



Statistical Learning II

Lecture 4 - Least squares

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

Summary of ERM

Let $\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, \dots, n\}$ denote training data sampled i.i.d. from p .

Given a choice of:

- Parametric hypothesis class $\mathcal{H} = \{f_\theta : \mathcal{X} \rightarrow \mathcal{Y} : \theta \in \Theta\}$
- Loss function $\ell : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$

Empirical Risk Minimisation consists of:

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_\theta(x_i))$$

Summary of ERM

Let $\mathcal{D} = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1, \dots, n\}$ denote training data sampled i.i.d. from p .

Given a choice of:

- Parametric hypothesis class $\mathcal{H} = \{f_\theta : \mathcal{X} \rightarrow \mathcal{Y} : \theta \in \Theta\}$
- Loss function $\ell : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R}_+$

Empirical Risk **Minimisation** consists of:

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_\theta(x_i))$$

Key questions

- What optimisation procedure to choose?

$$F(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_{\theta}(x_i))$$

Is typically a non-convex function of $\theta \in \Theta$.

Key questions

- What optimisation procedure to choose?

$$F(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_{\theta}(x_i))$$

Is typically a non-convex function of $\theta \in \Theta$.

- How large n needs to be (with respect to p, d) so that $\hat{\theta} \in \operatorname{argmin} F(\theta)$ has low training and/or test error?

Key questions

- What optimisation procedure to choose?

$$F(\theta) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f_{\theta}(x_i))$$

Is typically a non-convex function of $\theta \in \Theta$.

- How large n needs to be (with respect to p, d) so that $\hat{\theta} \in \operatorname{argmin} F(\theta)$ has low training and/or test error?
- What properties of the data distribution p makes the problem easier / harder?

Least-squares regression

Least-squares regression

Let $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} : i = 1, \dots, n\}$ denote the training data.

Ordinary least-squares (OLS) regression is defined as:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \hat{\mathcal{R}}_n(\boldsymbol{\theta}) := \frac{1}{2n} \sum_{i=1}^n (y_i - \langle \boldsymbol{\theta}, \mathbf{x}_i \rangle)^2$$

Least-squares regression

Let $\mathcal{D} = \{(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R} : i = 1, \dots, n\}$ denote the training data.

Ordinary least-squares (OLS) regression is defined as:

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \hat{\mathcal{R}}_n(\boldsymbol{\theta}) := \frac{1}{2n} ||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||_2^2$$

Where we have defined the data matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ and label vector $\mathbf{y} \in \mathbb{R}^n$:

$$\mathbf{X} = \begin{bmatrix} - & \mathbf{x}_1 & - \\ - & \mathbf{x}_2 & - \\ & \vdots & \\ - & \mathbf{x}_n & - \end{bmatrix} \in \mathbb{R}^{n \times d} \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Bayes risk for OLS

Remarks:

- This corresponds to an ERM problem on the class of linear functions:

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

with the square loss functions:

$$\ell(y, f_{\theta}(\mathbf{x})) = \frac{1}{2} (y - f_{\theta}(\mathbf{x}))^2$$

Bayes risk for OLS

Remarks:

- This corresponds to an ERM problem on the class of linear functions:

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

with the square loss functions:

$$\ell(y, f_{\theta}(\mathbf{x})) = \frac{1}{2} (y - f_{\theta}(\mathbf{x}))^2$$

- The Bayes predictor and risk are given by:

$$f_{\star}(\mathbf{x}) = \mathbb{E}[y | \mathbf{x}] \quad \mathcal{R}_{\star} = \mathbb{E} \left[\frac{1}{2} (y - \mathbb{E}[y | \mathbf{x}])^2 \right]$$



Exercise:
show this.

