**MATH888: High-dimensional probability and statistics** Fall 2021

# Lecture 15 — October 11, 2021

*Sebastien Roch, UW-Madison* *Scribe: Daniel Szabo*

## 1 Overview

In the last lecture we started to apply our tail bounds for the norm of sub-gaussian vectors by using the covariance of a random vector to bound how fast mean estimation of i.i.d sub-gaussians can converge.

In this lecture we begin by reviewing the basics of linear regression, and then discuss some applications of these bounds in the context of regression.

## 2 Linear Regression

The setting for linear regression is as follows.

- Let $\Theta^* \in \mathbb{R}^p$ be an unknown random vector that we are trying to recover.
- Let $f(X) = X^T\Theta^*$ for $X \in \mathbb{R}^p$.
- Make $n$ noisy observations $(X_i, y_i)$ where $y_i = f(X_i) + \varepsilon_i$ for $\varepsilon_i$ a mean zero real valued random variable for $i = 1, 2, \ldots, n$.

Our goal is then to use these observations to recover $\Theta^*$. For us, this will mean finding a $\hat{\Theta} \in \mathbb{R}^p$ that minimizes the mean squared error of $\hat{f}_n(X) = X^T\hat{\Theta}$ where

$$MSE(\hat{f}_n) = \frac{1}{n}\sum_{i=1}^{n}(\hat{f}_n(X_i) - f(X_i))^2. \tag{1}$$

There are other potential goals, such as restricting to sparse $\hat{\Theta}$ by looking at the 0 norm or the 1 norm, or even minimizing the expected MSE over a random input $X$ that we won't consider. There is an exact characterization of this $\hat{\Theta}$.

### 2.1 Least Squares Solution

Let $\mathbb{X} \in \mathbb{R}^{n\times p}$ be the matrix with $X_i$ in each row, and $Y = [y_1, y_2, \ldots, y_n]$. Then minimizing equation (1) is equivalent to finding

$$\min_{\hat{\Theta}\in\mathbb{R}} \|Y - \mathbb{X}\hat{\Theta}\|_2^2. \tag{2}$$

The following theorem characterizes the solutions to this equation.

**Theorem 1.** *Let $\hat{\Theta}^{LS}$ be a solution to (2). Then it must satisfy the normal equations, namely*

$$\mathbb{X}^T\mathbb{X}\hat{\Theta}^{LS} = \mathbb{X}^T Y. \tag{3}$$

*We can also choose a unique $\hat{\Theta}^{LS} = \mathbb{X}^\dagger Y$ that minimizes $\|\hat{\Theta}^{LS}\|_2$.*

Recall the pseudoinverse of a matrix $A \in \mathbb{R}^{n\times p}$ with singular value decomposition $A = USV^T$ is $A^\dagger = V^T S^\dagger U$ where $S^\dagger$ is the diagonal matrix made up of $s_{ii} = \begin{cases} 1/s_{ii} & \text{if } s_{ii} > 0 \\ 0 & \text{else} \end{cases}$ for $i = 1, \dots, n$. Here $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{p\times p}$ are orthogonal matrices.

*Proof.* Because of the convexity of $\|\cdot\|_2^2$, (2) is just minimizing a convex function composed with a linear one and therefore a convex problem itself. Thus the minimum satisfies $\nabla_\Theta \|Y - \mathbb{X}\Theta\|_2^2 = 0$. Simple calculations show

$$\nabla_\Theta \|Y - \mathbb{X}\Theta\|_2^2 = \nabla_\Theta(\|Y\|_2^2 - 2\mathbb{X}^T\Theta^T Y + \Theta^T\mathbb{X}^T\mathbb{X}\Theta) = 2\mathbb{X}^T\mathbb{X}\Theta - 2\mathbb{X}^T Y$$

and thus $\nabla_{\hat{\Theta}^{LS}} \|Y - \mathbb{X}\hat{\Theta}^{LS}\|_2^2 = 0 \iff \mathbb{X}^T\mathbb{X}\hat{\Theta}^{LS} = \mathbb{X}^T Y$, which is exactly the normal equations (3).

