CS 839: Theoretical Foundations of Deep Learning		Spring 2022
Lecture 5 Approximation III		
Instructor: Yingyu Liang	*Date: Feb 8th, 2022*	*Scriber: Zhenmei Shi*

1 Overview

Last lecture we discussed universal approximation. In this lecture, we will continue to introduce how to represent the target function as an infinite-width network via Fourier inversion. The material is based on Chapter 3 of [1] which is in turn based on the seminal paper of Barron [2].

2 Infinite-width Networks

Definition 1. 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 $\mathbf{w}$:

$$\mathbf{x} \mapsto \int \sigma(\mathbf{w}^\top \mathbf{x}) g(\mathbf{w}) \mathrm{d}\mathbf{w}, \tag{1}$$

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

In Definition 1, integral operator can be viewed as sum of all neurons, $\sigma(\mathbf{w}^\top \mathbf{x})$ can be viewed as a neuron and $g(\mathbf{w})$ can be viewed as the weight on the neuron or a_i defined in 2-layer-neural network of the previous lecture.

2.1 Review Fourier Transformation

Definition 2. We define L^p as the function class such that $f \in L^p$ if $[\int |f(\mathbf{x})|^p \mathrm{d}x]^{1/p} < +\infty$. If $f \in L^1$, $f : \mathbb{R}^d \mapsto \mathbb{C}$, 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:

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

Since $\exp(iz) = \cos(z) + i\sin(z)$, the real part of $\widetilde{f}$ is defined as:

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

In Definition 2, $f \in L^1$ and $\widehat{f} \in L^1$ can guarantee $\widetilde{f}(\mathbf{x}) = f(\mathbf{x})$ almost everywhere. If furthermore f is continuous, then $\widetilde{f}(\mathbf{x}) = f(\mathbf{x})$ for any $\mathbf{x}$. If f is a real-valued function with $f \in L^1$ and $\widehat{f} \in L^1$, then $\widetilde{f}(\mathbf{x}) = f(\mathbf{x})$ almost everywhere. In the following we will just say $\widetilde{f}(\mathbf{x}) = f(\mathbf{x})$.

$\widetilde{f}$ could be viewed as an infinite-width complex-valued neural network function.

2.2 Rewrite Target Function as Infinite-width Networks

We will rewrite the target function as two infinite-width networks with standard threshold activations, using the Fourier transforms in the weighting measure. First, we introduce a useful lemma.

Lemma 3. Suppose $g : \mathbb{R} \mapsto \mathbb{R}$ is differentiable. For $z \in [0, B]$, we have

$$g(z) - g(0) = \int_0^B \mathbb{I}[z \geqslant b] g'(b) \mathrm{d}b.$$

Proof. By the fundamental theorem of calculus:

$$\begin{aligned} g(z) - g(0) &= \int_0^z g'(b)\mathrm{d}b \\ &= \int_0^z 1 \cdot g'(b)\mathrm{d}b + \int_z^B 0 \cdot g'(b)\mathrm{d}b \\ &= \int_0^z \mathbb{I}[z \geqslant b] g'(b)\mathrm{d}b + \int_z^B \mathbb{I}[z \geqslant b] g'(b)\mathrm{d}b \\ &= \int_0^B \mathbb{I}[z \geqslant b] g'(b)\mathrm{d}b. \end{aligned}$$

□

Then we have the following theorem.

Theorem 4. Suppose $f \in L^1$, and $\widehat{f} \in L^1$, $\int \|\nabla \widehat{f}(\mathbf{w})\| \mathrm{d}\mathbf{w} < \infty$. We write $\widehat{f}(\mathbf{w}) = |\widehat{f}(\mathbf{w})| \exp(2\pi\theta(\mathbf{w}))$, where $\theta(\mathbf{w})$ is the phase of $\mathbf{w}$ in its polar representation. Then for any $\|\mathbf{x}\| \leqslant 1$, we have,

