Mathématiques pour l'intelligence artificielle

M2 informatique parcours IAAA, Aix-Marseille Université

Thomas Schatz
Vendredi 8 septembre 2023

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

Optimisation sans contraintes

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

$$\min_x f(x) \ ?$$

x^* est un minimum global de f ssi pour tout $x \in \mathbf{R}^n$, $f(x^*) \leq f(x)$

x^* est un minimum local de f ssi il existe un ensemble ouvert $U \subset \mathbf{R}^n$ contenant x^* tel que pour tout $x \in U$, $f(x^*) \leq f(x)$

Analyse

Concept central $\lim_{x \to a} f(x) = l \ ?$

Notions de base de topologie

Dans $\mathbf{R}^n$, La boule fermée $B_f(x, \epsilon)$, centrée sur x et de rayon ϵ associée à la norme Euclidienne est l'ensemble des points y de $\mathbf{R}^n$ tels que $\|x - y\|_2 \leq \epsilon$

Dans $\mathbf{R}^n$, La boule ouverte $B_o(x, \epsilon)$, centrée sur x et de rayon ϵ associée à la norme Euclidienne est l'ensemble des points y de $\mathbf{R}^n$ tels que $\|x - y\|_2 < \epsilon$

U est un ouvert de $\mathbf{R}^n$ ssi $U \subset \mathbf{R}^n$ et pour tout x dans U, il existe $\epsilon > 0$, tel que $B_f(x, \epsilon) \subset U$.

U est un fermé de $\mathbf{R}^n$ ssi son complément dans $\mathbf{R}^n$ est un ouvert de $\mathbf{R}^n$

Jacobienne, gradient

$$f : U \subset \mathbf{R}^n \to \mathbf{R}^m, \text{ de classe } \mathrm{C}^1$$

Dérivée partielle $f : x \mapsto (f_1(x), f_2(x), \ldots, f_m(x))$

$$\frac{\partial f_i}{\partial x_j} : (a_1, \ldots, a_n) \mapsto \lim_{h \to 0} \frac{f_i(a_1, \ldots, a_{j-1}, a_j + h, a_{j+1}, \ldots a_n) - f_i(a_1, \ldots, a_j, \ldots, a_n)}{h}$$

Matrice Jacobienne (transposée du gradient si m=1)

$$\mathbf{J} = \begin{bmatrix} \dfrac{\partial \mathbf{f}}{\partial x_1} & \cdots & \dfrac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^{\mathrm{T}} f_1 \\ \vdots \\ \nabla^{\mathrm{T}} f_m \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_1} & \cdots & \dfrac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f_m}{\partial x_1} & \cdots & \dfrac{\partial f_m}{\partial x_n} \end{bmatrix}$$

“Chain rule”

$$J_{f \circ g}(\mathbf{a}) = J_f(g(\mathbf{a}))J_g(\mathbf{a}),$$

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

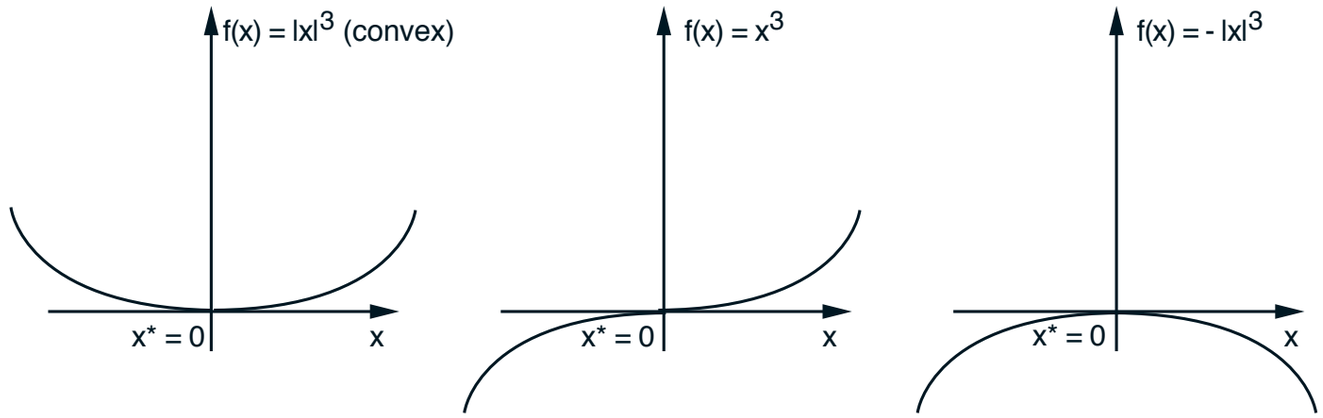
$$f : U \subset \mathbf{R}^n \mapsto \mathbf{R}$$

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

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 $\mathbf{R}^n$.

