IEMS 402 Statistical Learning - 2025 Winter
HW4

Tuorui Peng[1]

Exercise 1 Fast Rate Generalization Error in the Realizable Setting

For 0-1 loss, we notice that: if $L(h^*) = 0$ is reachable, then it means that $\exists h_* \in \mathcal{H}$ s.t. $y^{(i)} = h_*(x^{(i)})$ a.s. (i.e. h_* is the underlying truth, and the data generation $y|x$ has no randomness in the sense of a.s.). Then we have

$$\hat{L}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell\left((x^{(i)}, y^{(i)}), h\right) = \frac{1}{n}\sum_{i=1}^{n} \ell\left((x^{(i)}, h_*(x^{(i)}), h)\right)$$

can reach $\hat{L}(h) = 0$ in some non-empty set $H \subset \mathcal{H}$ (because we would have relation $h_* \in \{h^*\} \subseteq H \subseteq \mathcal{H}$). So the event of interest is that $\hat{h} \in H \backslash \{h^*\}$

For any $h \in \mathcal{H}$ s.t. $L(h) \geq t$, we have:

$$\begin{aligned}
\mathbb{P}\left(\hat{L}(h) = 0\right) =& \mathbb{P}\left(\forall i : \ \ell\big((x^{(i)}, y^{(i)}), h\big) = 0\right) \\
=& \mathbb{P}\left(\ell\big((x, h_*(x)), h\big) = 0\right)^n \\
=& \big(1 - \mathbb{E}\left[\ell\big((x, h_*(x)), h\big)\right]\big)^n \\
\leq& (1-t)^n \leq \exp\left(-nt\right)
\end{aligned}$$

So we have

$$\begin{aligned}
\mathbb{P}\left(L(\hat{h}) \geq t\right) =& \mathbb{P}\left(\exists h : \ L(h) \geq t, \ \hat{L}(h) = 0\right) \\
\leq& \sum_{h:\, L(h) \geq t} \mathbb{P}\left(\hat{L}(h) = 0\right) \\
\leq& |\mathcal{H}| \cdot \exp\left(-nt\right)
\end{aligned}$$

Now set $\delta = |\mathcal{H}| \cdot \exp\left(-nt\right)$ and we have

$$L(\hat{h}) \leq \frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{n}$$

Exercise 2 Generalization Error near Interpolate

2.(a)

Note that for any fixed h, we have $\mathbb{E}\left[\hat{L}_n(h)\right] = L(h) = \mathbb{E}\left[\ell(h)\right]$ and we have

$$var(\ell(h)) = \mathbb{E}\left[\ell(h)^2\right] - \mathbb{E}\left[\ell(h)\right]^2 \leq \mathbb{E}\left[\ell(h)\right] = L(h)$$

since $\ell(h) \in [0, 1]$. Now by Bernstein's inequality we proved in the previous homework, we have

$$\mathbb{P}\left(\hat{L}(h) - L(h) \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2(var(\ell(h)) + \varepsilon/3)}\right) \leq \exp\left(-\frac{n\varepsilon^2}{2(L(h) + \varepsilon/3)}\right)$$

[1] TuoruiPeng2028@u.northwestern.edu

2.(b)

By the same argument we also have inequality for the other side:

$$\mathbb{P}\left(\hat{L}(h) - L(h) \leq -\varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2(L(h) + \varepsilon/3)}\right)$$

Substitute $\hat{L}_n(h) \mapsto \hat{L}_n(h) - (L(h) - \varepsilon(h) - \varepsilon)$ and we have

$$\mathbb{P}\left(\hat{L}(h) \leq \varepsilon(h)\right) \leq \mathbb{P}\left(\hat{L}(h) - L(h) \leq -\varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2(\varepsilon(h) + \varepsilon + \varepsilon/3)}\right) = \exp\left(-\frac{n\varepsilon^2}{2(\varepsilon(h) + 4\varepsilon/3)}\right)$$

2.(c)

Note that by (b), we have for all h satisfying $L(h) \geq L(h^*) + 2\varepsilon$ that

$$\mathbb{P}\left(\hat{L}(h) \leq L(h^*) + \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2(\varepsilon + 4\varepsilon/3)}\right) = \exp\left(-\frac{n\varepsilon^2}{2(L(h^*) + 7\varepsilon/3)}\right)$$

