



Statistical Learning II

Lecture 6 - Bias-Variance decomposition

Bruno Loureiro
@ CSD, DI-ENS & CNRS

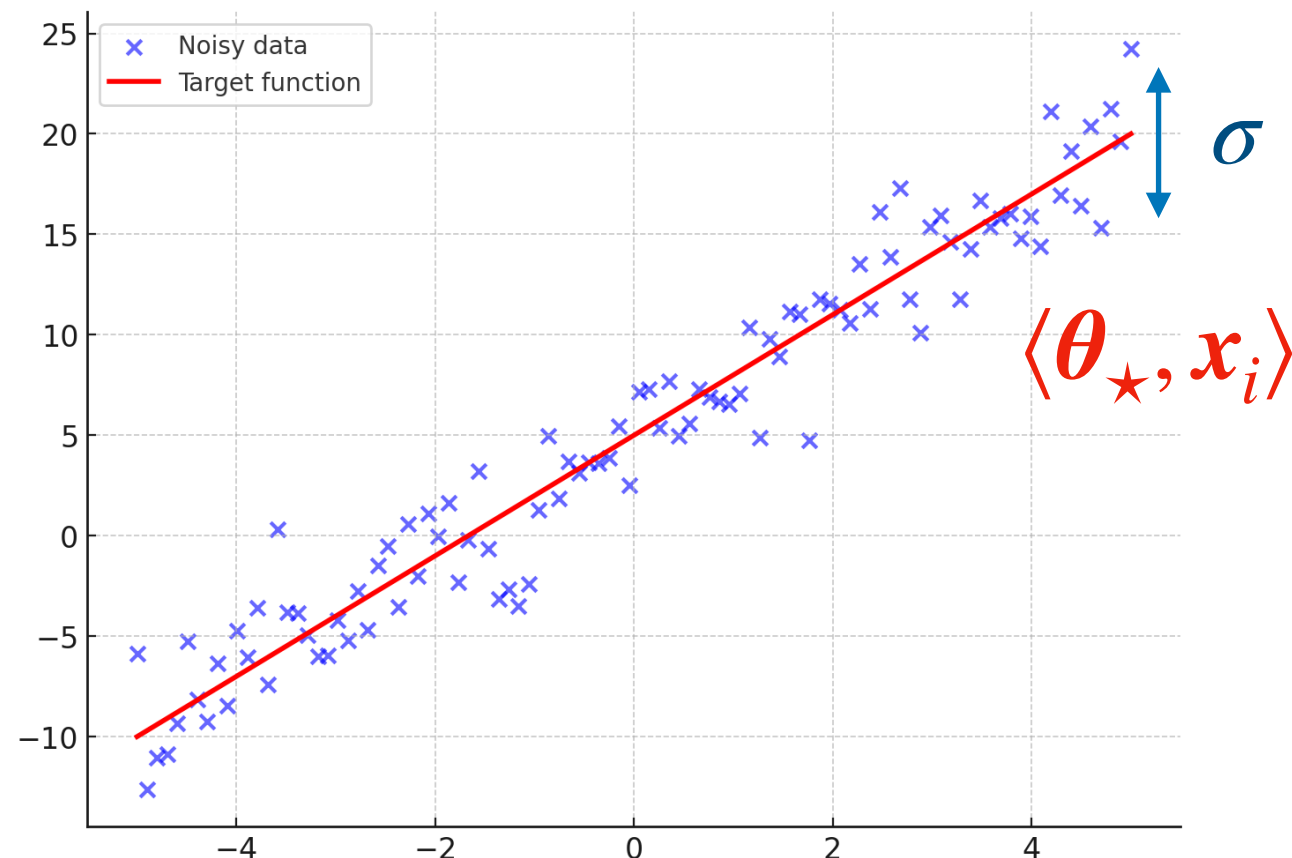
brloureiro@gmail.com

Assumptions

We now assume the following data generative model:

$$y_i = \langle \theta_\star, x_i \rangle + \varepsilon_i$$

- With:
- Fixed $\theta_\star \in \mathbb{R}^d$ and $x_i \in \mathbb{R}^d$ “fixed design”
 - $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$



