Mathematical Theory of Neural Network Models

Lecture 5: Error Estimates and Implicit Regularization

July 11

Lecturer: Lei Wu *Scribe: Ziheng Yan, Jinghang Chai, Chuhan Xie*

1 Error estimates for regularized random feature model

We denote

$$f(x;a) = \sum_{k=1}^{m} a_k \phi(x, w_k^0)$$

as the estimate function, and

$$f^*(x) = \int a(w)\phi(x,w)\,\mathrm{d}\pi(w)$$

as the target function. Then we take some steps for preparation:

1. Assume $|f^*| \leq 1$. Otherwise we denote $f := T \circ f = \min\{\max\{f,-1\},1\} \in [-1,1]$;
2. Define $\gamma(f^*) := max\{1, \sup_w a^*(w)\} < \infty$, and then $\|f^*\|_{\mathcal{H}_k} = \sqrt{\mathbb{E}_w(a^*(w))^2} \leq \gamma(f^*)$.

According to the approximation method taught in the last lecture, with probability $1-\delta$, $\exists a^*$, such that

$$\mathbb{E}_x \left(f(x;a^*) - f^*(x)\right)^2 \leq \frac{\gamma^2(f^*)}{m} C(\delta),$$

where $C(\delta) = 1 + \sqrt{log(\frac{1}{\delta})}$.

In the following discussion, we denote $a_k^* = \frac{a^*(w_k^0)}{m}$; and we assume that certain constants bares little importance in our discussion on error estimation, so sometimes our proposed inequalities hold true up to a constant.

We start by estimating the bound of $\|a^*\|$.

$$\|a^*\|^2 = \sum_{k=1}^{m} \frac{(a^*(w_k^0))^2}{m^2} \approx \frac{1}{m}\int a^*(w)^2\,\mathrm{d}\pi(w) \leq \frac{\gamma^2(f^*)}{m},$$

so

$$\|a^*\| \leq \frac{\gamma(f^*)}{\sqrt{m}} \Leftrightarrow \sqrt{m}\|a^*\| \leq \gamma(f^*). \tag{1}$$

Therefore, $\gamma(f^*)$ can control $\|a^*\|$.

Assume

$$\mathcal{F}_Q = \left\{ f(x;a) : \sqrt{m}\|a\| \leq Q \right\},$$

then

$$\begin{aligned}
Rad_n(\mathcal{F}_Q) &= \frac{1}{n}\mathbb{E}_\xi \sup_{\sqrt{m}\|a\|\le Q} \sum_{i=1}^{n} \xi_i \sum_{k=1}^{m} a_k \phi(x_i; w_k^0) \\
&= \frac{1}{n}\mathbb{E}_\xi \sup_{\sqrt{m}\|a\|\le Q} \sum_{k=1}^{m} a_k \sum_{i=1}^{n} \xi_i \phi(x_i; w_k^0) \\
&\le \frac{1}{n}\mathbb{E}_\xi \sup_{\sqrt{m}\|a\|\le Q} \sqrt{\sum_{k=1}^{m} a_k^2} \sqrt{\sum_{k=1}^{m} \left(\sum_{i=1}^{n} \xi_i \phi(x_i; w_k^0)\right)^2} \\
&\le \frac{Q}{n\sqrt{m}} \sqrt{\mathbb{E}_\xi \sum_{k=1}^{m} \left(\sum_{i=1}^{n} \xi_i \phi(x_i; w_k^0)\right)^2} \\
&= \frac{Q}{n\sqrt{m}} \sqrt{\sum_{k=1}^{m} \sum_{i=1}^{n} \mathbb{E}_\xi(\xi_i^2) \phi^2(x_i; w_k^0)} \\
&\le \frac{Q}{n\sqrt{m}} \sqrt{mn},
\end{aligned}$$

and we find an upper bound of $Rad_n(\mathcal{F}_Q)$:

$$Rad_n(\mathcal{F}_Q) \le \frac{Q}{\sqrt{n}}. \tag{2}$$

Then we can estimate the generation gap:

