Probability & Mathematical Needs

N Amanquah
Ashesi University

Acknowledgements:

S Antoine

Laboratoire d'Analyse, Topologie et Probabilités
Université Aix-Marseille 1

Pascal Bootcamp 2010

- **The big picture:**
- Machine learning ...
  - data mining -use historical data to improve decisions
  - apps that cannot be programmed easily eg speech recognition
  - self customizing program

- Learning =improving experience at some task
- How prior knowledge can help in learning

- Some methods:
  - Classification, clustering

  - Illustrate with linear regression(least squares)

Outline

1. Linear Algebra
   - Vector spaces
   - Orthogonality, dot product, norm
   - Matrices
   - Determinant
   - Matrix decompositions (SVD, Choleski, LU, QR)
2. Probabilities
   - Vocabulary, usual laws (discrete, continuous)
   - Conditional probabilities
   - Bayes rule, maximum likelihood, maximum a posteriori
   - Entropy, Kullback-Leibler divergence, perpexity
   - Bounds
3. Optimization
   - Minimima, maxima, saddle points
   - Convex fonctions
   - Primal and dual problems, Lagrange multipliers

Vector spaces

Example ($\mathbb{R}^n$)

$$\mathbb{R}^n = \{x = (x_1, \cdots, x_n)^T : x_i \in \mathbb{R}\ \forall i\}$$

- $x, y \in \mathbb{R}^n \Rightarrow x + y = (x_1 + y_1, \cdots, x_n + y_n)^T \in \mathbb{R}^n$
- $x \in \mathbb{R}^n, \lambda \in \mathbb{R} \Rightarrow \lambda x = (\lambda x_1, \cdots, \lambda x_n)^T \in \mathbb{R}^n$
- $\mathbb{R}^n = \{x : \exists(\lambda_1, \cdots, \lambda_n) \in \mathbb{R}^n\ s.t.\ x = \lambda_1 e_1 + \cdots + \lambda_n e_n\}$

  where $e_i = (0, \cdots, 0, \underset{\substack{\downarrow \\ i}}{1}, 0, \cdots, 0)$.

Example (Solutions of homogeneous differential equations)

$$\mathcal{S} = \{f : \mathbb{R} \to \mathbb{R} : \forall\ t,\ f''(t) + f(t) = 0\}$$

- $f \in \mathcal{S} \Rightarrow -f \in \mathcal{S}$
- $f, g \in \mathcal{S} \Rightarrow f + g \in \mathcal{S}$
- $f \in \mathcal{S}, \lambda \in \mathbb{R} \Rightarrow \lambda f \in \mathcal{S}$
- $\mathcal{S} = \{f : \mathbb{R} \to \mathbb{R} : \exists(\lambda_1, \lambda_2) \in \mathbb{R}^2 s.t.\ f = \lambda_1 \cos + \lambda_2 \sin\}$

Vector spaces

Example ($L^2(\mathbb{R})$)

$$L^2(\mathbb{R}) = \left\{ f : \mathbb{R} \to \mathbb{R} : \int_{\mathbb{R}} |f(x)|^2 dx < \infty \right\}$$

- $f \in L^2(\mathbb{R}) \Rightarrow -f \in L^2(\mathbb{R})$
- $f, g \in L^2(\mathbb{R}) \Rightarrow f + g \in L^2(\mathbb{R})$
- $f \in L^2(\mathbb{R}), \lambda \in \mathbb{R} \Rightarrow \lambda f \in L^2(\mathbb{R})$
- $L^2(\mathbb{R})$ *is not the span of any finite number of its elements.*
- Dot product : $f, g \in L^2(\mathbb{R})$, $\langle f, g \rangle = \int_{\mathbb{R}} f(x)g(x)dx$
- Norm : $\|f\|_{L^2(\mathbb{R})} = \left(\int_{\mathbb{R}} |f(x)|^2 dx\right)^{\frac{1}{2}}$
- Closeness :
  $\forall n,\ f_n \in L^2(\mathbb{R})$ and $\|f_n - f\|_{L^2(\mathbb{R})} \underset{n\to\infty}{\to} 0$ implies $f \in L^2(\mathbb{R})$.

Vector spaces

Example ($L^2(\mathbb{R})$)

$$L^2(\mathbb{R}) = \left\{ f : \mathbb{R} \to \mathbb{R} : \int_{\mathbb{R}} |f(x)|^2 dx < \infty \right\}$$

- $f \in L^2(\mathbb{R}) \Rightarrow -f \in L^2(\mathbb{R})$
- $f, g \in L^2(\mathbb{R}) \Rightarrow f + g \in L^2(\mathbb{R})$
- $f \in L^2(\mathbb{R}), \lambda \in \mathbb{R} \Rightarrow \lambda f \in L^2(\mathbb{R})$
- $L^2(\mathbb{R})$ *is not the span of any finite number of its elements.*
- Dot product : $f, g \in L^2(\mathbb{R})$, $\langle f, g \rangle = \int_{\mathbb{R}} f(x) g(x) dx$
- Norm : $\|f\|_{L^2(\mathbb{R})} = \left(\int_{\mathbb{R}} |f(x)|^2 dx \right)^{\frac{1}{2}}$
- Closeness :
  $\forall n,\ f_n \in L^2(\mathbb{R})$ and $\|f_n - f\|_{L^2(\mathbb{R})} \underset{n\to\infty}{\to} 0$ implies $f \in L^2(\mathbb{R})$.

Vector spaces

Definition (Vector space)

A set $\mathcal{S}$ is called a real vector space if it is endowed with

- an "addition" which is :
  - stable : $x, y \in \mathcal{S} \Rightarrow x + y \in \mathcal{S}$,
  - commutative and associative,
  - with an nul element $0 \in \mathcal{S}$ s.t. $\forall x \in \mathcal{S},\ 0 + x = x$,
  - for which all elements are invertible $x \in \mathcal{S} \Rightarrow\ -x \in \mathcal{S}$.
- the multiplication by a scalar in $\mathbb{R}$ which is :
  - stable : $x \in \mathcal{S},\ \lambda \in \mathbb{R} \Rightarrow\ \lambda x \in \mathcal{S}$.
  - associative and distributive over '+'.

Vector spaces may be decomposed into subspaces :

Definition (Subspace)

A subset F of a vector space $\mathcal{S}$ is a called a subspace of $\mathcal{S}$ if the previous properties are preserved in F.

Vector subspaces, family of vectors, dimension

- Supplementary subspaces :
  - F, G subspaces, $F \cap G = \{0\}$, $\mathcal{S} = F + G$.
  - Any $x \in \mathcal{S}$ has a unique decomposition $x = x_F + x_G$.
- Subspaces may be generated from a family of vectors :
  - $y \in \text{Span}\{x_1, \cdots, x_n\}$ iff $\exists \lambda_1 \cdots \lambda_n \in \mathbb{R}$ s.t. $y = \sum_{i=1}^n \lambda_i x_i$.
  - The family $\{x_i\}_{i=1..n}$ is linearly independent iff the decomposition $y = \sum_{i=1}^n \lambda_i x_i$ is unique.
  - Conversely if $F = \text{Span}\{\{x_i\}_{i=1..n}\}$ then the family $\{x_i\}_{i=1..n}$ is said to generate F.
- The dimension of a (sub)space F is the cardinal of its largest linearly independent family.
  - *Ex :* $\dim(\mathbb{R}^d) = d$, $\dim(\mathcal{S}_{\textit{diff. eq.}}) = 2$, $\dim(L^2(\mathbb{R})) = +\infty$.
  - A hyperplane is a subspace of which the supplementaries have dimension 1.
    - If $\dim(\mathcal{S}) = n$, an hyperplane is any subspace of dimension $n-1$. *Ex : lines in* $\mathbb{R}^2$, *planes in* $\mathbb{R}^3$.

Bases

- The family $\{x_i\}_{i=1..n}$ is a basis of $\mathcal{S}$ iff it is generative and linearly independent. Here n may be ∞ !
  - The cardinal of any basis is exactly the dimension of $\mathcal{S}$ (finite or not).
  - For $y \in \mathcal{S}$ there is a unique decomposition $y = \sum_{i=1..n} \lambda_i x_i$.

Example

- In $\mathbb{R}^d$:
  - $\{e_i\}_{i=1..d}$, where $e_i = (0, \cdots, 0, \overset{\overset{i}{\uparrow}}{1}, 0, \cdots, 0)$ is a basis.
  - $y = (y_1, \cdots, y_d)^T = \sum_{i=1..d} y_i e_i$.
- In $L^2([0, 2\pi])$:
  - $\{cos(mt), sin(mt)\}_{m\in\mathbb{N}}$ is a basis.
  - $f \in L^2([0, 2\pi]), f(t) = \sum_{m\in\mathbb{N}} (a_m \cos(mt) + b_m \cos(mt))$.

