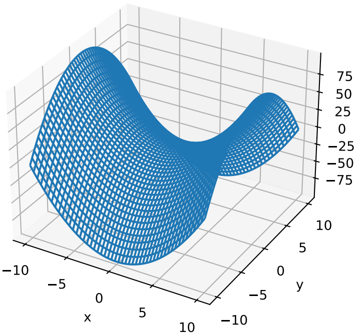
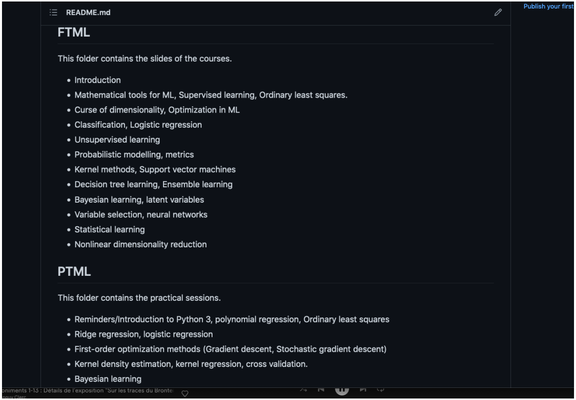
Fondamentaux théoriques du machine learning

Neither positive nor negative Hessian (saddle point)



`https://github.com/nlehir/FTML_PTML`

You have the planned overview of the course on the repo/



README.md

FTML

This folder contains the slides of the courses.

- Introduction
- Mathematical tools for ML, Supervised learning, Ordinary least squares.
- Curse of dimensionality, Optimization in ML
- Classification, Logistic regression
- Unsupervised learning
- Probabilistic modelling, metrics
- Kernel methods, Support vector machines
- Decision tree learning, Ensemble learning
- Bayesian learning, latent variables
- Variable selection, neural networks
- Statistical learning
- Nonlinear dimensionality reduction

PTML

This folder contains the practical sessions.

- Reminders/Introduction to Python 3, polynomial regression, Ordinary least squares
- Ridge regression, logistic regression
- First-order optimization methods (Gradient descent, Stochastic gradient descent)
- Kernel density estimation, kernel regression, cross validation.
- Bayesian learning

Some references have also been added to the repo.

FTML: References

Understanging machine learning : from theory to algorithms
[Shalev-Shwartz and Ben-David, 2013,]
https://www.cs.huji.ac.il/w~shais/UnderstandingMachineLearning/

Learning theory from first principles
[Bach, 2021,]
https://francisbach.com/i-am-writing-a-book/

Apprentissage artificiel : concepts et algorithmes
[Cornuéjols and Miclet, 2003,]
General reference on AI and ML.

Analyse numérique et optimisation : une introduction à la modélisation mathématique et à la simulation numérique
[Allaire, 2012,]
Chapters 9 and 10 are an introduction to optimization.

The elements of Statistical learning
[Hastie et al., 2009,]

RÉFÉRENCES

[Allaire, 2012] Allaire, G. (2012). Analyse numérique et optimisation Une introduc-

Overview of lecture 2

Mathematical tools for ML
- Linear algebra
- Statistics, probability theory
- Differential calculus

Supervised learning
- Excess risk
- Bayes predictor
- Bias-variance decomposition

Ordinary Least Squares I
- OLS estimator
- Statistical analysis of OLS

Objective

- The aim of the course if to give an introduction to **fundamental principles** in ML.
- To do so, we will need an adapted mathematical toolbox and a bag of important results.
- The first part of this lecture is dedicated to the presentation of this toolbox and to maths reminders.
- See also **FTML.pdf** on the repo.

Matricial calculus

In machine learning, optimization or statistics we often write the inner product of two vectors of $\mathbb{R}^d$ as a product of matrices. If $x \in \mathbb{R}^d$ writes :

$$x = \begin{pmatrix} x_1 \\ \dots \\ x_i \\ \dots \\ x_d \end{pmatrix}$$

And (with T denoting the transposition),

$$y^T = \big(y_1, \dots y_j, \dots, y_d\big)$$

Then we have that

$$\langle x, y \rangle = y^T x = x^T y$$

Moments of a distribution

Definition

Moments of a distribution

Let X be a real random variabe, and $k \in \mathbb{N}^*$. X is said to have a moment of order k if $E(|X|^k) < +\infty$, which means that :

- if X is discrete, with image $X(\Omega) = (x_i)_{i \in \mathbb{N}}$, the series

$$\sum (x_i)^k P(X = x_i)$$

  is **absolutely** convergent. The moment is then equal to the sum of that series (without absolute value).