Decomposition of OLS

$$\hat{\boldsymbol{\theta}}_{OLS}(X, \mathbf{y}) = \boldsymbol{\theta}_{\star} + \frac{1}{n} \hat{\boldsymbol{\Sigma}}_n^{-1} X^{\top} \boldsymbol{\varepsilon}$$

“signal”

“noise”

In particular:

- Bias: $\mathbb{E}_{\boldsymbol{\varepsilon}} \left[\hat{\boldsymbol{\theta}}_{OLS}(X, \mathbf{y}) \right] = \boldsymbol{\theta}_{\star}$ “Unbiased”
- Variance: $\text{Var}_{\boldsymbol{\varepsilon}} \left[\hat{\boldsymbol{\theta}}_{OLS}(X, \mathbf{y}) \right] = \frac{\sigma^2}{n} \hat{\boldsymbol{\Sigma}}_n^{-1}$

Informally, if $\hat{\boldsymbol{\Sigma}}_n \rightarrow \boldsymbol{\Sigma}$ a rank d matrix as $n \rightarrow \infty$, then:

$$\hat{\boldsymbol{\theta}}_{OLS} \rightarrow \boldsymbol{\theta}_{\star} \quad \text{as } n \rightarrow \infty \quad \text{“Consistency”}$$

Risk of OLS

Therefore, we have the following final result for the excess risk of OLS

$$\mathbb{E}_{\varepsilon} \left[\mathcal{R}(\hat{\boldsymbol{\theta}}_{OLS}) \right] - \sigma^2 = \sigma^2 \frac{d}{n}$$

Remarks:

- Excess risk is proportional to the noise level $\mathbb{E}[\varepsilon^2] = \sigma^2$.
- Excess risk is proportional to the data dimension.
- To achieve excess risk $\Delta \mathcal{R} < \delta$, need:

$$n > \frac{\sigma^2 d}{\delta}$$

samples.

Bias-variance decomposition

Generally, if we have a data generative model for the training data $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^{d+1} : i = 1, \dots, n\}$:

$$y_i = f_{\star}(\mathbf{x}) + \varepsilon_i = \text{signal} + \text{noise}$$

With $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2] = \sigma^2$

Bias-variance decomposition

Generally, if we have a data generative model for the training data $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^{d+1} : i = 1, \dots, n\}$:

$$y_i = f_{\star}(\mathbf{x}) + \varepsilon_i = \text{signal} + \text{noise}$$

With $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2] = \sigma^2$, we can decompose the excess risk:

$$\mathbb{E}_{\varepsilon}[\mathcal{R}(\hat{\boldsymbol{\theta}})] - \sigma^2 = \mathbb{E} \left[(f_{\star}(\mathbf{x}) - f_{\hat{\theta}}(\mathbf{x}))^2 \right]$$

Bias-variance decomposition

Generally, if we have a data generative model for the training data $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^{d+1} : i = 1, \dots, n\}$:

$$y_i = f_{\star}(\mathbf{x}) + \varepsilon_i = \text{signal} + \text{noise}$$

With $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2] = \sigma^2$, we can decompose the excess risk:

$$\begin{aligned}\mathbb{E}_{\varepsilon}[\mathcal{R}(\hat{\boldsymbol{\theta}})] - \sigma^2 &= \mathbb{E} \left[(f_{\star}(\mathbf{x}) - f_{\hat{\theta}}(\mathbf{x}))^2 \right] \\ &= \mathbb{E} \left[(f_{\star}(\mathbf{x}) - \mathbb{E}_{\varepsilon}[f_{\hat{\theta}}(\mathbf{x})] + \mathbb{E}_{\varepsilon}[f_{\hat{\theta}}(\mathbf{x})] - f_{\hat{\theta}}(\mathbf{x}))^2 \right]\end{aligned}$$

