Statistical Learning II

Lecture 5 - Least squares (continued)

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

DL3 IASO, Université Paris Dauphine-PSL
08.10.2025

Partiel

Wednesday 22/10/2025

08:30 - 10:00

Least-squares regression

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

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

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

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

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

Convexity of OLS

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

- <u>Gradient:</u> $\nabla_{\boldsymbol{\theta}} \hat{\mathscr{R}}_n = -\frac{1}{n} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) \in \mathbb{R}^d$

- <u>Hessian:</u> $\nabla^2_{\boldsymbol{\theta}} \hat{\mathscr{R}}_n = \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X} \in \mathbb{R}^{d \times d}$ $(:= \hat{\boldsymbol{\Sigma}}_n)$

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

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

Explicit solution

- <u>Gradient:</u> $\nabla_{\boldsymbol{\theta}} \hat{\mathscr{R}}_n = -\frac{1}{n} \boldsymbol{X}^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\theta}) \in \mathbb{R}^d$

If it exists, a minima must satisfy:

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

Explicit solution

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

If it exists, a minima must satisfy:

$$\nabla_{\boldsymbol{\theta}} \hat{\mathscr{R}}_n \stackrel{!}{=} 0 \qquad \Leftrightarrow \qquad \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{y}$$

Explicit solution

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

If it exists, a minima must satisfy:

$$\nabla_{\boldsymbol{\theta}} \hat{\mathscr{R}}_n \overset{!}{=} 0 \qquad \Leftrightarrow \qquad \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{y}$$

If $X^\top X$ is invertible, unique solution:

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

Explicit solution

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

If it exists, a minima must satisfy:

$$\nabla_{\boldsymbol{\theta}} \hat{\mathscr{R}}_n \overset{!}{=} 0 \qquad \Leftrightarrow \qquad \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{y}$$

If $X^\top X$ is invertible, unique solution:

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

Note this is consistent with strict convexity of Hessian!

Explicit solution

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

If it exists, a minima must satisfy:

$$\nabla_{\boldsymbol{\theta}} \hat{\mathscr{R}}_n \overset{!}{=} 0 \qquad \Leftrightarrow \qquad \boldsymbol{X}^\top \boldsymbol{X\theta} = \boldsymbol{X}^\top \boldsymbol{y}$$

If $X^\top X$ is invertible, unique solution:

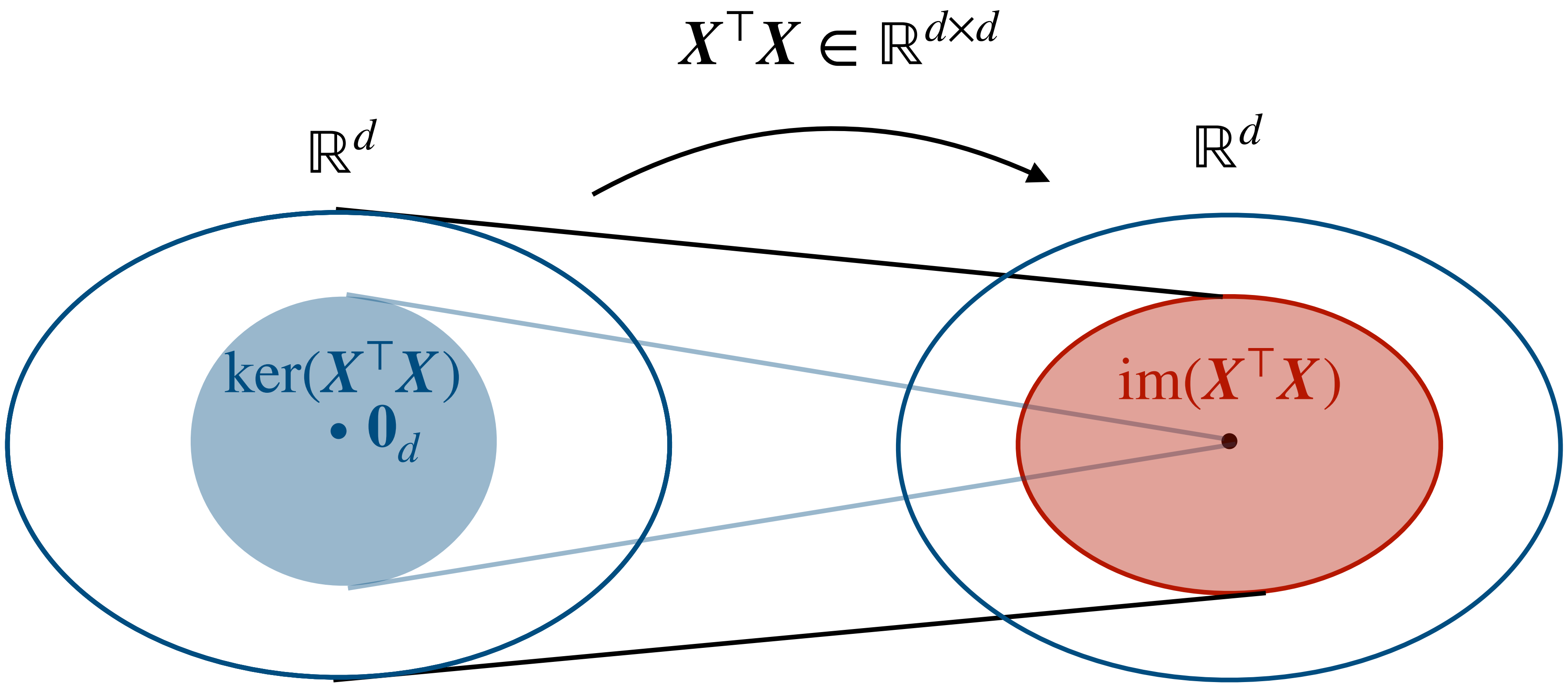
$$\boxed{\hat{\boldsymbol{\theta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}}$$