Intercept

Remarks:

- Without loss of generality, can add an intercept:

$$f_{\theta}(\mathbf{x}) = \langle \boldsymbol{\theta}, \mathbf{x} \rangle + b$$

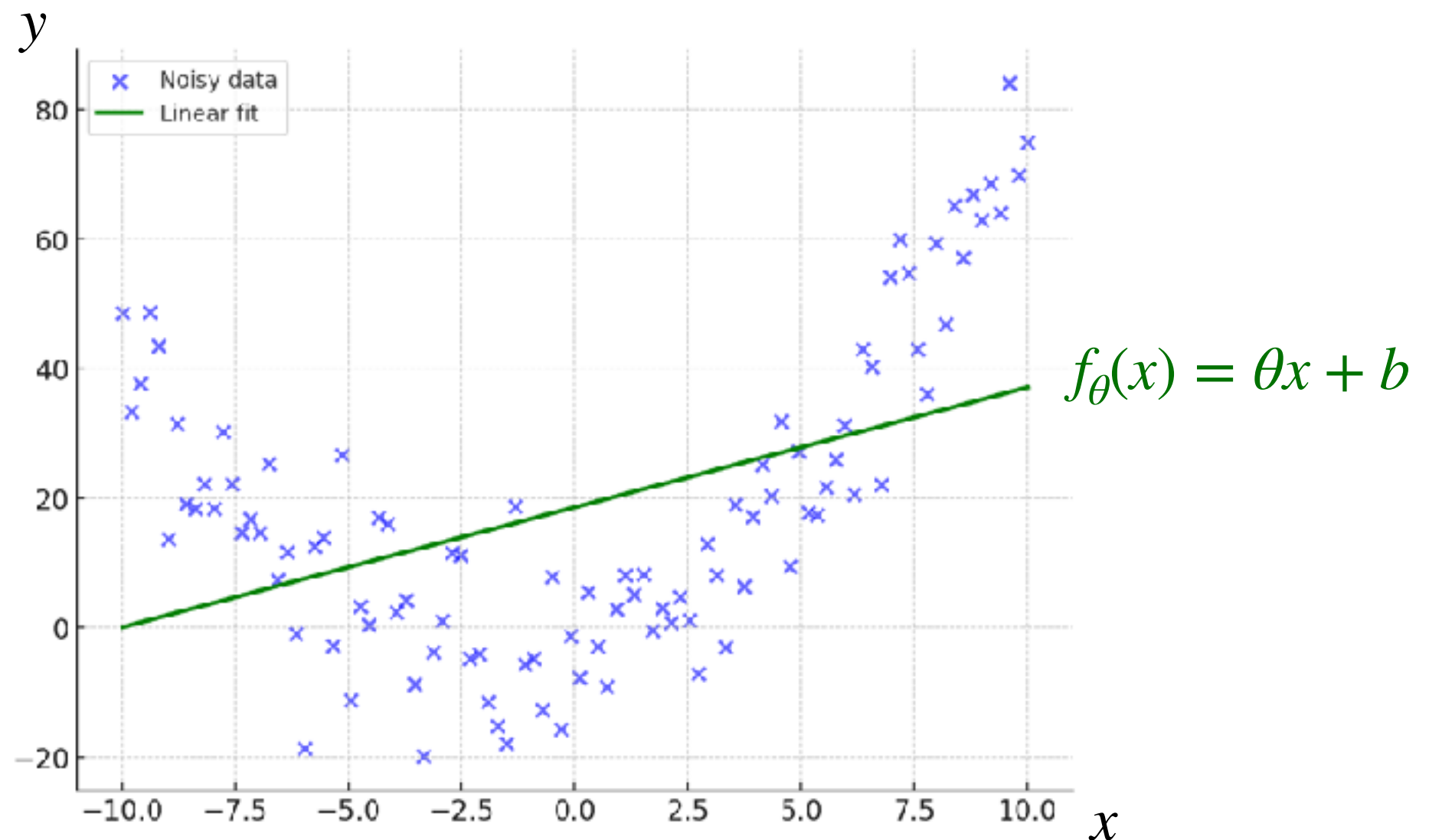
By redefining:

$$\tilde{\mathbf{X}} = \begin{bmatrix} - & \mathbf{x}_1 & - & 1 \\ - & \mathbf{x}_2 & - & 1 \\ & \vdots & & \\ - & \mathbf{x}_n & - & 1 \end{bmatrix} \in \mathbb{R}^{n \times (d+1)}$$

