Lecture 15

Linear Regression with Sub-Gaussian Noise

In many real-world scenarios, we aim to model the relationship between a set of input variables (features) and a continuous output variable. An important case is when this relationship is, or can be well-approximated by, a linear function. For instance, we might want to predict the sales price of a house (y) based on several features ($\mathbf{x}$), such as its square footage, the number of bedrooms, and its age under the assumption that these features contribute (almost) linearly to the final price.

Problem Definition

More formally, we assume there exists an underlying, unknown parameter vector $\boldsymbol{\beta}^\star \in \mathbb{R}^d$. This vector encapsulates the weights associated with each of the d features in $\mathbf{x}$ to determine the output y through a linear combination: $y = \langle \mathbf{x}, \boldsymbol{\beta}^\star \rangle := \mathbf{x}^\top \boldsymbol{\beta}^\star$. The task of *linear regression* refers to finding this unknown $\boldsymbol{\beta}^\star$ using a collection of past observations $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the feature vector for the i-th observation, and $y_i \in \mathbb{R}$ is the corresponding observed output. In this lecture note, we consider a "fixed design" of linear regression. Namely, $\mathbf{x}_i$'s are fixed[1].

If our observations were perfect and followed the exact linear relationship, determining $\boldsymbol{\beta}^\star$ would essentially involve solving a system of linear equations. In most non-degenerate cases, $d+1$ such perfect observations would suffice to uniquely determine $\boldsymbol{\beta}^\star$.

However, in practical settings, observed data is rarely exact. Various factors, such as measurement errors or unmodeled complexities, introduce noise into our observations. We model this by assuming that each observed output y_i is a noisy version of the true linear relationship:

$$y_i = \langle \mathbf{x}_i, \boldsymbol{\beta}^\star \rangle + \varepsilon_i = \mathbf{x}_i^\top \boldsymbol{\beta}^\star + \varepsilon_i; \qquad \forall\, i \in [n]. \tag{1}$$

Here, $\varepsilon_i \in \mathbb{R}$ represents the unknown additive noise component for the i-th observation. We make the following assumptions about this noise:

1. **Zero-mean:** $\mathbf{E}[\varepsilon_i] = 0$ for all $i \in [n]$. This implies that the noise does not introduce any systematic bias in our observations.

---

[1]The contrary is when $\mathbf{x}_i$'s are sampled from a certain distribution.

2. **Sub-Gaussianity:** $\varepsilon_i \in \mathsf{subG}\left(\sigma^2\right)$. This is a general assumption indicating that the tails of the noise distribution decay fast. The sub-Gaussianity assumption allows for a broader range of noise distributions beyond just the normal distribution.

Matrix representation: The above setup can be written in matrix form. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$, and $\mathbf{y}, \boldsymbol{\varepsilon} \in \mathbb{R}^n$ be given as below

$$
\mathbf{X} = \begin{bmatrix} \rule[0.5ex]{1em}{0.4pt} & \mathbf{x}_1^\top & \rule[0.5ex]{1em}{0.4pt} \\ & \vdots & \\ \rule[0.5ex]{1em}{0.4pt} & \mathbf{x}_n^\top & \rule[0.5ex]{1em}{0.4pt} \end{bmatrix} ; \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} ; \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}
$$

Then we can represent Equation (1) as

$$
\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon} \tag{2}
$$

We emphasize that both $\boldsymbol{\beta}^\star$ and $\boldsymbol{\varepsilon}$ are unknown, while $\mathbf{X}$ and $\mathbf{y}$ are observed.

Evaluation metric for the solution: Given a solution $\hat{\boldsymbol{\beta}}$ of the linear regression problem, there are various ways to evaluate the quality of $\hat{\boldsymbol{\beta}}$. For instance, one would want to find $\hat{\boldsymbol{\beta}}$ with a small error measured by $\left\|\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}\right\|_p$ for some $p \in [1, \infty)$. Another measure, which we use in this lecture note, is to focus on how well $\hat{\boldsymbol{\beta}}$ would have predicted the observed $\mathbf{y}$. More formally, we focus on the error measured by $(1/n) \cdot \left\|\mathbf{X}\boldsymbol{\beta}^\star - \mathbf{X}\hat{\boldsymbol{\beta}}\right\|_2^2$, a normalized version of the error.