Orthogonality, dot product, norm

In $\mathbb{R}^d$:

- The dot product is defined as :

$$\langle x, y \rangle_{\mathbb{R}^d} = \sum_{i=1}^{d} x_i y_i$$

- It is linked to the Euclidian norm :

$$\|x\| = \sqrt{\langle x, x \rangle_{\mathbb{R}^d}} = \sqrt{\sum_{i=1}^{d} |x_i|^2}$$

$$\langle x, y \rangle_{\mathbb{R}^d} = \|x\| \|y\| \cos(\theta)$$

- Any subspace has a unique orthogonal supplementary

Orthogonality, dot product, norm

Definition (norm, dot product, Hilbert space)

$\mathcal{S}$ a vector space.

- $\|.\| : \mathcal{S} \to \mathbb{R}^+$ is a norm iff
  1. $\|x\| = 0 \Leftrightarrow x = 0$
  2. $\lambda \in \mathbb{R},\ x \in \mathcal{S},\ \|\lambda x\| = |\lambda|\|x\|$
  3. $x, y \in \mathcal{S},\ \|x + y\| \leq \|x\| + \|y\|$
- a dot product is a bilinear symmetric application of $\mathcal{S}^2$ to $\mathbb{R}$.
  - then $x \to \sqrt{\langle x, x\rangle}$ is a norm.
  - x and y are orthogonal when $\langle x, y\rangle = 0$.
  - F has a unique orthogonal supplementary $F^\perp$.
  - For any x, the unique decomposition $x = x_F + x_{F^\perp}$ also verifies : $\|x\|^2 = \|x_F\|^2 + \|x_{F^\perp}\|^2$.
- a Hilbert space $\mathcal{H}$ is a vector space endowed with a dot product $\langle ., .\rangle_{\mathcal{H}}$, that is closed for the induced norm.

Orthonormal bases

- A basis $\{e_i\}_{i=1..n}$ is orthonormal of $\mathcal{H}$ iff $\langle e_i, e_j \rangle_{\mathcal{H}} = \delta_{\{i=j\}}$.
  - $y \in \mathcal{H}$, the unique decomposition $y = \sum_{i=1..n} \lambda_i x_i$ verifies :
    1. $\lambda_i = \langle y, e_i \rangle_{\mathcal{H}}$
    2. $\|y\|^2_{\mathcal{H}} = \sum_i |\lambda_i|^2$

Example

- In $\mathbb{R}^d$:
  - $\{e_i = (0, \cdots, 0, \overset{\overset{i}{\uparrow}}{1}, 0, \cdots, 0)\}_{i=1..d}$ is a an orthonormal basis.
  - $y = (y_1, \cdots, y_d)^T = \sum_{i=1..d} y_i e_i$ and $\|y\| = \sqrt{\sum_{i=1..d} y_i^2}$.
- In $L^2([0, 2\pi])$:
  - $\{cos(mt), sin(mt)\}_{m \in \mathbb{N}}$ is an orthonormal basis.
  - $f \in L^2([0, 2\pi]), f(t) = \sum_{m \in \mathbb{N}} (a_m \cos(mt) + b_m \cos(mt))$

    where $a_m = \int_0^{2\pi} f(t) \cos(mt) dt$, $b_m = \int_0^{2\pi} f(t) \sin(mt) dt$.
  - $\|f\|^2_{L^2} = \int_0^{2\pi} |f(t)|^2 dt = \sum_{m \in \mathbb{N}} (|a_m|^2 + |b_m|^2)$.

Hyperplanes

H a hyperplane then $\dim F^{\perp} = 1$ hence there is a vector $u \in \mathcal{H}$ such that :

$$F^{\perp} = \text{Span}\,\{u\} = \mathbb{R}u \quad \text{and} \quad \|u\|_{\mathcal{H}} = 1.$$

- Equation of H : $H = \{x \in \mathcal{H} : \langle x, u \rangle_{\mathcal{H}} = 0\}$.

$$H = \{x = (x_1, x_2)^T : x_1 u_1 + x_2 u_2 = 0\}$$

- The distance from x to H is : $d(x, H) = |\,\langle x, u \rangle_{\mathcal{H}}\,|$.

$$d(x, H) = |x_1 u_1 + x_2 u_2|$$

- The projection of x on H is : $P_H(x) = x - \langle x, u \rangle_{\mathcal{H}}\, u$.

$$P_H(x) = x - (x_1 u_1 + x_2 u_2)u$$

Hyperplanes

H a hyperplane then $\dim F^\perp = 1$ hence there is a vector $u \in \mathcal{H}$ such that :

$$F^\perp = \text{Span}\,\{u\} = \mathbb{R}u \quad \text{and} \quad \|u\|_{\mathcal{H}} = 1.$$

- Equation of H : $H = \{x \in \mathcal{H} : \langle x, u\rangle_{\mathcal{H}} = 0\}$.
  $$H = \{x = (x_1, x_2)^T : x_1 u_1 + x_2 u_2 = 0\}$$
- The distance from x to H is : $d(x, H) = |\,\langle x, u\rangle_{\mathcal{H}}\,|$.
  $$d(x, H) = |x_1 u_1 + x_2 u_2|$$
- The projection of x on H is : $P_H(x) = x - \langle x, u\rangle_{\mathcal{H}}\, u$.
  $$P_H(x) = x - (x_1 u_1 + x_2 u_2) u$$

Matrices

- Let $H_1 = \mathbb{R}u_1{}^{\perp}, H_2 = \mathbb{R}u_2{}^{\perp}, \cdots, H_m = \mathbb{R}u_m{}^{\perp}$ be m hyperplanes of $\mathbb{R}^d$ and $F = \bigcap_{i=1}^m H_i$.
- The equation of F is a system of m linear equations with d unknowns :

$$\left\{\begin{array}{cccc} u_1^1 x_1 + u_1^2 x_2 + & \cdots & u_1^d x_d & = 0 \\ u_2^1 x_1 + u_2^2 x_2 + & \cdots & u_2^d x_d & = 0 \\ \vdots & \ddots & \vdots & \vdots \\ u_m^1 x_1 + u_m^2 x_2 + & \cdots & u_m^d x_d & = 0 \end{array}\right.$$

which is equivalent to the matrix-vector equation :

$$Ux = 0 \Leftrightarrow \begin{pmatrix} u_1^1 & u_1^2 & \cdots & u_1^d \\ u_2^1 & u_2^2 & \cdots & u_2^d \\ \vdots & \vdots & \ddots & \vdots \\ u_m^1 & u_m^2 & \cdots & u_m^d \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Matrices

- Let $H_1 = \mathbb{R}{u_1}^{\perp}$, $H_2 = \mathbb{R}{u_2}^{\perp}, \cdots, H_m = \mathbb{R}{u_m}^{\perp}$ be m hyperplanes of $\mathbb{R}^d$ and $F = \bigcap_{i=1}^{m} H_i$.
- The equation of F is a system of m linear equations with d unknowns :

$$\left\{\begin{array}{cccc} u_1^1 x_1 + u_1^2 x_2 + & \cdots & u_1^d x_d & = b_1 \\ u_2^1 x_1 + u_2^2 x_2 + & \cdots & u_2^d x_d & = b_1 \\ \vdots & \ddots & \vdots & \vdots \\ u_m^1 x_1 + u_m^2 x_2 + & \cdots & u_m^d x_d & = b_m \end{array}\right.$$

which is equivalent to the matrix-vector equation :

$$Ux = b \Leftrightarrow \begin{pmatrix} u_1^1 & u_1^2 & \cdots & u_1^d \\ u_2^1 & u_2^2 & \cdots & u_2^d \\ \vdots & \vdots & \ddots & \vdots \\ u_m^1 & u_m^2 & \cdots & u_m^d \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

Matrices

- A matrix in $\mathbb{R}^{m\times d}$ is a an array made of m row-vectors of $\mathbb{R}^d$ or equiv. d column vectors of $\mathbb{R}^m$ (e.g. U).
- The matrix-vector product Ux may be seen as :
  1. Using column vectors $U^j = (u_1^j, u_2^j, \cdots, u_m^j)^T$:

$$Ux = \sum_{j=1}^{d} x_j U^j, \quad \text{where} U^j \in \mathbb{R}^m.$$

  2. Using row vectors $U_i = (u_i^1, u_i^2, \cdots, u_i^d)$:

$$Ux = \begin{pmatrix} \langle U_1^T, x\rangle_{\mathbb{R}^d} \\ \langle U_2^T, x\rangle_{\mathbb{R}^d} \\ \vdots \\ \langle U_m^T, x\rangle_{\mathbb{R}^d} \end{pmatrix} \in \mathbb{R}^m$$

Note : U is a representation of a linear operator : $x \in \mathbb{R}^d \to Ux \in \mathbb{R}^m$.

