S&DS 671 (Spring 2019) Selected Topics in Neural Nets **Lecture: 7**

Towards Understanding the Role of Over-Parametrization in Generalization of Neural Networks

Lecturer: Harrison Zhou *Scribe: Cynthia Yue, Feb 27, 2019*

1 Introduction

In their 2018 paper *Towards Understanding the Role of Over-Parametrization in Generalization of Neural Networks*, Neyshabur et al. introduce a way to understand why neural networks generalize better with over-parametrization. They show that the difference between test and training error keeps decreasing as the number of parameters is increased. Specifically, they find tighter upper and lower bounds for the difference between test and training error. To do this, they involve the Rademacher complexity in their upper bound and are able to find a matching lower bound for the Rademacher complexity, therefore showing not only that this complexity decreases to 0 but also the speed at which it decreases.

They prove these bounds for the particular model of a two layer ReLU network, where $\mathrm{V} \in \mathbb{R}^{1\mathrm{x}h}$, $\mathrm{U} \in \mathbb{R}^{h\mathrm{x}d}$, and $\mathrm{x} \in \mathbb{R}^{d\mathrm{x}1}$:

$$\begin{aligned} f_{\mathrm{V,U}}(\mathrm{x}) &= \mathrm{V}[\mathrm{Ux}]_+ \\ &= (v_1 \ \cdots \ v_h)\left(\begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_h \end{pmatrix} \mathrm{x}\right)_+. \end{aligned}$$

They consider using this network for the classification task with data (x_i, y_i) i.i.d. with $i = 1, \ldots, m$ where the label $y_i = \pm 1$. For this classification, they use the loss function of ramp loss:

$$l(f(\mathrm{x}), y) = \begin{cases} 0 & f(\mathrm{x})y \geq 1 \\ 1 - f(\mathrm{x})y & 0 \leq f(\mathrm{x})y \leq 1 \\ 1 & f(\mathrm{x})y < 0. \end{cases}$$

They have expected loss $L(f(\mathrm{x}), y) = \mathbb{E}(l(f(\mathrm{x}), y))$ and empirical loss $\hat{L}(f(\mathrm{x}), y) = \frac{1}{m}\sum_{i=1}^{m} l(f(\mathrm{x}_i), y_i)$.

When finding bounds for the difference between test and training error, they look at the difference between test error $L(f(\mathrm{x}), y)$ and training error $\hat{L}(f(\mathrm{x}), y)$: $L(f(\mathrm{x}), y) - \hat{L}(f(\mathrm{x}), y)$.

2 Upper Bound

For the upper bound, they find the bound with respect to the following parameter space:

$$\mathcal{W} = \{(\mathrm{V,U}) : \|\mathrm{v}_j\| \leq \alpha_j, \|\mathrm{u}_j - \mathrm{u}_j^0\|_2 \leq \beta_j\}.$$

The upper bound for this parameter space is found to be:

$$L(f(\mathrm{x}), y) - \hat{L}(f(\mathrm{x}), y) \lesssim \frac{1}{m}\sum_{j=1}^{h} \alpha_j(\beta_j\|\mathrm{X}\|_F + \|\mathrm{u}_j^0\mathrm{X}\|_2) + \sqrt{\frac{\log\frac{2}{s}}{m}}. \tag{1}$$

In order to find this upper bound, there are two parts. We first show that the the difference between test and training error is bounded above by the Rademacher complexity. We then show that the Rademacher complexity is bounded above by (1).

