MATH888: High-dimensional probability and statistics Fall 2021

Lecture 35 — Nov 29, 2021

Sebastien Roch, UW-Madison *Scribe: Susan Glenn, Sebastien Roch*

1 Overview

In the last lecture we focused on a constrained version of the least square estimator that is particularly applicable in the sparse setting.

In this lecture we continue with the proof of Theorem 2.6 in [1] which controls the MSE of $\hat{\theta}_K^{LS}$ when $K = \mathcal{B}_0(k)$

2 Sparse linear regression: Known k

We begin by describing the problem of sparse linear regression from last time.

2.1 Reminder

Recall from last time:

- Sparsity is defined as $\text{supp}(\boldsymbol{\theta}) = \{j : \boldsymbol{\theta}_j \neq 0\}$

$$\|\boldsymbol{\theta}\|_0 = |\text{supp}(\boldsymbol{\theta})|$$

$$\mathcal{B}_0(k) = \{\boldsymbol{\theta} \in \mathcal{R}^p : \|\boldsymbol{\theta}\|_0 \leq k\}$$

- $(\mathbb{X}, Y) \in \mathbb{R}^{n\times p} \times \mathbb{R}^n$

$$\hat{\boldsymbol{\theta}} := \text{ argmin}\{\|Y - \mathbb{X}\boldsymbol{\theta}\|^2 : \boldsymbol{\theta} \in \mathcal{B}_0(k)\}$$

Theorem 1 (Thm 2.6 in [1]). *Assume $2k \leq p$ and $\boldsymbol{\theta}^* \in \mathcal{B}_0(k)$ and $\|\boldsymbol{\varepsilon}\|_{\psi_2} \leq \sigma$ w.p. $1 - \delta$*

$$\frac{1}{n}\|\mathbb{X}\hat{\boldsymbol{\theta}} - \mathbb{X}\boldsymbol{\theta}^*\| \lesssim \frac{\sigma^2}{n}\left\{k \log\left(\frac{p}{k}\right) + \log\left(\frac{1}{\delta}\right)\right\}$$

2.2 Proof of Theorem 1

Recall that, for a subset s of $[p]$, the matrix $\mathbb{X}_s$ contains the columns of $\mathbb{X}$ in s. We let r_s be the rank of $\mathbb{X}_s$. Also let Φ_s be a matrix whose columns form an orthonormal basis of $\mathbb{X}_s$.

Lemma A:

$$\|\mathbb{X}\hat{\boldsymbol{\theta}} - \mathbb{X}\boldsymbol{\theta^*}\| \leq 4 \max_{s:|s|=2k} \sup_{\boldsymbol{u}\in\mathcal{B}_2^{r_s}} (\widetilde{\boldsymbol{\varepsilon}}_s^T \boldsymbol{u})^2$$

where $\widetilde{\varepsilon}_s = \Phi_s^T \boldsymbol{\varepsilon}$ with $\|\boldsymbol{\varepsilon}\|_{\psi_2} \leq \sigma$.

The vectors $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta^*}$ are sparse, the first one by construction and the second one by assumption. We proved this lemma in the last lecture but we need it for the proof of Theorem 1.

Lemma B: $P\left(\sup_{\boldsymbol{u}\in\mathcal{B}_2^{r_s}} (\widetilde{\boldsymbol{\varepsilon}}_s^T \boldsymbol{u})^2 > t\right) \leq 2 \cdot 5^{2k} \exp\left(\frac{-t}{4C\sigma^2}\right)$

Proof. Let N be a $\frac{1}{2}$-net of $\mathcal{B}_2^{r_s}$. By a previous bound (see Lecture 19)

$$|N| \leq \left(\frac{2}{\frac{1}{2}+1}\right)^{r_s} \leq 5^{2k}.$$

For any $\boldsymbol{u} \in \mathcal{B}_2^{r_s}$, there exists $\boldsymbol{z} \in N$, $\boldsymbol{x} \in \frac{1}{2}\mathcal{B}_2^{r_s}$ such that $\boldsymbol{u} = \boldsymbol{z} + \boldsymbol{x}$. By definition of an ε-net

$$\sup_{\boldsymbol{u}\in\mathcal{B}_2^{r_s}} |\widetilde{\boldsymbol{\varepsilon}}_s^T \boldsymbol{u}| \leq \max_{\boldsymbol{z}\in N} |\widetilde{\boldsymbol{\varepsilon}}_s^T \boldsymbol{z}| + \sup_{\boldsymbol{x}\in\frac{1}{2}\mathcal{B}_2^{r_s}} |\widetilde{\boldsymbol{\varepsilon}}_s^T \boldsymbol{x}|$$

$$\sup_{\boldsymbol{u}\in\mathcal{B}_2^{r_s}} |\widetilde{\boldsymbol{\varepsilon}}_s^T \boldsymbol{u}| \leq \max_{\boldsymbol{z}\in N} |\widetilde{\boldsymbol{\varepsilon}}_s^T \boldsymbol{z}| + \frac{1}{2} \sup_{\boldsymbol{u}\in\mathcal{B}_2^{r_s}} |\widetilde{\boldsymbol{\varepsilon}}_s^T \boldsymbol{u}| \text{ (by homogeneity)}$$

$$\Rightarrow \sup_{\boldsymbol{u}\in\mathcal{B}_2^{r_s}} |\widetilde{\boldsymbol{\varepsilon}}_s^T \boldsymbol{u}| \leq 2 \max_{\boldsymbol{z}\in N} |\widetilde{\boldsymbol{\varepsilon}}_s^T \boldsymbol{z}|$$

