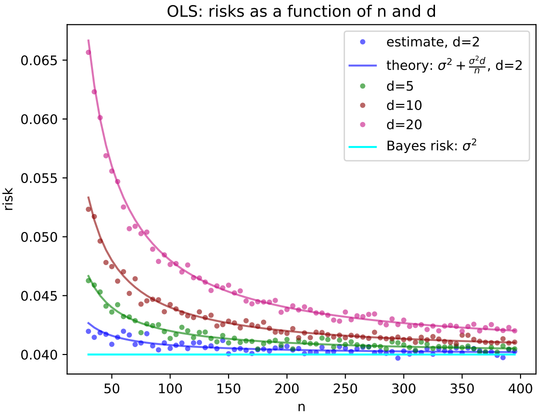
Fondamentaux théoriques du machine learning



Overview of lecture 3

Risks and risk decompositions
- Examples
- Expected value of empirical risk
- Risk decomposition
- Optimization error

Optimization in machine learning
- Existence results
- Convex analysis
- Gradients

Ordinary Least squares II
- OLS estimator
- Statistical analysis of OLS

Risks and risk decompositions
- Examples
- Expected value of empirical risk
- Risk decomposition
- Optimization error

Optimization in machine learning
- Existence results
- Convex analysis
- Gradients

Ordinary Least squares II
- OLS estimator
- Statistical analysis of OLS

Bayes rule

$$P(A \cap B) = P(A|B)P(B) \tag{1}$$

Law of total probability

If for instance $\Omega = A \cup B \cup C$ and A, B, C are mutually exclusive, then

$$P(X) = P(X \cap A) + P(X \cap B) + P(X \cap C) \tag{2}$$

Exercice 1 : Consider the following random variable (X, Y).

- $X \sim B(\frac{1}{2})$,

$$Y = \begin{cases} B(p) \text{ if } X = 1 \\ B(q) \text{ if } X = 0 \end{cases}$$

  With $B(p)$ a Bernoulli law with parameter p.
- Hence $\mathcal{X} = \{0, 1\}$, $\mathcal{Y} = \{0, 1\}$.

Exercice 1 : Consider the following random variable (X, Y).

- $X \sim B(\frac{1}{2})$,

$$Y = \begin{cases} B(p) \text{ if } X = 1 \\ B(q) \text{ if } X = 0 \end{cases}$$

  With $B(p)$ a Bernoulli law with parameter p.

- A predictor $f_1 : \{0, 1\} \rightarrow \{0, 1\}$:

$$f_1 = \begin{cases} 1 \text{ if } x = 1 \\ 0 \text{ if } x = 0 \end{cases}$$

With the "$0 - 1$" loss, what is the risk (generalization error) of f_1, $R(f_1)$?

Exercice 1 : Consider the following random variable (X, Y).

- $X \sim B(\frac{1}{2})$,

$$Y = \begin{cases} B(p) \text{ if } X = 1 \\ B(q) \text{ if } X = 0 \end{cases}$$

- $f_1 : \{0, 1\} \rightarrow \{0, 1\}$:

$$f = \begin{cases} 1 \text{ if } x = 1 \\ 0 \text{ if } x = 0 \end{cases}$$

$$\begin{aligned} R(f_1) &= E[l(Y, f(X))] \\ &= 1 \times P(Y \neq f(X)) + 0 \times P(Y = f(X)) \\ &= P(Y \neq f(X)) \end{aligned} \tag{3}$$

- $X \sim B(\frac{1}{2})$,

$$Y = \begin{cases} B(p) \text{ if } X = 1 \\ B(q) \text{ if } X = 0 \end{cases}$$