Inductive bias of OLS

Remarks:

- Inductive bias: can only fit affine functions of $\mathbf{x} \in \mathbb{R}^d$



Convexity of OLS

$$\hat{\mathcal{R}}_n(\boldsymbol{\theta}) := \frac{1}{2n} ||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||_2^2$$

- Gradient: $\nabla_{\boldsymbol{\theta}} \hat{\mathcal{R}}_n = -\frac{1}{n} \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \in \mathbb{R}^d$

Convexity of OLS

$$\hat{\mathcal{R}}_n(\boldsymbol{\theta}) := \frac{1}{2n} ||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||_2^2$$

- Gradient: $\nabla_{\boldsymbol{\theta}} \hat{\mathcal{R}}_n = -\frac{1}{n} \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \in \mathbb{R}^d$
- Hessian: $\nabla_{\boldsymbol{\theta}}^2 \hat{\mathcal{R}}_n = \frac{1}{n} \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{d \times d} \quad (:= \hat{\boldsymbol{\Sigma}}_n)$

Convexity of OLS

$$\hat{\mathcal{R}}_n(\boldsymbol{\theta}) := \frac{1}{2n} ||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||_2^2$$

- Gradient: $\nabla_{\boldsymbol{\theta}} \hat{\mathcal{R}}_n = -\frac{1}{n} \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \in \mathbb{R}^d$
- Hessian: $\nabla_{\boldsymbol{\theta}}^2 \hat{\mathcal{R}}_n = \frac{1}{n} \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{d \times d} \quad (:= \hat{\boldsymbol{\Sigma}}_n)$

Since $\mathbf{X}^\top \mathbf{X} \succeq 0$, $\hat{\mathcal{R}}_n$ is **convex** over \mathbb{R}^d . This implies that any minimum of $\hat{\mathcal{R}}_n$ is a global minimum.

Convexity of OLS

$$\hat{\mathcal{R}}_n(\boldsymbol{\theta}) := \frac{1}{2n} ||\mathbf{y} - \mathbf{X}\boldsymbol{\theta}||_2^2$$

- Gradient: $\nabla_{\boldsymbol{\theta}} \hat{\mathcal{R}}_n = -\frac{1}{n} \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \in \mathbb{R}^d$
- Hessian: $\nabla_{\boldsymbol{\theta}}^2 \hat{\mathcal{R}}_n = \frac{1}{n} \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{d \times d} \quad (:= \hat{\boldsymbol{\Sigma}}_n)$

Since $\mathbf{X}^\top \mathbf{X} \succeq 0$, $\hat{\mathcal{R}}_n$ is **convex** over \mathbb{R}^d . This implies that any minimum of $\hat{\mathcal{R}}_n$ is a global minimum.

For $n \geq d$, $\hat{\mathcal{R}}_n$ is **strictly convex** if and only if $\text{rank}(\mathbf{X}^\top \mathbf{X}) = d$. This implies that $\hat{\mathcal{R}}_n$ can have at most one global minimum.

Closed-form solution

- Gradient: $\nabla_{\boldsymbol{\theta}} \hat{\mathcal{R}}_n = -\frac{1}{n} \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \in \mathbb{R}^d$

If it exists, a minima must satisfy:

$$\nabla_{\boldsymbol{\theta}} \hat{\mathcal{R}}_n \stackrel{!}{=} 0$$

Closed-form solution

- Gradient: $\nabla_{\boldsymbol{\theta}} \hat{\mathcal{R}}_n = -\frac{1}{n} \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \in \mathbb{R}^d$

If it exists, a minima must satisfy:

$$\nabla_{\boldsymbol{\theta}} \hat{\mathcal{R}}_n \stackrel{!}{=} 0 \quad \Leftrightarrow \quad \mathbf{X}^\top \mathbf{X} \boldsymbol{\theta} = \mathbf{X}^\top \mathbf{y}$$

