MATH888: High-dimensional probability and statistics Fall 2021

Lecture 15 — Oct 11, 2021

Sebastien Roch, UW-Madison *Scribe: Angeliki Giannou*

1 Overview

In the last lecture we derived bounds for mean estimation of sub-Gaussian vectors.

In this lecture we will focus on the linear regression problem with noisy observations. Our goal would be to estimate the true parameter of the problem and derive error bounds of the loss. In order to do so, we will leverage tools from the previous lecture.

2 Main Section

2.1 Linear regression

The problem has the following components:

- Let $\boldsymbol{\theta}^* \in \mathbb{R}^p$ be an unknown vector.
- Let $f(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\theta}^*$, $\mathbf{x} \in \mathbb{R}^p$.
- We make n noisy observations $(\mathbf{x}_i, y_i)$, $y_i = f(\mathbf{x}_i) + \epsilon_i$. We consider $\epsilon_i \in \mathbb{R}$, $i = 1, \dots, n$ to be a random variable with mean zero.

Our *goal* is to find $\hat{\boldsymbol{\theta}} \in \mathbb{R}^p$ that minimizes the MSE loss, i.e.,

$$MSE(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}_n(\mathbf{x}_i) - f_n(\mathbf{x}_i)\right)^2 \tag{1}$$

where $\hat{f}_n(\mathbf{x}) = \mathbf{x}^T\hat{\boldsymbol{\theta}}$. This problem cannot be solved directly, since we don't have access to the values $f(\mathbf{x}_i)$.

2.1.1 Least square solution

Instead we will solve:

$$\min_{\hat{\boldsymbol{\theta}} \in \mathbb{R}^p} \|\mathbf{y} - \mathbb{X}\hat{\boldsymbol{\theta}}\|_2^2 \tag{*}$$

where $\mathbf{y} = (y_1, \dots, y_n)^T \in \mathbb{R}^n$, $\mathbb{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T \in \mathbb{R}^{n\times p}$.

Theorem 1. *Let $\hat{\boldsymbol{\theta}}^{LS}$ be a solution to (*). Then*

$$\mathbb{X}^T\mathbb{X}\hat{\boldsymbol{\theta}}^{LS} = \mathbb{X}^T\mathbf{y} \tag{**}$$

Furthermore, we can always choose $\hat{\boldsymbol{\theta}}^{LS} = \mathbb{X}^+\mathbf{y}$.

Recall. If $A \in \mathbb{R}^{n\times p}$ with Singular Value Decomposition (SVD): $A = USV^T$, where $U \in \mathbb{R}^{n\times n}$, $V \in \mathbb{R}^{p\times p}$ are orthogonal matrices and $S \in \mathbb{R}^{n\times p}$ is a diagonal matrix with all of its elements non-negative; then

$$A^+ = VS^+U^T$$

where

$$s^+_{ii} = \begin{cases} \dfrac{1}{s_{ii}} & , \text{ if } s_i > 0 \\ 0 & , \text{ otherwise} \end{cases}$$

We now proceed to the proof of theorem 1.

Proof. Notice that (*) is convex and its solution satisfies

$$\nabla_{\boldsymbol{\theta}} \|\mathbf{y} - \mathbb{X}\boldsymbol{\theta}\|_2^2 = 0$$

By doing the calculations we can derive (**).
Considering now the second part of the theorem, if the solution of the problem is not unique, one choice is to minimize also the norm of the parameter. Thus, we will solve:

$$\min_{\boldsymbol{\theta}} \|\boldsymbol{\theta}\|_2^2 \quad \text{s.t.} \quad \mathbb{X}^T\mathbb{X}\boldsymbol{\theta} = \mathbb{X}^T\mathbf{y}$$

Using SVD we have:

$$\mathbb{X}^T\mathbb{X} = VS^TU^TUSV^T = VS^2V^T$$

So, (**) becomes:

$$VS^2V^T\boldsymbol{\theta} = VSU^T\mathbf{y} \Rightarrow S^2V^T\boldsymbol{\theta} = SU^T\mathbf{y}$$

Let $\mathbf{z} = V^T\boldsymbol{\theta}$ and $\mathbf{w} = U^T\mathbf{y}$ and (**) becomes:

$$s^2_{ii}z_i = s_{ii}w_i \text{ for all } i = 1, \dots, n$$

Notice that since V is orthogonal, $\|\mathbf{z}\|_2^2 = \|\boldsymbol{\theta}\|_2^2$. So, we now want to minimize the norm of $\mathbf{z}$ or equivalently the expression

$$z_1^2 + \dots z_n^2 \quad \text{subject to} \quad s^2_{ii}z_i = s_{ii}w_i$$

In the case that $s_{ii} = 0$ we will choose $z_i = 0$, since we want to minimize the above expression. In any other case we will choose $z_i = \dfrac{w_i}{s_{ii}}$. We finally get that $\mathbf{z} = S^+\mathbf{w}$ or

$$\boldsymbol{\theta} = VS^+U^T\mathbf{y} = \mathbb{X}^+\mathbf{y}$$

□

Remark 2. *Of course if $\mathbb{X}$ is invertible, $\boldsymbol{\theta} = \mathbb{X}^{-1}\mathbf{y}$.*