- $f_1 : \{0,1\} \to \{0,1\}$:

$$f = \begin{cases} 1 \text{ if } x = 1 \\ 0 \text{ if } x = 0 \end{cases}$$

$$\begin{aligned} R(f_1) &= E[l(Y, f(X))] \\ &= 1 \times P(Y \neq f(X)) + 0 \times P(Y = f(X)) \\ &= P(Y \neq f(X)) \\ &= P((Y \neq f(X)) \cap (X = 1)) + P((Y \neq f(X)) \cap (X = 0)) \end{aligned} \tag{4}$$

$$
\begin{aligned}
R(f_1) &= E[l(Y, f(X))] \\
&= 1 \times P(Y \neq f(X)) + 0 \times P(Y = f(X)) \\
&= P(Y \neq f(X)) \\
&= P((Y \neq f(X)) \cap (X = 1)) + P((Y \neq f(X)) \cap (X = 0)) \\
&= P((Y \neq f(X))|X = 1)P(X = 1) \\
&+ P((Y \neq f(X))|X = 0)P(X = 0)
\end{aligned} \tag{5}
$$

$$
\begin{aligned}
R(f_1) &= E[l(Y, f(X))] \\
&= 1 \times P(Y \neq f(X)) + 0 \times P(Y = f(X)) \\
&= P(Y \neq f(X)) \\
&= P((Y \neq f(X)) \cap (X = 1)) + P((Y \neq f(X)) \cap (X = 0)) \\
&= P((Y \neq f(X))|X = 1)P(X = 1) \\
&+ P((Y \neq f(X))|X = 0)P(X = 0) \\
&= \frac{1}{2}P((Y \neq 1)|X = 1) + \frac{1}{2}P((Y \neq 0)|X = 0)
\end{aligned} \tag{6}
$$

$$
\begin{aligned}
R(f_1) &= E[l(Y, f(X))] \\
&= 1 \times P(Y \neq f(X)) + 0 \times P(Y = f(X)) \\
&= P(Y \neq f(X)) \\
&= P((Y \neq f(X)) \cap (X = 1)) + P((Y \neq f(X)) \cap (X = 0)) \\
&= P((Y \neq f(X))|X = 1)P(X = 1) \\
&+ P((Y \neq f(X))|X = 0)P(X = 0) \\
&= \frac{1}{2}P((Y = 0)|X = 1) + \frac{1}{2}P((Y = 1)|X = 0) \\
&= \frac{1}{2}(1 - p) + \frac{1}{2}q
\end{aligned} \tag{7}
$$

Exercice 2 : Now consider

$$f_2 = \begin{cases} 0 \text{ if } x = 1 \\ 1 \text{ if } x = 0 \end{cases}$$

What is $R(f_2)$?

Exercice 2 :

$$\forall x, f_2(x) = 1 - f_1(x) \tag{8}$$

Exercice 2 :

$$\forall x, f_2(x) = 1 - f_1(x) \tag{9}$$

Hence

$$\begin{aligned} R(f_2) &= P(Y \neq f_2(X)) \\ &= P(Y \neq (1 - f_1(X))) \\ &= P(Y = f_1(X)) \\ &= 1 - R(f_1) \end{aligned} \tag{10}$$

Exercice 3 : Third predictor :

$$\forall x, f_3(x) = 1 \tag{11}$$

What is $R(f_3)$?

Exercice 3 :

$$
\begin{aligned}
R(f_3) &= P(Y \neq f_3(X)) \\
&= P(Y = 0)
\end{aligned}
\tag{12}
$$

Exercice 3 :

$$
\begin{aligned}
R(f_3) &= P(Y \neq f_3(X)) \\
&= P(Y = 0) \\
&= P(Y = 0 \cap X = 0) + P(Y = 0 \cap X = 1) \\
&= P(Y = 0|X = 0)P(X = 0) + P(Y = 0|X = 1)P(X = 1) \\
&= \frac{1}{2}(1 - p) + \frac{1}{2}(1 - q)
\end{aligned}
\tag{13}
$$

Exercice 4 :

Now, we observe the following dataset :

$$D_4 = \{(0,1),(0,0),(0,0),(1,0)\} \tag{14}$$

Compute the empirical risks $R_4(f_1)$, $R_4(f_2)$, $R_4(f_3)$.

