Statistical Learning II

Lecture 5 - Least squares (continued)

Bruno Loureiro
@ CSD, DI-ENS & CNRS

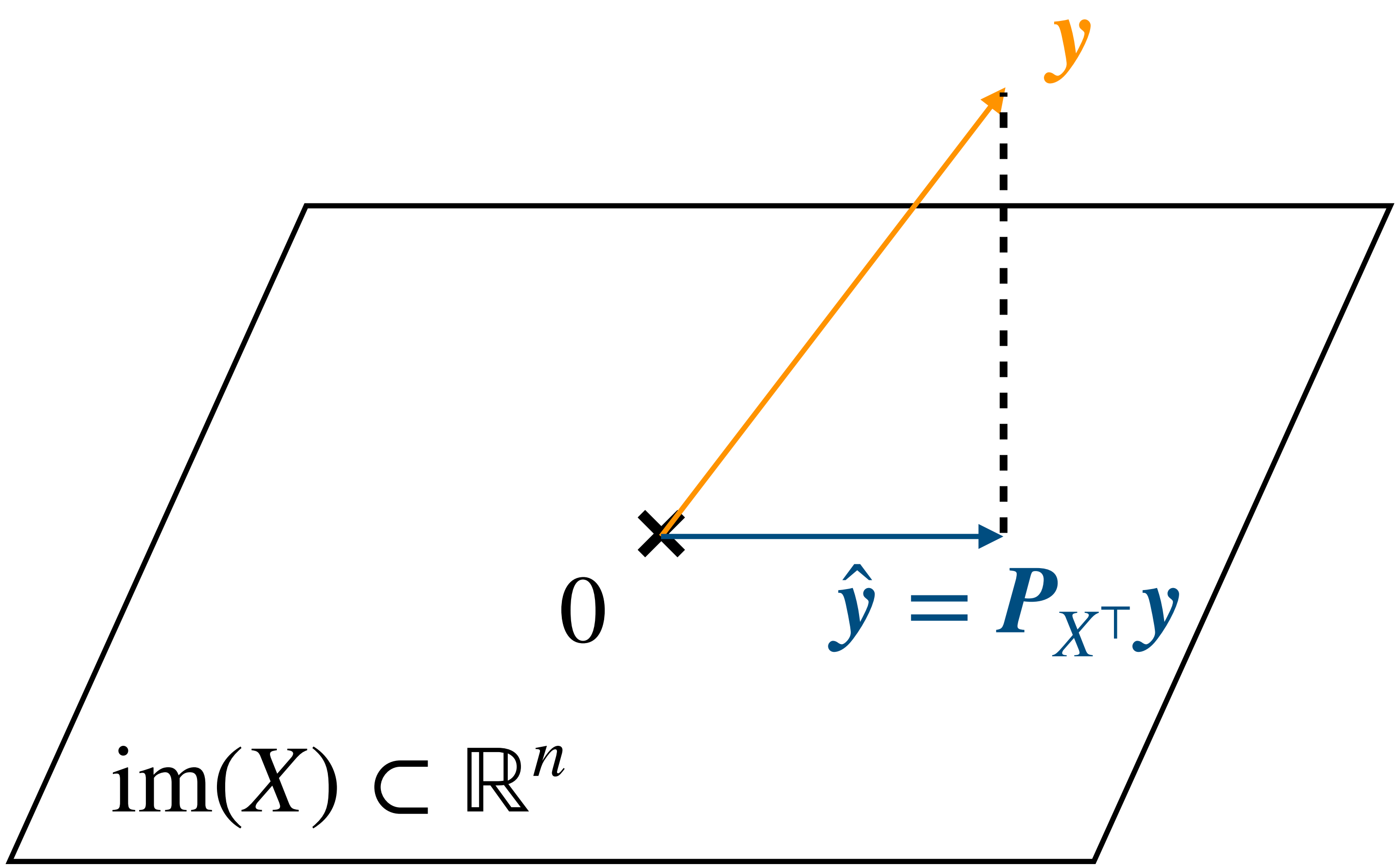
brloureiro@gmail.com

DL3 IASO, Université Paris Dauphine-PSL
09.10.2024

Geometrical interpretation

This gives a natural interpretation of the OLS predictor as an orthogonal projection of the labels in the row space of $\boldsymbol{X}$:

$$\hat{\boldsymbol{\theta}}_{OLS} = \boldsymbol{X}^{+}\boldsymbol{y} \quad \Rightarrow \quad \hat{\boldsymbol{y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\theta}}_{OLS} = \boldsymbol{X}\boldsymbol{X}^{+}\boldsymbol{y}$$



$$\min_{\boldsymbol{z} \in \mathrm{im}(\boldsymbol{X})} ||\boldsymbol{y} - \boldsymbol{z}||_2^2$$

Statistical analysis of OLS

Fixed-design analysis

Assumptions

We now assume the following data generative model:

$$y_i = \langle \boldsymbol{\theta}_\star, \boldsymbol{x}_i \rangle + \varepsilon_i$$

With:

- Fixed $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ and $\boldsymbol{x}_i \in \mathbb{R}^d$ "fixed design"
- $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$

Assumptions

We now assume the following data generative model:

$$y_i = \langle \boldsymbol{\theta}_\star, \boldsymbol{x}_i \rangle + \varepsilon_i$$

With:

- Fixed $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ and $\boldsymbol{x}_i \in \mathbb{R}^d$ "fixed design"
- $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$

<u>Remarks:</u>

- The Bayes predictor and error are given by

$$f_\star(\boldsymbol{x}) = \mathbb{E}[y \,|\, \boldsymbol{X} = \boldsymbol{x}] = \langle \boldsymbol{\theta}_\star, \boldsymbol{x} \rangle \qquad \mathscr{R}_\star = \mathscr{R}(\boldsymbol{\theta}_\star) = \sigma^2$$

Assumptions

We now assume the following data generative model:

$$y_i = \langle \boldsymbol{\theta}_\star, \boldsymbol{x}_i \rangle + \varepsilon_i$$

With:
- Fixed $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ and $\boldsymbol{x}_i \in \mathbb{R}^d$ "fixed design"
- $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$

Remarks:
- The Bayes predictor and error are given by

$$f_\star(\boldsymbol{x}) = \mathbb{E}[y \,|\, \boldsymbol{X} = \boldsymbol{x}] = \langle \boldsymbol{\theta}_\star, \boldsymbol{x} \rangle \qquad \mathscr{R}_\star = \mathscr{R}(\boldsymbol{\theta}_\star) = \sigma^2$$

- In particular

$$f_\star \in \mathscr{H} = \{ f(\boldsymbol{x}) = \langle \boldsymbol{\theta}, \boldsymbol{x} \rangle : \boldsymbol{\theta} \in \mathbb{R}^d \}$$