So we have

$$\begin{aligned}
\mathbb{P}\left(\hat{L}_n(\hat{h}_n) - L(h^*) \geq 2\varepsilon\right) \leq & \mathbb{P}\left(\exists h:\, L(h) \geq L(h^*) + 2\varepsilon,\, \hat{L}(h) \leq L(h^*) + \varepsilon\right) \\
\leq & \sum_{h: L(h) \geq L(h^*) + 2\varepsilon} \mathbb{P}\left(\hat{L}(h) \leq L(h^*) + \varepsilon\right) \\
\leq & |\mathcal{H}| \exp\left(-\frac{n\varepsilon^2}{2(L(h^*) + 7\varepsilon/3)}\right)
\end{aligned}$$

On the other hand, by Hoeffding's inequality we have

$$\mathbb{P}\left(\left|\hat{L}_n(\hat{h}_n) - L(\hat{h}_n)\right| \geq \varepsilon\right) \leq \mathbb{P}\left(\sup_{h \in \mathcal{H}} \left|\hat{L}_n(h) - L(h)\right| \geq \varepsilon\right) \leq |\mathcal{H}| \exp\left(-\frac{n\varepsilon^2}{2}\right)$$

Together, we have

$$\begin{aligned}
\mathbb{P}\left(L(\hat{h}_n) - L(h^*) \leq \varepsilon\right) \geq & \mathbb{P}\left(\hat{L}_n(\hat{h}_n) - L(h^*) \geq 2\varepsilon, L(\hat{h}_n) - \hat{L}_n(\hat{h}_n) \leq \varepsilon\right) \\
\geq & 1 - |\mathcal{H}| \exp\left(-\frac{n\varepsilon^2}{2(L(h^*) + 7\varepsilon/3)}\right) - |\mathcal{H}| \exp\left(-\frac{n\varepsilon^2}{2}\right)
\end{aligned}$$

Note that for large ε, the above probability bound is dominated by the second term, so we consider setting $\delta = |\mathcal{H}| \exp\left(-\frac{n\varepsilon^2}{2(L(h^*) + 7\varepsilon/3)}\right)$ and we have

$$w.p. \geq 1 - \delta, \quad L(\hat{h}_n) - L(h^*) \lesssim \sqrt{\frac{L(h^*) \log(|\mathcal{H}|/\delta)}{n}} + \frac{\log(|\mathcal{H}|/\delta)}{n}.$$

2.(d)

The naive approach based on Hoeffding's inequality would give us

$$L(\hat{h}_n) - L(h^*) \lesssim \sqrt{\frac{\log|H| + \log\frac{1}{\delta}}{n}}$$

By comparing the two bounds, we notice that: when we have $L(h^*) \ll 1/\log(|\mathcal{H}|/\delta)$ the bound is dominated by $O(n^{-1})$, which is a stronger bound than the naive approach $O(n^{-1/2})$, leading to a faster rate of convergence.

Exercise 3 Random Matrix

3.(a)

Denote the SVD of A as

$$A = U\Sigma V^T$$

and thus

$$\|A\|_{\text{op}} = \sigma_1 = \max_{u\in S^{m-1}, v\in S^{n-1}} \max_{u\in S^{m-1}, v\in S^{n-1}} u'Av = \max_{u\in S^{m-1}, v\in S^{n-1}} \langle Au, v\rangle$$

by SVD.

3.(b)

(For this part I feel that the best I can do is $\sim \dfrac{1}{1-2\varepsilon}$ in stead of $\sim \dfrac{1}{(1-\varepsilon)^2}$)

With ε-net of S^{m-1} and S^{n-1} denoted $\mathcal{N}_1$ and $\mathcal{N}_2$ respectively. And denote the maximizer of $\langle Au, v'\rangle$ as (u^*, v^*), we have some $u_1 \in \mathcal{N}_1$ and $v_1 \in \mathcal{N}_2$ s.t.

