Stats 300b: Theory of Statistics **Winter 2019**

Lecture 15 – February 26

Lecturer: John Duchi *Scribe: Dan Kluger*

Warning: *these notes may contain factual errors*

Reading: There is no reading corresponding to this lecture.

Oultline:

- Gaussian Sequence Models
  - Hard Thresholding
  - Soft Thresholding
- Basis Pursuit/Noiseless recovery
  - l_1 -relaxations
  - Isometry properties of matrices

1 Gaussian Sequence Model Recap

Recall the **Gaussian Sequence Model** that $Y = \theta + \sigma\epsilon$ where $\theta \in \mathbb{R}^n$ and $\epsilon \sim N(0, I_n)$.

Question: When can we recover θ to reasonable accuracy?

Answer: When using structural (sparsity) assumptions on θ.
Assume: θ is k-sparse, meaning that $||\theta||_0 \equiv \sum_{j=1}^{n} \mathbf{1}(\theta_j \neq 0) \leq k$.
Goal: Use k-sparse assumption on θ to achieve better MSE than the naive estimator $\hat{\theta}^{naive} = Y$

2 Hard Thresholding for Gaussian Sequence Model

For the Gaussian Sequence Model, a hard thresholding estimator is an estimator given by

$$\hat{\theta}_j = \begin{cases} Y_i & \text{if } |Y_j| > \tau \\ 0 & \text{if } |Y_j| \leq \tau \end{cases}$$

for some threshold $\tau \geq 0$.

Idea: Since $||\epsilon||_\infty \lesssim \sqrt{2log(n)}$ we can set $\tau \approx \sigma\sqrt{2log(n)}$, and in such a case any non-zero entries of $\hat{\theta}$ should be "true" non-zero entries in θ.

2.1 Upper bound on l_2 risk of the Hard Thresholding Estimator

We will now compute an upper bound on the l_2 risk of the Hard thresholding estimator for an arbitrary τ. To do this let $S = \{j \in [n] \ : \ \theta_j \neq 0\}$ (i.e let S be the support of θ). We will first find an upperbound on $E[(\hat{\theta}_j - \theta_j)^2]$ for $j \in S$. For $j \in S$,

$$E[(\hat{\theta}_j - \theta_j)^2] \leq \underbrace{E[(Y_j - \theta_j)^2]}_{=\sigma^2} + \theta_j^2 \underbrace{P(|Y_j| \leq \tau)}_{\equiv T_2}$$

We will now find an upper bound on T_2. To do so first consider the case where $\theta_j \geq \tau$. Note that in this case

$$|\theta_j + \sigma\epsilon_j| \leq \tau \Rightarrow \theta_j + \sigma\epsilon_j \leq \tau \Rightarrow \theta_j - \tau \leq -\sigma\epsilon_j \Rightarrow (|\theta_j| - \tau)_+ \leq -\sigma\epsilon_j$$

Thus noting that $-\sigma\epsilon_j \sim N(0, \sigma^2)$, and thus $-\sigma\epsilon_j$ is σ^2-subgaussian, by Chernoff's bound and the fact that $|\theta_j + \sigma\epsilon_j| \leq \tau \Rightarrow (|\theta_j| - \tau)_+ \leq -\sigma\epsilon_j$ in the case where $\theta_j \geq \tau$ we have that

$$P(|\theta_j + \sigma\epsilon_j| \leq \tau) \leq P\Big(-\sigma\epsilon_j \geq (|\theta_j| - \tau)_+\Big) \leq \exp\left(\frac{-(|\theta_j| - \tau)_+^2}{2\sigma^2}\right)$$

In the case where $\theta_j \leq -\tau$, by similar reasoning we can also show $P(|\theta_j + \sigma\epsilon_j| \leq \tau) \leq \exp\left(\frac{-(|\theta_j| - \tau)_+^2}{2\sigma^2}\right)$, and finally this inequality holds trivially in the case where $|\theta_j| < \tau$. Thus no matter what value θ_j takes on

$$T_2 \equiv P(|Y_j| \leq \tau) = P(|\theta_j + \sigma\epsilon_j| \leq \tau) \leq \exp\left(\frac{-(|\theta_j| - \tau)_+^2}{2\sigma^2}\right)$$

Fact 1. *For $u \geq 0$, there exists a constant C_1 such that* $u^2 \exp\left(\frac{-(u-\tau)_+^2}{2\sigma^2}\right) \leq C_1(\tau^2 + \sigma^2)$