$$\begin{aligned} f(\mathbf{x}) - f(\mathbf{0}) = &- 2\pi \int \int_0^{\|\mathbf{w}\|} \mathbb{I}[\mathbf{w}^\top \mathbf{x} - b \geqslant 0][\sin(2\pi b + 2\pi\theta(\mathbf{w}))]|\widehat{f}(\mathbf{w})| \mathrm{d}b \mathrm{d}\mathbf{w} \qquad (2) \\ &+ 2\pi \int \int_{-\|\mathbf{w}\|}^0 \mathbb{I}[-\mathbf{w}^\top \mathbf{x} + b \geqslant 0][\sin(2\pi b + 2\pi\theta(\mathbf{w}))]|\widehat{f}(\mathbf{w})| \mathrm{d}b \mathrm{d}\mathbf{w}. \qquad (3) \end{aligned}$$

Proof. Since $\exp(iz) = \cos(z) + i\sin(z)$, and f is real-valued,

$$\begin{aligned} f(\mathbf{x}) &= \int Re[\exp(2\pi i \mathbf{w}^\top \mathbf{x})]\widehat{f}(\mathbf{w})\mathrm{d}\mathbf{w} \\ &= \int Re[\exp(2\pi i \mathbf{w}^\top \mathbf{x})]|\widehat{f}(\mathbf{w})|\exp(2\pi i\theta(\mathbf{w}))\mathrm{d}\mathbf{w} \\ &= \int Re[\exp(2\pi i \mathbf{w}^\top \mathbf{x} + 2\pi i\theta(\mathbf{w}))]|\widehat{f}(\mathbf{w})|\mathrm{d}\mathbf{w} \\ &= \int \cos(2\pi \mathbf{w}^\top \mathbf{x} + 2\pi\theta(\mathbf{w}))|\widehat{f}(\mathbf{w})|\mathrm{d}\mathbf{w}. \end{aligned}$$

Then, $f(\mathbf{0}) = \int \cos(2\pi\theta(\mathbf{w})|\widehat{f}(\mathbf{w})|\mathrm{d}\mathbf{w}$,

$$f(\mathbf{x}) - f(\mathbf{0}) = \int [\cos(2\pi \mathbf{w}^\top \mathbf{x} + 2\pi\theta(\mathbf{w})) - \cos(2\pi\theta(\mathbf{w})]|\widehat{f}(\mathbf{w})|\mathrm{d}\mathbf{w} \tag{4}$$

$$= \int \left[-2\pi \int_0^{\mathbf{w}^\top \mathbf{x}} \sin(2\pi b + 2\pi\theta(\mathbf{w}))\mathrm{d}b\right] |\widehat{f}(\mathbf{w})|\mathrm{d}\mathbf{w}, \tag{5}$$

where the last equation is from fundamental theorem of calculus. Let

$$g(\mathbf{w}^\top \mathbf{x}) = \int_0^{\mathbf{w}^\top \mathbf{x}} \sin(2\pi b + 2\pi\theta(\mathbf{w}))\mathrm{d}b.$$

Then

$$\begin{aligned} g'(\mathbf{w}^\top \mathbf{x}) &= \sin(2\pi b + 2\pi\theta(\mathbf{w})) \\ g(0) &= 0. \end{aligned}$$

Since we only know $\|\mathbf{x}\| \leqslant 1$, which is not sufficient to determine the sign of $\mathbf{w}^\top \mathbf{x}$, we need to divide the integral into two separate cases: $\mathbf{w}^\top \mathbf{x} \geqslant 0$ (case A), and $\mathbf{w}^\top \mathbf{x} \leqslant 0$ (case B). Also, since $\mathbf{w}^\top \mathbf{x} \leqslant \|\mathbf{w}\| \cdot \|\mathbf{x}\| \leqslant \|\mathbf{w}\|$, by Lemma 3, we have,

$$g(\mathbf{w}^\top \mathbf{x}) = \begin{cases} \int_0^{\|\mathbf{w}\|} \mathbb{I}[\mathbf{w}^\top \mathbf{x} \geqslant b] \sin(2\pi b + 2\pi\theta(\mathbf{w}))\mathrm{d}b, & \text{if } \mathbf{w}^\top \mathbf{x} \geqslant 0, \\ -\int_{-\|\mathbf{w}\|}^0 \mathbb{I}[-\mathbf{w}^\top \mathbf{x} \geqslant -b] \sin(2\pi b + 2\pi\theta(\mathbf{w}))\mathrm{d}b, & \text{if } \mathbf{w}^\top \mathbf{x} < 0. \end{cases}$$