Théorème

Si f est convexe et minorée, alors f admet 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.

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.

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. $(X = \mathrm{dom} f)$

Optimisation

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

Optimisation sous contraintes

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

$$\Omega = \{x \in \mathbf{R}^n \mid \text{pour tout } i \in \mathcal{E},\ c_i(x) = 0,\ \text{pour tout } j \in \mathcal{I},\ c_j(x) \geq 0\}$$

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

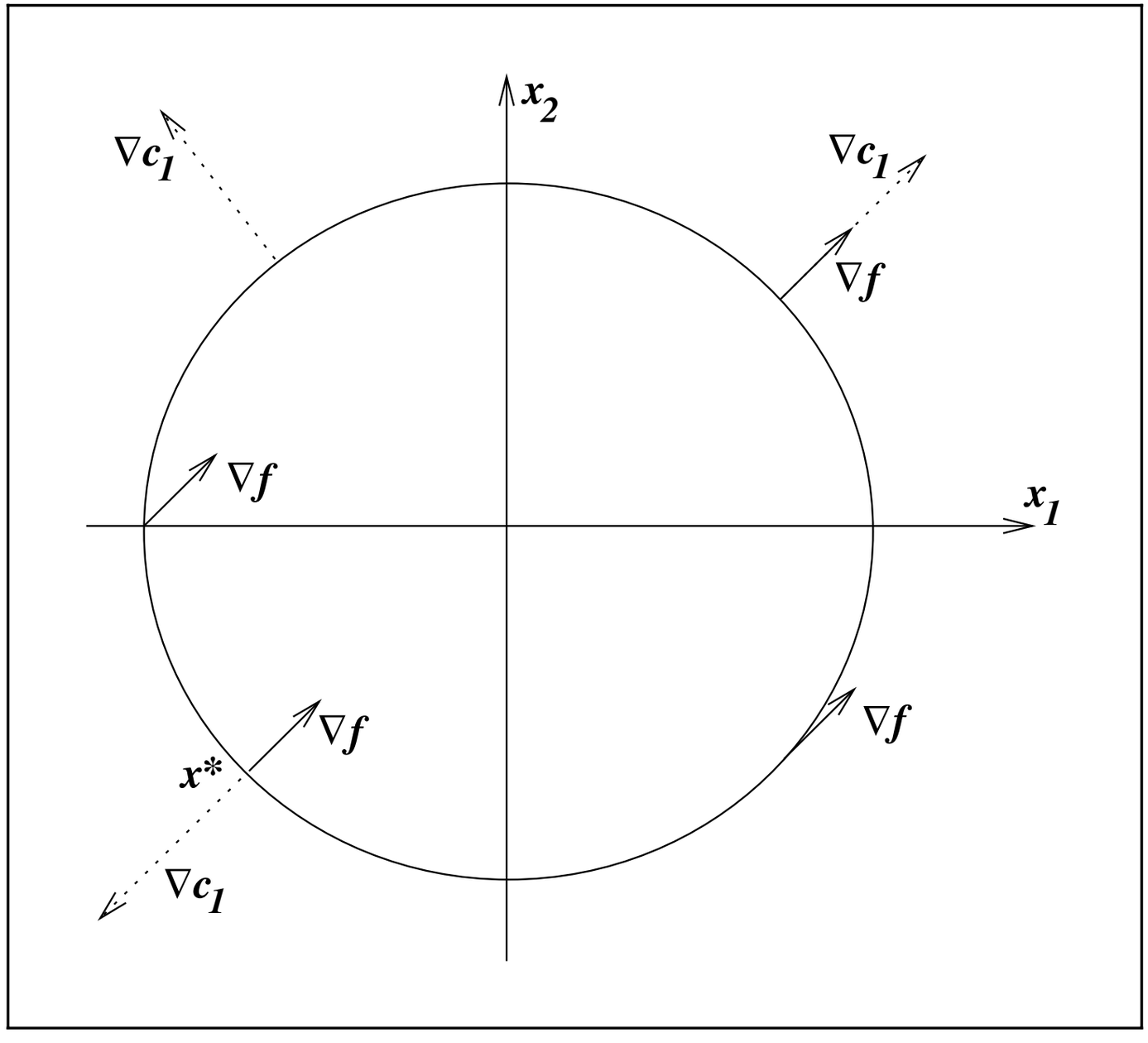
