36-755, Fall 2016
Homework 4

Due Wed Oct 19 by 5:00pm in Jisu's mailbox

1. **A sparse oracle inequality for the lasso**. Consider the following set-up, as described in class. We observe n pairs $(Y_1, x_1), \ldots, (Y_n, x_n)$, where each x_i is a fixed vector in $\mathbb{R}^d$ and

$$Y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

with $\epsilon_1, \ldots, \epsilon_n$ independent variables in $SG(\sigma^2)$. We have a dictionary $(f_1, \ldots, f_M)$ of M functions from $\mathbb{R}^d$ into $\mathbb{R}$ and would like to estimate f using a **sparse** linear combination of such functions. More precisely, for $\theta = (\theta_1, \ldots, \theta_M) \in \mathbb{R}^M$, let $f_\theta = \sum_{j=1}^M \theta_j f_j$. Then, we will estimate f with $f_{\hat\theta}$ where $\hat\theta$ is a lasso solution:

$$\hat\theta \in \text{argmin}_{\theta \in \mathbb{R}^M} \frac{1}{2n} \sum_{i=1}^n (Y_i - f_\theta(x_i))^2 + \lambda_n \|\theta\|_1,$$

for some $\lambda_n \geq 0$. To study the performance of this estimator, we will compared $f_{\hat\theta}$ to the **sparse oracle estimator** f_{θ^*}, where

$$\theta^* \in \text{argmin}_{\theta \in \mathbb{R}^M, \|\theta\|_0 \leq k} \sum_{i=1}^n (f(x_i) - f_\theta(x_i))^2, \tag{1}$$

and $0 < k < M$ is a fixed constant.

We will make the following assumption: let Φ be the $n \times M$ matrix with entries, $\Phi_{i,j} = f_j(x_i)$, and assume that, for some $\kappa > 0$, Φ satisfies the $RE(3, \kappa)$ condition with respect to all non-empty subsets S of $\{1, \ldots, M\}$ of cardinality no larger than k. Show that, if $\lambda_n \geq \frac{2}{n}\|\Phi^\top \epsilon\|_\infty$, then, for any $\alpha \in (0, 1)$,

$$MSE(f_{\hat\theta}) \leq \inf_{\theta \in \mathbb{R}^d, \|\theta\|_0 \leq k} \left\{ \frac{1+\alpha}{1-\alpha} MSE(f_\theta) + \frac{18}{\alpha(1-\alpha)\kappa} \|\theta\|_0 \lambda_n^2 \right\},$$

where, for $\theta \in \mathbb{R}^M$, $MSE(f_\theta) = \frac{1}{n}\sum_{i=1}^n (f_\theta(x_i) - f(x_i))^2$.

Since this proof is similar to the proof of the fast rate for the lasso, you may use facts proved in class and do not need to re-derive them.

Hint: this is an extension of the proof of the fast rates for the lasso. Here are some suggestions:

- *Consider any $\theta \in \mathbb{R}^M$ with $\|\theta\|_0 \leq k$ and start with the inequality*

$$\frac{1}{2n}\left[\sum_{i=1}^n (Y_i - f_{\hat\theta}(x_i))^2 - \sum_{i=1}^n (Y_i - f_\theta(x_i))^2\right] \leq \lambda_n(\|\theta\|_1 - \|\hat\theta\|_1).$$

- *Now substitute $Y_i = f(x_i) + \epsilon_i$ for all i and get a basic inequality.*
- *Continue following the proof of the fast rates for the lasso, using the RE condition.*
- *At some point you will need to use the variational inequality*

$$2xy = \inf_{\gamma > 0} \left(\frac{x^2}{\gamma} + y^2\gamma \right)$$

to get that $2xy \leq \frac{2}{\alpha}x^2 + \frac{\alpha}{2}y^2$, for $\alpha \in (0, 1)$.

- *You will also need to use the inequality* $(x-y)^2 \le 2x^2 + 2y^2$.

2. **Inference after model selection**. 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? If I correctly take into consideration the selection step, what would be a better test?

3. **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:

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

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 $\mathbb{P}\left(\max_i |\epsilon_i| > \tau/2\right) \le \delta$ (If this surprises you, refresh your memory on maximal inequalities and check out HW2!).

(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 $1-\delta$,

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

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

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

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 $i = 1, \ldots, d$.

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

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

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

4. **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. 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.

5. 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 $k > 0$,

$$\max_{i,j} \left| \frac{X_i^\top X_j}{n} - 1(i = j) \right| \leq \frac{1}{14k}$$

where X_i is the ith column of X, $i = 1, \ldots, d$. Think about what that means. Also, see Proposition 7.1 in the book.

(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 $\|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$,

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

(b) This is not quite the $RE(3, 1/2)$ condition given in class. First of all, it holds not just for a fixed $S \subset \{1, \ldots, d\}$ but for all subsets S of cardinality no larger than k. Secondly, in the definition from class the left hand side on (2) should be $\|\Delta\|^2$ instead. Nonetheless, show that a simple modification of the last few steps of the proof of the fast rates for the lasso covered in class will give slightly worse rate (up to constants) using the condition (2). The rate in this case is worse because the rate depends on the L_0-norm of the true regression parameters instead of its squared root. You don't have to reproduce the whole proof, just the last few steps. *Hint: you will need the inequality $\|\Delta\|_2 \leq \|\Delta\|_1$, valid for all vectors Δ.*