Theoretical Deep Learning

Lecture 7: The Barron space

August 2, 2021

Lecturer: Lei Wu *Scribe: Lei Wu*

Consider the two-layer ReLU nets. The Fourier analysis has shown that a sufficient condition for a function can be efficiently approximated is

$$\|f\|_{\mathbb{F}_2} = \inf_{f_e|_\Omega = f} \int_{\mathbb{R}^d} (1+\|\omega\|)^2 |\hat{f}_e(\omega)| \,\mathrm{d}\omega < \infty \tag{0.1}$$

We ask the question: Is the spectral Barron norm (0.1) also tight in characterize the "efficient" approximation characteristic of two-layer neural nets? Unfortunately, it is not. A counter example is given by the triangular function

Lemma 0.1. *Let* $f : [-2,2] \mapsto \mathbb{R}$ *be given by* $f(x) = \max(1-|x|, 1)$. *Then,* $\|f\|_{\mathbb{F}_2} = \infty$ *and* $f(x) = \sigma(x+1) + \sigma(x-1) - 2\sigma(x)$.

Proof. Let f_e be the zero extension of f, which is the triangular function in the whole space. Its Fourier transform is

$$\hat{f}_e(\omega) = \frac{\sin^2(\omega)}{\omega^2},$$

which leads to

$$\int_{\mathbb{R}} |\omega|^2 |\hat{f}_e(\omega)| \,\mathrm{d}\omega = \int_{\mathbb{R}} \sin^2(\omega) \,\mathrm{d}\omega = \infty.$$

Then, we still need to show that over all the extension, we still have $\int_{\mathbb{R}} |\omega|^2 |\hat{f}_e(\omega)| \,\mathrm{d}\omega = \infty$. We omit this part for simplicity. □

1 The Barron space

The previous study motivate us to consider all the functions that admit the following integral representation:

$$f_\pi(x) = \mathbb{E}_{(a,w)\sim\pi}[a\sigma(w^T x)]. \tag{1.1}$$

This can be viewed as an infinitely-wide two-layer net. It is the continuum limit of the *scaled* two-layer neural net:

$$f_m(x;\theta) = \frac{1}{m}\sum_{j=1}^m a_j \sigma(w_j^T x). \tag{1.2}$$

For any f that admit the representation (1.1), the representation π is usually not not unique. Define

$$R_f = \left\{ \pi \in \mathcal{P}(\mathbb{R}^1 \otimes \mathbb{R}^d) \,:\, f_\pi(x) = \mathbb{E}_{(a,w)\sim\pi}[a\sigma(w^T x)] \right\}. \tag{1.3}$$

Definition 1.1 (The Barron space)**.** Assume that σ is ReLU. Let

$$\|f\|_{\mathcal{B}_p}^p := \inf_{\pi \in R_f} \mathbb{E}_{(a,w)\sim\pi}[|a|^p \|w\|^p]. \tag{1.4}$$

The Barron space $\mathcal{B}_p := \{f : \|f\|_{\mathcal{B}_p} < \infty\}$.

- For a function f, one can think of π as the representation. Hence, the proceeding definition means that we use the moments of π to quantify the complexity of f_π.
- The specific moment (1.4) is scaling invariant, which is consistent with the fact that f_π is also scaling invariant. Note that the specific forms of "moment" may be different for different activation functions.
- The taking-infimum step in (1.4) is essential. First, it make the function norm well-defined in the sense that $\|\cdot\|_{\mathcal{B}_p}$ is independent of the choice of representations. Second, it means that the complexity of f is measured by choosing the best representation π (**adaptivity**). For instance, a single neuron, we can have two representations:

  $$\sigma(x_1) = \sigma(x_1) + r\sigma(x_2) - r\sigma(x_2). \tag{1.5}$$

  The according distributions π's are given by

  $$\pi_1(a, w) = \delta(a-1)\delta(w - e_1)$$
  $$\pi_2(a, w) = \delta(a-1)\delta(w - e_1) + r\delta(a-1)\delta(w - e_2) + r\delta(a+1)\delta(w - e_2),$$

  respectively. For the former, the moment is 1; for the latter, the moment is $(1 + 2r^p)^{1/p}$. The latter can be much larger than the former. This justifies why we must take the infimum. As shown latter, it is also the key to separate neural nets and random feature models.

