Topics in Deep Learning Theory (Spring 2025)

Lecture 7: Two-Layer Neural Networks

Instructor: Lei Wu Date: April 21, 2025

The set of functions that can be represented by two-layer neural nets is given by

$$\mathcal{F}_{\sigma,d} = \left\{ x \mapsto a^\top \sigma(Wx+b) : a \in \mathbb{R}^m, b \in \mathbb{R}^m, W \in \mathbb{R}^{m\times d}, m \in \mathbb{N} \right\}.$$

Next, we study the approximation power of two-layer neural nets.

1 Universal approximation properties

Definition 1.1 (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

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

Theorem 1.2 ([Siegel and Xu, 2020])**.** *Assume σ 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. First, we assume that $\sigma \in C^\infty(\mathbb{R})$. 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 $i = 1, \dots, d$. Similarly, for any $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$,

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

Since $\mathcal{F}_{\sigma,1}$ is dense in $C([0,1])$, σ cannot be a polynomial. Hence, we can choose $w = 0$ and $b \in \mathbb{R}$ such that $\sigma^k(b) \neq 0$ for any $k \in \mathbb{N}$. Therefore, all the polynomials of the form $x_1^{\alpha_1} \cdots x_d^{\alpha_d}$ are in $\bar{\mathcal{F}}_{\sigma,d}$. This implies that $\bar{\mathcal{F}}_{\sigma,d}$ contains all the polynomials. By Weierstrass-Stone theorem, $\bar{\mathcal{F}}_{\sigma,d}$ is dense in $C(\Omega)$.

For non-smooth σ, since $\mathcal{F}_{\sigma,1}$ is dense in $C([0,1])$, we can use a two-layer neural net to approximate a smooth one. Then, the same results follow. □

- The above proof implies that $\mathcal{F}_{\sigma,d}$ has UAP if σ is smooth and non-polynomial.
- For non-smooth networks, we only need to consider the one-dimensional case, where explicitly constructive proof is often doable. The following lemma concerns the ReLU activation function.

Lemma 1.3. *Assume $\sigma(z) = \max(0, z)$. For any Lipschitz continuous function f, there exits a two-layer neural network $f_m(\cdot;\theta)$ such that*

$$\sup_{x\in[0,1]} |f_m(x;\theta) - f(x)| \lesssim \frac{\mathrm{Lip}(f)}{m}.$$

Proof. Let $h = \frac{1}{m}$ and $\{x_j = jh\}_{j=0}^m$ be the uniform grids in $[0,1]$. Let $t(x) = \max(1-|x|, 0)$ be the triangular function. Then, the piecewise linear interpolator can be written as

$$\tilde{f}_m(x) = \sum_{j=0}^{m} f(x_i) t\left(\frac{x - x_i}{h}\right). \tag{1}$$

Consider the approximation error in the interval $[x_j, x_j + h]$: for $t \in [0, h]$,

$$\begin{aligned} |f(x_j + t) - \tilde{f}(x_j + t)| &= |f(x_j + t) - f(x_j) - \frac{f(x_j + h) - f(x_j)}{h} t| \\ &= |f'(\xi_1)t - f'(\xi_2)t| \lesssim \mathrm{Lip}(f) h. \end{aligned}$$

Hence,

$$\sup_{x \in [0,1]} |\tilde{f}_m(x) - f(x)| = \max_{j \in [m-1]} \sup_{t \in [0,h]} |f(x_j + t) - \tilde{f}(x_j + t)| \lesssim \mathrm{Lip}(f) h.$$

Notice that the triangular function can exactly represented with 3 ReLU neurons:

$$t(x) = \sigma(x+1) + \sigma(x-1) - 2\sigma(x).$$

Plugging it into (1), it shows that $\tilde{f}_m$ can be represented with a two-layer neural net with $3m$ neurons. □

Since the Lipschitz class is dense in $C([0,1])$, we thus prove the UAP for the ReLU activation function. For other activation functions, one can use other constructive proofs.

We remark that the seminal work [Cybenko, 1989] proved UAP only for networks activated by sigmoidal functions [1]. The function $\sigma : \mathbb{R} \mapsto \mathbb{R}$ is said to be sigmoidal if

$$\lim_{z \to -\infty} \sigma(z) = 0, \qquad \lim_{z \to \infty} \sigma(z) = 1. \tag{2}$$

A fast proof of [Cybenko, 1989] in our framework. Denote by H the Heaviside step function, which is a special sigmoidal function. Similar to the proof of Lemma 1.3, one can show that the two-layer neural network activated by H can mimic any piecewise constant function. Hence, $\mathcal{F}_{H,1}$ has UAP. Noticing that $\sigma(\beta z) \to H(z)$ as $\beta \to \infty$ if σ is sigmoidal, we have $\mathcal{F}_{\sigma,1}$ has also UAP. Applying Theorem 1.2 leads to that $\mathcal{F}_{\sigma,d}$ has UAP for any $d \geq 1$.