Finding a Solution

Ideally, we are looking for $\hat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} \left\|\mathbf{X}\boldsymbol{\beta}^\star - \mathbf{X}\hat{\boldsymbol{\beta}}\right\|_2^2$. However, since we do not know $\hat{\boldsymbol{\beta}}^\star$, we choose to adopt the approximation $\mathbf{y} \approx \mathbf{X}\boldsymbol{\beta}^\star$. In this way, we turn to solve the following minimization problem

$$
\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^d} f\left(\boldsymbol{\beta}\right) ; \quad f\left(\boldsymbol{\beta}\right) := \left\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\right\|_2^2
$$

Since the objective $f\left(\boldsymbol{\beta}\right)$ is convex, we can directly solve for $\hat{\boldsymbol{\beta}}$ from the equation $\nabla f\left(\hat{\boldsymbol{\beta}}\right) = \mathbf{0}$. Writing out the form of $\nabla f\left(\boldsymbol{\beta}\right)$ gives

$$
\nabla f\left(\hat{\boldsymbol{\beta}}\right) = 2\mathbf{X}^\top \mathbf{X}\hat{\boldsymbol{\beta}} - 2\mathbf{X}^\top \mathbf{y} = 0 .
$$

This equation gives the solution

$$
\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X}\right)^\dagger \mathbf{X}^\top \mathbf{y} , \tag{3}
$$

where $\left(\mathbf{X}^\top\mathbf{X}\right)^\dagger$ indicates the pseudoinverse of $\mathbf{X}^\top\mathbf{X}$.

Analyzing the Quality of $\hat{\boldsymbol{\beta}}$

In this section, we analyze the quality of $\hat{\boldsymbol{\beta}}$ by studying $\left\|\mathbf{X}\boldsymbol{\beta}^\star - \mathbf{X}\hat{\boldsymbol{\beta}}\right\|_2^2$ under the assumption that $\varepsilon_i \in \mathsf{subG}\left(\sigma^2\right)$. We start with a generic analysis that does not depend on the structure of $\mathbf{X}$.

Proposition 1. *Let* $\mathbf{X}, \mathbf{y}, \boldsymbol{\varepsilon}$*, and* $\hat{\boldsymbol{\beta}}$ *be defined in the linear regression set-up in Equation* (2) *and Equation* (3)*. Assume that* $\varepsilon_i \in \mathsf{subG}\left(\sigma^2\right)$*. Then we have that*

$$\left\|\mathbf{X}\boldsymbol{\beta}^\star - \mathbf{X}\hat{\boldsymbol{\beta}}\right\|_2 \leq 2\boldsymbol{\varepsilon}^\top\left(\frac{\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\right)}{\left\|\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\right)\right\|_2}\right) \tag{4}$$

Proof. By the definition of $\hat{\boldsymbol{\beta}}$, we have that

$$\left\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{y}\right\|_2^2 = \min_{\boldsymbol{\beta}\in\mathbb{R}^d}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{y}\|_2^2 \leq \|\mathbf{X}\boldsymbol{\beta}^\star - \mathbf{y}\|_2^2 = \|\boldsymbol{\varepsilon}\|_2^2 \tag{5}$$

where in the last equality we use $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^\star + \boldsymbol{\varepsilon}$. In the meantime, we also have that

$$\left\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{y}\right\|_2^2 = \left\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^\star - \boldsymbol{\varepsilon}\right\|_2^2 = \left\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^\star\right\|_2^2 - 2\boldsymbol{\varepsilon}^\top\mathbf{X}\left(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}\right) + \|\boldsymbol{\varepsilon}\|_2^2 \tag{6}$$