The following lemma actually show that $\mathcal{B}_q$ are the same for any $q = [1, \infty]$. Hence, we will simplify use $\mathcal{B}$ to denote the Barron space.

Lemma 1.2. *For any $p, q \geq 1$, we have $\|f\|_{\mathcal{B}_p} = \|f\|_{\mathcal{B}_q}$.*

Proof. By Holder's inequality, obviously $\|f\|_{\mathcal{B}_1} \leq \|f\|_{\mathcal{B}_\infty}$. To complete the proof, we only need to show that $\|f\|_{\mathcal{B}_\infty} \leq \|f\|_{\mathcal{B}_1}$ also holds.

- Assume $f \in \mathcal{B}_1$. For any $\varepsilon > 0$, there exists a ρ such that $f(x) = \mathbb{E}_{(a,w)\sim\rho}[a\sigma(w^T x)]$ and $\mathbb{E}_\rho[|a|\|w\|] < \|f\|_{\mathcal{B}_1} + \varepsilon$. For any w, let $\hat{w} = w/\|w\|$. Then,

  $$\begin{aligned} f(x) &= \int a\sigma(w^T x)\,\mathrm{d}\rho(a, w) = \int_{a>0} a\sigma(w^T x)d\rho(a, w) - \int_{a<0} (-a)\sigma(w^T x)d\rho(a, w) \\ &= \int_{a>0} a\|w\|\sigma(\hat{w}^T x)\,\mathrm{d}\rho(a, w) - \int_{a<0} (-a)\|w\|\sigma(\hat{w}^T x)\,\mathrm{d}\rho(a, w) \\ &= \int_{\mathbb{S}^{d-1}} \sigma(w^T x)\,\mathrm{d}\rho_+(w) - \int_{\mathbb{S}^{d-1}} \sigma(w^T x)\,\mathrm{d}\rho_-(w). \end{aligned}$$

  Here ρ_+, ρ_- are defined as follows. For any Borel set $A \subset \mathbb{S}^{d-1}$, define

  $$\rho_+(A) = \int_{\{(a,w):\hat{w}\in A, a>0\}} |a|\|w\|\,\mathrm{d}\rho(a, w),$$
  $$\rho_-(A) = \int_{\{(a,w):\hat{w}\in A, a<0\}} |a|\|w\|\,\mathrm{d}\rho(a, w),$$

- Let $M_+ = \rho_+(\mathbb{S}^{d-1}), M_- = \rho_-(\mathbb{S}^{d-1})$. Obviously $M = M_+ + M_- = \mathbb{E}_\rho[|a|\|w\|]$, and

$$\begin{aligned} f(x) &= \int_{\mathbb{S}^{d-1}} \sigma(w^T x)\,\mathrm{d}\rho_+(w) - \int_{\mathbb{S}^{d-1}} \sigma(w^T x)\,\mathrm{d}\rho_-(w) \\ &= \int_{\mathbb{S}^{d-1}} M\sigma(w^T x)\frac{\rho_+(w)}{M}\,\mathrm{d}w + \int_{\mathbb{S}^{d-1}} (-M)\sigma(w^T x)\frac{\rho_-(w)}{M}\,\mathrm{d}w = \mathbb{E}_{(a,w)\sim\tilde{\rho}}[a\sigma(w^T x)], \end{aligned}$$

  where $\tilde{\rho}(a,w) = \delta(a-M)\frac{\rho_+(w)}{M} + \delta(a+M)\frac{\rho_-(w)}{M}$.

- Hence, $\|f\|_{\mathcal{B}_\infty} \le M \le \|f\|_{\mathcal{B}_1} + \varepsilon$. Taking $\varepsilon \to 0$, we complete the proof.

□

Examples of Barron functions.

- We have shown that $\|f\|_{\mathcal{B}} \lesssim \|f\|_{\mathbb{F}_2}$. This contains a lot of functions.
- Finite-width neural nets: $f_m(x) = \sum_{j=1}^m a_j\sigma(b_j^T x)$. Obviously,

$$\|f_m\|_{\mathcal{B}} \le \sum_{j=1}^m |a_j|\|\boldsymbol{b}_j\|,$$

  where the right hand side is usually called the *path norm*, which can be used to regularize neural nets.