Moments of a distribution

Definition

Moments of a distribution

Let X be a real random variabe, and $k \in \mathbb{N}^*$. X is said to have a moment of order k if $E(|X|^k) < +\infty$, which means that :

- is X is continuous with density $p(x)$, the integral

$$\int_{-\infty}^{+\infty} x^k f(x) dx$$

is **absolutely** convergent. The moment is then equal to the sum of that series (without absolute value).

Moments of a distribution

Proposition

Let $k_1 < k_2$ be integers. Let X be a real random variable. Then if X has a moment of order k_2, X also has a moment of order k_1.

Moments of a distribution

Exercice 1 : **Prove the proposition**

Proposition

Let $k_1 < k_2$ be integers. Let X be a real random variable. Then if X has a moment of order k_2, X also has a moment of order k_1.

Expected value, variance

Definition

Expected value, variance

- If X has a moment of order 1, it is called the **expected value**
- If X has a moment of order 2, then $X - E(X)$ also has a moment of order 2. This moment is called the variance of X.

$$V(X) = E\left((X - E(X))^2\right)$$

We often note $\sigma(X) = \sqrt{Var(X)}$.

Expected value, variance

Proposition

Let a and b be real numbers, and X a random variable that admits a moment of order 2. Then

$$Var(aX + b) = a^2 Var(X)$$

Independence

Proposition

Let $(X_1, \dots, X_n)$ be n mutually independent real random variables. Then if they all admit a moment of order 1, then the product $X_1 X_2 \dots X_n$ also does admit a moment of order 1 and

$$E(X_1 X_2 \dots X_n) = \prod_{i=1}^{n} E(X_i)$$

If they also admit moments of order 2, then

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i)$$

Covariance

Lemma

Let $X, Y, Z \in \mathbb{R}$ be real random variables with a moment of order 2. We have :

$$Cov(X+Y, Z) = Cov(X, Z) + Cov(Y, Z)$$

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

$$|Cov(X, Y)| \leq \sigma(X)\sigma(Y)$$

Convention

From now on, if we write $E(X)$ or $Var(X)$, we implicitely assume that the quantities are correctly defined.

Random vectors

Definition

Let $X \in \mathbb{R}^d$ be a random vector.

$$X = \begin{pmatrix} X_1 \\ \dots \\ X_i \\ \dots \\ X_d \end{pmatrix}$$

The **expected value** of the vector writes

$$E(X) = \begin{pmatrix} E[X_1] \\ \dots \\ E[X_i] \\ \dots \\ E[X_d] \end{pmatrix}$$

Random vectors

Definition

$$X = \begin{pmatrix} X_1 \\ \dots \\ X_i \\ \dots \\ X_d \end{pmatrix}$$

The **variance matrix** (or **covariance matrix, variance-covariance, dispersion matrix**) $Var(X)$ is defined as

$$[Var(X)]_{ij} = Cov(X_i, X_j)$$

Random vector

Exercice 2 : **Random vector**
Whar does it mean to have a vector such that

$$Var(X) = \lambda I_d \tag{1}$$

?

Expected value as a minimization

Exercice 3 : **Expected value as minimization.**
Show that $E(X)$ is the value that minimizes the function

$$f(t) = E\left((X - t)^2\right) \tag{2}$$

Markov inequality

Proposition

Markov inequality
Let X ba a real non-negative random variable (variable aléatoire réelle positive), such that $E(|X|) < +\infty$. Let $a > 0$. Then

$$P(X \geq a) \leq \frac{E(X)}{a}$$

Chebychev inequality

Proposition

Chebyshev inequality Let X ba a real random variable, such that $E(|X|^2) < +\infty$. Let $a > 0$. Then

$$P(|X - E[X]| > a) \leq \frac{Var(X)}{a^2}$$

Weak law of large numbers

Theorem

Weak law of large numbers

Let $(X_n)_{n \in \mathbb{N}}$ be a sequence of i.i.d. variables that have a moment of order 2. We note m their expected value. Then

$$\forall \epsilon > 0, \lim_{n \to +\infty} P\Big(|\frac{1}{n}\sum_{i=1}^{n} X_i - m| \geq \epsilon\Big) = 0$$