Closed-form solution

- Gradient: $\nabla_{\boldsymbol{\theta}} \hat{\mathcal{R}}_n = -\frac{1}{n} \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \in \mathbb{R}^d$

If it exists, a minima must satisfy:

$$\nabla_{\boldsymbol{\theta}} \hat{\mathcal{R}}_n \stackrel{!}{=} 0 \quad \Leftrightarrow \quad \mathbf{X}^\top \mathbf{X} \boldsymbol{\theta} = \mathbf{X}^\top \mathbf{y}$$

This is precisely the definition of the pseudo-inverse:

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

Closed-form solution

- Gradient: $\nabla_{\boldsymbol{\theta}} \hat{\mathcal{R}}_n = -\frac{1}{n} \mathbf{X}^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \in \mathbb{R}^d$

If it exists, a minima must satisfy:

$$\nabla_{\boldsymbol{\theta}} \hat{\mathcal{R}}_n \stackrel{!}{=} 0 \quad \Leftrightarrow \quad \mathbf{X}^\top \mathbf{X} \boldsymbol{\theta} = \mathbf{X}^\top \mathbf{y}$$

This is precisely the definition of the pseudo-inverse:

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

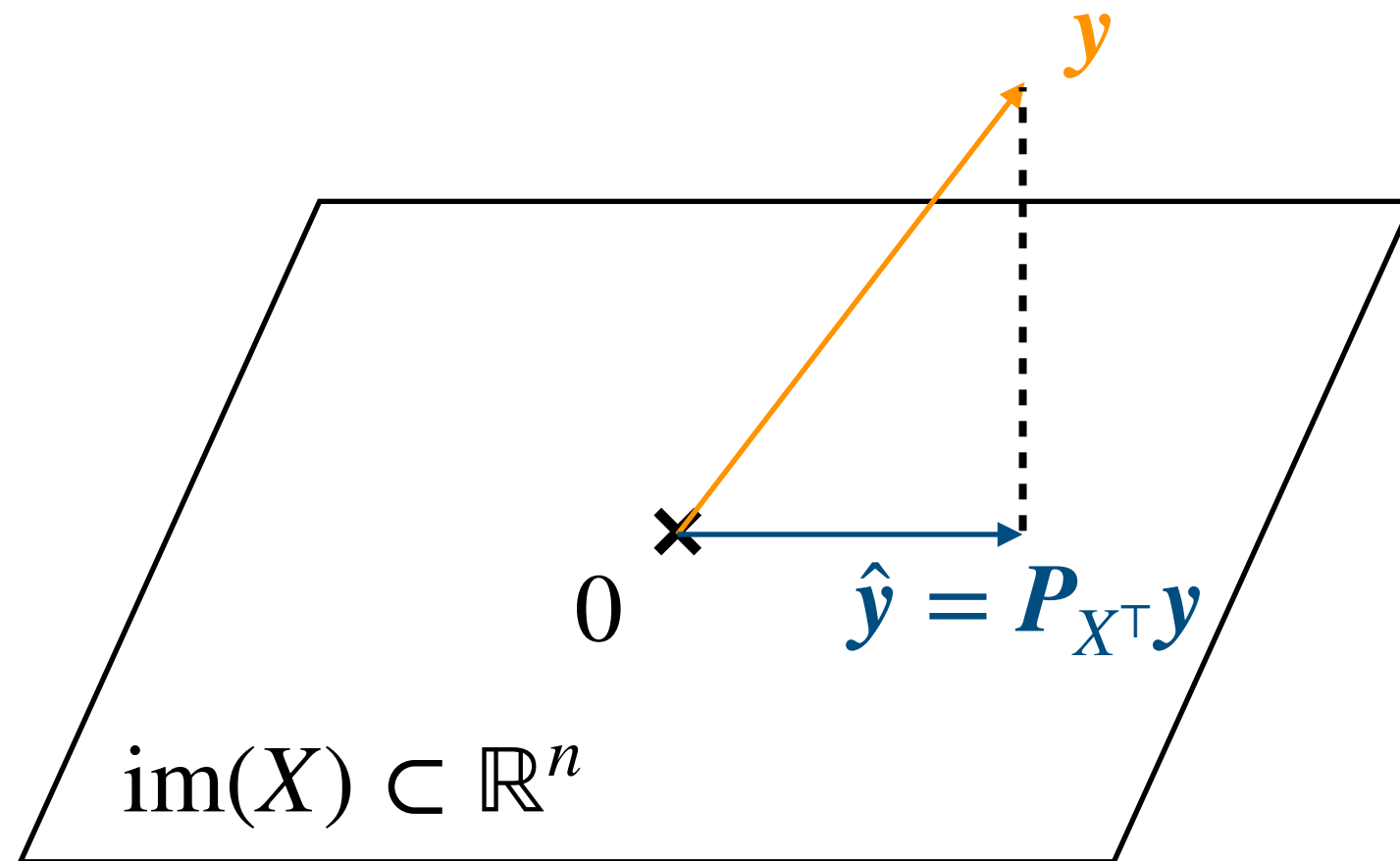
If $\text{rank}(\mathbf{X}) = \min(n, d)$:

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

Geometrical interpretation

This gives a natural interpretation of the OLS predictor as an orthogonal projection of the labels in the row space of X :

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

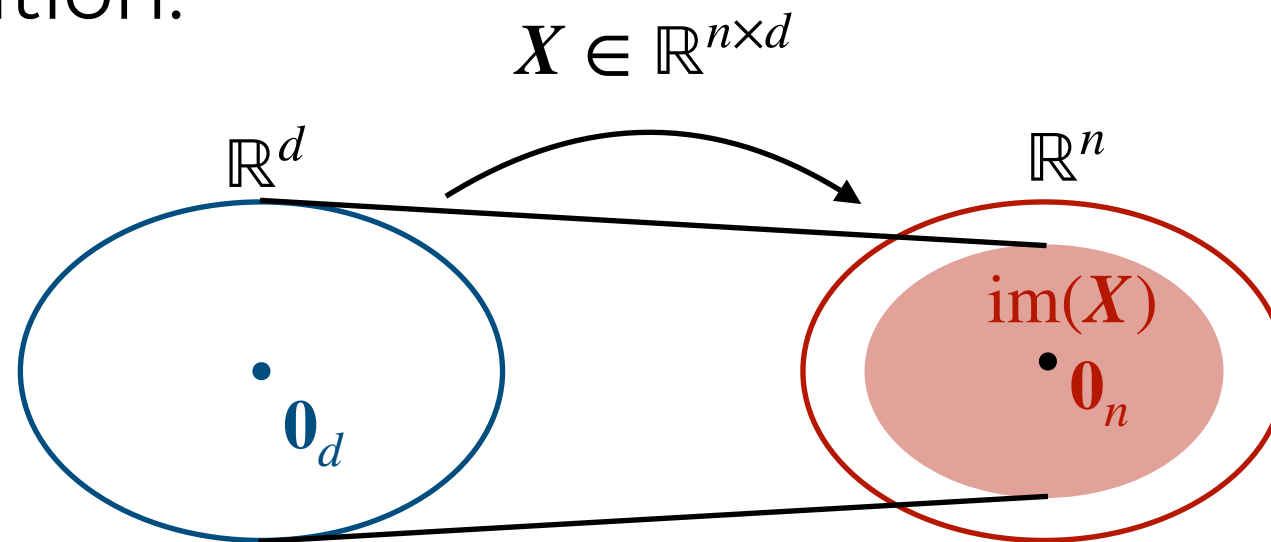


$$\min_{z \in \text{im}(X)} \|y - z\|_2^2$$

Two scenarios

From now on, let's **assume X is full-rank**.