2.1 Difference between test and training error bounded above by Rademacher complexity

In order to show $L(f(\mathrm{x}), y) - \hat{L}(f(\mathrm{x}), y) \lesssim \frac{1}{m} \sum_{j=1}^{h} \alpha_j(\beta_j \|\mathrm{X}\|_F + \|\mathrm{u}_j^0 \mathrm{X}\|_2) + \sqrt{\frac{\log \frac{2}{s}}{m}}$ (1), we first need to show that the difference between test and training error is bounded above by the Rademacher complexity:

$$L(f(\mathrm{x}), y) - \hat{L}(f(\mathrm{x}), y) \lesssim \frac{1}{m} \mathbb{E} \sup_{l} \sum_{i=1}^{m} \xi_i l(f(\mathrm{x}_i), y_i) + \sqrt{\frac{\log \frac{2}{s}}{m}}, \tag{2}$$

where $\frac{1}{m} \mathbb{E} \sup_{l} \sum_{i=1}^{m} \xi_i l(f(\mathrm{x}_i), y_i)$ is the Rademacher complexity, denoted $\mathcal{R}$, with ξ i.i.d. $= \begin{cases} 1 & \text{w.p. } \frac{1}{2} \\ -1 & \text{w.p. } \frac{1}{2}. \end{cases}$

Using a theorem about Rademacher complexity [2, Theorem 3.3]:

To do this, we first need the following theorem:

$$\mathbb{E} g(z) \leq \frac{1}{m} \sum_{i=1}^{m} g(z_i) + 2\mathcal{R}_m(\mathcal{G}) + \sqrt{\frac{\log \frac{1}{\delta}}{2m}}$$

$$\text{where } \mathcal{R}_m(\mathcal{G}) = \frac{1}{m} \mathbb{E} \sup_{g \in \mathcal{G}} \sum_{i=1}^{m} \xi_i g(z_i).$$

Let us prove this. However, the proof involves an inequality called McDiarmid's inequality, so before we prove this, let us briefly discuss McDiarmid's inequality [4].

McDiarmid's inequality is a concentration inequality. Concentration inequalities are used when we want to show that some random quantity is close to its mean with high probability [1]. Some examples of concentration inequalities are Markov's inequality, which states that, for $Z \geq 0$ and $\epsilon > 0, \mathbb{P}(Z > \epsilon) \leq \frac{\mathbb{E}(Z)}{\epsilon}$, and Chebyshev's inequality, which states that $\mathbb{P}(|Z - \mathbb{E}Z| > \epsilon) \leq \frac{\mathrm{Var}(Z)}{\epsilon^2}$. McDiarmid's inequality is stronger.

McDiarmid's inequality states: For $Z_1, \ldots, Z_m$ independent random variables and $i = 1, \ldots, m$, suppose we have a function f that satisfies

$$\sup_{z_1, \ldots, z_m, z_i'} \left| f(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_m) - f(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_m) \right| \leq c_i.$$

Then

$$\mathbb{P}\left(\left| f(Z_1, \ldots, Z_m) - \mathbb{E}(f(Z_1, \ldots, Z_m)) \right| \geq \epsilon \right) \leq 2 \exp\left(- \frac{2\epsilon^2}{\sum_{i=1}^{m} c_i^2} \right).$$

Now, let us use McDiarmid's inequality to actually prove the theorem about Rademacher complexity:

Proof. For any sample $S = (z_1, \ldots, z_m)$ and any $g \in \mathcal{G}$, we first choose function f to be

$$f(S) = \sup_{g \in \mathcal{G}} \left(\mathbb{E}[g] - \hat{\mathbb{E}}_S[g] \right).$$

Now, let us have two samples S and S', which differ by exactly one point, say z_m in S and z'_m in S'. Then,

$$\begin{aligned} f(S') - f(S) &\leq \sup_{g \in \mathcal{G}} \left(\hat{\mathbb{E}}_S[g] - \hat{\mathbb{E}}_{S'}[g] \right) \\ &= \sup_{g \in \mathcal{G}} \frac{g(z_m) - g(z'_m)}{m} \\ &\leq \frac{1}{m}. \end{aligned}$$

We can similarly show that $f(S) - f(S') \leq \frac{1}{m}$, which indicates that $|f(S) - f(S')| \leq \frac{1}{m}$. As a result, the c_i in $\sup_{z_1, \ldots, z_m, z'_i} \left| f(z_1, \ldots, z_{i-1}, z_i, z_{i+1}, \ldots, z_m) - f(z_1, \ldots, z_{i-1}, z'_i, z_{i+1}, \ldots, z_m) \right| \leq c_i$ can be considered to be $\frac{1}{m}$.