$$\begin{aligned}
|L(\hat{a}_n) - L_n(\hat{a}_n)| &\le \sup_{\sqrt{m}\|a\|\le Q} |L(a) - L_n(a)| \\
&\le 2Rad_n(\mathcal{H}_Q) + \sqrt{\frac{log(\frac{1}{\delta})}{n}},
\end{aligned} \tag{3}$$

where

$$\mathcal{H}_Q = \left\{ l\left(f(x; a), f^*(x)\right) : \sqrt{m}\|a\| \le Q \right\},$$
$$Rad_n(\mathcal{H}_Q) \le Lip(l) Rad_n(\mathcal{F}_Q).$$

Here $Lip(l)$ is a Lipschitz constant of l. We assume that l is first-order Lipschitz continuous. Then

$$Rad_n(\mathcal{H}_Q) \le Rad_n(\mathcal{F}_Q). \tag{4}$$

Lemma 1.1. *For any $\delta > 0$, with probability $1 - \delta$ over the sampling S, $\forall a$,*

$$|L(a) - L_n(a)| \le 2Rad_n(\mathcal{H}_{\sqrt{m}\|a\|+1}) + \sqrt{\frac{2log(\sqrt{m}\|a\| + 1)^2/\delta}{n}}.$$

Proof. It evident that $\mathcal{H} = \bigcup_{l=1}^{\infty} \mathcal{H}_l$, where $\mathcal{H}$ is the hypothesis space and $\mathcal{H}_l$ is defined similarly as $\mathcal{H}_Q$. We define $\delta_l = \frac{\delta}{Cl^2}$ and $C = \sum_{l=1}^{\infty} \frac{1}{l^2}$, and then $\sum_{l=1}^{\infty} \delta_l = \delta$. We know that with probability $1 - \delta_l$,

$$|L(a) - L_n(a)| \leq 2Rad_n(\mathcal{H}_l) + \sqrt{\frac{log(1/\delta_l)}{n}}.$$

Let $l_0 = \min\{l \in \mathbb{N}_+ : \sqrt{m}\|a\| \leq l\}$, and then $l_0 \leq \sqrt{m}\|a\| + 1 \Rightarrow Rad_n(\mathcal{H}_{l_0}) \leq 2Rad_n(\mathcal{H}_{\sqrt{m}\|a\|+1})$. Therefore, we have

$$\begin{aligned} |L(a) - L_n(a)| &\leq 2Rad_n(\mathcal{H}_{\sqrt{m}\|a\|+1}) + \sqrt{\frac{log(l^2/\delta)}{n}} \\ &\leq 2Rad_n(\mathcal{H}_{\sqrt{m}\|a\|+1}) + \sqrt{\frac{2log(\sqrt{m}\|a\|+1)^2/\delta}{n}}. \end{aligned} \tag{5}$$

Let S_l be the set in which (5) do not hold true, then

$$\mathbb{P}\left(\left(\bigcup_{l=1}^{\infty} S_l\right)^c\right) \geq 1 - \sum_l \mathbb{P}(S_l) = 1 - \sum_l \delta_l = 1 - \delta.$$

Hence, we have proved the lemma. □

Define $\hat{a}_n = \arg\min_a\{L_n(a) + \lambda\|a\|\}$, $\lambda = \frac{\sqrt{m}}{\sqrt{n}}t$, $t \geq 1$. Then

$$\begin{aligned} L(\hat{a}_n) &\leq L_n(\hat{a}_n) + \frac{2(\sqrt{m}\|\hat{a}_n\|+1)}{\sqrt{n}} + \sqrt{\frac{log(\sqrt{m}\|\hat{a}_n\|+1)^2/\delta}{n}} \\ &\leq L_n(\hat{a}_n) + \lambda\|\hat{a}_n\| + \sqrt{\frac{log(\sqrt{m}\|\hat{a}_n\|+1)^2/\delta}{n}} + \frac{1}{\sqrt{n}} \\ &\leq L_n(\tilde{a}_n^*) + \lambda\|\tilde{a}_n^*\| + \sqrt{\frac{log(\sqrt{m}\|\hat{a}_n\|+1)^2/\delta}{n}} + \frac{1}{\sqrt{n}} \\ &\leq L(a^*) + \frac{\sqrt{m}t\|a^*\|+1}{\sqrt{n}} + \sqrt{\frac{log(\sqrt{m}\|a^*\|+1)^2/\delta}{n}} + \lambda\|a^*\| + Q_n \\ &\leq \frac{\gamma^2(f^*)}{m} + \frac{\gamma(f^*)}{\sqrt{n}}(t+1) + \sqrt{\frac{log(\gamma(f^*)+1)^2/\delta}{n}} + Q_n, \end{aligned}$$