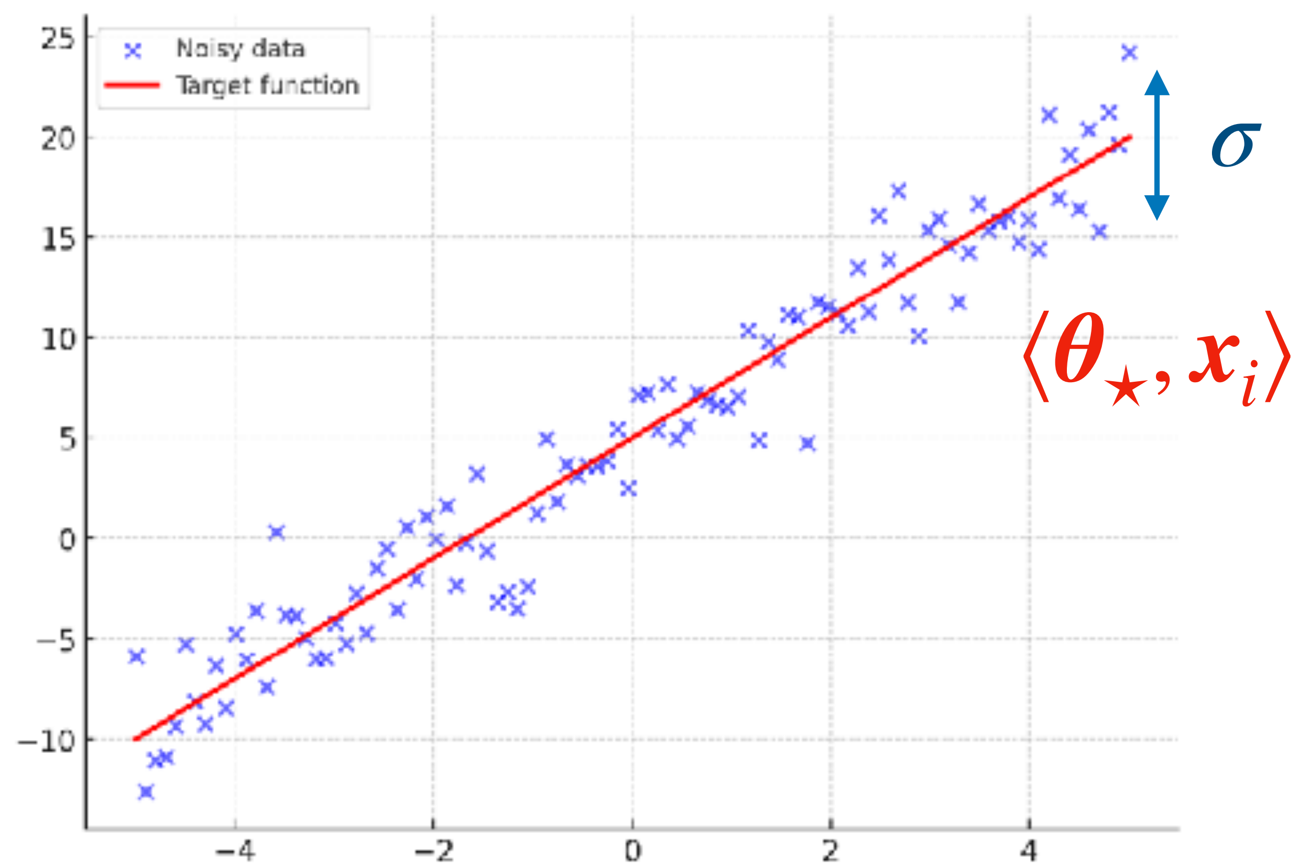
"Well-specified setting"

Assumptions

We now assume the following data generative model:

$$y_i = \langle \boldsymbol{\theta}_\star, \boldsymbol{x}_i \rangle + \varepsilon_i$$

With:
- Fixed $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ and $\boldsymbol{x}_i \in \mathbb{R}^d$ "fixed design"
- $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$



