

TD 5 — OLS and Bias–Variance

Reminder. We observe n independent pairs satisfying

$$y_i = x_i^\top \theta^* + \sigma \varepsilon_i, \quad \mathbb{E}[x_i] = 0, \quad \mathbb{E}[x_i x_i^\top] = \Sigma \succ 0, \quad \varepsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1),$$

where $x_i \in \mathbb{R}^d$, θ^* is fixed, and the errors are independent of all inputs. Let $X \in \mathbb{R}^{n \times d}$ have rows $x_i^\top$ and $y = (y_1, \dots, y_n)^\top$. We use the linear model parameterized by $\theta \in \mathbb{R}^d$: $\mathcal{F} = \{f_\theta : x \mapsto x^\top \theta : \theta \in \mathbb{R}^d\}$. Define

- The **empirical risk** as

$$\widehat{R}_n(\theta) = \frac{1}{2n} \|y - X\theta\|_2^2$$

- The **population risk** as

$$R(\theta) = \frac{1}{2} \mathbb{E}_{x_0, y_0} [(y_0 - f_\theta(x_0))^2] = \frac{1}{2} \mathbb{E}_{x_0, y_0} [(y_0 - x_0^\top \theta)^2]$$

where (x_0, y_0) is a fresh observation from the same model independent of the training set.

- The **excess risk** as

$$\mathcal{E}(\theta) = R(\theta) - R^*$$

where R^* is the Bayes risk for the quadratic loss. Only the randomness in the fresh observation is averaged; the training sample is held fixed.

- The **generalization error** (expected excess risk) of an estimator $\widehat{\theta}$ as

$$\mathcal{E}_{\text{gen}}(\widehat{\theta}) = \mathbb{E}_{\mathcal{D}}[\mathcal{E}(\widehat{\theta})] = \mathbb{E}_{\mathcal{D}}[R(\widehat{\theta})] - R^*.$$

The outer expectation averages over the training sample $\mathcal{D}$; the fresh observation is already averaged in R .

- The **OLS** estimator $\widehat{\theta}$ as minimizing the empirical risk. It must satisfy the normal equations

$$X^\top X \widehat{\theta} = X^\top y.$$

Exercise 1 — Bias–variance decomposition

Here we assume that X is full rank, i.e., $\text{rank}(X) = d$.

1. For a generic random estimator $\widetilde{\theta}$, define its bias and variance as

$$\text{Bias}(\widetilde{\theta}) = \mathbb{E}[\widetilde{\theta}] - \theta^*, \quad \text{Var}(\widetilde{\theta}) = \text{Cov}(\widetilde{\theta}) = \mathbb{E}[(\widetilde{\theta} - \mathbb{E}[\widetilde{\theta}])(\widetilde{\theta} - \mathbb{E}[\widetilde{\theta}])^\top]$$

Give the expression of the mean squared error $\mathbb{E} \left[\left\| \widetilde{\theta} - \theta^* \right\|_2^2 \right]$ as a function of the bias and variance of $\widetilde{\theta}$.

2. Give the expression of the OLS $\widehat{\theta}$. What is the source of randomness in $\widehat{\theta}$?
3. Compute the expectation $\mathbb{E}[\widehat{\theta}]$ of the OLS with respect of its source of randomness. What is its bias?
4. Compute the covariance $\text{Cov}(\widehat{\theta})$ of the OLS with respect of its source of randomness.

5. Deduce the mean squared error of the OLS estimator $\mathbb{E} \left[\left\| \hat{\theta} - \theta^* \right\|_2^2 \right]$.

Exercise 2 — Mean squared error in the degenerate case

This time we no longer assume that X is full rank.

1. Give the expression of the OLS estimator $\hat{\theta}$ in the degenerate case using the Moore–Penrose pseudoinverse of X .
2. Compute the expectation $\mathbb{E}[\hat{\theta}]$ of the OLS with respect of its source of randomness. What is its bias?
3. Compute the covariance $\text{Cov}(\hat{\theta})$ of the OLS with respect of its source of randomness.
4. Deduce the mean squared error of the OLS estimator $\mathbb{E} \left[\left\| \hat{\theta} - \theta^* \right\|_2^2 \right]$.

Exercise 3 — Bayes risk and bias–variance decomposition

1. Compute the Bayes risk R^* and show that for any parameter θ ,

$$\mathcal{E}(\theta) = \frac{1}{2}(\theta - \theta^*)^\top \Sigma (\theta - \theta^*).$$

Is there a similarity with the mean squared error?

2. For an estimator $\tilde{\theta}$ with finite second moments, express $\mathcal{E}_{\text{gen}}(\tilde{\theta})$ in terms of its bias and covariance, and identify the squared bias and variance contributions.
3. Apply this decomposition to OLS, assuming $\text{rank}(X) = d$ almost surely and finite second moments, and show that

$$\mathcal{E}_{\text{gen}}(\hat{\theta}) = \frac{1}{2} \text{tr}(\Sigma \text{Cov}_{\mathcal{D}}(\hat{\theta})).$$