MATH888: High-dimensional probability and statistics Fall 2021

Lecture 34 — Nov 24, 2021

Sebastien Roch, UW-Madison *Scribe: Yifei Ming, Sebastien Roch*

1 Overview

In Lecture 15, we discussed the properties of the (unconstrained) least square estimator for linear regression with sub-Gaussian noises. In this lecture, we turn our focus to a constrained version of the least square estimator that is particularly applicable in the sparse setting.

2 Main Section

We begin by formally introducing the setting of sparse linear regression, and then we consider two examples under different assumptions.

2.1 Sparse Linear Regression

Review of Linear Regression Recall that in the linear regression problem, we have n data points: $(\mathbb{X}, y)$, where $\mathbb{X} \in \mathbb{R}^{n\times p}$, and $y \in \mathbb{R}^n$. We assume $y = \mathbb{X}\boldsymbol{\theta}^* + \epsilon$, where $\boldsymbol{\theta}^* \in \mathbb{R}^p$ is the unknown parameter. The noise ϵ is assumed to be sub-Gaussian with $\|\epsilon\|_{\psi^2} \leq \sigma$. The mean squared error (MSE) of an estimator $\mathbb{X}\hat{\theta}$ is defined as $MSE(\mathbb{X}\hat{\theta}) = \frac{1}{n}\|\mathbb{X}\theta^* - \mathbb{X}\hat{\theta}\|_2^2$. In particular, the least square estimator is defined as:

$$\hat{\boldsymbol{\theta}}^{LS} \in \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \|y - \mathbb{X}\boldsymbol{\theta}\|_2^2$$

We proved in Lecture 15 that with probability at least $1-\delta$:

$$MSE(\mathbb{X}\hat{\boldsymbol{\theta}}^{LS}) \lesssim \frac{\sigma^2}{n}(rank(\mathbb{X}) + \log(1/\delta))$$

Sparsity Recall that the ℓ_0 norm is defined as the number of non-zero entries of a vector:

$$\|\boldsymbol{\theta}\|_0 = \sum_{j=1}^{p} \mathbb{1}\{\boldsymbol{\theta}_j \neq 0\}$$

Intuitively, sparsity corresponds to a "small" ℓ_0 norm. Specifically, $\boldsymbol{\theta}$ is a k-sparse vector if $\|\boldsymbol{\theta}\|_0 \leq k$. The support of $\boldsymbol{\theta}$ is defined as:

$$\mathrm{supp}(\boldsymbol{\theta}) = \{j : \boldsymbol{\theta}_j \neq 0\}$$

Therefore we have $\|\boldsymbol{\theta}\|_0 = |\,\mathrm{supp}(\boldsymbol{\theta})|$. Moreover, the ℓ_0 ball $\mathcal{B}_0(k)$ of all k-sparse vectors is denoted as:

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

where k is the sparsity level.

2.2 Benchmarks: Special cases

Case 1 After introducing the notations, now let's consider a simple case where we assume the non-zero entries of $\boldsymbol{\theta}$ are given. We denote $S := \text{supp}(\boldsymbol{\theta}^*)$ and $\Delta = |S|$. Let $\mathbb{X}_S$ denote the submatrix of $\mathbb{X}$ with columns $\mathbb{X}_j$ for $\forall j \in S$. The corresponding least square estimator is given by:

$$\hat{\boldsymbol{\theta}}_S^{LS} \in \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^\Delta} \|y - \mathbb{X}_S \boldsymbol{\theta}\|_2^2$$

Thus the full solution $\hat{\boldsymbol{\theta}}$ is:

$$\hat{\boldsymbol{\theta}}_i = \begin{cases} \hat{\boldsymbol{\theta}}_{S,i}^{LS} & i \in S \\ 0 & i \notin S \end{cases}$$

Then, by our previous results for linear regression, we have the following bound for MSE (w.p. at least $1 - \delta$):

$$MSE(\mathbb{X}\hat{\boldsymbol{\theta}}) \lesssim \frac{\sigma^2}{n}(\|\theta^*\|_0 + \log(1/\delta))$$

Case 2 Suppose $\text{supp}(\boldsymbol{\theta}^*)$ is unknown but we know the number of non-zero entries $k = \|\boldsymbol{\theta}^*\|_0$. A natural least square estimator for a fixed k is:

$$\hat{\boldsymbol{\theta}}_{\mathcal{B}_0(k)}^{LS} \in \arg\min\{\|y - \mathbb{X}\boldsymbol{\theta}\|_2^2 : \boldsymbol{\theta} \in \mathcal{B}_0(k)\}$$

Then we take the best estimator among $\hat{\boldsymbol{\theta}}_S^{LS}$ for each subset S with $|S| = k$. Note that this brute force approach is computationally expensive, as we need to compute $\begin{pmatrix} p \\ k \end{pmatrix}$ estimators for each k—we will come back later in the course to a computationally efficient approach. But for now, despite the computational difficulty, we analyze the statistical properties of this estimator.

