

# **COURS OPTIMISATION**

*Cours en Master M1 SITN*

Ionel Sorin CIUPERCA

# Table des matières

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Quelques rappels de calcul différentiel, analyse convexe et extremum</b>	<b>5</b>
2.1	Rappel calcul différentiel . . . . .	5
2.1.1	Quelques Notations . . . . .	5
2.1.2	Quelques rappels sur le calcul différentiel . . . . .	6
2.1.3	Rappel formules de Taylor . . . . .	10
2.1.4	Quelque rappels sur le matrices carrées réelles . . . . .	11
2.2	Convexité . . . . .	11
2.2.1	Fonctions convexes, strictement convexes, fortement convexes . . . . .	11
2.2.2	Exemples des fonctions convexes, strictement convexes et fortement convexes . . . . .	16
2.2.3	Fonctions coercives . . . . .	17
2.3	Conditions nécessaires et suffisantes de minimum . . . . .	17
2.4	Existence et unicité d'un point de minimum . . . . .	21
<b>3</b>	<b>Optimisation sans contraintes</b>	<b>23</b>
3.1	Méthodes de relaxation . . . . .	24
3.1.1	Description de la méthode . . . . .	24
3.1.2	Cas particulier des fonctions quadratiques . . . . .	27
3.2	Méthodes de gradient . . . . .	28
3.2.1	Méthodes de gradient à pas optimal . . . . .	29
3.2.2	Autres méthodes du type gradient . . . . .	30
3.3	La méthode des gradients conjugués . . . . .	33
3.3.1	Le cas quadratique . . . . .	33
3.3.2	Cas d'une fonction $J$ quelconque . . . . .	38
<b>4</b>	<b>Optimisation avec contraintes</b>	<b>39</b>
4.1	Rappel sur les multiplicateurs de Lagrange . . . . .	40
4.2	Optimisation sous contraintes d'inégalités . . . . .	41
4.2.1	Conditions d'optimalité de premier ordre : multiplicateurs de Karush-Kuhn-Tucker . . . . .	41
4.2.2	Théorie générale du point selle . . . . .	49

4.2.3	Applications de la théorie du point selle à l'optimisation . . . . .	51
4.2.4	Le cas convexe . . . . .	52
4.3	Algorithmes de minimisation avec contraintes . . . . .	53
4.3.1	Méthodes de relaxation . . . . .	54
4.3.2	Méthodes de projection . . . . .	54
4.3.3	Méthodes de pénalisation extérieure . . . . .	59
4.3.4	Méthode d'Uzawa . . . . .	61

# Chapitre 1

## Introduction

En général **optimiser** signifie le fait de chercher une configuration optimale d'un système, c'est à dire, chercher la meilleure configuration parmi tous les configurations possibles du système et ceci, par rapport à un critère donné.

Pour décrire (et éventuellement résoudre) un problème d'optimisation nous utilisons la **modélisation mathématique**. La démarche de modélisation comporte 3 étapes :

*Etape 1.* Choisir les **variables de décision**, qui sont les composantes du système sur lesquelles on peut agir. On supposera dans ce cours qu'il y a un nombre fini noté  $n \in \mathbb{N}^*$  de variables de décision, chacune de ces variables étant un nombre réel. Alors les variables de décision seront représentés par un vecteur  $x = (x_1, x_2, \dots, x_n)^T \in \mathbb{R}^n$  (vecteur colonne).

*Etape 2.* Décrire l'**état** du système, étant donnée une configuration des variables de décision. Ceci revient mathématiquement à se donner une fonction  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  qui s'appelle **fonction objectif** ou **fonction coût** et que nous voulons rendre la plus petite possible ou la plus grande possible.

*Etape 3.* Décrire les **contraintes** que les variables de décision satisfont. Ceci revient à définir un ensemble de contraintes  $U \subset \mathbb{R}^n$  et imposer d'avoir  $x \in U$ .

Pour résumer on peut dire que pour décrire un problème d'optimisation on se donne

1. Une fonction  $J : \mathbb{R}^n \mapsto \mathbb{R}$  (fonction coût)
2. Un ensemble  $U \subset \mathbb{R}^n$  (ensemble des contraintes)

**On cherche à minimiser**  $J$  sur  $U$ , c'est à dire, on cherche  $x^* \in U$  tel que

$$J(x^*) = \min_{x \in U} J(x)$$

ou équivalent

$$J(x^*) \leq J(x), \quad \forall x \in U.$$

**Motivation et exemples pratiques** : en classe

# Chapitre 2

## Quelques rappels de calcul différentiel, analyse convexe et extremum

### 2.1 Rappel calcul différentiel

#### 2.1.1 Quelques Notations

1. Pour tout  $n \in \mathbb{N}^*$ ,  $\mathbb{R}^n$  désigne l'espace **euclidien**  $\mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}$  ("produit  $n$  fois"). En général un vecteur  $x \in \mathbb{R}^n$  sera noté  $x = (x_1, x_2, \cdots, x_n)^T$  (vecteur colonne).
2. On note  $e_1, e_2, \cdots, e_n$  les éléments de la **base canonique** de  $\mathbb{R}^n$ , où  $e_i$  est le vecteur de  $\mathbb{R}^n$  donné par :

$$(e_i)_j = \delta_{ij} = \begin{cases} 0 & \text{si } j \neq i \\ 1 & \text{si } j = i \end{cases}, \quad \forall i, j = 1, 2, \cdots, n \quad (2.1)$$

( $\delta_{ij}$  = symboles de **Kronecker**).

3. Pour tous  $x, y \in \mathbb{R}^n$  on note par  $\langle x, y \rangle \in \mathbb{R}$  le **produit scalaire** de  $x$  et  $y$ , qui est donné par

$$\langle x, y \rangle = \sum_{i=1}^n x_i y_i.$$

Deux vecteurs  $x, y \in \mathbb{R}^n$  sont **orthogonaux** (on notera  $x \perp y$ ) si  $\langle x, y \rangle = 0$ .

4. Pour tout  $x \in \mathbb{R}^n$  on note par  $\|x\| \geq 0$  la **norme euclidienne** de  $x$ , donnée par

$$\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{i=1}^n x_i^2}.$$

Rappelons les **propriétés d'une norme** (donc aussi de la norme euclidienne) :

- $\|\lambda x\| = |\lambda| \|x\| \quad \forall \lambda \in \mathbb{R}, \forall x \in \mathbb{R}^n$
- $\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in \mathbb{R}^n$
- $\|0\| = 0$  et  $\|x\| > 0$  si  $x \in \mathbb{R}^n - \{0\}$ .

5. Pour tous  $x \in \mathbb{R}^n$  et  $r > 0$  on notera par  $B(x, r)$  la **boule ouverte** du centre  $x$  et rayon  $r$ , donnée par

$$B(x, r) = \{y \in \mathbb{R}^n, \quad \|y - x\| < r\}.$$

6. Si  $\{x^{(k)}\}_{k \in \mathbb{N}}$  est une suite dans  $\mathbb{R}^n$  et  $x$  est un élément de  $\mathbb{R}^n$  on dit que  $x^{(k)}$  **converge** vers  $x$  (notée  $x^{(k)} \rightarrow x$ ) si  $\|x^{(k)} - x\| \rightarrow 0$ .

Rappelons que nous avons :  $x^{(k)} \rightarrow x$  si et seulement si  $x_i^{(k)} \rightarrow x_i$  en  $\mathbb{R}$  où  $x_i^{(k)}$  (respectivement  $x_i$ ) est la  $i$ -ème composante de  $x^{(k)}$  (respectivement  $x$ ).

7. Soit  $U \subset \mathbb{R}^n$ .

- On définit **l'intérieur** de  $U$  comme l'ensemble des éléments  $x \in U$  pour lesquels il existe  $r > 0$  tel que  $B(x, r) \subset U$ .

- On dit que  $U$  est **ouvert** si  $\forall x \in U \exists r > 0$  tel que  $B(x, r) \subset U$ .

- On dit que  $U$  est **fermé** si pour toute suite  $\{x^{(k)}\} \subset U$  tel que  $x^{(k)} \rightarrow x \in \mathbb{R}^n$  on a  $x \in U$ .

8. Si  $a, b \in \mathbb{R}^n$  on note  $[a, b]$  le sous-ensemble de  $\mathbb{R}^n$  donné par

$$[a, b] = \{a + t(b - a) \equiv (1 - t)a + tb, \quad t \in [0, 1]\}.$$

L'ensemble  $[a, b]$  est aussi appelé **le segment** reliant  $a$  à  $b$ .

**Remarques :**

·  $[a, b] = [b, a]$  (Exo!)

· Si  $a, b \in \mathbb{R}$  avec  $a < b$  on retrouve la notation  $[a, b]$  pour l'intervalle des nombres  $x \in \mathbb{R}$  tels que  $a \leq x \leq b$ .

9. Rappelons aussi l'inégalité de Cauchy-Schwarz :

$$|\langle x, y \rangle| \leq \|x\| \cdot \|y\| \quad \forall x, y \in \mathbb{R}^n.$$

## 2.1.2 Quelques rappels sur le calcul différentiel

On considère dans cette partie  $m$  et  $n$  deux nombres de  $\mathbb{N}^*$  (très souvent dans ce cours on aura  $m = 1$ ).

1. Soit  $U$  un sous-ensemble de  $\mathbb{R}^n$  et  $f : U \mapsto \mathbb{R}^m$ .

On dit que  $f$  est **continu** en  $x \in U$  si  $f(x^{(k)}) \rightarrow f(x)$  pour toute suite  $x^{(k)} \subset U$  telle que  $x^{(k)} \rightarrow x$ .

On dit que  $f$  est continue sur  $U$  si  $f$  est continue en tout point  $x \in U$ .

**Remarque :** Si  $f = (f_1, f_2, \dots, f_m)$  avec  $f_1, f_2, \dots, f_m : U \rightarrow \mathbb{R}$  alors  $f$  est continu en  $x \in U$  si et seulement si  $f_1, f_2, \dots, f_m$  sont continues en  $x$ .

*Pour tous les points suivants on va supposer que  $\Omega$  est un ouvert de  $\mathbb{R}^n$  et  $f$  est une fonction  $f : \Omega \rightarrow \mathbb{R}^m$ .*

2. Pour tout  $x \in \Omega$  et  $h \in \mathbb{R}^n$  on note (quand  $\exists$ )

$$\frac{\partial f}{\partial h}(x) = \lim_{t \rightarrow 0} \frac{1}{t} [f(x + th) - f(x)]$$

(c'est la **dérivée directionnelle** de  $f$  en  $x$  dans la direction  $h$ ).

**Remarques :**

i)  $\frac{\partial f}{\partial 0}(x) = 0$ .

ii) Si  $f = (f_1, f_2, \dots, f_n)^T \in \mathbb{R}^n$  avec  $f_1, f_2, \dots, f_m : \Omega \rightarrow \mathbb{R}$  alors

$$\frac{\partial f}{\partial h}(x) = \left( \frac{\partial f_1}{\partial h}(x), \frac{\partial f_2}{\partial h}(x), \dots, \frac{\partial f_m}{\partial h}(x) \right)^T.$$

3. Pour tout  $x \in \Omega$  et tout  $i \in \{1, 2, \dots, n\}$  on note (quand  $\exists$ )

$$\frac{\partial f}{\partial x_i}(x) = \frac{\partial f}{\partial e_i}(x) = \lim_{t \rightarrow 0} \frac{1}{t} [f(x + te_i) - f(x)]$$

(c'est la **dérivée partielle** de  $f$  en  $x$  par rapport à la variable  $x_i$ .)

En particulier, si  $n = 1$  on note  $f'(x) = \frac{\partial f}{\partial x_1}(x) = \lim_{t \rightarrow 0} \frac{1}{t} [f(x + t) - f(x)] = \lim_{y \rightarrow x} \frac{1}{y-x} [f(y) - f(x)]$

4. Pour tout  $x \in \Omega$  on note (quand  $\exists$ )  $J_f(x) =$  la **matrice Jacobienne** de  $f$  en  $x$  qui est un élément de  $\mathcal{M}_{m,n}(\mathbb{R})$  définie par

$$(J_f(x))_{ij} = \frac{\partial f_i}{\partial x_j}(x) \in \mathbb{R} \quad \forall i = 1, \dots, m, \quad \forall j = 1, \dots, n.$$

Le **gradient** de  $f$  en  $x$  est défini comme la transposée de la matrice Jacobienne de  $f$  en  $x$  :

$$\nabla f(x) = (J_f(x))^T \in \mathcal{M}_{n,m}(\mathbb{R}).$$

**Remarque importante :** Dans le cas particulier  $m = 1$  (donc  $f : \Omega \rightarrow \mathbb{R}$ ) alors en considérant tout élément de  $\mathcal{M}_{n,1}$  comme un vecteur colonne de  $\mathbb{R}^n$ , on va dire que  $\nabla f(x)$  est le vecteur colonne

$$\nabla f(x) = \left( \frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots, \frac{\partial f}{\partial x_n} \right)^T \in \mathbb{R}^n.$$

Rappelons la formule :

$$\frac{\partial f}{\partial h}(x) = \langle \nabla f(x), h \rangle \quad \forall x \in \Omega \quad \forall h \in \mathbb{R}^n.$$

5. Si  $f : \Omega \rightarrow \mathbb{R}$  (ici  $m = 1$ ) on dit qu'un point  $x \in \Omega$  est un **point critique** pour la fonction  $f$  si  $\nabla f(x) = 0$ .

6. Pour tout  $x \in \Omega$  et  $i, j \in \{1, 2, \dots, n\}$  on note (quand  $\exists$ )

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial}{\partial x_i} \left( \frac{\partial f}{\partial x_j} \right) (x) \in \mathbb{R}^m$$

dérivée partielle à l'ordre 2.

**Notation :** pour  $i = j$  on écrira  $\frac{\partial^2 f}{\partial^2 x_i}(x)$  à la place de  $\frac{\partial^2 f}{\partial x_i \partial x_i}(x)$ .

7. Dans le cas  $m = 1$  on note pour tout  $x \in \Omega$  (quand  $\exists$ )  $\nabla^2 f(x) =$  la matrice carrée  $\in \mathcal{M}_n(\mathbb{R})$  donnée par

$$(\nabla^2 f(x))_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j}(x), \quad \forall i, j = 1, 2, \dots, n.$$

( $\nabla^2 f(x)$  s'appelle aussi **la matrice Hessienne** de  $f$  en  $x$ ).

8. On dit que  $f$  est de classe  $C^p$  sur  $\Omega$  (on notera  $f \in C^p(\Omega)$ ) pour  $p = 1$  ou  $p = 2$  si les dérivées partielles des  $f$  jusqu'à l'ordre  $p$  existent et sont continues sur  $\Omega$ . Par extension on dit que  $f$  est de classe  $C^0$  sur  $\Omega$  si  $f$  est continue sur  $\Omega$ .

9. On a le Théorème de Schwarz : si  $f \in C^2(\Omega)$  alors

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial^2 f}{\partial x_j \partial x_i}(x) \quad \forall x \in \Omega, \quad \forall i, j = 1, \dots, n$$

(c'est à dire, la matrice  $\nabla^2 f(x)$  est symétrique).

10. (*Lien entre  $\nabla$ ,  $J_f$  et  $\nabla^2$* ) : Si  $f : \Omega \rightarrow \mathbb{R}$  est de classe  $C^2$  alors

$$\nabla^2 f(x) = J_{\nabla f}(x) = \nabla J_f(x) \quad \forall x \in \Omega$$

(la matrice Hessienne de  $f$  est le Jacobien du gradient de  $f$  ou le gradient de la Jacobienne de  $f$ ).

11. (*Composition*) Soient  $\Omega \subset \mathbb{R}^n$ ,  $U \subset \mathbb{R}^m$  avec  $\Omega, U$  ouverts  $f : \Omega \rightarrow \mathbb{R}^m$ ,  $g : U \rightarrow \mathbb{R}^p$  avec  $p \in \mathbb{N}^*$  et  $f(\Omega) \subset U$ . Considérons la fonction composée  $g \circ f : \Omega \rightarrow \mathbb{R}^p$ .

*i)* Si  $f$  et  $g$  sont continues alors  $g \circ f$  est continue.

*ii)* Si  $f$  et  $g$  sont de classe  $C^1$  alors  $g \circ f$  est de classe  $C^1$  et on a l'égalité matricielle

$$J_{g \circ f}(x) = J_g(f(x)) \cdot J_f(x) \quad \forall x \in \Omega.$$

**Conséquences :**

*i)* Si  $m = p = 1$  alors

$$\nabla(g \circ f)(x) = g'(f(x)) \nabla f(x).$$

*i)* Si  $n = p = 1$  alors

$$(g \circ f)'(x) = \langle \nabla g(f(x)), f'(x) \rangle.$$



**Proposition 2.1.** *Nous avons*

$$\nabla^2 f(x)h = \nabla \langle \nabla f(x), h \rangle \quad \forall x \in \Omega, \forall h \in \mathbb{R}^n.$$

où le premier gradient dans le membre de droite de l'égalité est considéré par rapport à la variable  $x$ .