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

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

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

2 Approximation theorems

For a two-layer neural network $f_m(\cdot;\theta)$, define the path norm

$$\|\theta\|_{\mathcal{P}} := \frac{1}{m}\sum_{j=1}^m |a_j|\|b_j\|. \tag{2.1}$$

The path norm is a discrete analog of the $\mathcal{B}_1$ norm. It is very useful in analyzing two-layer neural networks.

Theorem 2.1 (Direct Approximation Theorem, L^2-version)**.** *For any $f \in \mathcal{B}$ and $m \in \mathbb{N}$, there exists a two-layer neural network $f_m(\cdot;\theta)$ such that*

$$\begin{gathered} \|f - f_m(\cdot;\theta)\|_{L^2(\rho)}^2 \lesssim \frac{\|f\|_{\mathcal{B}}^2}{m} \\ \|\theta\|_{\mathcal{P}} \le 2\|f\|_{\mathcal{B}}. \end{gathered}$$

Proof. For $f \in \mathcal{B}$, there exists a ρ such that $f(x) = \mathbb{E}_\pi[a\sigma(w \cdot x)]$ and $\mathbb{E}[a^2\|w\|^2] \leq 2\|f\|_{\mathcal{B}}^2$. Consider $\{(a_j, w_j)\}_j$ i.i.d. drawn from ρ. Then,

$$\begin{aligned}
\mathbb{E}_{(a_j,w_j)} \mathbb{E}_x |\frac{1}{m}\sum_{j=1}^m a_j\sigma(w_j \cdot x) - f(x)|^2 &= \mathbb{E}_x \mathbb{E}_{(a_j,w_j)} |\frac{1}{m}\sum_{j=1}^m a_j\sigma(w_j \cdot x) - f(x)|^2 \\
&= \mathbb{E}_x \frac{1}{m}\sum_{j=1}^m \mathbb{E}_{(a_j,w_j)} |a_j\sigma(w_j \cdot x) - f(x)|^2 \qquad \text{(Use the independence of } (a_j, w_j)\text{)} \\
&\leq \mathbb{E}_x \frac{1}{m}\sum_{j=1}^m \mathbb{E}_{(a_j,w_j)} a_j^2\sigma(w_j \cdot x)^2 \leq \frac{1}{m}\sum_{j=1}^m \mathbb{E}_{(a_j,w_j)} a_j^2\|w_j\|^2 \\
&\leq \frac{2\|f\|_{\mathcal{B}}^2}{m}.
\end{aligned}$$

Then, there must exist $\{(a_j, w_j)\}$ such that the theorem holds. □

Note that the control of path norm for the approximator is important for our later analysis of the generalization performance.

Theorem 2.2 (Inverse approximation)**.** *Let $f^* \in C(\Omega)$. If there exists a constant Q and a sequence of two-layer neural networks $f_m(\cdot; \theta^{(m)})$ with the path norm uniformly bounded, i.e., $\|\theta^{(m)}\|_{\mathcal{P}} \leq Q$ such that*

$$f_m(x; \theta^{(m)}) \to f^*(x), \quad \forall x \in \Omega.$$

Then, there exists $\rho^* \in \mathcal{P}(\mathbb{R} \otimes \mathbb{R}^d)$ such that

$$f^*(x) = \int a\sigma(w^T x)\, \mathrm{d}\rho^*(a, w), \quad \forall\, x \in \Omega.$$

Moreover, $f^* \in \mathcal{B}$ and $\|f^*\|_{\mathcal{B}} \leq Q$.

Inverse approximation theorem implies that the Barron space is sufficiently large, since the functions of interest lie in it.

Proof. Let $\theta^{(m)} = \{(a_k^{(m)}, w_k^{(m)})\}_{k=1}^m$. WLOG, assume $\|w_k^{(m)}\| = 1$. Let $A_m = \sum_{k=1}^m |a_k^{(m)}|$. Define the "weighted" empirical measure

$$\rho_m(a, w) = \sum_{k=1}^m \frac{|a_k^{(m)}|}{A_m} \delta\left(a - \text{sign}(a_k^{(m)})A_m\right) \delta(w - w_k^{(m)}).$$

- It is easy to verify that

$$\begin{aligned}
\mathbb{E}_{(a,w)\sim\rho_m}[a\sigma(w^T x)] = \sum_{k=1}^m \frac{|a_k^{(m)}|}{A_m}\text{sign}(a_k^{(m)})A_m\sigma(w_k^{(m)} \cdot x) = \sum_{k=1}^m a_k\sigma(w_k^{(m)} \cdot x) = f(x; \theta^{(m)}) \\
\text{supp}(\rho_m) \subset K_Q = \{(a, b) : |a| \leq Q, \|w\| \leq 1\}.
\end{aligned}$$

- Since K_Q is compact, (ρ_m) is tight. By Prokhorov's Theorem, there exists a subsequence (ρ_{m_k}) and ρ^* such that ρ_{m_k} converges to ρ^* weakly.

