Statistical Learning II

Lecture 4 - Least squares

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

DL3 IASO, Université Paris Dauphine-PSL
01.10.2025

Summary of ERM

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

Given a choice of:

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

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

Given a choice of:

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

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

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

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

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 $\boldsymbol{y} \in \mathbb{R}^n$:

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

Bayes risk for OLS

Remarks:

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

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

with the square loss functions:

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

Bayes risk for OLS

Remarks:

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

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

with the square loss functions:

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

- The Bayes predictor and risk are given by:

$$f_\star(\boldsymbol{x}) = \mathbb{E}[y\,|\,\boldsymbol{x}] \qquad \mathscr{R}_\star = \mathbb{E}\left[\frac{1}{2}(y - \mathbb{E}[y\,|\,\boldsymbol{x}])^2\right]$$

🧐 Exercise: show this.

Intercept

Remarks:

- Without loss of generality, can add an intercept:

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

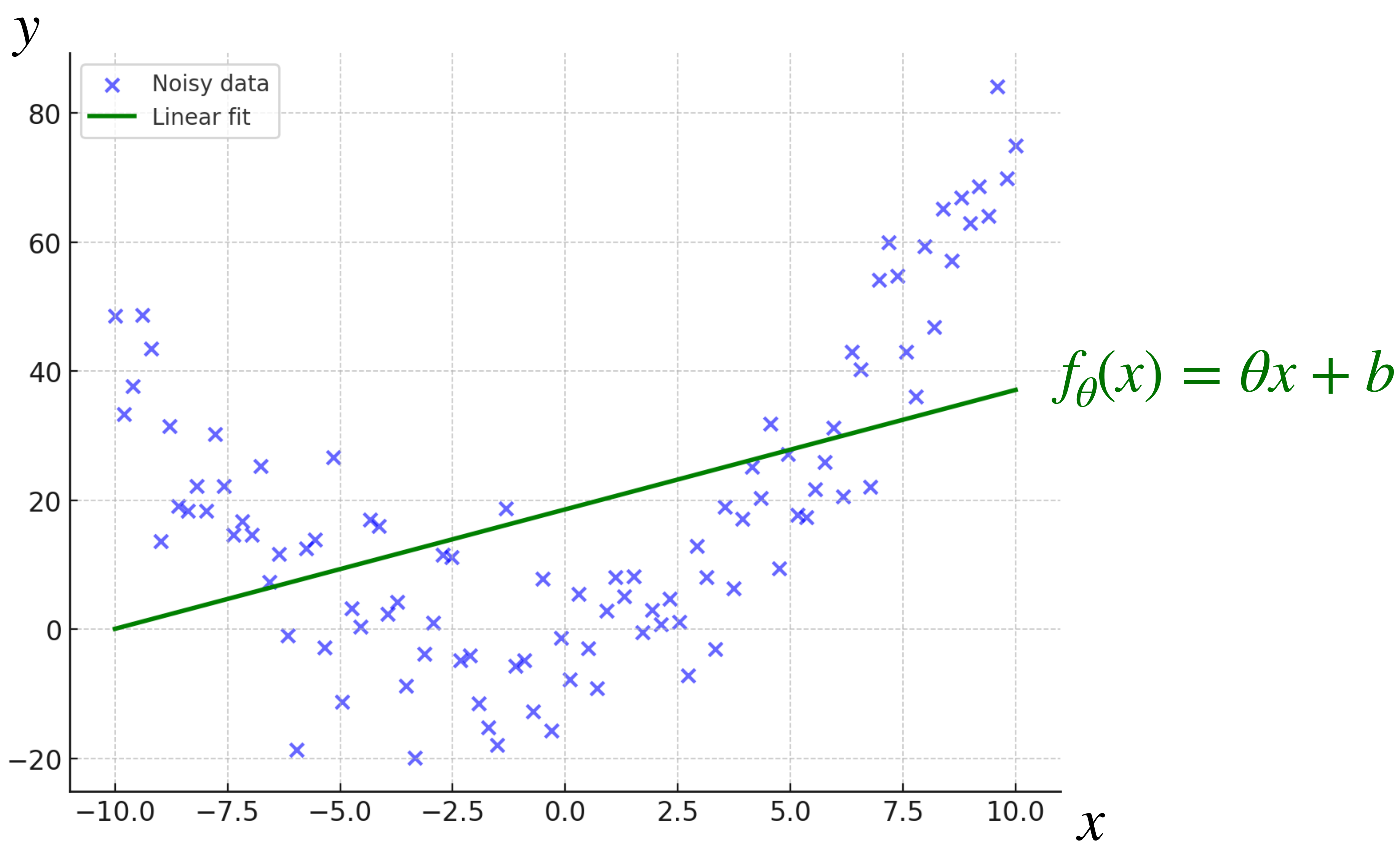
By redefining:

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

Inductive bias of OLS

Remarks:

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



Convexity of OLS

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

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

Convexity of OLS

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

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

Convexity of OLS

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

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

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

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

Convexity of OLS

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

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

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

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.