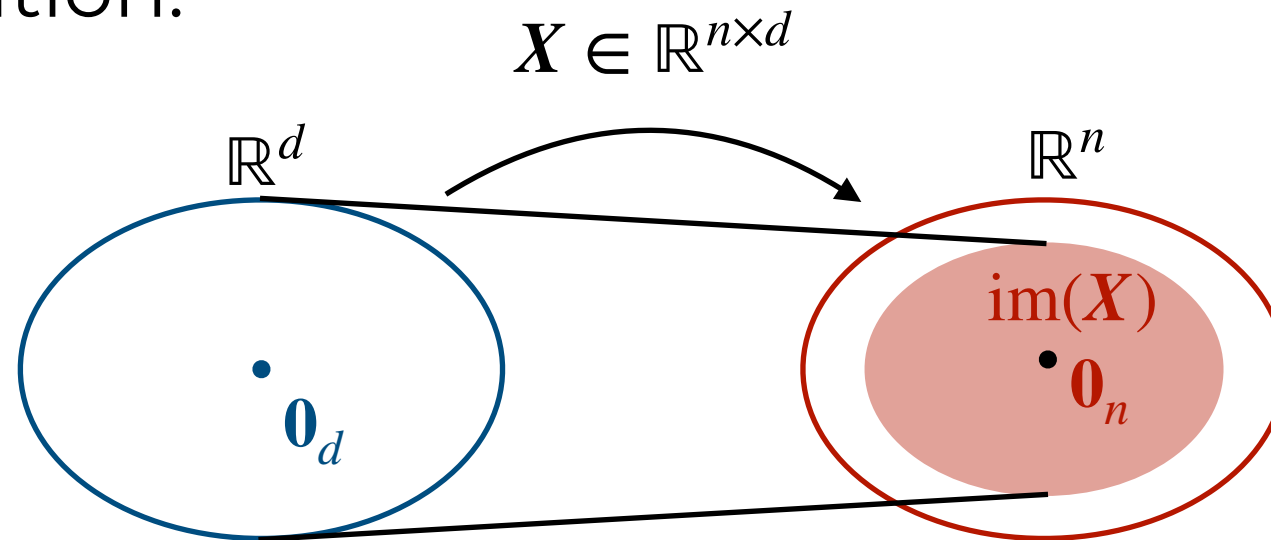
- For $n \geq d$: more equations than variables. $X\theta = y$ admits a unique solution.



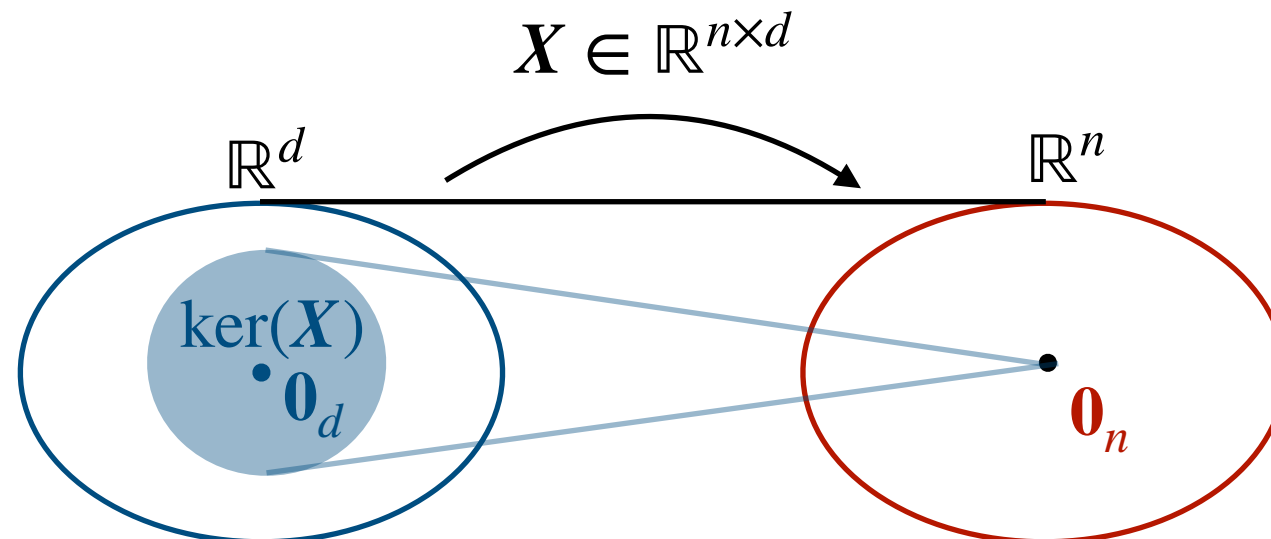
Two scenarios

From now on, let's **assume X is full-rank**.