Matrices

- Notation :

$$A = \in \mathbb{R}^{m\times d} = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,d} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,d} \end{pmatrix} = (a_{i,j})_{\substack{i=1\cdots m \\ j=1\cdots d}}$$

- Operations on matrices :
  - $\mathbb{R}^{m\times d}$ is a real vector space with $A + B = (a_{i,j} + b_{i,j})_{\substack{i=1\cdots m \\ j=1\cdots d}}$
  - Matrix product : $A \in \mathbb{R}^{m\times p}$, $B \in \mathbb{R}^{p\times d}$, then :

$$AB \in \mathbb{R}^{m\times d} \quad \text{s.t.} \quad (AB)_{i,j} = \sum_{k=1}^{p} a_{i,k} b_{k,j}$$

    Note : $AB \neq BA$!
  - Matrix transposition : $A \in \mathbb{R}^{m\times d}$, then :

$$A^T \in \mathbb{R}^{d\times m} = (a_{j,i})_{\substack{j=1\cdots d \\ i=1\cdots m}}$$

Square matrices (m=d)

- Matrix product is stable in $\mathbb{R}^{d\times d}$, so some are invertible !
- Remarquable matrices
  - Diagonal matrices.

$$D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_d \end{pmatrix}$$

  - Upper and Lower triangular matrices :

$$U = \begin{pmatrix} u_{1,1} & u_{1,2} & \cdots & u_{1,d} \\ 0 & u_{2,2} & \cdots & u_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{d,d} \end{pmatrix} \qquad L = \begin{pmatrix} l_{1,1} & 0 & \cdots & 0 \\ l_{2,1} & l_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{d,1} & l_{d,2} & \cdots & l_{d,d} \end{pmatrix}$$

  - Symmetric matrices : $A = A^T$.
  - Unitary matrices : $AA^T = A^TA = I$ (matrix of an orthonormal basis).

Inverting a matrix

- A is diagonal, lower or upper triangular then :
$$A \text{ invertible} \Leftrightarrow \prod_{i=1}^{d} a_{i,i} \neq 0$$
- Lower triangular systems

$$Ax = b \Leftrightarrow \begin{pmatrix} a_{1,1} & 0 & \cdots & 0 \\ a_{2,1} & a_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{d,1} & a_{d,2} & \cdots & a_{d,d} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{pmatrix} \text{ wh. } \prod_{i=1}^{d} a_{i,i} \neq 0$$

are solved recursively from the first to the last equation :

$$\left\{ \begin{array}{rcl} a_{1,1}x_1 & = & b_1 \\ a_{2,2}x_2 + a_{2,1}x_1 & = & b_1 \\ a_{3,3}x_3 + a_{3,2}x_2 + a_{3,1}x_d & = & b_2 \\ \vdots & & \vdots \\ a_{1,1}x_1 \; + \; a_{1,2}x_2 \; + \cdots\cdots\cdots + \; a_{1,d}x_d & = & b_d \end{array} \right.$$

Inverting a matrix

- A is diagonal, lower or upper triangular then :
$$A \text{ invertible} \Leftrightarrow \prod_{i=1}^{d} a_{i,i} \neq 0$$
- Lower triangular systems

$$Ax = b \Leftrightarrow \begin{pmatrix} a_{1,1} & 0 & \cdots & 0 \\ a_{2,1} & a_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{d,1} & a_{d,2} & \cdots & a_{d,d} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{pmatrix} \text{ wh. } \prod_{i=1}^{d} a_{i,i} \neq 0$$

are solved recursively from the first to the last equation :

$$\left\{ \begin{array}{rcl} x_1 & = & b_1 / a_{1,1} \\ a_{2,1} x_1 + a_{2,2} x_2 & = & b_2 \\ a_{3,1} x_1 + a_{3,2} x_2 + a_{3,3} x_3 & = & b_3 \\ \vdots & & \vdots \\ a_{1,1} x_1 \; + \; a_{1,2} x_2 \; + \cdots\cdots\cdots + \; a_{1,d} x_d & = & b_d \end{array} \right.$$

Inverting a matrix

- A is diagonal, lower or upper triangular then :

$$A \text{ invertible} \Leftrightarrow \prod_{i=1}^{d} a_{i,i} \neq 0$$

- Lower triangular systems

$$Ax = b \Leftrightarrow \begin{pmatrix} a_{1,1} & 0 & \cdots & 0 \\ a_{2,1} & a_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{d,1} & a_{d,2} & \cdots & a_{d,d} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{pmatrix} \text{ wh. } \prod_{i=1}^{d} a_{i,i} \neq 0$$

are solved recursively from the first to the last equation :

$$\left\{ \begin{array}{rcl} \color{red}{x_1} & \color{red}{=} & \color{red}{b_1 / a_{1,1}} \\ a_{2,1}\color{red}{x_1} + a_{2,2}x_2 & = & b_2 \\ a_{3,1}x_1 + a_{3,2}x_2 + a_{3,3}x_3 & = & b_3 \\ \vdots & & \vdots \\ a_{1,1}x_1 + a_{1,2}x_2 + \cdots\cdots\cdots + a_{1,d}x_d & = & b_d \end{array} \right.$$

Inverting a matrix

- A is diagonal, lower or upper triangular then :

$$A \text{ invertible} \Leftrightarrow \prod_{i=1}^{d} a_{i,i} \neq 0$$

- Lower triangular systems

$$Ax = b \Leftrightarrow \begin{pmatrix} a_{1,1} & 0 & \cdots & 0 \\ a_{2,1} & a_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{d,1} & a_{d,2} & \cdots & a_{d,d} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_d \end{pmatrix} \text{ wh. } \prod_{i=1}^{d} a_{i,i} \neq 0$$

are solved recursively from the first to the last equation :

$$\begin{cases} x_1 = b_1 / a_{1,1} \\ x_2 = \left(b_2 - a_{2,1} b_1 / a_{1,1}\right) / a_{2,2} \\ a_{3,1}x_1 + a_{3,2}x_2 + a_{3,3}x_3 = b_3 \\ \vdots \quad \vdots \\ a_{1,1}x_1 + a_{1,2}x_2 + \cdots\cdots + a_{1,d}x_d = b_d \end{cases}$$

Matrix determinant

- $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ is invertible iff $ad - bc \neq 0$ and $A^{-1} = \frac{1}{ad-bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$
- For lower/upper triangular and diagonal matrices :
  $$A \text{ is invertible iff } \prod_{i=1}^{d} a_{i,i} \neq 0.$$
- In general, $A \in \mathbb{R}^{d \times d}$ is invertible

  $\Leftrightarrow$ its d row (resp. column) vectors are linearly independent.

  $\Leftrightarrow$ its determinant $det(A) = \begin{vmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,d} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ a_{d,1} & a_{d,2} & \cdots & a_{d,d} \end{vmatrix} \neq 0.$
- The determinant is found recursively, developping on any row or column : $det(A) = \sum_{i=1}^{d} a_{i,j} Cof(A)_{i,j}$.
  - $Cof(A)_{i,j} = det((a_{k,l})_{k \in \{1 \cdots d\} \setminus \{i\}, l \in \{1 \cdots d\} \setminus \{j\}})$
  - if $det(A) \neq 0$ then $A^{-1} = \frac{1}{det(A)} Cof(A)^T$.

Eigenvalues, eigenvectors

A a square matrix.

Definition (Eigenvalues and eigenvectors)

- λ is an eigenvalue of A if there exists a vector $v \in \mathbb{R}^d$, $v \neq 0$ s.t. $Av = \lambda v$.
- Equivalently : λ is an eigenvalue of A if $det(A - \lambda I) = 0$.
- Any v verifying $Av = \lambda v$ is an eigenvector associated to the eigenvalue λ.

- Properties :
  - For diagonal matrices, the eigenvalues are the diagonal elements (not for triangular matrices !).
  - 0 is an eigenvalue iff A is not invertible.
- A is diagonalizable if there exists a basis of eigenvectors :

$$A = PDP^{-1} \text{ with } D \text{ diagonal.}$$

Singular value decomposition

Symmetric matrices and eigenvalues/eigenvectors :

- A symmetric matrix is diagonalizable on an orthonormal basis :

$$A = PDP^T \text{ with } D \text{ diagonal, } PP^T = I.$$

- A symmetric matrix is said
  - semi-definite positive if $\langle x, Ax \rangle \geq 0, \ \forall x$.
    Its eigenvalues are ≥ 0.

    Any diagonal matrix,
    $A = B^T B$ for any $B \in \mathbb{R}^{m,d}$.
  - definite positive if $\langle x, Ax \rangle \geq 0, \ \forall x$ and $\langle x, Ax \rangle = 0, \Rightarrow x = 0$.
    Its eigenvalues are > 0.

    Any diagonal matrix without zeros,
    $A = B^T B$ for any $B \in \mathbb{R}^{m,d}$ when A is invertible.

Note : a definite positive matrix defines a new norm on $\mathbb{R}^d$ via the scalar product $\langle x, x \rangle_A = \langle x, Ax \rangle$

Singular value decomposition

Symmetric matrices and eigenvalues/eigenvectors :

- A symmetric matrix is diagonalizable on an orthonormal basis :

$$A = PDP^T \text{ with } D \text{ diagonal, } PP^T = I.$$

- A symmetric matrix is said
  - semi-definite positive if $\langle x, Ax \rangle \geq 0, \ \forall x$.
    Its eigenvalues are ≥ 0.

    Any diagonal matrix,
    $A = B^T B$ for any $B \in \mathbb{R}^{m,d}$.
  - definite positive if $\langle x, Ax \rangle \geq 0, \ \forall x$ and $\langle x, Ax \rangle = 0, \Rightarrow x = 0$.
    Its eigenvalues are > 0.

    Any diagonal matrix without zeros,
    $A = B^T B$ for any $B \in \mathbb{R}^{m,d}$ when A is invertible.