2 Approximation with rates

UAP does not provide any quantitative information about the approximation process. In particular, it cannot explain the superiority of neural nets over the classical methods, such as polynomials, spline, finite element methods, since all these methods also have UAP.

We first review some classical results of approximation rates.

- Approximating functions in $C(\mathcal{X})$ does not have rate. Why?

[1] The proof in [Cybenko, 1989] is quite elegant by utilizing the Hahn-Banach separation theorem.

- Lemma 1.3 can be extended to $d > 1$, where the rate is $O(\frac{1}{m^{1/d}})$. This means that to reach the accuracy ε, the number of parameters needed is ε^{-d}, which depends on the input dimension exponentially. For instance, taking $\varepsilon = 0.1, d = 20$, the number of parameters needed is 10^{20}. The issue is referred to as the *curse of dimensionality* (CoD).

- **High-order smoothness.** To obtain a faster approximation rate, we need to consider a smaller target function space. The classical approach in applied math is to impose stronger smoothness by assuming the high-order differentiability. For example, consider the Sobolev space defined by the Sobolev norm:

$$\|f\|_{H_d^s} = \left(\sum_{|\alpha| \le s} |D^\alpha f|^2 \, \mathrm{d}x \right)^{1/2} < \infty.$$

  For H_d^s, it has been shown that the minimax rate of approximating this space is $O(m^{-s/d})$ regardless what model is utilized. This rate suffers from the CoD unless $s \gtrsim d$.

The above approximation rates obtained by assuming certain (classical) smoothness on target functions all suffer from the CoD. They are quantitative but not useful in high dimensions. The successs of ML in solving high-dimensional functions implies that ML models must be able to overcome CoD for certain class of functions. Therefore, the most fundamental problem in ML is to understand:

What kind of functions can be approximated/learned by a particular ML model without CoD.

We already proved that functions in RKHS can be learned without CoD. The question in this lecture is what kind of functions can be learned efficiently by two-layer neural networks?

Avoid CoD via Monte-Carlo approximation. The Monte-Carlo method for high-dimensional integration is only example in applied math that we can avoid CoD (do we have other examples?). Hence, we anticipate similar cases also happen to the approximating high-dimensional functions.

Consider the taking limit for the **scaled** two-layer neural networks:

$$f_m(x;\theta) = \frac{1}{m} \sum_{j=1}^{m} a_j \varphi(x; v_j) \to \mathbb{E}_{(a,v)\sim\rho}[a\varphi(x;v)] = f_\rho(x), \tag{3}$$

where $\varphi(x;v) = \sigma(w^\top x + b)$ but can also take general feature functions. In this way, the two-layer network $f_m(\cdot;\theta)$ is a Monte-Carlo approximation of f_ρ with the approximation error satisfying

$$f_m(x;\theta) - f_\rho(x) \sim \frac{\mathrm{Var}_{(a,v)\sim\rho}[a^2\varphi(x,v)^2]}{\sqrt{m}}$$

This suggests that if a function f has the probabilistic representation $f(x) = \mathbb{E}_{(a,v)\sim\rho}[a\varphi(x;v)]$, then it can be approximated by Monte-Carlo discretization and the resulting model is exactly a two-layer neural network. What remains is to identify what kind of functions admit this probabilistic representation.

2.1 The Jones' trick: probabilistic representation via Fourier transform

The following procedure was first developed in [Jones, 1992]. Let $\hat{f}$ be the Fourier transform of f:

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

The Fourier inversion theorem says

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

This gives a integral representation of f and we will impose some conditions such that it can be converted into a probabilistic representation.

Let $\hat{f}(\omega) = |\hat{f}(\omega)| e^{ib(\omega)}$ be the polar decomposition of $\hat{f}(\omega)$. 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 $\gamma_0(f) = \int |\hat{f}(\omega)| \, \mathrm{d}\omega$ and let $\mathrm{d}\pi(\omega) = \frac{|\hat{f}(\omega)|}{\gamma_0(f)} \, \mathrm{d}\omega$. Then,

$$f(x) = \gamma_0(f) \, \mathbb{E}_{\omega \sim \pi}[\cos(\omega^\top x + b(\omega))]. \tag{6}$$

Thus, we represent the function as an expectation. Recall that the property of Monte-Carlo integration:

$$\mathbb{E}_{x \sim \rho}[h(x)] - \frac{1}{m} \sum_{j=1}^m h(x_j) \sim \frac{\mathrm{Var}(h)}{m},$$

where $x_1, \ldots, x_m$ are i.i.d. sampled from ρ. The following theorem shows that the similar result also hold for function approximation.