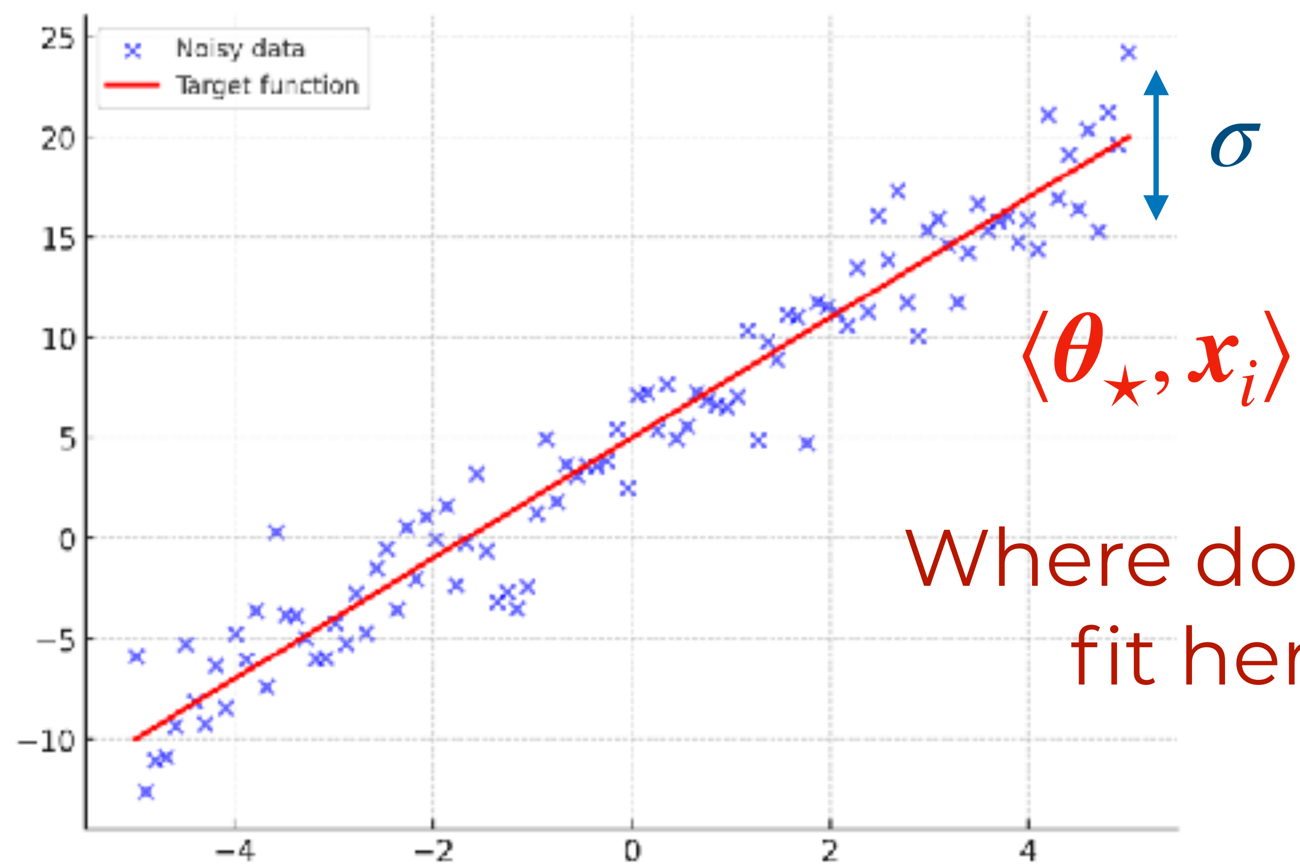
Assumptions

We now assume the following data generative model:

$$y_i = \langle \boldsymbol{\theta}_\star, \boldsymbol{x}_i \rangle + \varepsilon_i$$

With:

- Fixed $\boldsymbol{\theta}_\star \in \mathbb{R}^d$ and $\boldsymbol{x}_i \in \mathbb{R}^d$ "fixed design"
- $\mathbb{E}[\varepsilon_i] = 0$ and $\mathbb{E}[\varepsilon_i^2] = \sigma^2 < \infty$



Where does OLS fit here?