Note : a definite positive matrix defines a new norm on $\mathbb{R}^d$ via the scalar product $\langle x, x \rangle_A = \langle x, Ax \rangle$

Singular value decomposition

Symmetric matrices and eigenvalues/eigenvectors :

- A symmetric matrix is diagonalizable on an orthonormal basis :
$$A = PDP^T \text{ with } D \text{ diagonal, } PP^T = I.$$
- A symmetric matrix is said
  - semi-definite positive if $\langle x, Ax\rangle \geq 0, \ \forall x$.
    Its eigenvalues are ≥ 0.

    Any diagonal matrix,
    $A = B^T B$ for any $B \in \mathbb{R}^{m,d}$.
  - definite positive if $\langle x, Ax\rangle \geq 0, \ \forall x$ and $\langle x, Ax\rangle = 0, \Rightarrow x = 0$.
    Its eigenvalues are > 0.

    Any diagonal matrix without zeros,
    $A = B^T B$ for any $B \in \mathbb{R}^{m,d}$ when A is invertible.

Note : a definite positive matrix defines a new norm on $\mathbb{R}^d$ via the scalar product $\langle x, x\rangle_A = \langle x, Ax\rangle$

Singular value decomposition

Symmetric matrices and eigenvalues/eigenvectors :

- A symmetric matrix is diagonalizable on an orthonormal basis :

$$A = PDP^T \text{ with } D \text{ diagonal, } PP^T = I.$$

- A symmetric matrix is said
  - semi-definite positive if $\langle x, Ax \rangle \geq 0, \ \forall x$.
    Its eigenvalues are ≥ 0.

    Any diagonal matrix,
    $A = B^T B$ for any $B \in \mathbb{R}^{m,d}$.

  - definite positive if $\langle x, Ax \rangle \geq 0, \ \forall x$ and $\langle x, Ax \rangle = 0, \Rightarrow x = 0$.
    Its eigenvalues are > 0.

    Any diagonal matrix without zeros,
    $A = B^T B$ for any $B \in \mathbb{R}^{m,d}$ when A is invertible.

Note : a definite positive matrix defines a new norm on $\mathbb{R}^d$ via the scalar product $\langle x, x \rangle_A = \langle x, Ax \rangle$

Singular value decomposition

Fix $B \in \mathbb{R}^{m\times d}$, note that :

- $B^T B \in \mathbb{R}^{d\times d}$ and $BB^T \in \mathbb{R}^{m\times m}$ are symmetric semi-definite positive :
  - $B^T B = V\Delta_1 V^T$ with Δ_1 diagonal, $VV^T = I$ in $\mathbb{R}^{d\times d}$.
  - $BB^T = U\Delta_2 U^T$ with Δ_2 diagonal, $UU^T = I$ in $\mathbb{R}^{m\times m}$.
- One can show :
  - Δ_1 and Δ_2 have the same non-zero values $\lambda_1^2, \cdots, \lambda_k^2$.
  - $B = UDV^T$ with

$$D = diag(\lambda_1, \cdots, \lambda_k) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 & 0\cdots\cdots 0 \\ 0 & \lambda_2 & \cdots & 0 & 0\cdots\cdots 0 \\ \vdots & \vdots & \ddots & \vdots & 0\cdots\cdots 0 \\ 0 & 0 & \cdots & \lambda_k & 0\cdots\cdots 0 \\ 0 & 0 & \cdots & 0 & 0\cdots\cdots 0 \end{pmatrix} \in \mathbb{R}^{m,d}.$$

  - $B^T = VDU^T$ with $D = diag(l_1, \cdots, \lambda_k) = \in \mathbb{R}^{d,m}$.
- $B = UDV^T$ is its singular value decomposition and $\lambda_1, \cdots, \lambda_k$ its singular values.

Singular value decomposition

Fix $B \in \mathbb{R}^{m\times d}$, note that :

- $B^T B \in \mathbb{R}^{d\times d}$ and $BB^T \in \mathbb{R}^{m\times m}$ are symmetric semi-definite positive :
  - $B^T B = V\Delta_1 V^T$ with Δ_1 diagonal, $VV^T = I$ in $\mathbb{R}^{d\times d}$.
  - $BB^T = U\Delta_2 U^T$ with Δ_2 diagonal, $UU^T = I$ in $\mathbb{R}^{m\times m}$.
- One can show :
  - Δ_1 and Δ_2 have the same non-zero values $\lambda_1^2, \cdots, \lambda_k^2$.
  - $B = UDV^T$ with

$$D = diag(\lambda_1, \cdots, \lambda_k) = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 & 0\cdots\cdots 0 \\ 0 & \lambda_2 & \cdots & 0 & 0\cdots\cdots 0 \\ \vdots & \vdots & \ddots & \vdots & 0\cdots\cdots 0 \\ 0 & 0 & \cdots & \lambda_k & 0\cdots\cdots 0 \\ 0 & 0 & \cdots & 0 & 0\cdots\cdots 0 \end{pmatrix} \in \mathbb{R}^{m,d}.$$

  - $B^T = VDU^T$ with $D = diag(l_1, \cdots, \lambda_k) = \in \mathbb{R}^{d,m}$.
- $B = UDV^T$ is its singular value decomposition and $\lambda_1, \cdots, \lambda_k$ its singular values.

Other decompositions

- LU factorization
  - for a diagonally dominant matrix A ($|a_{i,i}| \geq \sum j \neq i |a_{i,j}|$)
  - $A = LU$, L is lower triangular, U is upper triangular with 1 on the diagonal.
  - $Ax = B$ solved in two steps : $Lz = b$ and $Ux = z$!
- Choleski decomposition
  - for symmetric semi-definite positive matrices
  - $A = U^T U$ with U upper triangular
  - again easy to solve $Ax = b$ in two steps.
- QR decomposition
  - for any matrix $A \in \mathbb{R}^{m \times d}$
  - $A = QR$ with Q unitary in $\mathbb{R}^{m \times m}$ and R upper triangular.

Framework

- Random Space
  - Ω is the set of random events.

    $$\Omega = \{heads, tails\}$$

  - $\mathcal{A}$ is the set of "measurable" collections of events.

    $$\mathcal{A} = \{\emptyset, \{heads\}, \{tails\}, \{heads, tails\}\}$$

  - $\mathbb{P} : \mathcal{A} \to [0, 1]$ is the probability.

    $$\mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(\{heads\}) = p,$$
    $$\mathbb{P}(\{tails\}) = 1 - p, \quad \mathbb{P}(\{heads, tails\}) = 1$$

- Properties of $\mathbb{P}$
  - $0 \leq \mathbb{P} \leq 1$,
  - $\mathbb{P}(\emptyset) = 0, \ \mathbb{P}(\Omega) = 1$,
  - $A, B \in \mathcal{A}$, $A \cup B = \emptyset \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ (chain rule).
  - Equivalently : $A, B \in \mathcal{A}$, $\mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B)$.

- Random events are observed only through measurable quantities called Random variables.

Framework

- Random Space
  - Ω is the set of random events.
    $$\Omega = \{heads, tails\}$$
  - $\mathcal{A}$ is the set of "measurable" collections of events.
    $$\mathcal{A} = \{\emptyset, \{heads\}, \{tails\}, \{heads, tails\}\}$$
  - $\mathbb{P} : \mathcal{A} \to [0,1]$ is the probability.
    $$\mathbb{P}(\emptyset) = 0, \quad \mathbb{P}(\{heads\}) = p,$$
    $$\mathbb{P}(\{tails\}) = 1 - p, \quad \mathbb{P}(\{heads, tails\}) = 1$$
- Properties of $\mathbb{P}$
  - $0 \leq \mathbb{P} \leq 1$,
  - $\mathbb{P}(\emptyset) = 0, \ \mathbb{P}(\Omega) = 1$,
  - $A, B \in \mathcal{A}, A \cup B = \emptyset \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ (chain rule).
  - Equivalently : $A, B \in \mathcal{A}, \mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B)$.
- Random events are observed only through measurable quantities called Random variables.

Random variables

- A Random variable is a measurable function $X : (\Omega, \mathcal{A}) \rightarrow (\mathcal{F}, \mathcal{B}(\mathcal{F}))$
  $\hookrightarrow$ the measurability means $F \subset \mathcal{F} \Rightarrow X^{-1}(F) \subset \mathcal{A}$.

- $X(\Omega) \subset \mathcal{F}$ may be
  - finite ($\{0, 1\}$) or infinite ($\mathbb{R}$), discrete ($\mathbb{N}$) or continuous($\mathbb{R}$)

    discrete/continuous random variables
  - have one or several variables ($\mathbb{R}^d$)

    random variables/ random vectors.

- The measurability of X implies that $\mathbb{P}$ may be transported to $\mathcal{F}$ through X :

$$\mathbb{P}(\{\omega / X(\omega) \in F\}) = \mathbb{P}(X \in F) \overset{def}{=} \mathbb{P}_X(F)$$

$\mathbb{P}$ is a probability on $(\Omega, \mathcal{A})$
$\mathbb{P}_X$ is a probability on $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$.