*Démonstration.* On a :

$$\frac{\partial}{\partial x_i} \langle \nabla f(x), h \rangle = \frac{\partial}{\partial x_i} \left( \sum_{j=1}^n \frac{\partial f}{\partial x_j}(x) h_j \right) = \sum_{j=1}^n \frac{\partial^2 f}{\partial x_i \partial x_j}(x) h_j = (\nabla^2 f(x)h)_i.$$

□

**Quelques exemples importants :**

1. Si  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  est une fonction **constante** alors  $\nabla f = 0$  et  $J_f = 0$ . On a aussi évidemment  $\nabla^2 f = 0$  dans le cas  $m = 1$ .
2. Soit  $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$  définie par

$$f(x) = Ax \quad \forall x \in \mathbb{R}^n$$

où  $A \in \mathcal{M}_{m,n}(\mathbb{R})$  est une matrice donnée (c'est à dire,  $f$  est une fonction **linéaire**). Il est facile de voir qu'on a

$$J_f(x) = A \quad \forall x \in \mathbb{R}^n$$

(la matrice Jacobienne est constante).

Dans la cas particulier  $m = 1$  une fonction linéaire générale peut être écrite sous la forme

$$f(x) = \langle a, x \rangle \quad \forall x \in \mathbb{R}^n$$

où  $a \in \mathbb{R}^n$  est un vecteur donné. Il est clair alors que

$$\nabla f = a$$

et

$$\nabla^2 f = 0.$$

3. Soit  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  donnée par

$$f(x) = \langle Ax, x \rangle \quad \forall x \in \mathbb{R}^n,$$

où  $A \in \mathcal{M}_n(\mathbb{R})$  est une matrice carrée, réelle, de taille  $n$  (c'est à dire,  $f$  est la **forme quadratique** associée à la matrice  $A$ ). Alors pour un  $p \in \{1, 2, \dots, n\}$  fixé, on peut écrire

$$f(x) = \sum_{i,j=1}^n A_{ij} x_i x_j = A_{pp} x_p^2 + \sum_{j=1, j \neq p}^n A_{pj} x_p x_j + \sum_{i=1, i \neq p}^n A_{ip} x_i x_p + \sum_{i,j=1, i \neq p, j \neq p}^n A_{ij} x_i x_j$$

ce qui nous donne

$$\frac{\partial f}{\partial x_p} = 2A_{pp}x_p + \sum_{j=1, j \neq p}^n A_{pj}x_j + \sum_{i=1, i \neq p}^n A_{ip}x_i = \sum_{j=1}^n A_{pj}x_j + \sum_{i=1}^n A_{ip}x_i = (Ax)_p + (A^T x)_p.$$

Nous avons donc obtenu :

$$\nabla f(x) = (A + A^T)x, \quad \forall x \in \mathbb{R}^n.$$

En utilisant la formule  $\nabla^2 f = J_{\nabla f}$  on déduit

$$\nabla^2 f(x) = A + A^T, \quad \forall x \in \mathbb{R}^n$$

(donc la hessienne de  $f$  est constante).

**Remarque :** En particulier, si  $A$  est **symétrique** (c'est à dire  $A = A^T$ ) alors

$$\nabla \langle Ax, x \rangle = 2Ax, \quad \forall x \in \mathbb{R}^n.$$

$$\nabla^2 \langle Ax, x \rangle = 2A, \quad \forall x \in \mathbb{R}^n.$$

### 2.1.3 Rappel formules de Taylor

**Proposition 2.2.** (sans preuve)

Soit  $\Omega \subset \mathbb{R}^n$  ouvert,  $f : \Omega \mapsto \mathbb{R}$ ,  $a \in \Omega$  et  $h \in \mathbb{R}^n$  tels que  $[a, a + h] \subset \Omega$ . Alors :

1. Si  $f \in C^1(\Omega)$  alors

$$i) f(a + h) = f(a) + \int_0^1 \langle \nabla f(a + th), h \rangle dt$$

(formule de Taylor à l'ordre 1 avec reste intégral).

$$ii) f(a + h) = f(a) + \langle \nabla f(a + \theta h), h \rangle \text{ avec } 0 < \theta < 1$$

(formule de Taylor - Maclaurin à l'ordre 1)

$$iii) f(a + h) = f(a) + \langle \nabla f(a), h \rangle + o(\|h\|)$$

(formule de Taylor - Young à l'ordre 1)

2. Si  $f \in C^2(\Omega)$  alors

$$i) f(a + h) = f(a) + \langle \nabla f(a), h \rangle + \int_0^1 (1-t) \langle \nabla^2 f(a + th)h, h \rangle dt$$

(formule de Taylor à l'ordre 2 avec reste intégral).

$$ii) f(a + h) = f(a) + \langle \nabla f(a), h \rangle + \frac{1}{2} \langle \nabla^2 f(a + \theta h)h, h \rangle \text{ avec } 0 < \theta < 1$$

(formule de Taylor - Maclaurin à l'ordre 2)

$$iii) f(a + h) = f(a) + \langle \nabla f(a), h \rangle + \frac{1}{2} \langle \nabla^2 f(a)h, h \rangle + o(\|h\|^2)$$

(formule de Taylor - Young à l'ordre 2).

**Remarque :** Dans la proposition précédente la notation  $o(\|h\|^k)$  pour  $k \in \mathbb{N}^*$  signifie une expression qui tend vers 0 plus vite que  $\|h\|^k$  (c'est à dire, si on la divise par  $\|h\|^k$ , le résultat tend vers 0 quand  $\|h\|$  tend vers 0).

## 2.1.4 Quelques rappels sur les matrices carrées réelles

Soit  $n \in \mathbb{N}^*$  et  $A \in \mathcal{M}_n(\mathbb{R})$  une matrice carrée réelle.

1. Soit  $C =$  l'ensemble des nombres complexes. On rappelle que  $\lambda \in C$  est une **valeur propre** de  $A$  s'il existe  $x \in C^n$  avec  $x \neq 0$  tel que  $Ax = \lambda x$ ; on appelle  $x$  **vecteur propre** de  $A$  associé à la valeur propre  $\lambda$ .
2. On dit que la matrice  $A$  est **semi-définie positive** si  $\langle Ax, x \rangle \geq 0, \quad \forall x \in \mathbb{R}^n$ .  
On dit que  $A$  est **définie positive** si  $\langle Ax, x \rangle > 0, \quad \forall x \in \mathbb{R}^n$  avec  $x \neq 0$ .
3. Rappelons que si  $A$  est symétrique alors toutes les valeurs propres de  $A$  sont réelles; en plus il existe  $n$  vecteurs propres de  $A$  appartenant à  $\mathbb{R}^n$  formant une base ortho-normée en  $\mathbb{R}^n$ .
4. Supposons que la matrice  $A$  est symétrique. Alors

$$\langle Ah, h \rangle \geq \lambda_{\min} \|h\|^2, \quad \forall h \in \mathbb{R}^n$$

où  $\lambda_{\min} \in \mathbb{R}$  est la plus petite valeur propre de  $A$ .

Rémarquons que l'inégalité précédente devient égalité si  $h$  est un vecteur propre associé à la valeur propre  $\lambda_{\min}$ .

5. Supposons que  $A$  est symétrique. Alors  $A$  est semi-définie positive si et seulement si  $\lambda_{\min} \geq 0$  et  $A$  est définie positive si et seulement si  $\lambda_{\min} > 0$ .
6. **Abréviation :** La notation SDP pour une matrice carrée réelle signifie "matrice symétrique et définie positive" (elle ne signifie pas "matrice semi-définie positive"!).

## 2.2 Convexité

### 2.2.1 Fonctions convexes, strictement convexes, fortement convexes

**Définition 2.3.** Un ensemble  $U \subset \mathbb{R}^n$  est dit **convexe** si  $\forall x, y \in U$  on a  $[x, y] \subset U$  (quelque soit deux points dans  $U$ , tout le segment qui les unit est dans  $U$ ).

**Définition 2.4.** Soit  $U \subset \mathbb{R}^n$  un ensemble convexe et  $f : U \rightarrow \mathbb{R}$  une fonction.

1. On dit que  $f$  est **convexe** sur  $U$  si

$$f(ty + (1-t)x) \leq tf(y) + (1-t)f(x), \quad \forall x, y \in U, \quad \forall t \in [0, 1]$$

2. On dit que  $f$  est **strictement convexe** sur  $U$  si

$$f(ty + (1-t)x) < tf(y) + (1-t)f(x), \quad \forall x, y \in U \text{ avec } x \neq y, \quad \forall t \in ]0, 1[.$$

3. On dit que  $f$  est **fortement convexe** sur  $U$  s'il existe  $\alpha > 0$  tel que

$$f(ty + (1-t)x) \leq tf(y) + (1-t)f(x) - \alpha t(1-t)\|y - x\|^2, \quad \forall x, y \in U, \quad \forall t \in [0, 1]$$

4. On dit que  $f$  est **concave** (respectivement **strictement concave**, respectivement **fortement concave**) si  $-f$  est convexe (respectivement strictement convexe, respectivement fortement convexe).

**Remarque :** Il est facile de voir qu'on a : fortement convexe  $\implies$  strictement convexe  $\implies$  convexe. Les réciproques ne sont pas vraies en général ; par exemple une application affine  $f(x) = Ax + b$  est convexe (et aussi concave) mais elle n'est pas strictement convexe (ni strictement concave) donc elle n'est pas fortement convexe (ni fortement concave).

On a le résultat utile suivant :

**Proposition 2.5.** Soit  $U \subset \mathbb{R}^n$  un ensemble convexe,  $p \in \mathbb{N}^*$ ,  $f_1, f_2, \dots, f_p : U \rightarrow \mathbb{R}$  des fonctions convexes et  $\gamma_1, \gamma_2, \dots, \gamma_n$  des constantes strictement positives.

Posons  $f = \gamma_1 f_1 + \gamma_2 f_2 + \dots + \gamma_p f_p$ . Alors on a :

1. La fonction  $f$  est convexe (donc toute combinaison linéaire avec des coefficients strictement positifs de fonctions convexes est convexe).
2. Si au moins l'une des fonctions  $f_1, \dots, f_p$  est strictement convexe alors  $f$  est strictement convexe.
3. Si au moins l'une des fonctions  $f_1, \dots, f_p$  est fortement convexe alors  $f$  est fortement convexe.

*Démonstration.* Laissée en exercice ! □

Il est en général difficile de vérifier la convexité d'une fonction en utilisant uniquement la définition (essayez avec  $f(x) = x^2$  ou avec  $f(x) = x^4$  !) Les propositions suivantes donnent des critères de convexité, convexité stricte et convexité forte, plus faciles à utiliser que les définitions respectives.

**Proposition 2.6.** (caractérisation de la convexité)

Soit  $\Omega \subset \mathbb{R}^n$  ouvert,  $U \subset \Omega$  avec  $U$  convexe et  $f : \Omega \rightarrow \mathbb{R}$  une fonction de classe  $C^1$ . Alors

a) Les 3 propositions suivantes sont équivalentes :

1.  $f$  est convexe sur  $U$

2.

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in U$$

3.  $\nabla f$  est **monotone** sur  $U$ , c'est à dire

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq 0 \quad \forall x, y \in U.$$

b) Si de plus  $f$  est de classe  $C^2$  sur  $\Omega$  alors  $f$  est convexe sur  $U$  si et seulement si

$$\langle \nabla^2 f(x)(y - x), y - x \rangle \geq 0, \quad \forall x, y \in U \tag{2.2}$$

*Démonstration.* **a)** On montre ici l'équivalence entre 1), 2) et 3).

1)  $\implies$  2) : Supposons  $f$  convexe ; la définition de la convexité peut s'écrire

$$f(x + t(y - x)) - f(x) \leq t[f(y) - f(x)]$$

En fixant  $x, y$  en divisant par  $t$  et en faisant  $t$  tendre vers 0 (ce qui est possible car  $t \in [0, 1]$ ) on obtient 2).

2)  $\implies$  3) : De 2) on déduit

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in U$$

et aussi (en inversant  $x$  et  $y$ ) :

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle, \quad \forall x, y \in U$$

En faisant la somme de ces 2 inégalités on obtient 3).

3)  $\implies$  1) : Soient  $x, y \in U$  fixés. On introduit la fonction  $g : I \rightarrow \mathbb{R}$  définie par

$$t \in I \rightarrow g(t) = f(ty + (1 - t)x) \in \mathbb{R}$$

où  $I$  est un intervalle ouvert qui contient  $[0, 1]$ . Il est facile de voir que  $g$  est de classe  $C^1$  et on a

$$g'(t) = \langle \nabla f(ty + (1 - t)x), y - x \rangle \quad \forall t \in I.$$

Soient  $t_1, t_2 \in [0, 1]$  avec  $t_1 < t_2$ . Alors

$$\begin{aligned} g'(t_2) - g'(t_1) &= \langle \nabla f(x + t_2(y - x)) - \nabla f(x + t_1(y - x)), y - x \rangle = \\ &\langle \nabla f(x + t_2(y - x)) - \nabla f(x + t_1(y - x)), (t_2 - t_1)(y - x) \rangle > \frac{1}{t_2 - t_1}. \end{aligned}$$

Par hypothèse 3) le dernier term de l'égalité précédente est  $\geq 0$ , ce qui montre que la fonction  $g'$  est une fonction croissante. On déduit alors que  $g$  est une fonction convexe sur  $[0, 1]$ , ce qui nous donne pour tout  $t \in [0, 1]$  :

$$g(t \cdot 1 + (1 - t) \cdot 0) \leq tg(1) + (1 - t)g(0)$$

c'est à dire

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

donc  $f$  est convexe.

**b)** On suppose ici  $f \in C^2(\Omega)$ .

" $\implies$ " Supposons que  $f$  est convexe et montrons (2.2). Soit  $h \in \mathbb{R}^n$  fixé et notons  $g : \Omega \mapsto \mathbb{R}$  la fonction donnée par  $g(x) = \langle \nabla f(x), h \rangle, \quad \forall x \in \Omega$ . Nous avons en utilisant aussi la Proposition 2.1 :

$$\langle \nabla^2 f(x)h, h \rangle = \langle \nabla g(x), h \rangle = \frac{\partial g}{\partial h}(x) = \lim_{t \rightarrow 0} \frac{\langle \nabla f(x + th), h \rangle - \langle \nabla f(x), h \rangle}{t}$$

ce qui nous donne

$$\langle \nabla^2 f(x)h, h \rangle = \lim_{t \rightarrow 0} \frac{\langle \nabla f(x + th) - \nabla f(x), th \rangle}{t^2}.$$

Considérons maintenant  $x, y \in U$  arbitraires et  $h = y - x$ . Comme  $x + t(y - x) \in U \quad \forall t \in [0, 1]$ , de l'égalité précédente on déduit à l'aide de la monotonie de  $\nabla f$  que  $\langle \nabla^2 f(x)h, h \rangle \geq 0$ , c'est à dire (2.2).

“ $\Leftarrow$ ” Supposons maintenant que (2.2) est satisfaite et montrons que  $f$  est convexe. Soient  $x, y \in U$  fixées arbitraires, et considérons la fonction  $g_1 : \Omega \mapsto \mathbb{R}$  donnée par  $g_1(z) = \langle \nabla f(z), x - y \rangle \quad \forall z \in \Omega$ . Alors

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle = g_1(x) - g_1(y) = \langle \nabla g_1(y + \theta(x - y)), x - y \rangle$$

avec  $\theta \in ]0, 1[$  (on a utilisé l'une des formules de Taylor).

D'autre part, nous avons

$$\nabla g_1(z) = \nabla^2 f(z)(x - y)$$

et ceci nous permet de déduire, en utilisant aussi (2.2) :

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle = \langle \nabla^2 f(y + \theta(x - y))(x - y), x - y \rangle \geq 0.$$

Ceci nous donne la monotonie de  $\nabla f$  donc la convexité de  $f$ . □

**Proposition 2.7.** (caractérisation de la convexité stricte)

Soit  $\Omega \subset \mathbb{R}^n$  ouvert,  $U \subset \Omega$  avec  $U$  convexe et  $f : \Omega \mapsto \mathbb{R}$  une fonction de classe  $C^1$ . Alors

a) Les 3 propositions suivantes sont équivalentes :

1.  $f$  est strictement convexe sur  $U$

2.

$$f(y) > f(x) + \langle \nabla f(x), y - x \rangle, \quad \forall x, y \in U \text{ avec } x \neq y$$

3.  $\nabla f$  est **strictement monotone sur**  $U$ , c'est à dire

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle > 0 \quad \forall x, y \in U \text{ avec } x \neq y.$$

b) Si de plus  $f$  est de classe  $C^2$  sur  $\Omega$  alors si

$$\langle \nabla^2 f(x)(y - x), y - x \rangle > 0, \quad \forall x, y \in U \text{ avec } x \neq y \tag{2.3}$$

alors  $f$  est strictement convexe sur  $U$ .