Bias-variance decomposition

Generally, if we have a data generative model for the training data $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^{d+1} : i = 1, \dots, n\}$:

$$y_i = f_{\star}(\mathbf{x}) + \varepsilon_i = \text{signal} + \text{noise}$$

With $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2] = \sigma^2$, we can decompose the excess risk:

$$\begin{aligned}\mathbb{E}_{\varepsilon}[\mathcal{R}(\hat{\boldsymbol{\theta}})] - \sigma^2 &= \mathbb{E} \left[(f_{\star}(\mathbf{x}) - f_{\hat{\theta}}(\mathbf{x}))^2 \right] \\ &= \mathbb{E} \left[(f_{\star}(\mathbf{x}) - \mathbb{E}_{\varepsilon}[f_{\hat{\theta}}(\mathbf{x})] + \mathbb{E}_{\varepsilon}[f_{\hat{\theta}}(\mathbf{x})] - f_{\hat{\theta}}(\mathbf{x}))^2 \right] \\ &= \mathbb{E} \left[(f_{\star}(\mathbf{x}) - \mathbb{E}_{\varepsilon}[f_{\hat{\theta}}(\mathbf{x})])^2 \right] + \mathbb{E} \left[(\mathbb{E}_{\varepsilon}[f_{\hat{\theta}}(\mathbf{x})] - f_{\hat{\theta}}(\mathbf{x}))^2 \right]\end{aligned}$$

Bias-variance decomposition

Generally, if we have a data generative model for the training data $\mathcal{D} = \{(\mathbf{x}_i, y_i) \in \mathbb{R}^{d+1} : i = 1, \dots, n\}$:

$$y_i = f_{\star}(\mathbf{x}) + \varepsilon_i = \text{signal} + \text{noise}$$

With $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2] = \sigma^2$, we can decompose the excess risk:

$$\begin{aligned}\mathbb{E}_{\varepsilon}[\mathcal{R}(\hat{\boldsymbol{\theta}})] - \sigma^2 &= \mathbb{E} \left[(f_{\star}(\mathbf{x}) - f_{\hat{\theta}}(\mathbf{x}))^2 \right] \\ &= \mathbb{E} \left[(f_{\star}(\mathbf{x}) - \mathbb{E}_{\varepsilon}[f_{\hat{\theta}}(\mathbf{x})] + \mathbb{E}_{\varepsilon}[f_{\hat{\theta}}(\mathbf{x})] - f_{\hat{\theta}}(\mathbf{x}))^2 \right] \\ &= \mathbb{E} \left[(f_{\star}(\mathbf{x}) - \mathbb{E}_{\varepsilon}[f_{\hat{\theta}}(\mathbf{x})])^2 \right] + \mathbb{E} \left[(\mathbb{E}_{\varepsilon}[f_{\hat{\theta}}(\mathbf{x})] - f_{\hat{\theta}}(\mathbf{x}))^2 \right] \\ &= \text{bias}^2 + \text{variance}\end{aligned}$$

Bias-variance decomposition

$$\mathbb{E}_{\boldsymbol{\varepsilon}}[\mathcal{R}(\hat{\boldsymbol{\theta}})] - \sigma^2 = \mathcal{B} + \mathcal{V}$$

$$\mathcal{B} = \mathbb{E} \left[(f_{\star}(\mathbf{x}) - \mathbb{E}_{\boldsymbol{\varepsilon}}[f_{\hat{\boldsymbol{\theta}}}(\mathbf{x})])^2 \right]$$

$$\mathcal{V} = \mathbb{E} \left[(\mathbb{E}_{\boldsymbol{\varepsilon}}[f_{\hat{\boldsymbol{\theta}}}(\mathbf{x})] - f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}))^2 \right]$$

Bias-variance decomposition

$$\mathbb{E}_{\epsilon}[\mathcal{R}(\hat{\theta})] - \sigma^2 = \mathcal{B} + \mathcal{V}$$
$$\mathcal{B} = \mathbb{E} \left[(f_{\star}(\mathbf{x}) - \mathbb{E}_{\epsilon}[f_{\hat{\theta}}(\mathbf{x})])^2 \right]$$
$$\mathcal{V} = \mathbb{E} \left[(\mathbb{E}_{\epsilon}[f_{\hat{\theta}}(\mathbf{x})] - f_{\hat{\theta}}(\mathbf{x}))^2 \right]$$



Recall the the **approximation + estimation decomposition** from lecture 3:

$$\mathcal{R}(\theta) - \mathcal{R}_{\star} = \left(\mathcal{R}(\theta) - \inf_{\theta' \in \Theta} \mathcal{R}(\theta') \right) + \left(\inf_{\theta' \in \Theta} \mathcal{R}(\theta') - \mathcal{R}_{\star} \right)$$