We can now apply McDiarmid's inequality to get: for any $\delta > 0$, with probability at least $1 - \delta/2$,

$$f(S) \leq \mathop{\mathbb{E}}_{S}[f(S)] + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

From here, let's break down the expectation on the right-hand side:

$$\begin{aligned} \mathop{\mathbb{E}}_{S}[f(S)] &= \mathop{\mathbb{E}}_{S}\left[\sup_{g \in \mathcal{G}} (\mathbb{E}[g] - \hat{\mathbb{E}}_S(g)) \right] \\ &= \mathop{\mathbb{E}}_{S}\left[\sup_{g \in \mathcal{G}} \mathop{\mathbb{E}}_{S'}[\hat{\mathbb{E}}_{S'}(g) - \hat{\mathbb{E}}_S(g)] \right] \\ &\leq \mathop{\mathbb{E}}_{S,S'}\left[\sup_{g \in \mathcal{G}} (\hat{\mathbb{E}}_{S'}(g) - \hat{\mathbb{E}}_S(g)) \right] \\ &= \mathop{\mathbb{E}}_{S,S'}\left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} (g(z'_i) - g(z_i)) \right] \\ &= \mathop{\mathbb{E}}_{\sigma,S,S'}\left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i (g(z'_i) - g(z_i)) \right] \\ &\leq \mathop{\mathbb{E}}_{\sigma,S'}\left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z'_i) \right] + \mathop{\mathbb{E}}_{\sigma,S}\left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} -\sigma_i g(z_i) \right] \\ &= 2 \mathop{\mathbb{E}}_{\sigma,S}\left[\sup_{g \in \mathcal{G}} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i) \right] \\ &= 2\mathcal{R}_m(\mathcal{G}). \end{aligned}$$

Now, replacing $\mathop{\mathbb{E}}_{S}[f(S)]$ with $2\mathcal{R}_m(\mathcal{G})$, we have:

$$f(S) \leq 2\mathcal{R}_m(\mathcal{G}) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}.$$

We chose f to be $\sup_{g\in\mathcal{G}}\left(\mathbb{E}[g]-\hat{\mathbb{E}}_S[g]\right)$, so:

$$\sup_{g\in\mathcal{G}}\left(\mathbb{E}[g]-\hat{\mathbb{E}}_S[g]\right)\leq 2\mathcal{R}_m(\mathcal{G})+\sqrt{\frac{\log\frac{2}{\delta}}{2m}}.$$

$\hat{\mathbb{E}}_S[g]=\frac{1}{m}\sum_{i=1}^{m}g(z_i)$, so:

$$\mathbb{E}g(z)\leq\frac{1}{m}\sum_{i=1}^{m}g(z_i)+2\mathcal{R}_m(\mathcal{G})+\sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

□

Proving difference between test and training error bounded above by Rademacher complexity:

We now have that

$$\mathbb{E}g(z)\leq\frac{1}{m}\sum_{i=1}^{m}g(z_i)+2\mathcal{R}_m(\mathcal{G})+\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$
$$\text{where }\mathcal{R}_m(\mathcal{G})=\frac{1}{m}\mathbb{E}\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\xi_i g(z_i).$$

Let us now use the above in order to prove that the difference between test and training error is bounded above by Rademacher complexity

$$L(f(\mathrm{x}),y)-\hat{L}(f(\mathrm{x}),y)\lesssim\frac{1}{m}\mathbb{E}\sup_{l}\sum_{i=1}^{m}\xi_i l(f(\mathrm{x}_i),y_i)+\sqrt{\frac{\log\frac{2}{s}}{m}}.$$

Proof. Choose $g=l(f(\mathrm{x}),y)$.