We say that we have ***convergence in probability****.*

Standard deviation of the average

If $E(S_n) = m$, then

$$\sqrt{Var\Big(S_n - m\Big)} = \frac{\sigma}{\sqrt{n}} \tag{3}$$

Differentiable function

Definition

Differentiable function

Let V and W be real Hilbert spaces (complete vector space with an inner product). Let $f : V \to W$. We say that f is differentiable in $x \in V$ if there exsists a continuous linear application $L_x : V \to \mathbb{R}$ such that

$$f(x+h) = f(x) + L_x(h) + o(h)$$

with $\lim_{h \to 0} \frac{|o(h)|}{||h||} = 0$.

Gradient

If $W = \mathbb{R}$.

$$\exists ! p_x \in V, \forall h \in V, L_x(h) = \langle p, h \rangle \tag{4}$$

p is sometimes noted $f'(x)$, $\nabla_x f$ or $\nabla f(x)$.

Two time differentiable functions

Definition

Two times differentiable function

$W = \mathbb{R}$. If $x \mapsto \nabla_x f$ is differentiable in x, the we say that f is two times differentiable in x. In that case we note $f"(x)$ the second-order derivative, that satisfies :

$$\nabla_{x+h} f = \nabla_x f + f"(x)(h) + o(h)$$

Two times differentiable function

Lemma

$\forall x \in V$, $f''(x)(h) \in V$, that can also be identified to an element of its dual space V^*. With the notation $f''(x)(h, h') = f''(x)(h)(h')$, we can show that

$$f(x+h) = f(x) + \nabla_x f(h) + \frac{1}{2} f''(x)(h, h) + o(||h||^2)$$

Jacobian matrix

- If $f : \mathbb{R}^d \to \mathbb{R}^p$ is differentiable on $\mathbb{R}^d$ we note $L_x^f : \mathbb{R}^d \to \mathbb{R}^p$ the differential in x. Its matrix is the **Jacobian** also noted $L_x^f \in \mathbb{R}^{p,d}$.
- If f has real values ($p = 1$), then

$$\nabla_x f = (L_x^f)^T \in \mathbb{R}^{d,1}$$

- If $g : \mathbb{R}^p \to \mathbb{R}^q$ is differentiable in $f(x)$:

$$L_x^{g \circ f} = L_{f(x)}^g L_x^f \in \mathbb{R}^{q,d} \tag{5}$$

Hessian

If $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is two times differentiable in x, then $\nabla f : \mathbb{R}^d \rightarrow \mathbb{R}^d$, $x \mapsto \nabla_x f$ has a matrix $H_x^f \in \mathbb{R}^{d,d}$, called the **Hessian**.

$$\nabla_{x+h} f = \nabla_x f + H_x^f h + o(h)$$

Then, the development of f around x can be written

$$f(x+h) = f(x) + L_x^f h + \frac{1}{2} h^T (H_x^f) h + o(||h||^2)$$

Explicit formulation of gradient

If f has real values ($p = 1$), then

$$\nabla_x f = \begin{pmatrix} \frac{\partial f}{\partial x_1}(x) \\ \dots \\ \frac{\partial f}{\partial x_i}(x) \\ \dots \\ \frac{\partial f}{\partial x_d}(x) \end{pmatrix}$$

Explicit formulation of the Hessian

if f is two times differentiable, then the Hessian reads :

$$H_x^f = \begin{pmatrix} \frac{\partial^2 f}{\partial^2 x_1}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \dots & \frac{\partial^2 f}{\partial x_d \partial x_1}(x) \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \frac{\partial^2 f}{\partial x_1 \partial x_d}(x) & \frac{\partial^2 f}{\partial x_2 \partial x_d}(x) & \dots & \frac{\partial^2 f}{\partial x_d^2}(x) \end{pmatrix}$$