Proof Let $u \geq 0$. Note that by convexity of $y \mapsto (y)_+^2$,

$$u^2 = (u - \tau + \tau)_+^2 = 4\Big(\frac{1}{2}(u - \tau) + \frac{1}{2}\tau\Big)_+^2 \leq 2(u - \tau)_+^2 + 2\tau^2$$

Thus

$$\begin{aligned} u^2 \exp\left(\frac{-(u - \tau)_+^2}{2\sigma^2}\right) \leq & 2(u - \tau)_+^2 \exp\left(\frac{-(u - \tau)_+^2}{2\sigma^2}\right) + 2\tau^2 \exp\left(\frac{-(u - \tau)_+^2}{2\sigma^2}\right) \\ \leq & 2\left(\sup_v v^2 \exp\left(\frac{-v^2}{2\sigma^2}\right)\right) + 2\tau^2 \exp\left(\frac{-(u - \tau)_+^2}{2\sigma^2}\right) \\ \leq & 4\sigma^2 e^{-1} + 2\tau^2 \end{aligned}$$

where the last inequality holds because we can show $\sup_v v^2 \exp(\frac{-v^2}{2\sigma^2}) = 2\sigma^2 e^{-1}$ by taking the log and taking derivatives and noting the expression is maximized for $v^2 = 2\sigma^2$. Letting $C_1 = 3$, we have thus shown for $u \geq 0$, $u^2 \exp\left(\frac{-(u-\tau)_+^2}{2\sigma^2}\right) \leq C_1(\sigma^2 + \tau^2)$

□

Putting the previous results together and using the above fact we have that for any $j \in S$,

$$E[(\hat{\theta}_j - \theta_j)^2] \leq \sigma^2 + T_2 \leq \sigma^2 + |\theta_j|^2 \exp\left(\frac{-(|\theta_j| - \tau)_+^2}{2\sigma^2}\right) \leq \sigma^2 + C_1(\sigma^2 + \tau^2)$$

Now for $j \notin S$ note

$$\begin{aligned}
E[(\hat{\theta}_j - \theta_j)^2] =& E\Big[|\sigma\epsilon_j|^2 \mathbf{1}(|\epsilon_j| \geq \frac{\tau}{\sigma})\Big] \\
\leq& \sqrt{E\Big[\sigma^4\epsilon_j^4\Big] P\Big(|\epsilon_j| \geq \frac{\tau}{\sigma}\Big)} && \text{(by Cauchy Schwartz)} \\
=& \sqrt{3}\sigma^2 \sqrt{P\Big(|\epsilon_j| \geq \frac{\tau}{\sigma}\Big)} && \text{(Using 4th moment of a Gaussian)} \\
\leq& \sqrt{3}\sigma^2 \sqrt{2\exp\left(\frac{-\tau^2}{2\sigma^2}\right)} && \text{(since } \epsilon_j \text{ is 1-sub-Gaussian)} \\
=& \sqrt{6}\sigma^2 \exp\left(\frac{-\tau^2}{4\sigma^2}\right)
\end{aligned}$$

Thus the complete l_2 risk (MSE) for hard thresholding is bounded above by

$$\begin{aligned}
E[||\hat{\theta} - \theta||_2^2] =& \sum_{j \in S} E[(\hat{\theta}_j - \theta_j)^2] + \sum_{j \in S^c} E[(\hat{\theta}_j - \theta_j)^2] \\
\leq& \sum_{j \in S} \Big(\sigma^2 + C_1(\sigma^2 + \tau^2)\Big) + \sum_{j \notin S^c} \sqrt{6}\sigma^2 \exp\left(\frac{-\tau^2}{4\sigma^2}\right) \\
\leq& \sigma^2|S| + C_1|S|(\sigma^2 + \tau^2) + C_1|S^c|\sigma^2 \exp\left(\frac{-\tau^2}{4\sigma^2}\right) \\
\leq& k\sigma^2 + C_1 k(\tau^2 + \sigma^2) + C_1 n\sigma^2 \exp\left(\frac{-\tau^2}{4\sigma^2}\right)
\end{aligned}$$

Thus we have an upper bound on $E[||\hat{\theta} - \theta||_2^2]$ when $\hat{\theta}$ is a hard thresholding estimator with $\tau \geq 0$. This upper bound can be used to immediately prove the following theorem.