$$\begin{aligned}
\mathbb{E}l(f(\mathrm{x}),y)&=L(f(\mathrm{x}),y)\\
L(f(\mathrm{x}),y)&\leq\frac{1}{m}\sum_{i=1}^{m}l(f(\mathrm{x}_i),y_i)+2\mathcal{R}_m(\mathcal{G})+\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\\
L(f(\mathrm{x}),y)&\leq\frac{1}{m}\sum_{i=1}^{m}l(f(\mathrm{x}_i),y_i)+2*\frac{1}{m}\mathbb{E}\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\xi_i l(f(\mathrm{x}_i),y_i)+\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\\
\hat{L}(f(\mathrm{x}),y)&=\frac{1}{m}\sum_{i=1}^{m}l(f(\mathrm{x}_i),y_i)\\
L(f(\mathrm{x}),y)-\hat{L}(f(\mathrm{x}),y)&\leq 2*\frac{1}{m}\mathbb{E}\sup_{g\in\mathcal{G}}\sum_{i=1}^{m}\xi_i l(f(\mathrm{x}_i),y_i)+\sqrt{\frac{\log\frac{1}{\delta}}{2m}}\\
&\lesssim\frac{1}{m}\mathbb{E}\sup_{l}\sum_{i=1}^{m}\xi_i l(f(\mathrm{x}_i),y_i)+\sqrt{\frac{\log\frac{2}{s}}{m}}.
\end{aligned}$$

□

2.2 Rademacher complexity bounded above by (1)

Now that we know that the difference between test and training error is bounded above by the Rademacher complexity, we can find an upper bound for the Rademacher complexity in order to get the proposed bound $L(f(\mathrm{x}), y) - \hat{L}(f(\mathrm{x}), y) \lesssim \frac{1}{m} \sum_{j=1}^{h} \alpha_j(\beta_j \|\mathrm{X}\|_F + \|\mathrm{u}_j^0 \mathrm{X}\|_2) + \sqrt{\frac{\log \frac{2}{s}}{m}}$.

In other words, if we assume $U_0 = 0$, we essentially need to show that

$$\mathbb{E} \sup_l \sum_{i=1}^{m} \xi_i l(f(\mathrm{x}_i), y_i) \leq \sum_{j=1}^{h} \alpha_j(\beta_j \|\mathrm{X}\|_F). \tag{3}$$

To do so, we will need to use the fact

$$\mathcal{R}(\phi \circ f) \leq a\mathcal{R}(f) \quad \text{if} \quad \mathrm{Lip}(\phi) \leq a.$$

Proof. Using this fact, we have:

$$\begin{aligned}
\mathbb{E} \sup_l \sum_{i=1}^{m} \xi_i l(f(\mathrm{x}_i), y_i) &\leq \mathbb{E} \sup_{V,U} \sum_{i=1}^{m} \xi_i \sum_{j=1}^{h} v_j (u_j^T x_i)_+ \\
&= \mathbb{E} \sup_{V,U} \sum_{i=1}^{m} \xi_i \sum_{j=1}^{h} v_j \|u_j\|_2 \frac{(u_j^T x_i)}{\|u_j\|_2}{}_+ \\
&= \mathbb{E} \sup_{V,U} \sum_{j=1}^{h} v_j \|u_j\|_2 \sum_{i=1}^{m} \xi_i \left(\frac{u_j^T}{\|u_j\|_2} x_i \right)_+ \\
&\leq \mathbb{E} \sum_{j=1}^{h} \alpha_j \beta_j \sup_{u_j} \left| \sum_{i=1}^{m} \xi_i \left(\frac{u_j^T}{\|u_j\|_2} x_i \right)_+ \right| \\
&\leq \mathbb{E} \sum_{j=1}^{h} \alpha_j \beta_j \sup_{\|q_j\|_2 \leq 1} \left| \sum_{i=1}^{m} \xi_i (q_j^T x_i)_+ \right| \qquad \text{where} \quad q_j^T = \frac{u_j^T}{\|u_j\|_2} \\
&= \mathbb{E} \sum_{j=1}^{h} \alpha_j \beta_j \left[\sup_{\|q_j\|_2 \leq 1} \left(\sum_{i=1}^{m} \xi_i (q_j^T x_i)_+ \right) - \inf_{\|q_j\|_2 \leq 1} \left(\sum_{i=1}^{m} \xi_i (q_j^T x_i)_+ \right) \right] \\
&= \mathbb{E} \sum_{j=1}^{h} \alpha_j \beta_j \left[\sup_{\|q_j\|_2 \leq 1} \left(\sum_{i=1}^{m} \xi_i (q_j^T x_i)_+ \right) + \sup_{\|q_j\|_2 \leq 1} \left(\sum_{i=1}^{m} (-\xi_i) (q_j^T x_i)_+ \right) \right] \\
&= \mathbb{E} \sum_{j=1}^{h} \alpha_j \beta_j \sup_{\|q_j\|_2 \leq 1} \left(\sum_{i=1}^{m} \xi_i (q_j^T x_i)_+ \right) + \mathbb{E} \sum_{j=1}^{h} \alpha_j \beta_j \sup_{\|q_j\|_2 \leq 1} \left(\sum_{i=1}^{m} (-\xi_i) (q_j^T x_i)_+ \right) \\
&= 2\mathbb{E} \sum_{j=1}^{h} \alpha_j \beta_j \sup_{\|q_j\|_2 \leq 1} \sum_{i=1}^{m} \xi_i (q_j^T x_i)_+.
\end{aligned}$$

From here, we can write:

$$2\mathbb{E}\sum_{j=1}^{h}\alpha_j\beta_j \sup_{\|q_j\|_2\le 1}\sum_{i=1}^{m}\xi_i(q_j^T x_i)_+ \le \sum_{j=1}^{h}\alpha_j\beta_j\mathbb{E}\sup_{\|q_j\|_2\le 1}\sum_{i=1}^{m}\xi_i(q_j^T x_i)$$
$$= \sum_{j=1}^{h}\alpha_j\beta_j\mathbb{E}\sup_{\|q_j\|_2\le 1} q_j^T\sum_{i=1}^{m}\xi_i x_i.$$

Now, we can look at the expectation on the right-hand side and show that it is the Frobenius norm:

$$\mathbb{E}\sup_{\|q_j\|_2\le 1} q_j^T\sum_{i=1}^{m}\xi_i x_i = \mathbb{E}\left\|\sum_{i=1}^{m}\xi_i x_i\right\|_2$$
$$\le \sqrt{\mathbb{E}\left\|\sum_{i=1}^{m}\xi_i x_i\right\|_2^2}$$
$$= \|\mathrm{X}\|_F.$$

As a result, we have

$$\mathbb{E}\sup_{l}\sum_{i=1}^{m}\xi_i l(f(\mathrm{x}_i), y_i) \le \sum_{j=1}^{h}\alpha_j\beta_j\|\mathrm{X}\|_F.$$

□

Hence, we find the overall upper bound to be

$$L(f(\mathrm{x}), y) - \hat{L}(f(\mathrm{x}), y) \lesssim \frac{1}{m}\sum_{j=1}^{h}\alpha_j(\beta_j\|\mathrm{X}\|_F + \|\mathrm{u}_j^0\mathrm{X}\|_2) + \sqrt{\frac{\log\frac{2}{s}}{m}}.$$

3 Lower Bound

For the lower bound, the paper finds a lower bound for the Rademacher complexity.