Combining Equations (5) and (6) gives

$$\left\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^\star\right\|_2^2 - 2\boldsymbol{\varepsilon}^\top\mathbf{X}\left(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}\right) + \|\boldsymbol{\varepsilon}\|_2^2 \leq \|\boldsymbol{\varepsilon}\|_2^2 \Rightarrow \left\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^\star\right\|_2^2 \leq 2\boldsymbol{\varepsilon}^\top\mathbf{X}\left(\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}}\right)$$

Dividing both sides by $\left\|\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^\star\right\|_2$ gives the desired result. $\qquad\square$

Now, we can assume specific structure of $\mathbf{X}$ and $\boldsymbol{\beta}$ to derive results based on this generic analysis. In the following parts, we consider two cases: 1) $\mathbf{X}$ is a low-rank matrix; 2) $\boldsymbol{\beta}^\star$ is sparse.

Low rank X

Here we assume that $\mathbf{X}$ is rank-r. In this case, $\mathbf{X}$ has the following singular value decomposition (SVD)

$$\mathbf{X} = \sum_{j=1}^{r}\sigma_j\mathbf{u}_j\mathbf{v}_j^\top$$

Let $\mathbf{U} \in \mathbb{R}^{n \times r}$ denote the matrix where the jth column is $\mathbf{u}_j$ for $j \in [r]$. In other word, $\mathbf{u}_j$ are forming a orthonormal basis for the column space of $\mathbf{X}$. Then for all $\mathbf{a} \in \mathbb{R}^d$ we have

$$\mathbf{X}\mathbf{a} = \sum_{j=1}^{r} \sigma_j \left(\mathbf{v}_j^\top \mathbf{a} \right) \mathbf{u}_j = \mathbf{U}\mathbf{b}$$

where $\mathbf{b} \in \mathbb{R}^r$ has jth entry defined by $b_j = \sigma_j \left(\mathbf{v}_j^\top \mathbf{a} \right) \mathbf{u}_j$.

Moreover, since $\mathbf{u}_j$'s are orthogonal to each other, it holds that $\mathbf{u}_{j_1}^\top \mathbf{u}_{j_2} = 0$ for $j_1 \neq j_2$ and $\mathbf{u}_{j_1}^\top \mathbf{u}_{j_2} = 1$ for $j_1 = j_2$. Therefore, we can compute that

$$\|\mathbf{X}\mathbf{a}\|_2^2 = \sum_{j_1, j_2 = 1}^{r} \sigma_{j_1} \left(\mathbf{v}_{j_1}^\top \mathbf{a} \right) \cdot \sigma_{j_2} \left(\mathbf{v}_{j_1}^\top \mathbf{a} \right) \cdot \left(\mathbf{u}_{j_1}^\top \mathbf{u}_{j_2} \right) = \sum_{j=1}^{r} \sigma_j^2 \left(\mathbf{v}_j^\top \mathbf{a} \right)^2 = \|\mathbf{b}\|_2^2$$

Another way to view the above equations is that since the columns of $\mathbf{U}$ form an orthonormal basis for the column space of $\mathbf{X}$, it naturally follows that any vector in this space, including $\mathbf{X}\mathbf{a}$, can be represented as a linear combination of $\mathbf{U}$'s columns. Furthermore, the squared length of $\mathbf{X}\mathbf{a}$ is just $\|\mathbf{b}\|_2^2$, which is obtained by squaring and summing its coordinates when written in the basis of the columns of $\mathbf{U}$.

