Mathématiques pour l'intelligence artificielle

UE MIA, M2 IAAA, AMU, 2022-2023

Thomas Schatz

Jeudi 3 novembre 2022

Plan du cours (prévisionnel)

9 séances de 3h

Partie 1 (DM1) : algèbre linéaire et probabilités

1. Notions de bases sur les preuves (+ Algèbre linéaire?)
2. Algèbre linéaire (+ Probabilités?)
3. Probabilités

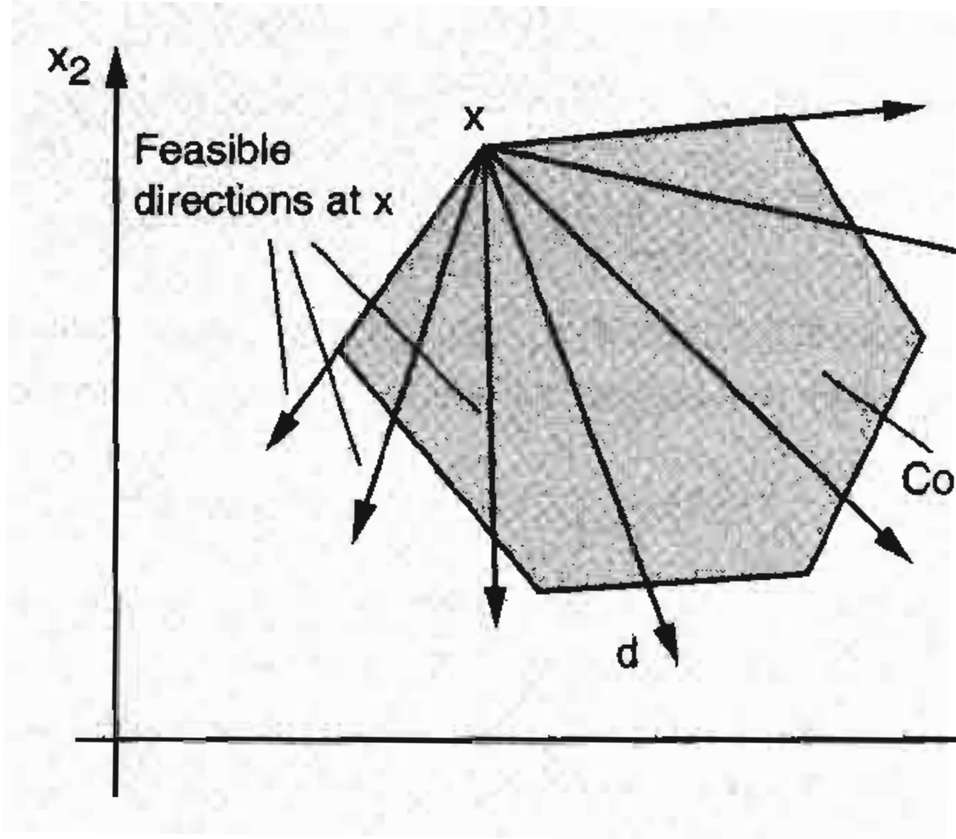
Partie 2 (DM2): statistique et optimisation

4. Statistiques
5. Optimisation

Partie 3 (DM3):

6. Optimisation sous contraintes
7. Optimisation stochastique
8. Théorie de l'apprentissage

Présence de contraintes: Descente de gradient projeté



Présence de contraintes: Descente de gradient projeté

‘Backtracking’ avec la règle d’Armijo le long de l’arc de projection

Input: $x_0 \in \mathbf{R}^n, f : U \subset \mathbf{R}^n \to \mathbf{R}$ de classe $\mathcal{C}^1, s > 0, 0 < \beta < 1, 0 < \sigma < 1.$

