ECE598: Information-theoretic methods in high-dimensional statistics Spring 2016

Lecture 19: Minimax rates for sparse linear regression

Lecturer: Yihong Wu *Scribe: Subhadeep Paul, April 13/14, 2016*

In the last lecture we analyzed the k-sparse Gaussian location model in high dimension (the denoising problem) and proved minimax rate for estimating the location parameter. In this lecture we extend the earlier ideas to sparse linear regression in high dimension. We prove a minimax lower bound and then obtain upper bounds on the risk of a few procedures.

19.1 Problem setup: Sparse linear regression

The sparse linear regression model is

$$Y_{n\times 1} = X_{n\times p}\theta_{p\times 1} + Z, \quad Z \sim \mathcal{N}(0, I_n), \tag{19.1}$$

where $X \in \mathbb{R}^{n\times p}$ is the design matrix and $\theta \in \mathbb{R}^p$ is an unknown k-sparse parameter vector. In this lecture we are concerned with the case when $n << p$ but $n \geq k$, i.e., we have more predictors in the design matrix than we have samples.

Interpretation: Y is a noisy linear combination of the columns of the design matrix X. The goal here is to estimate θ, given Y and X. Note that the system has more unknowns than the number of equations and hence is indeterminate even without the noise. Estimation is made possible due to the k-sparsity structure.

Note: We consider here random design matrices only. More precisely we have

$$X_{ij} \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, 1/n),$$

so that the columns have roughly unit norm.

The next theorem proves a lower bound on the minimax risk for estimating θ in the k-sparse regression model.

Theorem 19.1. *The minimax risk for estimating θ in the model defined by Equation (19.1) is lower bounded by*

$$R^* = R^*(p, k, n) = \inf_{\hat{\theta}} \sup_{\|\theta\|_0 \leq k} \mathbb{E}_\theta \|\hat{\theta} - \theta\|_2^2 \gtrsim k \log \frac{ep}{k}, \quad \forall n.$$

Proof. Note that (X_i, Y_i)'s are i.i.d sampled from a distribution P_θ. We show that the KL-diameter for two different θ is exactly same as that in p-dimensional gaussian location model, i.e.,

$$D(P_{\theta_0} \| P_{\theta_1}) = \frac{1}{2}\|\theta_0 - \theta_1\|^2.$$

We derive the result as follows,

$$
\begin{aligned}
D(P_{X_i,Y_i|\theta_0}\|P_{X_i,Y_i|\theta_1}) &= \mathbb{E}_{X_i}[D(P_{Y_i|X_i,\theta_0}\|P_{Y_i|X_i,\theta_1})] \\
&= \mathbb{E}_{X_i}[D(\mathcal{N}(\langle X_i,\theta_0\rangle,1)\|\mathcal{N}(\langle X_i,\theta_1\rangle,1))] \\
&= \mathbb{E}_{X_i}[\frac{1}{2}\langle X_i,\theta_0-\theta_1\rangle^2] \\
&= \frac{1}{2}\mathbb{E}_{X_i}[(\theta_0-\theta_1)'X_i'X_i(\theta_0-\theta_1)] \\
&= \frac{1}{2}(\theta_0-\theta_1)'\mathbb{E}[X_i'X_i](\theta_0-\theta_1) \\
&= \frac{1}{2n}\|\theta_0-\theta_1\|^2.
\end{aligned}
$$

Hence,

$$
D(P_{\theta_0}\|P_{\theta_1}) = D(P_{X_i,Y_i|\theta_0}^{\otimes n}\|P_{X_i,Y_i|\theta_1}^{\otimes n}) = \frac{1}{2}\|\theta_0-\theta_1\|^2.
$$

The result then follows from the analysis of the gaussian location model from previous lecture. □

From the previous discussion we have, $R^* \geq k\log\frac{ep}{k}$ for any n, which is identical to the minimax rate of the denoising problem for any n. This is reasonable, because even with full observation $n \gtrsim p$, which roughly corresponds to the denoising problem we cannot beat this rate. Surprisingly, as long as $n \gtrsim k\log\frac{ep}{k}$, the denoising rate is attainable and we have $R^* \asymp k\log\frac{ep}{k}$, achieved by, e.g., the maximum likelihood estimator. This is proved in Section 19.2.

