Mathématiques pour l'intelligence artificielle

UE MIA, M2 IAAA, AMU, 2022-2023

Thomas Schatz

Vendredi 7 octobre 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

9. Putting it all together

Conditions d'optimalité

Conditions nécessaires, premier ordre

Si x^* est un minimum local et f est de classe $\mathcal{C}^1$ sur un ouvert $U \subset \mathbf{R}^n$ contenant x, alors $\nabla f(x^*) = 0$

Analyse

Hessienne

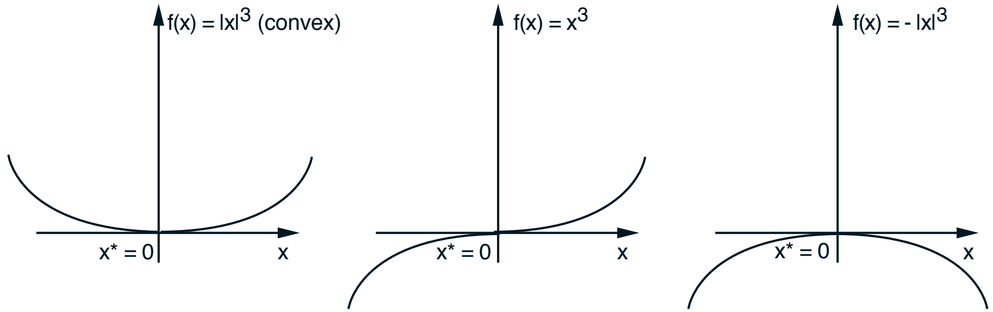
$$f : U \subset \mathbf{R}^n \mapsto \mathbf{R}$$

$$\nabla^2 f = \boldsymbol{H}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial {x_1}^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial {x_2}^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial {x_n}^2} \end{bmatrix}$$

Conditions d'optimalité

Conditions nécessaires, premier ordre

Si x^* est un minimum local et f est de classe $\mathcal{C}^1$ sur un ouvert $U \subset \mathbf{R}^n$ contenant x, alors $\nabla f(x^*) = 0$



Conditions nécessaires, second ordre

Si x^* est un minimum local et f est de classe $\mathcal{C}^2$ sur un ouvert $U \subset \mathbf{R}^n$ contenant x, alors $\nabla f(x^*) = 0$ et $\nabla^2 f(x^*)$ est semi-définie positive

Conditions suffisantes, second ordre

Si f est de classe $\mathcal{C}^2$ sur un ouvert $U \subset \mathbf{R}^n$ contenant x, si $\nabla f(x^*) = 0$ et si $\nabla^2 f(x^*)$ est définie positive, alors x^* est un minimum local.

Existence de minimum

Théorème de Weierstrass

Si f est définie et continue sur un sous-ensemble fermé et borné de $\mathbf{R}^n$, alors f admet un minimum global.

Théorème

Si f est définie et continue sur $\mathbf{R}^n$ et *coercive* (i.e. $f(x) \to +\infty$ when $\|x\| \to +\infty$, alors f admet un minimum global sur tout sous-ensemble fermé de $\mathbf{R}^n$.

Unicité du minimum

Si f est une fonction *strictement* convexe définie sur un ensemble convexe $X \subset \mathbf{R}^n$, alors f admet au plus un minimum global.

Convexité

L'ensemble X est convexe ssi pour tout x, y dans X, le segment $[x, y] := \{tx + (1-t)y \mid t \in [0,1]\}$ est inclus dans X

Une fonction $f : X \subset \mathbf{R}^n \to R$ est convexe ssi X est convexe et pour tout x, y dans X et pour tout $t \in [0,1]$,

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y).$$

Si l'inégalité est stricte pour $t \in]0,1[$, la fonction est dite *strictement* convexe.

Intérêt

Si f est une fonction *strictement* convexe définie sur un ensemble convexe $X \subset \mathbf{R}^n$, alors f admet au plus un minimum global.

