

Statistical Learning II - Lecture 6 (14/10/2026)

Q: How much data we need to learn (generalise)

Recall: $R(f_\theta) = \mathbb{E}[(y - f_\theta(x))^2]$

↳ w.r.t data distribution $(x, y) \sim p$ iid.

Need a model! For $f_\theta(x) = \langle \theta, x \rangle$ can only fit linear functions (inductive bias).

Therefore, assume that $y_i = \underbrace{\langle \theta_*, x_i \rangle}_{\text{signal}} + \underbrace{\varepsilon_i}_{\text{noise}}$ $\theta_* \in \mathbb{R}^d$

with $\mathbb{E}[\varepsilon_i] = 0$ ($\mathbb{E}_\varepsilon[y_i] = \langle \theta_*, x_i \rangle$)
 $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$

This defines distribution of $y|x$. Recall that $p(x, y) = p(y|x) \times p(x)$

To simplify problem, assume fixed design, i.e. $\{x_i\}_{i=1}^n$ is fixed, not random. fix this

Under fixed design, generalisation error is given by:

$$R(f_\theta) = \mathbb{E}_{y|x} \left[\frac{1}{n} \|y - X\theta\|_2^2 \right] \quad \left(\text{compare with } \mathbb{E}_{(x,y)} [(y - \langle \theta, x \rangle)^2] \right)$$

$y \in \mathbb{R}^n, X \in \mathbb{R}^{n \times d}$ $y \in \mathbb{R}, x \in \mathbb{R}^d$

Exercise: check that for $x_i \sim \text{iid } p_X$,

$$\mathbb{E} \left[\frac{1}{n} \|y - X\theta\|_2^2 \right] = \mathbb{E} [(y - \langle \theta, x \rangle)^2]$$

$\nearrow (x, y) = (x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \quad \nwarrow (x, y)$

⚠ ε_i is not Gaussian !!!

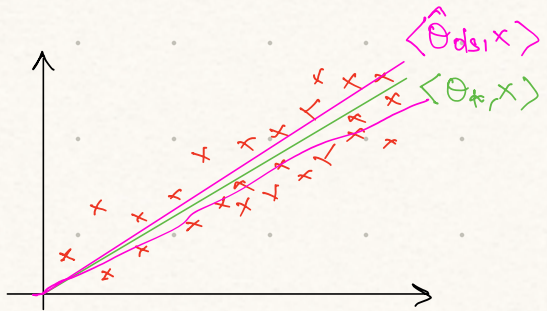
Bias - variance decomposition

Remember OLS predictor is given by :

$$\hat{\theta}_{OLS} = X^T y = \begin{cases} (X^T X)^{-1} X^T y & \text{for } n \geq d \\ X^T (X X^T)^{-1} y & \text{for } n < d \end{cases}$$

X full rank ($\text{rank}(X) = \min(n, d)$)

• For $n \geq d$: $\hat{\theta}_{OLS}(X, y) = (X^T X)^{-1} X^T y = (X^T X)^{-1} X^T (X \theta_* + \varepsilon)$



$$\begin{aligned} &= (X^T X)^{-1} X^T X \theta_* + (X^T X)^{-1} X^T \varepsilon \\ &= \underbrace{\theta_*}_{\text{Signal}} + \underbrace{(X^T X)^{-1} X^T \varepsilon}_{\text{Noise}} \end{aligned}$$

On average : $\mathbb{E}_{\varepsilon} [\hat{\theta}_{OLS}(X, y)] = \mathbb{E}_{\varepsilon} [\theta_* + (X^T X)^{-1} X^T \varepsilon]$

" Bias

$$\begin{aligned} &= \theta_* + (X^T X)^{-1} X^T \mathbb{E}[\varepsilon] \\ &= \theta_* \quad (=0) \end{aligned}$$

$\therefore \hat{\theta}_{OLS}(X, y)$ is unbiased for $n \geq d$.

What about the variance ?

$$\begin{aligned} \text{Cov}_{\varepsilon}(\hat{\theta}_{OLS}(X, y)) &= \mathbb{E}[(\hat{\theta}_{OLS} - \mathbb{E}[\hat{\theta}_{OLS}])(\hat{\theta}_{OLS} - \mathbb{E}[\hat{\theta}_{OLS}])^T] \\ &= \mathbb{E}[(\hat{\theta}_{OLS} - \theta_*)(\hat{\theta}_{OLS} - \theta_*)^T] \end{aligned}$$

$$\hat{\theta}_{OLS} - \theta_* = \theta_* + (X^T X)^{-1} X^T \varepsilon - \theta_* = (X^T X)^{-1} X^T \varepsilon$$

$$\therefore \text{Cov}(\hat{\theta}_{OLS}) = \mathbb{E}[(X^T X)^{-1} X^T \varepsilon (X^T X)^{-1} X^T \varepsilon^T]$$

Recall : $(AB)^T = B^T A^T$

$$X \in \mathbb{R}^{n \times d}$$

$$= \mathbb{E}_{\varepsilon} \left[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1} \right]$$

$$= (X^T X)^{-1} X^T \mathbb{E}(\varepsilon \varepsilon^T) X (X^T X)^{-1}$$

$\underbrace{\underbrace{d \times n \times n \times d}_{d \times d} \quad d \times n}_{d \times n} \quad \underbrace{\underbrace{n \times 1 \times 1 \times n}_{n \times n} \quad n \times d \quad d \times d}_{n \times d} \quad n \times d$
 $\underbrace{\hspace{10em}}_{d \times d}$

$$\left. \begin{aligned} & \bullet (A^T)^T = A \\ & \bullet (A^{-1})^T = (A^T)^{-1} \end{aligned} \right\}$$

$$A = \varepsilon \varepsilon^T \in \mathbb{R}^{n \times n}, \quad A_{ij} = \varepsilon_i \varepsilon_j \quad \mathbb{E}_{\varepsilon}[A_{ij}] = \mathbb{E}_{\varepsilon}[\varepsilon_i \varepsilon_j]$$

$$\text{Recall: } \varepsilon \text{ is iid, } \varepsilon_i \sim \rho_{\varepsilon} \quad \forall i=1, \dots, n \quad = \begin{cases} \sigma^2 & i=j \\ 0 & i \neq j \end{cases}$$

$$\therefore \mathbb{E}[\varepsilon \varepsilon^T] = \sigma^2 I_n \in \mathbb{R}^{n \times n}$$

$$\begin{aligned} \text{Cov}(\hat{\theta}_{ols}) &= (X^T X)^{-1} X^T (\sigma^2 I_n) X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \end{aligned}$$

$$\text{Since } \hat{\Sigma}_n = \frac{1}{n} X^T X \quad \text{I can also write } \text{Cov}(\hat{\theta}_{ols}) = \frac{\sigma^2}{n} \hat{\Sigma}_n^{-1}$$

$$\begin{aligned} \text{To summarize:} \quad \text{Bias: } \mathbb{E}[\hat{\theta}_{ols}] &= \theta_* \\ \text{Variance: } \text{Cov}(\hat{\theta}_{ols}) &= \frac{\sigma^2}{n} \hat{\Sigma}_n^{-1} \end{aligned}$$