Remark 3. *We can also think of the minimization of the norm of $\boldsymbol{\theta}$ as the projection to the column space of the matrix $\mathbb{X}$.*

2.2 The MSE loss

We now want to focus on the MSE loss meaning $\frac{1}{n}\|\mathbb{X}\hat{\boldsymbol{\theta}}^{LS} - \mathbb{X}\boldsymbol{\theta}^*\|_2^2$ and see whether we can bound it.

Theorem 4. *Suppose that $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)$ is sub-Gaussian with $\|\boldsymbol{\epsilon}\|_{\psi_2} \leq K$. Then, with probability at least $1-\delta$*

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

Remark 5. *If $p << n$ then we get a small bound. However, if $p >> n$ then $rank(\mathbb{X}) \simeq n$ and we just get a big constant as a bound.*

Proof. We have

$$\begin{aligned}
\mathbb{X}\hat{\boldsymbol{\theta}}^{LS} - \mathbb{X}\boldsymbol{\theta}^* &= \mathbb{X}\mathbb{X}^+\mathbf{y} - \mathbb{X}\boldsymbol{\theta}^* \\
&= \mathbb{X}\mathbb{X}^+(\mathbb{X}\boldsymbol{\theta}^* + \boldsymbol{\epsilon}) - \mathbb{X}\boldsymbol{\theta}^* \\
&= \mathbb{X}\mathbb{X}^+\mathbb{X}\boldsymbol{\theta}^* + \mathbb{X}\mathbb{X}^+\boldsymbol{\epsilon} - \mathbb{X}\boldsymbol{\theta}^* \\
&= USV^TVS^+U^TUSV^T\boldsymbol{\theta}^* + \mathbb{X}\mathbb{X}^+\boldsymbol{\epsilon} - \mathbb{X}\boldsymbol{\theta}^* \\
&= USV^T\boldsymbol{\theta}^* + \mathbb{X}\mathbb{X}^+\boldsymbol{\epsilon} - \mathbb{X}\boldsymbol{\theta}^* \\
&= \mathbb{X}\boldsymbol{\theta}^* + \mathbb{X}\mathbb{X}^+\boldsymbol{\epsilon} - \mathbb{X}\boldsymbol{\theta}^* \\
&= \mathbb{X}\mathbb{X}^+\boldsymbol{\epsilon}
\end{aligned}$$

So,

$$\frac{1}{n}\|\mathbb{X}\hat{\boldsymbol{\theta}}^{LS} - \mathbb{X}\boldsymbol{\theta}^*\|_2^2 = \frac{1}{n}\|X\mathbb{X}^+\boldsymbol{\epsilon}\|_2^2$$

In the previous lecture we saw that if B is a matrix and $\|X\|_{\psi_2} \leq K$, then for all $t \geq 0$

$$\mathbb{P}\left(\|BX\|_2^2 \geq CK^2\|B\|_F^2 + t\right) \leq \exp(-ct/K^2\|B\|_2^2)$$

Using the above inequality with $B = \frac{\mathbb{X}\mathbb{X}^+}{\sqrt{n}}$ and $X = \boldsymbol{\epsilon}$ we get

$$\mathbb{P}\left(\frac{1}{n}\|\mathbb{X}\mathbb{X}^T\boldsymbol{\epsilon}\|_2^2 \geq \frac{CK^2}{n}\|\mathbb{X}\mathbb{X}^+\|_F^2 + t\right) \leq \exp\left(\frac{-ctn}{K^2\|\mathbb{X}\mathbb{X}^+\|_2^2}\right)$$

By choosing $t = \frac{K^2\|\mathbb{X}\mathbb{X}^+\|_2^2}{cn}\log(1/\delta)$ we get that with probability at least $1-\delta$

$$\begin{aligned}
MSE(\mathbb{X}\hat{\boldsymbol{\theta}}^{LS}) &\leq \frac{CK^2}{n}rank\mathbb{X} + \frac{K^2}{cn} \\
&\lesssim \frac{K^2}{n}(rank(\mathbb{X}) + \log(1/\delta))
\end{aligned}$$

Notice that $\|\mathbb{X}\mathbb{X}^+\|_2^2$ is the maximum singular value of $\mathbb{X}\mathbb{X}^+$ which is equal to one and $\|\mathbb{X}\mathbb{X}^+\|_F^2$ is equal to the rank of $\mathbb{X}$. □