$$f : x_1, x_2 \mapsto x_1 + x_2$$

Cas d'une seule contrainte d'égalité

$$x_1^2 + x_2^2 = 2$$

Condition nécessaire d'optimalité

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



Condition nécessaire d'optimalité: intuition

$$f : x_1, x_2 \mapsto x_1 + x_2$$

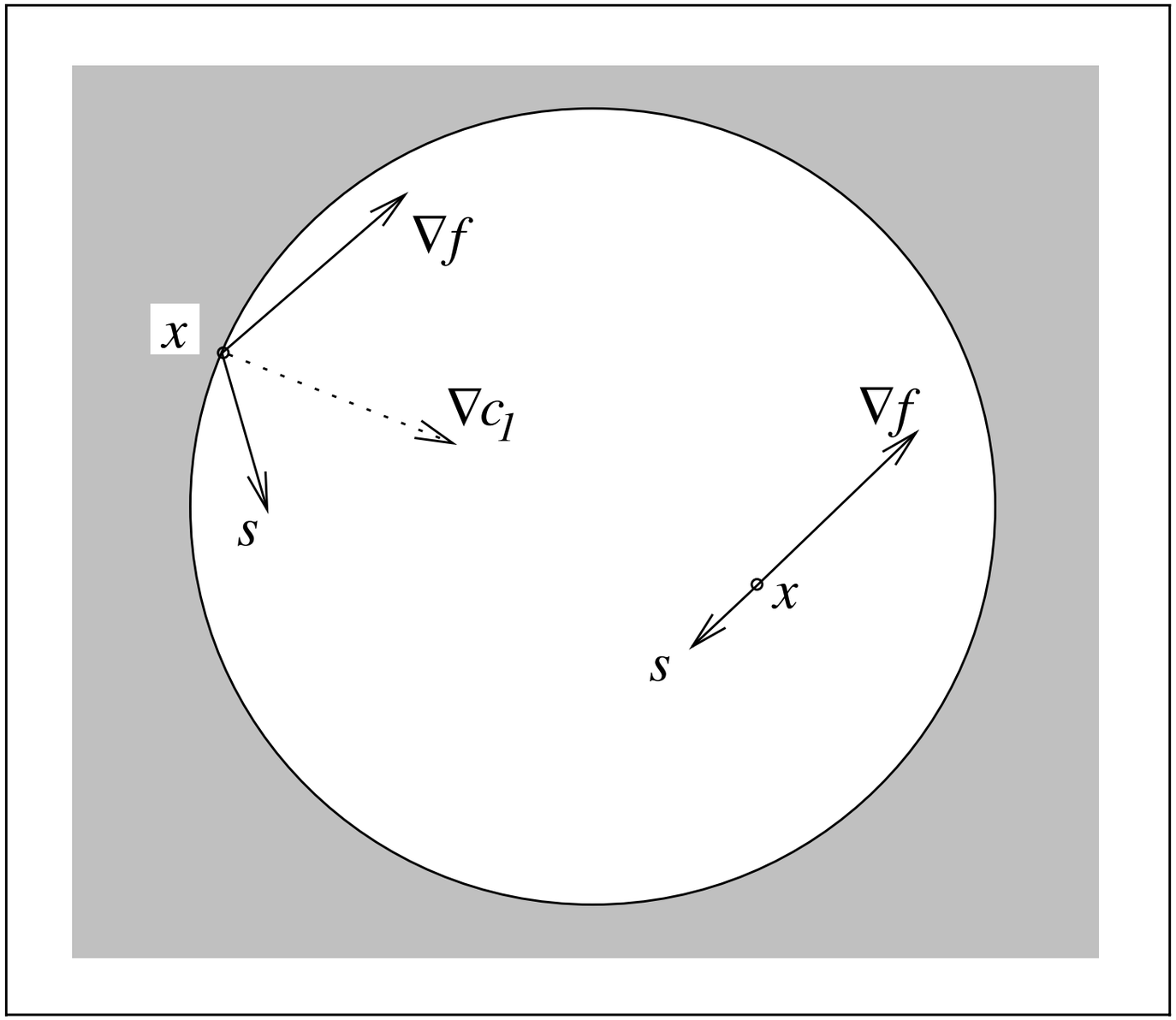
Cas d'une seule contrainte d'inégalité

$$x_1^2 + x_2^2 \leq 2$$

Condition nécessaire d'optimalité

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

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



Condition nécessaire d'optimalité: cas général

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

Conditions de Karush-Kuhn-Tucker (KKT)