- For $n \geq d$: more equations than variables. $X\theta = y$ admits a unique solution.



- For $n < d$: more variables than equations. $X\theta = y$ admits several solutions.



OLS as least norm solution

Assume $\text{rank}(\mathbf{X}) = n < d$. Then, OLS admits the following interpretation as the minimum ℓ_2 -norm solution:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{OLS} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} ||\boldsymbol{\theta}||_2 \\ \text{subject to} \quad \mathbf{X}\boldsymbol{\theta} = \mathbf{y} \end{aligned}$$

OLS as least norm solution

Assume $\text{rank}(\mathbf{X}) = n < d$. Then, OLS admits the following interpretation as the minimum ℓ_2 -norm solution:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{OLS} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} ||\boldsymbol{\theta}||_2 \\ \text{subject to} \quad \mathbf{X}\boldsymbol{\theta} = \mathbf{y} \end{aligned}$$

Proof: Let $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ denote a different solution from $\hat{\boldsymbol{\theta}}_{OLS}$.

OLS as least norm solution

Assume $\text{rank}(\mathbf{X}) = n < d$. Then, OLS admits the following interpretation as the minimum ℓ_2 -norm solution:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{OLS} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} ||\boldsymbol{\theta}||_2 \\ \text{subject to} \quad \mathbf{X}\boldsymbol{\theta} = \mathbf{y} \end{aligned}$$

Proof: Let $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ denote a different solution from $\hat{\boldsymbol{\theta}}_{OLS}$.

Then: $\langle \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}, \hat{\boldsymbol{\theta}}_{OLS} \rangle = \langle \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}, \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} \rangle$

OLS as least norm solution

Assume $\text{rank}(\mathbf{X}) = n < d$. Then, OLS admits the following interpretation as the minimum ℓ_2 -norm solution:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{OLS} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} ||\boldsymbol{\theta}||_2 \\ \text{subject to } \mathbf{X}\boldsymbol{\theta} = \mathbf{y} \end{aligned}$$

Proof: Let $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ denote a different solution from $\hat{\boldsymbol{\theta}}_{OLS}$.

$$\begin{aligned} \text{Then: } \langle \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}, \hat{\boldsymbol{\theta}}_{OLS} \rangle &= \langle \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}, \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} \rangle \\ &= \langle \mathbf{X}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}), (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} \rangle \end{aligned}$$

OLS as least norm solution

Assume $\text{rank}(\mathbf{X}) = n < d$. Then, OLS admits the following interpretation as the minimum ℓ_2 -norm solution:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{OLS} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} ||\boldsymbol{\theta}||_2 \\ \text{subject to } \mathbf{X}\boldsymbol{\theta} = \mathbf{y} \end{aligned}$$

Proof: Let $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ denote a different solution from $\hat{\boldsymbol{\theta}}_{OLS}$.

$$\begin{aligned} \text{Then: } \langle \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}, \hat{\boldsymbol{\theta}}_{OLS} \rangle &= \langle \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}, \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} \rangle \\ &= \langle \mathbf{X}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}), (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} \rangle \\ &= 0 \end{aligned}$$

OLS as least norm solution

Assume $\text{rank}(\mathbf{X}) = n < d$. Then, OLS admits the following interpretation as the minimum ℓ_2 -norm solution:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{OLS} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} ||\boldsymbol{\theta}||_2 \\ \text{subject to} \quad \mathbf{X}\boldsymbol{\theta} = \mathbf{y} \end{aligned}$$

Proof: Let $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ denote a different solution from $\hat{\boldsymbol{\theta}}_{OLS}$.

$$\begin{aligned} \text{Then: } \langle \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}, \hat{\boldsymbol{\theta}}_{OLS} \rangle &= \langle \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}, \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} \rangle \\ &= \langle \mathbf{X}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}), (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} \rangle \\ &= 0 \end{aligned}$$

Therefore $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS} \perp \hat{\boldsymbol{\theta}}_{OLS}$.