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

$$D_4 = \{(0,1),(0,0),(0,0),(1,0)\} \tag{15}$$

$$\begin{aligned} R_4(f_1) &= \frac{1}{4}\sum_{i=1}^{4} l(f_1(x_i), y_i) \\ &= \frac{1}{4}\Big(l(f_1(0),1) + l(f_1(0),0)) + l(f_1(0),0)) + l(f_1(1),0))\Big) \\ &= \frac{1}{4} \times 2 \\ &= \frac{1}{2} \end{aligned} \tag{16}$$

$$D_4 = \{(0,1),(0,0),(0,0),(1,0)\} \tag{17}$$

$$\begin{aligned} R_4(f_2) &= \frac{1}{4}\sum_{i=1}^{4} l(f_2(x_i), y_i) \\ &= \frac{1}{4}\Big(l(f_2(0),1) + l(f_2(0),0)) + l(f_2(0),0)) + l(f_2(1),0))\Big) \\ &= \frac{1}{4} \times 2 \\ &= \frac{1}{2} \end{aligned} \tag{18}$$

$$D_4 = \{(0,1),(0,0),(0,0),(1,0)\} \tag{19}$$

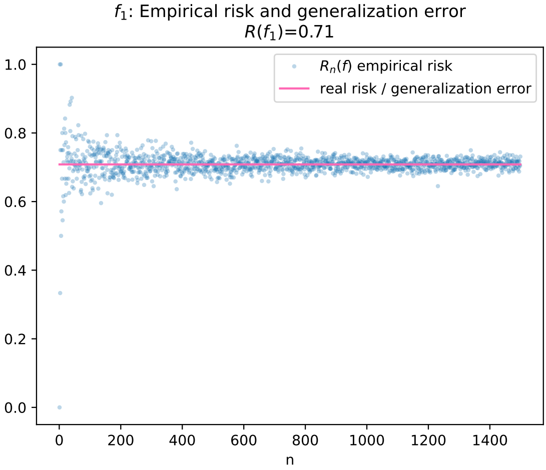
$$\begin{aligned} R_4(f_3) &= \frac{1}{4}\sum_{i=1}^{4} l(f_3(x_i), y_i) \\ &= \frac{1}{4}\Big(l(f_3(0),1) + l(f_3(0),0)) + l(f_3(0),0)) + l(f_3(1),0))\Big) \\ &= \frac{1}{4} \times 3 \\ &= \frac{3}{4} \end{aligned} \tag{20}$$

Random variable

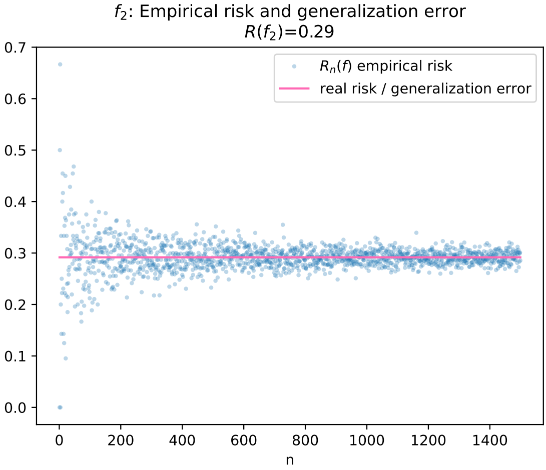
- $R_4(f)$ (empirical risk) **depends** on D_4. If we sample another dataset, $R_4(f)$ is likely to change, it is a **random variable**.
- $R(f)$ (generalization error) is **deterministic**, given the joint law of (X, Y).

Given a predictor f, a natural question arises :
Does $R_n(f)$ have a limit when $n \to +\infty$?

