36-755, Fall 2017
Homework 3

Due Wed Oct 18 by 5:00pm in Jisu's mailbox

1. **Median and sample quantiles.**

   (a) Suppose that $(X_1, \ldots, X_n)$ is an i.i.d. sample from a distribution P (if you like, you may assume P to be absolutely continuous). Let $X_{(1)} \leq X_{(2)} < \ldots < X_{(n)}$ be the order statistics and set $\alpha \in (0,1)$. Determine a $1-\alpha$ confidence interval for the median of P of the form

   $$\left(X_{(k_1)}, X_{(k_2)}\right)$$

   for some choice of $k_1 < k_2$. Determine k_1 and k_2 by relating this problem to a $\mathrm{Bin}(n, 1/2)$ distribution and use concentration.

   (b) **Bonus problem (it means this is optional).** Letting m be the median of P, assumed unique for convenience. Assume also that there exists an $\eta > 0$ such that the c.d.f. F of P is differentiable at all $x \in I = (m-\eta, m+\eta)$, with $\inf_{x\in I} F'(x) \geq C > 0$. Compute a high probability bound on the length of the confidence interval found in the previous point. You may use the following result, known as the DKW inequality. If $X_1, \ldots, X_n$ is an i.i.d. sample from a distribution over the real line with c.d.f. F and F_n denotes the corresponding empirical c.d.f., then

   $$\mathbb{P}\left(\|F - F_n\|_\infty \geq t\right) \leq 2e^{-2nt^2}.$$

   What happens when η or C gets smaller?

   (c) Consider the same setting as the previous exercise and let F be the c.d.f. of P and $p \in (0,1)$. The pth quantile and p-th sample quantile are, respectively,

   $$\xi_p = \inf\{x \colon F(x) \geq p\}$$

   and

   $$\hat{\xi}_p = \inf\{x \colon F_n(x) \geq p\},$$

   res[ectively, where F_n is the sample c.d.f. (i.e. $F_n(x) = \frac{1}{n}\sum_{i=1}^n 1(X_i \leq x)$). Show that, for any $\epsilon > 0$,

   $$\mathbb{P}\left(|\hat{\xi}_p - \xi_p| > \epsilon\right) \leq 2\exp\left\{-2n\delta_\epsilon^2\right\},$$

   where $\delta_\epsilon = \min\left\{F(x_p+\epsilon) - p, p - F(\xi_p - \epsilon)\right\}$.
   Write, for instance, $\mathbb{P}\left(\hat{\xi}_p > \xi_p + \epsilon\right) = \mathbb{P}\left(p > F_n(\xi_p + \epsilon)\right)$. *Then, notice that* $F_n(x)$ *is a sum of i.i.d. Bernoulli and use Hoeffding yet again...*

2. Consider the linear regression model

   $$Y = X\theta^* + \epsilon$$

   where $\theta \in \mathbb{R}^d$, X is fixed and $\epsilon \in \mathbb{R}^n$ consists of independent zero-mean variables with finite variance. The ridge estimator is defined as

   $$\hat{\theta}_{\mathrm{ridge}} = \hat{\theta}_{\mathrm{ridge}}(\lambda) = \mathrm{argmin}_{\theta\in\mathbb{R}^d}\left\{\frac{1}{n}\|Y - X\theta\|^2 + \lambda\|\theta\|^2\right\},$$

   where $\lambda > 0$.

(a) Show that $\hat{\theta}_{\text{ridge}}$ is uniquely defined for any $\lambda > 0$ and find a closed-form expression. Will the solution exist and be unique if $d > n$?

(b) Compute the bias of $\hat{\theta}_{\text{ridge}}$.

3. Consider the distribution-free framework for regression: the pair $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ has a distribution P on $\mathbb{R}^d$. For any $x \in \mathbb{R}^d$ in the support of X, let $\mu(x) = \mathbb{E}[Y|X = x]$ be the regression function. As we discussed in class, linear regression postulates that $\mu(x) = \beta^\top x$, for some $\beta \in \mathbb{R}^d$. This is a very strong assumption, which is unlikely to hold in most scenarios. What if one still fits a linear regression function?

(a) Let $\Sigma = \mathbb{V}[X]$, assumed to be invertible. Define

$$\beta^* = \text{argmin}_{\beta \in \mathbb{R}^d} \mathbb{E}\left[(Y - X^\top \beta)^2\right].$$

The vector β^* contains the coefficients of the best (in an L_2 sense) approximation of Y by linear functions of X (In fact, $X^\top \beta^*$ is the L_2 projection of Y into the linear space of linear functions on X). Show that

$$\beta^* = \Sigma^{-1}\alpha,$$

where $\alpha = \mathbb{E}[YX] \in \mathbb{R}^d$.

(b) Now observe data in the form of n pairs $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} P$. Assume for simplicty that $\mathbb{E}[X] = 0$. The plug-in estimator of β^* is the ordinary least squares estimator

$$\hat{\beta} = \hat{\Sigma}^{-1}\hat{\alpha},$$

where $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\top$ and $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^m Y_i X_i$. We assume that P belongs to a large non-parametric class of probability distributions satisfying the folowing assumptions:

i. each P in the class has a Lebesgue density (which implies that $\hat{\Sigma}$ is invertible almost surely if $n \geq d$; why?).

ii. Y and all the coordinates of X are bounded in absolute value by some constant K, almost surely (this could be relaxed to a sub-gaussian assumption, but let's keep things simple).

iii. the covariancer matrix of X, Σ, has a positve minimal eigenvalue bounded from below by $\lambda_{\min} > 0$.