Hence,

$$P\left(\sup_{\boldsymbol{u}\in\mathcal{B}_2^{r_s}} |\widetilde{\boldsymbol{\varepsilon}}_s^T \boldsymbol{u}| > \sqrt{t}\right) \leq P\left(\max_{\boldsymbol{z}\in N} |\widetilde{\boldsymbol{\varepsilon}}_s^T \boldsymbol{z}| > \frac{\sqrt{t}}{2}\right)$$

$$\leq \sum_{\boldsymbol{z}\in N} P\left(|\widetilde{\boldsymbol{\varepsilon}}_s^T \boldsymbol{z}| > \frac{\sqrt{t}}{2}\right) \text{ (union bound)}$$

$$= 5^{2k} \cdot 2 \exp\left(\frac{-(\frac{\sqrt{t}}{2})^2}{C\sigma^2}\right) \text{ (sub gaussian)}$$

which concludes the proof. □

We are ready to prove the main theorem.

Proof of Theorem 1.

$$P\left(\|\mathbb{X}\hat{\boldsymbol{\theta}} - \mathbb{X}\boldsymbol{\theta^*}\|_2^2 \leq 4t\right) \leq P\left(\max_{|S|=2k} \sup_{\boldsymbol{u}\in\mathcal{B}_2^{r_s}} (\widetilde{\boldsymbol{\varepsilon}}_s^T \boldsymbol{u})^2\right) \text{ (by lemma A)}$$

$$\leq \sum_{\max_{|S|=2k}} P\left(\sup_{\boldsymbol{u}\in\mathcal{B}_2^{rs}} (\widetilde{\boldsymbol{\varepsilon}}_s^T \boldsymbol{u})^2\right) \text{ since finite and using the union bound}$$

$$\leq \delta \text{ (by lemma B)}$$

for the choice $t = \frac{1}{4}C\sigma^2\{\log(\binom{p}{2k} \times 2 \times 5^k) + \log(\frac{1}{\delta})\}$.

By Stirling's approximation, we get that $P\left(\|\mathbb{X}\hat{\boldsymbol{\theta}} - \mathbb{X}\boldsymbol{\theta}^*\|_2^2 \leq C\sigma^2\{k\log(\frac{p}{k}) + \log(\frac{1}{\delta})\}\right) \leq \delta$. □

2.3 Another Special Case: Sub-Gaussian Sequence Model

We assume here all columns of $\mathbb{X}$ are orthogonal (which can only be satisfied when $p \leq n$). Formally:

- *Assumption ORT*: $\frac{\mathbb{X}^T\mathbb{X}}{n} = \boldsymbol{I}_p$
- from this consider $y := \frac{1}{n}\mathbb{X}^T Y = \frac{1}{n}\mathbb{X}^T[\mathbb{X}\boldsymbol{\theta}^* + \boldsymbol{\varepsilon}] = \boldsymbol{\theta}^* + \boldsymbol{\xi}$ where $\xi = \frac{1}{n}\mathbb{X}^T\boldsymbol{\varepsilon} \in \mathbb{R}^p$
- if $\|\varepsilon\|_{\psi_2}\| \leq \sigma$ the $\|\xi\|_{\psi_2} \leq \frac{\sigma}{\sqrt{n}}$
- *Normal Equations*: $\frac{1}{n}\mathbb{X}^T\mathbb{X}\hat{\boldsymbol{\theta}}^{LS} = \frac{1}{n}\mathbb{X}^T Y$ then $\hat{\boldsymbol{\theta}}^{LS} = y$
- $\text{MSE}(\mathbb{X}\hat{\boldsymbol{\theta}}) = \frac{1}{n}\|\mathbb{X}\hat{\boldsymbol{\theta}} - \mathbb{X}\boldsymbol{\theta}^*\|_2^2 = \frac{1}{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\mathbb{X}^T\mathbb{X}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\|_2^2$

Next time: assume that $\boldsymbol{\theta}^*$ is sparse. How can we improve over the least squares estimate (whose MSE we showed to be $\approx p\frac{\sigma^2}{n}$)? When there are lots of 0s in $\boldsymbol{\theta}^*$, we can hope to improve the MSE by taking small y values and setting them equal to 0. We will give the details next time.

References

[1] Philippe Rigollet and Jan-Christian Hütter 18.657: *High Dimensional Statistics Lecture Notes*, MIT, 2017