Math 281C Homework 5

Due: May 15, 5 pm

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 $\mathbb{E}[\|X\|] < \infty$ for some norm $\|\cdot\|$ on $\mathbb{R}^d$, then

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

(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 $\{\boldsymbol{x} \in \mathbb{R}^d : \|x\|_* \le M\}$. 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

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

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

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

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 $R_n(\phi \circ \mathcal{F}) \le L R_n(\mathcal{F})$.

(a) In one word, is the procedure (1) likely to give a reasonably good classifier?

(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 $\{\boldsymbol{x} \in \mathbb{R}^d : \|x\|_2 \le M\}$. 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?

(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 $\{\boldsymbol{x} \in \mathbb{R}^d : \|x\|_\infty \le M\}$. 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?