Compute a bound for

$$\|\hat{\beta} - \beta^*\|.$$

The bound should depend on d, K and $\lambda_{\min}$, all of which are allowed to change with n. Based on your bound, comment on the dependence on d.

Hint: Recall that $\|Ax\| \leq \|A\|_{\text{op}}\|x\|$, $\|AB\|_{\text{op}} \leq \|A\|_{\text{op}}\|B\|_{\text{op}}$ and that the maximal eigenvaue of Σ^{-1} (which is also its operator norm) is the reciprocal of the minimal eogenvale of Σ. Also, you may find the following result useful (see equation 5.8.2 in the book Matrix Analysis, by Horn and Johnson, 2012): letting $E = \hat{\Sigma} - \Sigma$, if $\|\Sigma^{-1}E\|_{\text{op}} < 1$, we have that

$$\|\hat{\Sigma}^{-1} - \Sigma^{-1}\|_{\text{op}} \leq \|\Sigma^{-1}\|_{\text{op}} \frac{\|\Sigma^{-1}E\|_{\text{op}}}{1 - \|\Sigma^{-1}E\|_{\text{op}}}.$$

You may want to use the matrix Bernstein inequality to get sharer rates.

Note: One should be able to infer this result from the main Theorem in the highly

recommended paper "Random design analysis of ridge regression", by Daniel Hsu, Sham M. Kakade and Tong Zhang, available here. However, presumably, if you follow the hint you should end up with a simpler proof. I am curious to see what rates you get...

4. **Hard thresholding in the sub-gaussian many means problem.** Suppose we observe the vector $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$, where

$$X = \theta^* + \epsilon,$$

with $\theta^* \in \mathbb{R}^d$ unknown and $\epsilon \in SG_d(\sigma^2)$. We would like to estimate θ^* using the hard thresholding estimator $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_d)$ with parameter $\tau > 0$, given by:

$$\hat{\theta}_i = \begin{cases} X_i & \text{if } |X_i| > \tau \\ 0 & \text{if } |X_i| \leq \tau. \end{cases}$$

This estimator either keeps or kills each coordinate of X.

For $\delta \in (0, 1)$, set

$$\tau = 2\sigma\sqrt{2\log(2d/\delta)}.$$

Notice that $\mathbb{P}\left(\max_i |\epsilon_i| > \tau/2\right) \leq \delta$ (If this surprises you, refresh your memory on maximal inequalities).

(a) Prove that the hard-thresholding estimator is the solution the optimization problem

$$\min_{\theta \in \mathbb{R}^d} \|X - \theta\|^2 + \tau^2 \|\theta\|_0.$$

(b) Prove that if $\|\theta^*\|_0 = k$, with probability at least $1 - \delta$,

$$\|\hat{\theta} - \theta^*\|^2 \leq C\sigma^2 k \log(2d/\delta),$$

for some universal constant $C > 0$. *Hint: show that, for each* $i = 1, \ldots, d$

$$|\hat{\theta}_i - \theta^*_i| \leq C' \min\{|\theta^*_i|, \tau\}$$

for some $C' > 0$, *with probability at least* $1 - \delta$.

(c) Compare with the oracle estimator $\hat{\theta}^{\text{or}}$, with coordinates given by

$$\hat{\theta}^{\text{or}}_i = \begin{cases} X_i & \text{if } i \in \text{supp}(\theta^*) \\ 0 & \text{otherwise.} \end{cases}$$

for $i = 1, \ldots, d$. This estimator is of course not computable, as it requires knwoledge of $\text{supp}(\theta^*)$. It is an estimator that an oracle, who has access to this additional knowledge, would be able to compute. Oracle estimators are idealized estimators, which perform at least as well as any computable estimators. Thus, in rder to show that a given estimator performs well, it is enoygh to show that it mimicks closely the performance of an oracle estimator.

(d) Show that if $\min_{i \in \text{supp}(\theta^*)} |\theta_i| > \frac{3}{2}\tau$, then, with probability at least $1 - \delta$,

$$\text{supp}(\hat{\theta}) = \text{supp}(\theta^*).$$

How does $\hat{\theta}$ compare now to the oracle estimator?

5. **Reading Exercise, graded for effort, not correctness.**

   The following paper outlines a general strategy, called primal dual witness construction, for showing model selection consistency for the lasso, aka sparsistency. It means that the LASSO selects the right set of non-zero covariates. It an be extended to other penalized likelihood procedures.

   - Wainwright, M. (2009). Sharp Thresholds for High-Dimensional and Noisy Sparsity Recovery Using ℓ_1-Constrained Quadratic Programming (Lasso), IEEE TRANSACTIONS ON INFORMATION THEORY, VOL. 55, NO. 5, 2183–2202.

   Reproduce the proof of Theorem 1. Notice that the incoherence condition, which is necessary for the result, is a very strong assumption, unlikely to be satisifed in practice. Since this assumption is nearly necessary, conclude that the LASSO in practice should not be expected to be sparistent (even if you believe a linear model!).