2.7 Fixed design linear regression

Prediction: f: input $\to$ output

Fixed design: Input $\{x_1, \cdots, x_n\}$ is fixed. $x_i \in \mathbb{R}^d$

Assumption: There is a true underlying parameter $\theta^* \in \mathbb{R}^d$:

$$y_i = x_i \cdot \theta^* + \varepsilon_i \qquad \forall i = 1, \cdots, n,$$

where we assume ε_i are i.i.d noise terms with $E[\varepsilon_i] = 0$ and $Var[\varepsilon_i] = \sigma^2$.

Training: observe $y_1, \cdots, y_n$ and take notations below:

(i) $X = [x_1, \cdots, x_n]^T \in \mathbb{R}^{n \times d}$

(ii) $\varepsilon = [\varepsilon_1, \cdots, \varepsilon_n]^T \in \mathbb{R}^d$

(iii) $Y = [y_1, \cdots, y_n]^T \in \mathbb{R}^d$

(iv) $\Sigma = \frac{1}{n} X^T X \in \mathbb{R}^{d \times d}$ (second moment matrix)

Optimising: minimise the expected risk defined as:

$$\min L(\theta) := \frac{1}{n} \sum_{i=1}^{n} E[(x_i \cdot \theta - y_i)^2] = \frac{1}{n} E[\|X\theta - Y\|_2^2]$$

the expectation is over the randomness in Y.

$$\hat{\theta} := \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \frac{1}{n} \|X\theta - Y\|_2^2 \qquad \text{(least square error)}$$

By taking derivative we have

$$\hat{\theta} = (X^T X)^{-1} X^T Y = \frac{1}{n} \Sigma^{-1} X^T Y \qquad (\text{assume } \Sigma \text{ is invertible})$$

For simplicity, we now study $E[L(\hat{\theta})]$. We start with an arbitrary θ:

$$\begin{aligned} L(\theta) &= \frac{1}{n} E[\|X\theta - Y\|_2^2] \\ &= \frac{1}{n} E[\|X\theta - X\theta^* - \varepsilon\|_2^2] \\ &= \frac{1}{n} E[\|X\theta - X\theta^*\|_2^2 + \|\varepsilon\|_2^2] \qquad (E[\varepsilon] = 0) \\ &= \frac{1}{n} (\theta - \theta^*)^T (X^T X)(\theta - \theta^*) + \sigma^2 \\ &= \|\theta - \theta^*\|_\Sigma^2 + \sigma^2 \end{aligned}$$

Note: The first term is the squared distance between θ and

θ^* as measured by Σ. If the data varies slowly in one direction, then the discrepancy of θ and θ^* in that direction will be downweighted. We assert $L(\theta^*) = \sigma^2$

Now we turn back to $\hat{\theta}$.

$$
\begin{aligned}
L(\hat{\theta}) - L(\theta^*) &= \|\hat{\theta} - \theta^*\|_\Sigma^2 \\
&= \frac{1}{n}\|X\hat{\theta} - X\theta^*\|_2^2 \\
&= \frac{1}{n}\|X(X^TX)^{-1}X^T(X\theta^* + \varepsilon) - X\theta^*\|_2^2 \\
&= \frac{1}{n}\|X(X^TX)^{-1}X^T\varepsilon\|_2^2 \\
&= \frac{1}{n}\mathrm{tr}(X(X^TX)^{-1}X^T\varepsilon\varepsilon^T) \\
&= \frac{1}{n}\mathrm{tr}(\varepsilon\varepsilon^T) \qquad (\mathrm{tr}(AB) = \mathrm{tr}(BA))
\end{aligned}
$$

Take expectation and using the fact $E[\varepsilon\varepsilon^T] = \sigma^2 I$:

$$E[L(\hat{\theta}) - L(\theta^*)] = \frac{d\sigma^2}{n}$$