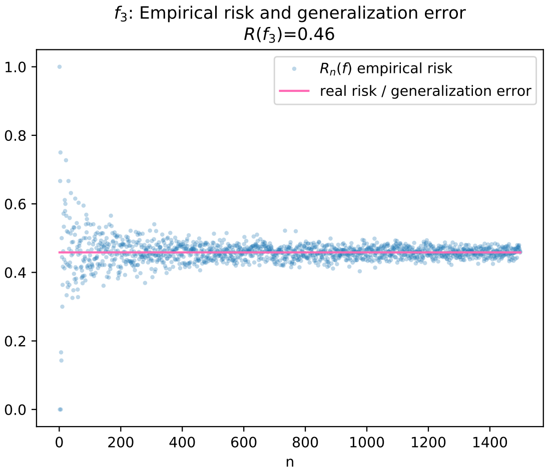
Simulations



Simulations



Simulations



Convergence of empirical risk

We fix $f \in H$ (hypothesis space). We assume that the samples (X_i, Y_i) are i.i.d, with the distribution of (X, Y), noted ρ. Then, under some assumptions (for instance, if the empirical risks are bounded), we have that in probability :

$$\lim_{n \to +\infty} R_n(f) = R(f) \tag{21}$$

The empirical risk of a fixed f conerges to its real risk.

Proof

- $$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, f(x_i))$$
- $$\forall i, E[l(f(X_i), Y_i)] = E[l(f(X), Y)] \tag{22}$$
- i.i.d. variables.
- Law of large numbers.

Also

$$
\begin{aligned}
E_{D_n \sim \rho}(R_n(h)) &= \frac{1}{n}\sum_{i=1}^{n} E_{D_n \sim \rho}(l(f(X_i), Y_i)) \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{(X,Y)\sim \rho}(l(f(X), Y)) \\
&= E_{(X,Y)\sim \rho}(l(f(X), Y)) \\
&= R(h)
\end{aligned}
$$

However, we do **not** have

$$E[R_n(\tilde{f}_n)] = R(\tilde{f}_n) \tag{23}$$

where $\tilde{f}_n$ is the minimizer of the empirical risk.
$\tilde{f}_n$ depends on the dataset D_n.

$$E_{D_n \sim \rho}(r(\tilde{f}_n(X_i), Y_i)) \neq E_{(X,Y) \sim \rho}(r(\tilde{f}_n(X), Y)) \tag{24}$$

Risk decomposition

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

$$E[R(\tilde{f}_n)] - R^* = \Big(E[R(\tilde{f}_n)] - \inf_{f \in F} R(f)\Big) + \Big(\inf_{f \in F} R(f) - R^*\Big) \quad (25)$$

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)

Deterministic bound on the estimation error

We consider the best estimator in hypothesis space

$$f_a = \arg\min_{h \in F} R(h)$$

We can show that

$$R(\tilde{f}_n) - R(f_a) \leq 2 \sup_{h \in F} |R(h) - R_n(h)| \tag{26}$$

Deterministic bound on the estimation error

$$f_a = \arg\min_{h \in F} R(h)$$

$$\begin{aligned} R(\tilde{f}_n) - R(f_a) = {} & \left(R(\tilde{f}_n) - R_n(\tilde{f}_n)\right) \\ & + \left(R_n(\tilde{f}_n) - R_n(f_a)\right) \\ & + \left(R_n(f_a) - R(f_a)\right) \end{aligned} \tag{27}$$

Deterministic bound on the estimation error

$$f_a = \arg\min_{h \in F} R(h)$$

$$\begin{aligned} R(\tilde{f}_n) - R(f_a) = & \left(R(\tilde{f}_n) - R_n(\tilde{f}_n)\right) \\ & + \left(R_n(\tilde{f}_n) - R_n(f_a)\right) \\ & + \left(R_n(f_a) - R(f_a)\right) \end{aligned} \tag{28}$$

But by definition $\tilde{f}_n$ minimizes R_n, so $\left(R_n(\tilde{f}_n) - R_n(f_a)\right) \leq 0$.