Theorem 2. *Let $\hat{\theta}$ be a hard thresholding estimator with $\tau = 2\sigma\sqrt{log(\frac{n}{k})}$. Then*

$$\sup_{||\theta||_0 \leq k} E[||\hat{\theta} - \theta||_2^2] \leq Ck\sigma^2\Big(1 + log(\frac{n}{k})\Big)$$

for some numerical constant $C < \infty$

Proof Letting $\tau = 2\sigma\sqrt{log(\frac{n}{k})}$, simply plug this value into the derived inequality that

$$E[||\hat{\theta} - \theta||_2^2] \leq k\sigma^2 + C_1 k(\tau^2 + \sigma^2) + C_1 n\sigma^2 \exp\left(\frac{-\tau^2}{4\sigma^2}\right)$$

and note that the above inequality holds for any θ such that $||\theta||_0 \leq k$. $\square$

Note: This hard thresholding estimator with $\tau = 2\sigma\sqrt{log(\frac{n}{k})}$ is unimprovable and minimax optimal.

3 Soft Thresholding for Gaussian Sequence Model

Idea: Instead of just chopping of observations in Y, let's shrink them.

Definition 3.1. *Define the **soft thresholding operator** to be given by*

$$S_\lambda(v) \equiv sgn(v)(|v| - \lambda)_+ = \underset{u\in\mathbb{R}}{\operatorname{argmin}}\Big\{\frac{1}{2}(u-v)^2 + \lambda|u|\Big\}$$

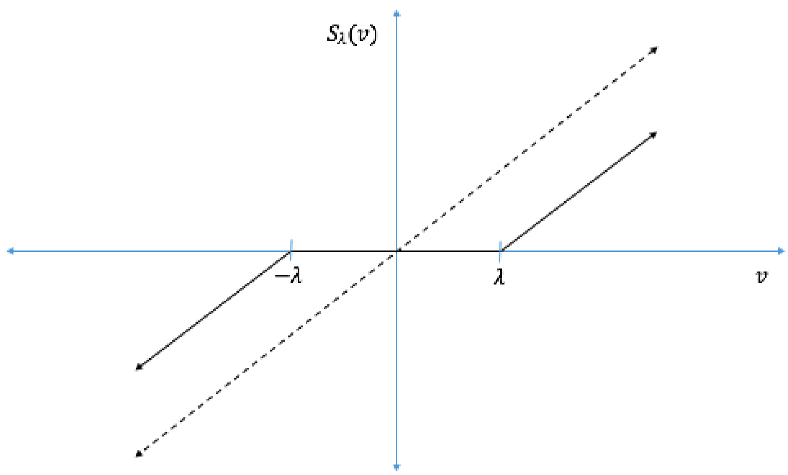


Figure 1: A plot of $S_\lambda(v)$

Definition 3.2. *Define the **soft thresholding estimator** to be given by*

$$\hat{\theta} \equiv S_\lambda(Y) = \underset{u\in\mathbb{R}^2}{\operatorname{argmin}}\Big\{\frac{1}{2}||u-Y||_2^2 + \lambda||u||_1\Big\}$$

Theorem 3. *If $\hat{\theta}$ is a soft thresholding estimator for the Gaussian Sequence Model, the choice $\lambda = \sqrt{2\sigma^2 log(\frac{n}{k})}$ yields $E[||\hat{\theta}-\theta||_2^2] \leq Ck\sigma^2(1+log(\frac{n}{k}))$ if θ is k-sparse. (For sharp constants, see Johnstone 2108 monograph)*

Proof

For $\theta_j = 0$,

$$\begin{aligned}
E[(\hat{\theta}_j - \theta_j)^2] =& E[(\sigma|\epsilon_j| - \lambda)_+^2] \\
=& \int_0^\infty P\Big((\sigma|\epsilon_j| - \lambda)_+^2 > a\Big)da \\
\leq& \int_0^\infty P\Big(\sigma|\epsilon_j| - \lambda \geq \sqrt{a}\Big)da \\
=& 2\int_\lambda^\infty (t-\lambda)P(\sigma|\epsilon_j| \geq t)dt \qquad \text{(letting } t = \sqrt{a}+\lambda) \\
\leq& 2\int_\lambda^\infty tP(\sigma|\epsilon_j| \geq t)dt \\
\leq& 4\int_\lambda^\infty t\exp\Big(\frac{-t^2}{2\sigma^2}\Big)dt \qquad \text{(since } \sigma\epsilon_j \sim N(0,\sigma^2)) \\
=& -4\sigma^2\exp\Big(\frac{-t^2}{2\sigma^2}\Big)\Big|_{t=\lambda}^{t=\infty} \\
=& 4\sigma^2\exp\Big(\frac{-\lambda^2}{2\sigma^2}\Big)
\end{aligned}$$