This implies that $\frac{\mathbf{X}\mathbf{a}}{\|\mathbf{X}\mathbf{a}\|_2} = \frac{\mathbf{U}\mathbf{b}}{\|\mathbf{b}\|_2}$. Therefore, for all $\mathbf{a} \in \mathbb{R}^d$, there must exists a vector $\mathbf{b} \in \mathbb{R}^r$ with $\|\mathbf{b}\|_2 = 1$ such that

$$\frac{\mathbf{X}\mathbf{a}}{\|\mathbf{X}\mathbf{a}\|_2} = \mathbf{U}\mathbf{b} \, .$$

By Proposition 1, we have that

$$\left\| \mathbf{X}\boldsymbol{\beta}^\star - \mathbf{X}\hat{\boldsymbol{\beta}} \right\|_2 \leq 2\boldsymbol{\varepsilon}^\top \left(\frac{\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star \right)}{\left\| \mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star \right) \right\|_2} \right)$$

$$\leq \sup_{\boldsymbol{\beta} \in \mathbb{R}^d} 2\boldsymbol{\varepsilon}^\top \left(\frac{\mathbf{X}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^\star \right)}{\left\| \mathbf{X}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^\star \right) \right\|_2} \right)$$

$$= \sup_{\mathbf{b} \in \mathbb{R}^r, \|\mathbf{b}\|_2 = 1} 2\boldsymbol{\varepsilon}^\top \mathbf{U}\mathbf{b}$$

$$= \sup_{\mathbf{b} \in \mathbb{R}^r, \|\mathbf{b}\|_2 = 1} 2\left(\mathbf{U}^\top \boldsymbol{\varepsilon} \right)^\top \mathbf{b}$$

$$= 2\left\| \mathbf{U}^\top \boldsymbol{\varepsilon} \right\|_2 \, .$$

The maximization of a dot product occurs when the two vectors are in the same direction. Consequently, the supremum in the last line is attained by the unit vector whose direction is identical to that of $\mathbf{U}^\top \boldsymbol{\varepsilon}$. Thus, the final equality is derived by setting $\mathbf{b} = \frac{\mathbf{U}^\top \boldsymbol{\varepsilon}}{\|\mathbf{U}^\top \boldsymbol{\varepsilon}\|_2}$.

Next, recall that the coordinates of the noise vector $\boldsymbol{\varepsilon}$ were assumed to be sub-Gaussian: $\varepsilon_i \in \mathsf{subG}\left(\sigma^2 \right)$. The sub-Gaussianity of $\boldsymbol{\varepsilon}$ has interesting implications for the concentration

of the length of $\mathbf{U}^\top \boldsymbol{\varepsilon}$. First, note that for the j-th coordinate of $\mathbf{U}^\top \boldsymbol{\varepsilon}$, we have:

$$\left[\mathbf{U}^\top \boldsymbol{\varepsilon} \right]_j = \mathbf{u}_j^\top \boldsymbol{\varepsilon} = \sum_{i=1}^n \mathbf{U}_{ij} \cdot \boldsymbol{\varepsilon}_i \, .$$

Since $\varepsilon_i \in \mathsf{subG}\left(\sigma^2\right)$, and the sum of independent sub-Gaussian random variables is also sub-Gaussian (as established in Lecture 9), each coordinate $\left[\mathbf{U}^\top \boldsymbol{\varepsilon}\right]_j$ is sub-Gaussian. The sub-Gaussianity parameter for this sum is scaled by the sum of the squares of the coefficients. In this case, since $\mathbf{u}_j$ is a column of an orthonormal matrix $\mathbf{U}$, its L_2 norm is 1 ($\|\mathbf{u}_j\|_2 = 1$). This implies that the sum $\sum_{i=1}^n \mathbf{U}_{ij} \cdot \boldsymbol{\varepsilon}_i$ is sub-Gaussian with a variance proxy of $\sigma^2 \|\mathbf{u}_j\|_2^2 = \sigma^2$:

$$\left[\mathbf{U}^\top \boldsymbol{\varepsilon} \right]_j \in \mathsf{subG}\left(\sigma^2\right) \, .$$

This implies that

$$\mathbf{E}\left[\left\| \mathbf{U}^\top \boldsymbol{\varepsilon} \right\|_2^2 \right] = \sum_{j=1}^r \mathbf{E}\left[\left[\mathbf{U}^\top \boldsymbol{\varepsilon} \right]_j^2 \right] \leq \Theta\left(r \cdot \sigma^2 \right) \, .$$