Decomposition of OLS

Given a batch of data sampled from this model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon} \in \mathbb{R}^n$$

Our goal is to understand the statistical properties of OLS.
For simplicity, assume that $\operatorname{rank}(\boldsymbol{X}) = d$ $(n > d)$:

$$\hat{\boldsymbol{\theta}}_{OLS}(\boldsymbol{X}, \boldsymbol{y}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Decomposition of OLS

Given a batch of data sampled from this model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon} \in \mathbb{R}^n$$

Our goal is to understand the statistical properties of OLS.
For simplicity, assume that $\mathrm{rank}(\boldsymbol{X}) = d \ (n > d)$:

$$\hat{\boldsymbol{\theta}}_{OLS}(\boldsymbol{X}, \boldsymbol{y}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$$

Decomposition of OLS

Given a batch of data sampled from this model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon} \in \mathbb{R}^n$$

Our goal is to understand the statistical properties of OLS.
For simplicity, assume that $\mathrm{rank}(\boldsymbol{X}) = d$ $(n > d)$:

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

Decomposition of OLS

Given a batch of data sampled from this model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon} \in \mathbb{R}^n$$

Our goal is to understand the statistical properties of OLS.
For simplicity, assume that $\mathrm{rank}(\boldsymbol{X}) = d$ $(n > d)$:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{OLS}(\boldsymbol{X}, \boldsymbol{y}) &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon}) \\ &= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta}_\star + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon}\end{aligned}$$

Decomposition of OLS

Given a batch of data sampled from this model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon} \in \mathbb{R}^n$$

Our goal is to understand the statistical properties of OLS.
For simplicity, assume that $\operatorname{rank}(\boldsymbol{X}) = d$ $(n > d)$:

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

Where we have defined $\hat{\boldsymbol{\Sigma}}_n = \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}^{d \times d}$ (Empirical covariance)

Decomposition of OLS

$$\hat{\boldsymbol{\theta}}_{OLS}(\boldsymbol{X}, \boldsymbol{y}) = \underbrace{\boldsymbol{\theta}_\star}_{\text{“signal”}} + \underbrace{\frac{1}{n}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}}_{\text{“noise”}}$$

Decomposition of OLS

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

"signal" "noise"

In particular:

- Bias: $\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\hat{\boldsymbol{\theta}}_{OLS}(\boldsymbol{X}, \boldsymbol{y})\right] = \boldsymbol{\theta}_\star$ "Unbiased"

Decomposition of OLS

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

"signal" "noise"

In particular:

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

Decomposition of OLS

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

"signal" "noise"

In particular:

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

Hence, informally:

$$\hat{\boldsymbol{\theta}}_{OLS} \to \boldsymbol{\theta}_\star \qquad \text{as } n \to \infty$$

"Consistency"

Risk of OLS

We now look at the population risk.

Risk of OLS

We now look at the population risk. First, note that for fixed $\boldsymbol{\theta} \in \mathbb{R}^d$:

$$\mathcal{R}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{y}}\left[\frac{1}{n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2\right]$$

Risk of OLS

We now look at the population risk. First, note that for fixed $\boldsymbol{\theta} \in \mathbb{R}^d$:

$$\mathscr{R}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{y}}\left[\frac{1}{n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2\right] = \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\frac{1}{n}||\boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{\theta}||_2^2\right]$$

Risk of OLS

We now look at the population risk. First, note that for fixed $\boldsymbol{\theta} \in \mathbb{R}^d$:

$$\mathscr{R}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{y}}\left[\frac{1}{n}||\boldsymbol{y} - \boldsymbol{X\theta}||_2^2\right] = \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\frac{1}{n}||\boldsymbol{X\theta}_\star + \boldsymbol{\varepsilon} - \boldsymbol{X\theta}||_2^2\right]$$

$$= (\boldsymbol{\theta}_\star - \boldsymbol{\theta})^\top \hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\theta}_\star - \boldsymbol{\theta}) + \sigma^2$$

Risk of OLS

We now look at the population risk. First, note that for fixed $\boldsymbol{\theta} \in \mathbb{R}^d$:

$$\mathscr{R}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{y}} \left[\frac{1}{n} ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 \right] = \mathbb{E}_{\boldsymbol{\varepsilon}} \left[\frac{1}{n} ||\boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{\theta}||_2^2 \right]$$

$$= (\boldsymbol{\theta}_\star - \boldsymbol{\theta})^\top \hat{\boldsymbol{\Sigma}}_n (\boldsymbol{\theta}_\star - \boldsymbol{\theta}) + \sigma^2$$

In the fixed design setting, $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is the same for training and testing.