Theorem 2.1. *Let ρ be any probability distribution over $\mathbb{R}^d$. Assume $\gamma_0(f) = \int |\hat{f}(\omega)| \, \mathrm{d}\omega < \infty$, then there exists a two-layer neural net $f_m(\cdot; \theta)$ activated by the cosine function such that*

$$\|f_m(\cdot; \theta) - f\|^2_{L^2(\mathbb{P}_x)} \lesssim \frac{\gamma_0(f)^2}{m}.$$

Proof. Let $W = (\omega_1, \ldots, \omega_m)$ with $\{\omega_j\}$ being i.i.d. random variable sampled from π. Let

$$f_m(x; \tilde{\theta}) = \frac{1}{m} \sum_{j=1}^m \gamma_0(f) \cos(w_j^\top x + b(w_j)) =: \frac{1}{m} \sum_{j=1}^m Z_j.$$

Moreover,

$$\begin{aligned} \mathbb{E}_W[Z_j - f(x)] &= 0 \\ \mathbb{E}_W[(Z_j - f(x))^2] &\le \mathbb{E}_W \, Z_j^2 \le \gamma_0(f)^2. \end{aligned} \tag{7}$$

Then, using the independence of Z_j, we have

$$\begin{aligned} \mathbb{E}_W[\|f_m(\cdot; \tilde{\theta}) - f\|^2_{L^2(\mathbb{P}_x)}] &= \mathbb{E}_x \, \mathbb{E}_W \, |\frac{1}{m} \sum_{j=1}^m (Z_j - f(x))|^2 \\ &= \mathbb{E}_x \, \frac{1}{m^2} \sum_{j=1}^m \mathbb{E} \, |Z_j - f(x)|^2 \le \frac{\gamma_0(f)^2}{m}, \end{aligned}$$

where the last inequality follows from (7). □

The preceding rate is a standard Monte-Carlo rate, which is independent of d. This explains the superiority of neural networks for approximating functions with $C_f < \infty$. Note that C_f may depend on d, althoutgh the rate is not.

Unfortunately, there are still two issues.

- The cosine activation function is not often used in practice, though it is recently found effective in solving some scientific computing problems [Sitzmann et al., 2020].
- The input domain is $\mathbb{R}^d$. In practice, it is more often to consider a compact domain, e.g., the image where the pixel value lies in $[0, 1]$.

2.2 The Barron's trick

Andrew R. Barron developed some tricks in [Barron, 1993] to resolve these issues. Let Ω be a compact domain and define the dual norm

$$\|w\|_\Omega = \sup_{x\in\Omega} |w^\top x|. \tag{8}$$

Let $\hat{w} = w/\|w\|_\Omega$. A particular example is that Ω is the ℓ_p ball, for which $\|\cdot\|_\Omega$ corresponds to the ℓ_q norm with q be the Holder conjugate of p, i.e., $1/p + 1/q = 1$. In the following, the dependence of Ω will be omitted for simplicity, but we will frequently use the property that $|\hat{w}^\top x| \le 1, \forall x \in \Omega$.

Consider $f \in C(\Omega)$ and let f_e be a $L^1(\mathbb{R})$ extension of f. Since, $f(0) = \int \hat{f}_e(\omega)\,\mathrm{d}\omega$, we can express f as follows

$$\begin{aligned} f(x) - f(0) &= \int (e^{i\omega^\top x} - 1)\hat{f}_e(\omega)\,\mathrm{d}\omega \\ &= \int \frac{e^{i\omega^\top x} - 1}{\|\omega\|}\|\omega\|\hat{f}_e(\omega)\,\mathrm{d}\omega \\ &= \int \frac{\cos(\omega^\top x + b(\omega)) - \cos(b(\omega))}{\|\omega\|}\|\omega\||\hat{f}_e(\omega)|\,\mathrm{d}\omega \\ &= \int g(\omega, x)\|\omega\||\hat{f}_e(\omega)|\,\mathrm{d}\omega, \end{aligned} \tag{9}$$

where

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

Assume that

$$\tilde{\gamma}_1(f) := \int \|\omega\||\hat{f}(\omega)|\,\mathrm{d}\omega < \infty.$$

Then,

$$f(x) - f(0) = \tilde{\gamma}_1(f)\,\mathbb{E}_{\omega\sim\pi}[g(x, \omega)] = \gamma_1(f)\,\mathbb{E}_{\omega\sim\pi}[h(\hat{w}^\top x, \omega)], \tag{10}$$

where $h(t, \omega) = (\cos(\|\omega\|t + b(\omega)) - \cos(b(\omega)))/\|\omega\|$ is Lipschitz with respect to t.