Si f est une fonction convexe définie sur un ensemble convexe $X \subset \mathbf{R}^n$, alors tout minimum local de f est aussi un minimum global de f. Si X est ouvert, alors $\nabla f(x^*) = 0$ est équivaut à x^* est un minimum global de f

Convexité

Reconnaître une fonction convexe

- les fonctions linéaires sont convexes
- les combinaisons linéaires positives de fonctions convexes sont convexes
- le maximum (point par point) d'un ensemble de fonctions convexes est convexe
- ...

Si f est de classe C^2, f est convexe si et seulement si $\nabla^2 f$ est semi-définie positive sur l'intérieur de X. Si $\nabla^2 f$ est définie positive sur l'intérieur de X, f est strictement convexe.

Application 1: Moindre carrés

A matrice réelle m par n, y matrice colonne de taille m.

$$\min_{x \in \mathbf{R}^n} \|Ax - y\|_2 \ ?$$

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

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

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

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

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

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

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

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

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

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

Algèbre linéaire

1. Espaces vectoriels et fonctions linéaires
2. Matrices
3. Angles et orthogonalité
4. Structure des applications linéaires et matrices
5. Espaces de matrices et normes
6. Algèbre linéaire numérique

Algèbre linéaire

1. Espaces vectoriels et fonctions linéaires
2. Matrices
3. Angles et orthogonalité
4. Structure des applications linéaires et matrices
5. Espaces de matrices et normes
6. Algèbre linéaire numérique

Vu l'année dernière

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 $\mathcal{C}^1, s > 0, 0 < \beta < 1, 0 < \sigma < 1.$