- Since $h(a,w) = a\sigma(w^T x), g(a,w) = |a|\|w\|$ are continuous with respect (a,w), we have

$$f^*(x) = \lim_{k\to\infty} \int a\sigma(w^T x)\,\mathrm{d}\rho_{m_k}(a,w) = \int a\sigma(w^T x)\,\mathrm{d}\rho^*(a,w)$$
$$\|f^*\|_{\mathcal{B}} \le \int |a|\|w\| d\rho^*(a,w) = \lim_{k\to\infty} \int |a|\|w\|\,\mathrm{d}\rho_{m_k}(a,w) \le Q.$$

□

Note that the naive choice of the empirical measure sequence is $(\tilde{\rho}_m)$ with

$$\tilde{\rho}_m(a,b) = \frac{1}{m}\sum_{j=1}^{m} \delta(a - a_k^{(m)})\delta(b - b_k^{(m)}).$$

However, $(\tilde{\rho}_m)$ may be not tight since $a_k^{(m)} = O(mQ)$ in the worst case.

3 Generalization analysis

Proposition 3.1. *Let $\mathcal{F}_Q = \{f \in \mathcal{B} : \|f\|_{\mathcal{B}} \le Q\}$. Then,*

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

C_d is a constant depending on the domain Ω. Assume Ω is a ℓ_p ball. If Ω is the L^∞ ball, then $C_d = \log d$, otherwise $C_d = 1$.

Proof. Let $\xi = (\xi_1, \ldots, \xi_n)$. By definition, we have

$$\begin{aligned}
n\widehat{\mathrm{Rad}}_n(\mathcal{F}_Q) &= \mathbb{E}_\xi[\sup_{f\in\mathcal{F}_Q} \sum_{i=1}^n \xi_i\, \mathbb{E}_\rho[a\sigma(w^T x_i)]] = \mathbb{E}_{\boldsymbol{\xi}}[\sup_{f\in\mathcal{F}_Q} \mathbb{E}_\rho[|a|\|w\| \sum_{i=1}^n \xi_i\sigma(\hat{w}^T x_i)]] \\
&\le \mathbb{E}_\xi[\sup_{f\in\mathcal{F}_Q} \mathbb{E}_\rho[|a|\|w\| \sup_{\|w\|\le 1} |\sum_{i=1}^n \xi_i\sigma(w^T x_i)|] \\
&\le Q\,\mathbb{E}_\xi[\sup_{\|w\|\le 1} |\sum_{i=1}^n \xi_i\sigma(w^T x_i)|] \\
&\le Q\,\mathbb{E}_\xi[\sup_{\|w\|\le 1} \sum_{i=1}^n \xi_i\sigma(w^T x_i)] + Q\,\mathbb{E}_\xi[\sup_{\|w\|\le 1} -\sum_{i=1}^n \xi_i\sigma(w^T x_i)] \\
&= 2Q\,\mathbb{E}_\xi[\sup_{\|w\|\le 1} \sum_{i=1}^n \xi_i\sigma(w^T x_i)] \qquad \text{(Use the symmetry of } \xi\text{)} \\
&\le 2Q\,\mathbb{E}_\xi[\sup_{\|w\|\le 1} \sum_{i=1}^n \xi_i w^T x_i] \qquad \text{(Use the contraction lemma).}
\end{aligned}$$

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

The regularized estimator. Consider the path norm-regularized estimator:

$$\hat{\theta}_n = \operatorname*{argmin}_{\theta} \hat{R}_n(\theta) + \frac{\lambda}{\sqrt{n}}\|\theta\|_{\mathcal{P}}. \tag{3.1}$$

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

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

Theorem 3.2. *Assume $\lambda \ge C$, where C is an absolute constant. For any $\delta \in (0,1)$, with probability $1-\delta$ over the choice of training samples, we have*

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

- The three terms of the RHS denote the approximation error, estimation error, and error caused by the exception set, respectively.
- The estimate does not suffer from the curse of dimensionality (CoD), and works well in the over-parameterized regime, i.e., $m > n$.