Thus, we express f as an expectation and for a fixed ω, $g(x, \omega)$ only depends on $\omega^\top x$. In other words, it is essentially an one-dimensional function. Different from the Jones' expression, here $h(\cdot; \omega)$ is a nicely behaved function. What remains is to show that $h(\cdot, \omega)$ can be further expressed in an expectation form, or approximated by two-layer neural networks.

Theorem 2.2. *Assume*

$$\gamma_1(f) = \inf_{f_e|_\Omega = f} \int (1 + \|\omega\|)|\hat{f}_e(\omega)| < \infty,$$

where the infimum is taken over all the $L^1(\mathbb{R})$ extensions of f. Consider the sigmoidal activation function (2). Then, there exits a two-layer neural nets such that

$$\|f_m(\cdot;\theta) - f\|_{L^2(\rho)}^2 \lesssim \frac{\gamma_1(f)^2}{m}.$$

Proof. First, write $g(x,\omega) = h(\hat{\omega}^\top x; w)$ with $h(\cdot; w) : [-1,1] \mapsto \mathbb{R}$ given by

$$h(t;w) = \frac{\cos(\|w\|t + b(w)) - \cos(b(w))}{\|w\|},$$

for which $\sup_{t\in[-1,1]} \max\{|h(t;w)|, |h'(t;w)|\} \le 1$. Let $H(t) = 1(t \ge 1)$ be the Heaviside step function. Then,

$$\begin{aligned} h(t;w) &= h(-1) + \int_{-1}^{\top} h'(s;w)\,\mathrm{d}s \\ &= h(-1) + \int_{-1}^{1} h'(s;w) H(t-s;w)\,\mathrm{d}s, \end{aligned}$$

which means h can be represented by a two-layer neural nets activated by the step function. Plugging it into (10) yields

$$f(x) = f(0) + \tilde{\gamma}_1(f)\,\mathbb{E}_{\omega\sim\pi}[h(-1;\omega)] + 2\tilde{\gamma}_1(f)\,\mathbb{E}_{\omega\sim\pi}\,\mathbb{E}_{s\sim\mathrm{Unif}[-1,1]}[h'(s;\omega)H(\hat{\omega}^\top x - s)], \tag{11}$$

where $\tilde{\gamma}_1(f) = \int \|\omega\||\hat{f}_e(\omega)|\,\mathrm{d}\omega$. Thus, we write f in an expectation form. Using the fact that $\max\{h(-1;\omega), h'(s;\omega)\} \le 1$ and $|H(\hat{w}^\top x - s)| \le 1$. The approximation error is bounded by

$$\begin{aligned} \text{app-err} &\lesssim \frac{\tilde{\gamma}_1(f)^2 + f^2(0)}{m} \lesssim \frac{1}{m}\left((\int |\hat{f}_e(\omega)|\,\mathrm{d}\omega)^2 + (\int \|\omega\||\hat{f}_e(\omega)|\,\mathrm{d}\omega)^2\right) \\ &\lesssim \frac{1}{m}\left(\int (1+\|\omega\|)|\hat{f}_e(\omega)|\,\mathrm{d}\omega\right)^2 = \frac{\gamma_1^2(f)}{m}. \end{aligned}$$

Taking over all the $L^1(\mathbb{R})$ extension f_e, we complete the proof for the Heaviside activation function.

For general sigmoidal activation functions, the result follows from the fact that $\sigma(\beta z) \mapsto H(z)$ as $\beta \to \infty$. Moreover, noticing that the above derivation holds for any extension f_e. Hence, it must hold for the one with the smallest moment. □

2.3 An alternative Fourier analysis

2.4 Step functions

Lemma 2.3. *Suppose $h \in C^2([-1,1])$. 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. □

Applying this lemma to e^{ict}, we discover

$$e^{ict} = 0 + ic\int_0^1 e^{is}H(t-s)\,\mathrm{d}s + ic\int_0^{-1} e^{is}H(s-t)\,\mathrm{d}s. \tag{12}$$

Using this identity, we have

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

where I_2 accounts for the negative part. Hence,

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

Hence, if $\int \|\omega\||\hat{f}_e(\omega)|\,\mathrm{d}\omega < \infty$, the I_1 as well as $f(x)$ can be written as an expectation form by applying the Jones' trick.

