Mathematical Deep Learning Theory

Lec 2: Approximation guarantees

Jinhee Paeng

Oct 19, 2023

Seoul National University

Table of Contents

Table of Contents

Review

In previous lecture we proved 2-layer Neural Network structure:

$$f_\theta(x) = \sum_{i=1}^{N} u_i \sigma(a_i^T x + b_i)$$

forms a *dense* subset of the set of continuous function. This gives the mathematical reason of why neural network structure may approximate the target function well.

Review

Question. However, we are computationally limited in the size of width. Size of the layer should be also considered. If so, how well could we approximate when the width is fixed?

Goal. In this lecture, we will show the result in the form of

$$\|f_\theta - f_\star\|^2_{L^2(B)} \leq \mathcal{O}(1/N), \quad \exists \theta \in \Theta_{(N)}.$$

The norm $\|f\|^2_{L^2(B)}$ is defined as $\int_{\mathcal{B}(0,B))} (f(x))^2 dx$

Table of Contents

Approximation guarantees

Theorem

Let $B \in (0, \infty)$ *and* $\sigma : \mathbb{R} \to \mathbb{R}$ *be continuous function satisfying*

$$\lim_{r \to -\infty} \sigma(r) = 0, \quad \lim_{r \to \infty} \sigma(r) = 1, \quad |\sigma(r)| \leq 1, \forall r \in \mathbb{R}.$$

Assume the target function $f_\star : \mathbb{R}^d \to \mathbb{R}$ *satisfies the condition*($\star$)*:*

- *has an absolutely integrable Fourier representation* $\hat{f}_\star : \mathbb{R}^d \to \mathbb{C}$*, i.e.*

$$f_\star = \int_{\mathbb{R}^d} e^{-iw^T x} \hat{f}_\star(w) dw, \forall x \in \mathbb{R}^d, \quad \int_{\mathbb{R}^d} |\hat{f}_\star(w)| dw < \infty.$$

- $\hat{f}_\star$ *satisfies* $Q = \int_{\mathbb{R}^d} \|w\| \, |\hat{f}_\star(w)| dw < \infty$*.*

Then for any $N \in \mathbb{N}$ *there exists* $\theta \in \Theta_{(N+1)}$ *such that*

$$\|f_\theta - f_\star\|^2_{L^2(B)} \leq \frac{5 Q^2 B^2 Vol(\mathcal{B}(0, B))}{N}.$$

Proof Idea - 1

Remark. The main idea of proof is to use Erdös' probabilistic method:

Paul Erdös' probabilistic method.

Consider a random variable X in $D \in \mathbb{R}_{\geq 0}$.
Then, there exists $x \in D$ such that

$$x \leq \mathbb{E}[X].$$

Proof Idea - 1

With such Idea in mind, here's a lemma we will use.

Lemma

Let $\mathcal{H}$ be a Hilbert space. Let $(\mathcal{W}, P)$ be a probability space.
Let $h : \mathcal{W} \to \mathcal{H}$ with $\|h(w)\| \leq H < \infty$ for P-almost all $w \in \mathcal{W}$. Define

$$f = \int_{\mathcal{W}} h(w) dP(w) = \mathbb{E}_{w \sim P}[h(w)].$$

Then, for any $N \in \mathbb{N}$, there exists $h_1, h_2, \cdots, h_N \in \mathcal{H}$ such that

$$\tilde{f} = \sum_{i=1}^{N} \frac{1}{N} h_i, \quad \left\| \tilde{f} - f \right\|^2 \leq \frac{H^2}{N}.$$

Proof Idea - 1

Proof.