**Proposition 2.8.** (caractérisation de la convexité forte)

Soit  $\Omega \subset \mathbb{R}^n$  ouvert,  $U \subset \Omega$  avec  $U$  convexe et  $f : \Omega \mapsto \mathbb{R}$  une fonction de classe  $C^1$ . Alors

a) Les 3 propositions suivantes sont équivalentes :

1.  $f$  est fortement convexe sur  $U$

2. Il existe  $\beta > 0$  tel que

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \beta \|y - x\|^2, \quad \forall x, y \in U$$

3. Il existe  $\gamma > 0$  tel que

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \gamma \|y - x\|^2, \quad \forall x, y \in U.$$

**b)** Si de plus  $f$  est de classe  $C^2$  sur  $\Omega$  alors  $f$  est fortement convexe sur  $U$  si et seulement si il existe  $\beta > 0$  tel que

$$\langle \nabla^2 f(x)(y - x), y - x \rangle \geq \beta \|y - x\|^2, \quad \forall x, y \in U \quad (2.4)$$

Les démonstrations des Propositions 2.7 et 2.8 sont admises. (certains points sont des simples adaptations des preuves des points analogues de la Proposition 2.6.

**Remarques :**

1. Dans la Proposition 2.7 **b)** il n'y a pas d'équivalence entre la stricte convexité de  $f$  et l'inégalité (2.3) de cette proposition dans le cas  $f \in C^2$  (par exemple la fonction  $f : \mathbb{R} \rightarrow \mathbb{R}$  donnée par  $f(x) = x^4$  est strictement convexe, mais l'inégalité n'est pas satisfaite si  $x = 0$ ).
2. Dans le cas particulier où l'ensemble  $U$  est ouvert alors pour les 3 propositions précédentes les inégalités (2.2), (2.3) et (2.4) concernant la matrice Hessienne  $\nabla^2 f$  s'écrivent respectivement :

$$\langle \nabla^2 f(x)h, h \rangle \geq 0 \quad \forall x \in U, \quad \forall h \in \mathbb{R}^n \quad (\text{pour la Proposition 2.6}),$$

$$\langle \nabla^2 f(x)h, h \rangle > 0 \quad \forall x \in U, \quad \forall h \in \mathbb{R}^n \quad (\text{pour la Proposition 2.7}),$$

$$\langle \nabla^2 f(x)h, h \rangle \geq \beta \|h\|^2, \quad \forall x \in U, \quad \forall h \in \mathbb{R}^n \quad (\text{pour la Proposition 2.8}).$$

Pour voir ceci il suffit de remarquer que d'une part, pour tout  $y \in U$  on peut considérer  $h = y - x \in \mathbb{R}^n$  et d'autre part, que pour tout  $h \in \mathbb{R}^n$  on peut considérer  $y = x + th \in U$  avec  $t > 0$  assez petit.

3. Dans le cas particulier  $n = 1$  et  $U$  un intervalle dans  $\mathbb{R}$ , on a  $\nabla^2 f(x) = f''(x)$ , donc  $\langle \nabla^2 f(x)(y - x), y - x \rangle = f''(x)(y - x)^2$ . Alors les 3 inégalités précédentes peuvent s'écrire respectivement :

$$f''(x) \geq 0 \quad \forall x \in U, \quad f''(x) > 0 \quad \forall x \in U \quad \text{et} \quad f''(x) \geq \beta \quad \forall x \in U.$$

On retrouve alors une caractérisation bien connue pour la convexité et la convexité stricte.

**Définition 2.9.** On appelle **fonction elliptique** une fonction  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  de classe  $C^1$  et fortement convexe.

## 2.2.2 Exemples des fonctions convexes, strictement convexes et fortement convexes

1. Soit  $f : \mathbb{R} \rightarrow \mathbb{R}$  donnée par  $f(x) = x^2$ . Comme  $\nabla f(x) = f'(x) = 2x$  on a  $\forall x, y \in \mathbb{R}$  :

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle = [f'(x) - f'(y)] \cdot (x - y) = 2(x - y)^2 \equiv \beta \|x - y\|^2$$

avec  $\beta = 2$ , donc  $f$  est fortement convexe.

C'est encore plus facile si on utilise  $f''$  :

$$f''(x) = 2 \equiv \beta > 0$$

donc  $f$  est fortement convexe.

2. Plus généralement, si  $n = 1$  et  $\Omega \subset \mathbb{R}$  intervalle ouvert :  
Toute fonction  $f : \Omega \rightarrow \mathbb{R}$  de classe  $C^2$  et satisfaisant :

$$\exists \alpha > 0, \quad f''(x) \geq \alpha, \quad \forall x \in \Omega$$

est une fonction fortement convexe (c'est une conséquence immédiate de la Proposition 2.8). Si  $\Omega = \mathbb{R}$  alors la fonction  $f$  est elliptique.

**Exemples :**

i)  $f(x) = ax^2 + bx + c$  avec  $a > 0$

ii)  $f(x) = x^2 + \sin(x)$  (car  $f''(x) = 2 - \sin(x) \geq 1, \forall x \in \mathbb{R}$ ).

3. **Le cas général** ( $n \in \mathbb{N}^*$ )

Soit  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  donnée par

$$f(x) = \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle + c, \quad \forall x \in \mathbb{R}^n \quad (2.5)$$

avec  $A \in \mathcal{M}_n(\mathbb{R})$  une matrice carrée réelle et **symétrique** de taille  $n$ , avec  $b \in \mathbb{R}^n$  un vecteur et  $c \in \mathbb{R}$  un scalaire (*on appelle encore, par abus de langage, fonction (ou forme) quadratique une fonction de ce type*). On calcule facilement :

$$\nabla f(x) = Ax - b$$

$$\nabla^2 f(x) = A$$

(donc la hessienne de  $f$  est constante).

En utilisant les rappels de la Section 2.1.4 et les Propositions 2.6, 2.7 et 2.8 on déduit :

i)  $f$  est une fonction convexe  $\iff A$  est une matrice semi-définie positive

ii)  $f$  est fortement convexe  $\iff f$  est strictement convexe  $\iff A$  est une matrice définie positive (matrice SDP).



### 2.2.3 Fonctions coercives

**Définition 2.10.** Soit  $\Omega \subset \mathbb{R}^n$  un ensemble **non borné** et  $f : \Omega \rightarrow \mathbb{R}$ . On dit que  $f$  est **coercive sur  $\Omega$**  si on a

$$\lim_{x \in \Omega, \|x\| \rightarrow +\infty} f(x) = +\infty.$$

(ceci peut aussi s'écrire :

$$f(x^k) \rightarrow +\infty \quad \text{pour toute suite } \{x^k\}_{k \in \mathbb{N}} \subset \Omega \text{ telle que } \|x^k\| \rightarrow +\infty.)$$

**Remarque :** Supposons qu'on a  $\Omega_0 \subset \Omega \subset \mathbb{R}^n$  avec  $\Omega_0$  non borné et  $f : \Omega \rightarrow \mathbb{R}$ . Si  $f$  est coercive sur  $\Omega$  alors  $f$  est coercive sur  $\Omega_0$ , la réciproque n'étant pas vraie en général.

**Proposition 2.11.** Soit  $\Omega \subset \mathbb{R}^n$  un ouvert et  $U \subset \Omega$  un ensemble convexe et non borné. Soit  $f : \Omega \rightarrow \mathbb{R}$  une fonction de classe  $C^1$  sur  $\Omega$  et fortement convexe sur  $U$ . Alors  $f$  est coercive sur  $U$ .

*Démonstration.* Ecrivons l'inégalité 2. de la Proposition 2.8 avec  $x = a$  où  $a$  est un élément fixé de  $U$  :

$$f(y) = f(a) + \langle \nabla f(a), y - a \rangle + \beta \|y - a\|^2 \quad \forall y \in U$$

où  $\beta > 0$  est un nombre fixé. D'autre part, nous avons

$$\langle \nabla f(a), y - a \rangle \geq -\|\nabla f(a)\| \cdot \|y - a\|$$

(conséquence de l'inégalité de Cauchy-Schwartz). Ceci donne

$$f(y) \geq f(a) - \|\nabla f(a)\| \cdot \|y - a\| + \beta \|y - a\|^2 \quad \forall y \in U.$$

En faisant  $\|y\| \rightarrow +\infty$  (ce qui donne  $\|y - a\| \rightarrow +\infty$ ) on obtient le résultat. □

## 2.3 Conditions nécessaires et suffisantes de minimum

**Définition 2.12.** Soit  $U \subset \mathbb{R}^n$ ,  $u^* \in U$  et  $f : U \rightarrow \mathbb{R}$ .

1. On dit que  $u^*$  est un point de minimum **absolu** (ou **global**) de  $f$  sur  $U$  si

$$f(u) \geq f(u^*), \quad \forall u \in U.$$

2. On dit que  $u^*$  est un point de minimum **relatif** (ou **local**) de  $f$  sur  $U$  si  $\exists V$  voisinage de  $u^*$  dans  $\mathbb{R}^n$ , tel que

$$f(u) \geq f(u^*), \quad \forall u \in U \cap V.$$

3. On dit que  $u^*$  est un point de maximum absolu (respectivement relatif) de  $f$  sur  $U$  si  $u^*$  est un point de minimum absolu (respectivement relatif) de  $-f$  sur  $U$ .

4. On dit que  $u^*$  est un point d'**extremum** absolu (respectivement relatif) de  $f$  sur  $U$  si  $u^*$  est : soit un point de minimum absolu (respectivement relatif) de  $f$  sur  $U$ , soit un point de maximum absolu (respectivement relatif) de  $f$  sur  $U$ .

Dans toute la suite du cours on parlera uniquement de la minimisation d'une fonction  $f$  ; pour la maximisation, faire la minimisation de la fonction  $-f$ .

**Remarques :**

- Un point de minimum absolu est clairement un point de minimum relatif.
- Si on dit simplement *minimum* on comprends *minimum absolu*.

**Définition 2.13.** Soit  $U \subset \mathbb{R}^n$  et  $u^* \in U$ . On dit que le vecteur  $w \in \mathbb{R}^n$  est une **direction admissible** pour  $u^*$  en  $U$  s'il existe  $t_0 > 0$  tel que  $u^* + tw \in U \quad \forall t \in [0, t_0]$ .

On a le résultat suivant :

**Lemme 2.14.** Soit  $\Omega \subset \mathbb{R}^n$  un ouvert,  $U \subset \Omega$ ,  $f : \Omega \rightarrow \mathbb{R}$  une fonction de classe  $C^1$  et  $u^* \in U$  un point de minimum relatif de  $f$  sur  $U$ . Soit  $w \in \mathbb{R}^n$  une direction admissible pour  $u^*$  en  $U$ . Alors

$$\langle \nabla f(u^*), w \rangle \geq 0.$$

*Démonstration.* On peut supposer  $w \neq 0$  (sinon l'inégalité demandée est évidente). Soit  $V$  un voisinage de  $u^*$  en  $\mathbb{R}^n$  tel que

$$f(u) \geq f(u^*), \quad \forall u \in U \cap V. \quad (2.6)$$

Soit  $t_1 > 0$  tel que

$$u^* + tw \in V, \quad \forall t \in [-t_1, t_1]$$

(il suffit de prendre  $t_1 = \frac{r}{\|w\|}$  où  $r > 0$  est tel que  $B(u^*, r) \subset V$ ).

On déduit alors de (2.6)

$$f(u^* + tw) - f(u^*) \geq 0, \quad \forall t \in [0, \min\{t_0, t_1\}].$$

avec  $t_0$  comme dans la Définition 2.13.

En divisant par  $t$  et en passant à la limite  $t \rightarrow 0_+$ , on obtient :

$$\frac{\partial f}{\partial w}(u^*) \geq 0$$

ce qui est exactement l'inégalité demandée. □

Nous pouvons alors énoncer :

**Théorème 2.15.** Soit  $\Omega \subset \mathbb{R}^n$  un ouvert,  $U \subset \Omega$  un ensemble convexe et  $f : \Omega \rightarrow \mathbb{R}$  une fonction de classe  $C^1$ . Soit  $u^* \in U$  un point de minimum relatif de  $f$  sur  $U$ . Alors

$$\langle \nabla f(u^*), u - u^* \rangle \geq 0, \quad \forall u \in U. \quad (2.7)$$

(c'est l'inéquation d'Euler).

*Démonstration.* Pour tout  $u \in U$  le vecteur  $u - u^*$  est une direction admissible pour  $u^*$  en  $U$  car la convexité de  $U$  nous donne  $u^* + t(u - u^*) \in U \quad \forall t \in [0, 1]$ . Alors (2.7) est une conséquence immédiate de Lemme 2.14 avec  $w = u - u^*$ .  $\square$

On va considérer deux cas particuliers : le cas  $u^*$  à l'intérieur de  $U$  et la cas  $U$  sous-espace affine.

**Lemme 2.16.** *Soit  $U \subset \mathbb{R}^n$ ,  $b \in \mathbb{R}^n$  et  $u^*$  un élément appartenant à l'intérieur de  $U$  ( $u^* \in \overset{\circ}{U}$ ). Alors les deux assertions suivantes sont équivalentes :*

1.

$$\langle b, u - u^* \rangle \geq 0 \quad \forall u \in U. \quad (2.8)$$

2.

$$b = 0. \quad (2.9)$$

*Démonstration.* L'implication (2.9)  $\Rightarrow$  (2.8) est évidente.

Pour montrer l'implication inverse, soit  $w \in \mathbb{R}^n$  arbitraire, avec  $w \neq 0$ . Alors il existe  $\theta_0 > 0$  telle que

$$u^* + \theta w \in U, \quad \forall \theta \in [-\theta_0, \theta_0]$$

(il suffit de prendre  $\theta_0 = \frac{r}{\|w\|}$  où  $r > 0$  est tel que  $B(u^*, r) \subset U$ .) Alors en prenant  $u^* + \theta w$  à la place de  $u$  dans (2.8) on déduit

$$\langle a, \theta w \rangle \geq 0, \quad \forall \theta \in [-\theta_0, \theta_0].$$

En prenant d'abord  $\theta = \theta_0$  et ensuite  $\theta = -\theta_0$  dans l'inégalité précédente, on déduit

$$\langle a, w \rangle = 0 \quad \text{et ceci quelque soit } w \in \mathbb{R}^n, w \neq 0.$$

Comme l'égalité précédente est évidemment valable aussi pour  $w = 0$  alors elle est valable pour tout  $w \in \mathbb{R}^n$ . Ceci donne immédiatement (2.9) (il suffit de prendre  $w = b$  dans cette égalité).  $\square$

**Lemme 2.17.** *Soit  $U_0$  un sous espace vectoriel de  $\mathbb{R}^n$ ,  $a \in \mathbb{R}^n$  et  $U = a + U_0$  (on dit que  $U$  est un **sous-espace affine** de  $\mathbb{R}^n$ .) Soit  $b \in \mathbb{R}^n$  et  $u^* \in U$ . Alors les deux assertions suivantes sont équivalentes :*

1.

$$\langle b, u - u^* \rangle \geq 0, \quad \forall u \in U. \quad (2.10)$$

2.

$$\langle b, h \rangle = 0 \quad \forall h \in U_0. \quad (2.11)$$

*Démonstration.* L'implication (2.11)  $\implies$  (2.10) est immédiate du fait que pour tout  $u \in U$  on a  $u - u^* \in U_0$ .

Pour montrer (2.10)  $\implies$  (2.11) soit  $h \in U_0$  fixé arbitraire et remarquons que nous pouvons prendre en (2.10)  $u = u^* \pm h$  car c'est un élément de  $U$ . Ceci nous donne

$$\langle b, \pm h \rangle \geq 0$$

ce qui nous donne (2.11). □

En appliquant ces deux lemmes avec  $b = \nabla f(u^*)$  on obtient immédiatement le corollaire suivant :

**Corollaire 2.18.** *Soit  $\Omega \subset \mathbb{R}^n$  un ouvert,  $U \subset \Omega$  un ensemble convexe et  $f : \Omega \rightarrow \mathbb{R}$  une fonction de classe  $C^1$ . Soit  $u^* \in U$  un point de minimum relatif de  $f$  sur  $U$ . Alors*

1. *Si  $u^*$  est dans l'intérieur de  $U$  ( $u^* \in \overset{\circ}{U}$ ) alors*

$$\nabla f(u^*) = 0 \tag{2.12}$$

*(c'est l'équation d'Euler).*

2. *Si  $U$  est un sous-espace affine de  $\mathbb{R}^n$  (donc  $U = a + U_0$  avec  $a \in \mathbb{R}^n$  et  $U_0$  un sous-espace vectoriel de  $\mathbb{R}^n$ ) alors*

$$\langle \nabla f(u^*), h \rangle = 0 \quad \forall h \in U_0. \tag{2.13}$$

**Remarques :**

1. Ces relations respectives : (2.7), (2.12) et (2.13) donnent seulement des conditions **nécessaires** de minimum relatif, qui ne sont pas en général suffisantes.
2. Si  $U$  est un ensemble **ouvert** (par exemple si  $U$  est l'espace entier,  $U = \mathbb{R}^n$ ) alors  $u^*$  est forcément dans l'intérieur de  $U$  et donc l'équation d'Euler (2.12) peut être utilisée comme condition nécessaire de minimum relatif.