Random variables

- A Random variable is a measurable function $X : (\Omega, \mathcal{A}) \rightarrow (\mathcal{F}, \mathcal{B}(\mathcal{F}))$
  $\hookrightarrow$ the measurability means $F \subset \mathcal{F} \Rightarrow X^{-1}(F) \subset \mathcal{A}$.

- $X(\Omega) \subset \mathcal{F}$ may be
  - finite ($\{0, 1\}$) or infinite ($\mathbb{R}$), discrete ($\mathbb{N}$) or continuous($\mathbb{R}$)
    discrete/continuous random variables
  - have one or several variables ($\mathbb{R}^d$)
    random variables/ random vectors.

- The measurability of X implies that $\mathbb{P}$ may be transported to $\mathcal{F}$ through X :

$$\mathbb{P}(\{\omega / X(\omega) \in F\}) = \mathbb{P}(X \in F) \overset{def}{=} \mathbb{P}_X(F)$$

$\mathbb{P}$ is a probability on $(\Omega, \mathcal{A})$
$\mathbb{P}_X$ is a probability on $(\mathcal{F}, \mathcal{B}(\mathcal{F}))$.

Discrete random variables

Examples

- A single coin toss is a Bernoulli variable with parameter p
  - $X : (\Omega, \mathcal{A}) \to (\{0, 1\}, 2^{\{0,1\}})$,
  - $\mathbb{P}(X = 1) = p$, (hence $\mathbb{P}(X = 0) = p$).
  - Notation : $X \sim B(p)$.

- The sum of n independent coin tosses is a multinomial with parameter n, p
  - $Y : (\Omega, \mathcal{A}) \to (\{0, 1, \cdots, n\}, 2^{\{0,1,\cdots,n\}})$,
  - $Y = X_1 + X_2 + \cdots + X_n$ where the X_i are independent copies $\equiv B(p)$.
  - $\mathbb{P}(Y = k) = \binom{n}{k} p^k (1 - p)^{n-k}$ for $k = 0 \cdots n$.
  - Notation : $Y \sim Bin(n, p)$.

Discrete random variables

- $\mathcal{F}$ is discrete $\mathcal{F} = \{x_1, x_2, \cdots, x_N\}$, N finite or not.
- $X : (\Omega, \mathcal{A}) \to (\mathcal{F}, 2^{\mathcal{F}})$,
  - Notation : $\mathbb{P}(X = x_i) = p_i$. *Note that $p_i \geq 0$ and $\sum_{i=1}^N p_i = 1$.*
- The mean value or expectation of X is :

$$\begin{aligned} \mathbb{E}[X] &= \textstyle\sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega) \\ \mathbb{E}[X] &= \textstyle\sum_{i=1}^N x_i \mathbb{P}_X(x_i) \end{aligned}$$

$$\text{Here,} \quad \mathbb{E}[X] = \textstyle\sum_{i=1}^N x_i p_i$$

- The variance of X is its deviation from its mean :

$$\begin{aligned} Var[X] &= \mathbb{E}[(X - E[X])^2] \\ Var[X] &= \mathbb{E}[X^2] - E[X]^2 \end{aligned}$$

$$\text{Here,} \quad Var[X] = \textstyle\sum_{i=1}^N x_i^2 p_i - (\sum_{i=1}^N x_i p_i)^2.$$

Discrete random variables

- $\mathcal{F}$ is discrete $\mathcal{F} = \{x_1, x_2, \cdots, x_N\}$, N finite or not.
- $X : (\Omega, \mathcal{A}) \to (\mathcal{F}, 2^{\mathcal{F}})$,
  - Notation : $\mathbb{P}(X = x_i) = p_i$. *Note that* $p_i \geq 0$ *and* $\sum_{i=1}^{N} p_i = 1$.
- The mean value or expectation of X is :

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega)$$
$$\mathbb{E}[X] = \sum_{i=1}^{N} x_i \mathbb{P}_X(x_i)$$

Here, $\mathbb{E}[X] = \sum_{i=1}^{N} x_i p_i$

- The variance of X is its deviation from its mean :

$$Var[X] = \mathbb{E}[(X - E[X])^2]$$
$$Var[X] = \mathbb{E}[X^2] - E[X]^2$$

Here, $Var[X] = \sum_{i=1}^{N} x_i^2 p_i - (\sum_{i=1}^{N} x_i p_i)^2$.

Discrete random variables

- $\mathcal{F}$ is discrete $\mathcal{F} = \{x_1, x_2, \cdots, x_N\}$, N finite or not.
- $X : (\Omega, \mathcal{A}) \to (\mathcal{F}, 2^{\mathcal{F}})$,
  - Notation : $\mathbb{P}(X = x_i) = p_i$. *Note that $p_i \geq 0$ and $\sum_{i=1}^{N} p_i = 1$.*
- The mean value or expectation of X is :

$$\begin{aligned}\mathbb{E}[X] &= \textstyle\sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega)\\ \mathbb{E}[X] &= \textstyle\sum_{i=1}^{N} x_i \mathbb{P}_X(x_i)\end{aligned}$$

Here, $\mathbb{E}[X] = \sum_{i=1}^{N} x_i p_i$

- The variance of X is its deviation from its mean :

$$\begin{aligned}Var[X] &= \mathbb{E}[(X - E[X])^2]\\ Var[X] &= \mathbb{E}[X^2] - E[X]^2\end{aligned}$$

Here, $Var[X] = \sum_{i=1}^{N} x_i^2 p_i - (\sum_{i=1}^{N} x_i p_i)^2.$

Discrete random variables

- More generally for any measurable function $f : \mathcal{F} \to \mathbb{R}^d$, the expectation of $f(X)$ is :

$$
\begin{aligned}
\mathbb{E}[f(X)] &= \textstyle\sum_{\omega \in \Omega} f(x)\mathbb{P}(X(\omega) = x) \\
\mathbb{E}[f(X)] &= \textstyle\sum_{i=1}^{N} f(x_i)\mathbb{P}_X(x_i)
\end{aligned}
$$

Here, $\mathbb{E}[f(X)] = \sum_{i=1}^{N} f(x_i)p_i$

Bernoulli variables

- $X \sim B(p)$, hence
  $\mathcal{F} = \{0, 1\}$, $p_1 = p$, $p_0 = 1 - p$.
- The expectation of X is :

$$\begin{aligned}
\mathbb{E}[X] &= \textstyle\sum_{i=1}^{N} x_i p_i \\
\mathbb{E}[X] &= 0 * (1 - p) + 1 * p \\
\mathbb{E}[X] &= p
\end{aligned}$$

- The variance of X is :

$$\begin{aligned}
Var[X] &= \textstyle\sum_{i=1}^{N} x_i^2 p_i - (\sum_{i=1}^{N} x_i p_i)^2 \\
Var[X] &= 0^2(1 - p) + 1^2 * p - p^2 \\
Var[X] &= p(1 - p).
\end{aligned}$$

- The expectation of $f(X)$ is :

$$\begin{aligned}
\mathbb{E}[f(X)] &= \textstyle\sum_{i=1}^{N} f(x_i) p_i \\
\mathbb{E}[f(X)] &= f(0) * (1 - p) + f(1) * p.
\end{aligned}$$

Discrete random vectors

- X has d coordinates, each of which is a discrete variable.
$$X = (X_1, \cdots, X_d)^T : (\Omega, \mathcal{A}) \to (\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_d, 2^{\mathcal{F}}),$$
- $\mathbb{P}(X = x_i) = p_i \Leftrightarrow \mathbb{P}(X = (x^1, \cdots, x^d))$, where $x^i \in \mathcal{F}_i$.
- The expectation of X is the vector of the expectation of each coordinate :
$$\mathbb{E}[X] = (\mathbb{E}[X_1], \cdots, \underset{\substack{\uparrow \\ \text{row i}}}{\mathbb{E}[X_i]}, \cdots \mathbb{E}[X_d])^T$$
- The variance is replaced by the covariance matrix :
  - $\text{Cov}(X)$ is a $d \times d$-matrix.
  - $\text{Cov}(X)_{i,i} = \text{Var}(X_i)$.
  - If $i \neq j$, $\text{Cov}(X)_{i,j} = \text{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$.

Discrete random vectors

Example

- $X = (X_1, X_2)$ with
  - $X_1 \sim B(p_1)$,
  - $X_1 \sim B(p_2)$,
  - X_1 and X_2 are decorrelated i.e. $\mathrm{Cov}(X_1, X_2) = 0$.
- The expectation of X is :

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$$

- The covariance matrix of X is :

$$\mathrm{Cov}[X] = \begin{pmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}[X_1, X_2] \\ \mathrm{Cov}[X_2, X_1] & \mathrm{Var}[X_2] \end{pmatrix} = \begin{pmatrix} p_1(1-p_1) & 0 \\ 0 & p_2(1-p_2) \end{pmatrix}$$

Note : independence $\Rightarrow$ decorrelation but the inverse is false !