Sample $w_1, w_2, \cdots, w_N \overset{i.i.d}{\sim} P$. Define $\hat{f}$ as:

$$\hat{f} = \sum_{i=1}^{N} \frac{1}{N} h(w_i).$$

Then, from $\mathbb{E}[\hat{f}] = f$, we have

$$\mathbb{E}\left[\left\|\hat{f} - f\right\|^2\right] = \frac{1}{N}\mathbb{E}\left[\|h(w_1) - f\|^2\right] \leq \frac{H^2}{N}.$$

Thus, there exists an instance $\tilde{f}$ such that satisfies

$$\tilde{f} = \sum_{i=1}^{N} \frac{1}{N} h_i, \quad \left\|\tilde{f} - f\right\|^2 \leq \frac{H^2}{N}.$$

□

Proof Idea - 2

Remark. Meaning of $Q < \infty$: It allows gradient to be evaluated by DCT

$$\begin{aligned}\nabla f_\star(x) &= \nabla \int_{\mathbb{R}^d} e^{-iw^T x} \hat{f}_\star(w) dw \\ &= \int_{\mathbb{R}^d} \nabla e^{-iw^T x} \hat{f}_\star(w) dw \\ &= \int_{\mathbb{R}^d} -iwe^{-iw^T x} \hat{f}_\star(w) dw.\end{aligned}$$

Also, $|\nabla f_\star(x)| \leq Q < \infty$.

Proof Idea - 2

Furthermore, we can obtain the following result.

Lemma

Let $B \in (0, \infty)$ and $f_\star$ satisfy the condition($\star$). Then, there exists $\phi : \mathbb{R}^d \to [0, 2\pi)$ and a probability measure P on $\mathbb{R}^d \times \mathbb{R}$ such that it is absolutely continuous with respect to the Lebesgue measure and

$$f_\star(x) - f_\star(0) = 2BQ \int_{\mathbb{R}^d \times \mathbb{R}} \sin(b - \phi(w)) \mathbf{1}_{\{w^t x + b \geq 0\}} dP(w, b)$$

for all $x \in \mathcal{B}(0, B)$.

Proof Idea - 2

Abstract proof.

Define ϕ as $\hat{f}_\star(w) = e^{-i\phi(w)}|\hat{f}_\star(w)|$. Then, by taking a real part,

$$f_\star(x) - f_\star(0) = \int_{\mathbb{R}^d} \left(\cos(w^T x + \phi(w)) - \cos(\phi(w))\right) |\hat{f}_\star(w)| dw.$$

Next, we rewrite $\cos(w^T x + \phi(w)) - \cos(\phi(w))$ as

$$-\int_0^{B\|w\|} \mathbf{1}_{\{w^T x \geq b\}} \sin(b+\phi(w)) db + \int_{-B\|w\|}^{0} \mathbf{1}_{\{b \geq -w^T x\}} \sin(b+\phi(w)) db.$$

Define $dP \propto \mathbf{1}_{-B\|w\| \leq b \leq 0} |\hat{f}_\star(w)| db dw$ and use $\hat{f}_\star(w) = \bar{\hat{f}}_\star(-w)$, then

$$f_\star(x) - f_\star(0) = 2BQ \int_{\mathbb{R}^d \times \mathbb{R}} \sin(b - \phi(w)) \mathbf{1}_{\{w^t x + b \geq 0\}} dP(w, b).$$

$\square$

Proof Idea - 3

For the last step of the proof, we will use a function f_δ to form a relation

$$\|f_\delta - f_\star + f_\star(0)\|^2 \leq \epsilon \left(= \frac{(\sqrt{5}-2)^2 B^2 Q^2 \mathrm{Vol}(\mathcal{B}(0,B)))}{N} \right)$$

alongside with a relation induced from the first lemma

$$\|f_\theta - f_\delta\|^2 \leq \frac{4B^2 Q^2 \mathrm{Vol}(\mathcal{B}(0,B)))}{N}.$$

Remark. Note that the number 5 is quite arbitrary. The theorem also holds for any number larger than 4.

