# Mathématiques pour l’intelligence artificielle

M2 informatique parcours IAAA, Aix-Marseille Université

Thomas Schatz

Mardi 19 septembre 2023

# Plan du cours (prévisionnel)

Introduction

1. Notions de bases sur les preuves

2. Algèbre linéaire

3. Optimisation

4. Probabilités

5. Statistique

Discussion

# Statistique

1. Cadre général

2. Statistique non-asymptotique: biais, variance, risque et concentration

3. Statistique asymptotique

# Modes de convergence

- Convergence en loi $X_n \to_d X$ si et seulement si

$$\lim_{n \to +\infty} E[f(X_n)] = E[f(X)]$$

pour toute fonction $f$ à valeurs réelles, continue et bornée

Définition équivalente pour des variables aléatoire réelles:

$$\lim_{n \to +\infty} F_{X_n}(x) = F_X(x) \text{ en tout point } X \text{ où } F_X \text{ est continue}$$

# Modes de convergence

- Convergence presque sûre ou presque partout $X_n \to_{a.s.} X$

$$P\left(\lim_{n\to+\infty} X_n = X\right) = 1$$

- Convergence en probabilité $X_n \to_p X$

$$\forall \epsilon > 0, \ \lim_{n\to+\infty} P(|X_n - X| > \epsilon) = 0$$

- Convergence en moyenne quadratique $X_n \to_{\mathcal{L}_2} X$

$$\lim_{n\to+\infty} E[|X_n - X|^2] = 0$$

# Loi forte des grands nombres

$$X_1, X_2, \ldots \text{ variables aléatoires i.i.d.}$$

(ii) (The SLLN). A necessary and sufficient condition for the existence of a constant $c$ for which

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to_{a.s.} c \tag{1.81}$$

is that $E|X_1| < \infty$, in which case $c = EX_1$

# Théorème de la limite centrale

(Multivariate CLT). Let $X_1, ..., X_n$ be i.i.d. random $k$-vectors with a finite $\Sigma = \mathrm{Var}(X_1)$. Then

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - EX_1) \to_d N_k(0, \Sigma). \quad \blacksquare$$

# Transformations continues et théorème de Slutsky

**Theorem 1.10.** Let $X, X_1, X_2, \ldots$ be random $k$-vectors defined on a probability space and $g$ be a measurable function from $(\mathcal{R}^k, \mathcal{B}^k)$ to $(\mathcal{R}^l, \mathcal{B}^l)$. Suppose that $g$ is continuous a.s. $P_X$. Then
(i) $X_n \to_{a.s.} X$ implies $g(X_n) \to_{a.s.} g(X)$;
(ii) $X_n \to_p X$ implies $g(X_n) \to_p g(X)$;
(iii) $X_n \to_d X$ implies $g(X_n) \to_d g(X)$. ∎

**Theorem 1.11** (Slutsky's theorem). Let $X, X_1, X_2, \ldots, Y_1, Y_2, \ldots$ be random variables on a probability space. Suppose that $X_n \to_d X$ and $Y_n \to_p c$, where $c$ is a fixed real number. Then
(i) $X_n + Y_n \to_d X + c$;
(ii) $Y_n X_n \to_d cX$;
(iii) $X_n / Y_n \to_d X/c$ if $c \neq 0$.

# Delta method

Soit $X_1$, $X_2$, ... and $Y$ des vecteurs aléatoires de dimension $k$, tels que :

$$a_n(X_n - c) \rightarrow_d Y,$$

pour $c \in \mathbf{R}^k$ et $(a_n)$ une suite de nombres positifs tendant vers $+\infty$. Alors pour toute fonction $g : \mathbf{R}^k \rightarrow \mathbf{R}$ différentiable en $c$, on a:

$$a_n[g(X_n) - g(c)] \rightarrow_d [\nabla g(c)]^T Y.$$

# Statistique

1. Cadre général

2. Statistique non-asymptotique: biais, variance, risque et concentration

3. Statistique asymptotique

# Plan du cours (prévisionnel)

Introduction

1. Notions de bases sur les preuves

2. Algèbre linéaire

3. Optimisation

4. Probabilités

5. Statistique

Discussion

# Plan du cours (prévisionnel)

Introduction

1. Notions de bases sur les preuves

2. Algèbre linéaire

3. Optimisation

4. Probabilités

5. Statistique

Discussion

# Optimisation

1. Optimisation sans contraintes
2. Optimisation sous contraintes
3. Analyse convexe et dualité
4. Analyse des algorithmes d'optimisation

# Descente de gradient

**Pas fixe**

Input: $x_0 \in \mathbf{R}^n$, $f : \mathbf{R}^n \to \mathbf{R}$ de classe $\mathcal{C}^1$, pas fixe $\gamma > 0$.

