Math 281C Homework 5 Solutions

1. Consider the pair $\boldsymbol{z} = (\boldsymbol{x}, y) \in \mathbb{R}^d \times \{-1, 1\}$. Recall from Math 281A that the logistic loss is

$$m_{\boldsymbol{\theta}}(\boldsymbol{z}) = \log(1 + \exp(-y \cdot \langle \boldsymbol{x}, \boldsymbol{\theta} \rangle)),$$

and the population expectation is $M(\boldsymbol{\theta}) = \mathbb{E}[m_{\boldsymbol{\theta}}(\boldsymbol{X}, Y)]$, for $(\boldsymbol{X}, Y) \sim P$.

(a) Show that if $\boldsymbol{\Theta} \in \mathbb{R}^d$ is a compact set and $\mathbb{E}[\|\boldsymbol{X}\|] < \infty$ for some norm $\|\cdot\|$ on $\mathbb{R}^d$, then

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |P_n m_{\boldsymbol{\theta}}(\boldsymbol{X}, Y) - M(\boldsymbol{\theta})| \xrightarrow{p} 0.$$

Solution: First we show that $m_{\boldsymbol{\theta}}(\boldsymbol{z})$ is $\|\boldsymbol{x}\|_*$-Lipschitz in $\boldsymbol{\theta}$. For any $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ in $\boldsymbol{\Theta}$,

$$|m_{\boldsymbol{\theta}}(\boldsymbol{x}, y) - m_{\boldsymbol{\theta}'}(\boldsymbol{x}, y)| \le |\langle \boldsymbol{\theta} - \boldsymbol{\theta}', \boldsymbol{x} \rangle| \le \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \cdot \|\boldsymbol{x}\|_*.$$

Then consider a ϵ-net $\{\boldsymbol{\theta}^i\}_{i=1}^N$ for $\boldsymbol{\Theta}$, for each $\boldsymbol{\theta}^i$ in the ϵ-net, construct function pairs $\{\ell_i, u_i\}$ with

$$\ell_i = m_{\boldsymbol{\theta}^i}(\boldsymbol{x}, y) - \epsilon\|\boldsymbol{x}\|_* \quad \text{and} \quad u_i = m_{\boldsymbol{\theta}^i}(\boldsymbol{x}, y) + \epsilon\|\boldsymbol{x}\|_*,$$

such that for any $m_{\boldsymbol{\theta}}(\boldsymbol{x}, y)$, we can find a pair $\{\ell_i, u_i\}$ satisfying $\ell_i(\boldsymbol{x}) \le m_{\boldsymbol{\theta}}(\boldsymbol{x}, y) \le u_i(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{R}^d$, and

$$\mathbb{E}[u_i(\boldsymbol{x}) - \ell_i(\boldsymbol{x})] \le 2\epsilon \mathbb{E}\|\boldsymbol{x}\|_*.$$

This means $\{\ell_i, u_i\}_{i=1}^N$ form a $2\epsilon\mathbb{E}\|\boldsymbol{x}\|_*$-bracketing of $\{m_{\boldsymbol{\theta}}(\cdot) | \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ with respect to ℓ_1-norm, and by construction, we have

$$N_{[]}(\{m_{\boldsymbol{\theta}}\}, \ell_1, 2\epsilon\mathbb{E}\|\boldsymbol{x}\|_*) \le N(\boldsymbol{\Theta}, \|\cdot\|, \epsilon) < \infty,$$

which implies the uniform consistency.