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

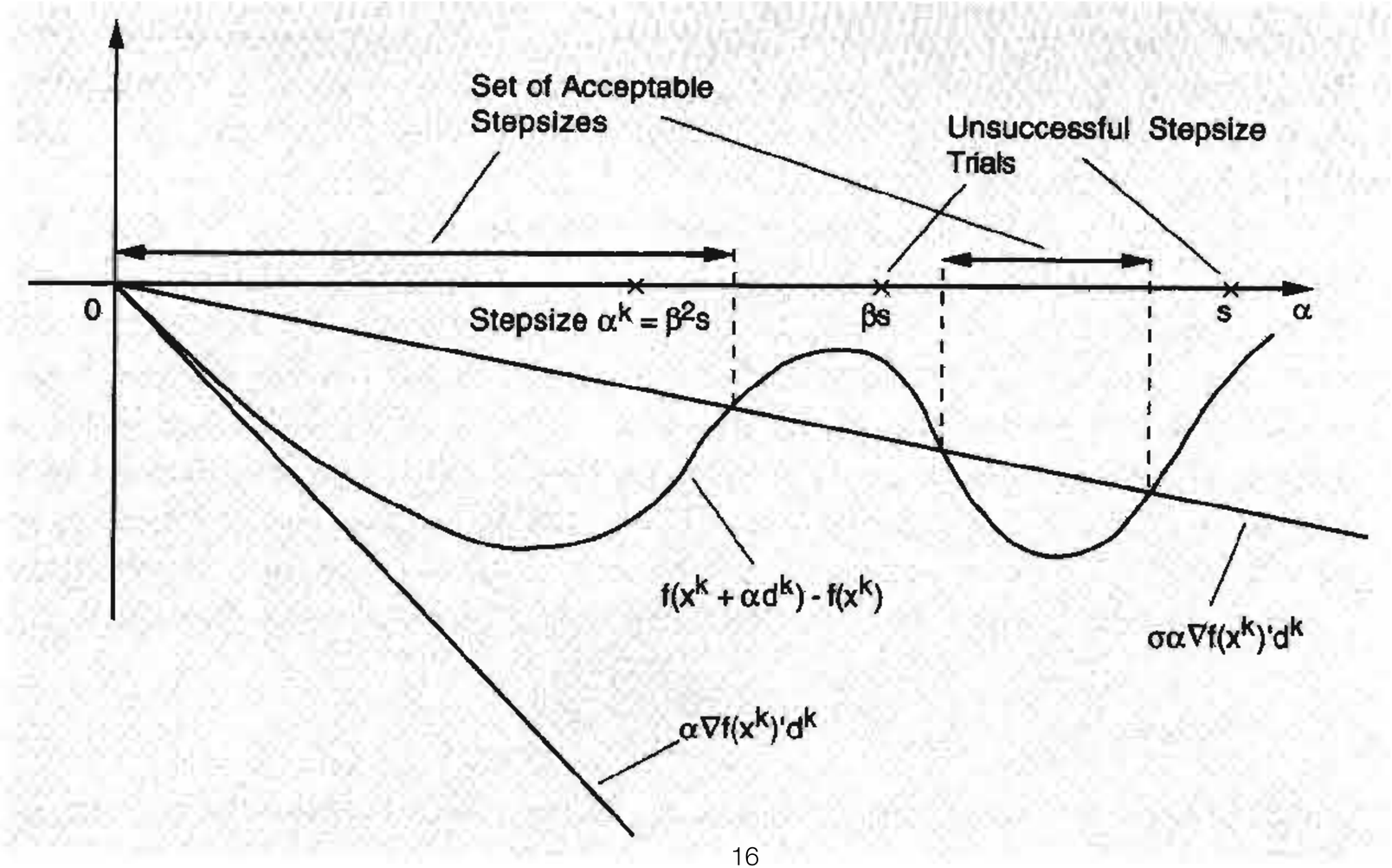
$$\alpha_k = \beta^{m_k} s$$

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

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

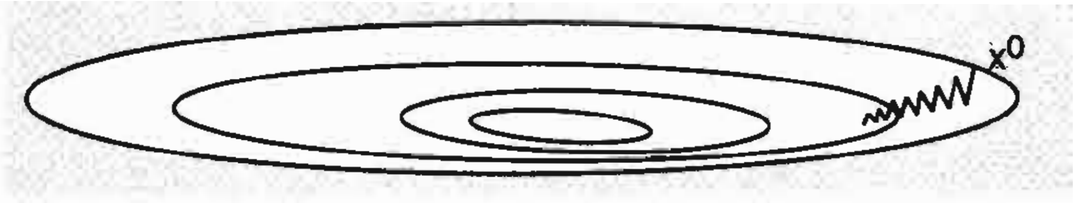
Descente de gradient

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 ?