2.5 ReLU activations

Lemma 2.4. *Suppose $h \in C^2([-1,1])$. Then, we have*

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

where σ is the ReLU function.

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

$$\begin{aligned} h(t) &= h(0) + \int_0^t h'(\tau)\,\mathrm{d}\tau \\ &= h(0) + \int_0^t \left(h'(0) + \int_0^s h''(s)\,\mathrm{d}s\right)\mathrm{d}\tau \\ &= h(0) + h'(0)t + \int_0^t\int_0^s h''(s)\,\mathrm{d}s\,\mathrm{d}\tau \end{aligned}$$

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

If $t < 0$, the proof is similar. □

Theorem 2.5. *Suppose* $\gamma_2(f) = \inf_{f_e|\Omega=f} \int (1+\|\omega\|)^2 |\hat{f}_e(\omega)|\,\mathrm{d}\omega < \infty$. *Then, there exists a two-layer ReLU network* $f_m(x;\theta)$ *such that*

$$
\mathbb{E}_x[|\sum_{j=1}^m a_j \,\mathrm{ReLU}(w_j^\top x + b_j) - f(x)|^2] \lesssim \frac{\gamma_2(f)^2}{m}.
$$

Moreover, for any $j \in [m]$, *we have*

$$
|a_j| \lesssim \frac{\gamma_2(f)}{m}, \quad \|w_j\|_\Omega \le 1, \quad |b_j| \le 1. \tag{13}
$$

Proof. Applying the above lemma to e^{ict}, we discover the following identity

$$
e^{ict} - ict - 1 = -c^2 \int_0^1 e^{ics}\sigma(t-s)\,\mathrm{d}s - c^2 \int_0^{-1} e^{ics}\sigma(-t+s)\,\mathrm{d}s. \tag{14}
$$

Then,

$$
\begin{aligned}
f(x) - \nabla f(0)^\top x - f(0) &= \int_{\mathbb{R}^d} (e^{i\omega^\top x} - i\omega^\top x - 1)\hat{f}_e(\omega)\,\mathrm{d}\omega \\
&= -\int_{\mathbb{R}^d} \int_0^1 \|\omega\|^2 \sigma(\hat{\omega}^\top x - s) e^{i\|\omega\|s}\,\mathrm{d}s \hat{f}_e(\omega)\,\mathrm{d}\omega + I_2 \\
&= \underbrace{-\int_{\mathbb{R}^d} \int_0^1 \cos(\|\omega\|t + b(\omega))\sigma(\hat{\omega}^\top x - t)\|\omega\|^2 |\hat{f}(\omega)|\,\mathrm{d}t\,\mathrm{d}\omega}_{I_1} + I_2,
\end{aligned} \tag{15}
$$

where the I_2 is similar to I_1, accounting for the case $\omega^\top x \le 0$. The explicit form of I_2 is omitted for notation simplicity. Hence, if $\int \|\omega\|^2 |\hat{f}(\omega)|\,\mathrm{d}\omega < \infty$, by using the Jones' trick, we can write (15) in an expectation form.

In addition, 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 $w = \nabla f(0)$. □

3 Generalization analysis

In this section, we assume $\Omega = \mathbb{S}^{d-1}$ for simplicity. In Lecture 12, we derive the Rademacher complexity of neural networks of the following class:

$$
\left\{ f_m(x;\theta) : \sum_{j=1}^m |a_j| \le A, \|w_j\|_2 + |b_j| \le B \right\},
$$

where the inner-layer and outer-layer weights are controlled independently. However, for ReLU networks, we only need to control the path norm

$$\|\theta\|_{\mathcal{P}} := \sum_{j=1}^{m} |a_j|(\|w_j\|_2 + |b_j|) \tag{16}$$

because of the positive homogeneity of ReLU. Specifically, we have

$$\begin{aligned}\mathcal{F}_Q &= \{f_m(x;\theta) : \|\theta\|_{\mathcal{P}} \le Q\} \\ &= \left\{ f_m(\cdot;\theta) : \sum_{j=1}^{m} |a_j| \le Q, \|w_j\| + |b_j| = 1 \text{ for } j = 1, 2, \ldots, m \right\}. \end{aligned} \tag{17}$$