Note that MLE is computationally expensive. A computable alternative procedure is the Dantzig selector [CT07]. As analyzed in Section 19.3 It is guaranteed to achieve the rate $R^* \lesssim k\log p$ as long as $n \geq k\log\frac{ep}{k}$. This falls slightly sort of the optimal rate. However unlike MLE, the procedure is completely adaptive and can be cast as a linear programming problem. More recently a procedure called SLOPE [SC15] has been proposed which achieves the optimal rate. In particular its risk coincides with the minimax rate with sharp constant, namely, $R^* \leq (2+o(1))k\log\frac{ep}{k}$ as $p\to\infty$ and $k = o(p)$, provided that $n \gtrsim k\log\frac{ep}{k}$.

19.2 Analysis of MLE

The MLE in this case is defined in terms of the solution (may or may not be unique) of an optimization problem,

$$
\hat{\theta}_{\text{MLE}} \in \arg\min_{\|\theta_0\|\leq k} \|Y - X\theta\|_2^2. \tag{19.2}
$$

Unfortunately the optimization problem can only be solved through exhaustive search which is NP-hard in the worst case.

Theorem 19.2. *Whenever $n \geq Ck\log\frac{ep}{k}$ for some sufficiently large constant C, $\forall\theta \in B_0(k)$.*

$$
\|\hat{\theta}_{\text{MLE}} - \theta\|_2^2 \lesssim k\log\frac{ep}{k}, \tag{19.3}
$$

$$
\|X(\hat{\theta}_{\text{MLE}} - \theta)\|_2^2 \lesssim k\log\frac{ep}{k}, \tag{19.4}
$$

hold with high probability.

Proof. Since $\hat{\theta}$ is a minimizer, we have

$$\|Y - X\hat{\theta}\|_2^2 \leq \|Y - X\theta\|_2^2 = \|Z\|_2^2.$$

On the left hand side we have,

$$\|Y - X\hat{\theta}\|_2^2 = \|Y - X\theta + X\theta - X\hat{\theta}\|_2^2 = \|Z - Xh\|_2^2,$$

where $h = \hat{\theta} - \theta$. Hence we have,

$$\|Z - Xh\|_2^2 \leq \|Z\|_2^2,$$

which leads to the basic inequality

$$\begin{aligned}\|Xh\|_2^2 &\leq 2\langle Z, Xh\rangle \\ &= 2Z'Xh \\ &\leq 2\|h\|_2 \sup_{u \in S^{p-1} \cap B_0(2k)} Z'Xu.\end{aligned}$$

Note that the left hand side is not the estimation error, instead it is the prediction error $\|Xh\|_2^2 = \|X\hat{\theta} - X\theta\|_2^2$. Hence to conclude both (19.2) and (19.4) from the basic inequality, it suffices to show

(a) $\|h\|_2 \lesssim \|Xh\|_2$, (Restricted isometry property)

(b) $\sup_{u \in S^{p-1} \cap B_0(2k)} Z'Xu \lesssim \sqrt{k \log \frac{ep}{k}}$, with high probability.

We first prove (b).

$$\sup_{u \in S^{p-1} \cap B_0(2k)} Z'xu = \|Z\| w(G) \lesssim \sqrt{k \log \frac{ep}{k}},$$

where $w(G)$ is the Gaussian width of the set $S^{p-1} \cap B_0(2k)$ and from last lecture we know, $w(G) \lesssim \sqrt{k \log \frac{ep}{k}}$.

For (a) we will show that

$$\inf_{\|h\|_0 \leq k} \frac{\|Xh\|_2}{\|h\|_2} \gtrsim c \quad \text{if } n \gtrsim k \log \frac{ep}{k},$$

where c is a constant. First note that,

$$\inf_{u \neq 0} \frac{\|Au\|_2}{\|u\|_2} = \sigma_{\min}(A).$$

For any feasible h, $Xh = X_J h_J$, where $J = supp(h)$ is the support of h and X_J is the $n \times k$ matrix whose columns are the columns of X that corresponds to the rows in the support J. Then we have

$$\inf_{\|h\|_0 \leq k} \frac{\|Xh\|_2}{\|h\|_2} = \min_{|J| \leq k} \sigma_{\min}(X_J).$$