Putting the two parts together,

$$\begin{aligned} g(\mathbf{w}^\top \mathbf{x}) = &\int_0^{\|\mathbf{w}\|} \mathbb{I}[\mathbf{w}^\top \mathbf{x} \geqslant b] \sin(2\pi b + 2\pi\theta(\mathbf{w}))\mathrm{d}b \\ &- \int_{-\|\mathbf{w}\|}^0 \mathbb{I}[-\mathbf{w}^\top \mathbf{x} \geqslant -b] \sin(2\pi b + 2\pi\theta(\mathbf{w}))\mathrm{d}b. \end{aligned}$$

Plug $g(\mathbf{w}^\top \mathbf{x})$ back into (5),

$$f(\mathbf{x}) - f(\mathbf{0}) = -2\pi \int \int_0^{\|\mathbf{w}\|} \mathbb{I}[\mathbf{w}^\top \mathbf{x} - b \geqslant 0][\sin(2\pi b + 2\pi\theta(\mathbf{w}))]|\widehat{f}(\mathbf{w})|\mathrm{d}b\mathrm{d}\mathbf{w} \tag{6}$$

$$+ 2\pi \int \int_{-\|\mathbf{w}\|}^0 \mathbb{I}[-\mathbf{w}^\top \mathbf{x} + b \geqslant 0][\sin(2\pi b + 2\pi\theta(\mathbf{w}))]|\widehat{f}(\mathbf{w})|\mathrm{d}b\mathrm{d}\mathbf{w}, \tag{7}$$

which completes the proof. □

2.3 Subsampling from infinite widths NNs

We now switch our focus to the issue of sampling a finite width neural network from our infinite width representation. We have shown that in many cases we can represent f using an infinite width neural network, but if we want a finite width network for use in practice, can we sample one to give a good approximation? In other words, will a finite neural network, obtained through averaging a bunch of activation functions in the layer before their output, drawn with probability proportional to the size of the weight measure, approximate the original function well?

Let us first consider the problem of approximating the mean by sampling in a Hilbert space. A Hilbert space is a complete vector space endowed with an inner product.

Suppose $X = \mathbb{E}[V]$, where random vector V is supported on a set S. A natural way to compute X is to consider $\widehat{X} := \frac{1}{k}\sum_{i=1}^{k} V_i$, where $\{V_1, \ldots, V_k\}$ are drawn i.i.d. from the distribution of V.

We want to show $\widehat{X} \approx X$ by showing $\|\widehat{X} - X\|$ is small, where $\|Z\| = \langle Z, Z\rangle^{1/2}$ is the norm induced by the inner product on the Hilbert space. A characterization can be seen in the following lemma:

Lemma 5 (Maurey)**.** Let $X = \mathbb{E}[V]$, with V supported on set S, and let $\{V_1, \ldots, V_k\}$ be drawn i.i.d. from the distribution of V. Then

$$\mathbb{E}_{V_1,\ldots,V_k}\left\|X - \frac{1}{k}\sum_{i=1}^{k} V_i\right\|^2 \leqslant \frac{\mathbb{E}\|V\|^2}{k} \leqslant \frac{\sup_{U\in S}\|U\|^2}{k},$$

and there exists $\{U_1, \ldots, U_k\}$ in S so that

$$\left\|X - \frac{1}{k}\sum_{i=1}^{k} U_i\right\|^2 \leqslant \mathbb{E}_{V_1,\ldots,V_k}\left\|X - \frac{1}{k}\sum_{i=1}^{k} V_i\right\|^2.$$