(b) Assume that $\boldsymbol{\Theta}$ is contained in the norm ball $\{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\| \le r\}$ and that $\boldsymbol{X}$ is supported on the dual norm ball $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_* \le M\}$. Show that there is a constant $C < \infty$ such that for all $0 < \delta < 1$,

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |P_n m_{\boldsymbol{\theta}}(\boldsymbol{X}, Y) - M(\boldsymbol{\theta})| \ge \epsilon_n(\delta) \right) \le \delta,$$

where

$$\epsilon_n(\delta) = C\sqrt{\frac{r^2 M^2}{n}\left(d \log n + \log \frac{1}{\delta} \right)}.$$

Solution: First we have

$$\log(1 + \exp(-Mr)) \le m_{\boldsymbol{\theta}}(\boldsymbol{x}, y) \le \log(1 + \exp(Mr)),$$

which means $m_{\boldsymbol{\theta}}(\boldsymbol{x}, y) - \log(2) \in [-Mr, Mr]$. Thus, for any fixed $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $t > 0$, applying Hoeffding's inequality gives

$$\mathbb{P}(|P_n m_{\boldsymbol{\theta}}(\boldsymbol{X}, Y) - M(\boldsymbol{\theta})| \ge t) \le 2\exp\left(-\frac{nt^2}{2M^2 r^2} \right). \tag{1}$$

Then consider a minimal ϵ-net $\{\boldsymbol{\theta}^i\}_{i=1}^N$ for $\boldsymbol{\Theta}$ satisfying $N \le (1 + 2r/\epsilon)^d$. For any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, there is $\boldsymbol{\theta}^i$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}^i\| \le \epsilon$, and

$$\begin{aligned}\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |P_n m_{\boldsymbol{\theta}} - M(\boldsymbol{\theta})| &\le \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |P_n m_{\boldsymbol{\theta}} - P_n m_{\boldsymbol{\theta}^i}| + \max_{i=1,\dots,N} |P_n m_{\boldsymbol{\theta}^i} - P m_{\boldsymbol{\theta}^i}| + \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |P m_{\boldsymbol{\theta}^i} - P m_{\boldsymbol{\theta}}| \\ &\le 2M\epsilon + \max_{i=1,\dots,N} |P_n m_{\boldsymbol{\theta}^i} - P m_{\boldsymbol{\theta}^i}|.\end{aligned}$$

Therefore, combining (1) with union bound, we have

$$\mathbb{P}\bigg(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|P_n m_{\boldsymbol{\theta}} - M(\boldsymbol{\theta})| \ge 2M\epsilon + t\bigg) \le \mathbb{P}\bigg(\max_{i=1,\dots,N}|P_n m_{\boldsymbol{\theta}^i} - P m_{\boldsymbol{\theta}^i}| \ge t\bigg)$$
$$\le \bigg(1 + \frac{2r}{\epsilon}\bigg)^d 2\exp\bigg(-\frac{nt^2}{2M^2r^2}\bigg).$$

Finally, choosing

$$t = \sqrt{\frac{2M^2r^2(d\log(1+2r/\epsilon) + \log(2/\delta))}{n}}$$

and $\epsilon = r/n$ gives the desired bound.

2. Consider a binary classification problem with data in pair $(\boldsymbol{x}, y) \in \mathbb{R}^d \times \{-1, 1\}$, and let $\phi : \mathbb{R} \to \mathbb{R}_+$ be a 1-Lipschitz non-increasing convex function, for example, $\phi(t) = \log(1 + e^{-t})$ or $\phi(t) = [1 - t]_+$. Define $m_{\boldsymbol{\theta}}(\boldsymbol{x}, y) = \phi(y \cdot \langle \boldsymbol{x}, \boldsymbol{\theta}\rangle)$. Given an i.i.d. sample $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$ and consider the empirical risk minimization procedure

$$\widehat{\boldsymbol{\theta}}_n = \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} P_n m_{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{n}\sum_{i=1}^n m_{\boldsymbol{\theta}}(\boldsymbol{X}_i, Y_i). \tag{2}$$

Ledoux-Talagrand contraction inequality may be useful. Let $\phi \circ \mathcal{F} = \{h : h(x) = \phi(f(x)), f \in \mathcal{F}\}$ denote the composition of $\phi(\cdot)$ with functions in $\mathcal{F}$. If $\phi(\cdot)$ is L-Lipschitz, then $\mathcal{R}_n(\phi \circ \mathcal{F}) \le L\mathcal{R}_n(\mathcal{F})$.