Note this is consistent with strict convexity of Hessian!

But what if $X^\top X$ is not invertible? For example, if $\mathrm{rank}(X) = n < d$?

Two scenarios

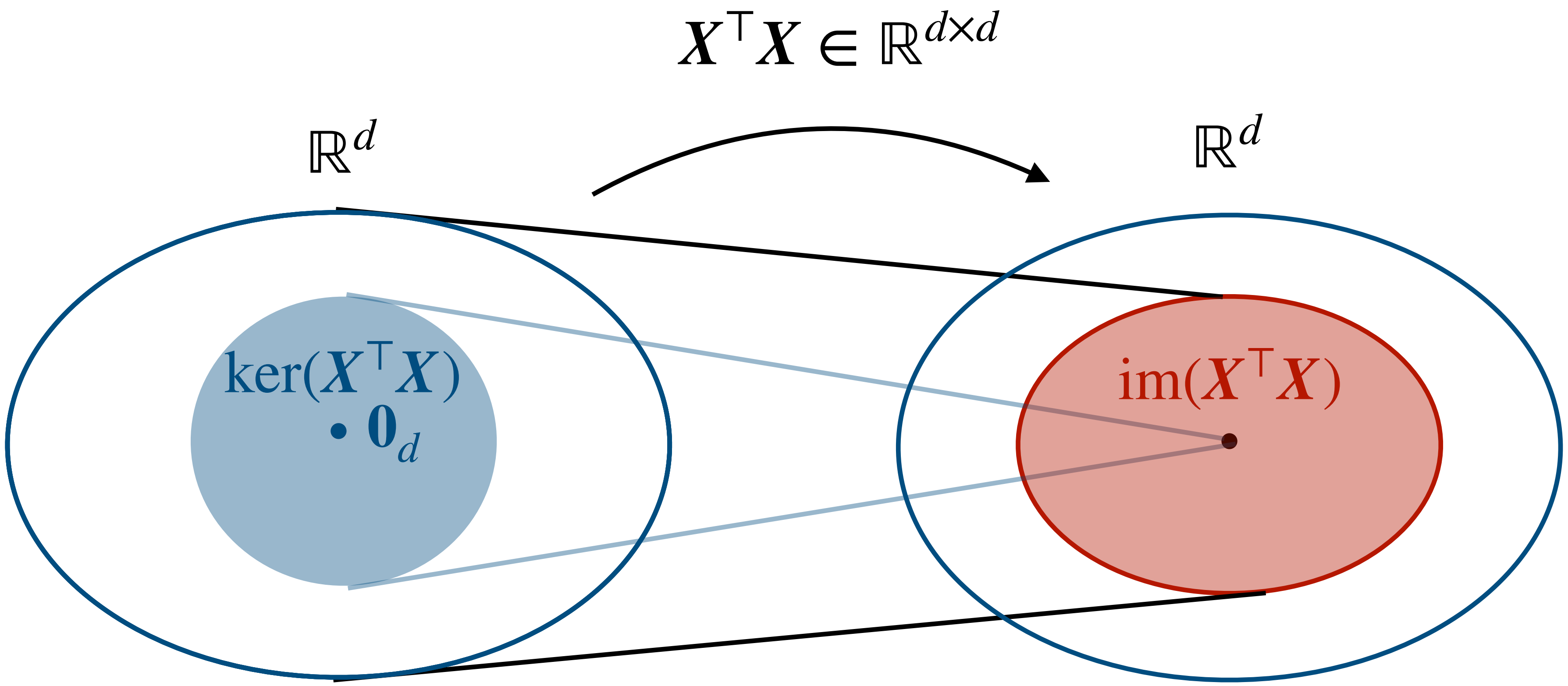
Focus on case $\mathrm{rank}(\boldsymbol{X}) = n < d$ (i.e. $\boldsymbol{X}$ is full-rank)