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

Optimisation

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

Dualité

Problème considéré

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in X, \qquad g_j(x) \leq 0, \quad j = 1, \ldots, r, \end{aligned}$$

$$f : \mathbf{R}^n \to \mathbf{R}$$

$$g_j : \mathbf{R}^n \to \mathbf{R}$$

$$X \subset \mathbf{R}^n$$

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

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

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

Dualité

Lagrangien associé au problème

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

Dualité

Fonction duale

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

Problème dual

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

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

Résultats

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

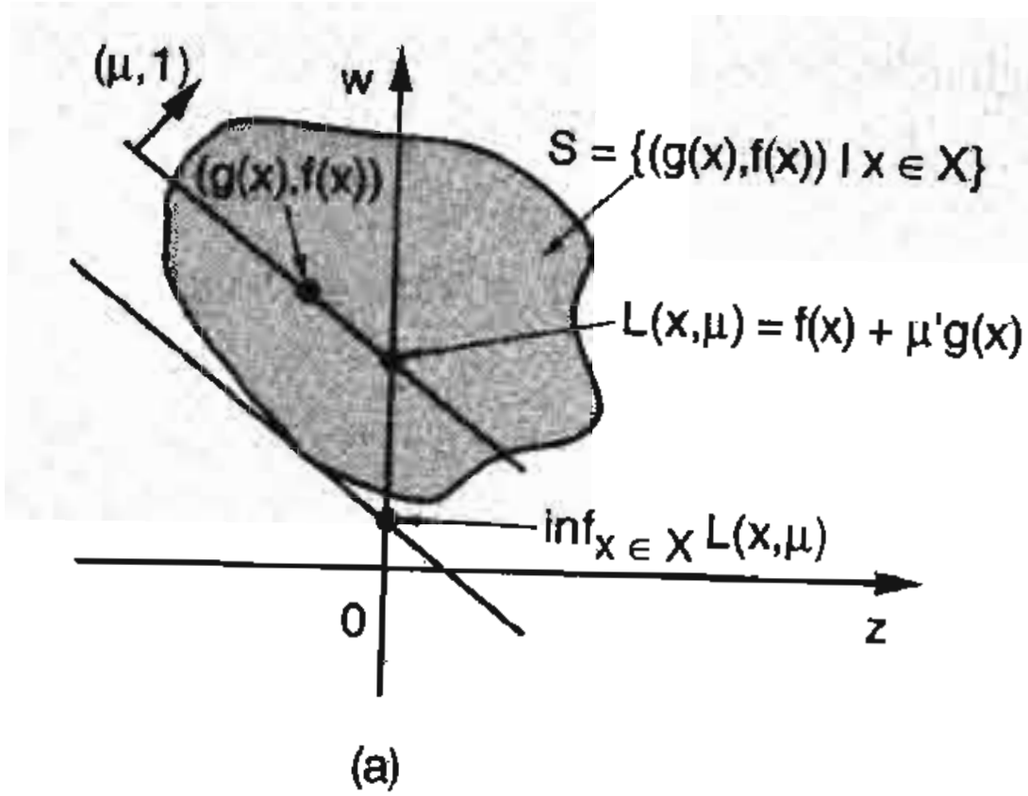
$$q^* \leq f^*.$$

Dualité

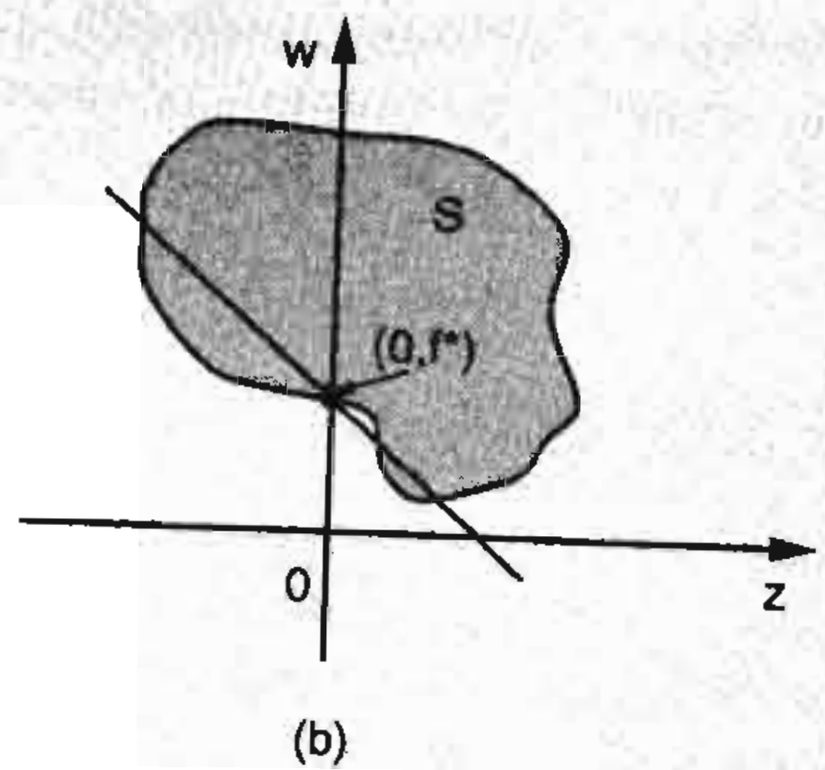
Dualité forte et qualification des contraintes