Proof. Let $Q = \|f^*\|_{\mathcal{B}}$.

(1) By the direct approximation theorem, there exits $\tilde{\theta}$ such that

$$\hat{R}_n(\tilde{\theta}) \le \frac{3Q^2}{m}, \qquad \|\tilde{\theta}\|_{\mathcal{P}} \le 2Q.$$

By definition,
$$\hat{R}_n(\hat{\theta}_n) + \frac{\lambda}{\sqrt{n}}\|\hat{\theta}_n\|_{\mathcal{P}} \le \hat{R}_n(\tilde{\theta}) + \frac{\lambda}{\sqrt{n}}\|\tilde{\theta}\|_{\mathcal{P}} \le \frac{3Q^2}{m} + 2\frac{\lambda}{\sqrt{n}}Q.$$

Hence,

$$\|\hat{\theta}_n\|_{\mathcal{P}} \le 2Q + \frac{3Q^2\sqrt{n}}{\lambda m} =: C(m,\lambda,Q)$$
$$\hat{R}_n(\hat{\theta}_n) \le \frac{3Q^2}{m} + \frac{2\lambda}{\sqrt{n}}Q. \tag{3.2}$$

(2) Let $\mathcal{H}_C = \{(\tilde{f}_m(x;\theta) - f^*(x))^2 : \|\theta\|_{\mathcal{P}} \le C\}$. Since t^2 is 2-Lipschitz continuous for $t \in [-1,1]$. By the contraction lemma,

$$\widehat{\mathrm{Rad}}_n(\mathcal{H}_C) \le 2\widehat{\mathrm{Rad}}_n(\mathcal{F}_C). \tag{3.3}$$

By (3.2), $\hat{f}_m(\cdot;\hat{\theta}_n) \in \mathcal{F}_{C(m,\lambda,Q)}$.

(3) Using the Rademacher complexity-based generalization bound, we have

$$\begin{aligned}\mathcal{R}(\hat{\theta}_n) &\le \hat{\mathcal{R}}_n(\hat{\theta}_n) + 2\widehat{\mathrm{Rad}}_n(\mathcal{H}_{C(m,\lambda,Q)}) + \sqrt{\frac{\log(2/\delta)}{n}} \\ &\le \hat{\mathcal{R}}_n(\hat{\theta}_n) + 4\widehat{\mathrm{Rad}}_n(\mathcal{F}_{C(m,\lambda,Q)}) + \sqrt{\frac{\log(2/\delta)}{n}} \qquad \text{(Use Eq.(3.3))}\end{aligned}$$

$$
\begin{aligned}
&\lesssim \hat{\mathcal{R}}_n(\hat{\theta}_n) + \frac{C(m,\lambda,Q)}{\sqrt{n}} + \sqrt{\frac{\log(2/\delta)}{n}} && \text{(Use Prop.3.1 and Eq.(3.2))} \\
&\leq \frac{3Q^2}{m} + \frac{2\lambda}{\sqrt{n}}Q + \frac{1}{\sqrt{n}}\left(2Q + \frac{3Q^2\sqrt{n}}{\lambda m}\right) + \sqrt{\frac{\log(2/\delta)}{n}} && \text{(Use Eq.(3.2))} \\
&\lesssim \frac{Q^2}{m} + \frac{Q}{\sqrt{n}} + \sqrt{\frac{\log(2/\delta)}{n}}.
\end{aligned}
$$

□

Remark 3.3.

- The generalization error analysis can be extended to the noisy case, i.e., $y_i = f^*(x_i) + \varepsilon_i$. As long as ε_i is sub-Gaussian, we can use the truncation method. But this will introduce a $\log(n)$ factor. See Theorem 4.2 of [E et al., 2019].
- Currently, we focus on the ReLU activation function. The theory of Barron space can be extended to more general activation functions (see Section 3&4 in [Li et al., 2020])

4 Connection with kernel methods

The Barron space is closely related to a family of RKHSs. WLOG, assume $w \in \mathbb{S}^{d-1}$. Then,

$$
f(x) = \int_{\mathbb{R}\times\mathbb{S}^{d-1}} a\sigma(w^T x)\,\mathrm{d}\rho(a,w) = \int_{\mathbb{S}^{d-1}} a(w)\sigma(w^T x)\,\mathrm{d}\pi(w), \tag{4.1}
$$

where

$$
\pi(w) = \int_{\mathbb{R}} \rho(a,w)\,\mathrm{d}a, \quad a(w) = \frac{\int_{\mathbb{R}} a\rho(a,w)\,\mathrm{d}a}{\pi(w)}.
$$

