CS 839: Theoretical Foundations of Deep Learning **Spring 2022**

Lecture 4 Approximation II

Instructor: Yingyu Liang *Date: Feb 3^{rd}, 2022* *Scriber: Zhenmei Shi*

1 Overview

In the previous lecture, we showed how 3-layer-neural-networks with ReLU activation function can approximate high-dimension Lipschitz function family with small approximation error. In this lecture, we will shift our attention to universal approximation.

2 Universal Approximation

Definition 1 (Universal Approximation)**.** For a class of functions $\mathcal{F}$ and a compact set $S \subset \mathbb{R}^d$, if for every continuous function g on S and for any $\epsilon > 0$, there exists $f \in \mathcal{F}$ such that $\|f - g\|_\infty := \max_{\mathbf{x} \in S} |f(\mathbf{x}) - g(\mathbf{x})| \leqslant \epsilon$. Then, the class of functions $\mathcal{F}$ is a universal approximator of all continuous functions on S.

The following theorem characterizes the universal approximator.

Theorem 2 (Stone-Weierstrauss Theorem (limited version))**.** Let $\mathcal{F}$ be a class of functions defined on a compact set $S \subset \mathbb{R}^d$. If $\mathcal{F}$ satifies:

1. Each $f \in \mathcal{F}$ is continuous.
2. For every $\mathbf{x}$, there exists $f \in \mathcal{F}$ such that $f(\mathbf{x}) \neq 0$.
3. For every $\mathbf{x}, \mathbf{x}'$ with $\mathbf{x} \neq \mathbf{x}'$, there exists $f \in \mathcal{F}$ such that $f(\mathbf{x}) \neq f(\mathbf{x}')$ ($\mathcal{F}$ separates points).
4. $\mathcal{F}$ is closed under multiplication ($\forall f, g \in \mathcal{F}$, we have $h \in \mathcal{F}$ and $h(\mathbf{x}) = f(\mathbf{x})g(\mathbf{x})$) and vector space operations ($\mathcal{F}$ is an algebra).

Then, for every continuous function $g : \mathbb{R}^d \mapsto \mathbb{R}$, and any $\epsilon > 0$, there exists $f \in \mathcal{F}$ such that $\|f - g\|_\infty \leqslant \epsilon$. In other words, $\mathcal{F}$ is a universal approximator.

Remark 3. It is easy to see that Conditions 2 and 3 are necessary. If remove Condition 2, there exist $\mathbf{x}$ such that $\forall f \in \mathcal{F}, f(\mathbf{x}) = 0$. Then we could not approximate functions g with $g(\mathbf{x}) \neq 0$. If remove Condition 3, there exist $\mathbf{x}, \mathbf{x}'$, with $\mathbf{x} \neq \mathbf{x}'$, so that $\forall f \in \mathcal{F}, f(\mathbf{x}) = f(\mathbf{x}')$. Then we could not approximate functions g with $g(\mathbf{x}) \neq g(\mathbf{x}')$ since $\|f - g\|_\infty \geqslant \max\{|f(\mathbf{x}) - g(\mathbf{x})|, |f(\mathbf{x}') - g(\mathbf{x}')|\} > 0$.

We now discuss universal approximation with infinitely wide neural networks with a single hidden layer, beginning with some preliminaries. Consider the following definition for

1-hidden layer neural network function classes with nonlinear activation σ, input dimensionality d, and hidden layer width m.

$$\mathcal{F}_{\sigma,d,m} = \{\mathbf{x} \mapsto \mathbf{a}\sigma(\mathbf{W}\mathbf{x} + \boldsymbol{b}), \mathbf{a} \in \mathbb{R}^{1\times m}, \mathbf{W} \in \mathbb{R}^{m\times d}, \boldsymbol{b} \in \mathbb{R}^m\}.$$

We then define the infinitely wide class of one hidden layer neural networks as follows:

$$\mathcal{F}_{\sigma,d} = \bigcup_{m\geqslant 0} \mathcal{F}_{\sigma,d,m}.$$

Now, we prove $\mathcal{F}_{\exp,d}$ and $\mathcal{F}_{\cos,d}$ are two universal approximators, by checking the Stone-Weierstrass conditions.

Example 4. Prove $\mathcal{F}_{\exp,d}$ is a universal approximator.

Proof. We need to verify the four conditions of the Stone-Weierstrass theorem.

1. Each $f \in \mathcal{F}_{\exp,d}$ is continuous.

2. $\forall \mathbf{x}$, $f_{\mathbf{x}}(\mathbf{z}) = \exp(\mathbf{x}^\top \mathbf{z}) \neq 0$ at $\mathbf{z} = \mathbf{x}$.