U convexe, fermé, non-vide

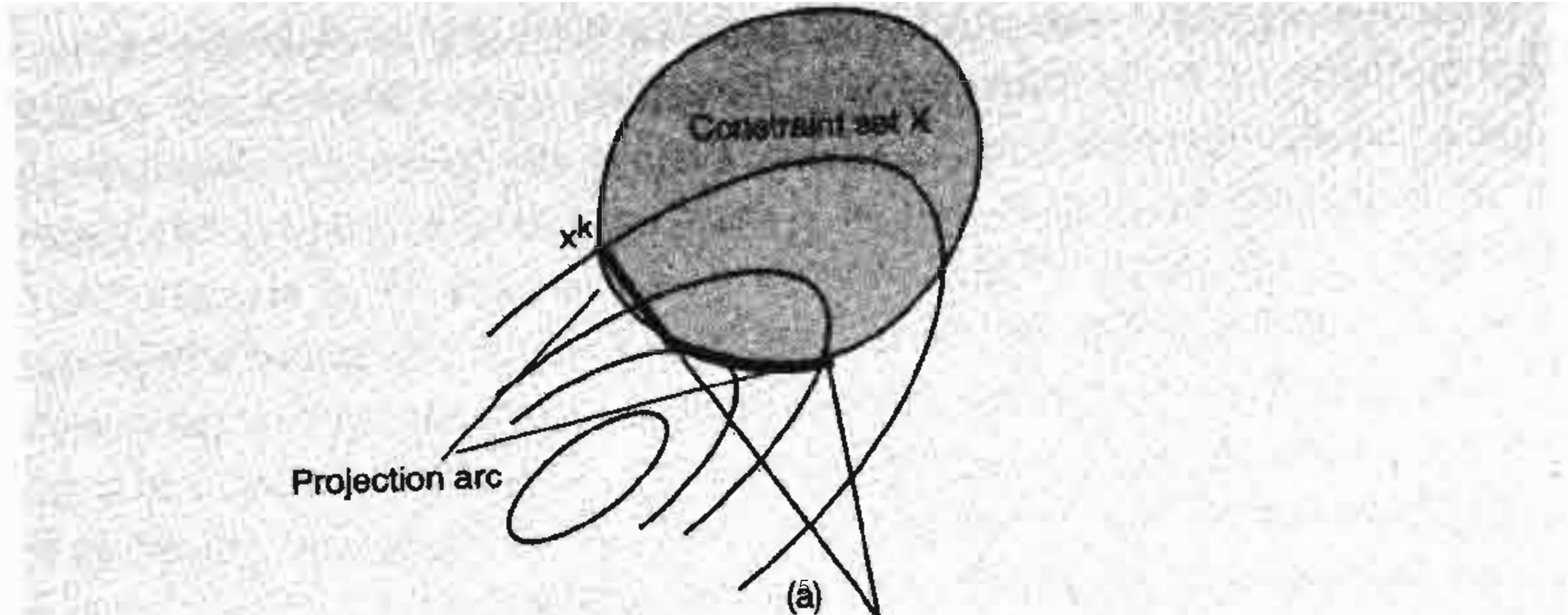
Iteration: $x_{k+1} := p_k(\beta^{m_k} s)$

$p_k(r) = [x_k - r\nabla f(x_k)]_U$ et m_k plus petit entier m tel que

$$f(x_k) - f(x_{k+1}) \geq \sigma \nabla f(x_k)^T (x_k - x_{k+1})$$

Présence de contraintes: Descente de gradient projeté

‘Backtracking’ avec la règle d’Armijo le long de l’arc de projection



Présence de contraintes: Descente de gradient projeté

‘Backtracking’ avec la règle d’Armijo le long de l’arc de projection

Input: $x_0 \in \mathbf{R}^n, f : U \subset \mathbf{R}^n \to \mathbf{R}$ de classe $\mathcal{C}^1, s > 0, 0 < \beta < 1, 0 < \sigma < 1.$

