IDS.160 – Mathematical Statistics: A Non-Asymptotic Approach

Lecturer: P Rigollet
Scribe: Kayhan Behdin, Wei Fang

Lecture 7
Feb. 27, 2020

Goals: In the last lecture we introduced the linear regression model with fixed design, and provided solutions using least-squares and constrained least-squares estimators. In this lecture, we continue to analyze linear regression. Specifically, we assume that the underlying model in the regression is sparse (i.e. has many zeros). We consider two scenarios where we know the true sparsity or we do not, and provide analysis for each case.

1. SPARSITY IN LINEAR REGRESSION

Recall that the linear regression model with subGaussian noise is

$$Y = \mathbb{X}\theta^* + \varepsilon,$$

where $\varepsilon^\top u \sim \mathsf{subG}(\sigma^2|u|_2^2)\ \forall u \in \mathbb{R}^d$. In this section, we consider the case that θ^* is s-sparse, meaning that it has s non-zero coordinates where $s \ll d$. Additionally, for this first section we assume that s is known.

In the last lecture, we introduced constrained least-squares estimators by restricting θ^* inside the ℓ_1 ball. In this lecture, rather than the ℓ_1 ball we consider the ℓ_0 "ball". Formally we define the ℓ_0 ball as

$$\mathcal{B}_0(s) \coloneqq \{\theta \in \mathbb{R}^d : \sum_{j=1}^{d} \mathbb{I}(\theta_j \neq 0) \leq s\},$$

and if $\theta^* \in \mathcal{B}_0(s)$ we say θ^* is s-sparse. $\mathcal{B}_0(s)$ includes all vectors in $\mathbb{R}^d$ with up to s non-zero coordinates. Another way to describe this space is to view it as an union of subspaces of dimension s that are aligned with the coordinate axes. As an example, we can see how this relates to linear regression $y = \sum_{j=1}^{d} \theta_j^* X_j + \varepsilon$ by observing that in this model, $\theta_j^* = 0 \iff$ "X_j does not enter the regression".

With the definition of $\mathcal{B}_0(s)$, we consider the estimator $\hat{\theta}_{\mathcal{B}_0(s)}$ where

$$\hat{\theta}_{\mathcal{B}_0(s)} \in \operatorname*{argmin}_{\theta \in \mathcal{B}_0(s)} |Y - \mathbb{X}\theta|_2^2.$$

To analyze this estimator, we again utilize the basic inequality

$$|Y - \mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)}|_2^2 \leq |Y - \mathbb{X}\theta^*|_2^2.$$

Rearranging we get

$$|\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)} - \mathbb{X}\theta^*|_2^2 \leq 2\langle \varepsilon, \mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)} - \mathbb{X}\theta^* \rangle.$$

Then we use the fixed point trick, as we did in the previous lecture, giving us

$$|\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)} - \mathbb{X}\theta^*|_2 \leq 2\langle \varepsilon, \frac{\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)} - \mathbb{X}\theta^*}{|\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)} - \mathbb{X}\theta^*|_2} \rangle.$$

Next, we control the supremum in the ℓ_0 ball:

$$|\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)} - \mathbb{X}\theta^*|_2^2 \leq 4\big(\max_{\substack{S\subset[d]\\ |S|\leq 2s}} \sup_{v:\ \mathrm{supp}(v)\subset S} \langle \varepsilon, \frac{\mathbb{X}v}{|\mathbb{X}v|_2}\rangle^2\big).$$

Now, similar to the previous lecture, we introduce $\Phi_S \in \mathbb{R}^{n\times 2s}$, consisting of an orthonormal basis for the span of the columns of $\mathbb{X}$, $\{\mathbb{X}_j : j \in S\}$. Note that when the span is rank-deficient, we pad the basis up to $2s$. We can now write $\mathbb{X}v = \Phi_S\nu$, where ν are the coordinates in this new coordinate system, and in particular, we know that $|\Phi_S\nu|_2 = |\nu|_2$, thus

$$\sup_{v:\ \mathrm{supp}(v)\subset S} \langle \varepsilon, \frac{\mathbb{X}v}{|\mathbb{X}v|_2}\rangle^2 \leq \sup_{\nu\in\mathcal{B}_2(\mathbb{R}^{2s})} \langle \Phi_S^\top\varepsilon, \nu\rangle^2.$$