Le résultat suivant nous donne aussi des conditions suffisantes de minimum :

**Théorème 2.19.** *Soit  $\Omega \subset \mathbb{R}^n$  ouvert,  $U \subset \Omega$  un ensemble convexe,  $f : \Omega \rightarrow \mathbb{R}$  une fonction de classe  $C^1$  et **convexe** et  $u^* \in U$ . Alors les trois assertions suivantes sont équivalentes :*

1.  *$u^*$  est un point de minimum absolu de  $f$  sur  $U$*
2.  *$u^*$  est un point de minimum relatif de  $f$  sur  $U$*
3.  *$\langle \nabla f(u^*), u - u^* \rangle \geq 0, \quad \forall u \in U.$*

*Démonstration.* 1)  $\implies$  2) est évidente et 2)  $\implies$  3) est une conséquence immédiate du Théorème 2.15.

Montrons 3)  $\implies$  1). Grâce à la convexité de  $f$  et à la Partie I de la Proposition 2.6 nous avons

$$f(u) - f(u^*) \geq \langle \nabla f(u^*), u - u^* \rangle \quad \forall u \in U.$$

De 3) nous obtenons alors 1). □

**Remarque :** Dans le cas où  $u^* \in \overset{\circ}{U}$  alors l'assertion 3. du Théorème 2.19 peut être remplacée par l'équation d'Euler (2.12). Si  $U$  est un sous espace affine alors cette assertion peut être remplacée par (2.13).

Rappelons enfin le résultat bien connu sur les conditions d'optimalité de second ordre (c'est à dire, faisant intervenir la matrice hessienne).

**Théorème 2.20.** Soit  $\Omega \subset \mathbb{R}^n$  ouvert,  $U \subset \Omega$ ,  $f : \Omega \rightarrow \mathbb{R}$  une fonction de classe  $C^2$  et  $u^* \in \overset{\circ}{U}$ . Supposons que  $u^*$  est un point critique de  $f$  (c'est à dire  $\nabla f(u^*) = 0$ ). On a les résultats suivants :

1. Si  $u^*$  est un point de minimum local de  $f$  sur  $U$  alors la matrice  $\nabla^2 f(u^*)$  est semi-définie positive (condition nécessaire).
2. Si la matrice  $\nabla^2 f(u^*)$  est définie positive alors  $u^*$  est un point de minimum local de  $f$  sur  $U$  (condition suffisante).

## 2.4 Existence et unicité d'un point de minimum

**Théorème 2.21.** (Existence)

Soit  $U \subset \mathbb{R}^n$  un ensemble non-vide et fermé et  $f : U \rightarrow \mathbb{R}$  une fonction continue. On suppose

- Soit  $U$  est borné
- Soit  $U$  est non bornée et  $f$  est une fonction coercive.

Alors il existe **au moins** un point de minimum de  $f$  sur  $U$  (c'est à dire,  $\exists u^* \in U$  tel que  $f(u^*) \leq f(u)$ ,  $\forall u \in U$ )

*Démonstration.* On distingue deux cas :

**Cas 1)** L'ensemble  $U$  est borné.

Alors comme  $U$  est aussi fermé,  $U$  est compact. Comme  $f$  est continue, le Théorème de Weierstrass nous assure que  $f$  est bornée sur  $U$  et elle atteint ses bornes. Donc il existe au moins un point de minimum absolu de  $f$  sur  $U$ .

**Cas 2)** L'ensemble  $U$  est non borné.

Soit  $a \in U$  et considérons l'ensemble

$$E = \{x \in U, f(x) \leq f(a)\} \quad \text{Remarque : } a \in E.$$

Il est facile de montrer :

1.  **$E$  est fermé**  
(car  $E = f^{-1}(]-\infty, f(a)])$ , donc  $E$  est l'image inverse d'un intervalle fermé par une fonction continue)
2.  **$E$  est borné**  
(supposons le contraire : alors il existe une suite  $x_k \in E$  avec  $\|x_k\| \rightarrow +\infty$  pour  $k \rightarrow +\infty$ . Comme  $f$  est coercive, ceci entraîne  $f(x_k) \rightarrow +\infty$  ce qui est absurde, car  $f(x_k) \leq f(a)$ ,  $\forall k \in \mathbb{N}$ .)

On déduit alors que  $E$  est un ensemble compact dans  $\mathbb{R}^n$ . Du Théorème de Weierstrass,  $\exists u^* \in E$  tel que

$$f(u^*) \leq f(u), \quad \forall u \in E.$$

Mais d'autre part, on a

$$f(u^*) < f(u), \quad \forall u \in U - E$$

(car  $f(u^*) \leq f(a) < f(u)$ ,  $\forall u \in U - E$ ).

Ceci prouve que  $u^*$  est un point de minimum absolu de  $f$  sur  $U$ , ce qui finit la preuve.  $\square$

**Théorème 2.22.** (*Unicité*)

Soit  $U \subset \mathbb{R}^n$  un ensemble convexe et  $f : U \rightarrow \mathbb{R}$  une fonction strictement convexe. Alors il existe **au plus** un point de minimum de  $f$  sur  $U$ .

*Démonstration.* On va raisonner par absurd. Soient  $u_1, u_2 \in U$  avec  $u_1 \neq u_2$  deux points de minimum de  $f$  sur  $U$ . Nous avons donc :

$$f(u_1) = f(u_2) \leq f(u), \quad \forall u \in U. \quad (2.14)$$

Comme  $f$  est strictement convexe, on a

$$f\left(\frac{1}{2}u_1 + \frac{1}{2}u_2\right) < \frac{1}{2}f(u_1) + \frac{1}{2}f(u_2) = f(u_1)$$

(car  $f(u_1) = f(u_2)$ ) et ceci contredit (2.14).  $\square$

**Corollaire 2.23.** (*Existence et unicité*)

Soit  $\Omega \subset \mathbb{R}^n$  un ouvert,  $U \subset \Omega$  un ensemble fermé et convexe et  $f : \Omega \rightarrow \mathbb{R}$  une fonction de classe  $C^1$  sur  $\Omega$  et fortement convexe sur  $U$ . Alors **il existe un unique** point de minimum de  $f$  sur  $U$ .

*Démonstration.* **Existence** : Il y a deux possibilités :

1. Soit  $U$  est borné et alors l'existence est immédiate du Théorème 2.21.
2. Soit  $U$  est non borné et alors comme conséquence de la Proposition 2.11  $f$  est coercive sur  $U$  et donc l'existence résulte encore du Théorème 2.21.

**Unicité** : Comme  $f$  est strictement convexe sur  $U$ , l'unicité est une conséquence immédiate du Théorème 2.22.  $\square$

**Remarque** : On utilisera très souvent ce corollaire avec  $\Omega = \mathbb{R}^n$  et  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  une fonction de classe  $C^1$  et fortement convexe sur  $\mathbb{R}^n$  (c'est à dire  $f$  est une fonction elliptique). Alors pour tout ensemble  $U \subset \mathbb{R}^n$  convexe et fermé on a l'existence et l'unicité d'un point de minimum de  $f$  sur  $U$ .

# Chapitre 3

## Optimisation sans contraintes

On considère ici le cas particulier où l'ensemble de contraintes  $U$  est l'espace entier  $\mathbb{R}^n$ . On va donc considérer le problème de minimisation :

$$\min_{u \in \mathbb{R}^n} J(u) \quad (3.1)$$

où  $J : \mathbb{R}^n \mapsto \mathbb{R}$  est une fonction donnée.

C'est un problème de **minimisation sans contraintes**.

Le problème (3.1) peut s'écrire

$$\text{Trouver } u^* \in \mathbb{R}^n \text{ tel que } J(u^*) \leq J(u), \quad \forall u \in \mathbb{R}^n. \quad (3.2)$$

On supposera que  $u^*$  existe (éventuellement qu'il est unique) et on se propose de trouver une **approximation numérique** de  $u^*$ , en construisant une suite  $\{u^{(k)}\}_{k \in \mathbb{N}} \subset \mathbb{R}^n$  telle que

$$u^{(k)} \mapsto u^* \quad \text{pour } k \mapsto +\infty.$$

**Remarque :** Si  $J \in C^1$  alors  $u^*$  satisfait nécessairement l'équation d'Euler

$$\nabla J(u^*) = 0$$

(vu en Chapitre 2).

*Rappel :* Si en plus  $J$  est convexe alors l'équation d'Euler est aussi une condition **suffisante** de minimum. Si en plus  $J$  est strictement convexe, alors  $u^*$  est unique.

Alors une méthode numérique qui peut être envisagée c'est de résoudre numériquement l'équation d'Euler, qui est en fait un système de  $n$  équations avec  $n$  inconnues, du type

$$\text{Trouver } u \in \mathbb{R}^n \text{ tel que } F(u) = 0$$

où  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  est une fonction donnée (ici  $F = \nabla J$ ).

On peut utiliser la méthode numérique de **Newton-Raphson**

**En particulier** si  $J$  est donnée par

$$J(x) = \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle + c$$

avec  $A$  matrice symétrique,  $b$  vecteur et  $c$  scalaire (c'est à dire,  $J$  est quadratique) alors nous avons  $\nabla J(u) = Au - b$ , et l'équation d'Euler devient un **système algébrique linéaire**

$$Au = b.$$

On peut utiliser alors toute méthode numérique connue pour la résolution de ce système linéaire.

On se propose ici de donner des méthodes numériques qui sont spécifiques à l'optimisation. Ces méthodes consistent à construire la suite approximante  $\{u^{(k)}\}$  par **recurrence**, c'est à dire : on se donne un point initial  $u^{(0)} \in \mathbb{R}^n$  et on donnera  $u^{(k+1)}$  comme une fonction de  $u^{(k)}$ . L'expression générale de  $u^{(k+1)}$  sera

$$u^{(k+1)} = u^{(k)} + \rho_k d^{(k)} \quad (3.3)$$

avec  $d^{(k)} \in \mathbb{R}^n$  des vecteurs qu'on appelle **directions de descente** et  $\rho_k \in \mathbb{R}$  des scalaires qu'on appelle **facteurs de descente**. Ces noms viennent du fait qu'on cherchera toujours à avoir (entre autres)

$$J(u^{(k+1)}) < J(u^{(k)}), \quad \forall k \in \mathbb{N} \quad (3.4)$$

(la suite  $\{J(u^{(k)})\}$  doit être strictement décroissante).

L'algorithme associé est en général de la forme :

**pas 1.** Faire  $k = 0$ , choisir  $u^{(0)} \in \mathbb{R}^n$  (par exemple  $u^{(0)} = 0$ ), choisir  $k_{max} \in \mathbb{N}$  assez grand (par exemple  $k_{max} = 1000$ ).

**pas 2.** Tant que ((“test arrêt” est faux) et ( $k < k_{max}$ )) faire  $u^{(k+1)} = u^{(k)} + \rho_k d^{(k)}$  et  $k = k + 1$

**pas 3.** Si (“test arrêt” est vrai) alors  $u^{(k)}$  est une approximation du point de minimum recherché. Sinon, la méthode n'a pas convergé.

Comme “test arrêt” on choisit le plus souvent :  $\nabla J(u^{(k)}) = 0$  (en pratique le test sera :  $\|\nabla J(u^{(k)})\| \leq \epsilon$  où  $\epsilon$  est un nombre strictement positif et petit, qui représente un niveau de tolérance admis, à fixer au début (par exemple  $\epsilon = 10^{-6}$ )).

**Remarque :** On écrira souvent simplement (sans détailler) :

$$\begin{cases} u^{(k+1)} = u^{(k)} + \rho_k d^{(k)} & k \in \mathbb{N} \\ u^{(0)} & \text{donné.} \end{cases} \quad (3.5)$$

Il y a un grand nombre de méthodes numériques de minimisation, suivant le choix qui est fait pour  $d^{(k)}$  et  $\rho_k$ .

## 3.1 Méthodes de relaxation

### 3.1.1 Description de la méthode

Cette méthode consiste à faire le choix suivant pour les directions de descente :

$$d^{(0)} = e_1, \quad d^{(1)} = e_2, \quad \dots \quad d^{(n-1)} = e_n$$



ensuite on recommence

$$d^{(n)} = e_1, \quad d^{(n+1)} = e_2, \quad \dots \quad d^{(2n-1)} = e_n$$

et ainsi de suite ...

(rappelons que  $\{e_1, e_2, \dots, e_n\}$  sont les vecteurs de la base canonique de  $\mathbb{R}^n$ ).

Donc en général on a :

$$d^{(k)} = e_l$$

si et seulement si  $l$  est le rest de la division de  $k + 1$  par  $n$ .

On prend ensuite les facteurs  $\rho_k \in \mathbb{R}$  tels que

$$J(u^{(k)} + \rho_k d^{(k)}) = \min_{\rho \in \mathbb{R}} J(u^{(k)} + \rho d^{(k)}), \quad \forall k \in \mathbb{N} \quad (3.6)$$

(en supposant que ces minimum existent).

Finalement on pose

$$u^{(k+1)} = u^{(k)} + \rho_k d^{(k)}.$$

C'est la méthode (ou l'algorithme) de **relaxation**.

On peut écrire cet algorithme de la manière équivalente suivante :

On suppose connu le vecteur  $u^{(k)} = (u_1^{(k)}, u_2^{(k)}, \dots, u_n^{(k)})^T$  et on calcule

$u^{(k+1)} = (u_1^{(k+1)}, u_2^{(k+1)}, \dots, u_n^{(k+1)})^T$  en  $n$  pas successifs par les formules suivantes :

$$J(u_1^{(k+1)}, u_2^{(k)}, \dots, u_n^{(k)}) = \min_{y \in \mathbb{R}} J(y, u_2^{(k)}, \dots, u_n^{(k)})$$

$$J(u_1^{(k+1)}, u_2^{(k+1)}, u_3^{(k)}, \dots, u_n^{(k)}) = \min_{y \in \mathbb{R}} J(u_1^{(k+1)}, y, u_3^{(k)}, \dots, u_n^{(k)})$$

...

$$J(u_1^{(k+1)}, \dots, u_n^{(k+1)}) = \min_{y \in \mathbb{R}} J(u_1^{(k+1)}, \dots, u_{n-1}^{(k+1)}, y)$$

Ci-dessus on a utilisé le fait que (par exemple)

$$\min_{\rho \in \mathbb{R}} J(u_1^{(k)} + \rho, u_2^{(k)}, \dots, u_n^{(k)}) = \min_{y \in \mathbb{R}} J(y, u_2^{(k)}, \dots, u_n^{(k)}),$$

égalité obtenue en faisant le changement de variable  $y = u_1^{(k)} + \rho$ .

Le but dans la suite est de démontrer que l'algorithme de relaxation est bien défini, au moins dans le cas d'une fonction elliptique. Pour cela montrons d'abord le résultat suivant :

**Proposition 3.1.** *Soit  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  et  $a, b \in \mathbb{R}^n$ . Définissons la fonction  $g : \mathbb{R} \mapsto \mathbb{R}$ ,  $g(t) = f(a + tb)$ ,  $\forall t \in \mathbb{R}$ . Alors on a :*

1. *Si  $f$  est convexe alors  $g$  est convexe.*
2. *Si  $f$  est strictement convexe et  $b \neq 0$  alors  $g$  est strictement convexe.*
3. *Si  $f$  est fortement convexe et  $b \neq 0$  alors  $g$  est fortement convexe.*

*Démonstration.* Remarquons d'abord que les hypothèses de convexité et de convexité forte peuvent s'écrire :

$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y) - \alpha\theta(1 - \theta)\|x - y\|^2 \quad \forall x, y \in \mathbb{R}^n, \quad \forall \theta \in [0, 1] \quad (3.7)$$

avec  $\alpha = 0$  pour la convexité et  $\alpha > 0$  pour la convexité forte.

*Preuve de 1) et 3) :* Soient  $t_1, t_2 \in \mathbb{R}$ ,  $\theta \in [0, 1]$ . Alors

$$g(\theta t_1 + (1 - \theta)t_2) = f(a + [\theta t_1 + (1 - \theta)t_2]b) = f(\theta(a + t_1b) + (1 - \theta)(a + t_2b))$$

(on a utilisé l'égalité :  $a = \theta a + (1 - \theta)a$ ).

En utilisant (3.7) on obtient

$$g(\theta t_1 + (1 - \theta)t_2) \leq \theta f(a + t_1b) + (1 - \theta)f(a + t_2b) - \alpha\theta(1 - \theta)\|a + t_1b - a - t_2b\|^2$$

ce qui donne

$$g(\theta t_1 + (1 - \theta)t_2) \leq \theta g(t_1) + (1 - \theta)g(t_2) - \alpha\|b\|^2\theta(1 - \theta)|t_1 - t_2|^2$$

ce qui prouve 1) et 3).

*Preuve de 2).* La preuve est analogue à celle de 1); il faut prendre  $t_1 \neq t_2$  et  $\theta \in ]0, 1[$ . Comme  $b \neq 0$  on déduit que  $a + t_1b \neq a + t_2b$ . Alors on peut reprendre la démonstration précédente avec  $\alpha = 0$  en remplaçant " $\leq$ " par " $<$ ".  $\square$

**Corollaire 3.2.** Soit  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  et  $a_1, a_2, \dots, a_{i-1}, a_{i+1}, \dots, a_n \in \mathbb{R}, i \in \{1, 2, \dots, n\}$  données. Soit  $g : \mathbb{R} \rightarrow \mathbb{R}$  définie par

$$g(t) = f(a_1, a_2, \dots, a_{i-1}, t, a_{i+1}, \dots, a_n)$$

(fonction partielle par rapport à la  $i$ -ème variable).