For a fixed J, $\sigma_{\min}(X_J)$ concentrates to $1 - \sqrt{\frac{k}{n}}$. Hence an union bound gives,

$$\mathbb{P}\left[\min_{|J| \leq k} \sigma_{\min}(X_J) < t\right] \leq \binom{p}{k} \mathbb{P}[\sigma_{\min}(X_{[k]}) < t].$$

Using the tail bound

$$\mathbb{P}\left[\sigma_{\min}(X_{[k]}) < 1 - \sqrt{\frac{k}{n}} - \frac{t}{\sqrt{n}}\right] \le \exp(-t^2/2),$$

and choosing $t = 4k \log \frac{ep}{k}$ and $n = 100k \log \frac{ep}{k}$, we have $1 - \sqrt{\frac{k}{n}} - \frac{t}{\sqrt{n}} \ge 0.5$ and consequently, $\mathbb{P}[\min_{|J|\le k} \sigma_{\min}(X_J) < 0.5] \to 0$. This completes the proof. □

19.3 Dantzig selector

The Dantzig selector can written as the following optimization problem,

$$\min \|\theta\|_1, \quad \text{s.t } \|X'(Y - X\theta)\|_\infty \le \tau. \tag{19.5}$$

This optimization problem can be efficiently solved as a linear programming. Another computable procedure for k-sparse regression is the Lasso, which can be written as the following optimization problem

$$\min \|Y - X\theta\|_2^2 + \lambda\|\theta\|_1, \quad \theta \in \mathbb{R}^n. \tag{19.6}$$

Note that X' is added to the constraint in the Dantzig selector in (19.5) to make the solution rotation-invariant. Precisely, if $U \in O(n)$ be a $n \times n$ orthogonal rotation matrix, then $UY = UX\theta + UZ$. Note that in this case, $\hat{\theta}(X, Y) = \hat{\theta}(UX, UY)$.

Theorem 19.3. *Let the Dantzig selector $\hat{\theta}_{\text{DS}}$ denote a minimizer of (19.5) we have*

$$\|\hat{\theta}_{\text{DS}} - \theta\|_2^2 \lesssim k \log p,$$

w.h.p as long as $n \ge Ck \log \frac{ep}{k}$ for some sufficiently large constant C.

Proof. Similar to the proof of Theorem **??** in the denoising problem, the proof is divided in three steps.

- *Step 1 : set τ to guarantee θ is feasible*

  We choose,

  $$\|X'Z\|_\infty \le \tau = \sqrt{2 \log p},$$

  so that ground truth is feasible.

- *Step 2: Structure of the error $h = \hat{\theta} - \theta$*

  Let J be the support of θ. Define the cone

  $$C_J \triangleq \{h : \|h_{J^c}\|_1 \le \|h_J\|_1\}. \tag{19.7}$$

  Since $\|\hat{\theta}\|_1 \le \|\theta\|_1$, we have $h \in C_J$.

  Claim 19.1.

  $$\|X'Xh\|_\infty \le 2\tau$$

To see this note that,

$$\begin{aligned}\|X'Xh\|_\infty &= \|X'X(\hat{\theta}-\theta)\|_\infty \\ &= \|X'(Y-X\theta)-X'(Y-X\hat{\theta})\|_\infty \\ &\le \|X'(Y-X\theta)\|_\infty + \|X'(Y-X\hat{\theta})\|_\infty \\ &\le 2\tau.\end{aligned}$$

- *Step 3: The risk*

  We have,

$$\begin{aligned}\|Xh\|_2^2 &= \langle Xh, Xh\rangle \\ &= h'X'Xh \\ &= \left\langle X'Xh, h\right\rangle \\ &\le \|X'Xh\|_\infty \|h\|_1 \quad \text{(Holder)} \\ &\le 2\tau 2\|h_J\|_1 \\ &\le 4\sqrt{k}\tau\|h_J\|_2 \quad \text{(Cauchy-Schwarz)} \\ &\le 4\sqrt{k}\tau\|h\|_2.\end{aligned}$$

  Now we need to show one last thing to complete the proof.

  Claim 19.2.

$$\|h\|_2^2 \lesssim \|Xh\|_2^2 \quad w.h.p \ \forall h \in C_J$$

  To prove this claim we use the special feature of the cone C_J defined in (19.7), that for any $h \in C_J$, half of the energy of h is on $2k$ co-ordinates, i.e. h is almost $2k$- sparse.