Risk of OLS

We now look at the population risk. First, note that for fixed $\boldsymbol{\theta} \in \mathbb{R}^d$:

$$\mathscr{R}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{y}}\left[\frac{1}{n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2\right] = \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\frac{1}{n}||\boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{\theta}||_2^2\right]$$

$$= (\boldsymbol{\theta}_\star - \boldsymbol{\theta})^\top \hat{\boldsymbol{\Sigma}}_n (\boldsymbol{\theta}_\star - \boldsymbol{\theta}) + \sigma^2$$

In the fixed design setting, $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is the same for training and testing.

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

$$\mathscr{R}(\hat{\boldsymbol{\theta}}_{OLS}) - \sigma^2 = \frac{1}{n^2}\boldsymbol{\varepsilon}^\top \boldsymbol{X} \hat{\boldsymbol{\Sigma}}_n^{-1} \boldsymbol{X}^\top \boldsymbol{\varepsilon}$$

Risk of OLS

We now look at the population risk. First, note that for fixed $\boldsymbol{\theta} \in \mathbb{R}^d$:

$$\mathscr{R}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{y}}\left[\frac{1}{n}||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}||_2^2\right] = \mathbb{E}_{\boldsymbol{\varepsilon}}\left[\frac{1}{n}||\boldsymbol{X}\boldsymbol{\theta}_\star + \boldsymbol{\varepsilon} - \boldsymbol{X}\boldsymbol{\theta}||_2^2\right]$$

$$= (\boldsymbol{\theta}_\star - \boldsymbol{\theta})^\top \hat{\boldsymbol{\Sigma}}_n(\boldsymbol{\theta}_\star - \boldsymbol{\theta}) + \sigma^2$$

In the fixed design setting, $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ is the same for training and testing.

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

$$\mathscr{R}(\hat{\boldsymbol{\theta}}_{OLS}) - \sigma^2 = \frac{1}{n^2}\boldsymbol{\varepsilon}^\top\boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}$$

This is a random variable since $\hat{\boldsymbol{\theta}}_{OLS}$ is random!

Risk of OLS

First, note we can rewrite:

$$\mathscr{R}(\hat{\boldsymbol{\theta}}_{OLS}) - \sigma^2 = \frac{1}{n^2}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon}$$

Risk of OLS

First, note we can rewrite:

$$\mathscr{R}(\hat{\boldsymbol{\theta}}_{OLS}) - \sigma^2 = \frac{1}{n^2}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon} = \frac{1}{n^2}\mathrm{Tr}\left[\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon}\right]$$

Risk of OLS

First, note we can rewrite:

$$\mathscr{R}(\hat{\boldsymbol{\theta}}_{OLS}) - \sigma^2 = \frac{1}{n^2}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon} = \frac{1}{n^2}\mathrm{Tr}\left[\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top \boldsymbol{\varepsilon}\right]$$

$$= \frac{1}{n^2}\mathrm{Tr}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right]$$

Risk of OLS

First, note we can rewrite:

$$\mathscr{R}(\hat{\boldsymbol{\theta}}_{OLS}) - \sigma^2 = \frac{1}{n^2}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon} = \frac{1}{n^2}\mathrm{Tr}\left[\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right]$$

$$= \frac{1}{n^2}\mathrm{Tr}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right]$$

Now taking the expectation:

Risk of OLS

First, note we can rewrite:

$$\mathscr{R}(\hat{\boldsymbol{\theta}}_{OLS}) - \sigma^2 = \frac{1}{n^2}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon} = \frac{1}{n^2}\mathrm{Tr}\left[\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right]$$
$$= \frac{1}{n^2}\mathrm{Tr}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right]$$