Bias-variance decomposition

$$\mathbb{E}_{\varepsilon}[\mathcal{R}(\hat{\theta})] - \sigma^2 = \mathcal{B} + \mathcal{V}$$
$$\mathcal{B} = \mathbb{E} \left[(f_{\star}(\mathbf{x}) - \mathbb{E}_{\varepsilon}[f_{\hat{\theta}}(\mathbf{x})])^2 \right]$$
$$\mathcal{V} = \mathbb{E} \left[(\mathbb{E}_{\varepsilon}[f_{\hat{\theta}}(\mathbf{x})] - f_{\hat{\theta}}(\mathbf{x}))^2 \right]$$



Recall the the **approximation + estimation decomposition** from lecture 3:

$$\mathcal{R}(\theta) - \mathcal{R}_{\star} = \left(\mathcal{R}(\theta) - \inf_{\theta' \in \Theta} \mathcal{R}(\theta') \right) + \left(\inf_{\theta' \in \Theta} \mathcal{R}(\theta') - \mathcal{R}_{\star} \right)$$

For the OLS setting from before ($\text{rank}(X) = d < n$):

$$\mathbb{E}[f_{\hat{\theta}}(\mathbf{x})] = \langle \boldsymbol{\theta}_{\star}, \mathbf{x} \rangle = f_{\star}(\mathbf{x}) \quad \Rightarrow \quad \mathcal{B} = 0 \quad \mathcal{V} = \sigma^2 \frac{d}{n}$$

Marvels and pitfalls of OLS

To summarise, the OLS estimator $\hat{\theta}_{\text{OLS}}(\mathbf{X}, \mathbf{y}) = \mathbf{X}^+ \mathbf{y}$:

- Can only fit **linear functions**.
- For $n > d$, **has low bias** $\mathcal{B} = 0$
- When, $n \gg d$, has **low variance** $\mathcal{V} = \sigma^2 \frac{d}{n}$

Marvels and pitfalls of OLS

To summarise, the OLS estimator $\hat{\theta}_{\text{OLS}}(X, y) = X^+y$:

- Can only fit **linear functions**.
- For $n > d$, **has low bias** $\mathcal{B} = 0$
- When, $n \gg d$, has **low variance** $\mathcal{V} = \sigma^2 \frac{d}{n}$

But what about $n \approx d$? Consider for instance $n = d$.

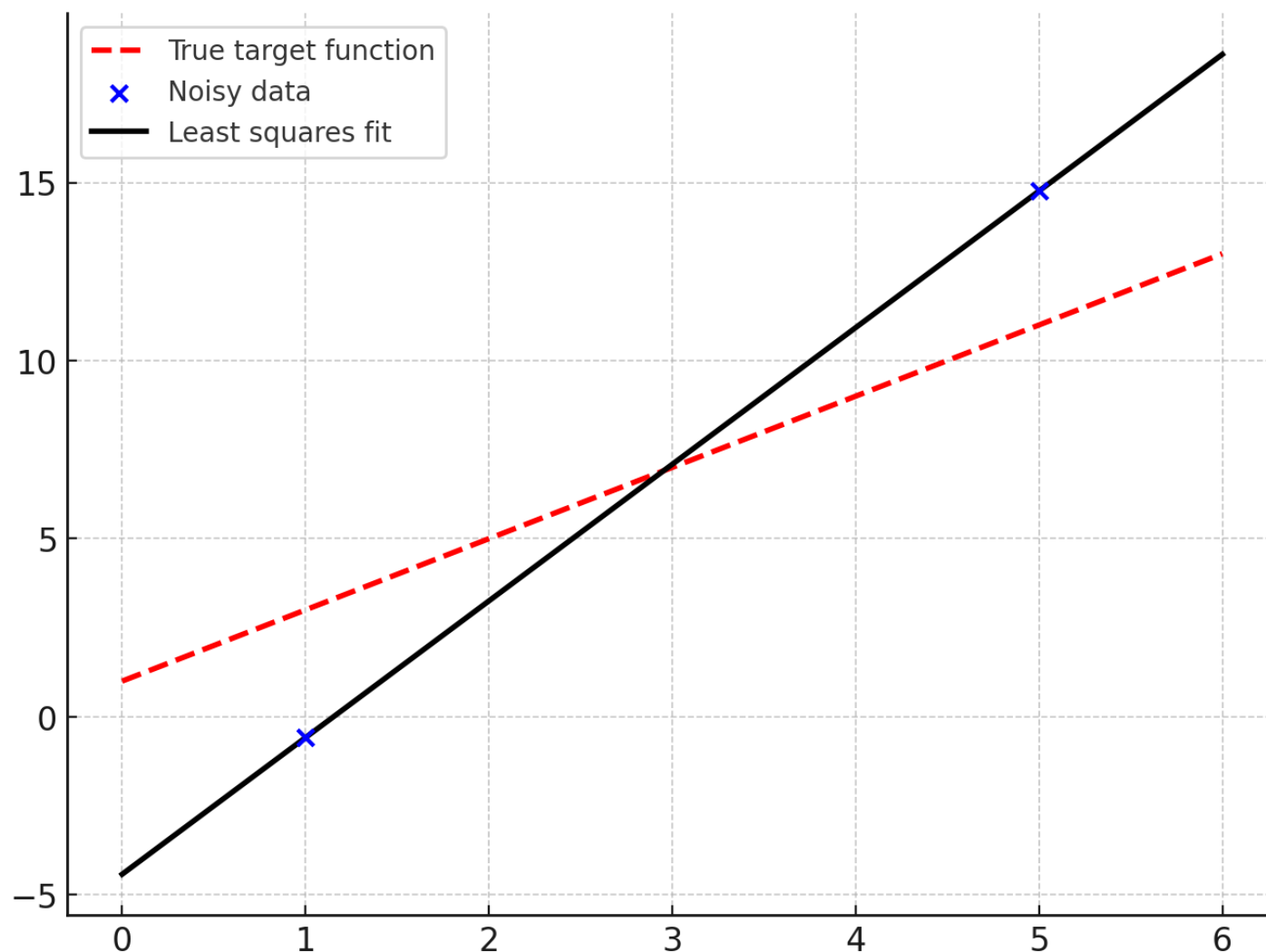
$X \in \mathbb{R}^{d \times d}$ is invertible $\Rightarrow y = X\hat{\theta}_{\text{OLS}}$ **interpolates the training data.**