(a) In one word, is the procedure (2) likely to give a reasonably good classifier?

Solution: Yes. Intuitively, we will pick $\widehat{\boldsymbol{\theta}}_n$ such that $Y_i \cdot \langle \boldsymbol{X}_i, \widehat{\boldsymbol{\theta}}_n\rangle > 0$ for most cases.

Before we proceed to parts (b) and (c), let us prove some general results. Let $\|\cdot\|$ be an arbitrary norm and define $\mathcal{F} = \{f(x) = \langle \boldsymbol{x}, \boldsymbol{\theta}\rangle \mid \|\boldsymbol{\theta}\| \le r\}$, then

$$\mathcal{R}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}\bigg[\sup_{\|\boldsymbol{\theta}\|\le r}\sum_{i=1}^n \epsilon_i \langle \boldsymbol{X}_i, \boldsymbol{\theta}\rangle\bigg] \le \frac{r}{n}\mathbb{E}\bigg[\bigg\|\sum_{i=1}^n \epsilon_i \boldsymbol{X}_i\bigg\|_*\bigg]. \tag{3}$$

Moreover, suppose a function class $\mathcal{F}$ is b-uniformly bounded, then for any $t > 0$, we have

$$\mathbb{P}(\|P_n - P\|_{\mathcal{F}} \ge 2\mathcal{R}_n(\mathcal{F}) + t) \le \exp\bigg(-\frac{nt^2}{2b^2}\bigg). \tag{4}$$

To prove (4), first we show concentration around mean. If we define

$$G(x_1, \dots, x_n) := \sup_{f\in\mathcal{F}}\bigg|\frac{1}{n}\sum_{i=1}^n (f(x_i) - \mathbb{E}f(X))\bigg|,$$

then it can be checked that

$$|G(x_1, \dots, x_i, \dots, x_n) - G(x_1, \dots, x_i', \dots, x_n)| \le \frac{2b}{n}.$$

Therefore, by bounded difference inequality, we have

$$\mathbb{P}(\|P_n - P\|_{\mathcal{F}} \ge \mathbb{E}\|P_n - P\|_{\mathcal{F}} + t) \le \exp\bigg(-\frac{nt^2}{2b^2}\bigg).$$

Furthermore, by Theorem 1.1 of Lecture 5, $\mathbb{E}\|P_n - P\|_{\mathcal{F}} \le 2\mathcal{R}_n(\mathcal{F})$. Combining this with the above display completes the proof of (4).

(b) Let $\boldsymbol{\Theta} \subset \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_2 \le r\}$ and let $\{\boldsymbol{X}_i\}_{i=1}^n$ be supported on the ℓ_2-ball $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 \le M\}$. Give the smallest $\epsilon_n(\delta, d, r, M)$ you can (ignoring the constants) such that

$$\mathbb{P}\bigg(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}|P_n m_{\boldsymbol{\theta}} - P m_{\boldsymbol{\theta}}| \ge \epsilon_n(\delta, d, r, M)\bigg) \le \delta.$$

How does your ϵ_n compare with Question 1?

Solution: The dual norm of ℓ_2-norm is ℓ_2-norm. Consider $\mathcal{F} = \{f(x) = \langle \boldsymbol{x}, \boldsymbol{\theta}\rangle | \boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. By the independence of Rademacher variables and Jensen's inequality,

$$\mathbb{E}\left[\left\|\sum_{i=1}^n \epsilon_i \boldsymbol{X}_i\right\|_2\right] \le \sqrt{\mathbb{E}\left[\left\|\sum_{i=1}^n \epsilon_i \boldsymbol{X}_i\right\|_2^2\right]} = \sqrt{\mathbb{E}\left[\sum_{i=1}^n \|\boldsymbol{X}_i\|_2^2\right]} \le \sqrt{n}M,$$

which, together with (3), implies

$$\mathcal{R}_n(\mathcal{F}) \le \frac{Mr}{\sqrt{n}}.$$

Applying Ledoux-Talagrand contraction inequality yields

$$\mathcal{R}_n(\{m_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}) \le \frac{Mr}{\sqrt{n}}.$$

