B9145: Problem Set 1

Due: Feb 17, 11:59pm

Carefully follow submission instructions announced on Canvas.

Question 1.1 (Tail bound for sub-Gaussian RVs and Lasso)**:** For a class of functions $\mathcal{H} \subset \{h : \mathcal{Z} \to \mathbb{R}\}$, recall the definition of (empirical) Rademacher complexity

$$\mathfrak{R}_n(\mathcal{H}) := \mathbb{E}\left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i h(Z_i) \mid Z_1, \ldots, Z_n\right],$$

where ε_i's are i.i.d. random signs (Rademacher variables), independent of everything else.

(a) Let X_j be sub-Gaussian random variables with parameter c_j^2 for $j = 1, \ldots, N$. Show that for any $N \geq 3$,

$$\mathbb{E}[\max_{1 \leq j \leq N} X_j] \leq \max_{1 \leq j \leq N} c_j \cdot \sqrt{2 \log N}.$$

(b) For any finite $\mathcal{H}$, show that $\mathfrak{R}_n(\mathcal{H}) \leq \left(\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} h(Z_i)^2\right)^{\frac{1}{2}} \sqrt{\frac{2 \log |\mathcal{H}|}{n}}$.

(c) Consider L^1-regularized linear models $\mathcal{H}_s := \{z \mapsto \theta^\top z : \|\theta\|_1 \leq s\}$. Assume there exists $C_\infty > 0$ such that $\|Z\|_\infty \leq C_\infty$ almost surely. Derive the following scale-sensitive bound

$$\mathfrak{R}_n(\mathcal{H}_s) \leq s C_\infty \sqrt{\frac{2 \log(2d)}{n}}.$$

Hint For finite $\mathcal{G}$, $\mathfrak{R}_n(\mathcal{G}) = \mathfrak{R}_n(\text{convex-hull}(\mathcal{G}))$.

Question 1.2 (Two-layer neural networks)**:** Consider a neural network with two layers and activation function $a : \mathbb{R} \to \mathbb{R}$. Let $Z \in \mathbb{R}^d$ be an input vector with $\|Z\|_2 \leq R_2$ almost surely, and let $a : \mathbb{R} \to \mathbb{R}$ be a 1-Lipschitz activation function with $a(0) = 0$. For example, the rectified linear unit (ReLU) $a(x) := \max(x, 0)$, or hyperbolic tangent $a(x) := \tanh(x)$ are common choices that satisfy this condition.

Let m be the number of hidden units in the two-layer neural network. We denote by $w_j \in \mathbb{R}^d$ the weights of the first layer connecting to the j-th hidden unit, for $j = 1, \ldots, m$, and use $v \in \mathbb{R}^m$ to denote the weights of the second layer. Consider L^2-regularized two-layer neural networks

$$\mathcal{H} := \left\{ z \mapsto \sum_{j=1}^{m} v_j a(w_j^\top z) : \|v\|_2 \leq C_{2,v}, \text{ and } \|w_j\|_2 \leq C_{2,w} \text{ for all } j = 1, \ldots, m \right\}.$$

Show the scale-sensitive bound $\mathfrak{R}_n(\mathcal{H}) \leq 2 R_2 C_{2,v} C_{2,w} \sqrt{\frac{m}{n}}$.

Hint Use the contraction principle: for a 1-Lipschitz function $a : \mathbb{R} \to \mathbb{R}$ with $a(0) = 0$,

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i a(h(Z_i))\right| \mid Z_1, \ldots, Z_n\right] \le 2\mathbb{E}\left[\sup_{h \in \mathcal{H}} \left|\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i h(Z_i)\right| \mid Z_1, \ldots, Z_n\right].$$

Question 1.3 (Fast rates under curvature)**:** In this problem, we will show losses with curvature achieves faster rates of convergence. To do this, we study a localized Rademacher process around the population optimum.