$$\|u_1 - u^*\| \le \varepsilon, \quad \|v_1 - v^*\| \le \varepsilon$$

Then we have for $u^* \in S^{m-1}, v^* \in S^{n-1}$ and their corresponding u_1, v_1 we have

$$\begin{aligned}
\|A\|_{\text{op}} &= \langle Au, v\rangle \\
&= \langle A(u - u_1 + u_1), (v - v_1 + v_1)\rangle \\
&= \langle Au_1, v_1\rangle + \langle A(u - u_1), v_1\rangle + \langle Au, (v - v_1)\rangle \\
&\le \max_{u\in\mathcal{U}, v\in\mathcal{V}} \langle Au, v\rangle + \|A\|_{\text{op}}\,\varepsilon + \|A\|_{\text{op}}\,\varepsilon \\
\Rightarrow \|A\|_{\text{op}} &\le \frac{1}{1-2\varepsilon} \max_{u\in\mathcal{U}, v\in\mathcal{V}} \langle Au, v\rangle
\end{aligned}$$

3.(c)

Note that for any $u \in S^{m-1}, v \in S^{n-1}$, we have

$$\langle Au, v\rangle \sim N(0, 1)$$

And for ε covering set of S^{m-1} and S^{n-1}, we have

$$|\mathcal{U}| \le |\mathcal{N}(B_m(1)), \varepsilon| \le \left(1 + \frac{2}{\varepsilon}\right)^m$$
$$|\mathcal{V}| \le |\mathcal{N}(B_n(1)), \varepsilon| \le \left(1 + \frac{2}{\varepsilon}\right)^n$$

We have

$$\begin{aligned}\mathbb{E}\left[\|A\|_{\text{op}}\right] \leq & \frac{1}{1-2\varepsilon}\mathbb{E}\left[\max_{u\in\mathcal{U},v\in\mathcal{V}}\langle Au,v\rangle\right]\\ \lesssim & \frac{1}{1-2\varepsilon}\sqrt{\log|U|+\log|V|}\\ \lesssim & \sqrt{m}+\sqrt{n}\end{aligned}$$

where the last step can be done by taking some small constant ε.

3.(d)

For S^{n-1}, we construct a $\sqrt{2}$-net $\mathcal{N}$ formed by all unit vectors. This can be easily verified to be a covering by noticing that for $\sum_{i=1}^n x_i^2 \leq 1$ with WLOG $x_1 \geq 0$ we have

$$(1-x_1^2)+\sum_{i=2}^n x_i^2 \leq 2$$

Using this method we obtain covering for S^{m-1} and S^{n-1}respectively, and we have their size bounded by $(1+2/\sqrt{2})^m$ and $(1+2/\sqrt{2})^n$ respectively.

With this construction, we have

$$\langle u_i,u_j\rangle=\delta_{ij},\quad \langle v_i,v_j\rangle=\delta_{ij},\quad u_i,u_j\in\mathcal{N}_1;\ v_i,v_j\in\mathcal{N}_2$$

and thus for any $u_1,u_2\in\mathcal{N}_1$ and $v_1,v_2\in\mathcal{N}_2$ we have

$$\begin{aligned}cov\left(\langle Au_1,v_1\rangle,\langle Au_2,v_2\rangle\right) =& cov(\sum_{i=1}^m u_{1i}v_{1j}A_{ij},\sum_{i=1}^n u_{2i}v_{2j}A_{ij})\\ =&\sum_{i=1}^m\sum_{j=1}^n u_{1i}v_{1j}u_{2i}v_{2j}\\ =&\langle u_1,u_2\rangle\langle v_1,v_2\rangle\\ =&0\end{aligned}$$

i.e. any two $\langle Au_1,v_1\rangle,\langle Au_2,v_2\rangle$ are uncorrelated. Using this covering set, we have

$$\begin{aligned}\mathbb{E}\left[\|A\|_{\text{op}}\right] \geq & \mathbb{E}\left[\sup_{u\in\mathcal{U},v\in\mathcal{V}}\langle Au,v\rangle\right]\\ \overset{(*)}{\gtrsim} & \sqrt{\log|\mathcal{U}|+\log|\mathcal{V}|}\\ \gtrsim & \sqrt{m}+\sqrt{n}\end{aligned}$$

where $(*)$ by Martin J. Wainwright's book, page 53, Exercise 2.11 (Upper and lower bounds for Gaussian maxima).