OLS as least norm solution

Assume $\text{rank}(\mathbf{X}) = n < d$. Then, OLS admits the following interpretation as the minimum ℓ_2 -norm solution:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{OLS} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} ||\boldsymbol{\theta}||_2 \\ \text{subject to } \mathbf{X}\boldsymbol{\theta} = \mathbf{y} \end{aligned}$$

Proof: Let $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ denote a different solution from $\hat{\boldsymbol{\theta}}_{OLS}$.

$$\begin{aligned} \text{Then: } \langle \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}, \hat{\boldsymbol{\theta}}_{OLS} \rangle &= \langle \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}, \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} \rangle \\ &= \langle \mathbf{X}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}), (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} \rangle \\ &= 0 \end{aligned}$$

Therefore $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS} \perp \hat{\boldsymbol{\theta}}_{OLS}$. Hence:

$$||\hat{\boldsymbol{\theta}}||_2^2 = ||\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS} + \hat{\boldsymbol{\theta}}_{OLS}||_2^2$$

OLS as least norm solution

Assume $\text{rank}(\mathbf{X}) = n < d$. Then, OLS admits the following interpretation as the minimum ℓ_2 -norm solution:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{OLS} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} ||\boldsymbol{\theta}||_2 \\ \text{subject to} \quad \mathbf{X}\boldsymbol{\theta} = \mathbf{y}\end{aligned}$$

Proof: Let $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ denote a different solution from $\hat{\boldsymbol{\theta}}_{OLS}$.

$$\begin{aligned}\text{Then: } \langle \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}, \hat{\boldsymbol{\theta}}_{OLS} \rangle &= \langle \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}, \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} \rangle \\ &= \langle \mathbf{X}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}), (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} \rangle \\ &= 0\end{aligned}$$

Therefore $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS} \perp \hat{\boldsymbol{\theta}}_{OLS}$. Hence:

$$||\hat{\boldsymbol{\theta}}||_2^2 = ||\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS} + \hat{\boldsymbol{\theta}}_{OLS}||_2^2 = ||\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}||_2^2 + ||\hat{\boldsymbol{\theta}}_{OLS}||_2^2$$

OLS as least norm solution

Assume $\text{rank}(\mathbf{X}) = n < d$. Then, OLS admits the following interpretation as the minimum ℓ_2 -norm solution:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{OLS} = \operatorname{argmin}_{\boldsymbol{\theta} \in \mathbb{R}^d} ||\boldsymbol{\theta}||_2 \\ \text{subject to } \mathbf{X}\boldsymbol{\theta} = \mathbf{y} \end{aligned}$$

Proof: Let $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$ denote a different solution from $\hat{\boldsymbol{\theta}}_{OLS}$.

$$\begin{aligned} \text{Then: } \langle \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}, \hat{\boldsymbol{\theta}}_{OLS} \rangle &= \langle \hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}, \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} \rangle \\ &= \langle \mathbf{X}(\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}), (\mathbf{X}\mathbf{X}^\top)^{-1} \mathbf{y} \rangle \\ &= 0 \end{aligned}$$

Therefore $\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS} \perp \hat{\boldsymbol{\theta}}_{OLS}$. Hence:

$$||\hat{\boldsymbol{\theta}}||_2^2 = ||\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS} + \hat{\boldsymbol{\theta}}_{OLS}||_2^2 = ||\hat{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{OLS}||_2^2 + ||\hat{\boldsymbol{\theta}}_{OLS}||_2^2 \geq ||\hat{\boldsymbol{\theta}}_{OLS}||_2^2$$