Alors on a :

1. Si  $f$  est convexe alors  $g$  est convexe
2. Si  $f$  est strictement convexe alors  $g$  est strictement convexe
3. Si  $f$  est fortement convexe alors  $g$  est fortement convexe.

*Démonstration.* Il suffit d'appliquer la Proposition 3.1 avec  $a = (a_1, a_2, \dots, a_{i-1}, 0, a_{i+1}, \dots, a_n)$  et  $b = e_i \neq 0$ .  $\square$

Le résultat théorique principal de cette section est le suivant :

**Théorème 3.3.** Soit  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  une fonction elliptique. Alors la méthode de relaxation est bien définie et elle converge (c'est à dire, pour tout  $u^{(0)} \in \mathbb{R}^n$  la suite  $u^{(k)}$  construit par cet algorithme converge vers l'unique point de minimum de  $J$ ).

*Démonstration.* Du Corollaire 3.2 on déduit que la fonction

$$y \in \mathbb{R} \rightarrow J(u_1^{(k+1)}, \dots, u_{i-1}^{(k+1)}, y, u_{i+1}^k, \dots, u_n^k) \in \mathbb{R}$$

est une fonction elliptique (car  $f$  est elliptique). On déduit alors qu'il existe un unique point de minimum de cette fonction, ce qui montre que l'algorithme de relaxation est bien défini. La convergence de  $u^{(k)}$  est admise.  $\square$

**Remarque :** Cette méthode de relaxation est intéressante si ces minimisations sur  $\mathbb{R}$  peuvent se faire de manière exacte.

### 3.1.2 Cas particulier des fonctions quadratiques

Dans ce paragraphe nous supposons que  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  est donnée par

$$J(x) = \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle + c \quad (3.8)$$

avec  $A \in \mathcal{M}_n(\mathbb{R})$  matrice *SDP* (c'est à dire matrice symétrique et définie positive),  $b \in \mathbb{R}^n$ ,  $c \in \mathbb{R}$  (donc c'est une forme quadratique associée à une matrice *SDP*).

On a besoin de calculer  $y^* \in \mathbb{R}$  tel que

$$\min_{y \in \mathbb{R}} g(y) = g(y^*)$$

où la fonction  $g : \mathbb{R} \mapsto \mathbb{R}$  est définie par

$$g(y) = J(u_1^{(k+1)}, \dots, u_{i-1}^{(k+1)}, y, u_{i+1}^{(k)}, \dots, u_n^{(k)})$$

avec  $u_1^{(k+1)}, \dots, u_{i-1}^{(k+1)}, u_{i+1}^{(k)}, \dots, u_n^{(k)} \in \mathbb{R}$  données.

Pour faciliter l'écriture, faisons les notations suivantes :

$$a_1 = u_1^{(k+1)}, \quad a_2 = u_2^{(k+1)}, \quad \dots, \quad a_{i-1} = u_{i-1}^{(k+1)}, \quad a_{i+1} = u_{i+1}^{(k)}, \quad \dots, \quad a_n = u_n^{(k)}$$

Comme  $J$  est elliptique alors  $g$  est aussi elliptique, donc  $y^*$  existe et est unique;  $y^*$  est l'unique solution de l'équation d'Euler

$$g'(y^*) = 0.$$

Mais  $g'(y)$  n'est autre que  $\frac{\partial J}{\partial x_i}(a_1, a_2, \dots, a_{i-1}, y, a_{i+1}, \dots, a_n)$  ce qui nous donne

$$g'(y) = \sum_{j=1, j \neq i}^n A_{ij} a_j + A_{ii} y - b_i.$$

Donc on doit résoudre en  $y^* \in \mathbb{R}$  :

$$\sum_{j=1, j \neq i}^n A_{ij} a_j + A_{ii} y^* - b_i = 0$$

ce qui nous donne

$$y^* = \frac{1}{A_{ii}} \left( b_i - \sum_{j=1, j \neq i}^n A_{ij} a_j \right)$$

**Remarque :** On a  $A_{ii} > 0$  car  $A_{ii} = \langle Ae_i, e_i \rangle > 0$  ( $A$  est *SDP*).

On a donc montré :

**Proposition 3.4.** Dans le cas où  $J$  est donnée par (3.8) ( $J$  quadratique avec  $A$  une matrice SDP) la méthode de relaxation s'écrit :

$$\begin{aligned} u_1^{(k+1)} &= \frac{1}{A_{11}} \left( b_1 - \sum_{j>1}^n A_{1j} u_j^{(k)} \right) \\ u_2^{(k+1)} &= \frac{1}{A_{22}} \left( b_2 - A_{21} u_1^{(k+1)} - \sum_{j>2}^n A_{2j} u_j^{(k)} \right) \\ &\dots \\ u_i^{(k+1)} &= \frac{1}{A_{ii}} \left( b_i - \sum_{j<i}^n A_{ij} u_j^{(k+1)} - \sum_{j>i}^n A_{ij} u_j^{(k)} \right) \\ &\dots \\ u_n^{(k+1)} &= \frac{1}{A_{nn}} \left( b_n - \sum_{j<n}^n A_{nj} u_j^{(k+1)} \right) \end{aligned}$$

En plus la méthode converge (c'est une conséquence du Théorème 3.3).

**Remarque :** L'algorithme ci-dessus n'est autre que l'algorithme de Gauss-Seidel (vu en L3) pour la résolution numérique du système linéaire

$$Ax = b.$$

Ceci s'explique par le fait que le point de minimum  $u^*$  de  $J$  satisfait l'équation d'Euler  $\nabla J(u^*) = 0$  qui devient dans le cas  $J$  quadratique :  $Au^* = b$ .

## 3.2 Méthodes de gradient

**Idée :** prendre comme direction de descente :

$$d^{(k)} = -\nabla J(u^{(k)})$$

*Justification :*  $J(u^{(k)})$  est orthogonal à la ligne du niveau de  $J$  dans le point  $u^{(k)}$  et la fonction  $J$  diminue dans la direction  $-\nabla J(u^{(k)})$ . Cette affirmation se justifie par le calcul suivant, où on utilise un développement de Taylor :

$$J(u^{(k)} - \rho \nabla J(u^{(k)})) = J(u^{(k)}) + \langle \nabla J(u^{(k)}), -\rho \nabla J(u^{(k)}) \rangle + o(\rho)$$

(on suppose que  $J$  est de classe  $C^1$ ). Nous avons alors :

$$J(u^{(k)} - \rho \nabla J(u^{(k)})) - J(u^{(k)}) = -\rho \|\nabla J(u^{(k)})\|^2 + o(\rho).$$

Le membre de droite de l'égalité précédente est  $< 0$  si  $\rho > 0$  avec  $\rho$  "assez petit" et  $\nabla J(u^{(k)}) \neq 0$ .

Les **méthodes de gradient** sont définis alors par les relations :

$$u^{(k+1)} = u^{(k)} - \rho_k \nabla J(u^{(k)}) \tag{3.9}$$

où les facteurs de descente  $\rho_k \in \mathbb{R}$  sont à choisir.

Il y a plusieurs méthodes de gradient suivant le choix que nous faisons pour  $\rho_k$ .

### 3.2.1 Méthodes de gradient à pas optimal

La méthode est la suivante :  $u^{(k+1)}$  est donnée par (3.9) où  $\rho_k \in \mathbb{R}$  est tel que

$$J(u^{(k)} - \rho_k \nabla J(u^{(k)})) = \min_{\rho \in \mathbb{R}} J(u^{(k)} - \rho \nabla J(u^{(k)})) \quad (3.10)$$

(en supposant qu'un tel minimum existe).

**Remarque :** Nous faisons l'itération (3.9) dans le cas où  $\nabla J(u^{(k)}) \neq 0$ , donc le problème de minimisation (3.10) a un sens.

On a le résultat suivant :

**Théorème 3.5.** *Soit  $J : \mathbb{R}^n \mapsto \mathbb{R}$  une fonction elliptique. Alors la méthode de gradient à pas optimal donnée par (3.9) et (3.10) est bien définie et converge.*

*Démonstration.* Comme conséquence de la Proposition 3.1 la fonction

$$\rho \in \mathbb{R} \mapsto J(u^{(k)} - \rho \nabla J(u^{(k)})) \in \mathbb{R}$$

est une fonction elliptique, donc il existe un unique point de minimum de cette fonction. Ceci montre que la méthode de gradient à pas optimal est bien définie.

La convergence est admise! □

#### Cas particulier : fonctions quadratiques.

On suppose ici  $J$  comme dans le paragraph 3.1.2 (donc  $J$  quadratique associée à une matrice SDP).

Nous devons calculer  $\rho_k \in \mathbb{R}$  qui minimise la fonction  $g : \mathbb{R} \mapsto \mathbb{R}$  donnée par

$$g(\rho) = J(u^{(k)} - \rho \nabla J(u^{(k)})).$$

Alors  $\rho_k$  satisfait nécessairement

$$g'(\rho_k) = 0.$$

Un calcul simple nous donne

$$g'(\rho) = - \langle \nabla J(u^{(k)} - \rho \nabla J(u^{(k)})), \nabla J(u^{(k)}) \rangle$$

c'est à dire, comme  $\nabla J(u^{(k)}) = Au^{(k)} - b$

$$g'(\rho) = - \langle A(u^{(k)} - \rho \nabla J(u^{(k)})) - b, Au^{(k)} - b \rangle = -\|Au^{(k)} - b\|^2 + \rho \langle A(Au^{(k)} - b), Au^{(k)} - b \rangle$$

On obtient alors

$$\rho_k = \frac{\|Au^{(k)} - b\|^2}{\langle A(Au^{(k)} - b), Au^{(k)} - b \rangle} \quad (3.11)$$

Rémarquons qu'on a  $\langle A(Au^{(k)} - b), Au^{(k)} - b \rangle > 0$  (car  $A$  est SDP et  $Au^{(k)} - b = \nabla J(u^{(k)}) \neq 0$ ).

Donc la méthode de gradient à pas optimal dans le cas  $J$  quadratique est :

$$u^{(k+1)} = u^{(k)} - \rho_k (Au^{(k)} - b)$$

avec  $\rho_k$  données par (3.11) (valable uniquement pour  $Au^{(k)} - b \neq 0$ ).

**Remarque :** Comme  $g'(\rho_k) = 0$ , on déduit immédiatement

$$\langle \nabla J(u^{(k+1)}), \nabla J(u^{(k)}) \rangle = 0. \quad (3.12)$$

### 3.2.2 Autres méthodes du type gradient

Si  $J$  n'est pas quadratique, il peut être difficile de trouver  $\rho_k$  solution de (3.10). Quand on ne peut pas calculer  $\rho_k$  de manière exacte, il faut utiliser à chaque pas  $k$  une méthode numérique pour approcher  $\rho_k$ , ce qui alourdit les calculs. On peut se satisfaire uniquement d'une estimation "assez large" pour  $\rho_k$ . Le résultat suivant nous donne un intervalle tel que si tous les  $\rho_k$  se trouvent dans cet intervalle, alors la méthode du gradient converge :

**Théorème 3.6.** *Soit  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  une fonction elliptique (c'est à dire,  $f \in C^1$  et  $\exists \alpha > 0$  telle que*

$$\langle \nabla J(x) - \nabla J(y), x - y \rangle \geq \alpha \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n \quad )$$

*et telle que  $\nabla J$  soit lipschitzienne (c'est à dire :  $\exists M > 0$  tel que*

$$\|\nabla J(x) - \nabla J(y)\| \leq M \|x - y\|, \quad \forall x, y \in \mathbb{R}^n \quad ).$$

*Supposons que la suite  $\{\rho_k\}_{k \in \mathbb{N}} \subset \mathbb{R}$  satisfait la propriété suivante : il existe  $a_1, a_2$  avec  $0 < a_1 \leq a_2 < \frac{2\alpha}{M^2}$  tels que*

$$a_1 \leq \rho_k \leq a_2, \quad \forall k \in \mathbb{N}. \quad (3.13)$$

*Alors la méthode générale de gradient (3.9) converge et la convergence est au moins géométrique, c'est à dire :  $\exists \beta \in [0, 1[$  tel que*

$$\|u^{(k)} - u^*\| \leq \beta^k \|u^{(0)} - u^*\|, \quad \forall k \in \mathbb{N}$$

*où  $u^*$  est l'unique point de minimum de  $J$  sur  $\mathbb{R}^n$ .*

*Démonstration.* Nous savons que  $u^*$  est l'unique solution de l'équation

$$\nabla J(u^*) = 0.$$

Notre but est de montrer que  $u^{(k)} - u^* \rightarrow 0$  pour  $k \rightarrow +\infty$ . Nous avons :

$$u^{(k+1)} - u^* = u^{(k)} - \rho_k \nabla J(u^{(k)}) - u^* = u^{(k)} - u^* - \rho_k [\nabla J(u^{(k)}) - \nabla J(u^*)].$$

Utilisons la formule :  $\|x - y\|^2 = \|x\|^2 + \|y\|^2 - 2 \langle x, y \rangle$  pour tous  $x, y \in \mathbb{R}^n$ . Nous avons

$$\|u^{(k+1)} - u^*\|^2 = \|u^{(k)} - u^*\|^2 + \rho_k^2 \|\nabla J(u^{(k)}) - \nabla J(u^*)\|^2 - 2\rho_k \langle u^{(k)} - u^*, \nabla J(u^{(k)}) - \nabla J(u^*) \rangle.$$

En utilisant l'ellipticité de  $J$  et le fait que  $\nabla J$  est lipschitzienne, nous obtenons

$$\|u^{(k+1)} - u^*\|^2 \leq \|u^{(k)} - u^*\|^2 + M^2 \rho_k^2 \|u^{(k)} - u^*\|^2 - 2\rho_k \alpha \|u^{(k)} - u^*\|^2$$

c'est à dire

$$\|u^{(k+1)} - u^*\|^2 \leq (1 - 2\alpha\rho_k + M^2\rho_k^2) \|u^{(k)} - u^*\|^2. \quad (3.14)$$

Considérons maintenant la fonction  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$  donnée par

$$\varphi(\rho) = 1 - 2\alpha\rho + M^2\rho^2.$$

Comme  $a_1 \leq \rho_k \leq a_2$ , il est facile de voir que

$$\varphi(\rho_k) \leq \max\{\varphi(a_1), \varphi(a_2)\} \quad (3.15)$$

D'autre part, le minimum de  $\varphi$  est atteint en  $\frac{\alpha}{M^2}$ , ce qui nous donne

$$\varphi(\rho_k) \geq 1 - \frac{\alpha^2}{M^2}. \quad (3.16)$$

En utilisant l'inégalité de Cauchy-Schwarz nous avons pour tous  $x, y \in \mathbb{R}^n$  :

$$\alpha \|x - y\|^2 \leq \langle \nabla J(x) - \nabla J(y), x - y \rangle \leq \|\nabla J(x) - \nabla J(y)\| \cdot \|x - y\| \leq M \|x - y\|^2$$

ce qui nous donne, en prenant  $x \neq y$

$$\alpha \leq M. \quad (3.17)$$

Comme  $\varphi(0) = \varphi(\frac{2\alpha}{M^2}) = 1$ , on déduit grâce aux inégalités (3.16) et (3.17) que

$$\varphi(y) \in [0, 1[ \quad \forall y \in \left] 0, \frac{2\alpha}{M^2} \right[.$$

Posons alors  $\beta = \sqrt{\max\{\varphi(a_1), \varphi(a_2)\}}$ . Comme  $a_1, a_2 \in \left] 0, \frac{2\alpha}{M^2} \right[$  on a clairement :

$$0 \leq \beta < 1.$$

A l'aide de (3.15) on obtient

$$\varphi(\rho_k) \leq \beta^2$$

ce qui avec (3.14) nous donne

$$\|u^{(k+1)} - u^*\| \leq \beta \|u^{(k)} - u^*\|, \quad \forall k \in \mathbb{N}.$$

Par recurrence, nous déduisons  $\forall k \in \mathbb{N}^*$  :

$$\|u^{(k)} - u^*\| \leq \beta \|u^{(k-1)} - u^*\| \leq \beta^2 \|u^{(k-2)} - u^*\| \leq \dots \leq \beta^k \|u^{(0)} - u^*\|$$

ce qui nous donne le résultat attendu. □

**Définition :**

Une méthode de gradient du type (3.9) est dite **à pas constant (ou fixe)** si  $\rho_k$  est constant (indépendant de  $k$ ). Dans le cas contraire, elle est dite **à pas variable**.

**Remarque :**

On déduit du Théorème 3.6 que si  $\rho$  est tel que

$$0 < \rho < \frac{2\alpha}{M^2}$$

alors la méthode de gradient à pas fixe

$$u^{(k+1)} = u^{(k)} - \rho \nabla J(u^{(k)})$$

est convergente (choisir  $a_1 = a_2 = \rho$  dans le Théorème 3.6).

C'est la méthode de gradient la plus simple à utiliser.