U convexe, fermé, non-vide

Iteration: $x_{k+1} := p_k(\beta^{m_k} s)$

$p_k(r) = [x_k - r\nabla f(x_k)]_U$ et m_k plus petit entier m tel que

$$f(x_k) - f(x_{k+1}) \geq \sigma \nabla f(x_k)^T (x_k - x_{k+1})$$

Dualité

minimize $f(x)$

subject to $x \in X, \qquad g_j(x) \leq 0, \quad j = 1, \ldots, r,$

$$f : \mathbf{R}^n \to \mathbf{R}$$
$$g_j : \mathbf{R}^n \to \mathbf{R}$$
$$X \subset \mathbf{R}^n$$

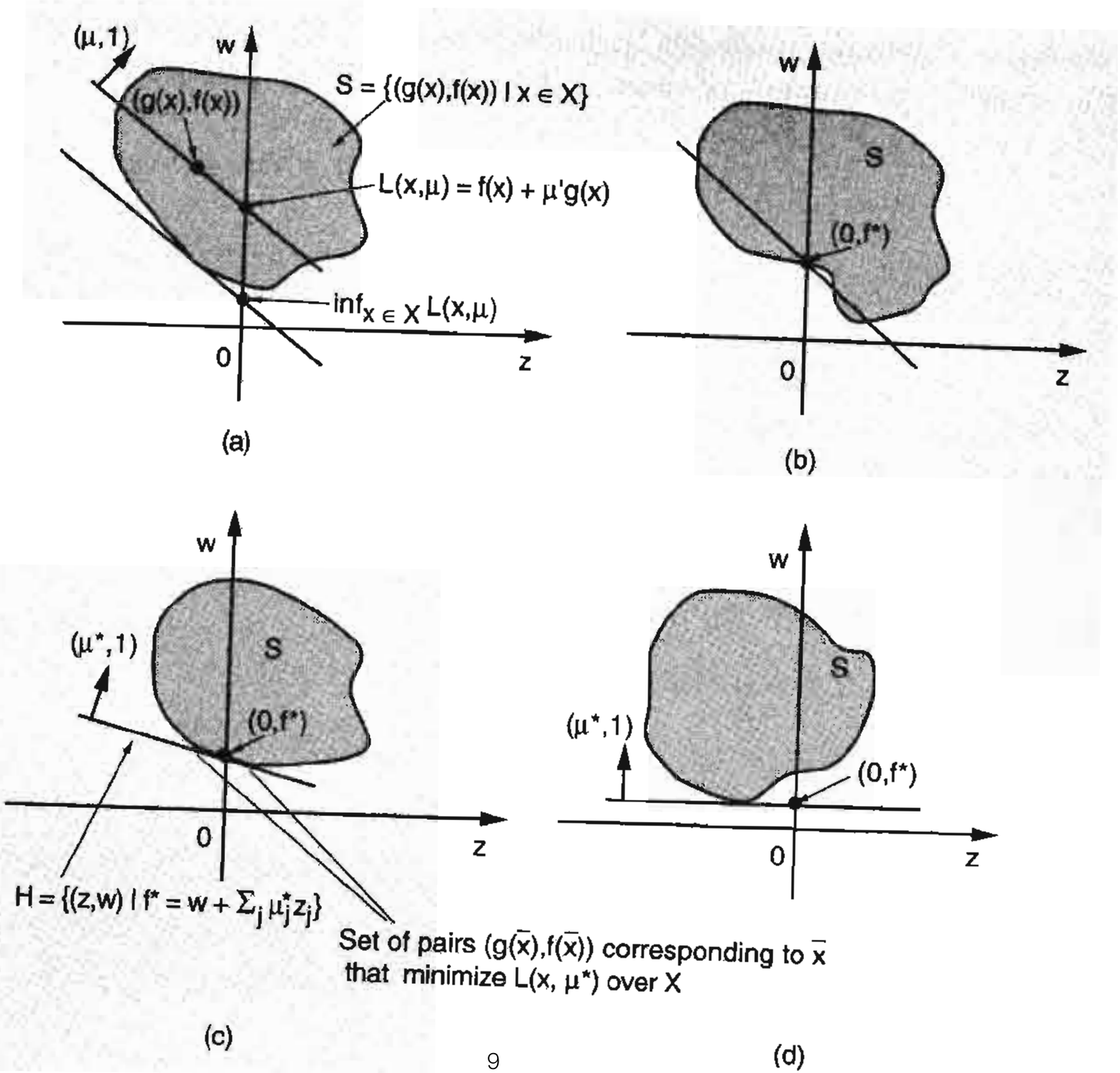
$$f^* = \inf_{\substack{x \in X \\ g_j(x) \leq 0,\, j=1,\ldots,r}} f(x).$$