Proposition 3.1. $\widehat{\mathrm{Rad}}_n(\mathcal{F}_Q) \lesssim Q/\sqrt{n}$

Proof. Follow exactly the proof of Lemma 4.9 in Lecture 12. □

The regularized estimator. Let the empirical risk

$$\hat{\mathcal{R}}_n(\theta) = \frac{1}{2}\sum_{i=1}^{n}(f_m(x_i;\theta) - f^*(x_i))^2.$$

Consider the path norm-regularized estimator:

$$\hat{\theta}_n = \operatorname*{argmin}_{\theta} \hat{\mathcal{R}}_n(\theta) + \frac{\lambda}{\sqrt{n}}\|\theta\|_{\mathcal{P}}. \tag{18}$$

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*

$$\mathcal{R}(\hat{\theta}_n) \lesssim \frac{\gamma_2^2(f^*)}{m} + \frac{\gamma_2(f^*)}{\sqrt{n}} + \sqrt{\frac{\log(1/\delta)}{n}}.$$

- 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 $Q = \gamma_2(f^*)$.

(1) By Theorem 2.5, there exits $\tilde{\theta}$ such that

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

By definition,

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

Hence,

$$\begin{aligned}\|\hat{\theta}_n\|_{\mathcal{P}} &\le 2Q + \frac{3Q^2\sqrt{n}}{\lambda m} =: C(m,\lambda,Q)\\ \hat{\mathcal{R}}_n(\hat{\theta}_n) &\le \frac{3Q^2}{m} + \frac{2\lambda}{\sqrt{n}}Q.\end{aligned} \tag{19}$$

(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,

$$\widehat{\mathrm{Rad}}_n(\mathcal{H}_C) \le 2\widehat{\mathrm{Rad}}_n(\mathcal{F}_C). \tag{20}$$

By (32), $\hat{f}_m(\cdot;\hat{\theta}_n) \in \mathcal{F}_{C(m,\lambda,Q)}$.

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

$$\begin{aligned}\mathcal{R}(\hat{\theta}_n) &\le \hat{\mathcal{R}}(\hat{\theta}_n) + 2\widehat{\mathrm{Rad}}_n(\mathcal{H}_{C(m,\lambda,Q)}) + \sqrt{\frac{\log(2/\delta)}{n}}\\ &\le \hat{\mathcal{R}}(\hat{\theta}_n) + 4\widehat{\mathrm{Rad}}_n(\mathcal{F}_{C(m,\lambda,Q)}) + \sqrt{\frac{\log(2/\delta)}{n}} && \text{(Use Eq.(33))}\\ &\lesssim \hat{\mathcal{R}}(\hat{\theta}_n) + \frac{C(m,\lambda,Q)}{\sqrt{n}} + \sqrt{\frac{\log(2/\delta)}{n}} && \text{(Use Prop.7.1 and Eq.(32))}\\ &\le \frac{3Q^2}{m} + \frac{2\lambda}{\sqrt{n}}Q + \frac{1}{\sqrt{n}}\left(2Q + \frac{3Q^2\sqrt{n}}{\lambda m}\right) + \sqrt{\frac{\log(2/\delta)}{n}} && \text{(Use Eq.(32))}\\ &\lesssim \frac{Q^2}{m} + \frac{Q}{\sqrt{n}} + \sqrt{\frac{\log(2/\delta)}{n}}.\end{aligned}$$

□

4 A brief overview

Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact set. For $f : \mathcal{X} \mapsto \mathbb{R}$, define

$$\|f\|_{\mathbb{F}_2} = \inf_{f_e|_{\mathcal{X}}=f} \int_{\mathbb{R}^d} (1 + \|\omega\|_{\mathcal{X}}^2)|\hat{f}_e(\omega)|\,\mathrm{d}\omega, \tag{21}$$

where $\|\omega\|_{\mathcal{X}} = \sup_{x\in\mathcal{X}} |\omega^\top x|$.

Assumption 4.1. *Throughout this lecture, we let $\mathcal{X} = [0,1]^d$ and $\|\cdot\|_{\mathcal{X}} = \|\cdot\|_1$.*

We have proved that if $\|f\|_{\mathbb{F}_2} < \infty$, then f can be expressed in an expectation form:

$$f(x) - f(0) - \nabla f(0)\cdot x = \mathbb{E}_{(a,w,b)\sim\rho}[a\sigma(w^\top x + b)], \quad x \in \mathcal{X}, \tag{22}$$

with $|a|(\|w\|_1 + |b|) \lesssim \|f\|_{\mathbb{F}_2}$ for any $(a,w,b) \in \mathbb{R}^{d+2}$. Here σ is the ReLU activation function. A direct consequence of this expectation-form expression is that f can be approximated by two-layer ReLU nets without CoD.