Deterministic bound on the estimation error

$$R(\tilde{f}_n) - R(f_a) \leq 2 \sup_{h \in F} |R(h) - R_n(h)| \tag{29}$$

Later in the course, based on **concentration inequalities** we will further build on this result and prove a probabilistic bound of the form

$$R(\tilde{f}_n) - R(f_a) \leq \frac{C}{\sqrt{n}} \tag{30}$$

(remember that by definition $0 \leq R(\tilde{f}_n) - R(f_a)$)

Order of magnitude of estimation error

We keep in mind that

$$R(\tilde{f}_n) - R(f_a) = \mathcal{O}(\frac{C}{\sqrt{n}}) \tag{31}$$

Approximate solution

- In machine learning, it is often not necessary to find the actual minimizer of the empirical risk , as there is an estimation error of $\mathcal{O}(\frac{1}{\sqrt{n}})$. [Bottou and Bousquet, 2009,]
- We can use an approximate solution $\hat{f}_n$, such that

$$R_n(\hat{f}_n) \leq R_n(\tilde{f}_n) + \rho \qquad (32)$$

with ρ a predefined tolerance.
- This important because in large-scale ML, the **computation time** need to be optimized.

Approximate solution

This gives a new risk decomposition :

$$E[R(\hat{f}_n)] - R^* = \Big(E[R(\hat{f}_n)] - E[R(\tilde{f}_n)]\Big) + \Big(E[R(\tilde{f}_n)] - \inf_{f \in F} R(f)\Big) + \Big(\inf_{f \in F} R(f) - R^*\Big) \tag{33}$$

Approximate solution

This gives a new risk decomposition :

$$E[R(\hat{f}_n)] - R^* = \Big(E[R(\hat{f}_n)] - E[R(\tilde{f}_n)]\Big) + \Big(E[R(\tilde{f}_n)] - \inf_{f \in F} R(f)\Big) + \Big(\inf_{f \in F} R(f) - R^*\Big) \quad (34)$$

$E[R(\hat{f}_n)] - E[R(\tilde{f}_n)]$ is the **optimization error.**
To conclude, we have :

- an approximation error
- an estimation error
- an optimization error

Risks and risk decompositions
- Examples
- Expected value of empirical risk
- Risk decomposition
- Optimization error

Optimization in machine learning
- Existence results
- Convex analysis
- Gradients

Ordinary Least squares II
- OLS estimator
- Statistical analysis of OLS

Minimizers

Definition

Let $f : \mathbb{R}^d \to \mathbb{R}$ be defined on $K \subset \mathbb{R}^d$.
$x \in K$ is a local minimum of f on K if and only if

$$\exists \delta > 0, \forall y \in K, ||y - x|| < \delta \Rightarrow f(x) \leq f(y)$$

$x \in K$ is a global minimum of f on K if and only if

$$\forall y \in K, f(x) \leq f(y)$$

Existence result

Theorem

Existence of a global minimum in $\mathbb{R}^d$

Let K *be a closed non-empty subset of* $\mathbb{R}^d$*, and* $f : \mathbb{R}^d \to \mathbb{R}$ *a continuous coercive function. Then, there exists at least a global minimum of* f *on* K*.*

Convexity

Definition

The function $f : \Omega \to \mathbb{R}$ with Ω convex is :

- **convex** if $\forall x, y \in \Omega, \alpha \in [0, 1]$,

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y)$$

- **strictly convex** if $\forall x, y \in \Omega, \alpha \in [0, 1]$,

$$f(\alpha x + (1-\alpha)y) < \alpha f(x) + (1-\alpha)f(y)$$

- μ-**strongly convex** if $\forall x, y \in \Omega, \alpha \in [0, 1]$,

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) - \frac{\mu}{2}\alpha(1-\alpha)||x-y||^2$$

