STAT240 Problem Set 5

Due April 29th in class

Regular problems:

1. Suppose that for a ridge regression problem, instead of the eigenvalues of S decaying as a power law $\mu_j^* = j^{-\alpha}$, they decay exponentially: $\mu_j^* = \exp(-\beta j)$. In the setting where $Y_i = \langle \beta^*, X_i \rangle + Z_i$, where β^* is drawn from a Gaussian prior with scale parameter ρ and $Z_i \sim \mathcal{N}(0, \sigma^2)$, compute the bias and variance up to constant factors, as a function of σ, ρ, and λ. For the optimal regularization $\lambda_n^* = \frac{\sigma^2}{\rho^2 n}$, what is the expected squared error of the ridge regression estimator?

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 $\frac{1}{n}S^{-1}\Omega S^{-1} \succeq t \cdot \frac{\sigma^2}{n}S^{-1}$.

4. Show that the following quantity also yields a doubly robust estimate, in the sense that it is correct if either $q = p$ or $\bar{Y} = \mathbb{E}[Y \mid X]$:
$$\mu_t = \frac{\mathbb{E}_{X,T,Y\sim p}\Big[\mathbb{I}[T = t](Y(T) - \bar{Y}_T(X))/q(t \mid X)\Big]}{\mathbb{E}_{X,T,Y\sim p}[\frac{\mathbb{I}[T=t]}{q(t|X)}]} + \mathbb{E}_{X\sim p}[\bar{Y}_t(X)], \text{ for } t = 0, 1, \tag{6}$$
$$DRE' = \mu_1 - \mu_0. \tag{7}$$
   Qualitatively, how does this estimate compare to the doubly robust estimator from lecture?

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

5. Call a set of points $S = \{x_1, \dots, x_s\}$ (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 $(x_1, y_1), \dots, (x_n, y_n)$. 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

   $$\frac{1}{|S^*|} \sum_{i \in S^*} (\langle \hat{w}_l, x_i \rangle - y_i)^2 - (\langle w^*, x_i \rangle - y_i)^2 = \mathcal{O}\Big(\kappa \sigma^2 \frac{\log(1/\alpha)}{\alpha}\Big). \tag{8}$$

   [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.]

6. Consider a causal inference setting where instead of two treatment conditions $T = 0, 1$, the treatment is represented by a non-negative real number $T \in \mathbb{R}_{\geq 0}$ (this might be the case for instance if we are estimating the effect of vitamin D consumption on catching the flu, and for each patient we observe their daily vitamin D consumption in milligrams).

   We will still define $Y(t)$ to be the outcome that "would have" occurred if the treatment were $T = t$, and unconfoundedness means that $(Y(t))_{t \geq 0} \perp\!\!\!\perp T \mid X$.

   We will define the marginal average treatment effect to be

   $$MATE = \mathbb{E}_{X,T \sim p}\left[\frac{d}{dt}\mathbb{E}[Y(t) \mid X]\Big|_{t=T}\right]. \tag{9}$$

   In other words, this is the infinitesimal amount by which Y would increase on average, if we increased everyone's treatment level by a small amount.

   (i) Show that under unconfoundedness, and assuming the propensity to treat $p(t \mid X)$ is differentiable in t, the MATE is equal to

   $$MATE = -\mathbb{E}_p\Big[\Big(\frac{d}{dt}\log p(t \mid X)\Big|_{t=T}\Big) \cdot Y\Big] \tag{10}$$

   (ii) Assuming we have an estimate q for p and an estimate $\bar{Y}(t, x)$ for Y, design a doubly robust estimator based on the above equation. Show that it yields the correct answer as long as either $q(t \mid x) = p(t \mid x)$, or $\bar{Y}(t, x) = \mathbb{E}[Y(t) \mid X = x]$.