Lecture 12

Linear regression

Suppose we have an unknown vector $\beta^* \in \mathbb{R}^d$.

We observe linear observation of β^* of the form:

$$\underbrace{Y_i}_{\text{known}} = \langle \underbrace{x_i}_{\text{known}}, \underbrace{\overset{\text{unknown}}{\beta^*}}_{\text{in } \mathbb{R}^d} \rangle + \underbrace{\varepsilon_i}_{\text{noise}} \quad i=1,\dots,n$$

Assume ε_i's are zero-mean and in $\mathrm{SubG}(\sigma^2)$

$$Y = \underbrace{X}^{\mathbb{R}^{n\times d}} \beta^* + \varepsilon$$

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \text{---} & x_1 & \text{---} \\ & \vdots & \\ \text{---} & x_n & \text{---} \end{bmatrix} \begin{bmatrix} \beta^* \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Fix design: we assume X is fixed.

[Another interesting regime is when x_i's are random.]

Goal:

find $\hat{\beta}$ such that $\hat{\beta}$ is close to β

what does close mean?

- small distance to $\hat{\beta}$ and β^*

  say $\|\hat{\beta} - \beta\|_2^2$ is small

- small "de-noising objective"

  $Y = X\hat{\beta}$ is similar to $Y^* = X\beta^*$

$$\frac{1}{n} \sum_{i=1}^{n} \left(\langle x_i, \hat{\beta} \rangle - \langle x_i, \beta^* \rangle \right)^2$$

$$= \frac{1}{n} \| X\hat{\beta} - X\beta^* \|_2^2$$

Clearly, we do not have β^*. Thus, it is difficult to measure the quality of $\hat{\beta}$. What we usually do is to pick a "proxy" quantity for these measurses and find $\hat{\beta}$ that minimize them.

While a great deal of effort is dedicated to finding solutions. It is always important to look back and see the solution we have found via optimizing the proxy is indeed a good solution for the original objective as well.

Solution

$$\hat{\beta} \in \arg\min_{\beta} \| X\beta - Y \|_2^2 \quad *$$

$$= \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left(\langle x_i, \beta \rangle - y_i \right)^2$$

the gradient of $\| X\beta - Y \|_2^2 = 0$

$$\| X\beta - Y \|_2^2 = (\beta^T X^T - Y^T) . (X\beta - Y)$$

$$= \beta^T X^T X \beta - 2 \beta^T X^T Y + Y^T Y$$

$$\Rightarrow \quad X^T X \beta = X^T Y$$

$$\Rightarrow \quad \beta = \underbrace{(X^T X)^{+}}_{\text{pseudoinverse}} X^T Y$$

[not the focus of this lecture]

Given Sub-Gaussianity assumption on ε,

what can we say about the error

of $\hat{\beta}$?

Can we exploit any structure in X ?

such as low rank X ? or sparsity of β^* ?

Since β^* is an arg min in * we have:

$$\|X\hat{\beta} - Y\|_2^2 \le \|X\beta^* - Y\|_2^2 = \|\varepsilon\|_2^2 \quad ①$$

on the other hand:

$$\|X\hat{\beta} - Y\|_2^2 = \|X\hat{\beta} - X\beta^* - \varepsilon\|_2^2$$

$$= \|X\hat{\beta} - X\beta^*\|_2^2 - 2\langle \varepsilon. X\hat{\beta} - X\beta^*\rangle + \|\varepsilon\|_2^2 \quad ②$$

$$①, ② \Rightarrow \frac{1}{n}\|X\hat{\beta} - X\beta^*\|_2^2 \le \frac{2}{n}\langle \varepsilon. X\hat{\beta} - X\beta^*\rangle$$

(basic inequality)

$$\Rightarrow \quad \| X\hat{\beta} - X\beta^* \|_2 \leq 2 \left\langle \varepsilon, \frac{X\hat{\beta} - X\beta^*}{\| X\hat{\beta} - X\beta^* \|} \right\rangle$$

$$\leq 2 \sup_{\beta} \left\langle \varepsilon, \frac{X\beta - X\beta^*}{\| X\beta - X\beta^* \|_2} \right\rangle \quad **$$

$\hookrightarrow$ does not on $\hat{\beta}$ any more.