where $Q_n = \sqrt{\frac{log(\sqrt{m}\|\hat{a}_n\|+1)^2/\delta}{n}}$. The last inequality holds true because $\|a^*\| \leq \frac{\gamma(f^*)}{\sqrt{m}}$. Then we are going to estimate Q_n.

Because

$$\begin{aligned} \sqrt{m}\|\hat{a}_n\| &\leq \left(\frac{L_n(a^*)}{\lambda} + \|a^*\|\right)\sqrt{m} \\ &\leq \frac{\sqrt{n}}{t}\left(\frac{\gamma^2(f^*)}{m} + \gamma(f^*)\right) \\ &\leq \frac{\sqrt{n}}{t}C, \end{aligned} \tag{6}$$

so

$$Q_n \leq \sqrt{\frac{log(n^2/\delta)}{n}}.$$

The first inequality in (6) is based on the definition of $\hat{a}_n$, and the second is based on the estimation on population risk taught in the last lecture.

Finally, we derive an upper bound of $L(\hat{a}_n)$: with probability $1-\delta$,

$$L(\hat{a}_n) \leq \frac{\gamma^2(f^*)}{m} + (1+t)\frac{\gamma(f^*)}{\sqrt{n}} + \sqrt{\frac{log(n^2/\delta)}{n}} + \sqrt{\frac{log(1/\delta)}{n}}\gamma(f^*), \tag{7}$$

which is called *explicit regularization*.

2 Error estimates for kernel methods with implicit regularization

In this section we estimate errors with implicit regularization. We have

$$L_n(a) = \frac{1}{n}\sum_{i=1}^{n}(\sum_{k=1}^{m} a_k \phi(x_i; w_k^0) - y_i)^2 = \|\Phi a - Y\|^2,$$

where $\Phi = (\phi(x_i; w_j^0))_{i,j}$.

We use gradient descent method to minimize $L_n(a)$. Consider the following differential equation:

$$\begin{aligned} \dot{a}_t &= -\nabla L_n(a) \\ &= -\Phi^T(\Phi a - Y), \end{aligned} \tag{8}$$

where $\Phi = U\Sigma V^T \in \mathbb{R}^{n\times m}$,$U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{m\times n}$, $\Sigma = diag\{\sigma_1, \sigma_2, ...\sigma_n\} \in \mathbb{R}^{n\times n}$. Then (8) turns into

$$\dot{a}_t = -V\Sigma^2 V^T a_t + V\Sigma U^T Y.$$

Define

$$\alpha = V^T a_t,$$

and then

$$\dot{\alpha}_t = -\Sigma^2 \alpha_t + \Sigma \tilde{Y}, \tag{9}$$

where

$$\tilde{Y} = U^T Y.$$

We can easily derive the solution of (9):

$$\alpha_i(t) = e^{-\sigma_i^2 t}\alpha_i(0) + \int_0^t e^{-\sigma^2 s}\,\mathrm{d}s \cdot \sigma_i \tilde{y}_i,$$

$$\alpha(\infty) = \Sigma^{-1} U^T Y.$$

Now we decompose a into two parts, one parallel to $span\{V_1, V_2, ..., V_n\}$, where V_i represents the i-th row in V, another perpendicular to it. When $t \to \infty$,

$$
\begin{aligned}
a &= a^{\parallel} + a^{\perp} \\
&= V\alpha + a^{\perp} \\
&= V\Sigma^{-1}U^T Y + \mathbb{P}_{V^\perp}(a_0) \\
&= \Phi(\Phi\Phi^T)^{-1}Y + \mathbb{P}_{V^\perp}(a_0) \\
&= a^* + a_0^{\perp}.
\end{aligned}
\tag{10}
$$