The lower bound is found to be:

$$\frac{1}{m}\mathbb{E}\sup_{f}\sum_{i=1}^{m}\xi_i f(x_i) \gtrsim \sum_{j=1}^{h}\alpha_j\beta_j\|\mathrm{X}\|_F. \tag{4}$$

Recall that we have

$$f_{\mathrm{V,U}}(\mathrm{x}) = \mathrm{V}[\mathrm{Ux}]_+.$$

This lower bound is obtained for a smaller parameter space with the specific case when $m = h = d = 2^k$ and x_i is the standard coordinate vector e_i for $i = 1, \ldots, 2^k$. We have

$$V = (\alpha_1, \ldots, \alpha_{2^k})$$
$$U = \begin{pmatrix}\beta_1 & & \\ & \ddots & \\ & & \beta_{2^k}\end{pmatrix} A_{2^k \mathrm{x} 2^k}$$

where $A_{2^k \mathrm{x} 2^k}$ is the Hadamard matrix.

The Hadamard matrix is a matrix, the elements of which are all $\pm 2^{-k/2}$. All of the columns of the Hadamard matrix are orthogonal and have norm 1, so they form an orthonormal basis of $\mathbb{R}^{2^k}$. Let us denote the jth column of $A_{2^k \mathrm{x} 2^k}$ as a_j. From this, we see that, as U is given by the columns $a_1, \ldots, a_{2^k}$, we can say f is also given by the columns $a_1, \ldots, a_{2^k}$.

Proof. To prove the lower bound, with this setup, we have:

$$\begin{aligned} f_{\mathrm{V,U}}(\mathrm{x}) &= \mathrm{V}[\mathrm{Ux}]_+ \\ &= (\alpha_1, \ldots, \alpha_{2^k}) \left(\begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_{2^k} \end{pmatrix} (a_1, \ldots, a_{2^k}) e_i \right)_+ . \end{aligned}$$

If we let $i = 1$, this becomes

$$\begin{aligned} &= (\alpha_1, \ldots, \alpha_{2^k}) \left(\begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_{2^k} \end{pmatrix} a_1 \right)_+ \\ &= \sum_{j=1}^{h} \alpha_j \beta_j (a_{1j})_+ . \end{aligned}$$

We can now look at the following:

$$(\alpha_1, \ldots, \alpha_{2^k}) \left(\begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_{2^k} \end{pmatrix} a_1 \right)_+ + (\alpha_1, \ldots, \alpha_{2^k}) \left(\begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_{2^k} \end{pmatrix} -a_1 \right)_+$$

$$\begin{aligned} &= \sum_{j=1}^{h} \alpha_j \beta_j \Big[(a_{1j})_+ + (-a_{1j})_+ \Big] \\ &= 2^{-k/2} \sum_{j=1}^{h} \alpha_j \beta_j . \end{aligned}$$

From this, we can see that either $\sum_{j=1}^{h} \alpha_j \beta_j (a_{1j})_+$ or $\sum_{j=1}^{h} \alpha_j \beta_j (-a_{1j})_+$ is greater than $2^{-k/2-1} \sum_{j=1}^{h} \alpha_j \beta_j$.

Now, let us pick a new $A^* = (a_1^*, \ldots, a_{2^k}^*)$ for which

$$a_i^* = \begin{cases} a_i & \text{if } \sum_{j=1}^{h} \alpha_j \beta_j (a_{1j})_+ \geq 2^{-k/2-1} \sum_{j=1}^{h} \alpha_j \beta_j \\ -a_i & \text{if } \sum_{j=1}^{h} \alpha_j \beta_j (-a_{1j})_+ \geq 2^{-k/2-1} \sum_{j=1}^{h} \alpha_j \beta_j . \end{cases}$$

Now, with Rademacher complexity, when ξ_i is fixed, we can choose a new matrix $\widetilde{A}$ with columns $\tilde{a}_1, \ldots, \tilde{a}_{2^k}$ to be used in a new function

$$\tilde{f}_{\mathrm{V,U}}(\mathrm{x}) = \mathrm{V}[\widetilde{\mathrm{U}}\mathrm{x}]_+$$

where

$$\widetilde{U} = \begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_{2^k} \end{pmatrix} \widetilde{A}_{2^k \mathrm{x} 2^k}$$

such that

$$\tilde{a_i} = \begin{cases} a_i^* & \text{if } \xi_i > 0 \\ 0 & \text{if } \xi_i \leq 0. \end{cases}$$