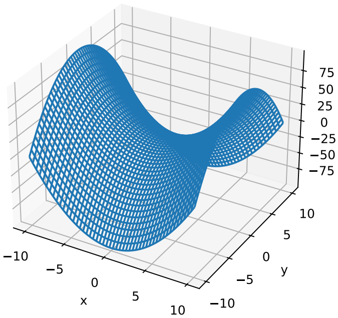
Exercice 4 : Hessian

Hessian of $f : (x, y) \mapsto x^2 - y^2$?

$$f : (x, y) \mapsto x^2 - y^2 \tag{6}$$

$$H_x^f = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} \tag{7}$$

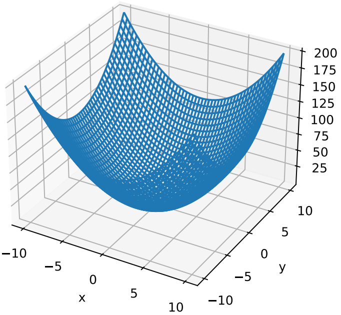
Neither positive nor negative Hessian (saddle point)



$$f : (x, y) \mapsto x^2 + y^2 \tag{8}$$

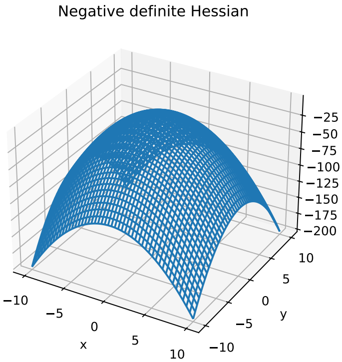
$$H_x^f = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \tag{9}$$

Positive definite Hessian



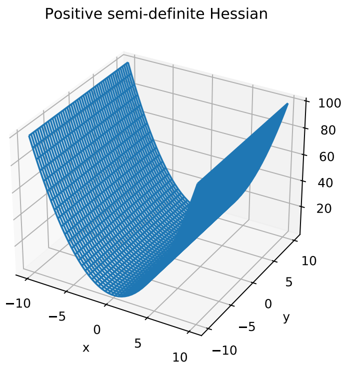
$$f : (x, y) \mapsto -x^2 - y^2 \tag{10}$$

$$H_x^f = \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix} \tag{11}$$



$$f : (x, y) \mapsto x^2 \tag{12}$$

$$H_x^f = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \tag{13}$$



Lipshitz continuity

Definition

L-Lipschitz continuous function

f differentiable, $L > 0$. f is L-Lipschitz continuous if $\forall x, y \in \mathbb{R}^d$,

$$||f(x) - f(y)|| \leq L||x - y||$$

Definition

L-Lipschitz continuous gradients

f differentiable, $L > 0$. f has L-Lipschitz continuous gradients if $\forall x, y \in \mathbb{R}^d$,

$$||\nabla_x f - \nabla_y f|| \leq L||x - y||$$

Quadratic function

Let $A \in \mathbb{R}^{d,d}$ be a symmetric real matrix. If $f(x) = \frac{1}{2}x^T Ax - b^T x$.
Exercice 5 : **Compute** $\nabla_x f$ and H_x^f.

Quadratic function

Let $A \in \mathbb{R}^{d,d}$ be a symmetric real matrix. If $f(x) = \frac{1}{2}x^T A x - b^T x$.

- $\nabla_x f = Ax - b$
- $H_x^f = A$.

Supervised learning

- The dataset D_n is a collection of n samples $\{(x_i, y_i)\}_{1 \leq i \leq n}$, that are **independent and identically distributed** draws of a joint random variable (X, Y).
- the law of (X, Y) is unknown, we can note it ρ. We assume there exists an unknown function f that relates X and Y (not necessary deterministic).
- we look for an estimator $\tilde{f}_n$ of f. n refers to the fact that we have n samples.

A **learning rule** $\mathcal{A}$ is a application that associates a **prediction function**, or **estimator** $\tilde{f}_n$, to D_n.

$$\mathcal{A} : \begin{cases} \cup_{n \in \mathbb{N}} (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}} \\ D_n \mapsto \tilde{f}_n \end{cases}$$

Risks

Let l be a loss.
The **risk** (or **statistical risk, generalization error, test error)** of estimator f writes

$$E_{(X,Y)\sim\rho}[l(Y, f(X))]$$

The **empirical risk (ER)** of an estimator f writes

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, f(x_i))$$

The risks depend on the loss l.

