\mathrm{d}a$. Then,

$$\int a\varphi(x, w)\,\mathrm{d}\rho(a, w) = \int \left(\int a\varphi(x, w) \frac{\rho(a, w)}{\pi(w)}\,\mathrm{d}a \right) \pi(w)\,\mathrm{d}w = \int a(w)\varphi(x, w)\,\mathrm{d}\pi(w).$$

- Let $k_\pi = \mathbb{E}_{w\in\pi}[\varphi(x, w)\varphi(x', w)]$ be the RF kernel.

The Perspective of Adaptive Kernel Methods

- Let $\pi(w) = \int \rho(a, w)\,\mathrm{d}a$ and $a(w) = \int a\rho(a|w)\,\mathrm{d}a$. Then,

$$\int a\varphi(x, w)\,\mathrm{d}\rho(a, w) = \int \left(\int a\varphi(x, w) \frac{\rho(a, w)}{\pi(w)}\,\mathrm{d}a \right) \pi(w)\,\mathrm{d}w = \int a(w)\varphi(x, w)\,\mathrm{d}\pi(w).$$

- Let $k_\pi = \mathbb{E}_{w\in\pi}[\varphi(x, w)\varphi(x', w)]$ be the RF kernel.

Lemma 15

$$\mathcal{B} = \cup_{\pi\in\mathcal{P}(\Omega)}\mathcal{H}_{k_\pi} \quad \|f\|_{\mathcal{B}} = \inf_{\pi\in\mathcal{P}(\Omega)} \|f\|_{\mathcal{H}_{k_\pi}}.$$

- This lemma shows that 2LNN can be viewed as adaptive RF/kernel method.
- The adaptivity plays the role of **variance reduction**:

$$\|f\|_{\mathcal{B}}^2 = \inf_{f=f_{a,\pi}} \mathbb{E}_{w\sim\pi}[a^2(w)].$$

This means that 2LNN takes the "feature" $\pi \in \mathcal{P}(\Omega)$ such that the representation coefficients $a \in L^2(\pi)$ to have the smallest variance.

Proof of Lemma 15

By the theory of random feature models,

$$\|f\|_{\mathcal{H}_{k_\pi}}^2 = \inf_{f=\int a(w)\varphi(\cdot,w)\,\mathrm{d}\pi(w)} \mathbb{E}[a(w)^2].$$

Then,

$$\begin{aligned}\|f\|_{\mathcal{B}}^2 &= \inf_{f=\mathbb{E}_{(a,w)\sim\rho}[a\varphi(\cdot,w)]} \mathbb{E}[a^2] = \inf_{\pi\in\mathcal{P}(\Omega)} \inf_{f=\mathbb{E}_{w\sim\pi}[a(w)\varphi(\cdot,w)]} \mathbb{E}[a(w)^2] \\ &= \inf_{\pi\in\mathcal{P}(\Omega)} \|f\|_{\mathcal{H}_{k_\pi}}^2,\end{aligned}$$

which implies that $\mathcal{B} = \cup_{\pi\in P(\Omega)}\mathcal{H}_{k_\pi}$.

A Separation Result

Theorem 16 (Modifying from Barron (1993), Theorem 6)

Let $\mathcal{B}$ be the Barron space associated with $\varphi(x, \tilde{w}) = \sigma_{\mathrm{ReLU}}(w^\top x + b)$, $\mathcal{X} = [0, 2\pi]^d$, and $h_1, h_2, \ldots, h_m$ be m arbitrary fixed functions. Then,

$$\sup_{\|f\|_{\mathcal{B}} \leq 1} \inf_{h \in \mathrm{span}\{h_1, \ldots, h_m\}} \|h - f\|_{L^2(\mathbb{P}_x)} \gtrsim \frac{1}{d^2 m^{2/d}}.$$

A Separation Result

Theorem 16 (Modifying from Barron (1993), Theorem 6)

Let $\mathcal{B}$ be the Barron space associated with $\varphi(x, \tilde{w}) = \sigma_{\mathrm{ReLU}}(w^\top x + b)$, $\mathcal{X} = [0, 2\pi]^d$, and $h_1, h_2, \ldots, h_m$ be m arbitrary fixed functions. Then,

$$\sup_{\|f\|_{\mathcal{B}} \leq 1} \inf_{h \in \mathrm{span}\{h_1, \ldots, h_m\}} \|h - f\|_{L^2(\mathbb{P}_x)} \gtrsim \frac{1}{d^2 m^{2/d}}.$$

- Any linear methods, including RFMs, suffer from CoD in learning Barron functions.

A Separation Result

Theorem 16 (Modifying from Barron (1993), Theorem 6)

Let $\mathcal{B}$ be the Barron space associated with $\varphi(x, \tilde{w}) = \sigma_{\mathrm{ReLU}}(w^\top x + b)$, $\mathcal{X} = [0, 2\pi]^d$, and $h_1, h_2, \ldots, h_m$ be m arbitrary fixed functions. Then,

$$\sup_{\|f\|_{\mathcal{B}} \leq 1} \inf_{h \in \mathrm{span}\{h_1, \ldots, h_m\}} \|h - f\|_{L^2(\mathbb{P}_x)} \gtrsim \frac{1}{d^2 m^{2/d}}.$$

- Any linear methods, including RFMs, suffer from CoD in learning Barron functions.
- 2LNNs can learn the Barron space without CoD.