Proof Idea - 3

Lemma

Let σ satisfy the assumptions of the theorem, and assume $|s(w,b)| \leq 1$ for all w, b. A function h follows a form of

$$h(x) = \int_{\mathbb{R}^d \times \mathbb{R}} s(w,b) \mathbf{1}_{\{w^T x + b \geq 0\}} dP(w,b),$$

where P is a probability measure that is absolutely continuous with respect to the Lebesgue measure.
Then for any $\delta > 0$, there are s^δ and a probability measure P^δ such that

$$h_\delta(x) = \int_{\mathbb{R}^d \times \mathbb{R}} s^\delta(w,b) \sigma(w^T x + b) dP^\delta(w,b), \quad |s^\delta(w,b)| \leq 1, \forall w, b,$$

where h_δ satisfies $\|h_\delta - h_\star\|_{L^2(B)} \xrightarrow{\delta \to 0} 0$.

Proof Idea - 3

Proof.

By dominated convergence theorem, we have

$$\int_{\mathbb{R}^d \times \mathbb{R}} (s(w,b))^2 \left(\sigma\left(\frac{w^T x}{\delta} + \frac{b}{\delta} \right) - \mathbf{1}_{\{w^T x + b \geq 0\}} \right)^2 dP(w,b) \xrightarrow{\delta \to 0} 0.$$

Define s^δ and a probability measure P^δ using the change of variables

$$\tilde{w} = w/\delta, \tilde{b} = b/\delta, \quad s^\delta(\tilde{w}, \tilde{b}) = s(\delta \tilde{w}, \delta \tilde{b}),$$

as in conclusion,

$$\begin{aligned} h_\delta(x) &= \int_{\mathbb{R}^d \times \mathbb{R}} s^\delta(\tilde{w}, \tilde{b}) \sigma(\tilde{w}^T x + \tilde{b}) dP^\delta(\tilde{w}, \tilde{b}) \\ &= \int_{\mathbb{R}^d \times \mathbb{R}} s(w,b) \sigma\left(\frac{w^T x}{\delta} + \frac{b}{\delta} \right) dP(w,b). \end{aligned}$$

□

Proof of the Theorem

Now let's sum up the results we've shown

Proof of the main theorem.

- With $s(w,b) = \sin(b - \phi(w))$, $|s(w,b)| \leq 1$ for all w, b and

$$f_\star(x) - f_\star(0) = 2BQ \int_{\mathbb{R}^d \times \mathbb{R}} s(w,b) \mathbf{1}_{\{w^t x + b \geq 0\}} dP(w,b).$$

- There exists $f_\delta = 2BQ \int s^\delta(w,b)\sigma(w^T x + b) dP^\delta(w,b)$ that satisfies

$$\|f_\delta - (f_\star - f_\star(0))\|^2 \leq \frac{(\sqrt{5}-2)^2 B^2 Q^2 \mathrm{Vol}(\mathcal{B}(0,B)))}{N}.$$

- We know that there exists $f_{\theta'}(x) = \sum_{i=1}^N s^\delta(w_i, b_i)\sigma(w_i^T x + b_i)$ with

$$\|f_\delta - f_{\theta'}\|^2 \leq \frac{4B^2 Q^2 \mathrm{Vol}(\mathcal{B}(0,B)))}{N}.$$

Proof of the Theorem

Proof of the main theorem.

Thus, choose the coefficients λ_i as $\lambda_i = s^\delta(w_i, b_i)$ with

$$\lambda_{N+1} = \frac{f_\star(0)}{\sigma(b_{N+1})}, \quad w_{N+1} = 0, \quad \sigma(b_{N+1}) \neq 0.$$

Then with the triangular inequality, f_θ defined as

$$f_\theta(x) = \sum_{i=1}^{N+1} \lambda_i \sigma(w_i^T x + b_i)$$

satisfies

$$\|f_\theta - f\star\|^2_{L^2(B)} \leq \frac{5B^2 Q^2 \mathrm{Vol}(\mathcal{B}(0,B)))}{N}.$$

□