Note $\mathrm{rank}(\boldsymbol{X}) = \mathrm{rank}(\boldsymbol{X}^\top \boldsymbol{X}) = \mathrm{rank}(\boldsymbol{X}\boldsymbol{X}^\top)$

Two scenarios

Focus on case $\mathrm{rank}(\boldsymbol{X}) = n < d$ (i.e. $\boldsymbol{X}$ is full-rank)



Note $\mathrm{rank}(\boldsymbol{X}) = \mathrm{rank}(\boldsymbol{X}^\top \boldsymbol{X}) = \mathrm{rank}(\boldsymbol{X}\boldsymbol{X}^\top)$

All solutions of $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{y}$ can be written as:

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_0 + \boldsymbol{k}$$

Where: $\boldsymbol{k} \in \ker(\boldsymbol{X}^\top \boldsymbol{X}) \simeq \mathbb{R}^{d-n}$ and $\hat{\boldsymbol{\theta}}_0 \in \mathrm{im}(\boldsymbol{X}^\top \boldsymbol{X}) \simeq \mathbb{R}^n$

Explicit solution

For $\mathrm{rank}(\boldsymbol{X}) = n < d$, a particular solution of $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{y}$ is:

$$\hat{\boldsymbol{\theta}}_0 = \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top)^{-1} \boldsymbol{y}$$

(Check this!)

Explicit solution

For $\operatorname{rank}(\boldsymbol{X}) = n < d$, a particular solution of $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{y}$ is:

$$\hat{\boldsymbol{\theta}}_0 = \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top)^{-1} \boldsymbol{y}$$

(Check this!)

Together, in the full-rank case $\operatorname{rank}(\boldsymbol{X}) = \min(n, d)$ solution is:

$$\hat{\boldsymbol{\theta}} = \begin{cases} (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} & \text{for } n \geq d \\ \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top)^{-1} \boldsymbol{y} + \boldsymbol{k} & \text{for } n < d \end{cases}$$

For any $\boldsymbol{k} \in \ker(\boldsymbol{X}^\top \boldsymbol{X})$.

Explicit solution

For $\mathrm{rank}(\boldsymbol{X}) = n < d$, a particular solution of $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{X}^\top \boldsymbol{y}$ is:

$$\hat{\boldsymbol{\theta}}_0 = \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y}$$ (Check this!)

Together, in the full-rank case $\mathrm{rank}(\boldsymbol{X}) = \min(n, d)$ solution is:

$$\hat{\boldsymbol{\theta}} = \begin{cases} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y} & \text{for } n \geq d \\ \boldsymbol{X}^\top (\boldsymbol{X}\boldsymbol{X}^\top)^{-1}\boldsymbol{y} + \boldsymbol{k} & \text{for } n < d \end{cases}$$

For any $\boldsymbol{k} \in \ker(\boldsymbol{X}^\top \boldsymbol{X})$.