**Inconvénient :** En général il est difficile de connaître  $\alpha$  et  $M$  (en supposant qu'ils existent). Alors dans la pratique on prends un  $\rho$  assez petit pour être sûr d'avoir la convergence. Mais dans ce cas la convergence peut être très lente !

**Un exemple où on peut calculer  $\alpha$  et  $M$  :** Si  $J$  est quadratique avec une matrice  $A$  qui est SDP. Alors on peut prendre  $\alpha = \lambda_{\min} > 0$  (la plus petite valeur propre de  $A$ ) et  $M = \|A\|_2 = \rho(A) = \lambda_{\max} > 0$  (la plus grande valeur propre de  $A$ ).

**Une alternative : un algorithme basé sur une recherche linéaire.**

Cette méthode pour trouver  $\rho_k$  est une méthode intermédiaire entre l'algorithme de gradient à pas optimal et un algorithme de gradient à pas constant. Le principe en est le suivant :

On note  $g : \mathbb{R} \rightarrow \mathbb{R}$  la fonction donnée par

$$g(\rho) = J(u^{(k)} - \rho \nabla J(u^{(k)})).$$

On cherche  $r > 0$  et  $s \in \mathbb{N}$ ,  $s \geq 2$ , tels que

$$g(0) > g(r) > g(2r) > \dots > g(sr)$$

et

$$g(sr) \leq g((s+1)r)$$

(pour tout  $r > 0$  assez petit il existe au moins un nombre naturel  $s$  avec cette propriété, si on fait une hypothèse de coercivité sur  $J$ ).

On pose alors

$$\rho_k = sr.$$

*Idée intuitive :* bien exploiter les possibilités de "descendre" dans la direction  $-\nabla J(u^{(k)})$ .

L'algorithme associé est le suivant :

faire  $r = 1$

tant que  $(g(2r) < g(r) < g(0))$  est faux faire  $r = r/2$

$\rho = 2r$

tant que  $(g(\rho+r) < g(\rho))$  faire  $\rho = \rho+r$

$\rho_k = \rho$ .



### 3.3 La méthode des gradients conjugués

**Rappels notations :**

1. Si  $v_1, v_2, \dots, v_l \in \mathbb{R}^n$  on note par  $\mathcal{L}(v_1, v_2, \dots, v_l) = \{\sum_{i=1}^l \alpha_i v_i, \alpha_1, \dots, \alpha_l \in \mathbb{R}\}$  l'espace vectoriel engendré par les vecteurs  $v_1, v_2, \dots, v_l$ ; c'est un sous-espace vectoriel de  $\mathbb{R}^n$ .
2. Si  $a \in \mathbb{R}^n$  et  $M \subset \mathbb{R}^n$  alors  $a + M$  désigne l'ensemble  $\{a + x, x \in M\}$ .

#### 3.3.1 Le cas quadratique

Dans ce paragraphe on considère  $J : \mathbb{R}^n \mapsto \mathbb{R}$  une fonction quadratique

$$J(u) = \frac{1}{2} \langle Au, u \rangle - \langle b, u \rangle + c, \quad \forall u \in \mathbb{R}^n$$

avec  $A \in \mathcal{M}_n(\mathbb{R})$ ,  $b \in \mathbb{R}^n$ ,  $c \in \mathbb{R}$ . On suppose que la matrice  $A$  est SDP.

**Idée de la méthode :** Rappelons que la méthode de gradient à pas optimal consiste à faire :

$$u^{(k+1)} = u^{(k)} - \rho_k \nabla J(u^{(k)}) = \min_{\rho \in \mathbb{R}} J(u^{(k)} - \rho \nabla J(u^{(k)}))$$

Ceci est équivalent avec :

$$u^{(k+1)} \in u^{(k)} + \mathcal{L}(\nabla J(u^{(k)})) \text{ est l'élément qui minimise } J \text{ sur } u^{(k)} + \mathcal{L}(\nabla J(u^{(k)}))$$

Dans la suite on va procéder de la manière suivante : On va noter pour tout  $k \in \mathbb{N}$  :

$$G_k = \mathcal{L}(\nabla J(u^{(0)}), \nabla J(u^{(1)}), \dots, \nabla J(u^{(k)})) \subset \mathbb{R}^n$$

La méthode des **gradients conjugués** consiste à chercher  $u^{(k+1)} \in u^{(k)} + G_k$  tel que

$$J(u^{(k+1)}) = \min_{v \in u^{(k)} + G_k} J(v) \tag{3.18}$$

(en supposant qu'un tel minimum existe).

On minimise donc sur un espace plus "grand" que dans la méthode de gradient à pas optimal. On s'attend alors à un "meilleur" minimum. Il reste à montrer que cette méthode est facile à implémenter et qu'elle donne le résultat attendu.

Comme  $J$  est elliptique et que  $u^{(k)} + G_k$  est un ensemble fermé et convexe (car c'est un espace affine), alors il existe une solution unique du problème de minimisation (3.18).

En plus, le Théorème 2.15 nous donne

$$\langle \nabla J(u^{(k+1)}), v - u^{(k+1)} \rangle \geq 0, \quad \forall v \in u^{(k)} + G_k. \tag{3.19}$$

Soit maintenant  $w \in G_k$  arbitraire. Remarquons qu'on a :

$$u^{(k+1)} + w = u^{(k)} + [u^{(k+1)} - u^{(k)}] + w \in u^{(k)} + G_k \text{ car } u^{(k+1)} - u^{(k)} \in G_k.$$

On peut alors prendre  $v = u^{(k+1)} + w$  en (3.19) et aussi  $v = u^{(k+1)} - w$ . On obtient

$$\langle \nabla J(u^{(k+1)}), w \rangle = 0, \quad \forall w \in G_k$$

(c'est à dire :  $\nabla J(u^{(k+1)})$  est **orthogonal** sur tous les vecteurs de  $G_k$ ). Ceci nous donne

$$\langle \nabla J(u^{(k+1)}), \nabla J(u^{(l)}) \rangle = 0, \quad \forall l = 0, 1, \dots, k. \quad (3.20)$$

(rappelons que dans l'algorithme de gradient à pas optimal on avait seulement  $\langle \nabla J(u^{(k+1)}), \nabla J(u^{(k)}) \rangle = 0$ .)

**Conséquences :**

1. Si les vecteurs  $\nabla J(u^{(0)}), \nabla J(u^{(1)}), \dots, \nabla J(u^{(k)})$  sont tous  $\neq 0$  alors ils sont indépendants (c'est une conséquence immédiate du résultat bien connu : si des vecteurs non-nuls sont orthogonaux par rapport à un produit scalaire, alors ils sont indépendants (facile à montrer)).
2. L'algorithme s'arrête en au plus  $n$  itérations, car il existe  $k \in \{0, 1, \dots, n\}$  tel que

$$\nabla J(u^{(k)}) = 0$$

(sinon, on aurait  $n+1$  vecteurs non-nuls indépendants en  $\mathbb{R}^{n+1}$ , ce qui est impossible). Alors  $u^{(k)}$  est la solution recherchée.

La question qui se pose maintenant est : **comment calculer  $u^{(k+1)}$  à partir de  $u^{(k)}$  ?** Supposons qu'on a

$$\nabla J(u^{(i)}) \neq 0, \quad \forall i = 0, 1, \dots, k$$

(sinon, l'algorithme s'arrêterait avant d'avoir à calculer  $u^{(k+1)}$ ).

Nous avons l'expression suivante :

$$u^{(l+1)} = u^{(l)} + \Delta_l \quad \forall l = 0, 1, \dots, k. \quad (3.21)$$

avec  $\Delta_l \in G_l$  que nous écrivons sous la forme

$$\Delta_l = \sum_{i=0}^l \alpha_i^l \nabla J(u^{(i)}) \quad (3.22)$$

avec  $\alpha_i^l \in \mathbb{R}$  des coefficients à trouver.

**Proposition 3.7.** *On a*

1.

$$\Delta_l \neq 0, \quad \forall l = 0, 1, \dots, k.$$

2.

$$\langle A\Delta_l, \Delta_m \rangle = 0 \quad \text{si } 0 \leq m < l \leq k.$$

*Démonstration.* Rappelons que  $\nabla J(u^{(i)}) = Au^{(i)} - b$ .

Nous avons :

$$\nabla J(u^{(l+1)}) = Au^{(l+1)} - b = A(u^{(l)} + \Delta_l) - b$$

et ceci nous donne

$$\nabla J(u^{(l+1)}) = \nabla J(u^{(l)}) + A\Delta_l. \quad (3.23)$$

En faisant le produit scalaire de l'égalité précédente par  $\nabla J(u^{(l)})$  on obtient

$$\|\nabla J(u^{(l)})\|^2 + \langle A\Delta_l, \nabla J(u^{(l)}) \rangle = 0.$$

Comme  $\|\nabla J(u^{(l)})\|^2 \neq 0$  on déduit  $\langle A\Delta_l, \nabla J(u^{(l)}) \rangle \neq 0$  donc  $\Delta_l \neq 0$ , ce qui finit la partie 1.

Pour montrer la partie 2. nous faisons le produit scalaire de (3.23) avec  $\nabla J(u^{(i)})$ ,  $i < l$ . On obtient

$$\langle A\Delta_l, \nabla J(u^{(i)}) \rangle = 0, \quad \forall i = 0, 1, \dots, l-1, \quad \text{donc } \forall i = 0, 1, \dots, m.$$

En multipliant par  $\alpha_i^m$  et en sommant pour  $i$  de 0 à  $m$  on obtient l'égalité de la partie 2.  $\square$

Ceci nous amène à donner la définition suivante :

**Définition 3.8.** On dit que deux vecteurs  $x, y \in \mathbb{R}^n$  sont **conjugués** par rapport à une matrice  $B \in \mathcal{M}_n(\mathbb{R})$  si

$$\langle Bx, y \rangle = 0.$$

**Remarque :** Si  $B$  est une matrice SDP alors l'application

$$(x, y) \in \mathbb{R}^n \times \mathbb{R}^n \rightarrow \langle Bx, y \rangle \in \mathbb{R}$$

est un produit scalaire en  $\mathbb{R}^n$  (résultat admis!) qu'on appelle *produit scalaire associé à la matrice  $B$* . (à noter que le produit scalaire habituel est un fait un produit scalaire associé à la matrice identité  $I_n$ ). Alors la définition précédente nous dit que, dans le cas où  $B$  est SDP, deux vecteurs sont conjugus par rapport à la matrice  $B$  s'ils sont orthogonaux par rapport au produit scalaire associé à  $B$ .

Comme les vecteurs  $\Delta_0, \Delta_1, \dots, \Delta_k$  sont orthogonaux par rapport au produit scalaire associé à la matrice  $A$  (Proposition 3.7) on déduit immédiatement :

**Proposition 3.9.** *Les vecteurs  $\Delta_0, \Delta_1, \dots, \Delta_k$  sont indépendants.*

Il est facile de voir qu'on a

$$\mathcal{L}(\Delta_0, \Delta_1, \dots, \Delta_l) = \mathcal{L}(\nabla J(u^{(0)}), \nabla J(u^{(1)}), \dots, \nabla J(u^{(l)})), \quad \forall l = 0, 1, \dots, k$$

(car l'inclusion " $\subset$ " est évidente de (3.22) et en plus les deux espaces ont la même dimension :  $l + 1$ ).

On déduit alors :

$$\alpha_l^l \neq 0, \quad \forall l = 0, 1, \dots, k \quad (3.24)$$

(car sinon, on aurait  $\Delta_l \in \mathcal{L}(\nabla J(u^{(0)}), \nabla J(u^{(1)}), \dots, \nabla J(u^{(l-1)})) = \mathcal{L}(\Delta_0, \Delta_1, \dots, \Delta_{l-1})$ , ce qui contredirait l'indépendance des  $\Delta_0, \dots, \Delta_l$ ).

On peut alors écrire

$$\Delta_k = \rho_k d^{(k)}$$

avec

$$d^{(k)} = \nabla J(u^{(k)}) + \sum_{i=0}^{k-1} \lambda_i^k \nabla J(u^{(i)}) \quad (3.25)$$

où on a posé

$$\begin{cases} \rho_k = \alpha_k^k \\ \lambda_i^k = \alpha_i^k / \alpha_k^k \end{cases} \quad (3.26)$$

(remarquons que dans (3.25) la somme n'existe pas si  $k = 0$ ).

Nous avons donc :

$$u^{(k+1)} = u^{(k)} + \rho_k d^{(k)}$$

avec  $d^{(k)}$  donné par (3.25) et (3.26).

Il nous reste à calculer les  $\rho_k$  et  $d^{(k)}$ .

Comme  $\langle A\Delta_k, \Delta_l \rangle = 0$  pour  $l < k$ , on déduit que  $\langle Ad^{(k)}, \Delta_l \rangle = 0$  ce qui nous donne, en utilisant aussi (3.23) :

$$0 = \langle d^{(k)}, A\Delta_l \rangle = \langle d^{(k)}, \nabla J(u^{(l+1)}) - \nabla J(u^{(l)}) \rangle \quad \text{pour } 0 \leq l \leq k-1.$$

A l'aide de (3.25) cela nous donne

$$\langle \nabla J(u^{(k)}) + \sum_{j=0}^{k-1} \lambda_j^k \nabla J(u^{(j)}), \nabla J(u^{(l+1)}) - \nabla J(u^{(l)}) \rangle = 0 \quad \text{pour } 0 \leq l \leq k-1. \quad (3.27)$$

En prenant  $l = k-1$  dans cette égalité on trouve

$$\|\nabla J(u^{(k)})\|^2 - \lambda_{k-1}^k \|\nabla J(u^{(k-1)})\|^2 = 0$$

ce qui nous donne

$$\lambda_{k-1}^k = \frac{\|\nabla J(u^{(k)})\|^2}{\|\nabla J(u^{(k-1)})\|^2}. \quad (3.28)$$

En prenant  $l \leq k-2$  en (3.27), on trouve la relation de récurrence

$$\lambda_{l+1}^k \|\nabla J(u^{(l+1)})\|^2 - \lambda_l^k \|\nabla J(u^{(l)})\|^2 = 0$$

ce qui donne

$$\lambda_l^k = \lambda_{l+1}^k \frac{\|\nabla J(u^{(l+1)})\|^2}{\|\nabla J(u^{(l)})\|^2}.$$

On en déduit facilement, en utilisant aussi (3.28)

$$\lambda_i^k = \frac{\|\nabla J(u^{(k)})\|^2}{\|\nabla J(u^{(i)})\|^2}, \quad i = 0, 1, \dots, k-1.$$

On obtient alors, à l'aide de (3.25) :

$$\begin{aligned} d^{(k)} &= \nabla J(u^{(k)}) + \sum_{i=0}^{k-1} \frac{\|\nabla J(u^{(k)})\|^2}{\|\nabla J(u^{(i)})\|^2} \nabla J(u^{(i)}) = \\ &\nabla J(u^{(k)}) + \frac{\|\nabla J(u^{(k)})\|^2}{\|\nabla J(u^{(k-1)})\|^2} \left[ \nabla J(u^{(k-1)}) + \sum_{i=0}^{k-2} \frac{\|\nabla J(u^{(k-1)})\|^2}{\|\nabla J(u^{(i)})\|^2} \nabla J(u^{(i)}) \right] \end{aligned}$$

Ceci nous donne la suite  $\{d^{(k)}\}$  par recurrence sous la forme

$$\begin{cases} d^{(0)} = \nabla J(u^{(0)}) \\ d^{(k)} = \nabla J(u^{(k)}) + \frac{\|\nabla J(u^{(k)})\|^2}{\|\nabla J(u^{(k-1)})\|^2} d_{k-1}. \end{cases} \quad (3.29)$$

Il reste à déterminer les  $\rho_k$ .

Rappelons qu'on a la relation de recurrence :

$$u^{(k+1)} = u^{(k)} + \rho_k d^{(k)}$$

et que nous avons

$$J(u^{(k)} + \rho_k d^{(k)}) \leq J(u^{(k)} + y), \quad \forall y \in G_k$$

ce qui nous donne

$$J(u^{(k)} + \rho_k d^{(k)}) \leq J(u^{(k)} + \rho d^{(k)}), \quad \forall \rho \in \mathbb{R}$$

car  $\rho d^{(k)} \in G_k$ . On en déduit que  $\rho_k$  est un point de minimum sur  $\mathbb{R}$  de l'application

$$\rho \in \mathbb{R} \rightarrow J(u^{(k)} + \rho d^{(k)}) \in \mathbb{R}.$$

donc  $\rho_k$  est un point où la dérivée de cette application s'annule, ce qui donne

$$\langle \nabla J(u^{(k)} + \rho_k d^{(k)}), d^{(k)} \rangle = 0$$

c'est à dire

$$\langle Au^{(k)} + \rho_k Ad^{(k)} - b, d^{(k)} \rangle = 0.$$

On obtient alors

$$\rho_k = -\frac{\langle \nabla J(u^{(k)}), d^{(k)} \rangle}{\langle Ad^{(k)}, d^{(k)} \rangle} \quad (3.30)$$

**Remarque :** La Proposition 3.7 nous dit que  $\Delta_k \neq 0$ , ce qui nous donne  $d^{(k)} \neq 0$ . Ceci implique  $\langle Ad^{(k)}, d^{(k)} \rangle > 0$ , car la matrice  $A$  est SDP.