Iteration: $x_{k+1} = x_k - \gamma \nabla f(x_k)$

# Descente de gradient

**'Backtracking' avec la règle d'Armijo**

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

$$\text{Iteration : } \; x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

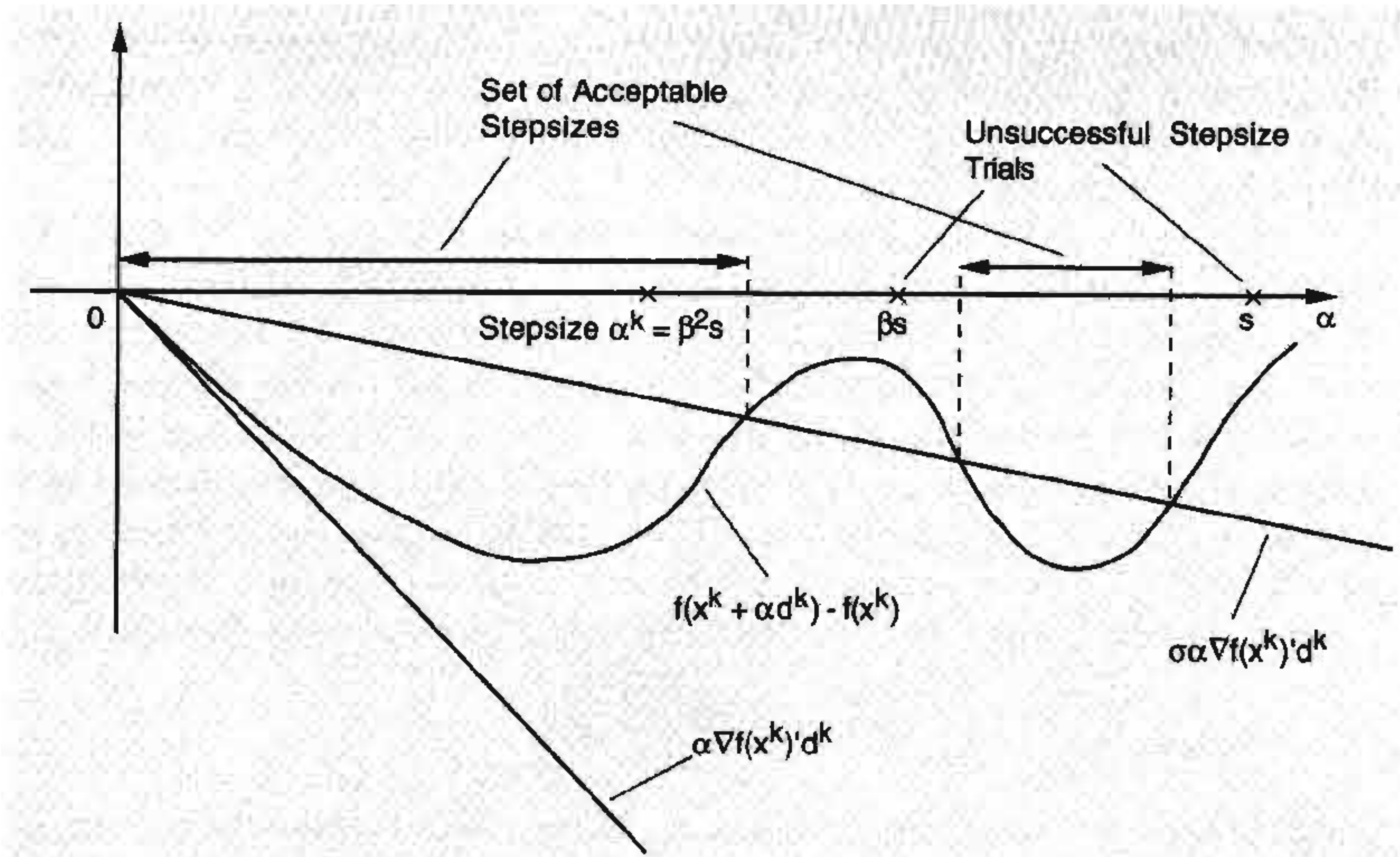
$$\alpha_k = \beta^{m_k} s$$

$m_k$ plus petit entier positif tel que

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