- $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.
- $\mathbb{P}(X = x_i) = p_i \leftrightarrow \mathbb{P}(X \in [a, b]) = P_X([a, b])$.
  Note : $P_X \geq 0$ and $\int_{\mathbb{R}} dP_X(x) = 1$.
- The expectations and variances are defined as previsouly :

$$
\begin{aligned}
\mathbb{E}[X] &= \int_{\Omega} X(\omega) d\mathbb{P}(\omega) \\
\mathbb{E}[X] &= \int_{\mathbb{R}} x d\mathbb{P}_X(x)
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[f(X)] &= \int_{\Omega} f(X(\omega)) d\mathbb{P}(\omega) \\
\mathbb{E}[f(X)] &= \int_{\mathbb{R}} f(x) d\mathbb{P}_X(x)
\end{aligned}
$$

$$
\mathbb{E}[\mathrm{Var}(X)] = \mathbb{E}[X^2] - E[X]^2
$$

- If $d\mathbb{P}_X(x) = f_X(x)dx$ then f_X is the probability density function of X (pdf).

Continuous random variables

Uniform distribution on $[a, b]$

- $X \sim \mathcal{U}_{[a,b]}$
- $\mathbb{E}[f(X)] = \int_{\mathbb{R}} f(x) dP_X(x) = \frac{1}{b-a} \int_{[a,b]} f(x) dx$
- pdf : $f_X(x) = \frac{1}{b-a} \delta_{[a,b]}(x)$

Gaussian distribution

of mean m and variance σ^2 :

- $X \sim \mathcal{N}_{m,\sigma^2}$
- $\mathbb{E}[f(X)] = \int_{\mathbb{R}} f(x) dP_X(x) = \int_{\mathbb{R}} f(x) * \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-m)^2}{2\sigma^2(x)} dx$
- pdf : $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{(x-m)^2}{2\sigma^2(x)}$

Continuous random variables

All we have seen previously extends to continuous random vectors such as :

Gaussian vector of mean $\mathbf{m}$ and covariance matrix Σ^2 :

- $X = (X_1, \cdots, X_d) \sim \mathcal{N}_{\mathbf{m},\Sigma^2}$
- pdf : $f_X(x) = \frac{1}{(2\pi \det(\Sigma))^{d/2}} \exp\left\{-\frac{(x-\mathbf{m})^T\Sigma^{-1}(x-\mathbf{m})}{2}\right\}$

$$\begin{aligned}\mathbb{E}[f(X)] &= \int_{\mathbb{R}^d} f(x_1, \cdots, x_d) dP_X(x_1, \cdots, x_d) \\ &= \int_{\mathbb{R}^d} f(x) * \frac{1}{(2\pi \det(\Sigma))^{d/2}} \exp\left\{-\frac{(x-\mathbf{m})^T\Sigma^{-1}(x-\mathbf{m})}{2}\right\} dx\end{aligned}$$

Joint probabilities

Two simultaneaous coin tosses :

- Each coin is fair $\mathbb{P}(heads) = \frac{1}{2}$
- All the possible outcomes of both draws ($\{heads, heads\}$,$\{heads, tails\}$,$\{tails, heads\}$,$\{tails, tails\}$) are equiprobable with $\mathbb{P}(\{heads, heads\}) = \frac{1}{4}$.
- Consider $Z = (X_1, X_2)$, X_i the random variable for tossing coin i. This means that :

$$\mathbb{P}(Z \in A \times B) = \mathbb{P}(X_1 \in A)\mathbb{P}(X_2 \in B)$$

or in other words :

$$P_{(X_1,X_2)} = P_{X_1} P_{X_2}$$

X_1 and X_2 are independent.

Joint probabilities

But this is not always the case :

Example

X/Y	Sick (S)	Sane (A)	*Total*
Positive test (P)	90	100	*190*
Negative test (N)	10	900	*910*
Total	*100*	*1000*	*1100*

- $\mathbb{P}(X = positive) = 190/1100$
- $\mathbb{P}(Y = sick) = 100/1100$
- Clearly :

$$\mathbb{P}((X, Y) = (positive, sick)) = 90/1100$$

$$\neq$$

$$\mathbb{P}(X = positive)\mathbb{P}(Y = sick) = 100 * 190/1100^2$$

Joint probabilities

But this is not always the case :

Example

X/Y	Sick (S)	Sane (A)	*Total*
Positive test (P)	90	100	*190*
Negative test (N)	10	900	*910*
Total	*100*	*1000*	*1100*

- $\mathbb{P}(X = \textit{positive}) = 190/1100$
- $\mathbb{P}(Y = \textit{sick}) = 100/1100$
- Clearly :

$$\mathbb{P}((X, Y) = (\textit{positive}, \textit{sick})) = 90/1100$$

$$\neq$$

$$\mathbb{P}(X = \textit{positive})\mathbb{P}(Y = \textit{sick}) = 100 * 190/1100^2$$

Independence

Definition (Independence)

X and Y are independent random variables ($X \perp\!\!\!\perp Y$) if and only if their joint probability $\mathbb{P}_{X,Y}$ is the product of their marginal probabilities : $\mathbb{P}_{X,Y} = \mathbb{P}_X \mathbb{P}_Y$.

Also, X_1,..X_n are independent iff $\mathbb{P}_{X_1,\cdots,X_n} = \prod_{i=1}^n P_{X_i}$.

- Equivalently :
  - $\forall A, B \;\; \mathbb{P}((X, Y) \in A \times B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$
  - $\forall f, g \;\; \mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$
- If X and Y are independent then $\mathrm{Cov}(X, Y) = 0$.
- For Gaussian variables only : $\mathrm{Cov}(X, Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y$.

If X and Y are indepedent, knowing X does not give any information on Y, what if they are not independent ?

Independence

Definition (Independence)

X and Y are independent random variables ($X \perp\!\!\!\perp Y$) if and only if their joint probability $\mathbb{P}_{X,Y}$ is the product of their marginal probabilities : $\mathbb{P}_{X,Y} = \mathbb{P}_X \mathbb{P}_Y$.

Also, X_1,..X_n are independent iff $\mathbb{P}_{X_1,\cdots,X_n} = \prod_{i=1}^n P_{X_i}$.

- Equivalently :
  - $\forall A, B \;\; \mathbb{P}((X, Y) \in A \times B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B)$
  - $\forall f, g \;\; \mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$
- If X and Y are independent then $\mathrm{Cov}(X, Y) = 0$.
- For Gaussian variables only : $\mathrm{Cov}(X, Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y$.

If X and Y are indepedent, knowing X does not give any information on Y, what if they are not independent ?

Conditional probabilities

Example

X/Y	Sick (S)	Fit (F)	*Total*
Positive test (P)	90	100	*190*
Negative test (N)	10	900	*910*
Total	*100*	*1000*	*1100*

- Amongst all people :
  $\mathbb{P}(Y = sick) = 100/1100,$
  $\mathbb{P}(Y = fit) = 1000/1100$
- Amongst people with a positive test :
  $\mathbb{P}(Y = sick|X = positive) = 90/190,$
  $\mathbb{P}(Y = fit|X = positive) = 100/190,$
- Amongst people with a negative test :
  $\mathbb{P}(Y = sick|X = negative) = 10/910,$
  $\mathbb{P}(Y = fit|X = negative) = 900/910,$

Conditional probabilities

Example

X/Y	Sick (S)	Fit (F)	*Total*
Positive test (P)	90	100	*190*
Negative test (N)	10	900	*910*
Total	*100*	*1000*	*1100*

- Amongst all people :

  $\mathbb{P}(Y = \mathit{sick}) = 100/1100,$
  $\mathbb{P}(Y = \mathit{fit}) = 1000/1100$

- Amongst people with a positive test :

  $\mathbb{P}(Y = \mathit{sick}|X = \mathit{positive}) = 90/190,$
  $\mathbb{P}(Y = \mathit{fit}|X = \mathit{positive}) = 100/190,$

- Amongst people with a negative test :

  $\mathbb{P}(Y = \mathit{sick}|X = \mathit{negative}) = 10/910,$
  $\mathbb{P}(Y = \mathit{fit}|X = \mathit{negative}) = 900/910,$

Conditional probabilities

Example

X/Y	Sick (S)	Fit (F)	*Total*
Positive test (P)	90	100	*190*
Negative test (N)	10	900	*910*
Total	*100*	*1000*	*1100*

- Amongst people with a positive test :

$$\mathbb{P}(Y = sick | X = positive) = 90/190,$$
$$\mathbb{P}(Y = fit | X = positive) = 100/190,$$

- Note :
$\mathbb{P}(Y = sick | X = negative)\mathbb{P}(X = negative) = \mathbb{P}((Y, X) = (sick, negative)),$

Definition (Conditional probabilities)

$\mathbb{P}(A \text{ and } B) = \mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A)$

Conditional probabilities

More generally :

Definition

The conditional probability $\mathbb{P}_{X|Y}$ is the probability s.t. :

$$\forall f, \mathbb{E}[f(X,Y)] = \int f(X,Y)dP_{X,Y} = \int dP_Y \int f(X,Y)dP_{X|Y}$$

- For discrete random variables : $\mathbb{P}((X,Y)=(x,y)) = \mathbb{P}(Y=y|X=x)\mathbb{P}(X=x)$
- If (X,Y) and Y have pdf $p_{(X,Y)}$ and p_Y, then $P_{X|Y}$ is a the correspoding pdf : $p_{X|Y} = \frac{p_{(X,Y)}}{p_Y}$
- $\mathbb{E}[X|Y]$ is the conditional esperance of X given Y is a random variable. It is the projection of X on the set of rndom variables of the form $g(Y)$.