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

9. Putting it all together

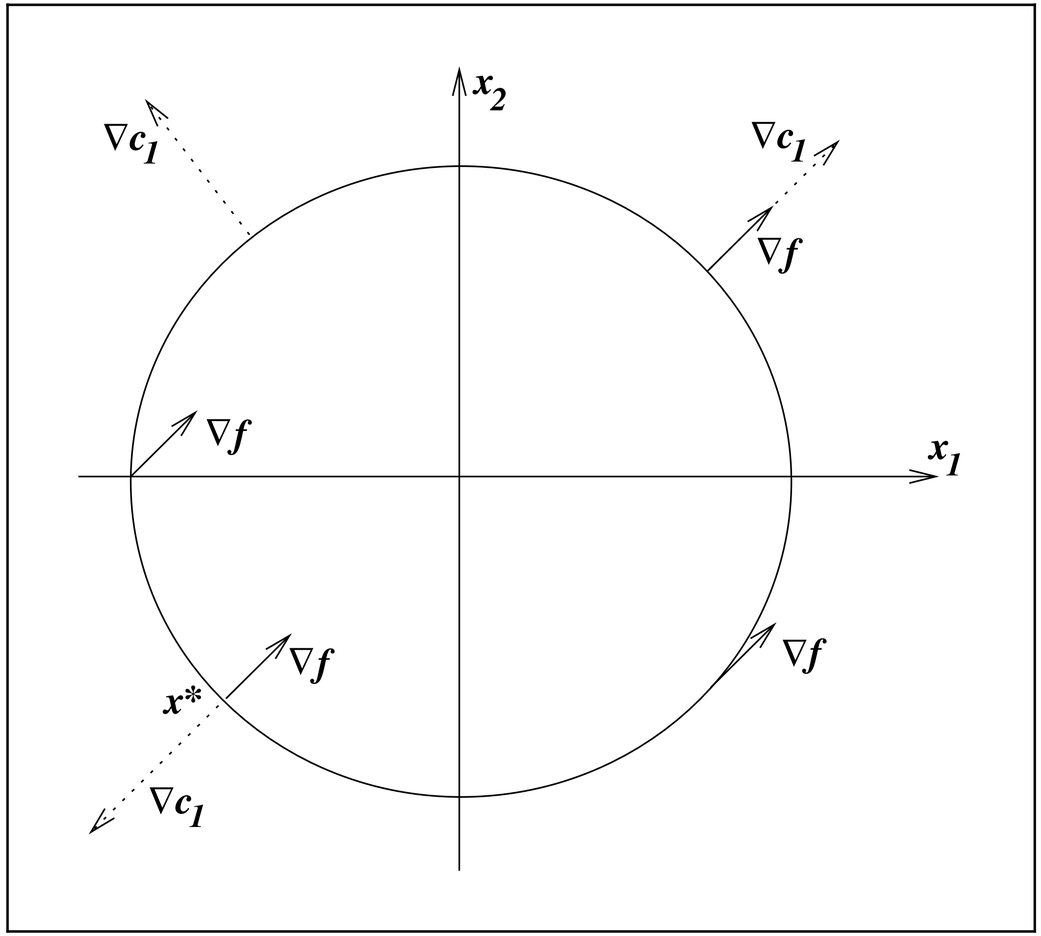
Optimisation sous contraintes

$$\min_{x \in \mathbf{R}^n} f(x) \text{ subject to } \begin{cases} c_i(x) = 0, & i \in \mathcal{E} \\ c_i(x) \geq 0, & i \in \mathcal{I} \end{cases}$$

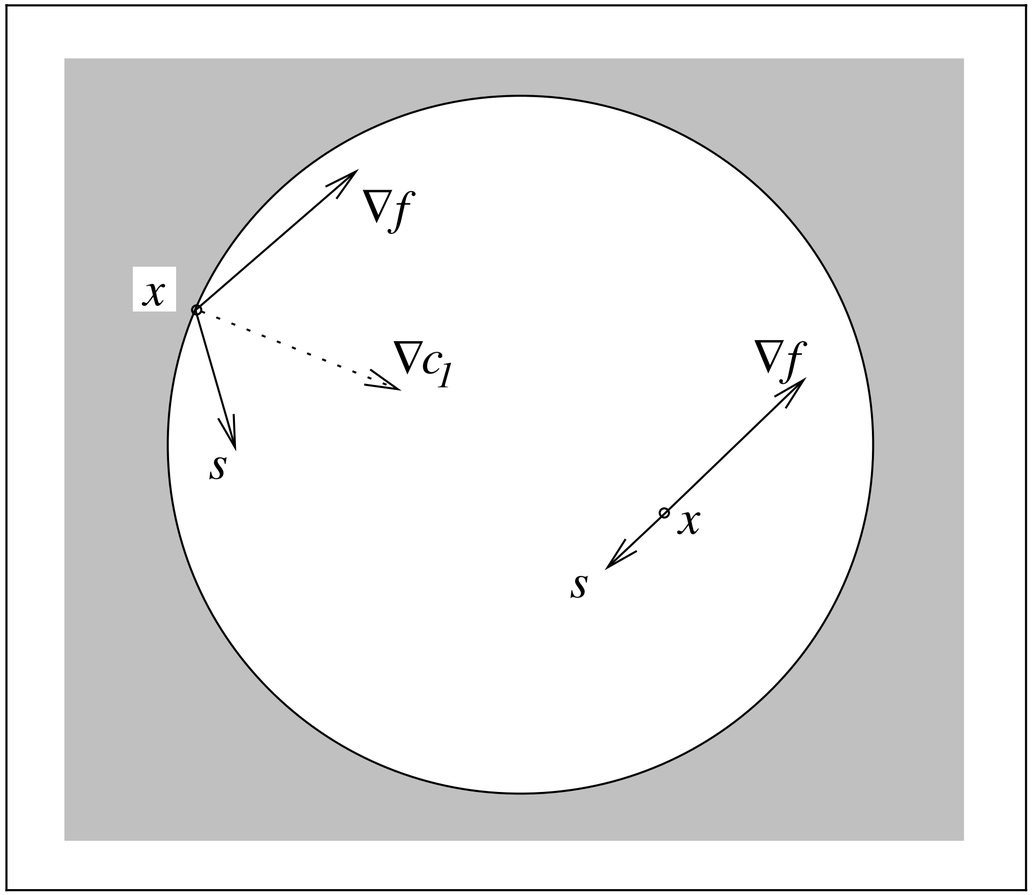
x^* est une solution locale du problème ssi $x^* \in \Omega$ et il existe un ensemble ouvert $U \subset \mathbf{R}^n$ contenant x^* tel que pour tout $x \in U \cap \Omega$, $f(x^*) \leq f(x)$