Marvels and pitfalls of OLS

But what about $n \approx d$? Consider for instance $n = d$.

$$X \in \mathbb{R}^{d \times d} \text{ is invertible} \Rightarrow y = X\hat{\theta}_{\text{OLS}}$$

interpolates the
training data.



$$\mathbb{E}_{\epsilon}[\mathcal{R}(\hat{\theta}_{\text{OLS}})] = 2\sigma^2$$

$$\hat{\mathcal{R}}_n(\hat{\theta}_{\text{OLS}}) = 0$$



The test error above is
valid for the fixed design.

Marvels and pitfalls of OLS

Recall that:

$$\hat{\theta}_{OLS}(X, y) = \theta_{\star} + \frac{1}{n} \hat{\Sigma}_n^{-1} X^{\top} \epsilon$$

Marvels and pitfalls of OLS

Recall that:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{OLS}(X, y) &= \boldsymbol{\theta}_{\star} + \frac{1}{n} \hat{\boldsymbol{\Sigma}}_n^{-1} X^{\top} \boldsymbol{\varepsilon} \\ &= \boldsymbol{\theta}_{\star} + \sum_{j=1}^d \frac{1}{\sigma_j} \langle \boldsymbol{u}_j, \boldsymbol{\varepsilon} \rangle \boldsymbol{v}_j\end{aligned}$$

Marvels and pitfalls of OLS

Recall that:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{OLS}(X, y) &= \boldsymbol{\theta}_{\star} + \frac{1}{n} \hat{\boldsymbol{\Sigma}}_n^{-1} X^{\top} \boldsymbol{\varepsilon} \\ &= \boldsymbol{\theta}_{\star} + \sum_{j=1}^d \frac{1}{\sigma_j} \langle \mathbf{u}_j, \boldsymbol{\varepsilon} \rangle \mathbf{v}_j\end{aligned}$$

- Hence:
- **signal** is stronger in directions with larger s.v.
 - **noise** dominates directions with smaller s.v.

OLS has larger variance for data with small “**effective dimension**”.

What to do?

Classical strategies to mitigate variance:

- Dimensionality reduction: PCA, random projections (sketching), etc.
- Variable subset selection: Stepwise selection, best Subset Selection, etc.
- Regularisation: ridge, LASSO, etc.