Theorem 4.2. *Suppose $\|f\|_{\mathbb{F}_2} < \infty$ and $f(0) = 0, \nabla f(0) = 0$. Then, there exists a two-layer ReLU nets $f(x;\theta) = \frac{1}{m}\sum_{j=1}^m a_j\sigma(w_j^\top x + b_j)$ such that*

$$\|f(\cdot;\theta) - f\|_{L^2(\mu)} \lesssim \frac{\|f\|_{\mathbb{F}_2}}{\sqrt{m}} \tag{23}$$

$$\|\theta\|_{\mathcal{P}}^2 := \frac{1}{m}\sum_{j=1}^m |a_j|^2(\|w_j\|_{\mathcal{X}} + |b_j|)^2 \lesssim \|f\|_{\mathbb{F}_2}^2. \tag{24}$$

Here $\|\cdot\|_{\mathcal{P}}$ is known as the *path norm* of the network, which is the sum of norms of all paths.

- It is worth noting that the control of path norm of that approximator is important for obtaining the estimation error, as shown later.

5 The Barron space

We ask the question: Is the spectral Barron norm (21) tight in characterizing the "efficient" approximation of two-layer neural nets? Unfortunately, it is not. A counter example is given by the triangular function

Lemma 5.1. *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. □

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

$$f_\pi(x) = \mathbb{E}_{(a,w)\sim\pi}[a\sigma(w^\top x)]. \tag{25}$$

Here we omit the bias term for brevity. This can be viewed as an infinitely-wide two-layer net. It is the continuum limit of the *scaled* two-layer neural net:

$$f_m(x;\theta) = \frac{1}{m}\sum_{j=1}^m a_j\sigma(w_j^\top x). \tag{26}$$

For any f that admit the representation (25), the representation π is usually not not unique. Define

$$R_f = \left\{\pi \in \mathcal{P}(\mathbb{R}^1 \otimes \mathbb{R}^d) \,:\, f_\pi(x) = \mathbb{E}_{(a,w)\sim\pi}[a\sigma(w^\top x)]\right\}. \tag{27}$$

Definition 5.2 (The Barron space)**.** Assume that σ is ReLU. Let

$$\|f\|_{\mathcal{B}}^2 := \inf_{\pi \in R_f} \mathbb{E}_{(a,w)\sim\pi}[|a|^2 \|w\|_1^2]. \tag{28}$$

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

- 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 taking-infimum step in (28) is essential. First, it makes 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:

  $$\sigma(x_1) = \sigma(x_1) + r\sigma(x_2) - r\sigma(x_2). \tag{29}$$

  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 $(1 + 2r^2)^{1/2}$. 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.

Examples of Barron functions.

- We have shown that $\|f\|_{\mathcal{B}} \lesssim \|f\|_{\mathbb{F}_2}$. This contains a lot of functions.
- General functions with a linear low-dimensional structure: $f(x) = g(W^\top x)$ with $g : \mathbb{R}^k \mapsto \mathbb{R}$. Obviously,

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

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

6 Capacity-Controlled Approximation

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

$$\|\theta\|_{\mathcal{P}} := \frac{1}{m}\sum_{j=1}^m |a_j|^2 \|w_j\|_1^2. \tag{30}$$

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

Theorem 6.1 (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*

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

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,

$$\begin{aligned}
\mathbb{E}_{(a_j,w_j)}\,\mathbb{E}_x|\frac{1}{m}\sum_{j=1}^m a_j\sigma(w_j\cdot x) - f(x)|^2 &= \mathbb{E}_x\,\mathbb{E}_{(a_j,w_j)}\,|\frac{1}{m}\sum_{j=1}^m a_j\sigma(w_j\cdot x) - f(x)|^2 \\
&= \mathbb{E}_x\,\frac{1}{m}\sum_{j=1}^m \mathbb{E}_{(a_j,w_j)}\,|a_j\sigma(w_j\cdot x) - f(x)|^2 \qquad \text{(Use the independence of } (a_j, w_j)\text{)} \\
&\leq \mathbb{E}_x\,\frac{1}{m}\sum_{j=1}^m \mathbb{E}_{(a_j,w_j)}\,a_j^2\sigma(w_j\cdot x)^2 \leq \frac{1}{m}\sum_{j=1}^m \mathbb{E}_{(a_j,w_j)}\,a_j^2\|w_j\|_1^2 \\
&\leq \frac{2\|f\|_{\mathcal{B}}^2}{m}.
\end{aligned}$$

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.