# Descente de gradient

**'Backtracking' avec la règle d'Armijo**



# Descente de gradient

**'Backtracking' avec la règle d'Armijo**

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

$$\text{Iteration : } \; x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

$$\alpha_k = \beta^{m_k} s$$

$m_k$ plus petit entier positif tel que

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

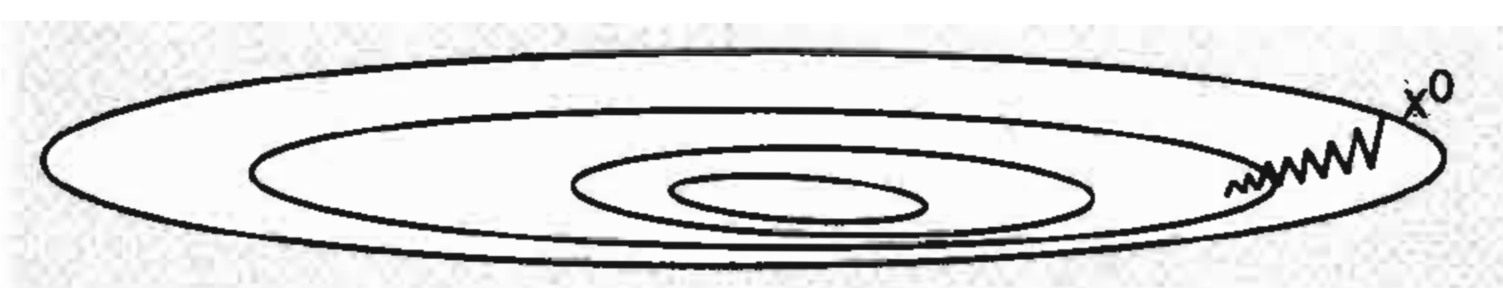
# Descente de gradient

**Convergence**

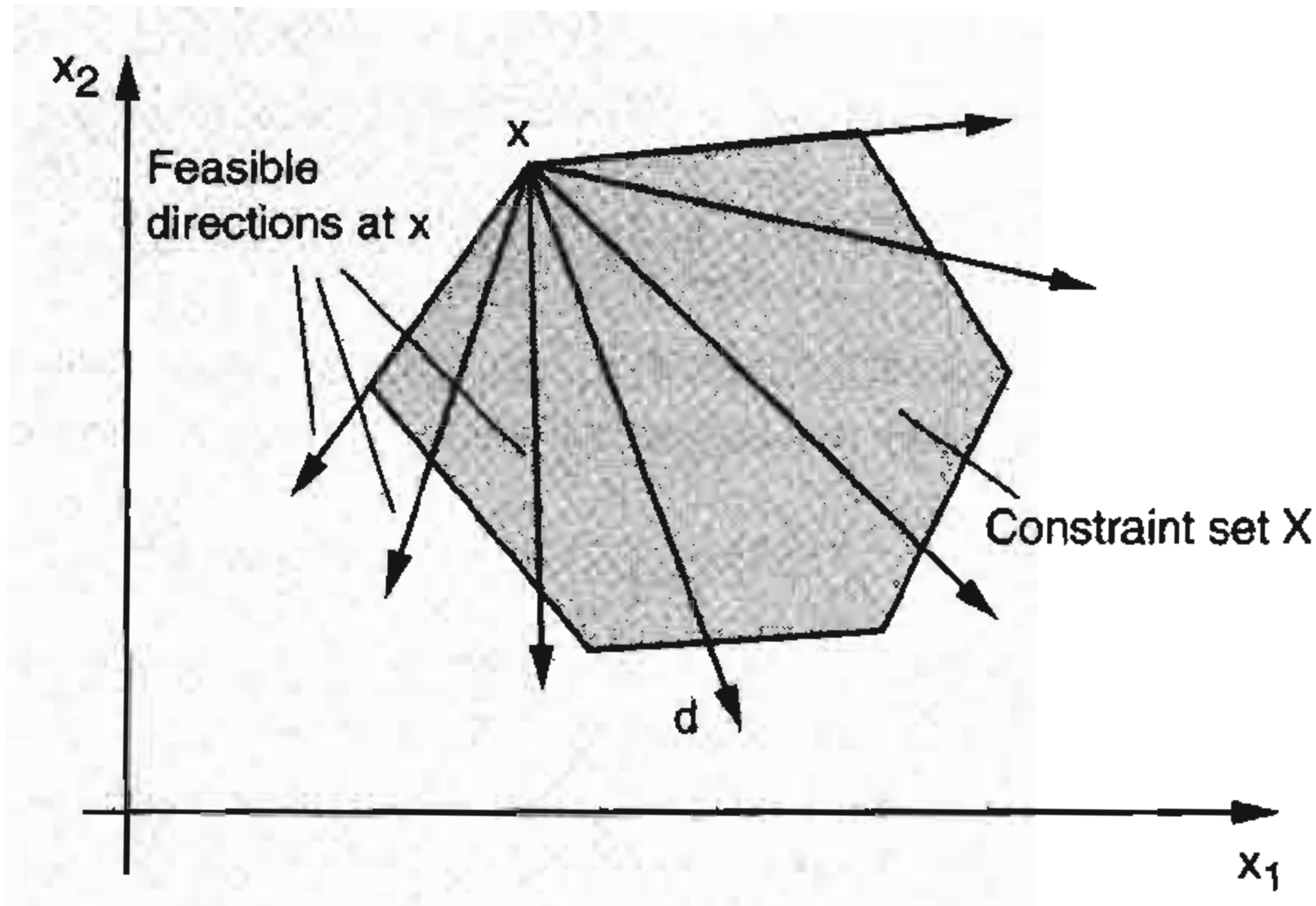
Soit $(x_k)$ la suite des points générés par l'algorithme de descente de gradient avec pas choisi par la règle d'Armijo. Alors, tout point limite (i.e. valeur d'adhérence) de $(x_k)$ est un point stationnaire.

De plus, si $x^*$ est le seul point stationnaire de $f$ dans un ensemble ouvert, il existe un ensemble ouvert $S$ contenant $x^*$ tel que si il existe $k_0$ tel que $x_{k_0} \in S$, alors $x_k \in S$ pour tout $k \geq k_0$ et $x_k \to x^*$.