The last inequality above comes from the moment bound for sub-Gaussians. Therefore, we can conclude that

$$\mathbf{E}\left[\frac{1}{n} \left\| \mathbf{X}\boldsymbol{\beta}^\star - \mathbf{X}\hat{\boldsymbol{\beta}} \right\|_2^2 \right] \leq \mathbf{E}\left[\frac{4}{n} \left\| \mathbf{U}^\top \boldsymbol{\varepsilon} \right\|_2^2 \right] = \frac{4}{n} \sum_{j=1}^r \mathbf{E}\left[\left[\mathbf{U}^\top \boldsymbol{\varepsilon} \right]_j^2 \right] \leq O\left(\frac{r\sigma^2}{n} \right) \, .$$

Sparse $\boldsymbol{\beta}^\star$

Let $\mathcal{B}_0^d(k)$ denote the set of k-sparse vectors in $\mathbb{R}^d$:

$$\mathcal{B}_0^d(k) = \left\{ \mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_0 \leq k \right\}$$

where $\|\mathbf{x}\|_0$ counts the number of non-zero entries in $\mathbf{x}$. Assume that $\boldsymbol{\beta}^\star, \hat{\boldsymbol{\beta}} \in \mathcal{B}_0^d(k)$. Then we have that $\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}} \in \mathcal{B}_0^d(2k)$. Define the support set S as

$$S := \left\{ i \in [d] : \left[\boldsymbol{\beta}^\star - \hat{\boldsymbol{\beta}} \right]_i \neq 0 \right\}$$

Then we have that $|S| \leq 2k$. Given a fixed S, let Γ_S denote the subspace of vectors whose support is S

$$\Gamma_S := \left\{ \mathbf{a} \in \mathbb{R}^d : x_i = 0 \; \forall i \notin S \right\}$$

For $\mathbf{a} \in \Gamma_S$, let $\mathbf{a}_S \in \mathbb{R}^{|S|}$ denote the vector that only contains entries of $\mathbf{a}$ with index in S. In the meantime, let $\mathbf{X}_S \in \mathbb{R}^{n \times |S|}$ denote the matrix that only contains jth column with $j \in S$. Therefore, we have that for all $\mathbf{a} \in \Gamma_S$, it must hold that

$$\mathbf{X}\mathbf{a} = \mathbf{X}_S \mathbf{a}_S$$

Since $\mathbf{X}_S \in \mathbb{R}^{n \times |S|}$, we have that $\mathbf{X}_S$ is at most rank-$|S|$. Let $\mathbf{U}_S \in \mathbb{R}^{n \times |S|}$ denote the matrix of the left singular vectors of $\mathbf{X}_S$. Similar to the low-rank assumption, there must exist $\mathbf{b} \in \mathbb{R}^{|S|}$ with $\|\mathbf{b}\|_2 = 1$ such that

$$\frac{\mathbf{Xa}}{\|\mathbf{Xa}\|_2} = \frac{\mathbf{X}_S \mathbf{a}_S}{\|\mathbf{X}_S \mathbf{a}_S\|_2} = \mathbf{U}_S \mathbf{b}$$

Similar to what we have had earlier, we get the following via Proposition 1:

$$
\begin{aligned}
\left\| \mathbf{X}\boldsymbol{\beta}^\star - \mathbf{X}\hat{\boldsymbol{\beta}} \right\|_2 &\leq 2\boldsymbol{\varepsilon}^\top \left(\frac{\mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\right)}{\left\| \mathbf{X}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\right) \right\|_2} \right) \\
&\leq \sup_{\boldsymbol{\beta} \in \mathcal{B}_0^d(k)} 2\boldsymbol{\varepsilon}^\top \left(\frac{\mathbf{X}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^\star\right)}{\left\| \mathbf{X}\left(\boldsymbol{\beta} - \boldsymbol{\beta}^\star\right) \right\|_2} \right) \\
&= \sup_{\mathbf{a} \in \mathcal{B}_0^d(2k)} 2\boldsymbol{\varepsilon}^\top \left(\frac{\mathbf{Xa}}{\|\mathbf{Xa}\|_2} \right) \\
&= \sup_{S \subseteq [d]; |S| = 2k} \sup_{\mathbf{a} \in \Gamma_S} 2\boldsymbol{\varepsilon}^\top \left(\frac{\mathbf{Xa}}{\|\mathbf{Xa}\|_2} \right) \\
&= \sup_{S \subseteq [d]; |S| = 2k} \sup_{\mathbf{b} \in \mathbb{R}^{2k}; \|\mathbf{b}\|_2 = 1} 2\boldsymbol{\varepsilon}^\top \mathbf{U}_S \mathbf{b} \\
&= \sup_{S \subseteq [d]; |S| = 2k} 2 \left\| \mathbf{U}_S^\top \boldsymbol{\varepsilon} \right\|_2
\end{aligned}
$$

Taking the expectation from the squared of both sides leads to:

$$\mathbf{E}\left[\left\| \mathbf{X}\boldsymbol{\beta}^\star - \mathbf{X}\hat{\boldsymbol{\beta}} \right\|_2^2 \right] \leq 4\,\mathbf{E}\left[\sup_{S \subseteq [d]; |S| = 2k} \left\| \mathbf{U}_S^\top \boldsymbol{\varepsilon} \right\|_2^2 \right],$$

where the expectation is taken over the randomness in ε. Now, let's focus on the centered version of $\left\| \mathbf{U}_S^\top \boldsymbol{\varepsilon} \right\|_2^2$:

$$\mathbf{E}\left[\left\| \mathbf{X}\boldsymbol{\beta}^\star - \mathbf{X}\hat{\boldsymbol{\beta}} \right\|_2^2 \right] \leq 4\,\mathbf{E}\left[\sup_{S \subseteq [d]; |S| = 2k} \left(\left\| \mathbf{U}_S^\top \boldsymbol{\varepsilon} \right\|_2^2 - \mathbf{E}\left[\left\| \mathbf{U}_S^\top \boldsymbol{\varepsilon} \right\|_2^2 \right] \right) \right] + \sup_{S \subseteq [d]; |S| = 2k} \mathbf{E}\left[\left\| \mathbf{U}_S^\top \boldsymbol{\varepsilon} \right\|_2^2 \right],$$

The first expression represents the expected maximum deviation from the mean of a set of centered random variables, $\left\| \mathbf{U}_S^\top \boldsymbol{\varepsilon} \right\|_2^2$, which are indexed by S. There are at most $\binom{d}{2k}$ such variables. Earlier in this lecture, we have shown that every coordinate of $\mathbf{U}^\top \boldsymbol{\varepsilon}$ is a sub-Gaussian random variable. It is not too difficult to show that leads to proving that

$$\left[\mathbf{U}^\top \boldsymbol{\varepsilon} \right]_j \in \mathsf{subG}\left(\sigma^2 \right) \qquad \Rightarrow \qquad \left[\mathbf{U}^\top \boldsymbol{\varepsilon} \right]_j^2 - \mathbf{E}\left[\left[\mathbf{U}^\top \boldsymbol{\varepsilon} \right]_j^2 \right] \in \mathsf{subE}\left(c \cdot \sigma^2 \right),$$

For some absolute constant c. We formally prove this in Lemma 2.[2]

Now, $\left\|\mathbf{U}_S^\top \boldsymbol{\varepsilon}\right\|_2^2$ represent sum of at most $|S| = 2k$ sub-exponential random variables, and therefore it is a sub-exponential random variable itself with parameters $(k\sigma^2)$. This leads to an upper bound for the expected maximum, given by Lemma 3:

$$4\,\mathbf{E}\left[\sup_{S \subseteq [d]; |S|=2k} \left(\left\|\mathbf{U}_S^\top \boldsymbol{\varepsilon}\right\|_2^2 - \mathbf{E}\left[\left\|\mathbf{U}_S^\top \boldsymbol{\varepsilon}\right\|_2^2\right]\right)\right] \leq O\left(k \cdot \sigma^2 \log \binom{d}{2k}\right)$$

where C is an absolute constant.