As a result, we have

$$\mathbb{E}\left[\|A\|_{\text{op}}\right] \asymp \sqrt{m}+\sqrt{n}$$

Exercise 4 Rademacher Complexity Leads to Suboptimal Bounds

4.(a)

We have by CLT:

$$\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N(0,I)$$

and we have

$$\mathbb{E}\left[\left\|\hat{\theta}-\theta\right\|_2^2\right] = \frac{d}{n}$$

4.(b)

We have the hypothesis space $\Theta = \mathbb{R}^d$. We consider the following:

1. Bound covering number: We consider the covering of $\mathbb{B}_n(\delta,\Theta)$ under norm $\frac{1}{n}\sum_{i=1}^n (X_i - \cdot)^2$:

$$\log N(t, \mathbb{B}_n(\delta,\Theta), \|\cdot\|_2) \leq d\log(1+\frac{2\delta}{t}).$$

2. We determine some δ satisfying the following:

$$\begin{aligned}
&\frac{16}{\sqrt{n}}\int_0^\delta \sqrt{\log N(t, \mathbb{B}_n(\delta,\Theta), \|\cdot\|_2)}\,\mathrm{d}t \leq \frac{\delta^2}{4}\\
\Leftarrow &\frac{1}{\sqrt{n}}\int_0^\delta \sqrt{d\log(1+\frac{2t}{\delta})}\,\mathrm{d}t \lesssim \delta^2\\
\Leftarrow &\frac{\sqrt{d}}{\sqrt{n}}\delta \lesssim \delta^2\\
\Leftarrow &\delta \gtrsim \sqrt{\frac{d}{n}}
\end{aligned}$$

3. By Martin J. Wainwright's book, page 426, Corollary 13.7, we have

$$\mathbb{E}\left[\left\|\hat{\theta}-\theta\right\|_n^2\right] \lesssim \frac{d}{n}$$

So we see that we have the same optimal rate of convergence, but the constant is undetermined.

Exercise 5 Curse of Dimensionality

We consider the following segment to $[-B,+B]$ and partition of $[-R,+R]^d$:

- $[-B,+B]$ into segments of length ε, which gives $\lceil\frac{2B}{\varepsilon}\rceil$ segments (with trivial edges omitted). The set is denoted as $\mathcal{S} = \{S_1,\ldots,S_{\lceil\frac{2B}{\varepsilon}\rceil}\}$.
- $[-R,+R]^d$ into cubes of side length $\varepsilon/2\rho$, which gives $\lceil\frac{4\rho(R+\varepsilon)}{\varepsilon}\rceil^d$ cubes (with trivial edges omitted). So we haved a lattice $\mathcal{C}$ of side length ε/ρ.

For any function f in function class $\mathcal{F} : [-R, +R]^d \to [-B, +B]$, there exists some $S_f \in \mathcal{S}$ s.t. $|f(0) - S_f| \leq \varepsilon/2$. Next, for any $x \in \mathcal{C}$, we do the following:

1. Find the $S_x \in \mathcal{S}$ s.t. $|f(x) - S_x| \leq \varepsilon/2$.
2. For any $x' \in x + \varepsilon/2\rho \cdot [-1, 1]^d$, we have

$$\left|f(x') - S_x\right| \leq \left|f(x') - f(x)\right| + |f(x) - S_x| \leq \varepsilon/2 + \varepsilon/2 = \varepsilon$$

So at least one of $S_x, S_x + \varepsilon, S_x - \varepsilon$ would satisfy:

$$\forall x' \in x + \varepsilon/2\rho \cdot [-1, 1]^d, \quad \exists S_{x'} \in \{S_x, S_x + \varepsilon, S_x - \varepsilon\} \quad \left|f(x') - S_{x'}\right| \leq \varepsilon$$

By doing so iteratively, we can determine that there are $3^{|\mathcal{C}|}$ possible functions in the covering set (after $f(0)$ is given).

So in total, we have the size of the covering set begin:

$$\begin{aligned}
\log \mathcal{N}(\mathcal{F}, \varepsilon, \|\cdot\|_\infty) \leq & \log \left[\lceil \frac{2B}{\varepsilon} \rceil \times 3^{|\mathcal{C}|} \vee 1 \right] \\
\leq & \log \left[\lceil \frac{2B}{\varepsilon} \rceil \times 3^{\lceil \frac{4\rho(R+\varepsilon)}{\varepsilon} \rceil^d} \vee 1 \right] \\
= & 0 \vee \left[\lceil \frac{4\rho(R+\varepsilon)}{\varepsilon} \rceil^d \log 3 + \log \lceil \frac{2B}{\varepsilon} \rceil \right]
\end{aligned}$$

(which is a stronger bound than required in the homework question).