From here, we have that

$$\begin{aligned} \xi_i \tilde{f}(x_i) &= \xi_i \tilde{f}(e_i) \\ &= \xi_i(\alpha_1, \ldots, \alpha_{2^k}) \left(\begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_{2^k} \end{pmatrix} \widetilde{A}_{2^k \mathrm{x} 2^k} e_i \right)_+ \\ &= \xi_i \sum_{j=1}^{h} \alpha_j \beta_j (\tilde{a_{ij}})_+ \\ &= \begin{cases} \sum_{j=1}^{h} \alpha_j \beta_j (a_{ij}^*)_+ & \text{when } \xi_i > 0, \text{ which has probability } \frac{1}{2} \text{ by Rademacher} \\ 0 & \text{when } \xi_i \leq 0, \text{ which has probability } \frac{1}{2} \text{ by Rademacher.} \end{cases} \end{aligned}$$

Now, we can show:

$$\begin{aligned} \mathbb{E} \sup_f \sum_{i=1}^{m} \xi_i f(x_i) &\geq \mathbb{E} \sum_{i=1}^{m} \xi_i \tilde{f}(x_i) \\ &= \sum_{i=1}^{m} \mathbb{E} \xi_i \tilde{f}(x_i) \\ &= \sum_{i=1}^{m} \left(\frac{1}{2} * 0 + \frac{1}{2} * \sum_{j=1}^{h} \alpha_j \beta_j (a_{ij}^*)_+ \right) \\ &= \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{h} \alpha_j \beta_j (a_{ij}^*)_+ \\ &\geq \frac{1}{2} \sum_{i=1}^{m} 2^{-k/2-1} \sum_{j=1}^{h} \alpha_j \beta_j \\ &= 2^{-k/2-2} m \sum_{j=1}^{h} \alpha_j \beta_j \\ &= 2^{-2} m \sum_{j=1}^{h} \alpha_j \beta_j 2^{-k/2} \qquad \text{(Note that } 2^{-k/2} = \frac{\|\mathrm{X}\|_F}{m}.\text{)} \\ &= \frac{1}{4} m \sum_{j=1}^{h} \alpha_j \beta_j \frac{\|\mathrm{X}\|_F}{m} \\ &= \frac{1}{4} \sum_{j=1}^{h} \alpha_j \beta_j \|\mathrm{X}\|_F. \end{aligned}$$

As a result, we have

$$\frac{1}{m}\mathbb{E}\sup_f \sum_{i=1}^{m} \xi_i f(x_i) \gtrsim \sum_{j=1}^{h} \alpha_j \beta_j \|\mathrm{X}\|_F.$$

□

Overall, we have an upper bound

$$L(f(\mathrm{x}), y) - \hat{L}(f(\mathrm{x}), y) \lesssim \frac{1}{m} \sum_{j=1}^{h} \alpha_j (\beta_j \|\mathrm{X}\|_F + \|\mathrm{u}_j^0 \mathrm{X}\|_2) + \sqrt{\frac{\log \frac{2}{s}}{m}}$$

and a lower bound (for the Rademacher complexity)

$$\frac{1}{m}\mathbb{E}\sup_f \sum_{i=1}^{m} \xi_i f(x_i) \gtrsim \sum_{j=1}^{h} \alpha_j \beta_j \|\mathrm{X}\|_F.$$

We can also see that we have matching upper and lower bounds for the Rademacher complexity.

References

[1] John Lafferty, Han Liu, and Larry Wasserman. *Concentration of Measure*, chapter 7. Carnegie Mellon University, 2008-2010.

[2] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2012.

[3] Behnam Neyshabur, Zhiyuan Li, Srinadh Bhojanapalli, Yann LeCun, and Nathan Srebro. Towards understanding the role of over-parametrization in generalization of neural networks, 2018.

[4] Galen Reeves. Concentration inequalities: Hoeffding and mcdiarmid, 2008.