Condition nécessaire d'optimalité: intuition

$$\nabla f(x^*) = \lambda_1^* \nabla c_1(x^*).$$



Condition nécessaire d'optimalité: intuition



$$\nabla f(x^*) = \lambda_1^* \nabla c_1(x^*).$$

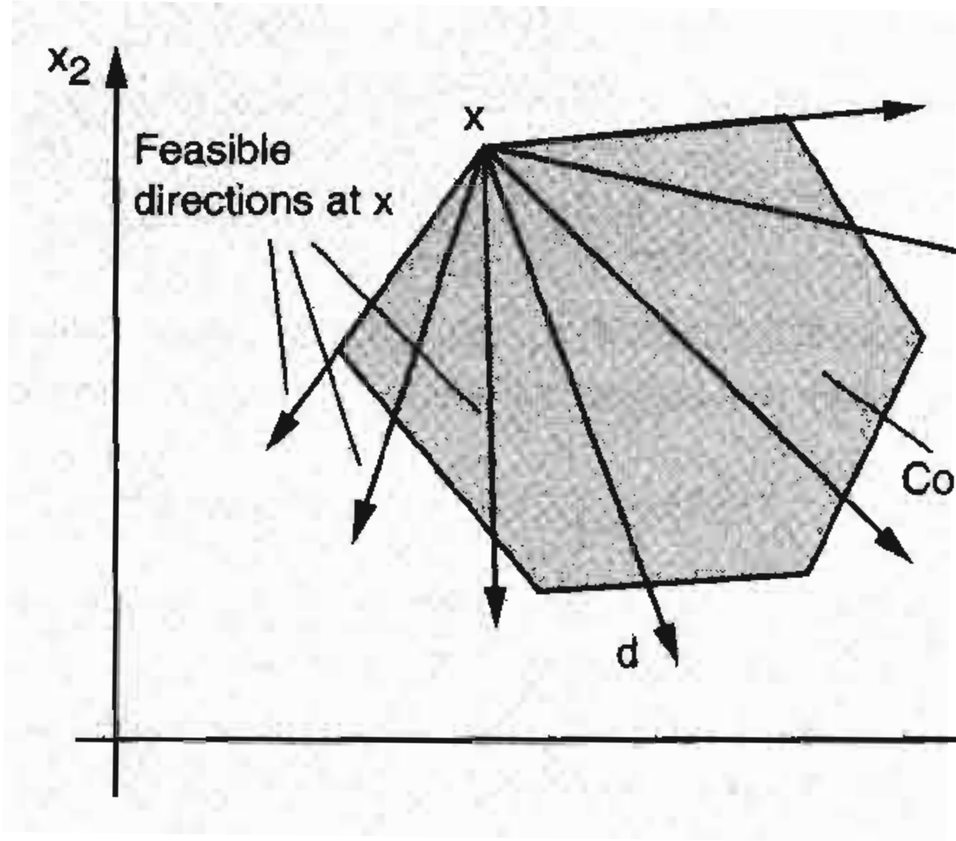
$$\lambda_1^* c_1(x^*) = 0.$$

Condition nécessaire d'optimalité: KKT

$$\mathcal{L}(x, \lambda) = f(x) - \sum_{i \in \mathcal{E} \cup \mathcal{I}} \lambda_i c_i(x).$$