$$-\infty < f^* < +\infty$$

Il existe x^*, tel que $f(x^*) = f^*$

Dualité

$$L(x, \mu) = f(x) + \sum_{j=1}^{r} \mu_j g_j(x) = f(x) + \mu' g(x).$$


(μ,1)
w
(g(x),f(x))
S = {(g(x),f(x)) | x ∈ X}
L(x,μ) = f(x) + μ'g(x)
inf_{x ∈ X} L(x,μ)
0
z
(a)
w
S
(0,f*)
0
z
(b)
w
(μ*,1)
S
(0,f*)
0
z
H = {(z,w) | f* = w + Σ_j μ*_j z_j}
(c)
w
S
(μ*,1)
(0,f*)
0
z
(d)
Set of pairs (g(x̄),f(x̄)) corresponding to x̄
that minimize L(x, μ*) over X

Dualité

Fonction duale

$$q(\mu) = \inf_{x \in X} L(x, \mu).$$

Problème dual

$$\begin{aligned} &\text{maximize} \quad q(\mu) \\ &\text{subject to} \quad \mu \geq 0, \ \mu \in D, \end{aligned}$$

$$D = \{\mu \mid q(\mu) > -\infty\}.$$

Proposition 5.1.2: The domain D of the dual function q is convex and q is concave over D.

Proposition 5.1.3: (Weak Duality Theorem) We have

$$q^* \leq f^*.$$

Dualité

Dualité forte et qualification des contraintes

Example :

Si f est convexe et les contraintes g_j sont linéaires, alors il n'y a pas d'écart dual: $q^* = f^*$

Dualité

Proposition 5.1.5: (Optimality Conditions) (x^*, μ^*) is an optimal solution-Lagrange multiplier pair if and only if

$$x^* \in X, \quad g(x^*) \leq 0, \qquad \text{(Primal Feasibility)}, \tag{5.4}$$

$$\mu^* \geq 0, \qquad \text{(Dual Feasibility)}, \tag{5.5}$$

$$x^* = \arg\min_{x \in X} L(x, \mu^*), \qquad \text{(Lagrangian Optimality)}, \tag{5.6}$$

$$\mu_j^* g_j(x^*) = 0, \quad j = 1, \ldots, r, \qquad \text{(Complementary Slackness)}. \tag{5.7}$$

Plan du cours (prévisionnel)

9 séances de 3h

Partie 1 (DM1) : algèbre linéaire et probabilités

1. Notions de bases sur les preuves (+ Algèbre linéaire?)
2. Algèbre linéaire (+ Probabilités?)
3. Probabilités

Partie 2 (DM2): statistique et optimisation

4. Statistiques
5. Optimisation

Partie 3 (DM3):

6. Optimisation sous contraintes
7. Optimisation stochastique
8. Théorie de l'apprentissage

Plan du cours (prévisionnel)

9 séances de 3h

Partie 1 (DM1) : algèbre linéaire et probabilités

1. Notions de bases sur les preuves (+ Algèbre linéaire?)
2. Algèbre linéaire (+ Probabilités?)
3. Probabilités

Partie 2 (DM2): statistique et optimisation

4. Statistiques
5. Optimisation

Partie 3 (DM3):