  Suppose h is ordered in the following fashion: The vector of length k that corresponds to $J = supp(\theta)$, h_J comes first. The rest of h is ordered in terms of decreasing magnitude. We divide the remaining h after first block into blocks of size k and name the blocks $K_1, K_2, \ldots$ and the vectors $h_1, h_2, \ldots$, such that $h_{K_i} = h_i$.

  Let $a = h_{J\cup K_1} = h_J + h_1$. By construction, it has more than $1/2$ of the energy, i.e.,

$$\|a\|_2^2 \ge \frac{1}{2}\|h\|_2^2$$

  and define,

$$b \triangleq h_{(J\cup K_1)^c} = \sum_{i\ge 2} h_i.$$

  Then

$$\begin{aligned}\|Xh\|_2^2 &= \|Xa + Xb\|_2^2 \\ &\ge \|Xa\|_2^2 + 2\langle Xa, Xb\rangle \end{aligned} \tag{19.8}$$

  Since X satisfies the restricted isometry property, for $n \ge ck\log\frac{ep}{k}$, there exists $c_1(c)$ with $c_1 \to 1$ if $c \to \infty$, such that

$$\|Xa\|_2^2 \ge c_1\|a\|_2^2 \ge \frac{c_1}{2}\|h\|_2^2.$$

  Now we need to just show that the cross term $\langle Xa, Xb\rangle$ is small in magnitude. For this we use the following restricted decorrelation lemma,

Lemma 19.1. *Let $n \geq ck \log \frac{ep}{k}$. Then, with high probability, for all $u, v \in B_0(2k)$ we have,*

$$|\langle Xa, Xb \rangle| \leq c_2 \|u\| \|v\|,$$

where $c_2 = c_2(c)$ and $c_2 \to 0$ if $c \to \infty$.

Then we have,

$$\begin{aligned}
\langle Xa, Xb \rangle &= \sum_{j \geq 2} \langle Xa, Xh_j \rangle \\
&\leq \sum_{j \geq 2} \langle Xa, hX_j \rangle \\
&\leq \sum_{j \geq 2} \|a\|_2 \|h_j\|_2 \quad \text{(Lemma 19.1)} \\
&\leq c_2 \|h\|_2 \sum_{j \geq 2} \sqrt{k} \|h_j\|_\infty \\
&\leq c_2 \|h\|_2 \sum_{j \geq 2} \sqrt{k} \frac{\|h_{j-1}\|_1}{k} \quad \text{(By ordering)} \\
&\leq \frac{c_2}{\sqrt{k}} \|h\|_2 (\sum_{j \geq 2} \|h_{j-1}\|_1) \\
&\leq \frac{c_2}{\sqrt{k}} \|h\|_2 \|h_{J^c}\|_1 \quad \text{Property of cone} \\
&\leq \frac{c_2}{\sqrt{k}} \|h\|_2 \|h_J\|_1 \\
&\leq \frac{c_2}{\sqrt{k}} \|h\|_2^2.
\end{aligned}$$

Reverting back to (19.8) we have,

$$\|Xh\|_2^2 \geq (\frac{c_1}{2} - c_2) \|h\|_2^2 \gtrsim \|h\|_2^2.$$

This completes the proof.

□

Remark 19.1 (Adaptivity issues)**.** Note that the Dantzig selector procedure is adaptive to k, but not to σ. To see this consider the following high dimensional k-sparse regression problem,

$$Y = X\theta + Z, \quad Z \sim \mathcal{N}(0, \sigma^2 I_n).$$

If σ is known then we can set $\tau = \sigma\sqrt{2 \log p}$, but typically σ is not known.

A similar problem arises with Lasso as well. In (19.6), if σ is known then the optimal $\lambda = 2\sigma\sqrt{\log p}$, but if σ is unknown then λ is a tuning parameter. As a remedy for this another procedure called square root Lasso was proposed which can be written as the following optimization problem,

$$\min \|Y - X\theta\|_2 + \lambda \|\theta\|_1, \quad \theta \in \mathbb{R}^n.$$

The optimal $\lambda = \sqrt{\log p}$ even when σ is unknown. However the downside is that this optimization problem is not easy to solve.

References

[CT07] Emmanuel Candés and Terence Tao. The Dantzig selector: statistical estimation when p is much larger than n. pages 2313–2351, 2007.

[SC15] Weijie Su and Emmanuel Candés. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *arXiv preprint arXiv:1503.08393*, 2015.