In this form, we have

$$
\|f\|_{\mathcal{B}_2}^2 = \inf_{\pi\in P(\mathbb{S}^{d-1})} \inf_{f(x)=\mathbb{E}_\pi[a(w)\sigma(w^T x)]} \mathbb{E}[a(w)^2].
$$

Given a fixed π, we can define a kernel:

$$
k_\pi(x,x') = \mathbb{E}_{w\sim\pi}[\sigma(w^T x)\sigma(w^T x')]
$$

Let $\mathcal{H}_{k_\pi}$ denote the RKHS induced by k_π.

Proposition 4.1.

$$
\mathcal{B} = \bigcup_{\pi\in\mathcal{P}(\mathbb{S}^{d-1})} \mathcal{H}_{k_\pi}, \quad \|f\|_{\mathcal{B}} = \inf_{\pi\in\mathcal{P}(\mathbb{S}^{d-1})} \|f\|_{\mathcal{H}_{k_\pi}}.
$$

Proof. By the theory of random feature models,

$$
\|f\|_{\mathcal{H}_{k_\pi}}^2 = \inf_{f(x)=\mathbb{E}_{w\sim\pi}[a(w)\sigma(w^T x)]} \mathbb{E}[a(w)^2].
$$

Then,

$$
\begin{aligned}
\|f\|_{\mathcal{B}_2}^2 &= \inf_{f(x)=\mathbb{E}_\rho[a\sigma(w^T x)]} \mathbb{E}[a^2\|w\|^2] = \inf_{\pi\in\mathcal{P}(\mathbb{S}^{d-1})} \inf_{f(x)=\mathbb{E}_\pi[a(w)\sigma(w^T x)]} \mathbb{E}[a(w)^2] \\
&= \inf_{\pi\in\mathbb{P}(\mathbb{S}^{d-1})} \|f\|_{\mathcal{H}_{k_\pi}}^2,
\end{aligned}
$$

which implies that $\mathcal{B} = \cup_{\pi\in P(\mathbb{S}^{d-1})}\mathcal{H}_{k_\pi}$. □

This proposition means that two-layer neural networks can be viewed adaptive kernel method with the kernel adaptively chosen from the kernel family: $K = \{k_\pi : \pi \in \mathcal{P}(\mathbb{S}^{d-1})\}$.

4.1 Separation results

Theorem 4.2 (Modifying from [Barron, 1993], Theorem 6)**.** *Let $\Omega = [0, 2\pi]^d$ and $h_1, h_2, \ldots, h_n$ be n arbitrary fixed functions. Then,*

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

It states that any linear methods, including the random feature models, suffer from CoD in learning functions in Barron space. On the contrary, two-layer neural networks can learn functions in Barron space without CoD. The comparison between the proceeding lower bound and Theorem 3.2 provides a clear separation between two methods.

Before proceeding to the proof, we first need the following lemma.

Lemma 4.3. *Suppose $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ to be a Hilbert space. Let $\{g_1, \ldots, g_{2n}\}$ be $2n$ orthonormal functions in $\mathcal{H}$. For any linear subspace V_n with $\dim(V_n) = n$, we have*

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

where $d^2(g, V_n) := \inf_{c_1, \ldots, c_n \in \mathbb{R}} \|g - \sum_{i=1}^n c_i e_i\|^2$ with assuming $\{e_i\}_{i=1}^n$ to be the orthonormal basis V_n.

Proof. WLOG, assume $e_1, \ldots, e_n$ to be an orthonormal basis of V_n. Then, for any $\|g\| = 1$, $d(g, V_n) = 1 - \sum_{i=1}^n \langle g, e_i \rangle^2$.

$$\begin{aligned}
\sup_{j \in [2n]} d^2(g_j, V_n) &\geq \frac{1}{2n} \sum_{j=1}^{2n} d^2(g_j, V_n) = 1 - \frac{1}{2n} \sum_{j=1}^{2n} \sum_{i=1}^{n} \langle g_j, e_i \rangle^2 \\
&= 1 - \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{2n} \langle g_j, e_i \rangle^2 \\
&\geq 1 - \frac{1}{2n} \sum_{i=1}^{n} \|e_i\|^2 = 1 - \frac{1}{2} = \frac{1}{2}
\end{aligned}$$

