**36-709, Spring 2018**
**Homework 4**

Due Friday April 5 by 5:00pm in JaeHyeok's mailbox

1. In earlier works on the lasso, people have used a even stronger assumptions than the restricted eigenvalue property. Here is one. Suppose that the design matrix $X$ is such that, for some integer $k > 0$,

$$\max_{i,j} \left| \frac{X_i^\top X_j}{n} - 1(i = j) \right| \leq \frac{1}{23k} \tag{1}$$

where $X_i$ is the $i$th column of $X$, $i = 1, \ldots, d$. Think about what that means.

   (a) Show that this condition implies that, for any subset $S$ of $\{1, \ldots, d\}$ of cardinality no larger than $k < d$ and any $\Delta \in \mathbb{R}^d$ with $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$,

$$\|\Delta\|^2 \leq \frac{2}{n}\|X\Delta\|^2.$$

   That is, show that this condition implies the $RE(3, 1/2)$ condition given in class for all non-empty subsets $S$ of $\{1, \ldots, d\}$ of size no larger than $k$. *Instead of the constant* 23 *you may take a larger one if it simplifies your calculations.*

   (b) Suppose that the entries of $X$ are now populated by independent Rademacher variables (a Rademacher variable is one that that takes the values $+1$ and $-1$ with equal probability). Show that, for any $\delta \in (0, 1)$, if

$$n \geq Ck^2(\log(d) + \log(1/\delta)),$$

   for some constant $C > 0$, then $X$ satisfies the condition (1), with probability at least $1 - \delta$. *Again, instead of* 23 *feel free to show the result for a different constant if it helps with the calculations.*

2. Exercise 7.13

3. Read the paper "Assumptionless consistency of the lasso", by S. Chatterjee. The paper is available at https://arxiv.org/pdf/1303.5817.pdf. Reproduce the proofs of Theorem 1 and 2. Theorem 1 in particular shows that the lasso is a good method for prediction.

4. **The Lasso and Fals Discoveries.** Read Sections 1-4 of the paper "False Discoveries occur Early on the Lasso Path", by Weijie Su, Malgorzata Bogda and Emmanuel J. Candés, available at https://statweb.stanford.edu/~candes/papers/LassoFDR.pdf. You are not expected to read the proofs, which are based on advanced techniques not covered in the course. Write a paragraph to summarize their findings.

5. **Inference after model selection**.

   (a) Suppose that we observe $n$ independent random variables $(X_1, \ldots, X_n)$ where $X_i \sim N(\mu_i, 1)$ for all $i$. The means $\mu_1, \ldots, \mu_n$ are unknown but we suspect that most of them are zero and some are large in absolute value. We first perform a naive model selection procedure by computing the random set of indexes

$$\hat{I} = \{i \colon |X_i| > 1\},$$

corresponding to the variables that presumably have the largest means in absolute value. This is the model selection part. Then, for any one $i \in \hat{I}$ (assumed non-empty), we test the null hypothesis that $\mu_i = 0$ at the significance level of $\alpha = 0.05$. This the inference part. We decide to ignore the selection step, and use the test that rejects if $|X_i| > z_{\alpha/2}$, the $1 - \alpha/2$ quantile of a standard normal. What is the problem with this choice? What would you suggest to do in order to correctly take into account the selection step?

(b) The bive problems examplifies a much more general phenomenon. Read the paper *A Note on Screening Regression Equation,* by David A. Freedman, published in 1983 in the American Statistician, available here.

6. 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

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

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 $\hat{\theta}_{\text{ridge}}$.

7. **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 $\theta^*$ 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

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

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?

8. 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 $\beta^*$ 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 $\beta^*$ 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 continuous distribution, which implies that $\hat{\Sigma}$ is invertible almost surely if $n \geq d$ (no need to show this fact).

ii. $Y$ takes values in $[-K, K]$ and $X$ is a sub-gassuian random vector with parameter $\sigma^2$;

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

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

The bound should depend on $d$, $K$, $\sigma^2$, $\lambda_{\min}$ and $\lambda_{\max}(\Sigma)$ 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* $\Sigma^{-1}$ *(which is also its operator norm) is the reciprocal of the minimal eogenvale of* $\Sigma$*. 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 andom design analysis of ridge regression", by iel Hsu, Sham M. Kakade and Tong Zhang, available efhttps://arxiv.org/pdf/1106.2363.pdfhere. However, presumably, if you follow the hint you should end up with a simpler proof. I am curious to see what rates you get...*