Excess risk

We define the **target function** f^* by

$$f^* \in \arg\min_{f:X \to Y} R(f)$$

with $f : X \to Y$ set of measurable functions. Notation : $R(f^*) = R^*$.

Definition

Fundamental problem of Supervised Learning
Estimate f^* given only D_n and l.

$\tilde{f}_n$ is the minimizer of the empirical risk.

Definition

Excess risk

The **excess risk** $\mathcal{R}(\tilde{f}_n)$ measures how close $\tilde{f}_n$ is to the best possible f^*, in terms of expected risk (average / expecter) error on new examples.

$$\mathcal{R}(\tilde{f}_n) = R(\tilde{f}_n) - R(f^*)$$

Definition

Consistency

The algorithm $\mathcal{A}$ is said to be **consistent** if

$$\lim_{n \to +\infty} E_{D_n} \mathcal{R}(\tilde{f}_n) = 0$$

Bayes predictor

Under some conditions, we can give an explicit formulation of f^*, the best predictor in $\mathcal{Y}^{\mathcal{X}}$, although we can not compute it without the knowledge of the distribution of (X, Y).
In this section we assume we have access to ρ and we approximately ignore measurability issues.
Decision theory : "if we have a perfect knowledge of the underlying probability distribution of the data, what should be done ?"

Bayes predictor

$$f^*(x) = \arg\min_{z \in \mathcal{Y}} E[l(Y, z)|X = x] \tag{14}$$

$E[l(Y, z)|X = x]$ denotes the **conditional expectation** of $l(Y, z)$ given that $X = x$.

$$E[l(Y, z)|X = x] = \int_{y \in \mathbb{R}} l(y, z) p_{Y|X=x}(y) dy \tag{15}$$

Bayes predictor for binary classification

- $\mathcal{Y} = \{0, 1\}$.
- $l(y, z) = 1_{y \neq z}$.

Exercice 6 : **What is the Bayes predictor ?**

Bayes predictor for binary classification

- $\mathcal{Y} = \{0, 1\}$.
- $l(y, z) = 1_{y \neq z}$.
- If $\eta(x) = P(Y = 1|X = x)$, then

$$R^* = E[min(\eta(x), 1 - \eta(x))] \tag{16}$$

Bayes predictor for binary classification

- $\mathcal{Y} = \{0, 1\}$.
- $l(y, z) = 1_{y \neq z}$.
- If $\eta(x) = P(Y = 1|X = x)$, then

$$R^* = E[min(\eta(x), 1 - \eta(x))] \tag{17}$$

Exercice 7 : **What is the meaning of having $R^* = 0$ in that context ?**

Bayes predictor for regression, squared loss

- $\mathcal{Y} = \mathbb{R}$, $\mathcal{X} = \mathbb{R}$.
- $l(y, z) = (y - z)^2$

Exercice 8 : **What is the Bayes predictor ?**

Conditional expectation

Definition

Conditional expectation

$$f^*(x) = E[Y|X = x] \tag{18}$$

Risk decomposition

We will introduce the concept of risk decomposition.

- f^* : Bayes predictor
- F : Hypothesis space
- $\tilde{f}_n$: estimated predictor (hence in F).

$$R(\tilde{f}_n) - R^* = \left(R(\tilde{f}_n) - \inf_{f \in F} R(f) \right) + \left(\inf_{f \in F} R(f) - R^* \right) \tag{19}$$

Risk decomposition

We will introduce the concept of risk decomposition.

- f^* : Bayes predictor
- F : Hypothesis space
- $\tilde{f}_n$: estimated predictor ($\in F$).

$$R(\tilde{f}_n) - R^* = \Big(R(\tilde{f}_n) - \inf_{f \in F} R(f)\Big) + \Big(\inf_{f \in F} R(f) - R^*\Big) \tag{20}$$

However : $\tilde{f}_n$ is a **random variable**, and so is $R(\tilde{f}_n)$. We can also consider the expected value of this quantity.