□

An important fact. Define a set of functions:

$$\mathcal{G}_m = \left\{ \cos(b^T x) \, : \, \sum_{i=1}^{n} b_i \leq m, b_i \in \mathbb{N} \right\}.$$

Then, $|\mathcal{G}_m| = \binom{m+d}{d}$. By Stirling formula and letting $m = sd$, we have

$$\binom{m+d}{d} \sim \sqrt{\frac{d+m}{md}} \frac{(m+d)^{m+d}}{d^d m^m} \tag{4.2}$$

$$\sim \frac{1}{\sqrt{d}}\left((s+1)\left(1+\frac{1}{s}\right)^s\right)^d \gtrsim \frac{1}{\sqrt{d}}(1+s)^d. \tag{4.3}$$

Meanwhile, 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 $\widehat{\cos(b^T \cdot)}(\omega) = \frac{1}{2}(\delta(\omega - b) + \delta(\omega + b))$. Then, for any $g \in \mathcal{G}_m$,

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

Hence, the important fact is:

Let $\mathcal{B}_s = \{f \in \mathcal{B} : \|f\|_{\mathcal{B}} \leq s^2 d^2\}$. Then, $\mathcal{B}_s$ contains $\frac{1}{\sqrt{d}}(1+s)^d$ (exponentially many) orthonormal functions.

Remark 4.4. The above fact will be also used later to show that the training of two-layer neural networks suffer from CoD for target functions in the Barron space.

Proof of Theorem 4.2 Choose $\bar{m}$ to be the smallest m such that $|\mathcal{G}_m| \geq 2n$. For any $\dim(V_n) = n$,

$$\sup_{\|f\|_{\mathcal{B}} \leq 1} d(f, V_n) \gtrsim \sup_{f \in \frac{1}{\bar{m}^2}\mathcal{G}_{\bar{m}}} d(f, V_n) \geq \frac{1}{\bar{m}^2} \sup_{f \in \mathcal{G}_{\bar{m}}} d(f, V_n) \gtrsim \frac{1}{\bar{m}^2}, \tag{4.4}$$

Let $\bar{m} = sd$. Let $\frac{1}{\sqrt{d}}(1+s)^d \sim 2n$. Then, $s \sim n^{1/d}$. Plugging it into (4.4), we have

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

□

Single neurons Theorem 4.2 provides a lower bound, showing the exponential separation. Can we construct some concrete hard functions? Fortunately, the simple single neuron can serve this purpose. Let $\sigma_v(x) = \sigma(v^T x)$ with $\|v\|_1 = 1$.

- On the one hand, $\|\sigma_v\|_{\mathcal{B}} \lesssim 1$ since $\sigma_v(x) = \int \sigma(w^T x) \, \mathrm{d}\pi(w)$ with $\pi(w) = \delta(w - v)$.
- On the other hand, we can show that σ_v is hard to approximate by using the random feature model:

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

  where $w_j \sim \pi_0$, and π_0 denotes the uniform distribution over the ball $\{w : \|w\|_1 = 1\}$. We can write

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

  where $a(w) = \delta(w - v)$. Obviously, $\|f\|_{\mathcal{H}_{k_{\pi_0}}}^2 = \mathbb{E}[a(w)^2] = \infty$.
- One can actually prove that, approximating σ_v with the above random features require exponentially (in terms d) many features and See [Yehudai and Shamir, 2019]. [Caveat: The result of [Yehudai and Shamir, 2019] is still restricted. For examples, it requires that the coefficient magnitudes can not be exponential in d.]

References

[Barron, 1993] Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945.

[E et al., 2019] E, W., Ma, C., and Wu, L. (2019). A priori estimates of the population risk for two-layer neural networks. *Communications in Mathematical Sciences*, 17(5):1407–1425.

[Li et al., 2020] Li, Z., Ma, C., and Wu, L. (2020). Complexity measures for neural networks with general activation functions using path-based norms. *arXiv preprint arXiv:2009.06132*.

[Yehudai and Shamir, 2019] Yehudai, G. and Shamir, O. (2019). On the power and limitations of random features for understanding neural networks. *Advances in Neural Information Processing Systems*, 32:6598–6608.