7 Generalization analysis

Proposition 7.1. *Let $\mathcal{F}_Q = \{f \in \mathcal{B} : \|f\|_{\mathcal{B}} \leq Q\}$. Then,*

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

Proof. By definition, there exist ρ such that $f(x) = \mathbb{E}_\rho[a\sigma(w^\top x)]$ for all $x \in \mathcal{X}$ and $\mathbb{E}_\rho[|a|^2\|w\|_1^2] < \|f\|_{\mathcal{B}}^2$. By Cauchy-Schwart inequality, we have $\mathbb{E}_\rho[|a|\|w\|_1] \leq \sqrt{\mathbb{E}_\rho[|a|^2\|w\|^2]} \leq \|f\|_{\mathcal{B}}$.

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

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

$$= 2Q\,\mathbb{E}_\xi[\sup_{\|w\|_1\le 1}\sum_{i=1}^n \xi_i\sigma(w^\top x_i)] \qquad \text{(Use the symmetry of } \xi\text{)}$$

$$\le 2Q\,\mathbb{E}_\xi[\sup_{\|w\|_1\le 1}\sum_{i=1}^n \xi_i w^\top x_i] \qquad \text{(Use the contraction lemma).}$$

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

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

$$\hat{\theta}_n = \operatorname*{argmin}_\theta \hat{R}_n(\theta) + \frac{\lambda}{\sqrt{n}}\|\theta\|_{\mathcal{P}}. \tag{31}$$

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 7.2. *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,

$$\|\hat{\theta}_n\|_{\mathcal{P}} \le 2Q + \frac{3Q^2\sqrt{n}}{\lambda m} =: C(m,\lambda,Q)$$
$$\hat{R}_n(\hat{\theta}_n) \le \frac{3Q^2}{m} + \frac{2\lambda}{\sqrt{n}}Q. \tag{32}$$

(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,

$$\widehat{\mathrm{Rad}}_n(\mathcal{H}_C) \le 2\widehat{\mathrm{Rad}}_n(\mathcal{F}_C). \tag{33}$$

By (32), $\hat{f}_m(\cdot;\hat{\theta}_n) \in \mathcal{F}_{C(m,\lambda,Q)}$.

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

$$
\begin{aligned}
\mathcal{R}(\hat{\theta}_n) &\leq \hat{\mathcal{R}}(\hat{\theta}_n) + 2\widehat{\mathrm{Rad}}_n(\mathcal{H}_{C(m,\lambda,Q)}) + \sqrt{\frac{\log(2/\delta)}{n}} \\
&\leq \hat{\mathcal{R}}(\hat{\theta}_n) + 4\widehat{\mathrm{Rad}}_n(\mathcal{F}_{C(m,\lambda,Q)}) + \sqrt{\frac{\log(2/\delta)}{n}} && \text{(Use Eq.(33))} \\
&\lesssim \hat{\mathcal{R}}(\hat{\theta}_n) + \frac{C(m,\lambda,Q)}{\sqrt{n}} + \sqrt{\frac{\log(2/\delta)}{n}} && \text{(Use Prop.7.1 and Eq.(32))} \\
&\leq \frac{3Q^2}{m} + \frac{2\lambda}{\sqrt{n}}Q + \frac{1}{\sqrt{n}}\left(2Q + \frac{3Q^2\sqrt{n}}{\lambda m}\right) + \sqrt{\frac{\log(2/\delta)}{n}} && \text{(Use Eq.(32))} \\
&\lesssim \frac{Q^2}{m} + \frac{Q}{\sqrt{n}} + \sqrt{\frac{\log(2/\delta)}{n}}.
\end{aligned}
$$

□

8 Final remarks

We present a function space viewpoint for understanding two-layer neural networks. Similar approaches can be extended to many other neural network models. We refer interested readers to https://leiwu0.github.io/teach/pku-summer2021/lecture-note/lec-7.pdf for more details.

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.

[Cybenko, 1989] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

[Jones, 1992] 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.

[Siegel and Xu, 2020] Siegel, J. W. and Xu, J. (2020). Approximation rates for neural networks with general activation functions. *Neural Networks*, 128:313–321.

[Sitzmann et al., 2020] Sitzmann, V., Martel, J., Bergman, A., Lindell, D., and Wetzstein, G. (2020). Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473.