A Separation Result

Theorem 16 (Modifying from Barron (1993), Theorem 6)

Let $\mathcal{B}$ be the Barron space associated with $\varphi(x, \tilde{w}) = \sigma_{\mathrm{ReLU}}(w^\top x + b)$, $\mathcal{X} = [0, 2\pi]^d$, and $h_1, h_2, \ldots, h_m$ be m arbitrary fixed functions. Then,

$$\sup_{\|f\|_{\mathcal{B}} \leq 1} \inf_{h \in \mathrm{span}\{h_1, \ldots, h_m\}} \|h - f\|_{L^2(\mathbb{P}_x)} \gtrsim \frac{1}{d^2 m^{2/d}}.$$

- Any linear methods, including RFMs, suffer from CoD in learning Barron functions.
- 2LNNs can learn the Barron space without CoD.
- The above theorem is not only an approximation but also generalization lower bound. For KRR, the hypothesis is $\hat{f} = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i)$, for which $m = n$. Thus, the above theorem yield a lower bound of sample size for learning the Barron space with kernel methods.

A Separation Result

Theorem 16 (Modifying from Barron (1993), Theorem 6)

Let $\mathcal{B}$ be the Barron space associated with $\varphi(x, \tilde{w}) = \sigma_{\mathrm{ReLU}}(w^\top x + b)$, $\mathcal{X} = [0, 2\pi]^d$, and $h_1, h_2, \ldots, h_m$ be m arbitrary fixed functions. Then,

$$\sup_{\|f\|_{\mathcal{B}} \leq 1} \inf_{h \in \mathrm{span}\{h_1, \ldots, h_m\}} \|h - f\|_{L^2(\mathbb{P}_x)} \gtrsim \frac{1}{d^2 m^{2/d}}.$$

- Any linear methods, including RFMs, suffer from CoD in learning Barron functions.
- 2LNNs can learn the Barron space without CoD.
- The above theorem is not only an approximation but also generalization lower bound. For KRR, the hypothesis is $\hat{f} = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$, for which $m = n$. Thus, the above theorem yield a lower bound of sample size for learning the Barron space with kernel methods.
- In Barron (1993), an analogous lower bound is proved for the Barron space associated with the Heaviside activation function (consequently all sigmoidal activation functions).

A Lower Bound of Approximating Orthogonal Functions

Lemma 17

Suppose $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ *to be a Hilbert space. Let* $\{g_1, \ldots, g_{2m}\}$ *be* m *orthonormal functions in* $\mathcal{H}$*. For any linear subspace* V_m *with* $\dim(V_m) = m$*, we have*

$$\sup_{j \in [2m]} d^2(g_j, V_m) \geq \frac{1}{2},$$

where $d^2(g, V_m) := \inf_{f \in V_m} \|g - f\|_{\mathcal{H}}^2$.

A Lower Bound of Approximating Orthogonal Functions

Lemma 17

Suppose $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ to be a Hilbert space. Let $\{g_1, \ldots, g_{2m}\}$ be m orthonormal functions in $\mathcal{H}$. For any linear subspace V_m with $\dim(V_m) = m$, we have

$$\sup_{j\in[2m]} d^2(g_j, V_m) \geq \frac{1}{2},$$

where $d^2(g, V_m) := \inf_{f\in V_m} \|g - f\|_{\mathcal{H}}^2$.

Proof: WLOG, let $e_1, \ldots, e_m$ be an orthonormal basis of V_m. Then, for any $\|g\|_{\mathcal{H}} = 1$, $d(g, V_m) = 1 - \sum_{i=1}^m \langle g, e_i\rangle_{\mathcal{H}}^2$.

$$\begin{aligned}
\sup_{j\in[2m]} d^2(g_j, V_m) &\geq \frac{1}{2m}\sum_{j=1}^{2m} d^2(g_j, V_m) = 1 - \frac{1}{2m}\sum_{j=1}^{2m}\sum_{i=1}^{m}\langle g_j, e_i\rangle_{\mathcal{H}}^2 \\
&= 1 - \frac{1}{2m}\sum_{i=1}^{m}\sum_{j=1}^{2m}\langle g_j, e_i\rangle_{\mathcal{H}}^2 \\
&\geq 1 - \frac{1}{2m}\sum_{i=1}^{m}\|e_i\|_{\mathcal{H}}^2 = 1 - \frac{1}{2} = \frac{1}{2}
\end{aligned}$$

The Barron Space Contains Exponential Orthgonal Functions

Lemma 18

Let $\mathcal{X} = [0, 2\pi]^d$ and consider the ReLU activation function. Let $\mathcal{B}(s) = \{f \in \mathcal{B} : \|f\|_{\mathcal{B}} \leq s^2 d^2\}$. Then, $\mathcal{B}(s)$ contains at least $(1+s)^d$ orthonormal functions.