Theorem 1. *(Thm 2.6 in [1]) Fix a positive integer $k \leq p/2$. Let $K = \mathcal{B}_0(k)$ be set of k-sparse vectors of $\mathbb{R}^p$ and assume that $\boldsymbol{\theta}^* \in \mathcal{B}_0(k)$. Then, for any $\delta > 0$, with probability $1 - \delta$, it holds*

$$\text{MSE}\left(\mathbb{X}\hat{\theta}_{\mathcal{B}_0(k)}^{LS}\right) \lesssim \frac{\sigma^2}{n} \log \begin{pmatrix} p \\ 2k \end{pmatrix} + \frac{\sigma^2 k}{n} + \frac{\sigma^2}{n} \log(1/\delta)$$

Before proving the theorem, we prove a lemma. First, some notation. For any subset $S \in \{1, \ldots, p\}$, denote $r_S = \text{rank}(\mathbb{X}_S) \leq |S|$. Further, let $\Phi_S = [\phi_1, \ldots, \phi_{r_S}] \in \mathbb{R}^{n \times r_S}$ be the collection of orthonormal basis of the column space of $\mathbb{X}_S$.

Lemma 2. *Let $\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_{\mathcal{B}_0(k)}^{\text{LS}}$. We have*

$$\|\mathbb{X}\tilde{\boldsymbol{\theta}} - \mathbb{X}\boldsymbol{\theta}^*\|_2^2 \leq 4 \max_{|S|=2k} \sup_{u \in \mathcal{B}_2^{r_S}} (\tilde{\epsilon}_S^\top u)^2$$

where $\tilde{\epsilon}_S = \epsilon^\top \Phi_S \sim \text{subG}_{r_S}(\sigma^2)$.

Proof. By definition,

$$\|y - \mathbb{X}\tilde{\boldsymbol{\theta}}\|_2^2 \leq \|y - \mathbb{X}\boldsymbol{\theta}^*\|_2^2 = \|\epsilon\|_2^2.$$

Moreover,

$$\|y - \mathbb{X}\tilde{\boldsymbol{\theta}}\|_2^2 = \|\mathbb{X}\boldsymbol{\theta}^* + \epsilon - \mathbb{X}\tilde{\boldsymbol{\theta}}\|_2^2 = \|\mathbb{X}\tilde{\boldsymbol{\theta}} - \mathbb{X}\boldsymbol{\theta}^*\|_2^2 - 2\epsilon^\top \mathbb{X}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) + \|\epsilon\|_2^2.$$

Rearranging the terms, we have

$$\|\mathbb{X}\tilde{\boldsymbol{\theta}} - \mathbb{X}\boldsymbol{\theta}^*\|_2^2 \leq 2\epsilon^\top \mathbb{X}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) = 2\|\mathbb{X}\tilde{\boldsymbol{\theta}} - \mathbb{X}\boldsymbol{\theta}^*\|_2 \frac{\epsilon^\top \mathbb{X}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)}{\|\mathbb{X}\tilde{\boldsymbol{\theta}} - \mathbb{X}\boldsymbol{\theta}^*\|_2}.$$

Next we aim to bound

$$\frac{\epsilon^\top \mathbb{X}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)}{\|\mathbb{X}\tilde{\boldsymbol{\theta}} - \mathbb{X}\boldsymbol{\theta}^*\|_2}.$$

Let $\hat{S} = \operatorname{supp}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$. As both $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ have at most k non-zero entries, we have $|\hat{S}| \leq 2k$. Then we know that there exists some vector $\nu \in \mathbb{R}^{r_{\hat{S}}}$ such that

$$\mathbb{X}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right) = \Phi_{\hat{S}}\nu.$$

Let $\tilde{\epsilon} = \epsilon^\top \Phi_{\hat{S}}$. Then we have

$$r = \frac{\epsilon^\top \mathbb{X}\left(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right)}{\|\mathbb{X}\tilde{\boldsymbol{\theta}} - \mathbb{X}\boldsymbol{\theta}^*\|_2} = \frac{\epsilon^\top \Phi_{\hat{S}}\nu}{\|\nu\|_2} = \tilde{\epsilon}^\top \frac{\nu}{\|\nu\|_2} \leq \max_{|S|=2k} \sup_{u \in \mathcal{B}_2^{r_S}} \left[\epsilon^\top \Phi_S\right] u,$$

where $\mathcal{B}_2^{r_S}$ is the unit ball of $\mathbb{R}^{r_S}$. Here we need to "sup out" the support of $\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$ and the vector ν as they depend in a subtle way on ϵ (and therefore would prevent us form applying the sub-Gaussianity of ϵ). Therefore, we have:

$$\|\mathbb{X}\tilde{\boldsymbol{\theta}} - \mathbb{X}\boldsymbol{\theta}^*\|_2^2 \leq 4 \max_{|S|=2k} \sup_{u \in \mathcal{B}_2^{r_S}} (\tilde{\epsilon}_S^\top u)^2,$$

where $\tilde{\epsilon}_S = \epsilon^\top \Phi_S \sim \operatorname{subG}_{r_S}(\sigma^2)$, as claimed. This last claim on the sub-Gaussian norm of $\epsilon^\top \Phi_S$ follows from the fact that the columns of Φ_S are orthonormal by construction and from the defintion of the sub-Gaussian vectors. $\square$

We will conclude the proof of the theorem next time.

References

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