Examples

- All norms are convex.
- $x \mapsto \theta^T x$ is convex on $\mathbb{R}^d$ with $\theta \in \mathbb{R}^d$ (linear form)
- if Q is a symmetric semidefinite positive matrix, then $x \mapsto x^T Qx$ is convex.
- if Q is a symmetric definite positive matrix (matrice définie positive) with smallest eigenvalue $\lambda_{min} > 0$, then $x \mapsto x^T Qx$ is $2\lambda_{min}$- strongly convex.
- If f is increasing and convex and g is convex, then $f \circ g$ is convex.
- Is f in convex and g is linear, then $f \circ g$ is convex.

Differential formulation of convexity

Proposition

Let $f : V \to \mathbb{R}$ be a differentiable function. The following conditions are equivalent.

- *f is convex.*
- $\forall x, y \in V, f(y) \geq f(x) + (f'(x)|y - x)$ *(f is above its tangent space)*
- $\forall x, y \in V, (f'(x) - f'(y)|x - y) \geq 0$ *(f' grows)*

Differential formulation of strong convexity

Proposition

Let $f : V \to \mathbb{R}$ be a differentiable function, and $\mu > 0$. The following conditions are equivalent.

- f *is* μ*-convex*
- $\forall x, y \in V, f(y) \geq f(x) + (f'(x)|y - x) + \frac{\mu}{2}||y - x||^2$
- $\forall x, y \in V, (f'(x) - f'(y)|x - y) \geq \mu||x - y||^2$

Convexity of two-times differentiable functions

- f is convex if anf only if

$$\forall x, h \in y, J''(x)(h, h) \geq 0$$

- f is μ-strongly convex if and only if

$$\forall x, h \in y, J''(x)(h, h) \geq \mu ||h||^2$$

Convexity and Hessian

If $V = \mathbb{R}^d$, this translates into

$$\forall x, h \in y, h^T(H_x f)h \geq 0 \tag{35}$$

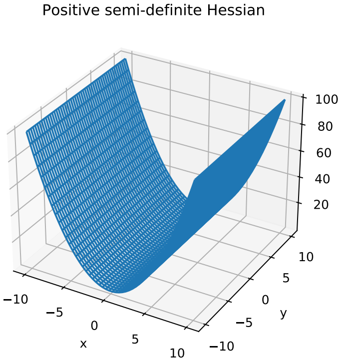
and

$$\forall x, h \in y, h^T(H_x f)h \geq \mu ||h||^2 \tag{36}$$

- 35 means that $\forall x \in \mathbb{R}^d$, all eigenvalues of $H_x f$ are non-negative (positive semi-definite Hessian)
- 36 means that they all are $\geq \mu$ (positive definite Hessian).

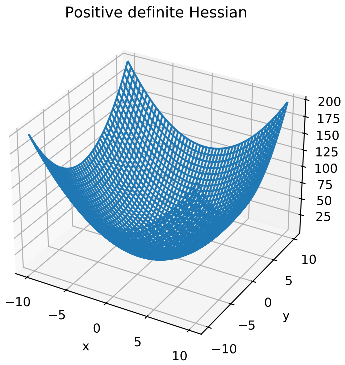
Positive semi-definite Hessian

$$H_x^f = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \tag{37}$$



Positive definite Hessian

$$H_x^f = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \tag{38}$$



Minima of convex functions

Proposition

- *If f is convex, any local minimum is a global minimum. The set of global minimizers is a convex set.*
- *If f is strictly convex, there exists at most one local minimum (that is thus global).*
- *If f is convex and C^1 (differentiable, $a \mapsto df_a$ continuous), then x is a minimum (thus global) of f on V if and only if the gradient cancels in x, $\nabla_x f = 0$. V need not be finite-dimensional.*

Risks and risk decompositions
- Examples
- Expected value of empirical risk
- Risk decomposition
- Optimization error

Optimization in machine learning
- Existence results
- Convex analysis
- Gradients

Ordinary Least squares II
- OLS estimator
- Statistical analysis of OLS

OLS

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

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

Exercice 5 : **Convexity**

Show that the objective function $R_n(\theta)$ is convex in θ.

$$\theta \mapsto ||Y - X\theta||_2^2 \tag{39}$$

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*

$$\hat{\theta} = (X^T X)^{-1} X^T Y \tag{40}$$

Setup

Assumptions :

- **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.