Example :

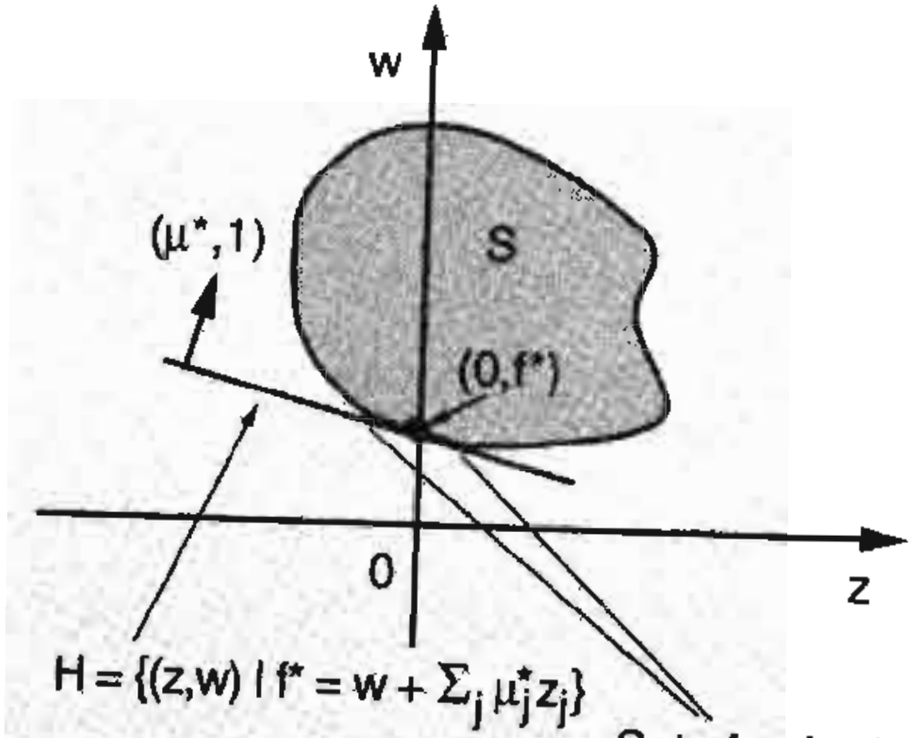
Si f est convexe et les contraintes g_j sont linéaires, alors il n'y a pas d'écart dual: $q^* = f^*$



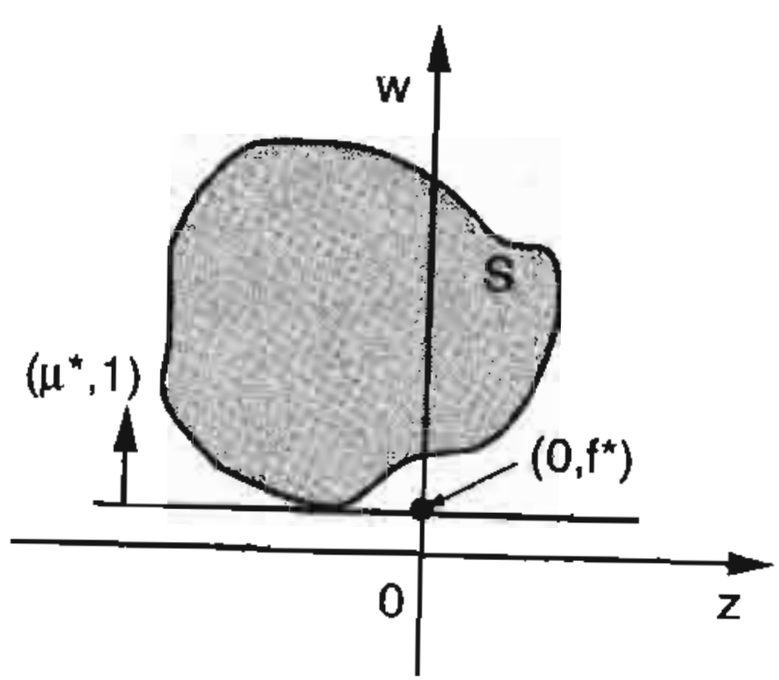
(a)



(b)



(c)



(d)

Set of pairs $(g(\bar{x}),f(\bar{x}))$ corresponding to $\bar{x}$ that minimize $L(x, \mu^*)$ over X

Dualité

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

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

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

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

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

Dualité

Perspective plus générale sur la dualité en optimisation convexe (et notamment la dualité Lagrangienne comme cas particulier de la dualité de Fenchel)

Chapter 3 from Ekeland, I., & Temam, R. (1999). *Convex analysis and variational problems*. SIAM

Références générales sur l'optimisation et l'analyse convexe

Boyd, S. P., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Rockafellar, R. T. (1997). *Convex analysis*. Princeton university press.