Now taking the expectation:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\mathscr{R}(\hat{\boldsymbol{\theta}}_{OLS})\right] - \sigma^2 = \frac{1}{n^2}\mathrm{Tr}\left[\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right]\boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right]$$

Risk of OLS

First, note we can rewrite:

$$
\begin{aligned}
\mathscr{R}(\hat{\boldsymbol{\theta}}_{OLS}) - \sigma^2 &= \frac{1}{n^2}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon} = \frac{1}{n^2}\mathrm{Tr}\left[\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right] \\
&= \frac{1}{n^2}\mathrm{Tr}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right]
\end{aligned}
$$

Now taking the expectation:

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\mathscr{R}(\hat{\boldsymbol{\theta}}_{OLS})\right] - \sigma^2 &= \frac{1}{n^2}\mathrm{Tr}\left[\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right] \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right] \\
&= \frac{\sigma^2}{n^2}\mathrm{Tr}\left[\boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right]
\end{aligned}
$$

Risk of OLS

First, note we can rewrite:

$$\mathscr{R}(\hat{\boldsymbol{\theta}}_{OLS}) - \sigma^2 = \frac{1}{n^2}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon} = \frac{1}{n^2}\mathrm{Tr}\left[\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right]$$

$$= \frac{1}{n^2}\mathrm{Tr}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right]$$

Now taking the expectation:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\mathscr{R}(\hat{\boldsymbol{\theta}}_{OLS})\right] - \sigma^2 = \frac{1}{n^2}\mathrm{Tr}\left[\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right]\boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right]$$

$$= \frac{\sigma^2}{n^2}\mathrm{Tr}\left[\boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right] = \frac{\sigma^2}{n^2}\mathrm{Tr}\left[\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\right]$$

Risk of OLS

First, note we can rewrite:

$$
\begin{aligned}
\mathscr{R}(\hat{\boldsymbol{\theta}}_{OLS}) - \sigma^2 &= \frac{1}{n^2}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon} = \frac{1}{n^2}\mathrm{Tr}\left[\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right] \\
&= \frac{1}{n^2}\mathrm{Tr}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right]
\end{aligned}
$$

Now taking the expectation:

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\mathscr{R}(\hat{\boldsymbol{\theta}}_{OLS})\right] - \sigma^2 &= \frac{1}{n^2}\mathrm{Tr}\left[\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right]\boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right] \\
&= \frac{\sigma^2}{n^2}\mathrm{Tr}\left[\boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right] = \frac{\sigma^2}{n^2}\mathrm{Tr}\left[\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\right] \\
&= \frac{\sigma^2}{n}\mathrm{Tr}[I_d]
\end{aligned}
$$

Risk of OLS

First, note we can rewrite:

$$\mathcal{R}(\hat{\boldsymbol{\theta}}_{OLS}) - \sigma^2 = \frac{1}{n^2}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon} = \frac{1}{n^2}\mathrm{Tr}\left[\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\boldsymbol{\varepsilon}\right]$$

$$= \frac{1}{n^2}\mathrm{Tr}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top \boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right]$$

Now taking the expectation:

$$\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\mathcal{R}(\hat{\boldsymbol{\theta}}_{OLS})\right] - \sigma^2 = \frac{1}{n^2}\mathrm{Tr}\left[\mathbb{E}_{\boldsymbol{\varepsilon}}\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\right]\boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right]$$

$$= \frac{\sigma^2}{n^2}\mathrm{Tr}\left[\boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\boldsymbol{X}^\top\right] = \frac{\sigma^2}{n^2}\mathrm{Tr}\left[\boldsymbol{X}^\top\boldsymbol{X}\hat{\boldsymbol{\Sigma}}_n^{-1}\right]$$

$$= \frac{\sigma^2}{n}\mathrm{Tr}[I_d] = \sigma^2\frac{d}{n}$$

Risk of OLS

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

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

Risk of OLS

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

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

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 $\Delta\mathscr{R} < \delta$, need:

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

  samples.