While for $\theta_j \neq 0$, since S_λ is 1-Lipschitz,

$$\begin{aligned}
E[(\hat{\theta}_j - \theta_j)^2] =& E\Big[\Big(\hat{\theta}_j - S_\lambda(\theta_j) + S_\lambda(\theta_j) - \theta_j\Big)^2\Big] \\
\leq& 2E\Big[(\hat{\theta}_j - S_\lambda(\theta_j))^2\Big] + 2(S_\lambda(\theta_j) - \theta_j)^2 && \text{(since } (a+b)^2 \leq 2a^2 + 2b^2) \\
=& 2E\Big[(S_\lambda(Y_j) - S_\lambda(\theta_j))^2\Big] + 2(S_\lambda(\theta_j) - \theta_j)^2 \\
\leq& 2E\Big[(Y_j - \theta_j)^2\Big] + 2\lambda^2 && \text{(since } S_\lambda \text{ is 1 -Lipschitz)} \\
=& 2\sigma^2 + 2\lambda^2
\end{aligned}$$

Thus combining these two cases and using our choice $\lambda = \sqrt{2\sigma^2 log(\frac{n}{k})}$, we get

$$\begin{aligned}
E[||\hat{\theta} - \theta||_2^2] \leq& 2k(\sigma^2 + \lambda^2) + 4n\sigma^2 \exp\Big(\frac{-\lambda^2}{2\sigma^2}\Big) \\
=& 2k\Big(\sigma^2 + 2\sigma^2 log(\frac{n}{k})\Big) + 4k\sigma^2 \\
=& Ck\sigma^2(1 + log(\frac{n}{k}))
\end{aligned}$$

for some constant $C < \infty$. $\square$

4 Sparse Solutions to Linear Equations

Suppose we have observations Y given by $Y = X\theta$, where $X \in \mathbb{R}^{n \times d}$, $d \gg n$.

Hope : If θ is structured (i.e. sparse), it can hopefully be recovered.

Example 1: : (Signal Processing) Consider an example with observation points $t_1, t_2, ..., t_n$ and frequencies $\omega_1, \omega_2, ..., \omega_d$ and a matrix $X \in \mathbb{R}^{n \times d}$ given by

$$X = \begin{bmatrix} cos(\omega_1 t_1) & cos(\omega_2 t_1) & \dots & cos(\omega_d t_1) \\ cos(\omega_1 t_2) & cos(\omega_2 t_2) & \dots & cos(\omega_d t_2) \\ \vdots & \vdots & \ddots & \vdots \\ cos(\omega_1 t_n) & cos(\omega_2 t_n) & \dots & cos(\omega_d t_n) \end{bmatrix}$$

Our observation $Y = X\theta$ will be the observation of superpositions of sinusoids at times $t_1, t_2, ..., t_n$. Note that for a true continuous signal, $Y(t) = \sum_{j=1}^{d} \theta_j cos(\omega_j t)$. If $||\theta||_0 \leq n$, maybe it is possible to recover θ. ♣

Idea: Find the sparsest solution to $Y = X\theta$. This is equivalent to solving the optimization problem

$$\begin{aligned} &\text{minimize} && ||\theta||_0 \\ &\text{subject to} && Y = X\theta \end{aligned}$$

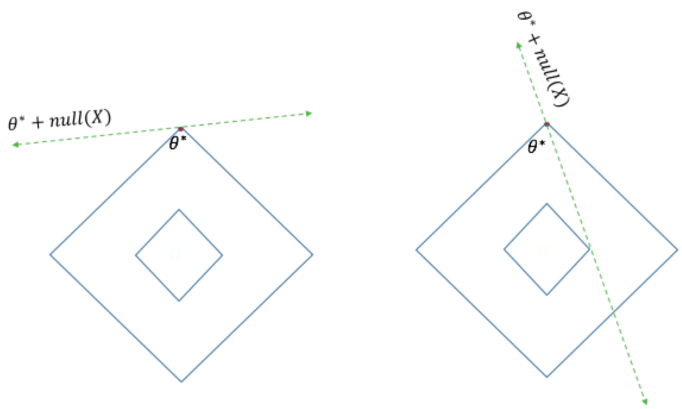
The problem is that this optimization problem is computationally intractable. One possible solution to this issue is to replace $|| \cdot ||_0$ with a convex approximation such as $|| \cdot ||_1$.