Risk decomposition

- f^* : Bayes predictor
- F : Hypothesis space
- $\tilde{f}_n$: estimated predictor ($\in F$).

$$E[R(\tilde{f}_n)] - R^* = \Big(E[R(\tilde{f}_n)] - \inf_{f \in F} R(f)\Big) + \Big(\inf_{f \in F} R(f) - R^*\Big) \quad (21)$$

Risk decomposition : bias term

- f^* : Bayes predictor
- F : Hypothesis space
- $\tilde{f}_n$: estimated predictor ($\in F$).

$$E[R(\tilde{f}_n)] - R^* = \Big(E[R(\tilde{f}_n)] - \inf_{f \in F} R(f)\Big) + \Big(\inf_{f \in F} R(f) - R^*\Big) \quad (22)$$

Approximation error (bias term) : depends on f^* and F, not on $\tilde{f}_n$, D_n.

$$\inf_{f \in F} R(f) - R^* \geq 0$$

Risk decomposition : bias term

- f^* : Bayes predictor
- F : Hypothesis space
- $\tilde{f}_n$: estimated predictor ($\in F$).

$$E[R(\tilde{f}_n)] - R^* = \Big(E[R(\tilde{f}_n)] - \inf_{f \in F} R(f)\Big) + \Big(\inf_{f \in F} R(f) - R^*\Big) \quad (23)$$

Estimation error (variance term, fluctuation error, stochastic error) : depends on D_n, F, $\tilde{f}_n$.

$$E(R(\tilde{f}_n)) - \inf_{f \in F} R(f) \geq 0$$

Underfitting and overfitting

Approximation error (bias term) : depends on f^* and F, not on $\tilde{f}_n$, D_n.

$$\inf_{f \in F} R(f) - R^* \geq 0$$

Estimation error (variance term, fluctuation error, stochastic error) : depends on D_n, F, $\tilde{f}_n$.

$$E(R(\tilde{f}_n)) - \inf_{f \in F} R(f) \geq 0$$

- too small F : underfitting (large bias, small variance)
- too large F : overfitting (small bias, large variance)

Expected value of empirical risk

If $h \in F$ is fixed (not $\tilde{f}_n$), then $R_n(h)$ is an **unbiased estimator** of the generalization error $R(h)$.

$$E[R_n(h)] = R(h) \tag{24}$$

But

$$E[R_n(\tilde{f}_n)] \neq R(\tilde{f}_n) \tag{25}$$

OLS

We will introduce the Ordinary Least-squares (OLS) problem.

- $\mathcal{X} = \mathbb{R}^d$
- $\mathcal{Y} = \mathbb{R}$.
- $l(y, y') = (y - y')^2$
- $$F = \{x \mapsto \theta^T x, \theta \in \mathbb{R}^d\}$$

OLS

The dataset is stored in the **design matrix** $X \in \mathbb{R}^{n \times d}$.

$$X = \begin{pmatrix} x_1^T \\ \dots \\ x_i^T \\ \dots \\ x_n^T \end{pmatrix} = \begin{pmatrix} x_{11}, & \dots, & x_{1j}, & \dots & x_{1d} \\ & & \dots & & \\ x_{i1}, & \dots, & x_{ij}, & \dots & x_{id} \\ & & \dots & & \\ x_{n1}, & \dots, & x_{nj}, & \dots & x_{nd} \end{pmatrix}$$

The vector of predictions of the estimator writes $Y = X\theta$. Hence,

$$\begin{aligned} R_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta^T x_i)^2 \\ &= \frac{1}{n}||Y - X\theta||_2^2 \end{aligned}$$

OLS estimator

We assume that X is **injective.** Necessary, $d \leq n$.

Proposition

Closed form solution
*We X is injective, there exists a unique minimiser of $R_n(\theta)$, called the **OLS estimator**, given by*

$$\hat{\theta} = (X^T X)^{-1} X^T Y \tag{26}$$

Setup

- **Linear model** : $\exists \theta^* \in \mathbb{R}^d$,

$$Y_i = {\theta^*}^T x_i + Z_i, \forall i \in [1, n]$$

  and Z_i is a centered noise (or error) ($E[Z_i] = 0$) with variance σ^2.
- **Fixed design** : X deterministic.

Then :

- $\hat{\theta}$ is **unbiased** : $E[\hat{\theta}] = \theta^*$.
- $Var(\hat{\theta}) = \frac{\sigma^2}{n} \Sigma^{-1}$.

with $\Sigma = X^T X \in \mathbb{R}^{d \times d}$.