In this setup, we wonder :

- 1) what is the Bayes predictor ? What is the Bayes risk ?
- 2) is the expected value of OLS equal to the Bayes predictor ?
- 3) what is the excess risk of the OLS estimator ?

1) Bayes predictor

With the square loss, we always have that the Bayes predictor is the conditional expectation, see FTML.pdf section 3.1.3.

$$f^*(x) = E[Y|X = x] \tag{41}$$

1) Bayes predictor

$$\begin{aligned} f^*(x) &= E[Y|X=x] \\ &= E[X^T\theta^* + \epsilon|X=x] \\ &= E[X^T\theta^*|X=x] + E[\epsilon|X=x] \\ &= X^T\theta^* \end{aligned} \quad (42)$$

1) Bayes risk

$$
\begin{aligned}
R^* &= E_{X,Y}[(Y - f^*(X))^2] \\
&= E_{X,\epsilon}[(X^T\theta^* + \epsilon - X^T\theta^*)^2] \\
&= E_{X,\epsilon}[\epsilon^2] \\
&= \sigma^2
\end{aligned}
\tag{43}
$$

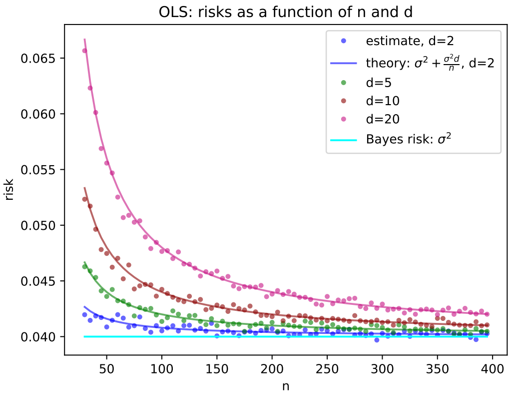
2) Expected value of $\hat{\theta}$

$$
\begin{aligned}
E[\hat{\theta}] &= E[(X^T X)^{-1} X^T Y] \\
&= E[(X^T X)^{-1} X^T (X\theta^* + \epsilon)] \\
&= E[(X^T X)^{-1} X^T (X\theta^*)] + E[(X^T X)^{-1} X^T \epsilon)] \\
&= E[(X^T X)^{-1} (X^T X)\theta^*] + (X^T X)^{-1} X^T E[\epsilon)] \\
&= E[\theta^*] \\
&= \theta^*
\end{aligned}
\tag{44}
$$

We conclude that the OLS estimator is an **unbiased estimator** of θ^*.

3) Excess risk + variance

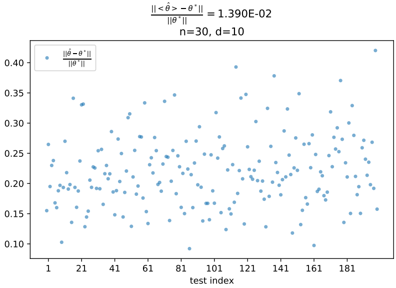
$$R(\hat{\theta}) - R(\theta^*) = \frac{\sigma^2 d}{n} \quad (45)$$



4) Variance

$$Var(\hat{\theta}) = \frac{\sigma^2}{n}\Sigma^{-1} \tag{46}$$

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



References I

Bottou, L. and Bousquet, O. (2009).
The tradeoffs of large scale learning.
Advances in Neural Information Processing Systems 20 - Proceedings of the 2007 Conference, (January 2007).