3. For every $\mathbf{x}, \mathbf{x}'$ with $\mathbf{x} \neq \mathbf{x}'$, consider the linear function h:

$$h(\mathbf{z}) = \frac{(\mathbf{z} - \mathbf{x})^\top (\mathbf{x}' - \mathbf{x})}{\|\mathbf{x}' - \mathbf{x}\|_2^2}.$$

   Then $h(\mathbf{x}) = 0$ and $h(\mathbf{x}') = 1$. Now let

$$f(\mathbf{z}) = \exp(h(\mathbf{z})) = \exp\left(\frac{(\mathbf{z} - \mathbf{x})^\top (\mathbf{x}' - \mathbf{x})}{\|\mathbf{x}' - \mathbf{x}\|_2^2}\right).$$

   Thus, $f(\mathbf{x}) = 1 \neq e = f(\mathbf{x}')$.

4. $\forall f, g \in \mathcal{F}_{\exp,d}, \forall \alpha \in \mathbb{R}$, suppose $f(\mathbf{x}) = a_f\sigma(\mathbf{W}_f\mathbf{x} + \boldsymbol{b}_f), g(\mathbf{x}) = a_g\sigma(\mathbf{W}_g\mathbf{x} + \boldsymbol{b}_g)$.
   (i) We have $\alpha f \in \mathcal{F}_{\exp,d}$.
   (ii)

$$f + g = [a_f, a_g]\sigma\left(\begin{bmatrix}\mathbf{W}_f \\ \mathbf{W}_g\end{bmatrix}\mathbf{x} + \begin{bmatrix}\boldsymbol{b}_f \\ \boldsymbol{b}_g\end{bmatrix}\right).$$

   Thus, $f + g \in \mathcal{F}_{\exp,d}$.
   (iii)

$$\begin{aligned} f \cdot g(\mathbf{x}) &= \left(\sum_{i=1}^{m_f} a_{fi} \exp(\langle \mathbf{W}_{fi}, \mathbf{x}\rangle + \boldsymbol{b}_{fi})\right)\left(\sum_{j=1}^{m_g} a_{gj} \exp(\langle \mathbf{W}_{gj}, \mathbf{x}\rangle + \boldsymbol{b}_{gj})\right) \\ &= \left(\sum_{i=1}^{m_f}\sum_{j=1}^{m_g} a_{fi}a_{gj} \exp(\langle \mathbf{W}_{fi} + \mathbf{W}_{gj}, \mathbf{x}\rangle + \boldsymbol{b}_{fi} + \boldsymbol{b}_{gj})\right). \end{aligned}$$

   Thus, $f \cdot g \in \mathcal{F}_{\exp,d}$.

Based on the above four conditions, as a result of the Stone-Weierstrass theorem, $\mathcal{F}_{\exp,d}$ is a universal approximator. □

Example 5. Prove $\mathcal{F}_{\cos,d}$ is a universal approximator. In particular, the cosine function has the helpful property $2\cos(\alpha)\cos(\beta) = \cos(\alpha+\beta) + \cos(\alpha-\beta)$. This allows for multiplicative closure of elements in $\mathcal{F}_{\cos,d}$: by multiplying two neural networks together, we obtain a third neural network, which implies that $\forall f, g \in \mathcal{F}_{\cos,d}, f \cdot g \in \mathcal{F}_{\cos,d}$.

Proof. We only prove multiplicative closure for $\mathcal{F}_{\cos,d}$. The proof of all other conditions is similar in Example 4.

$\forall f, g \in \mathcal{F}_{\cos,d}$, suppose $f(\mathbf{x}) = a_f\sigma(\mathbf{W}_f\mathbf{x} + \boldsymbol{b}_f), g(\mathbf{x}) = a_g\sigma(\mathbf{W}_g\mathbf{x} + \boldsymbol{b}_g)$,

$$
\begin{aligned}
f \cdot g(\mathbf{x}) &= \left(\sum_{i=1}^{m_f} a_{fi}\cos(\langle \mathbf{W}_{fi}, \mathbf{x}\rangle + \boldsymbol{b}_{fi})\right)\left(\sum_{j=1}^{m_g} a_{gj}\cos(\langle \mathbf{W}_{gj}, \mathbf{x}\rangle + \boldsymbol{b}_{gj})\right) \\
&= \left(\sum_{i=1}^{m_f}\sum_{j=1}^{m_g} a_{fi}a_{gj}\frac{1}{2}\left(\cos(\langle \mathbf{W}_{fi} + \mathbf{W}_{gj}, \mathbf{x}\rangle + \boldsymbol{b}_{fi} + \boldsymbol{b}_{gj}) + \cos(\langle \mathbf{W}_{fi} - \mathbf{W}_{gj}, \mathbf{x}\rangle + \boldsymbol{b}_{fi} - \boldsymbol{b}_{gj})\right)\right).
\end{aligned}
$$

Thus, $f \cdot g \in \mathcal{F}_{\cos,d}$. □

For arbitrary activation functions, we have the following theorem.