**Vitesse de convergence ?**



# 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

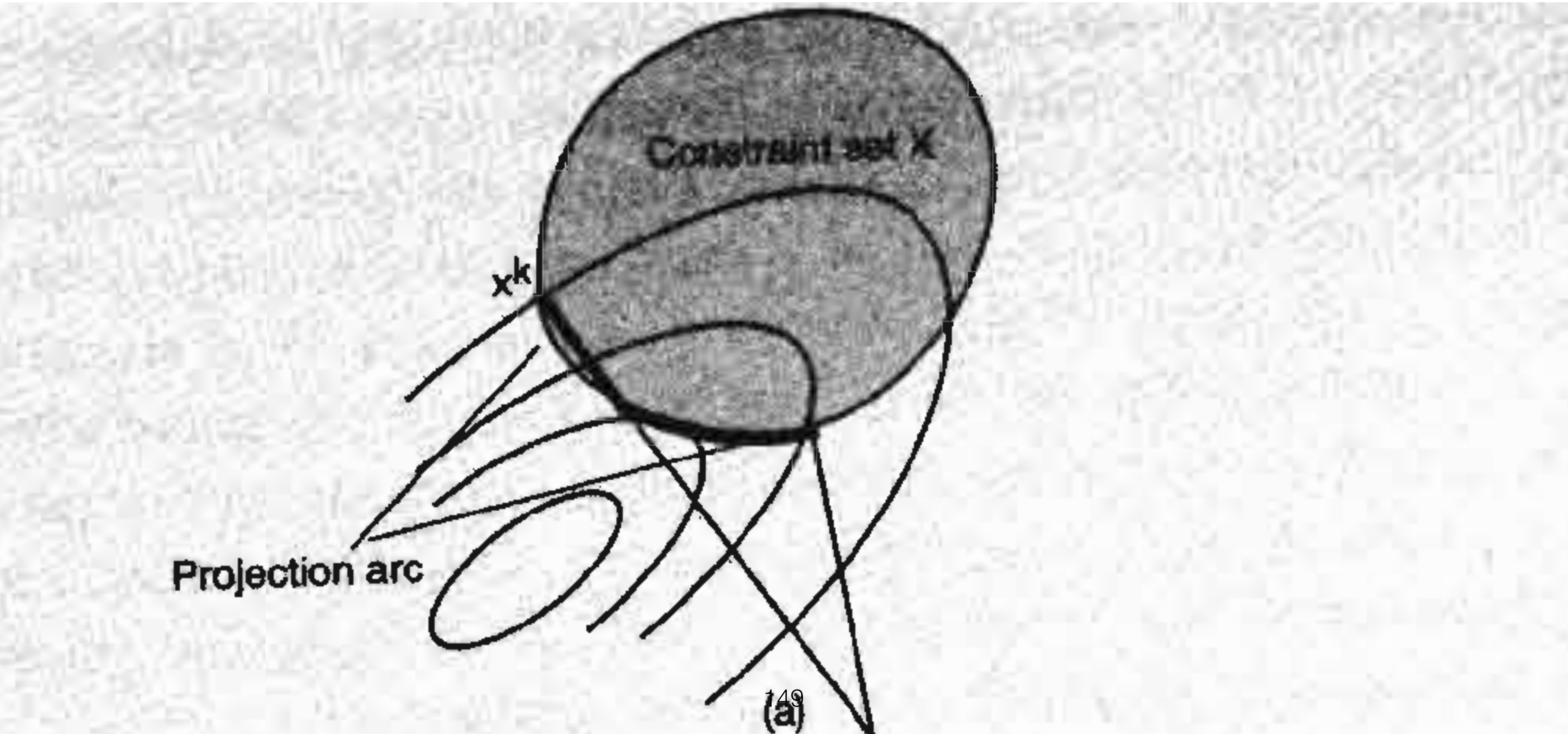
Iteration: $x_{k+1} := p_k(\beta^{m_k} s)$

$p_k(r) = [x_k - r\nabla f(x_k)]_U$ 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

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

# 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 $\alpha_k$ 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.,*

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

plus des conditions de régularité pas très contraignantes

**Alors**

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

# Optimisation

1. Optimisation sans contraintes
2. Optimisation sous contraintes
3. Analyse convexe et dualité
4. Analyse des algorithmes d'optimisation

# Optimisation

1. Optimisation sans contraintes
2. Optimisation sous contraintes
3. Analyse convexe et dualité
4. Analyse des algorithmes d'optimisation
5. Bonus : optimisation et algèbre linéaire

# Normes de vecteurs

A function $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ is called a *vector norm* if it has the following properties:

1. $\|\mathbf{x}\| \geq 0$ for any vector $\mathbf{x} \in \mathbb{R}^n$, and $\|\mathbf{x}\| = 0$ if and only if $\mathbf{x} = \mathbf{0}$
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$ for any vector $\mathbf{x} \in \mathbb{R}^n$ and any scalar $\alpha \in \mathbb{R}$
3. $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for any vectors $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n$.

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}.$$

$$|\mathbf{x}^T\mathbf{y}| \leq \|\mathbf{x}\|_2\|\mathbf{y}\|_2$$

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$$

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\mathbf{x}^T\mathbf{x}}$$

$$\|\mathbf{x}\|_\infty = \max_{1\leq i\leq n} |x_i|$$

Si Q est une matrice orthogonale,
$\|Qx\|_2 = \|x\|_2$

# 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 $\alpha$
- $\|A + B\| \leq \|A\| + \|B\|$ 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 de la notion de norme: lien valeurs propres, valeurs singulières