Optimisation

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

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

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

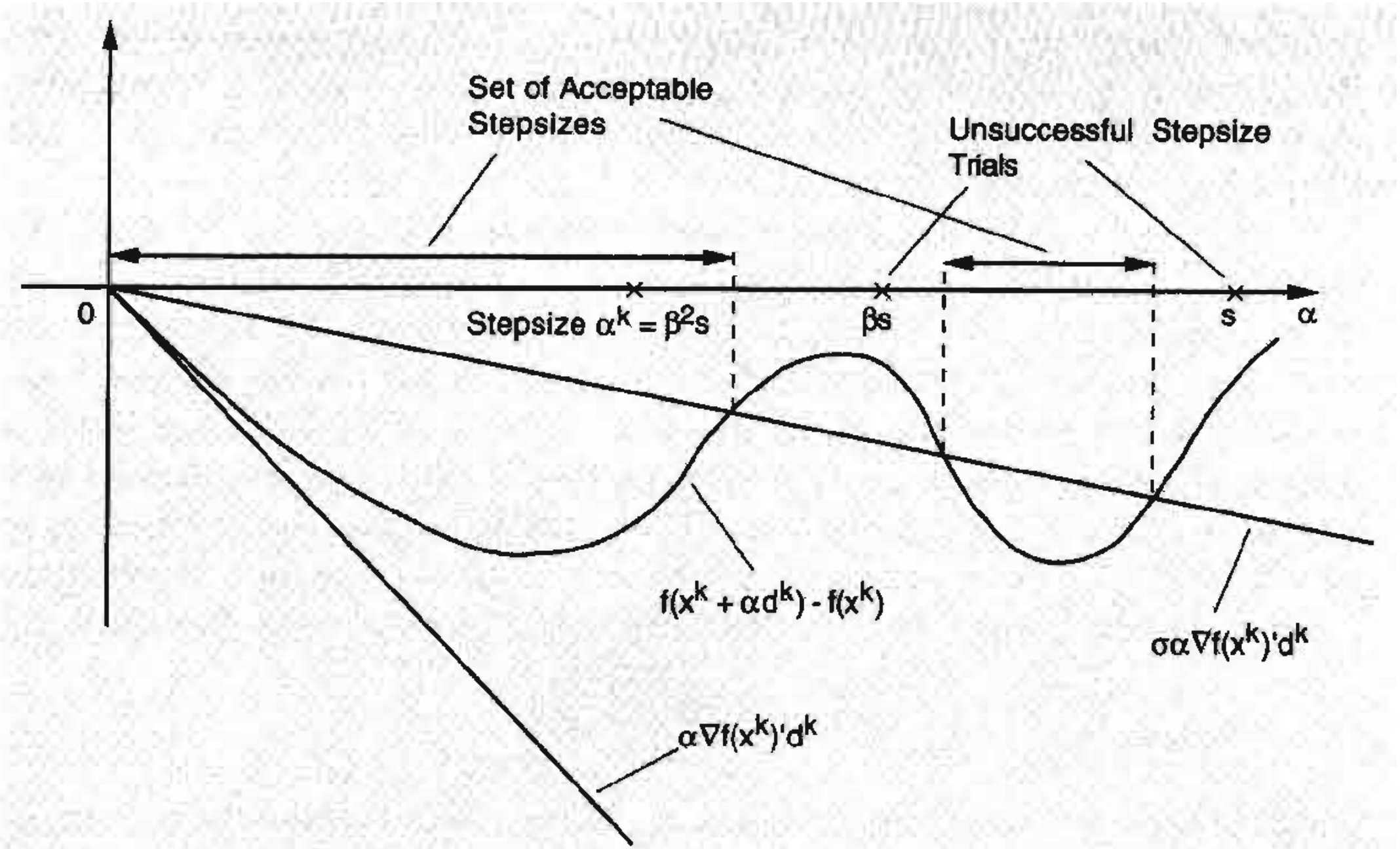
$$\alpha_k = \beta^{m_k} s$$

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

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