If the solution to this is not unique, we want to find $\min \|\Theta\|^2$ subject to (3). To do so we can use the SVD of $\mathbb{X} = USV^T$ to see

$$\begin{aligned} &\mathbb{X}^T\mathbb{X}\Theta = \mathbb{X}^T Y \\ \iff &VSU^T USV^T\Theta = VSU^T Y \\ \iff &V^T VS^2V^T\Theta = SU^T Y \\ \iff &S^2V^T\Theta = SU^T Y. \end{aligned}$$

Substituting $Z = V^T\Theta, W = U^T Y$ we can see (3) is equivalent to $S^2 Z = SW$ with $\|Z\|_2^2 = \|\Theta\|_2^2$ because $V$ is orthogonal. For any $i = 1, \dots, n$, if $s_{ii}$ is 0, we have no constraint on the corresponding $z_i$ which means we would choose $z_i = 0$ to minimize $\|\Theta\|_2^2$. Otherwise we need $z_i = \frac{1}{s_{ii}} w_i$, which is exactly the formula for $S^\dagger$. Thus

$$Z = S^\dagger W \implies \Theta = VS^\dagger U^T Y = \mathbb{X}^\dagger Y.$$

□

## 3 Assessing Least Squares on Sub-Gaussian Errors

We can now start applying the bounds from previous lectures to analyze how accurate the MSE of $\hat{\Theta}$ will be. Recall for our setting equation (1) is equivalent to

$$MSE(\mathbb{X}\hat{\Theta}^{LS}) = \frac{1}{n}\|\mathbb{X}\hat{\Theta}^{LS} - \mathbb{X}\Theta^*\|_2^2. \tag{4}$$

**Theorem 2.** *Suppose $\varepsilon = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n]^T$ is sub-gaussian with $\|\varepsilon\|_{\psi_2} \le K$. Then*

$$MSE(\mathbb{X}\hat{\Theta}^{LS}) \lesssim \frac{K^2}{n}\left(rank(\mathbb{X}) + \log(1/\delta)\right) \quad w.p.\ 1-\delta$$

Note that in the high dimensional case when $p \gg n$, $\text{rank}(\mathbb{X}) \approx n$, so this bound is not very meaningful; all it says is the error is bounded by a large constant that does not converge to 0.

*Proof.* Plugging the result of Theorem (1) into equation (4),

$$\begin{aligned}
\mathbb{X}\hat{\Theta}^{LS} - \mathbb{X}\Theta^* &= \mathbb{X}\mathbb{X}^\dagger Y - \mathbb{X}\Theta^* \\
&= \mathbb{X}\mathbb{X}^\dagger(\mathbb{X}\Theta^* + \varepsilon) - \mathbb{X}\Theta^* \\
&= (USV^T VS^\dagger U^T USV^T - \mathbb{X})\Theta^* + \mathbb{X}\mathbb{X}^\dagger\varepsilon \\
&= (USV^T - \mathbb{X})\Theta^* + \mathbb{X}\mathbb{X}^\dagger\varepsilon \\
&= \mathbb{X}\mathbb{X}^\dagger\varepsilon.
\end{aligned}$$

In other words, minimizing (4) is equivalent to minimizing $\|\mathbb{X}\mathbb{X}^\dagger\varepsilon\|_2^2$, which is exactly the form we saw in the previous two lectures. There we showed

$$\mathbb{P}\left(\frac{1}{n}\|\mathbb{X}\mathbb{X}^\dagger\varepsilon\|_2^2 \geq \frac{CK^2}{n}\|\mathbb{X}\mathbb{X}^\dagger\|_F^2 + \frac{t}{n}\right) \leq \exp\left(-\frac{ct}{K^2\|\mathbb{X}\mathbb{X}^\dagger\|_2^2}\right)$$

which is equal to $\delta$ when $t = \frac{1}{c}K^2\|\mathbb{X}\mathbb{X}^\dagger\|_2^2\log(1/\delta)$. Recall that $\|\mathbb{X}\mathbb{X}^\dagger\|_F^2$ is just the sum of the squared singular values, which for $\mathbb{X}\mathbb{X}^\dagger = USS^\dagger U^T$ just counts each nonzero element of $S$ which is exactly the rank of $\mathbb{X}$. Meanwhile $\|\mathbb{X}\mathbb{X}^\dagger\|_2^2 = 1$ is just the maximum singular value. Plugging in we have that

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\|\mathbb{X}\hat{\Theta}^{LS} - \mathbb{X}\Theta^*\|_2^2 \geq \frac{CK^2}{n}\text{rank}(\mathbb{X}) + \frac{\frac{1}{c}K^2\log(1/\delta)}{n}\right) \leq \delta \\
\implies &\mathbb{P}\left(MSE(\mathbb{X}\hat{\Theta}^{LS}) < \frac{CK^2}{n}\left(\text{rank}(\mathbb{X}) + \log(1/\delta)\right)\right) > 1 - \delta.
\end{aligned}$$

□