Define the minimum norm solution:

$$\hat{a} = \arg\min_{\Phi a = Y} \|a\|^2,$$

with constraint

$$\sqrt{m}\|\hat{a}\| \leq C\gamma(f^*).$$

We can see the perpendicular term $a^{\perp}$ remains unchanged, and the parallel term $a^{\parallel}$ shrinks and converges to that of the target function.

Lemma 2.1. *$\forall a$,*

$$\hat{L}_n(a_t) \leq L_n(a^*) + \frac{\|a_0 - a\|^2}{2t},$$

$$\|a_t - a^*\|^2 \leq \|a_0 - a^*\|^2 + 2t\hat{L}_n(a^*).$$

Specifically, if a^* is the minimum norm solution, then

$$\|a_t - a^*\| \leq \|a_0 - a^*\|.$$

We suppose $a_0 = 0$ in case that $a^{\perp}$ is small, then

$$\|a_t\| \leq 2\|a^*\|,$$

$$L(a_t) \leq L_n(a_t) + \frac{\sqrt{m}\|a_t\|}{\sqrt{n}}.$$

Proof. Define

$$J(t) = t(L_n(a_t) - L_n(a^*)) + \frac{1}{2}\|a_t - a^*\|^2,$$

then

$$
\begin{aligned}
\frac{\mathrm{d}J(t)}{\mathrm{d}t} &= L_n(a_t) - L_n(a^*) + t\langle \nabla L_n(a_t), -\nabla L_n(a_t)\rangle + \langle a_t - a^*, -\nabla L_n(a_t)\rangle \\
&= L_n(a_t) - L_n(a^*) + \langle a^* - a_t, \nabla L_n(a_t)\rangle - t\|\nabla L_n(a_t)\|^2 \\
&\leq 0.
\end{aligned}
$$

This implies $J(t) \leq J(0)$, a.e.

$$t(L_n(a_t) - L_n(a^*)) + \frac{1}{2}\|a_t - a^*\|^2 \leq \frac{1}{2}\|a_0 - a^*\|^2. \tag{11}$$

From (11) we can easily prove the lemma. $\square$

Now we let $a_0 = 0$, then

$$\begin{aligned} L(a_t) &\le |L(a_t) - L_n(a_t)| + L_n(a_t) \\ &= gen(a_t) + L_n(a_t) \\ &\le gen(a_t) + L_n(a^*) + \frac{\|a^*\|^2}{2t} \\ &\le \frac{\sqrt{m}\|a_t\|}{\sqrt{n}} + \sqrt{\frac{log(\sqrt{m}\|a_t\| + 1)^2/\delta}{n}} + \frac{\|a^*\|^2}{2t}. \end{aligned} \tag{12}$$

From *Lemma 2.1* we know that $\|a_t\| \le 2\|a^*\| + tL_n(a^*)$, so $\|a_t\| \le \frac{1}{\sqrt{m}} + t\left(\frac{1}{m} + \frac{1}{\sqrt{n}}\right)$. Then

$$L(a_t) \le \frac{t}{\sqrt{n}}\left(\frac{1}{\sqrt{m}} + \sqrt{\frac{m}{n}}\right) + \frac{1}{mt} + O\left(\frac{1}{\sqrt{n}} + \frac{1}{m}\right) + \sqrt{\frac{log\frac{\left(1+t\left(\frac{1}{\sqrt{m}}+\sqrt{\frac{m}{n}}\right)\right)^2}{\delta}}{n}}.$$

Take $T = \frac{\sqrt{n}}{m}$, then

$$\begin{aligned} L(a_T) &\le \frac{1}{m}\left(\frac{1}{\sqrt{m}} + \sqrt{\frac{m}{n}}\right) + \frac{1}{\sqrt{n}} + O\left(\frac{1}{\sqrt{n}} + \frac{1}{m}\right) + \sqrt{\frac{log(n/\delta)}{n}} \\ &\le O\left(\frac{1}{\sqrt{n}}\right) + O\left(\frac{1}{\sqrt{n}} + \frac{1}{m}\right) + \sqrt{\frac{log(n/\delta)}{n}}. \end{aligned} \tag{13}$$

Hence, we derive an upper bound of $L(a_T)$.