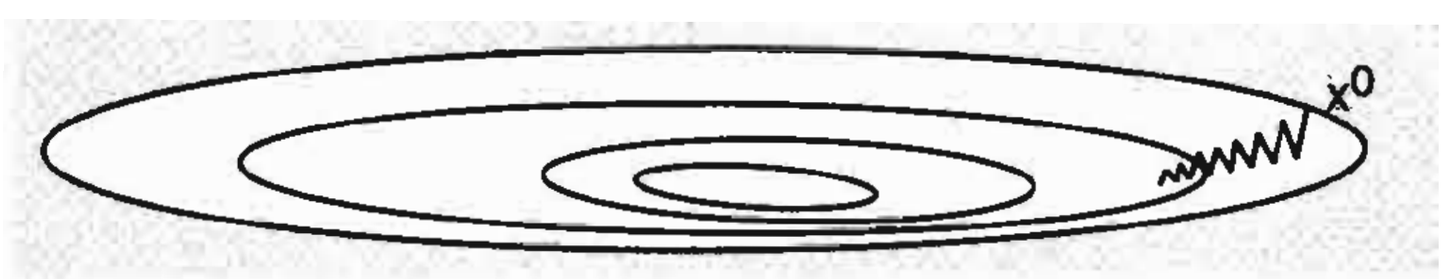
Descente de gradient

Convergence

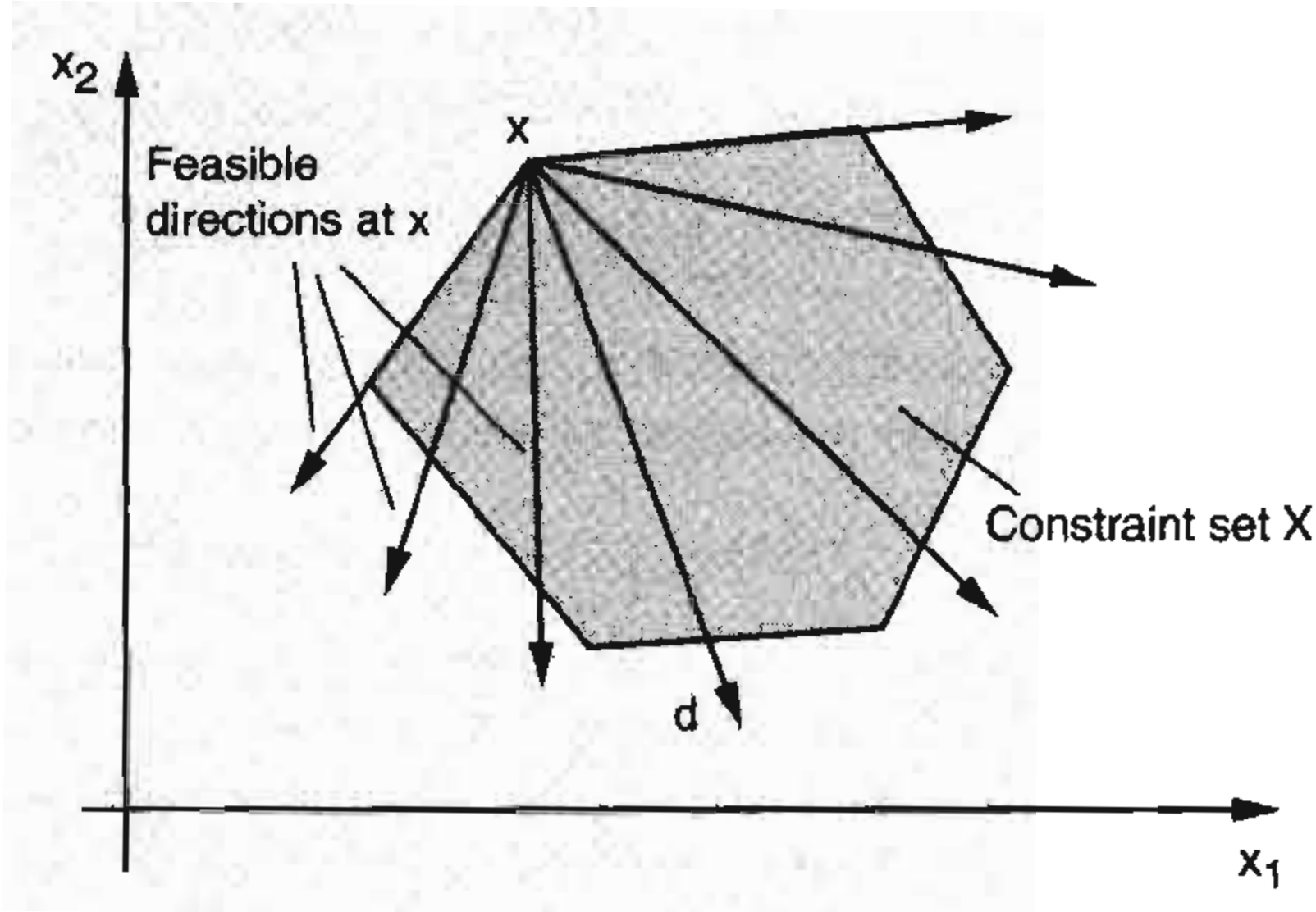
Soit (x_k) la suite des points générés par l'algorithme de descente de gradient avec pas choisi par la règle d'Armijo. Alors, tout point limite (i.e. valeur d'adhérence) de (x_k) est un point stationnaire.