Additionally, a second term in the overall expression is a lower-order term bounded by $O(k\sigma^2)$ (by the moment bounds of sub-exponentials). By applying the standard estimate for the binomial coefficient, $\binom{d}{k} \leq \left(\frac{e\,d}{k}\right)^k$, we can conclude that:

$$\mathbf{E}\left[\frac{1}{n}\left\|\mathbf{X}\boldsymbol{\beta}^\star - \mathbf{X}\hat{\boldsymbol{\beta}}\right\|_2^2\right] \leq O\left(\frac{\sigma^2 k^2 \log(d/k)}{n}\right)$$

Some useful lemmas

Lemma 2. *If X is a sub-Gaussian random variable with parameter σ^2, then $X^2 - \mathbf{E}[X^2]$ is a sub-exponential random variable with parameter $K' = \Theta(\sigma^2)$.*

Proof. We show that the moments of $X^2 - \mathbf{E}[X^2]$ satisfy the growth condition for sub-exponential variables. In particular, we will bound the L_p-norm of $X^2 - \mathbf{E}[X^2]$.

By the triangle inequality for the L_p-norm (Minkowski's inequality), we have:

$$\left\|X^2 - \mathbf{E}[X^2]\right\|_{L_p} \leq \left\|X^2\right\|_{L_p} + \left\|\mathbf{E}[X^2]\right\|_{L_p} \tag{7}$$

The second term is the L_p-norm of a constant, so $\left\|\mathbf{E}[X^2]\right\|_{L_p} = \mathbf{E}[X^2]$. The first term is $(\mathbf{E}[X^{2p}])^{1/p}$ by definition. Thus,

$$\left\|X^2 - \mathbf{E}[X^2]\right\|_{L_p} \leq (\mathbf{E}[X^{2p}])^{1/p} + \mathbf{E}[X^2] \tag{8}$$

Now, we use the sub-Gaussian moment property to bound each term on the right-hand side. From the sub-Gaussian moment bound, we know $(\mathbf{E}[|X|^q])^{1/q} \leq \sigma\sqrt{q}$. Let $q = 2p$.

$$(\mathbf{E}[X^{2p}])^{1/(2p)} \leq \sigma\sqrt{2p}$$

To get the desired form, we raise both sides to the power of 2:

$$(\mathbf{E}[X^{2p}])^{1/p} \leq 2\sigma^2 p \tag{9}$$

---

[2]It is important to note here, we are using the Vershynin's book [Ver18] notations of sub-exponential we discussed in Lecture 11.

We use the sub-Gaussian moment bound with $p = 2$ to bound the second term:

$$(\mathbf{E}[X^2])^{1/2} \leq \sigma\sqrt{2}\,.$$

Squaring both sides gives:

$$\mathbf{E}[X^2] \leq 2\sigma^2\,. \tag{10}$$

Substitute the results from (9) and (10) back into (8):

$$\left\| X^2 - \mathbf{E}[X^2] \right\|_{L_p} \leq 2\sigma^2 p + 2\sigma^2 = 2\sigma^2(p+1)$$

For any $p \geq 1$, we have $p + 1 \leq 2p$. Therefore,

$$\left\| X^2 - \mathbf{E}[X^2] \right\|_{L_p} \leq 4\sigma^2 p$$

Let $K = 4\sigma^2$. We have shown that for all $p \geq 1$,

$$(\mathbf{E}[|X^2 - \mathbf{E}[X^2]|^p])^{1/p} \leq Kp$$

This is the moment condition for a random variable to be sub-exponential.