Bayes rule, maximum likelihood, maximum a posteriori

Framework :

- Y is a random variable, Y is observed
- Θ is a random variable, Θ is the parameter.
- Goal : given observed data Y, find the best guess for Θ.

Probabilities

- The conditional probability of the observations : $\mathbb{P}_{Y|\Theta}$.
- The prior : $\mathbb{P}_{\Theta}$.
- The posterior : $\mathbb{P}_{\Theta|Y}$.

Bayes rule

$$\mathbb{P}_{\Theta|Y}(\Theta, y) = \frac{\mathbb{P}_{Y|\Theta}(y,\theta)\mathbb{P}_{\Theta}(\theta)}{\int P_{Y|\Theta}(\theta',y)\mathbb{P}_{\Theta}(\theta')d\theta}$$

Estimator

- Maximum likelihood : $\theta_{ML} = \text{argmax}_{\theta}\, \mathbb{P}_{Y|\Theta}(y, \theta)$.
- Maximum a posteriori : $\theta_{MAP} = \text{argmax}_{\theta}\, \mathbb{P}_{\Theta|Y}(\theta, y)$.
- Bayes mean square estimator : $\theta_{M} = \mathbb{E}[\Theta|Y]$.

Information theory

- Entropy measures the amount of disorder of X :
  - $H(X) = -\int P_X(x)\log(P_X(x))dx$. Note : $H(X) \geq 0$.
  - For discrete random variables :
    - $X \sim \mathcal{U}$ maximizes the entropy $H = \log(N)$.
    - $X \sim \delta x_i$ minimizes the entropy $H = \frac{1}{N}\log(N)$.
- The Kullback-Leibler divergence compares the laws of X and Y :
  - $D(X||Y) = \int P_X(x)\log\left(\frac{P_X(x)}{P_Y(x)}\right)dx$. Note : $D(X||Y) \neq D(Y||X)$.
  - $D(X||Y) \geq 0$ and $[D(X||Y) = 0 \Leftrightarrow P_X = P_Y]$.
- The mutual information measures the amount of shared information between X and Y :
  - $I(X,Y) = D(P_{(X,Y)}||P_X P_Y)$. Note : $I(X,Y) = I(Y,X)$.
  - $I(X,Y) \geq 0$ and $[I(X,Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y]$.
- The perplexity is a measure of complexity of a distribution :
  - $P(X) = 2^{H(X)}$.
  - this is a common way of evaluating language models.

Approximations and confidence intervals

- Statistical learning (classification) :
  - Goal : from i.i.d[1] samples $(x_i, y_i)_{i=1\cdots n}$, find a hypothesis f that minimizes the risk : $\mathbb{E}[loss(f(X), Y)]$
  - $\mathbb{E}[loss(f(x), Y)]$ is not known, only its empirical version is accessible : $\frac{1}{n}\sum loss(f(x_i), y_i)$

$\hookrightarrow$ need to control how far is the empirical loss to the true one.

- Some tools to do so are :
  - Markov inequality : $\mathbb{P}(X > \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}$
  - Chebicheff inequality : $\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{Var}[X]}{\epsilon^2}$

    Apply this to $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, with X_i i.i.d X, one gets :

$$\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{Var}[X]}{n\epsilon^2}$$

    (S_n is the empirical risk, $\mathbb{E}[X]$ the true one.)
  - Chernoff-Hoeffding bound : $\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq e^{-2n\epsilon^2}$

[1] independent identically distributed

Approximations and confidence intervals

- Statistical learning (classification) :
  - Goal : from i.i.d[1] samples $(x_i, y_i)_{i=1\cdots n}$, find a hypothesis f that minimizes the risk : $\mathbb{E}[loss(f(X), Y)]$
  - $\mathbb{E}[loss(f(x), Y)]$ is not known, only its empirical version is accessible : $\frac{1}{n}\sum loss(f(x_i), y_i)$

$\hookrightarrow$ need to control how far is the empirical loss to the true one.

- Some tools to do so are :
  - Markov inequality : $\mathbb{P}(X > \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}$
  - Chebicheff inequality : $\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{Var}[X]}{\epsilon^2}$

    Apply this to $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, with X_i i.i.d X, one gets :

$$\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{Var}[X]}{n\epsilon^2}$$

    (S_n is the empirical risk, $\mathbb{E}[X]$ the true one.)
  - Chernoff-Hoeffding bound : $\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq e^{-2n\epsilon^2}$

[1] independent identically distributed

Approximations and confidence intervals

- Statistical learning (classification) :
  - Goal : from i.i.d[1] samples $(x_i, y_i)_{i=1\cdots n}$, find a hypothesis f that minimizes the risk : $\mathbb{E}[loss(f(X), Y)]$
  - $\mathbb{E}[loss(f(x), Y)]$ is not known, only its empirical version is accessible : $\frac{1}{n}\sum loss(f(x_i), y_i)$

$\hookrightarrow$ need to control how far is the empirical loss to the true one.

- Some tools to do so are :
  - Markov inequality : $\mathbb{P}(X > \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}$
  - Chebicheff inequality : $\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{Var}[X]}{\epsilon^2}$

    Apply this to $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, with X_i i.i.d X, one gets :

    $$\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{Var}[X]}{n\epsilon^2}$$

    (S_n is the empirical risk, $\mathbb{E}[X]$ the true one.)
  - Chernoff-Hoeffding bound : $\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq e^{-2n\epsilon^2}$

[1] independent identically distributed

Approximations and confidence intervals

- Statistical learning (classification) :
  - Goal : from i.i.d[1] samples $(x_i, y_i)_{i=1\cdots n}$, find a hypothesis f that minimizes the risk : $\mathbb{E}[loss(f(X), Y)]$
  - $\mathbb{E}[loss(f(x), Y)]$ is not known, only its empirical version is accessible : $\frac{1}{n}\sum loss(f(x_i), y_i)$

$\hookrightarrow$ need to control how far is the empirical loss to the true one.

- Some tools to do so are :
  - Markov inequality : $\mathbb{P}(X > \epsilon) \leq \frac{\mathbb{E}[X]}{\epsilon}$
  - Chebicheff inequality : $\mathbb{P}(|X - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\mathrm{Var}[X]}{\epsilon^2}$

    Apply this to $S_n = \frac{1}{n}\sum_{i=1}^n X_i$, with X_i i.i.d X, one gets :

$$\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\mathrm{Var}[X]}{n\epsilon^2}$$

    (S_n is the empirical risk, $\mathbb{E}[X]$ the true one.)
  - Chernoff-Hoeffding bound : $\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq e^{-2n\epsilon^2}$

[1] independent identically distributed

Approximations and confidence intervals

- Proof of Markov inequality

$$\mathbb{E}[X] = \int x d\mathbb{P}_X(x) = \int_{x\geq\epsilon} x d\mathbb{P}_X(x) + \int_{x<\epsilon} x d\mathbb{P}_X(x)$$

$$\mathbb{E}[X] \leq \int_{x\geq\epsilon} x d\mathbb{P}_X(x) \leq \epsilon \int_{x\geq\epsilon} d\mathbb{P}_X(x)$$

$$\mathbb{E}[X] \leq \epsilon\mathbb{P}(X \geq \epsilon)$$

- From bounds to confidence intervals

  Chebicheff inequality : $\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{Var}[X]}{n\epsilon^2}$

  - $\frac{\text{Var}[X]}{n\epsilon^2} \leq \delta$ implies : $\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq \delta$ or

  If $n \geq \frac{\text{Var}[X]}{\delta\epsilon^2}$ then with probability at least $1-\delta$, $|S_n - \mathbb{E}[X]| \leq \epsilon$.

  - Then if $n = \frac{\text{Var}[X]}{\delta\epsilon^2}$, we obtain :

  For all n, with probability at least $1-\delta$, $|S_n - \mathbb{E}[X]| \leq \sqrt{\frac{\text{Var}[X]}{n\delta}}$.

$$\mathbb{E}[loss(f(X), Y)] \in \mathbb{E}_{emp}[loss(f(X), Y)] + \left[-\sqrt{\frac{\text{Var}[X]}{n\delta}}, \sqrt{\frac{\text{Var}[X]}{n\delta}}\right]$$

Approximations and confidence intervals

- Proof of Markov inequality

$$\begin{aligned}
\mathbb{E}[X] &= \int x d\mathbb{P}_X(x) = \int_{x \geq \epsilon} x d\mathbb{P}_X(x) + \int_{x<\epsilon} x d\mathbb{P}_X(x) \\
\mathbb{E}[X] &\leq \int_{x \geq \epsilon} x d\mathbb{P}_X(x) \leq \epsilon \int_{x \geq \epsilon} d\mathbb{P}_X(x) \\
\mathbb{E}[X] &\leq \epsilon \mathbb{P}(X \geq \epsilon)
\end{aligned}$$

- From bounds to confidence intervals

  Chebicheff inequality : $\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{Var}[X]}{n\epsilon^2}$

  - $\frac{\text{Var}[X]}{n\epsilon^2} \leq \delta$ implies : $\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq \delta$ or

  If $n \geq \frac{\text{Var}[X]}{\delta\epsilon^2}$ then with probability at least $1-\delta$, $|S_n - \mathbb{E}[X]| \leq \epsilon$.