$$\begin{aligned}
\nabla_x \mathcal{L}(x^*, \lambda^*) &= 0, \\
c_i(x^*) &= 0, \quad \text{for all } i \in \mathcal{E}, \\
c_i(x^*) &\geq 0, \quad \text{for all } i \in \mathcal{I}, \\
\lambda_i^* &\geq 0, \quad \text{for all } i \in \mathcal{I}, \\
\lambda_i^* c_i(x^*) &= 0, \quad \text{for all } i \in \mathcal{E} \cup \mathcal{I}.
\end{aligned}$$

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

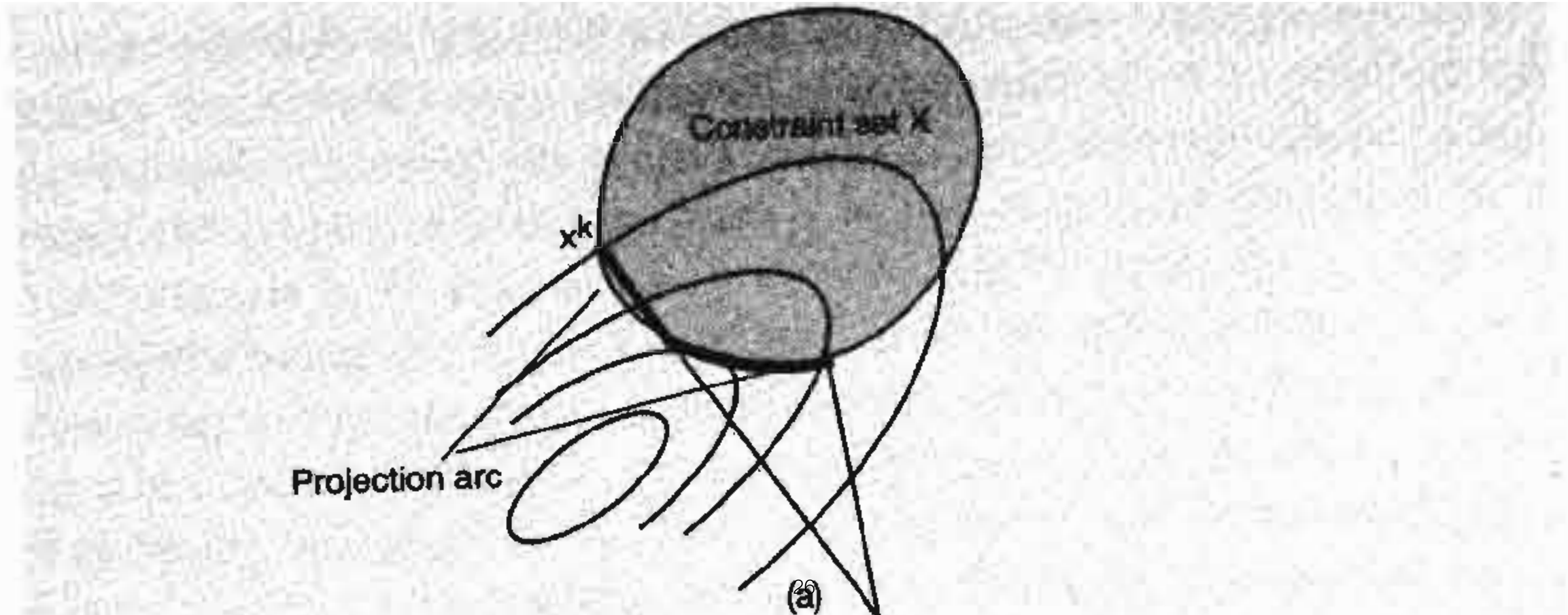
Iteration: $x_{k+1} := p_k(\beta^{m_k} s)$

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

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