L'algorithme des gradients conjugués (AGC) pour une fonction quadratique avec une matrice  $A$  qui est SDP est alors le suivant :

1. **pas 1.** On pose  $k = 0$ , on choisit  $u^{(0)} \in \mathbb{R}^n$  et on pose  $d^{(0)} = \nabla J(u^{(0)}) = Au^{(0)} - b$ .
2. **pas 2.** Si  $\nabla J(u^{(k)}) = 0$  *STOP* "La solution  $u^*$  est  $u^{(k)}$ ". Sinon, va au **pas 3**.
3. **pas 3.** On pose

$$\rho_k = -\frac{\langle \nabla J(u^{(k)}), d^{(k)} \rangle}{\langle Ad^{(k)}, d^{(k)} \rangle}$$

$$u^{(k+1)} = u^{(k)} + \rho_k d^{(k)}$$

$$\beta_k = \frac{\|\nabla J(u^{(k+1)})\|^2}{\|\nabla J(u^{(k)})\|^2}$$

$$d^{(k+1)} = \nabla J(u^{(k+1)}) + \beta_k d^{(k)}$$

faire  $k = k + 1$   
retour au **pas 2**.

### 3.3.2 Cas d'une fonction $J$ quelconque

On suppose ici  $J : \mathbb{R}^n \mapsto \mathbb{R}$  une fonction elliptique. Dans ce cas l'algorithme des gradients conjugués est le suivant (c'est une "petite" modification de l'algorithme précédent) :

1. **pas 1.** On pose  $k = 0$ , on choisit  $u^{(0)} \in \mathbb{R}^n$  et on pose  $d^{(0)} = \nabla J(u^{(0)})$ .
2. **pas 2.** Si  $\nabla J(u^{(k)}) = 0$  *STOP* "La solution  $u^*$  est  $u^{(k)}$ ". Sinon, va au **pas 3**.
3. **pas 3.** On pose

$$u^{(k+1)} = u^{(k)} + \rho_k d^{(k)}$$

où  $\rho_k \in \mathbb{R}$  est l'unique élément qui minimise la fonction

$$\rho \in \mathbb{R} \rightarrow J(u^{(k)} + \rho d^{(k)}) \in \mathbb{R}$$

et ensuite

$$\beta_k = \frac{\|\nabla J(u^{(k+1)})\|^2}{\|\nabla J(u^{(k)})\|^2}$$

$$d^{(k+1)} = \nabla J(u^{(k+1)}) + \beta_k d^{(k)}$$

faire  $k = k + 1$   
retour au **pas 2**.

Ceci est la variante dite de **Fletcher-Reeves**.

Il y a aussi la variante dite de **Polack-Ribiere** avec comme seul changement

$$\beta_k = \frac{\langle \nabla J(u^{(k+1)}) - \nabla J(u^{(k)}), \nabla J(u^{(k+1)}) \rangle}{\|\nabla J(u^{(k)})\|^2}$$

Cette dernière variante donne des meilleurs résultats en pratique.

**Remarque :** Ces deux versions coïncident dans le cas d'une fonction  $J$  quadratique, car dans ce cas on a

$$\langle \nabla J(u^{(k)}), \nabla J(u^{(k+1)}) \rangle = 0.$$

# Chapitre 4

## Optimisation avec contraintes

On rappelle qu'on se donne  $U \subset \mathbb{R}^n$  où  $U$  est un ensemble fermé des contraintes, avec  $U \neq \emptyset$  et  $U \neq \mathbb{R}^n$ . On se donne aussi la fonction

$$J : \mathbb{R}^n \mapsto \mathbb{R}.$$

On cherche à résoudre le problème de minimisation

$$\min_{u \in U} J(u) \tag{4.1}$$

(c'est à dire, on cherche  $u^* \in U$  tel que

$$J(u^*) \leq J(u), \quad \forall u \in U \quad ) \tag{4.2}$$

On va considérer deux cas pour  $U$  :

1.  $U$  est un **pavé**, c'est à dire, il est de la forme  $U = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_n, b_n]$  (avec la convention qu'on peut avoir  $]-\infty[$  à la place de  $]a_i[$  ou  $] +\infty[$  à la place de  $]b_i[$ , donc le pavé peut être de volume infini).
2.  $U$  est de la forme

$$U = \{x \in \mathbb{R}^n, \varphi_i(x) \leq 0, \quad i = 1, 2, \dots, m\} \tag{4.3}$$

avec  $m \in \mathbb{N}^*$  et  $\varphi_1, \varphi_2, \dots, \varphi_m$  des fonctions de  $\mathbb{R}^n$  dans  $\mathbb{R}$  données (on dit dans ce cas que  $U$  est donné par des **contraintes inégalités larges**).

### Remarques :

1. Le premier cas est un cas particulier du deuxième
2. Dans le cas des contraintes inégalités larges, le problème de minimisation associé est aussi appelé **problème de programmation nonlinéaire** si au moins une des fonctions  $\varphi_1, \varphi_2, \dots, \varphi_m$  ou  $J$  est non affine. Si toutes ces fonctions sont affines, alors on a un **problème de programmation linéaire** (vu en L3).

3. Dans la pratique on peut rencontrer des problèmes de minimisation avec contraintes égalités, c'est à dire, des contraintes de la forme

$$\tilde{U} = \{x \in \mathbb{R}^n, \varphi_i(x) = 0, \quad i = 1, 2, \dots, m\}$$

ou des problèmes avec des contraintes mélangées (égalités et inégalités larges). Dans ce cas on peut toujours se ramener à des problèmes avec contraintes inégalités larges, ceci par deux méthodes :

- soit en éliminant des contraintes à l'aides des égalités
- soit en écrivant une égalité  $\varphi_i(x) = 0$  comme une double inégalité :  $\varphi_i(x) \leq 0$  et  $-\varphi_i(x) \leq 0$ .

On peut montrer facilement le résultat suivant :

**Proposition 4.1.** *Soit  $U$  donné par (4.3). Si toutes les fonctions  $\varphi_1, \varphi_2, \dots, \varphi_m$  sont convexes alors  $U$  est un ensemble convexe.*

*Démonstration.* Laissée en exercice. □

## 4.1 Rappel sur les multiplicateurs de Lagrange

Soient  $J, \theta_1, \theta_2, \dots, \theta_m : \mathbb{R}^n \mapsto \mathbb{R}$  des fonctions de classe  $C^1$ , avec  $m \in \mathbb{N}^*$ . On note

$$\tilde{O} = \{x \in \mathbb{R}^n, \theta_i(x) = 0, \quad i = 1, 2, \dots, m\}.$$

On dit que  $\tilde{O}$  est une **variété** de  $\mathbb{R}^n$ .

**Définition 4.2.** 1. Si  $x \in \tilde{O}$  est tel que la famille des vecteurs  $\{\nabla\theta_i(x)\}_{i=1, \dots, m}$  forme un **système libre** en  $\mathbb{R}^n$  alors on dit que  $x$  est un **point régulier** de  $\tilde{O}$ .

2. La variété  $\tilde{O}$  est dite **régulière** si tous les points de  $\tilde{O}$  sont réguliers.

**Remarque :** Une condition équivalente de régularité de  $\tilde{O}$  en  $x$  est :  $\text{rang}(B) = m$  où  $B$  est la matrice Jacobienne donnée par

$$B_{ij} = \frac{\partial\theta_i}{\partial x_j}(x), \quad i = 1, \dots, m, \quad j = 1, \dots, n.$$

(on a alors nécessairement :  $m \leq n$ ).

On a le résultat suivant :

**Théorème 4.3.** *(Admis sans preuve !)*

*Soit  $x^*$  un point régulier de  $\tilde{O}$  tel que  $x^*$  soit un extremum local de  $J$  sur  $\tilde{O}$  (minimum local ou maximum local). Alors il existe  $\lambda_1^*, \lambda_2^*, \dots, \lambda_m^* \in \mathbb{R}$  (appelés **multiplicateurs de Lagrange**) tels que*

$$\nabla J(x^*) + \sum_{i=1}^m \lambda_i^* \nabla\theta_i(x^*) = 0 \tag{4.4}$$



**Remarque :** Le système (4.4) donne des conditions **nécessaires** d'optimalité (appelés aussi **conditions nécessaires d'optimalité de premier ordre**, car elles font intervenir des dérivées une fois uniquement). Ces conditions ne sont pas en général suffisantes. Le système (4.4) nous donne  $n$  équations avec  $n + m$  inconnues. Mais le fait que  $x^* \in \tilde{O}$  nous donne encore  $m$  équations :  $\theta_i(x^*) = 0$ ,  $i = 1, \dots, m$ , ce qui nous fait au total  $n + m$  équations avec  $n + m$  inconnues.

## 4.2 Optimisation sous contraintes d'inégalités

On se donne de nouveau les fonctions  $J, \theta_1, \theta_2, \dots, \theta_m : \mathbb{R}^n \mapsto \mathbb{R}$  de classe  $C^1$ , avec  $m \in \mathbb{N}^*$ .

Notons maintenant :

$$O = \{x \in \mathbb{R}^n, \theta_i(x) \leq 0, \forall i = 1, 2, \dots, m\}.$$

**Notations :**

On introduit la fonction à valeurs vectorielles  $\theta : \mathbb{R}^n \mapsto \mathbb{R}^m$  donnée par

$$\theta(x) = (\theta_1(x), \theta_2(x), \dots, \theta_m(x))^T.$$

Pour un vecteur  $b = (b_1, \dots, b_m)^T \in \mathbb{R}^m$  la notation  $b \leq 0$  signifie  $b_i \leq 0$ ,  $\forall i = 1, \dots, m$ . On peut donc écrire

$$O = \{x \in \mathbb{R}^n, \theta(x) \leq 0\}.$$

On va s'intéresser au problème de minimisation :

$$\min_{x \in O} J(x). \tag{4.5}$$

**Hypothèse :** On va supposer que les fonctions  $\theta_i$  sont de telle manière que  $O \neq \emptyset$  et aussi que  $O$  ne se réduit pas à un seul point ou à un nombre finit de points (dans quel cas le problème de minimisation (4.5) est très banal!)

### 4.2.1 Conditions d'optimalité de premier ordre : multiplicateurs de Karush-Kuhn-Tucker

On donnera des conditions d'optimalités similaires au système (4.4), mais pour des contraintes inégalités.

**Définition 4.4.** 1. On dit que la contrainte  $\theta_i(u) \leq 0$  est **active** en  $v \in O$  si on a

$$\theta_i(v) = 0.$$

On introduit alors l'ensemble

$$I(v) = \{i \in \{1, 2, \dots, m\}, \theta_i(v) = 0\}$$

et la variété (définie uniquement si  $I(v) \neq \emptyset$ )

$$\tilde{O}(v) = \{u \in \mathbb{R}^n, \quad \theta_i(u) = 0, \quad \forall i \in I(v)\}$$

(remarquons que  $v \in \tilde{O}(v)$ ).

2. On dira que  $v \in O$  est un **point régulier** de  $O$  si :

- soit  $I(v) = \emptyset$
- soit  $v$  est un point régulier de la variété  $\tilde{O}(v)$ .

Le résultat principal de ce paragraphe est le suivant :

**Théorème 4.5.** *Soit  $x^*$  un point régulier de  $O$ . Si  $x^*$  est un point de minimum local de  $J$  sur  $O$  alors il existe  $p_1^*, p_2^*, \dots, p_m^* \in [0, +\infty[$  (appelés **multiplicateurs de Karush-Kuhn-Tucker**) tels que :*

$$\nabla J(x^*) + \sum_{i=1}^m p_i^* \nabla \theta_i(x^*) = 0 \quad (4.6)$$

$$p_i^* \theta_i(x^*) = 0, \quad \forall i = 1, 2, \dots, m. \quad (4.7)$$

**Remarque :** Ce sont encore des conditions nécessaires d'optimalité de premier ordre appelés **conditions de Karush-Kuhn-Tucker**.

La preuve du Théorème 4.5 fait appel aux résultats suivants :

**Lemme 4.6.** *Soit  $v \in O$  tel que  $I(v) \neq \emptyset$  et soit  $w \in \mathbb{R}^n$  tel que*

$$\langle \nabla \theta_j(v), w \rangle < 0, \quad \forall j \in I(v).$$

*Alors  $w$  est une direction admissible pour  $v$  en  $O$ , c'est à dire, il existe  $t_0 > 0$  tel que*

$$v + tw \in O, \quad \forall t \in [0, t_0]$$

*(on a même plus :  $v + tw \in \overset{\circ}{O} \quad \forall t \in ]0, t_0[.$ )*

*Démonstration.* Il faut montrer :  $\theta_j(v + tw) \leq 0, \quad \forall j = 1, 2, \dots, m$  si  $t$  est "proche" de 0.

**Cas 1.** Si  $j \notin I(v)$ .

Alors  $\theta_j(v) < 0$  ce qui donne  $\theta_j(v + tw) < 0$  si  $t$  est "proche" de 0 (on utilise la continuité de  $\theta_j$ ).

**Cas 2.** Si  $j \in I(v)$ .

Alors  $\theta_j(v) = 0$  et en utilisant le développement de Taylor à l'ordre 1 on obtient

$$\theta_j(v + tw) = \theta_j(v) + t \langle \nabla \theta_j(v), w \rangle + o(t). \quad (4.8)$$

Comme  $\theta_j(v) = 0$  et en utilisant l'hypothèse du lemme, on obtient :  $\theta_j(v + tw) < 0$  si  $t > 0$  est "proche" de 0. Ceci nous donne le résultat.  $\square$

**Remarque :** On peut affaiblir l'hypothèse du lemme ci-dessus en prenant :  $\langle \nabla \theta_j(v), w \rangle \leq 0$  si  $j \in I(v)$  et si  $\theta_j$  est **affine** (car dans ce cas dans (4.8) on a  $o(t) = 0$ ).

Le corollaire suivant est une conséquence immédiate des lemmes 4.6 et 2.14 :

**Corollaire 4.7.** Soit  $v \in O$  tel que  $I(v) \neq \emptyset$  et soit  $w \in \mathbb{R}^n$  tel que

$$\langle \nabla \theta_j(v), w \rangle < 0, \quad \forall j \in I(v).$$

Supposons que  $v$  est un minimum local de  $J$  sur  $O$ . Alors

$$\langle \nabla J(v), w \rangle \geq 0.$$

**Lemme 4.8.** Soient  $d_1, d_2, \dots, d_l \in \mathbb{R}^n$  des vecteurs linéairement indépendents (donc forcément  $l \leq n$ ). Alors la matrice  $B \in \mathcal{M}(\mathbb{R})$  telle que  $B_{ij} = \langle d_i, d_j \rangle$ ,  $\forall i, j = 1, \dots, l$  est inversible.

*Démonstration.* Soit  $y = (y_1, y_2, \dots, y_l)^T \in \mathbb{R}^l$  telle que  $By = 0$  c'est à dire

$$\sum_{j=1}^l \langle d_i, d_j \rangle y_j = 0, \quad \forall i = 1, 2, \dots, l.$$

En multipliant chacune de ces égalités par  $y_i$  et en sommant sur  $i$  on obtient

$$\sum_{i=1}^l \sum_{j=1}^l \langle y_i d_i, y_j d_j \rangle = 0 \quad \text{c'est à dire} \quad \left\| \sum_{i=1}^l y_i d_i \right\|^2 = 0.$$

Ceci nous donne  $\sum_{i=1}^l y_i d_i = 0$ . Avec l'indépendance des vecteurs  $d_1, d_2, \dots, d_l$  on déduit  $y_1 = y_2 = \dots = y_l = 0$ , donc  $y = 0$ .

On a donc montré que  $By = 0$  entraîne  $y = 0$ , donc  $B$  est inversible.  $\square$

### Preuve du Théorème 4.5.

Soit  $x^* \in O$  un point de minimum local de  $J$  sur  $O$ . Il y a deux situations :

**Cas a)** Si  $I(x^*) = \emptyset$  (aucune contrainte n'est active en  $x^*$ ). Alors  $x^*$  appartient à l'intérieur de  $O$  ce qui implique

$$\nabla J(x^*) = 0$$

et le Théorème 4.5 est vraie avec  $p_1^* = p_2^* = \dots = p_m^* = 0$ .

**Cas b)** Si  $I(x^*) \neq \emptyset$ .

Pour simplifier supposons que  $I(x^*) = \{1, 2, \dots, l\}$  avec  $1 \leq l \leq m$ . Cela veut dire :

$$\begin{aligned} \theta_i(x^*) &= 0 & \text{si} & \quad 1 \leq i \leq l \\ \theta_i(x^*) &< 0 & \text{si} & \quad l+1 \leq i \leq m \end{aligned}$$

(avec la convention évidente que la dernière ligne est absente si  $l = m$ ).

Introduisons les ensembles

$$\tilde{O}(x^*) = \{x \in \mathbb{R}^n, \theta_i(x) = 0 \text{ si } 1 \leq i \leq l\}$$

et

$$\tilde{V}(x^*) = \{x \in \mathbb{R}^n, \theta_i(x) < 0 \text{ si } l+1 \leq i \leq m\}$$

(prendre  $\tilde{V}(x^*) = \mathbb{R}^n$  si  $l = m$ ).