Theorem 6 (Hornik, Stinchcombe, and White 1989)**.** Suppose $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is continuous, and satisfies

$$
\lim_{z\to-\infty}\sigma(z) = 0, \lim_{z\to+\infty}\sigma(z) = 1.
$$

Then $\mathcal{F}_{\sigma,d}$ is a universal approximator.

This theorem provides us with a useful tool to prove a function class with arbitrary activation to be universal, not directly via the Stone-Weierstrass theorem.

Since its proof is part of the homework, we skip the proof here. A sketch of the proof could be: Given $\epsilon > 0$ and continuous g, pick $h \in \mathcal{F}_{\cos,d}$ (or, $\mathcal{F}_{\exp,d}$) with $\sup_{\mathbf{x}\in[0,1]^d} h(\mathbf{x}) - g(\mathbf{x}) \leqslant \epsilon/2$. To finish, replace all appearances of cos with an element of $\mathcal{F}_{\sigma,1}$.

Remark 7. Note that $\mathcal{F}_{\mathrm{ReLU},d}$ is also a universal approximator based on Theorem 6. In particular, we can build an intermediate activation $\sigma_1(z) = \mathrm{ReLU}(z) - \mathrm{ReLU}(z-1)$, which satisfies the conditions of the above theorem. By $\mathcal{F}_{\sigma_1,d} \subset \mathcal{F}_{\mathrm{ReLU},d}$, we have $\mathcal{F}_{\mathrm{ReLU},d}$ is a universal approximator.

3 Infinite-width Networks

In the next section of this lecture, we introduced how to represent the target function as an infinite-width network via Fourier inversion. Before that, we first provide a definition for integral representation of infinite-width networks and then take a brief review of the Fourier transform.

Definition 8. An infinite-width shallow network is characterized by a signed measure ν (can be negative) over weight vectors in $\mathbb{R}^P$:

$$\mathbf{x} \mapsto \int \sigma(\mathbf{w}^\top \mathbf{x}) \mathrm{d}\nu(\mathbf{w}).$$

We can alternatively write the derivative of the measure as a function of $\mathbf{w}$:

$$\mathbf{x} \mapsto \int \sigma(\mathbf{w}^\top \mathbf{x}) g(\mathbf{w}) \mathrm{d}\mathbf{w},$$

where $\mathrm{d}\nu(\mathbf{w}) = g(\mathbf{w})\mathrm{d}\mathbf{w}$.

Example 9. Suppose $\mathbf{w} \in \{\mathbf{w}_1, \mathbf{w}_2\}$ and $g(\mathbf{w}_1) = \frac{1}{2}, g(\mathbf{w}_2) = -1$. Then $\int \sigma(\mathbf{w}^\top \mathbf{x}) g(\mathbf{w}) \mathrm{d}\mathbf{w} = \frac{1}{2}\sigma(\mathbf{w}_1^\top \mathbf{x}) - \sigma(\mathbf{w}_2^\top \mathbf{x})$.

3.1 Review Fourier Transformation

Definition 10. Let L^p be the function class such that $f \in L^p$ iff $[\int |f(x)|^p \mathrm{d}x]^{1/p} < +\infty$. If $f \in L^1$, the Fourier transform of f is:

$$\widehat{f}(\mathbf{w}) := \int \exp(-2\pi i \mathbf{w}^\top \mathbf{x}) f(\mathbf{x}) \mathrm{d}\mathbf{x}.$$

If $f \in L^1$, and $\widehat{f} \in L^1$, the Fourier inversion is defined as:

$$f'(\mathbf{x}) := \int \exp(2\pi i \mathbf{w}^\top \mathbf{x}) \widehat{f}(\mathbf{w}) \mathrm{d}\mathbf{w}.$$

In Definition 10, $f(x)$ could be viewed as an infinite-width complex-valued neural network function. Since $\exp(iz) = \cos(z) + i\sin(z)$, the real part of $f(x)$ is defined as:

$$\bar{f}(x) = Re(f'(x)) = \int \cos(2\pi \mathbf{w}^\top \mathbf{x}) \widehat{f}(\mathbf{w}) \mathrm{d}\mathbf{w}.$$

Next lecture, we will rewrite the target function as two infinite-width networks with standard threshold activations, using the Fourier transforms in the weighting measure.