Now, by the Lipschitz continuity of $\phi(\cdot)$,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}, \|\boldsymbol{x}\|_2\le M, y\in\{-1,1\}} |\phi(y\cdot\langle \boldsymbol{x}, \boldsymbol{\theta}\rangle) - \phi(0)| \le Mr,$$

and it follows by (4) that for any $t > 0$,

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |P_n m_{\boldsymbol{\theta}} - P m_{\boldsymbol{\theta}}| \ge \frac{2Mr}{\sqrt{n}} + t\right) \le \exp\left(-\frac{nt^2}{2M^2r^2}\right).$$

Finally, setting

$$t = \sqrt{\frac{2M^2r^2\log(1/\delta)}{n}}$$

gives

$$\epsilon_n = \frac{Mr}{\sqrt{n}}(2 + \sqrt{2\log(1/\delta)}).$$

This bound is apparently sharper than the bound in Question 1, since it's independent of d.

(c) Let $\boldsymbol{\Theta} \subset \{\boldsymbol{\theta} \in \mathbb{R}^d : \|\boldsymbol{\theta}\|_1 \le r\}$ and let $\{\boldsymbol{X}_i\}_{i=1}^n$ be supported on the ℓ_∞-ball $\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_\infty \le M\}$. Give the smallest $\epsilon_n(\delta, d, r, M)$ you can (ignoring the constants) such that

$$\mathbb{P}\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |P_n m_{\boldsymbol{\theta}} - P m_{\boldsymbol{\theta}}| \ge \epsilon_n(\delta, d, r, M)\right) \le \delta.$$

How does your ϵ_n compare with Question 1?

Solution: Let $\{X_i\}_{i=1}^n$ be a sequence of centered sub-Gaussian random variables with parameter σ, then

$$\mathbb{E}\left[\max_{i=1,\dots,n} |X_i|\right] \le \sigma\sqrt{2\log(2n)} \tag{5}$$

To prove (5), by the definition of sub-Gaussian and Jensen's inequality, for any $\lambda > 0$,

$$\begin{aligned}
\mathbb{E}\left[\max_{i=1,\dots,n} |X_i|\right] &= \frac{1}{\lambda}\mathbb{E}\log\exp\left(\lambda \max_{i=1,\dots,n} |X_i|\right) \\
&\le \frac{1}{\lambda}\log\mathbb{E}\exp\left(\lambda \max_{i=1,\dots,n} |X_i|\right) \\
&\le \frac{1}{\lambda}\log\left[2n\exp\left(\frac{\lambda^2\sigma^2}{2}\right)\right] \\
&= \frac{\log 2n}{\lambda} + \frac{\lambda\sigma^2}{2}.
\end{aligned}$$

Taking $\lambda = \sqrt{2\log(2n)}/\sigma$ completes the proof of (5).

Now notice that the dual norm of ℓ_1-norm is ℓ_∞-norm, and the j-th coordinate of $\sum_{i=1}^n \epsilon_i \boldsymbol{X}_i$ is sub-Gaussian with parameter $\sqrt{\sum_{i=1}^n \boldsymbol{X}_{ij}^2} \le M\sqrt{n}$. By (5), we have

$$\mathbb{E}\left[\left\|\sum_{i=1}^n \epsilon_i \boldsymbol{X}_i\right\|_\infty\right] \le M\sqrt{2n\log(2d)}.$$

Applying this to (3) gives

$$R_n(\mathcal{F}) \le \frac{Mr\sqrt{2\log(2d)}}{\sqrt{n}}.$$

The remaining arguments are the same as part (b), and we can achieve

$$\epsilon_n = \frac{Mr}{\sqrt{n}}(2\sqrt{2\log(2d)} + \sqrt{2\log(1/\delta)}).$$

This bound is also sharper than the one in Question 1.