  - Then if $n = \frac{\text{Var}[X]}{\delta\epsilon^2}$, we obtain :

  For all n, with probability at least $1 - \delta$, $|S_n - \mathbb{E}[X]| \leq \sqrt{\frac{\text{Var}[X]}{n\delta}}$.

  $$\mathbb{E}[loss(f(X), Y)] \in \mathbb{E}_{emp}[loss(f(X), Y)] + \left[-\sqrt{\frac{\text{Var}[X]}{n\delta}}, \sqrt{\frac{\text{Var}[X]}{n\delta}}\right]$$

Approximations and confidence intervals

- Proof of Markov inequality

$$
\begin{aligned}
\mathbb{E}[X] &= \int x d\mathbb{P}_X(x) = \int_{x \geq \epsilon} x d\mathbb{P}_X(x) + \int_{x < \epsilon} x d\mathbb{P}_X(x) \\
\mathbb{E}[X] &\leq \int_{x \geq \epsilon} x d\mathbb{P}_X(x) \leq \epsilon \int_{x \geq \epsilon} d\mathbb{P}_X(x) \\
\mathbb{E}[X] &\leq \epsilon \mathbb{P}(X \geq \epsilon)
\end{aligned}
$$

- From bounds to confidence intervals

  Chebicheff inequality : $\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{Var}[X]}{n\epsilon^2}$

  - $\frac{\text{Var}[X]}{n\epsilon^2} \leq \delta$ implies : $\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq \delta$ or

  If $n \geq \frac{\text{Var}[X]}{\delta\epsilon^2}$ then with probability at least $1-\delta$, $|S_n - \mathbb{E}[X]| \leq \epsilon$.

  - Then if $n = \frac{\text{Var}[X]}{\delta\epsilon^2}$, we obtain :

  For all n, with probability at least $1 - \delta$, $|S_n - \mathbb{E}[X]| \leq \sqrt{\frac{\text{Var}[X]}{n\delta}}$.

$$\mathbb{E}[loss(f(X), Y)] \in \mathbb{E}_{emp}[loss(f(X), Y)] + \left[-\sqrt{\frac{\text{Var}[X]}{n\delta}}, \sqrt{\frac{\text{Var}[X]}{n\delta}}\right]$$

Approximations and confidence intervals

- Proof of Markov inequality

$$\begin{aligned} \mathbb{E}[X] &= \int x d\mathbb{P}_X(x) = \int_{x\geq\epsilon} x d\mathbb{P}_X(x) + \int_{x<\epsilon} x d\mathbb{P}_X(x) \\ \mathbb{E}[X] &\leq \int_{x\geq\epsilon} x d\mathbb{P}_X(x) \leq \epsilon \int_{x\geq\epsilon} d\mathbb{P}_X(x) \\ \mathbb{E}[X] &\leq \epsilon\mathbb{P}(X\geq\epsilon) \end{aligned}$$

- From bounds to confidence intervals

Chebicheff inequality : $\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\text{Var}[X]}{n\epsilon^2}$

- $\frac{\text{Var}[X]}{n\epsilon^2} \leq \delta$ implies : $\mathbb{P}(|S_n - \mathbb{E}[X]| \geq \epsilon) \leq \delta$ or

If $n \geq \frac{\text{Var}[X]}{\delta\epsilon^2}$ then with probability at least $1-\delta$, $|S_n - \mathbb{E}[X]| \leq \epsilon$.

- Then if $n = \frac{\text{Var}[X]}{\delta\epsilon^2}$, we obtain :

For all n, with probability at least $1-\delta$, $|S_n - \mathbb{E}[X]| \leq \sqrt{\frac{\text{Var}[X]}{n\delta}}$.

$$\mathbb{E}[loss(f(X),Y)] \in \mathbb{E}_{emp}[loss(f(X),Y)] + \left[-\sqrt{\frac{\text{Var}[X]}{n\delta}}, \sqrt{\frac{\text{Var}[X]}{n\delta}}\right]$$

Minimizing a function

Goal : find the global minimimum/minimizer of $f : \mathbb{R}^d \to \mathbb{R}$.

Potentials problems / partial solutions :

- Existence of a global minimum ?
  - $\hookrightarrow$ f is continuous and coercive ($f(x) \to \infty$ when $\|x\| \to \infty$).
- Characterization of the minimizers ?
  - $\hookrightarrow$ f is C^1. If x^* is a local minimizer then its gradient $\nabla f(x) = 0_{\mathbb{R}^d}$.
  - $\hookrightarrow$ f is C^2. x^* is a local minimizer iff its gradient $\nabla f(x) = 0_{\mathbb{R}^d}$ and its hessian $\nabla^2 f(x)$ is a non-negative matrix.
- Characterization of the global minimizers ?

Zeroing the gradient is not sufficient (maxima, saddle points,...) !

Minimizing a function

Goal : find the global minimimum/minimizer of $f : \mathbb{R}^d \to \mathbb{R}$ for $x \in Q$.

- Constrained minimization ($Q \neq \mathbb{R}^d$) : characterization of the minimizers ?
  $\hookrightarrow$ minimizers may be on the border of Q : $\nabla f(x^*) \neq 0$!
- Gradient descents :
  - Algorithms of the form : $x^{t+1} = x^t - \gamma_t \nabla f(x^t)$
  - Ex : Gauss-Newton, conjuguate gradient descent,...
  - Convergence ?
- What if f is not differentiable ?

Convex fonctions

Definition (convex functions)

$f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex iff $\forall \lambda \in [0,1]$, $\forall x, y \in \mathbb{R}^d$,
$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$$

$f : \mathbb{R}^d \rightarrow \mathbb{R}$ is stricly convex iff $\forall \lambda \in [0,1]$, $\forall x, y \in \mathbb{R}^d$, s.t $x \neq y$
$$\text{(resp. } f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y)\text{)}$$

- Other characterizations
  - If $f \in C^2$, f convex iff its $\nabla^2 f$ is non-negative.
  - $f : \mathbb{R}^d \rightarrow \mathbb{R}$, fconvex iff f' is non-decreasing iff $f'' \geq 0$
  - f lies over all its tangents.
- *Ex. : affine fonctions, square loss, exp,...*
- Properties
  - no maxima, no saddle points and non local minima !
  - $\nabla f(x) = 0 \Rightarrow x$ is a global minimizer.

Convex functions are easier to minimize !

Convex fonctions

Definition (convex functions)

$f : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex iff $\forall \lambda \in [0,1]$, $\forall x, y \in \mathbb{R}^d$,

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

$f : \mathbb{R}^d \rightarrow \mathbb{R}$ is stricly convex iff $\forall \lambda \in [0,1]$, $\forall x, y \in \mathbb{R}^d$, s.t $x \neq y$

$$\text{(resp. } f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y))$$

- Other characterizations
  - If $f \in C^2$, f convex iff its $\nabla^2 f$ is non-negative.
  - $f : \mathbb{R}^d \rightarrow \mathbb{R}$, fconvex iff f' is non-decreasing iff $f'' \geq 0$
  - f lies over all its tangents.
- *Ex. : affine fonctions, square loss, exp,...*
- Properties
  - no maxima, no saddle points and non local minima !
  - $\nabla f(x) = 0 \Rightarrow x$ is a global minimizer.

Convex functions are easier to minimize !

A convex and constrained problem in classification

Problem

- Inputs : $\{x_i, y_i\}_{i=1..n}$, $x_i \in \mathbb{R}^d$, $y_i \in \{0, 1\}$.
- Goal : (P) Min $J(w, b) = \frac{1}{2}\|w\|^2 + \sum_1^n max(0, 1 - y_i(wx_i + b))$

Resolution :

- Rewrite (P) as :
  Min $J(w, b, \xi) = \frac{1}{2}w^2 + \sum_1^n \xi_i$ s.t. $y_i(wx_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$
- Introduce a Lagrange multiplier for each constraint :
  $L(w, b, \xi, \alpha, \eta) = \frac{1}{2}\|w\|^2 + \sum_1^n \xi_i + \sum_i \alpha_i(1 - \xi_i - y_i(wx_i + b)) + \sum_i \eta_i \xi_i$,
  $\alpha_i \geq 0$, $\eta_i > 0$.
- The first order conditions $\partial_w J = 0$, $\partial_\xi J = 0$, $\partial_b f = 0$ yield :
  $w = \sum_i \alpha_i y_i x_i \qquad \sum_i \alpha_i y_i = 0 \qquad \forall i, 1 = \alpha_i + \eta_i$
- Which substituted in (P) gives the dual problem :
  Maximize $J(\alpha) = \frac{1}{2}\|\sum_i \alpha_i y_i x_i\|^2 - \alpha^T \mathbf{1}$ s.t. $0 \leq \alpha \leq 1$