6. Optimisation sous contraintes
7. Optimisation stochastique
8. Théorie de l'apprentissage

Optimisation stochastique

Contexte : minimisation du risque empirique pour une fonction de coût “séparable par point de donnée”

$$R_n(w) = \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Descente de gradient stochastique

$$w_1 \in \mathbb{R}^d \text{ given}$$

$$w_{k+1} \leftarrow w_k - \alpha_k \nabla f_{i_k}(w_k)$$

i_k is chosen *randomly* from $\{1, \ldots, n\}$ and α_k is a positive stepsize

Optimisation stochastique

Exemple de garantie de convergence

(cf. Bottou, Curtis et Nocedal (2018) Optimisation Methods for Large-Scale Machine Learning)

Si

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty$$

Assumption 4.1 (**Lipschitz-continuous objective gradients**). *The objective function* $F : \mathbb{R}^d \to \mathbb{R}$ *is continuously differentiable and the gradient function of* F*, namely,* $\nabla F : \mathbb{R}^d \to \mathbb{R}^d$*, is Lipschitz continuous with Lipschitz constant* $L > 0$*, i.e.,*

$$\|\nabla F(w) - \nabla F(\overline{w})\|_2 \leq L\|w - \overline{w}\|_2 \quad \textit{for all} \quad \{w, \overline{w}\} \subset \mathbb{R}^d.$$

plus des conditions de régularité pas très contraignantes

Alors

$$\liminf_{k \to \infty} \mathbb{E}[\|\nabla F(w_k)\|_2^2] = 0$$

Plan du cours (prévisionnel)

9 séances de 3h

Partie 1 (DM1) : algèbre linéaire et probabilités

1. Notions de bases sur les preuves (+ Algèbre linéaire?)
2. Algèbre linéaire (+ Probabilités?)
3. Probabilités

Partie 2 (DM2): statistique et optimisation

4. Statistiques
5. Optimisation

Partie 3 (DM3):

6. Optimisation sous contraintes
7. Optimisation stochastique
8. Théorie de l'apprentissage

Plan du cours (prévisionnel)

9 séances de 3h

Partie 1 (DM1) : algèbre linéaire et probabilités

1. Notions de bases sur les preuves (+ Algèbre linéaire?)
2. Algèbre linéaire (+ Probabilités?)
3. Probabilités

Partie 2 (DM2): statistique et optimisation

4. Statistiques
5. Optimisation

Partie 3 (DM3):

6. Optimisation sous contraintes
7. Optimisation stochastique
8. Théorie de l'apprentissage

Théorie de l’apprentissage : TLDR

High-dimensional statistics : non asymptotic

- Bias-variance decomposition of risk (same approach as in classical case)
- Explicit computation impossible for typical ML problems
- Upper bounding :
  - bias : Rademacher (or Gaussian) averages
  - variance : concentration inequalities

High-dimensional statistics : asymptotic

- Different asymptotic regimes, e.g. n -> +∞ et n/p -> constante

Deep learning theory?

- Over-parametrisation + SGD -> computationally tractable minimum-norm interpolation?
- Good generalisation properties of minimum-norm interpolations?

Application 2: Normes de matrices

A matrix norm is a function $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ that has the following properties:

- $\|A\| \geq 0$ for any $A \in \mathbb{R}^{m\times n}$, and $\|A\| = 0$ if and only if $A = 0$
- $\|\alpha A\| = |\alpha|\|A\|$ for any $m \times n$ matrix A and scalar α
- $\|A + B\| \leq \|A\| + \|B\|$ for any $m \times n$ matrices A and B

$$\|A\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$$

$$\|A\|_2 = \sigma_1$$

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2}.$$

$$\|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$$

Application 2: Normes de matrices

A matrix norm is a function $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ that has the following properties:

- $\|A\| \geq 0$ for any $A \in \mathbb{R}^{m\times n}$, and $\|A\| = 0$ if and only if $A = 0$
- $\|\alpha A\| = |\alpha|\|A\|$ for any $m \times n$ matrix A and scalar α
- $\|A + B\| \leq \|A\| + \|B\|$ for any $m \times n$ matrices A and B

$$\|A\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$$

$$\|A\|_2 = \sigma_1$$

$$\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2\right)^{1/2}.$$

$$\|A\|_F = \sqrt{\sigma_1^2 + \cdots + \sigma_r^2}$$