Definition 4.1. *The* ***basis pursuit linear program*** *(Chen, Donoho, Saunders 1998) is the following optimization problem*

$$\begin{aligned} &\textit{minimize} \quad ||\theta||_1 \\ &\textit{subject to} \quad Y = X\theta \end{aligned}$$

Question: If θ^* minimizes $||\theta||_0$ subject to $Y = X\theta$, does the basis pursuit linear program recover θ^*?

Answer: Sometimes it works. In figure 2a and figure 2b, the diamonds are l_1 balls, and θ^* lies on the corner of an l_1 ball to indicate it is sparse. Whether or not the the basis pursuit linear program recovers θ^* depends on the null space of the matrix X, because the output to the basis pursuit linear program will find the minimizer of the l_1 norm in the affine subspace $\theta^* + null(X)$. Thus figure 2a represents the cases in which the basis pursuit linear program will succeed in recovering θ^*, while figure 2b represents the cases in which the basis pursuit linear program will fail to recover θ^*.



We will formalize this with some definitions and a theorm.

Definition 4.2. *For a set* $S \subseteq \{1, ..., d\}$ *the* ***critical cone*** *is the subset of* $\mathbb{R}^d$ *given by*

$$\mathbb{C}(S) \equiv \left\{ \Delta \in \mathbb{R}^d \ : \ ||\Delta_{S^c}||_1 \leq ||\Delta_S||_1 \right\}$$

Definition 4.3. *A matrix X is said to satisfy* ***the restricted null spaces property*** *with respect to S if*

$$Null(X) \cap \mathbb{C}(S) = \{0\}$$

where $Null(X) \equiv \left\{ \Delta \in \mathbb{R}^d \ : \ X\Delta = 0 \right\}$

Intuition: If S is the support of θ^* (i.e. $S = \{j \ : \ \theta_j \neq 0\}$) and X satisfies the restricted null spaces property (w.r.t. S) moving from θ^* along $null(X)$ increases the l_1 norm. Figure 2a corresponds to the case where X satisfies restricted null spaces property with respect to S, where S is the support of θ^*.

Theorem 4. *The following two statements are equivalent:*

(1) X satisfies the restricted null spaces property with respect to S

(2) For any θ^* such that $\operatorname{supp} \theta^* = S$ and $Y = X\theta^*, \theta^*$ is the unique solution to basis pursuit linear program

Proof To show (1) $\Rightarrow$ (2), assume (1) holds and let $\hat{\theta}$ be a solution to the basis pursuit linear program and let θ^* satisfy $\operatorname{supp}\theta^* = S$ and $Y = X\theta^*$. Now define Δ so that $\hat{\theta} = \theta^* + \Delta$. We will show that $\Delta \in Null(X) \cap \mathbb{C}(S)$ and hence by (1), $\Delta = 0$. To show this first note that

$$\begin{aligned}
||\theta^*_S||_1 =& ||\theta^*||_1 \\
\geq& ||\hat{\theta}||_1 && \text{(since } \hat{\theta} \text{ minimizes } ||\theta||_1 \text{ subject to } Y = X\theta) \\
=& ||\theta^* + \Delta||_1 \\
=& ||\theta^*_S + \Delta_S||_1 + ||\Delta_{S^c}||_1 && \text{(by decomposition of } l_1\text{-norm)} \\
\geq& ||\theta^*_S||_1 - ||\Delta_S||_1 + ||\Delta_{S^c}||_1 && \text{(by the triangle inequality)}
\end{aligned}$$

Adding $||\Delta_S||_1 - ||\theta^*_S||_1$ to each side we get $||\Delta_{S^c}||_1 \leq ||\Delta_S||_1$, and thus $\Delta \in \mathbb{C}(S)$.

To show $\Delta \in Null(X)$ note that $Y = X\theta^*$ and also $Y = X\hat{\theta}$. Thus

$$0 = Y - Y = X(\hat{\theta} - \theta^*) = X\Delta \Rightarrow \Delta \in Null(X)$$

Thus since $\Delta \in Null(X) \cap \mathbb{C}(S)$, and since by (1), $Null(X) \cap \mathbb{C}(S) = \{0\}$, we have that $\Delta = 0$. Thus $\theta^* = \hat{\theta}$. Hence if (1) holds then for any θ^* such that $\operatorname{supp}\theta^* = S$ and $Y = X\theta^*, \theta^*$ is the unique solution to basis pursuit linear program.

Showing that (2) $\Rightarrow$ (1) will be an exercise left to the reader.

$\square$