With this inequality, we can now bound the MSE using the union bound:

$$\begin{aligned}
\mathbb{P}[|X\hat{\theta}_{\mathcal{B}_0(s)} - X\theta^*|_2^2 > t] &\leq \mathbb{P}[4\max_{|S|=2s}\sup_{\nu\in\mathcal{B}_2(\mathbb{R}^{2s})}\langle\Phi_s^\top\varepsilon,\nu\rangle^2 > t] \\
&\leq \sum_{|S|=2s}\mathbb{P}[\sup_{\nu\in\mathcal{B}_2(\mathbb{R}^{2s})}\langle\Phi_s^\top\varepsilon,\nu\rangle > \frac{\sqrt{t}}{2}] \\
&\leq \binom{d}{2s}\exp(-\frac{t}{32\sigma^2} + 2s\log 5) \\
&= \exp\big(-\frac{t}{32\sigma^2} + 2s\log 5 + \log\binom{d}{2s}\big).
\end{aligned}$$

In the third inequality, notice that $\tilde{\varepsilon} = \Phi_s^\top\varepsilon$, $\tilde{\varepsilon}^\top u \sim \mathsf{subG}(\sigma^2|u|_2^2)\ \forall u \in \mathbb{R}^{2s}$, so the tail $\mathbb{P}[\sup_{\nu\in\mathcal{B}_2(\mathbb{R}^{2s})}\langle\Phi_s^\top\varepsilon,\nu\rangle > \frac{\sqrt{t}}{2}]$ is bounded by $5^{2s}\exp(-\frac{(\frac{\sqrt{t}}{4})^2}{2\sigma^2})$ using the theorem found in Section 3, Lecture 5, with the term 5^{2s} corresponding to the half-net of $\mathcal{B}_2(\mathbb{R}^{2s})$. Additionally, note that in the first inequality we set $|S| = 2s$ and not $|S| < 2s$ since we padded the orthonormal basis.

Before bounding the log term, notice that without the log term this is exactly the least-squares bound that we would get if we were told which of the coordinates of θ^*, up to $2s$ dimensions, are nonzero. The price we are paying for not knowing which of those coordinates are nonzero is exactly the log of the choices that we have.

In order to bound the term $\log\binom{d}{2s}$, we claim that $\binom{d}{k} \leq (\frac{ed}{k})^k$ and prove by induction. For $k = 1$, $d \leq (ed)$ holds. Suppose true for k, so for $k+1$ we have

$$\begin{aligned}
\binom{d}{k+1} &= \frac{d!}{(k+1)!(d-k-1)!} = \binom{d}{k}\frac{(d-k)}{(k+1)} \\
&\leq \frac{e^k d^k}{k^k}\cdot\frac{(d-k)}{(k+1)} \leq \frac{e^k d^{k+1}}{k^k(k+1)} \\
&\leq \frac{e^k d^{k+1}}{(k+1)^{k+1}}\cdot \underbrace{\frac{(k+1)^{k+1}}{k^k(k+1)}}_{=(1+\frac{1}{k})^k = e^{k\log(1+\frac{1}{k})} \leq e^{\frac{k}{k}} = e} \\
&\leq (\frac{ed}{k+1})^{k+1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{P}[|\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)} - \mathbb{X}\theta^*|_2^2 > t] &\leq \exp\big(-\frac{t}{32\sigma^2} + 2s\log 5 + \log\binom{d}{2s}\big) \\ &\leq \exp\big(-\frac{t}{32\sigma^2} + 2s\log 5 + 2s\log(\frac{ed}{2s})\big).\end{aligned}$$

Setting $\delta \coloneqq \exp\big(-\frac{t}{8\sigma^2} + 2s\log 5 + 2s\log(\frac{ed}{2s})\big)$ results in

$$t \lesssim \sigma^2 s + \sigma^2 s\log(\frac{d}{s}) + \sigma^2\log(1/\delta) \lesssim \sigma^2 s\log(\frac{d}{s\delta}).$$

Notice that the first term comes from the cardinality of the ℓ_2 ball of size $2s$, and the second term comes from the number of subsets of size s. Thus we are not paying only a log factor for not knowing where the sparsity is, but instead a full additional term that is a log factor larger than the original.