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

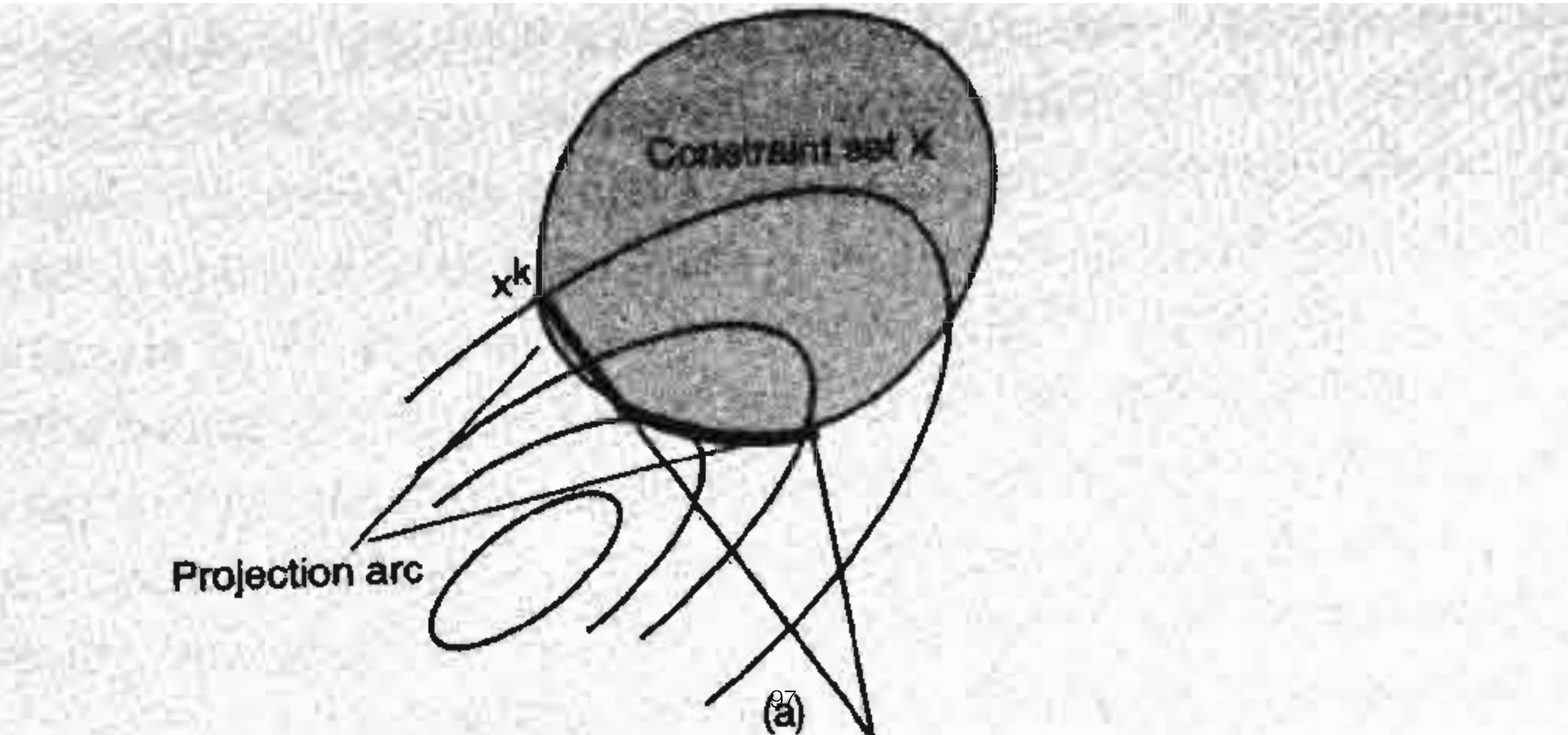
Iteration: $x_{k+1} := p_k(\beta^{m_k} s)$

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

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

Plan du cours (prévisionnel)

Introduction

1. Notions de bases sur les preuves

2. Algèbre linéaire

3. Optimisation

4. Probabilités

5. Statistique

Discussion