Thus, $X^2 - \mathbf{E}[X^2]$ is sub-exponential. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\square$

Lemma 3. *Let $X_1, \ldots, X_n$ be random variables such that each X_i is sub-exponential with a parameter $K > 0$, i.e., for any $t \geq 0$,*

$$\mathbf{Pr}[|X_i| \geq t] \leq 2\exp(-t/K).$$

Then the expected value of their maximum is bounded as:

$$\mathbf{E}\left[\max_{i=1}^{n} X_i\right] \leq K(1 + \ln(2n)).$$

Proof. Let $M_n = \max_{i=1}^{n} X_i$. We aim to show that $\mathbf{E}[M_n]$ has an upper bound proportional to $\ln n$. The plan is to find a tail bound for M_n, and then use the integral identity for the expectation.

Since the integral identity works for positive values only, we bound the expectation of the positive part, $\mathbf{E}[M_n^+] = \mathbf{E}[\max(M_n, 0)]$, using the integral identity formula:

$$\mathbf{E}[M_n^+] = \int_0^\infty \mathbf{Pr}[M_n > t]dt.$$

We split this integral into two parts at a threshold t_0, which we will define shortly.

$$\mathbf{E}[M_n^+] = \int_0^{t_0} \mathbf{Pr}[M_n > t]dt + \int_{t_0}^\infty \mathbf{Pr}[M_n > t]dt.$$

Let's choose $t_0 = K\ln(2n)$. This choice is motivated to balance the two parts of the integral.

For the first part of the integral, we use the simple fact that any probability is at most 1:

$$\int_0^{t_0} \mathbf{Pr}[M_n > t]dt \leq \int_0^{t_0} 1\, dt = t_0 = K\ln(2n).$$

For the second part, we use the union bound on the tail probability of the maximum:

$$\mathbf{Pr}[M_n > t] = \mathbf{Pr}[\cup_{i=1}^n \{X_i > t\}] \leq \sum_{i=1}^n \mathbf{Pr}[X_i > t] \leq \sum_{i=1}^n 2\exp(-t/K) = 2n\exp(-t/K).$$

Now we integrate this tail bound from t_0 to infinity:

$$\begin{aligned}
\int_{t_0}^\infty \mathbf{Pr}[M_n > t]dt &\leq \int_{t_0}^\infty 2n\exp(-t/K)dt \\
&= 2n\left[-K\exp(-t/K)\right]_{t_0}^\infty \\
&= 2nK\exp(-t_0/K).
\end{aligned}$$

By substituting our choice of $t_0 = K\ln(2n)$, we get:

$$\exp(-t_0/K) = \exp\left(-\frac{K\ln(2n)}{K}\right) = \exp(-\ln(2n)) = \frac{1}{2n}.$$

Therefore, the second integral is bounded by a constant:

$$\int_{t_0}^\infty \mathbf{Pr}[M_n > t]dt \leq 2nK\left(\frac{1}{2n}\right) = K.$$

Finally, we combine the bounds on the two parts of the integral:

$$\mathbf{E}[M_n] \leq \mathbf{E}[M_n^+] = \int_0^{t_0} \mathbf{Pr}[M_n > t]dt + \int_{t_0}^\infty \mathbf{Pr}[M_n > t]dt \leq K\ln(2n) + K.$$

This gives us the desired result:

$$\mathbf{E}\left[\max_{i=1}^n X_i\right] \leq K(1 + \ln(2n)).$$

This completes the proof.   □

Bibliographic Note

The content of this lecture was adapted from the lecture notes of Prof. Sasha Rakhlin for "Mathematical Statistics: A Non-Asymptotic Approach", which can be found here [Rak22].

References

[Rak22] Alexander Rakhlin. Mathematical statistics: A non-asymptotic approach, 2022. Lecture notes for MIT course IDS.160, Spring 2022.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science.* Cambridge University Press, 2018.