Finally, with probability $1-\delta$, we have

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)}) \leq \frac{\sigma^2 s}{n}\log(\frac{d}{s\delta}).$$

2. GAUSSIAN SEQUENCE MODEL

For this section, we assume $\frac{\mathbb{X}^\top}{\sqrt{n}}\frac{\mathbb{X}}{\sqrt{n}} = I_d$ which we call orthogonal design. Let's consider the linear regression model with Gaussian noise as

$$Y = \mathbb{X}\theta^* + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_d)$. By multiplying both sides by $\mathbb{X}^\top/n$, we have

$$\frac{\mathbb{X}^\top Y}{n} = \frac{\mathbb{X}^\top\mathbb{X}}{n}\theta^* + \frac{\mathbb{X}^\top\varepsilon}{n}$$

or equivalently,

$$\tilde{Y} = \theta^* + \frac{\mathbb{X}^\top\varepsilon}{n}$$

where $\tilde{Y} = \frac{\mathbb{X}^\top Y}{n} \in \mathbb{R}^d$ is the new observations vector and $\frac{\mathbb{X}^\top\varepsilon}{n} \sim \mathcal{N}(0, \frac{\mathbb{X}^\top\mathbb{X}}{n^2}\sigma^2) = \mathcal{N}(0, \frac{\sigma^2}{n}I_d)$. Therefore, under the orthogonal design assumption, we can consider the following equivalent model:

$$Y = \theta^* + \varepsilon \in \mathbb{R}^d$$

where $\varepsilon \sim \mathcal{N}(0, \frac{\sigma^2}{n}I_d)$. Note that this is estimating the mean of a Gaussian random variable and the variance of noise is the only location n or number of observations appears. As n goes to infinity, the variance of noise goes to zero and we can observe θ^* exactly. Such a model is called Gaussian Sequence Model (GSM). The literature of GSM is quite rich, e.g. see Tsybakov (09) chapter 3 or Johnstone ('20+) which is a 467-page long draft about GSMs. In addition, this approach to linear regression is also called the direct (observation) model, in contrast to inverse problem where the goal is to estimate the inverse of an operator A in the model $Y = A\theta^* + \varepsilon$.

We also consider a slight generalization of GSM, namely SubGaussian Sequence Models. We assume

$$Y = \theta^* + \varepsilon$$

where $\varepsilon^\top u \sim \mathsf{subG}(\frac{\sigma^2}{n}|u|_2^2)$ for any $u \in \mathbb{R}^d$. Under this model,

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}) = (\hat{\theta} - \theta^*)^\top \frac{\mathbb{X}^\top \mathbb{X}}{n}(\hat{\theta} - \theta^*) = |\hat{\theta} - \theta^*|_2^2$$

which we denote by $\mathsf{MSE}(\hat{\theta})$. Under the direct model, we have $\hat{\theta}^{\mathrm{LS}} = Y$ and for any $j \in [d]$, $\hat{\theta}^{(j)}_{\mathcal{B}_0(s)}$ is equal to $Y^{(j)}$ if $Y^{(j)}$ is among the s largest elements of Y and otherwise, $\hat{\theta}^{(j)}_{\mathcal{B}_0(s)} = 0$.

3. SPARSITY ADAPTIVE THRESHOLDING ESTIMATION

In this part of lecture, we try to solve the direct model linear regression with sparse underlying model. However, we no longer assume that the sparsity s of the solution is known. The algorithm we use here is a hard thresholding algorithm. To be more specific, if an observation is smaller than the threshold, we decide that the observation is just noise and we set it to zero. Otherwise, we decide to keep the observation as probably the additive noise in the observation is not too high. Mathematically,

$$\hat{\theta}_j^{\mathrm{HRD}} = \begin{cases} Y_j & \text{if } |Y_j| > 2\tau \\ 0 & \text{if } |Y_j| \leq 2\tau \end{cases}$$

where τ is the threshold.