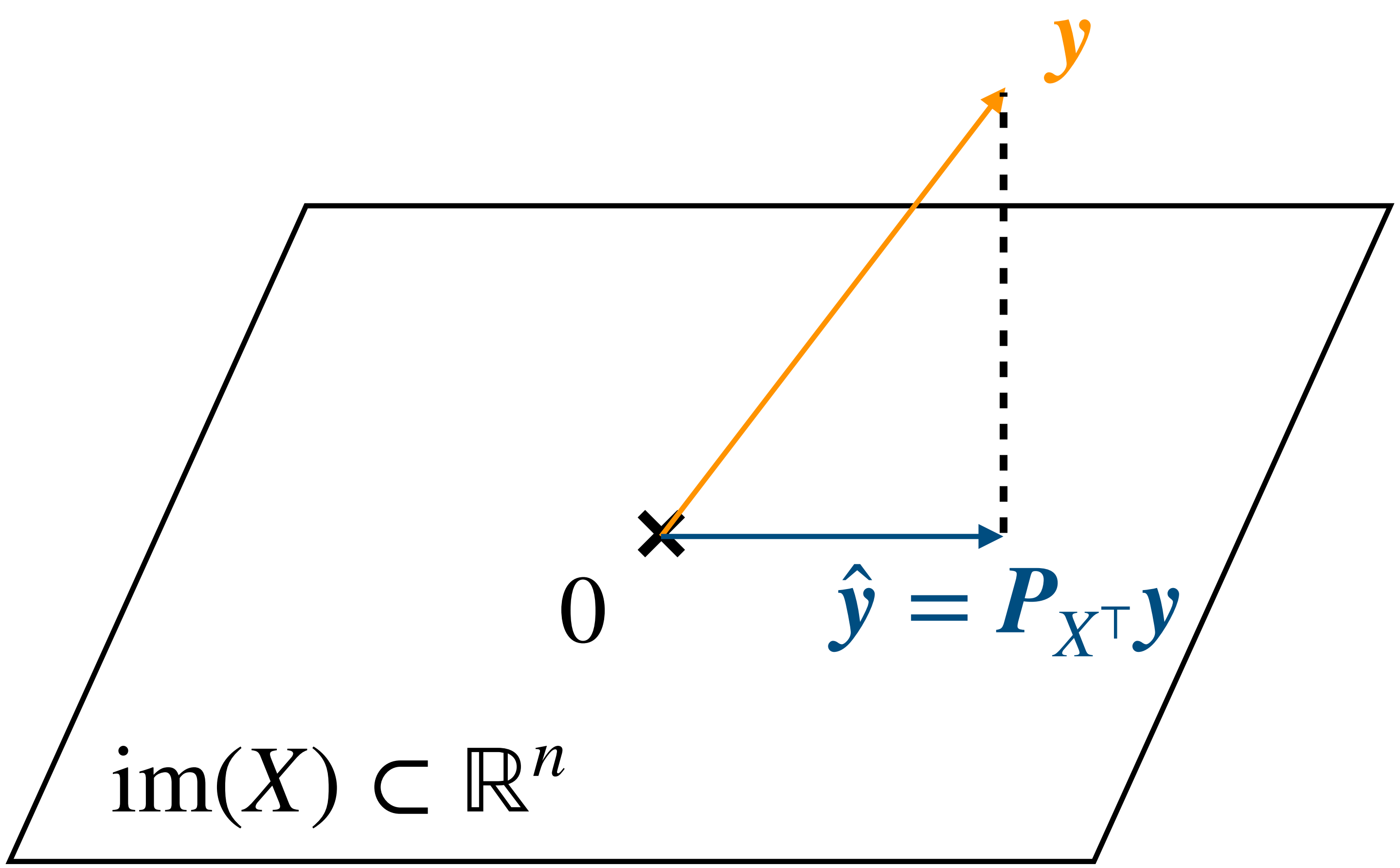
In particular, for $\boldsymbol{k} = \boldsymbol{0} \in \ker(\boldsymbol{X}^\top \boldsymbol{X})$ this is the Moore-Penrose inverse:

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

Geometrical interpretation

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

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



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

OLS as least norm solution

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

$$\hat{\boldsymbol{\theta}}_{OLS} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}} \; ||\boldsymbol{\theta}||_2$$

$$\text{subject to} \quad \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{y}$$

OLS as least norm solution

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

$$\hat{\boldsymbol{\theta}}_{OLS} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\mathrm{argmin}}\ ||\boldsymbol{\theta}||_2$$

subject to $\quad \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{y}$

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

OLS as least norm solution

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

$$\hat{\boldsymbol{\theta}}_{OLS} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\operatorname{argmin}} \; ||\boldsymbol{\theta}||_2$$

$$\text{subject to} \quad \boldsymbol{X\theta} = \boldsymbol{y}$$

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

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

OLS as least norm solution

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

$$\hat{\boldsymbol{\theta}}_{OLS} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\mathrm{argmin}} \ ||\boldsymbol{\theta}||_2$$

$$\text{subject to} \quad \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{y}$$

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

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

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

OLS as least norm solution

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

$$\hat{\boldsymbol{\theta}}_{OLS} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\mathrm{argmin}} \; ||\boldsymbol{\theta}||_2$$

$$\text{subject to} \quad \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{y}$$

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

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

OLS as least norm solution

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

$$\hat{\boldsymbol{\theta}}_{OLS} = \underset{\boldsymbol{\theta}\in\mathbb{R}^d}{\operatorname{argmin}} \; ||\boldsymbol{\theta}||_2$$

$$\text{subject to} \quad \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{y}$$

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

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

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

OLS as least norm solution

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

$$\hat{\boldsymbol{\theta}}_{OLS} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\mathrm{argmin}} \; ||\boldsymbol{\theta}||_2$$

$$\text{subject to} \quad \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{y}$$

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

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

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

$$\hat{\boldsymbol{\theta}}_{OLS} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\text{argmin}} \; ||\boldsymbol{\theta}||_2$$

subject to $\quad \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{y}$

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

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

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

$$\hat{\boldsymbol{\theta}}_{OLS} = \underset{\boldsymbol{\theta} \in \mathbb{R}^d}{\mathrm{argmin}} \; ||\boldsymbol{\theta}||_2$$

subject to $\quad \boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{y}$

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

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

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

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