$$\sigma_r \leq |\lambda| \leq \sigma_1$$

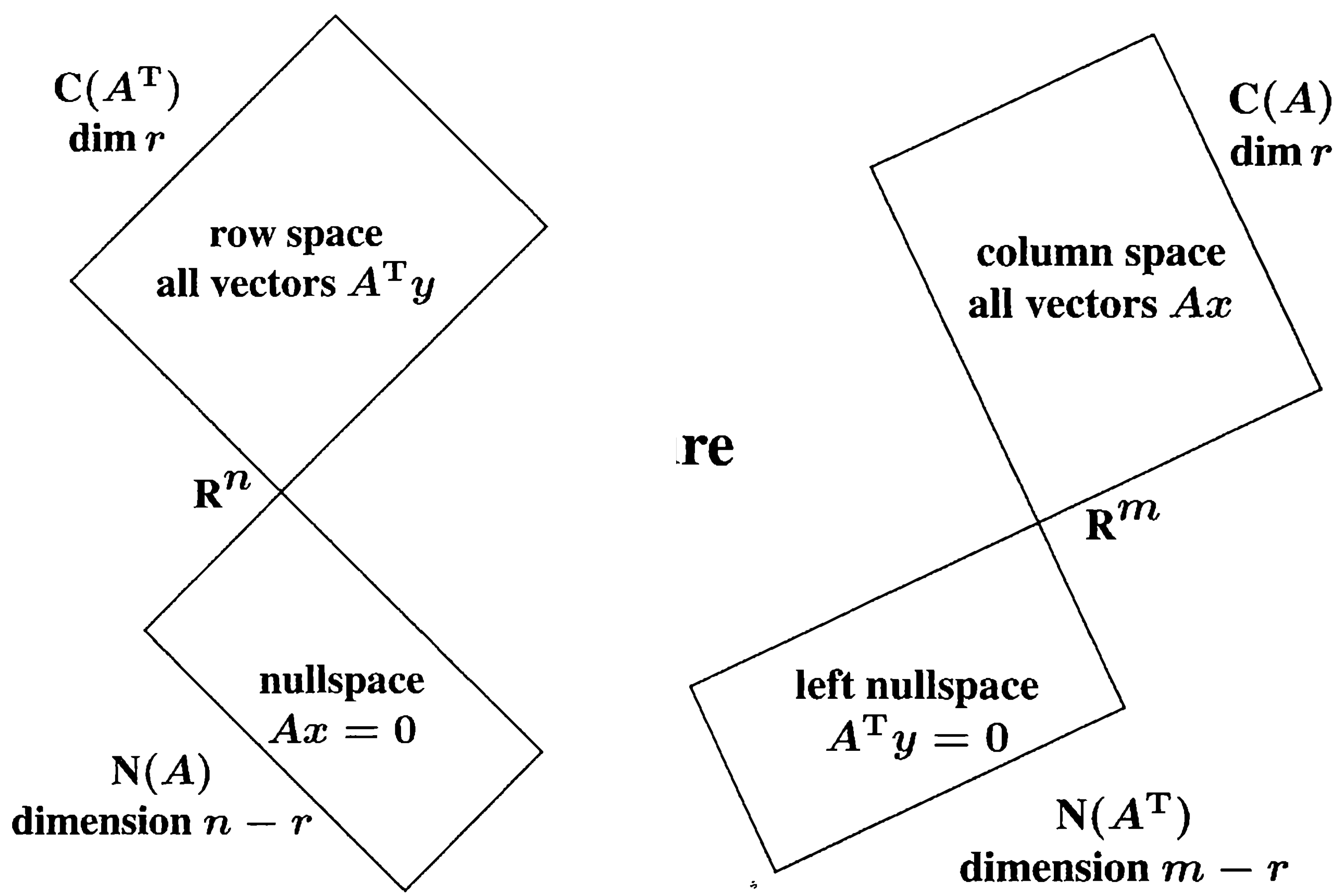
# Application de la notion de norme: lien valeurs propres, valeurs singulières

$$\sigma_r \leq |\lambda| \leq \sigma_1$$

Éléments propres de $A^T A$ et $AA^T$ ?

# Moindre carrés linéaires

$$AV = U\Sigma \qquad A\begin{bmatrix} \\ v_1 \,..\, v_r \,..\, v_n \\ \\ \end{bmatrix} = \begin{bmatrix} \\ u_1 \,..\, u_r \,..\, u_m \\ \\ \end{bmatrix} \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right]$$



# Optimisation

1. Optimisation sans contraintes
2. Optimisation sous contraintes
3. Analyse convexe et dualité
4. Analyse des algorithmes d’optimisation
5. Bonus : optimisation et algèbre linéaire

# Plan du cours (prévisionnel)

Introduction

1. Notions de bases sur les preuves

2. Algèbre linéaire

3. Optimisation

4. Probabilités

5. Statistique

Discussion