Theorem: If $\tau = \sigma\sqrt{\frac{2\log(2d/\delta)}{n}}$ and $\theta^* \in B_0(s)$, then with probability at least $1 - \delta$,

$$\mathsf{MSE}(\hat{\theta}^{\mathrm{HRD}}) = |\hat{\theta}^{\mathrm{HRD}} - \theta^*|_2^2 \lesssim \frac{\sigma^2 s}{n}\log(\frac{2d}{\delta}).$$

Proof. For the sake of simplicity, we use θ and $\hat{\theta}$ instead of θ^* and $\hat{\theta}^{\mathrm{HRD}}$, respectively. Let

$$\mathcal{A} = \{\max_{j\in[d]} |\varepsilon_j| \leq \tau\}.$$

From maximal inequalities, we have $\mathbb{P}(\mathcal{A}) \geq 1 - \delta$. On this event, we know

1. $|Y_j| > 2\tau \Rightarrow |\theta_j| \geq |Y_j| - |\varepsilon_j| > \tau$.
2. $|Y_j| \leq 2\tau \Rightarrow |\theta_j| \leq |Y_j| + |\varepsilon_j| \leq 3\tau$.

Therefore, one can write

$$
\begin{aligned}
|\theta-\hat{\theta}|_2^2 = \sum_{j\in[d]}(\hat{\theta}_j-\theta_j)^2 &= \sum_{\substack{j\in[d]\\ |Y_j|>2\tau}}(Y_j-\theta_j)^2 + \sum_{\substack{j\in[d]\\ |Y_j|\le 2\tau}}\theta_j^2 \\
&= \sum_{\substack{j\in[d]\\ |Y_j|>2\tau}}\varepsilon_j^2 + \sum_{\substack{j\in[d]\\ |Y_j|\le 2\tau}}\theta_j^2 \\
&\le \tau^2\sum_{j\in[d]}\mathbb{1}(|Y_j|>2\tau) + \sum_{j\in[d]}\theta_j^2\mathbb{1}(|Y_j|\le 2\tau) \\
&\le \tau^2\sum_{j\in[d]}\mathbb{1}(|\theta_j|>\tau) + \sum_{j\in[d]}\theta_j^2\mathbb{1}(|\theta_j|\le 3\tau) \\
&\le \sum_{j\in[d]}\min(\tau,|\theta_j|)^2 + \sum_{j\in[d]}(3\min(\tau,|\theta_j|))^2 \\
&= 10\sum_{j\in[d]}\min(\tau^2,|\theta_j|^2) \\
&\le 10s\tau^2 = 20\frac{s\sigma^2}{n}\log(\frac{2d}{\delta})
\end{aligned}
$$

where the last inequality results from the fact that $\theta\in\mathcal{B}_0(s)$. □

Note that τ introduced above does not depend on s which is what we require. In addition, comparing this result to the known s result, we lose $1/s$ in the logarithm here which considering $s\ll d$, does not change the final result much.

Summary: We considered sparse linear regression with $\mathsf{subG}(\sigma^2)$ noise in this lecture. First, we assumed that the underlying sparsity s is known, and define s-sparsity:

$$\mathcal{B}_0(s) := \{\theta\in\mathbb{R}^d : \sum_{j=1}^d \mathbb{1}(\theta_j\neq 0)\le s\},$$

By defining the estimator

$$\hat{\theta}_{\mathcal{B}_0(s)}\in\arg\min_{\theta\in\mathcal{B}_0(s)}|Y-\mathbb{X}\theta|_2^2,$$

we showed with probability $1-\delta$,

$$\mathsf{MSE}(\mathbb{X}\hat{\theta}_{\mathcal{B}_0(s)})\lesssim\frac{\sigma^2 s}{n}\log(\frac{d}{s}).$$

In the next part, we assumed $\frac{\mathbb{X}^\top\mathbb{X}}{n}=I_d$ and under orthogonality assumption, we showed our problem is equivalent to a direct model which can be solved by hard thresholding as

$$\hat{\theta}_j^{\text{HRD}} = \begin{cases} Y_j & \text{if } |(\mathbb{X}^\top Y/n)_j|>2\tau \\ 0 & \text{if } |(\mathbb{X}^\top Y/n)_j|\le 2\tau \end{cases}$$

with the guarantee

$$\mathsf{MSE}(\hat{\theta}^{\mathrm{HRD}}) \lesssim \frac{\sigma^2 s}{n} \log(\frac{2d}{\delta})$$

with probability $1 - \delta$ for $\tau = \sigma\sqrt{\frac{2\log(2d/\delta)}{n}}$.