Let $U = [u_1, \dots, u_r]$ be a matrix with orthonormal columns. a basis for column space of X

(where r is the rank of $X^T X$)

$\dfrac{X\beta - X\beta^*}{\| X\beta - X\beta^* \|_2}$ is a vector in column space of X

Hence, it can be written in the basis u_i's

$$\exists\, a : \quad \frac{X\beta - X\beta^*}{\| X\beta - X\beta^* \|} = \frac{U a}{\| a \|}$$

$$\|X\hat{\beta} - X\beta^*\|_2 \le 2 \sup_{\|a\| \le 1} \langle \varepsilon, Ua \rangle$$

$$\le 2 \sup_{\|a\| \le 1} \langle U^T \varepsilon, a \rangle$$

$$= 2 \|U^T \varepsilon\|$$

$$\Rightarrow \frac{1}{n} \|X\hat{\beta} - X\beta^*\|_2^2 \le 4 \|U^T \varepsilon\|_2^2$$

$U \in \mathbb{R}^{n \times r}$

$$U^T \cdot \varepsilon = \begin{bmatrix} \text{———} & u_1 & \text{———} \\ & \vdots & \\ \text{———} & u_r & \text{———} \end{bmatrix} \begin{bmatrix} \\ \varepsilon \\ \\ \end{bmatrix}$$

Let $v = U^T \varepsilon$. $\Rightarrow$ $v_i = \langle u_i, \varepsilon \rangle$

$$\Rightarrow v_i = \sum_{j=1}^{n} u_{ij} \cdot \varepsilon_j$$

$$\Rightarrow v_i \in \mathrm{SubG}\left(\underbrace{\sum u_{ij}^2}_{=1} \sigma^2 \right) \in \mathrm{SubG}(\sigma^2)$$

$$\mathbb{E}_{\varepsilon}\left[\frac{1}{n}\|X\hat{\beta}-X\beta^*\|_2^2\right] \leq \frac{4}{n}\mathbb{E}_{\varepsilon}\left[\|U^T\varepsilon\|_2^2\right]$$

$$= \frac{4}{n}\mathbb{E}\left[\sum_{i=1}^{r} v_i^2\right] \leq \frac{4}{n}\sum_{i=1}^{r}\mathbb{E}[v_i^2]$$

$$\leq \frac{4}{n}\sum_{i=1}^{r}(\sqrt{2}\sigma)^2 \lesssim \theta\left(\frac{r\sigma^2}{n}\right)$$

depends on r not d

$r \leq \min\{n, d\}$

Sparsity

Consider the case where all but k coordinate of β^* is zero.

$$B_0^d(k) := \{ x \in \mathbb{R}^d . \|x\|_0 \leq k \}$$

↳ unit ball

we also pick $\hat{\beta}$ from $B_0^d(k)$

$$\hat{\beta} = \underset{\beta \in B_0^d(k)}{\arg\min} \| X\beta - Y \|_2^2$$

Now we focus on bounding

$$\| X\hat{\beta} - X\beta^* \|_2^2 .$$

Eerlier, we have shown

$$\| X\hat{\beta} - X\beta^* \|_2 \leq 2 \left\langle \varepsilon, \frac{X\hat{\beta} - X\beta^*}{\| X\hat{\beta} - X\beta^* \|_2} \right\rangle$$

$$\leq 2 \sup_{\beta \in B_0^d(k)} \left\langle \varepsilon, \frac{X(\beta - \beta^*)}{\| X(\beta - \beta^*) \|_2} \right\rangle$$

Now $\beta - \beta^*$ is a vector in $\mathbb{R}^d$ with at most $2k$ non-zero entries. Let S denote the set of indices that are not zero.

we know $|S| \leq 2k$

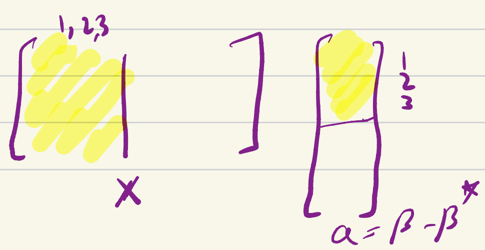
we can continue our bound by:

$$2 \sup_{\beta \in \mathcal{B}_0^d(k)} \left\langle \varepsilon, \frac{X(\beta-\beta^*)}{\|X(\beta-\beta^*)\|_2} \right\rangle$$

$$\le 2 \max_{\substack{S \subseteq [d] \\ |S| \le 2k}} \sup_{a_s} \left\langle \varepsilon, \frac{X a_s}{\|X a_s\|} \right\rangle$$

where a_i is zero for all $i \notin S$

Note that $X a_s$ lies in the column space of X_S restricted to columns that are in S.



for example $S = \{1, 2, 3\}$

Let $U_S = [u_1, \ldots, u_r]$ be a matrix where its columns form an orthonormal basis for the column space of X_S.

here $r \leq 2k$

with a very similar argument as before

$$\| X\hat{\beta} - X\beta^* \|_2 \leq 2 \max_{\substack{S \subseteq [d] \\ |S| \leq 2k}} \| U_S^T \varepsilon \|_2$$

$$\Rightarrow \mathbb{E}_\varepsilon \left[\frac{1}{n} \| X\hat{\beta} - X\beta^* \|_2^2 \right] \leq 2 \, \mathbb{E}_\varepsilon \left[\max_{\substack{S \subseteq [d] \\ |S| \leq 2k}} \| U_S^T \varepsilon \|_2^2 \right]$$

$$\leq \mathcal{O}\left(\frac{\sigma^2 k \log(d)}{n} \right)$$

next lecture