3 Two-layer neural network and Barron space

Now we consider a two-layer neural network, where the estimate function is

$$f(x) = \frac{1}{m}\sum_{k=1}^{m} a_k \sigma(b_k^T x), \qquad b_k \sim \pi(\cdot),$$

and the function space is

$$\Phi_f = \left\{ f(x) : f(x) = \int a(w)\sigma(w^T x)\, \mathrm{d}\pi(w) \right\}.$$

We define

$$\|f\|_{\mathcal{B}_p} = \inf_{(a,\pi)\in\Phi_f} \left(\int |a(w)|^p \, \mathrm{d}\pi(w) \right)^{\frac{1}{p}},$$

and thus

$$\|f\|_{\mathcal{B}_2}^2 = \inf_{(a,\pi)\in\Phi_f} \left(\int a^2(w)\, \mathrm{d}\pi(w) \right).$$

Define *Barron space*

$$\mathcal{B}_2 = \left\{ f \in C(X) : \|f\|_{\mathcal{B}_2} < +\infty \right\}, \qquad X = [-1, 1]^d.$$

Theorem 3.1.
$$\mathcal{B}_2 = \bigcup_{\pi} \mathcal{H}_{k_\pi},$$
where $k_\pi(x, x') = \int \sigma(w^T x)\sigma(w^T x') \,\mathrm{d}\pi(w)$ *and* $\mathcal{H}_{k_\pi}$ *is the reproducing kernel Hilbert space generated by* k_π*.*

Proof. $\forall f \in \mathcal{H}_{k_\pi}$,
$$\int a^2(w) \,\mathrm{d}\pi(w) < +\infty \Rightarrow \|f\|_{\mathcal{B}_2} < +\infty \Rightarrow f \in \mathcal{B}_2,$$
so $\bigcup_\pi \mathcal{H}_{k_\pi} \subset \mathcal{B}_2$;

$\forall f \in \mathcal{B}_2$, $\exists \tilde{\pi}$, such that
$$\int a^2(w) \,\mathrm{d}\tilde{\pi}(w) < 2\|f\|^2_{\mathcal{B}_2} < +\infty,$$
so $f \in \mathcal{H}_{k_{\tilde{\pi}}} \Rightarrow \mathcal{B}_2 \subset \bigcup_\pi \mathcal{H}_{k_\pi}$.

Therefore, we have proved $\mathcal{B}_2 = \bigcup_\pi \mathcal{H}_{k_\pi}$. □

Theorem 3.2. *$\mathcal{B}_2$ is a Barron space, and $f : X \to \mathbb{R}$ is a function in $\mathcal{B}_2$. Then*
$$\|f\|_{\mathcal{B}_2} \le \inf_{F|_X = f} \int \|w\|_1^2 |\hat{F}(w)| \,\mathrm{d}w < \infty.$$

References

[1] A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. Foundations of Computational Mathematics, 7(3):331–368, Jul 2007.

[2] L. Carratino, A. Rudi, and L. Rosasco. Learning with SGD and Random Features. arXiv e-prints, page arXiv:1807.06343, Jul 2018.

[3] W. E, C. Ma, and L. Wu. A Priori Estimates for Two-layer Neural Networks. arXiv e-prints, page arXiv:1810.06397, Oct 2018.

[4] W. E, C. Ma, and L. Wu. Barron spaces and the compositional function spaces for neural network models, 2019.

[5] W. E, C. Ma, and L. Wu. A comparative analysis of the optimization and generalization property of two-layer neural network and random feature models under gradient descent dynamics, 2019.