Let us now see if we can extend the results of this lemma to infinite width neural networks. To do so, we must first define a Hilbert space in which the functions that can be approximated by neural networks. Let us denote a Hilbert space on such functions as $\mathcal{F}$, with an inner product defined as follows: $\forall f, g \in \mathcal{F}, \langle f, g\rangle = \int f(\mathbf{x})g(\mathbf{x})\mathrm{d}p(\mathbf{x})$. We denote the norm induced by this inner product as $\|\cdot\|_{L_2(p)}$, and $\|f\|^2_{L_2(p)} = \langle f, f\rangle$, $\|f\|_{L_2(p)} = \sqrt{\int f^2(\mathbf{x})\mathrm{d}p(\mathbf{x})}$.

However, an issue with the infinite width neural networks we have constructed so far is: the measures are not nice enough such that the expectation is calculated easily. As an example, consider $x \in [0, 1]$ and $\sin(2\pi x) = \int_0^1 \mathbb{I}[x \geqslant b]2\pi cos(2\pi x)\mathrm{d}b$. We see two issues:

1. $\cos(2\pi b)$ is not always postive or negative.
2. $\int_0^1 |2\pi\cos(2\pi x)|\mathrm{d}b \neq 1$.

So our "distribution" on $\mathbf{w}$ (the neural network weight) is not a probability.

2.4 Modifications to the neural network

We would want these two conditions to be satisfied. To this end, we can modify the neural network a little bit. Let us introduce the following modifications to a measure μ.

1. Given a signed, non-identically zero measure μ, we decompose $\mu = \mu_+ - \mu_-$, where μ_+ , μ_- are non-negative measures with disjoint support.
2. For non-negative measures μ_+ , μ_-, define the total mass as $\|\mu_+\|_1 = \int |\mathrm{d}\mu_+|$, $\|\mu_-\|_1 = \int |\mathrm{d}\mu_-|$. As the measures are disjoint, it is easy to see that $\|\mu\|_1 = \|\mu_+\|_1 + \|\mu_-\|_1 = \int |\mathrm{d}\mu|$.
3. Denote $\widetilde{\mu}_s := \frac{\mu_s}{\|\mu\|_1}$, where $s \in \{+1, -1\}$.

We write the general expression for infinite width neural networks as $\mathbf{x} \mapsto \int \sigma(\mathbf{w}^\top \mathbf{x} - b) g(\mathbf{w}, b) \mathrm{d}(\mathbf{w}, b)$, we can rewrite it to be:

$$\begin{aligned}
\int \sigma(\mathbf{w}^\top \mathbf{x} - b) g(\mathbf{w}, b) \mathrm{d}(\mathbf{w}, b) =& \|\mu\|_1 \int \sigma(\mathbf{w}^\top \mathbf{x} - b) s(\mathbf{w}, b) \frac{s(\mathbf{w}, b) g(\mathbf{w}, b) \mathrm{d}(\mathbf{w}, b)}{\|\mu\|_1} \\
=& \|\mu\|_1 \int \sigma(\mathbf{w}^\top \mathbf{x} - b) s(\mathbf{w}, b) \frac{s(\mathbf{w}, b) \mathrm{d}\mu}{\|\mu\|_1} \\
=& \|\mu\|_1 \int \sigma(\mathbf{w}^\top \mathbf{x} - b) s(\mathbf{w}, b) \frac{\mathrm{d}\mu_s}{\|\mu\|_1} \\
=& \int \|\mu\|_1 \sigma(\mathbf{w}^\top \mathbf{x} - b) s(\mathbf{w}, b) \mathrm{d}\widetilde{\mu}_s,
\end{aligned}$$

where $g(\mathbf{w}, b)\mathrm{d}(\mathbf{w}, b) = \mathrm{d}\mu$, $s(\mathbf{w}, b) = \mathrm{sign}(g(\mathbf{w}, b))$.