The Barron Space Contains Exponential Orthgonal Functions

Lemma 18

Let $\mathcal{X} = [0, 2\pi]^d$ and consider the ReLU activation function. Let $\mathcal{B}(s) = \{f \in \mathcal{B} : \|f\|_{\mathcal{B}} \leq s^2 d^2\}$. Then, $\mathcal{B}(s)$ contains at least $(1 + s)^d$ orthonormal functions.

- This lemma implies that the Barron space contain exponentially many orthnormal functions as long as the norm is polynomially large.
- This is a very crucial fact about the Barron space and 2LNNs. We will use this property later to establish the hardness of training.

Proof of Lemma 18

Define a set of functions:

$$\mathcal{G}_M = \left\{ x \mapsto \cos(b^\top x) \,:\, \sum_{i=1}^{n} b_i \leq M, b_i \in \mathbb{N} \right\}.$$

Proof of Lemma 18

Define a set of functions:

$$\mathcal{G}_M = \left\{ x \mapsto \cos(b^\top x) \,:\, \sum_{i=1}^{n} b_i \le M, b_i \in \mathbb{N} \right\}.$$

- For any $g, g' \in \mathcal{G}_M$ with $g \neq g'$, we have $\langle g, g' \rangle_{L^2(\mathbb{P}_x)} = 0$. Moreover, notice that $\widehat{\cos(b^\top \cdot)}(\omega) = \frac{1}{2}(\delta(\omega - b) + \delta(\omega + b))$. Then, for any $g \in \mathcal{G}_M$,

$$\|g\|_{\mathcal{B}} \le \|g\|_{\mathbb{F}_2} = \int_{\mathbb{R}} \|\omega\|_1^2 |\hat{q}(\omega)| \, \mathrm{d}\omega \lesssim \|b\|_1^2 \le M^2.$$

Proof of Lemma 18

Define a set of functions:

$$\mathcal{G}_M = \left\{ x \mapsto \cos(b^\top x) \,:\, \sum_{i=1}^{n} b_i \le M, b_i \in \mathbb{N} \right\}.$$

- For any $g, g' \in \mathcal{G}_M$ with $g \neq g'$, we have $\langle g, g' \rangle_{L^2(\mathbb{P}_x)} = 0$. Moreover, notice that $\widehat{\cos(b^\top \cdot)}(\omega) = \frac{1}{2}(\delta(\omega - b) + \delta(\omega + b))$. Then, for any $g \in \mathcal{G}_M$,

$$\|g\|_{\mathcal{B}} \le \|g\|_{\mathbb{F}_2} = \int_{\mathbb{R}} \|\omega\|_1^2 |\hat{q}(\omega)| \, \mathrm{d}\omega \lesssim \|b\|_1^2 \le M^2.$$

- Let $M = sd$. Then,

$$|\mathcal{G}_M| = \binom{M+d}{d} \ge \left(\frac{M+d}{d} \right)^d = (1+s)^d. \tag{10}$$

Proof of Theorem 16

Choose $\bar{M}$ to be the smallest M such that $|\mathcal{G}_M| \geq 2m$. For any $\dim(V_m) = m$,

$$\sup_{\|f\|_{\mathcal{B}} \leq 1} d(f, V_m) \gtrsim \sup_{f \in \frac{1}{\bar{M}^2}\mathcal{G}_{\bar{M}}} d(f, V_m) \geq \frac{1}{\bar{M}^2} \sup_{f \in \mathcal{G}_{\bar{M}}} d(f, V_m) \gtrsim \frac{1}{\bar{M}^2}, \tag{11}$$

Let $\bar{M} = sd$. Let $(1+s)^d \sim 2m$. Then, $s \sim m^{1/d}$. Plugging it into (11), we have

$$\sup_{\|f\|_{\mathcal{B}} \leq 1} d(f, V_m) \gtrsim \frac{1}{s^2 d^2} \gtrsim \frac{1}{d^2 m^{2/d}}.$$

A Explicit Hard Example: A Single Neuron

A single neuron is given by $\sigma_v(x) := \sigma(v^\top x)$ with $\|v\|_1 = 1$.

- σ_v is hard to approximate by using RFM:

$$f_m(x; a) = \frac{1}{m} \sum_j a_j \sigma(w_j^\top x),$$

where $w_j \sim \pi_0$, and $\pi_0 = \mathrm{Unif}(\mathbb{S}^{d-1})$. We can write

$$\sigma_v(x) = \int a(w)\sigma(w^\top x) d\pi_0(w),$$

where $a(w) = \delta(w - v)$. Obviously, $\|f\|^2_{\mathcal{H}_{k_{\pi_0}}} = \mathbb{E}[a(w)^2] = \infty$.

- $\|\sigma_v\|_{\mathcal{B}} \lesssim 1$ since $\sigma_v(x) = \int \sigma(w^\top x)\, \mathrm{d}\pi(w)$ with $\pi(w) = \delta(w - v)$. This gives an example how 2LNNs perform "variance reduction"!
- One can prove that approximating σ_v with RF requires $m \geq \exp(d)$; See [Yehudai and Shamir, 2019] and [Wu and Long, 2022].