STAT260 Problem Set 5

Due December 11th via e-mail to jsteinhardt+pset5@berkeley.edu

Regular problems:

1. Consider a logistic regression model with loss $\ell(\theta; x, y) = -\log \sigma(y\langle \theta, x\rangle)$, where $\sigma(z) = \frac{1}{1+\exp(-z)}$. Show that $\max_{\bar{x} : \|\bar{x} - x\|_\infty \leq \epsilon} \ell(\theta; \bar{x}, y)$ is equal to $-\log \sigma(y\langle \theta, x\rangle - \epsilon\|\theta\|_1)$.

   (Observe that this shows that for linear models, robustness in ℓ_∞ is asking for some combination of maximizing the margin of classification and minimizing the ℓ_1-norm of θ.)

2. Suppose we observe data $(x_1, t_1, y_1), \ldots, (x_n, t_n, y_n)$ drawn i.i.d. from p and satisfying the unconfoundedness assumption, with known true propensity scores $\pi_i = \pi(x_i)$ (i.e. it is known that $p(T = 1 \mid x_i) = \pi_i$). Consider the clipped inverse-propensity weighted estimator for the average treatment effect:

$$\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\mathbb{I}[t_i = 1]}{\max(\pi_i, 1/M)} - \frac{\mathbb{I}[t_i = 0]}{\max(1 - \pi_i, 1/M)}\Big)y_i, \tag{1}$$

   where the clipping parameter M ensures that the clipped inverse propensity weights are all at most M. Assuming that $y \in [-1, 1]$ almost surely, show that the bias of the estimator is at most

$$\mathbb{E}_{x\sim p}[\max(1 - \pi(x)M, 0) + \max(1 - (1 - \pi(x))M, 0)], \tag{2}$$

   while the variance is at most M^2/n.

3. Recall that for a regression problem, the (non-robust) standard error is given by $\frac{\sigma^2}{n}S^{-1}$, while the robust standard error is given by $\frac{1}{n}S^{-1}\Omega S^{-1}$, where

$$S = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top, \tag{3}$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \langle \hat{w}, x_i\rangle)^2, \tag{4}$$

$$\Omega = \frac{1}{n}\sum_{i=1}^{n} x_i (y_i - \langle \hat{w}, x_i\rangle)^2 x_i^\top, \tag{5}$$

   and $\hat{w}$ is the ordinary least squares estimate from $(x_1, y_1), \ldots, (x_n, y_n)$.

   Show that the robust standard error can be arbitrarily larger than the standard error. In other words, show that for any real number t there is a collection of points (x_i, y_i) such that $\frac{1}{n}S^{-1}\Omega S^{-1} \succeq t \cdot \frac{\sigma^2}{n}S^{-1}$.

Challenge problems (turn in as a separate document typset in LaTeX):

4. Call a set of points $S = \{x_1, \ldots, x_s\}$ (ϵ, κ)-*dimension-preserving* if $\frac{1}{|T|}\sum_{i\in T} x_i x_i^\top \succeq \kappa^{-1}\frac{1}{|S|}\sum_{i\in S} x_i x_i^\top$ for all $T \subseteq S$ with $|T| \geq \epsilon|S|$.

   Consider a linear-regression setting where we observe $(x_1, y_1), \ldots, (x_n, y_n)$. Suppose that there is a set S^* of αn of the x_i that are $(\alpha/4, \kappa)$-dimension-preserving, and that for these points we have $y_i = \langle w^*, x_i\rangle + z_i$, where $z_i \sim \mathcal{N}(0, \sigma^2 I)$. Show that with high probability it is possible to output a set of $m = \mathcal{O}(1/\alpha)$ candidates $\hat{w}_1, \dots, \hat{w}_m$ such that, for at least one of the elements $\hat{w}_l$, the excess prediction loss on S^* satisfies

$$\frac{1}{|S^*|} \sum_{i \in S^*} (\langle \hat{w}_l, x_i \rangle - y_i)^2 - (\langle w^*, x_i \rangle - y_i)^2 = \mathcal{O}\Big(\kappa \sigma^2 \frac{\log(1/\alpha)}{\alpha}\Big). \tag{6}$$

[Note: This should be true as stated, but you will get full points for any bound that is polynomial in κ, σ, and α, as long as it is independent of the dimension d for n sufficiently large.]

5. Consider a two-layer neural network $f(x) = c^\top \max(Wx, 0)$, where $x \in \mathbb{R}^d$, $W \in \mathbb{R}^{m \times d}$, and $c \in \mathbb{R}^m$. Take c to be the all-1s vector and each entry of W to be drawn independently and uniformly from $\{-1, +1\}$. Let f_{LP} be the upper bound on $\max\{f(x) \mid \|x\|_\infty \le 1\}$ certified by the LP, and f_{SDP} be the same upper bound certified by the SDP. Show that $f_{LP} = \Omega(md)$ almost surely, while $f_{SDP} = \mathcal{O}(m\sqrt{d} + d\sqrt{m})$ with probability $1 - \exp(-\Omega(m + d))$.

   For reference, the SDP relaxation in this case would be

$$\begin{aligned} &\text{maximize } c^\top z \\ &\text{subject to } \begin{bmatrix} 1 & x^\top & z^\top \\ x & X & Y^\top \\ z & Y & Z \end{bmatrix} \succeq 0, \\ &\quad \operatorname{diag}(X) \le 1, \\ &\quad z \ge 0, z \ge Wx, \\ &\quad \operatorname{diag}(Z) = \operatorname{diag}(WY^\top). \end{aligned} \tag{7}$$