Let $\Theta \subset \mathbb{R}^d$ be a compact, convex set, and let $\ell(\cdot; z) : \mathbb{R}^d \to \mathbb{R}$ be a convex function for P-almost surely all z. We assume that the population optimum $\theta^\star = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}[\ell(\theta; Z)]$ is unique. Consider Lipschitz losses (in some norm $\|\cdot\|$) that grow sufficiently fast near the optimum: for constants $r, c, L > 0$, and all θ, θ' satisfying $\|\theta - \theta^\star\| \le r$, $\|\theta' - \theta^\star\| \le r$,

$$\begin{aligned} &\left|\ell(\theta; z) - \ell(\theta'; z)\right| \le L \left\|\theta - \theta'\right\| \quad \text{for } P\text{-almost surely all } z, \\ &\text{and } \ \mathbb{E}[\ell(\theta; Z)] \ge \mathbb{E}[\ell(\theta^\star; Z)] + \frac{c}{2}\left\|\theta - \theta^\star\right\|^2. \end{aligned}$$

(e.g. think about a linear regression problem with bounded data.)

Define the set of empirical and population approximate optimizers

$$\begin{aligned} \widehat{S}_\epsilon &:= \left\{\theta \in \Theta : \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; Z_i) \le \inf_{\theta' \in \Theta} \frac{1}{n}\sum_{i=1}^{n} \ell(\theta'; Z_i) + \epsilon\right\} \\ S_\epsilon &:= \left\{\theta \in \Theta : \mathbb{E}[\ell(\theta; Z)] \le \inf_{\theta' \in \Theta} \mathbb{E}[\ell(\theta'; Z)] + \epsilon\right\}. \end{aligned}$$

Let $0 < \epsilon \le cr^2/4$ in the following.

(a) Argue that $\widehat{S}_\epsilon \not\subseteq S_{2\epsilon}$ implies

$$\sup_{\theta \in S_{2\epsilon}} \left\{\mathbb{E}[\ell(\theta; Z) - \ell(\theta^\star; Z)] - \frac{1}{n}\sum_{i=1}^{n} \left(\ell(\theta; Z_i) - \ell(\theta^\star; Z_i)\right)\right\} \ge \epsilon.$$

Hint Construct a $\theta \in \Theta$ with $\mathbb{E}[\ell(\theta; Z)] = \mathbb{E}[\ell(\theta^\star; Z)] + 2\epsilon$, $\frac{1}{n}\sum_{i=1}^{n} \ell(\theta; Z_i) \le \frac{1}{n}\sum_{i=1}^{n} \ell(\theta^\star; Z_i) + \epsilon$.

(b) Using results from class, prove that with probability at least $1 - e^{-t}$,

$$\begin{aligned} &\sup_{\theta \in S_{2\epsilon}} \left\{\mathbb{E}[\ell(\theta; Z) - \ell(\theta^\star; Z)] - \frac{1}{n}\sum_{i=1}^{n} \left(\ell(\theta; Z_i) - \ell(\theta^\star; Z_i)\right)\right\} \\ &\le 2\mathbb{E}\left[\sup_{\theta \in S_{2\epsilon}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \left(\ell(\theta; Z_i) - \ell(\theta^\star; Z_i)\right)\right] + 2L\sqrt{\frac{2t\epsilon}{cn}}. \end{aligned}$$

(c) Show the following: for some numerical constant $C > 0$,

$$\mathbb{E}\left[\sup_{\theta \in S_{2\epsilon}} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i \left(\ell(\theta; Z_i) - \ell(\theta^\star; Z_i)\right) \mid Z_1, \ldots, Z_n\right] \le CL\sqrt{\frac{d\epsilon}{cn}}.$$

(You don't need to find the constant.)

(d) Conclude that for a numerical constant $C > 0$ (which may differ from the one above), setting $\epsilon_t = CL^2 \frac{d+t}{cn}$ yields $\mathbb{P}\left(\widehat{S}_{\epsilon_t} \not\subseteq S_{2\epsilon_t}\right) \le e^{-t}$.