As we can see from the above formula, we sample $\mathbf{w}$ from $\widetilde{\mu}_s$ and denote the new output $\widetilde{\sigma}(\mathbf{w}^\top \mathbf{x} - b) = \|\mu\|_1 \sigma(\mathbf{w}^\top \mathbf{x} - b) s(\mathbf{w}, b)$. As a result, our neural network can be written as an expectation over a new measure $\widetilde{\mu}_s$. We can now state our result on sampling from infinite width neural networks by Maurey Lemma, by setting V to $\widetilde{\sigma}(\mathbf{w}^\top \mathbf{x} - b)$, $\mathbb{E}[V]$ to $\int \|\mu\|_1 \sigma(\mathbf{w}^\top \mathbf{x} - b) s(\mathbf{w}, b) \mathrm{d}\widetilde{\mu}_s$ and $\|\cdot\|$ to $\|\cdot\|_{L_2(p)}$.

For Fourier transformation, we have,

$$\nabla \widehat{f}(\mathbf{w}) = 2\pi i \mathbf{w} \widehat{f}(\mathbf{w}).$$

Thus, for infinite-width networks in Barron's Theorem,

$$\begin{aligned}
\|\mu\|_1 &= 2\pi \int \int_0^{\|\mathbf{w}\|} |\sin(2\pi b + 2\pi\theta(\mathbf{w})) \widehat{f}(\mathbf{w})| \mathrm{d}b \mathrm{d}\mathbf{w} \\
&\leqslant 2\pi \int \|\mathbf{w}\| |\widehat{f}(\mathbf{w})| \mathrm{d}\mathbf{w} \\
&= \int \|\nabla \widehat{f}(\mathbf{w})\| \mathrm{d}\mathbf{w}.
\end{aligned}$$

Now, consider Fourier representation in Barron's theorem,

$$\begin{aligned}
f(\mathbf{x}) - f(\mathbf{0}) =& -2\pi \int \int_0^{\|\mathbf{w}\|} \mathbb{I}[\mathbf{w}^\top \mathbf{x} - b \geqslant 0][\sin(2\pi b + 2\pi\theta(\mathbf{w}))] |\widehat{f}(\mathbf{w})| \mathrm{d}b \mathrm{d}\mathbf{w} \\
&+ 2\pi \int \int_{-\|\mathbf{w}\|}^0 \mathbb{I}[-\mathbf{w}^\top \mathbf{x} + b \geqslant 0][\sin(2\pi b + 2\pi\theta(\mathbf{w}))] |\widehat{f}(\mathbf{w})| \mathrm{d}b \mathrm{d}\mathbf{w},
\end{aligned}$$

by Maurey's lemma there exist samples $\{(\mathbf{w}_i, b_i, s_i)\}_{i=1}^k$ such that for any probability measure P supported on $\mathbf{x}$, where $\|\mathbf{x}\| \leqslant 1$,

$$\left\| f(\mathbf{x}) - f(\mathbf{0}) - \frac{1}{k}\sum_{i=1}^{k} s_i \|\mu\|_1 \mathbb{I}[\mathbf{w}_i^\top \mathbf{x} - b \geqslant 0] \right\|_{L_2(P)} \leqslant \frac{\|\mu\|_1^2 \sup_{\mathbf{w},b} \|\mathbb{I}[\mathbf{w}^\top \mathbf{x} - b \geqslant 0]\|_{L_2(P)}^2}{k}$$
$$\leqslant \frac{\left(\int \|\nabla \widehat{f}(\mathbf{w})\| \mathrm{d}\mathbf{w}\right)^2}{k}.$$

Note that this means our sampled neural network approximates the original function with order $O(\frac{1}{k})$. Thus, we see that the infinite width neural networks we have constructed are not just for theoretical understanding. In practice by Maurey's Lemma, when an infinite representation exists, there is a sampled finite neural network that approximates the output function better and better as the size of the sample (or number of neurons) increases.

References

[1] Deep learning theory lecture note, 2021, version 2021-10-27 v0.0-e7150f2d.

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