Il est clair que  $\tilde{V}(x^*)$  est un ouvert et que  $x^* \in \tilde{V}(x^*)$ , ce qui nous dit que  $\tilde{V}(x^*)$  est un voisinage de  $x^*$ .

D'autre part nous avons  $\tilde{O}(x^*) \cap \tilde{V}(x^*) \subset O$ . En plus il existe  $W \subset \mathbb{R}^n$  avec  $W$  voisinage de  $x^*$  tel que  $x^*$  est un point de minimum de  $J$  sur  $O \cap W$ . Mais  $\tilde{O}(x^*) \cap \tilde{V}(x^*) \subset O$  donc  $\tilde{O}(x^*) \cap \tilde{V}(x^*) \cap W \subset O \cap W$  ce qui nous donne que  $x^*$  est un point de minimum de  $J$  sur  $\tilde{O}(x^*) \cap \tilde{V}(x^*) \cap W$ . Comme  $\tilde{V}(x^*) \cap W$  est un voisinage de  $x^*$  alors  $x^*$  est un point de minimum local de  $J$  sur  $\tilde{O}(x^*)$ .

On utilise alors le Théorème des multiplicateurs de Lagrange, donc il existe  $\tilde{p}_1, \dots, \tilde{p}_l \in \mathbb{R}$  tels que

$$\nabla J(x^*) + \sum_{k=1}^l \tilde{p}_k \nabla \theta_k(x^*) = 0. \quad (4.9)$$

Nous posons alors  $p_1^*, p_2^*, \dots, p_m^* \in \mathbb{R}$  tels que

$$\begin{aligned} p_i^* &= \tilde{p}_i & \text{si } & 1 \leq i \leq l \\ p_i^* &= 0 & \text{si } & l+1 \leq i \leq m. \end{aligned}$$

Alors les égalités (4.6) et (4.7) du Théorème 4.5 sont satisfaites (observer qu'on a  $p_i^* \theta_i(x^*) = 0, \forall i = 1, \dots, m$ ).

Pour finir la preuve du Théorème, il reste encore à montrer :

$$p_i^* \geq 0, \quad \forall i = 1, \dots, l. \quad (4.10)$$

Nous considérons deux sous-cas :

**Cas b1)**  $l = 1$  (c'est à dire :  $I(u^*) = \{1\}$ ).

Nous avons alors de (4.9) :

$$\nabla J(x^*) + p_1^* \nabla \theta_1(x^*) = 0.$$

En faisant le produit scalaire de cette égalité par  $\nabla \theta_1(x^*)$  (remarquer que  $\nabla \theta_1(x^*) \neq 0$ ) on obtient

$$p_1^* \|\nabla \theta_1(x^*)\|^2 = - \langle \nabla J(x^*), \nabla \theta_1(x^*) \rangle.$$

D'autre part, du Corollaire 4.7 avec  $v = x^*$  et  $w = -\nabla \theta_1(x^*)$  on déduit

$$\langle \nabla J(x^*), -\nabla \theta_1(x^*) \rangle \geq 0$$

On déduit alors immédiatement

$$p_1^* \geq 0.$$

**Cas b2)**  $l \geq 2$ .

On va montrer uniquement

$$p_1^* \geq 0, \quad (4.11)$$

les preuves des autre inégalités étant identiques.

L'idée de la preuve est de construire une suite  $\{w^{(q)}\}_{q \in \mathbb{N}^*} \subset \mathbb{R}^n$  telle que

$$\begin{aligned} \langle w^{(q)}, \nabla\theta_1(x^*) \rangle &= -1 \\ \langle w^{(q)}, \nabla\theta_k(x^*) \rangle &= -\frac{1}{q} \quad \forall k = 2, \dots, l \end{aligned} \quad (4.12)$$

Pour cela, on construit  $w^{(q)}$  sous la forme

$$w^{(q)} = \sum_{j=1}^l \alpha_j \nabla\theta_j(x^*)$$

avec  $\alpha_j \in \mathbb{R}$  à choisir de sorte que (4.12) soit vraie. Il est clair qu'il faut alors :

$$\sum_{j=1}^l \alpha_j \langle \nabla\theta_k(x^*), \nabla\theta_j(x^*) \rangle = \begin{cases} -1 & \text{si } k = 1 \\ -1/q & \text{si } k = 2, \dots, l \end{cases}$$

Ceci est une égalité du type  $B\alpha = b$  où  $B$  est la matrice carrée de taille  $l$  donnée par

$$B_{kj} = \langle \nabla\theta_k(x^*), \nabla\theta_j(x^*) \rangle$$

$\alpha$  est le vecteur inconnu  $\alpha = (\alpha_1, \dots, \alpha_l)^T$  et  $b$  est le vecteur dont les éléments  $b_k$  sont donnés par

$$b_k = \begin{cases} -1 & \text{si } k = 1 \\ -1/q & \text{si } k = 2, \dots, l \end{cases}$$

Comme les vecteurs  $\nabla\theta_1(x^*), \dots, \nabla\theta_l(x^*)$  sont indépendents par hypothèse, la matrice  $B$  est inversible (voir Lemme 4.8) donc l'équation précédente admet une solution  $\alpha$  unique. Donc la suite  $w^{(q)}$  avec la propriété (4.12) existe.

D'autre part, l'égalité (4.9) nous donne

$$\nabla J(x^*) + \sum_{k=1}^l p_k^* \nabla\theta_k(x^*) = 0.$$

En multipliant cette égalité par  $w^{(q)}$  on obtient

$$\langle \nabla J(x^*), w^{(q)} \rangle = - \sum_{k=1}^l p_k^* \langle \nabla\theta_k(x^*), w^{(q)} \rangle$$

ce qui donne à l'aide de (4.12)

$$\langle \nabla J(x^*), w^{(q)} \rangle = p_1^* + \frac{1}{q} \sum_{k=2}^l p_k^*. \quad (4.13)$$

Nous avons aussi

$$\langle \nabla J(x^*), w^{(q)} \rangle \geq 0$$

comme conséquence du Corollaire 4.7 avec  $w = w^{(q)}$ . En passant alors à la limite  $q \rightarrow +\infty$  en (4.13), on obtient l'inégalité (4.11), ce qui finit la preuve du Théorème 4.5.

**Remarque :** Si on considère l'ensemble

$$O = \{x \in \mathbb{R}^n, \theta_i(x) \leq 0, i = 1, \dots, m_1, \quad \varphi_j(x) = 0, j = 1, \dots, m_2\}$$

avec  $m_2 \geq 1$  alors on peut considérer  $O$  comme donné uniquement par des contraintes inégalités, en écrivant

$$O = \{x \in \mathbb{R}^n, \theta_i(x) \leq 0, i = 1, \dots, m_1, \quad \varphi_j(x) \leq 0, j = 1, \dots, m_2, \quad -\varphi_j(x) \leq 0, j = 1, \dots, m_2\}$$

Il est facile de voir qu'aucun point  $x$  de  $O$  n'est régulier, car d'une part les deux contraintes  $\varphi_j(x) \leq 0$  et  $-\varphi_j(x) \leq 0$  sont actives et d'autre part, aucune famille de vecteurs qui contient à la fois  $\nabla \varphi_j(x)$  et  $-\nabla \varphi_j(x)$  ne peut être indépendante.

**Conclusion :** L'hypothèse de régularité n'est satisfaite pour aucun  $x \in O$  si  $O$  comporte au moins une contrainte égalité transformée artificiellement en deux contraintes inégalités.

**Exemple :** L'ensemble

$$O = \{x \in \mathbb{R}^2, x_1^2 - x_2 \leq 0, x_2 - x_1 - 4 = 0\}$$

peut s'écrire sous la forme

$$O = \{x \in \mathbb{R}^2, x_1^2 - x_2 \leq 0, x_2 - x_1 - 4 \leq 0, -x_2 + x_1 + 4 \leq 0\}.$$

En introduisant les fonctions :  $\theta_1, \theta_2, \theta_3 : \mathbb{R}^2 \rightarrow \mathbb{R}$  données par

$$\theta_1(x) = x_1^2 - x_2, \theta_2(x) = x_2 - x_1 - 4, \theta_3(x) = -x_2 + x_1 + 4$$

on peut écrire  $O$  sous la forme standard

$$O = \{x \in \mathbb{R}^2, \theta_i(x) \leq 0, i = 1, 2, 3\}.$$

Il est clair que pour tout  $x \in O$  nous avons  $\{2, 3\} \subset I(x)$ . D'autre part,  $\nabla \theta_2(x) = (-1, 1)^T$  et  $\nabla \theta_3(x) = (1, -1)^T = -\nabla \theta_2(x)$ . Donc aucune famille de vecteurs qui inclut  $\nabla \theta_2(x)$  et  $\nabla \theta_3(x)$  ne peut être indépendante, donc aucun  $x \in O$  n'est régulier.

Pour remédier à ce genre d'inconvénient, il y a deux solutions :

1. Eliminer les contraintes égalités en éliminant des variables (quand cela est possible)
2. Affaiblir l'hypothèse de régularité du Théorème 4.5.

Dans la suite on choisira cette dernière possibilité.

**Définition 4.9.** On dit que les contraintes de  $O$  sont **qualifiés** en  $v \in O$  si

a) Soit  $I(v) = \emptyset$

b) Soit il existe  $w \in \mathbb{R}^n$  tel qu'on a  $\forall i \in I(v)$  :

$$\langle \nabla \theta_i(v), w \rangle \leq 0,$$

avec en plus

$$\langle \nabla \theta_i(v), w \rangle < 0 \quad \text{si } \theta_i \text{ est non-affine.}$$

**Remarques :**

1. Si  $\theta_i$  est affine pour tout  $i \in I(v)$  alors les contraintes de  $O$  sont qualifiés en  $v$  (prendre  $w = 0$  dans la partie **b**) de la Définition 4.9). En particulier si les fonctions  $\theta_i$  sont affines pour tout  $i = 1, \dots, m$  alors les contraintes sont qualifiées en tout point de  $O$ .
2. Comme conséquence de la Définition 4.9 le vecteur  $w$  "pointe" vers l'ensemble  $O$  (donc  $\exists t_0 > 0$  tel que  $v + tw \in O, \forall t \in [0, t_0]$ ) (voir Lemme 4.6 et la Remarque après).

Nous avons :

**Proposition 4.10.** *Si un point  $v \in O$  est régulier alors les contraintes de  $O$  sont qualifiés en  $v$ . Autrement dit, la condition de qualification de la Définition 4.9 est plus **faible** que la condition de régularité de la Définition 4.4.*

*Démonstration.* Nous avons construit dans la preuve du Théorème 4.5 une suite  $w^{(q)}$  telle que  $\langle \nabla \theta_i(v), w^{(q)} \rangle < 0, i = 1, \dots, l$ . Ceci prouve le résultat.  $\square$

On a alors le résultat suivant, plus forte que le Théorème 4.5 ; ce sera un résultat admis !

**Théorème 4.11.** *Soit  $x^* \in O$  tel que les contraintes de  $O$  sont qualifiés en  $x^*$ . Si  $x^*$  est un point de minimum local de  $J$  sur  $O$  alors il existe  $p_1^*, p_2^*, \dots, p_m^* \in [0, +\infty[$  (appelés **multiplicateurs de Karush-Kuhn-Tucker**) tels que :*

$$\nabla J(x^*) + \sum_{i=1}^m p_i^* \nabla \theta_i(x^*) = 0 \tag{4.14}$$

$$p_i^* \theta_i(x^*) = 0, \quad \forall i = 1, 2, \dots, m. \tag{4.15}$$

**Remarque :**

l'énoncé de ce dernier résultat s'obtient à partir de l'énoncé du Théorème 4.5 en remplaçant l'hypothèse " $x^*$  point régulier de  $O$ " par l'hypothèse plus faible "*les contraintes de  $O$  sont qualifiés en  $x^*$* ".

Le résultat suivant nous donne une condition suffisante pour la qualification en tout point de  $O$  :

**Proposition 4.12.** *Supposons que les fonctions  $\theta_1, \theta_2, \dots, \theta_m$  sont convexes. Supposons aussi que*

- ou bien toutes les fonctions  $\theta_1, \theta_2, \dots, \theta_m$  sont affines
- ou bien il existe  $y \in \mathbb{R}^n$  tel que pour tout  $i = 1, \dots, m$  on a

$$\theta_i(y) \leq 0$$

$$\theta_i(y) < 0 \quad \text{si } \theta_i \text{ n'est pas affine.}$$

Alors pour tout  $x \in O$  les contraintes de  $O$  sont qualifiées en  $x$ .

*Démonstration.* Soit  $x \in O$ . Supposons que  $I(x) \neq \emptyset$  (sinon OK,  $x$  est qualifié). Supposons aussi que toutes les fonctions  $\{\theta_i\}_{i=1, \dots, m}$  ne sont pas affines (sinon OK,  $x$  est qualifié, voir la première remarque après la Définition 4.9). Alors il existe  $y \in \mathbb{R}^n$  avec la propriété :

$$\theta_i(y) \leq 0, \quad \forall i = 1, \dots, m \text{ et}$$

pour tout  $i = 1, \dots, m$  tel que  $\theta_i$  n'est pas affine, on a  $\theta_i(y) < 0$ .

Alors par convexité de  $\theta_i$  on a

$$\theta_i(y) \geq \theta_i(x) + \langle \nabla \theta_i(x), y - x \rangle.$$

En utilisant l'égalité  $\theta_i(x) = 0$  pour tout  $i \in I(x)$  on obtient

$$\langle \nabla \theta_i(x), y - x \rangle \leq \theta_i(y), \quad \forall i \in I(x).$$

En utilisant les propriétés de  $\theta_i(y)$  on déduit

$$\langle \nabla \theta_i(x), y - x \rangle \leq 0, \quad \forall i \in I(x)$$

et

$$\langle \nabla \theta_i(x), y - x \rangle < 0, \quad \forall i \in I(x) \text{ si } \theta_i \text{ est non-affine.}$$

Donc la partie **b)** de la Définition 4.9 est satisfaite avec  $w = y - x$ , ce qui finit la preuve.  $\square$

### Exemple 1

$$O_1 = \{x \in \mathbb{R}^2, x_2 \geq x_1^2, x_2 - x_1 = 4\}.$$

Il est clair qu'on peut écrire  $O_1$  sous la forme

$$O_1 = \{x \in \mathbb{R}^2, \theta_i(x) \leq 0, i = 1, 2, 3\}$$

avec  $\theta_1, \theta_2, \theta_3 : \mathbb{R}^2 \rightarrow \mathbb{R}$  définies par

$$\theta_1(x) = x_1^2 - x_2, \quad \theta_2(x) = x_2 - x_1 - 4, \quad \theta_3(x) = -x_2 + x_1 + 4.$$

On observe que  $\theta_1, \theta_2$  et  $\theta_3$  sont convexes, avec en plus  $\theta_2$  et  $\theta_3$  affines. En prenant  $y = (0, 4)^T$  nous avons :

$$\theta_1(y) < 0, \quad \theta_2(y) = \theta_3(y) = 0.$$



Alors les hypothèses de la Proposition 4.12 sont satisfaites, donc tous les points de  $O_1$  sont qualifiés.

**Exemple 2**

$$O_2 = \{x \in \mathbb{R}^2, x_2 = x_1^2, x_2 - x_1 \leq 4\}.$$

On écrit  $O_2$  sous la forme

$$O_2 = \{x \in \mathbb{R}^2, \theta_i(x) \leq 0, i = 1, 2, 3\}$$

avec

$$\theta_1(x) = x_1^2 - x_2, \quad \theta_2(x) = -x_1^2 + x_2, \quad \theta_3(x) = x_2 - x_1 - 4.$$

Comme  $\theta_2$  n'est pas convexe, la Proposition 4.12 ne s'applique pas (ceci n'implique pas automatiquement la non-qualification!) Soit  $x \in O_2$ . Il est clair que  $\{1, 2\} \subset I(x)$ . Pour avoir la qualification, il faudrait qu'il existe  $w \in \mathbb{R}^2$  tel que

$$\langle \nabla\theta_1(x), w \rangle < 0 \quad \text{et} \quad \langle \nabla\theta_2(x), w \rangle < 0$$

or ceci est impossible car  $\nabla\theta_2(x) = -\nabla\theta_1(x)$ . Donc aucun point de  $O_2$  n'est qualifié.

### 4.2.2 Théorie générale du point selle

Soit  $U$  et  $P$  deux ensembles quelconques et l'application  $\mathcal{L} : U \times P \rightarrow \mathbb{R}$

**Définition 4.13.** On dit que le couple  $(u^*, p^*) \in U \times P$  est un **point selle** de  $\mathcal{L}$  si on a

$$\mathcal{L}(u^*, p) \leq \mathcal{L}(u^*, p^*) \leq \mathcal{L}(u, p^*), \quad \forall u \in U, \quad \forall p \in P$$

(Autrement dit :  $u^*$  est un point de minimum de la fonction  $u \in U \rightarrow \mathcal{L}(u, p^*) \in \mathbb{R}$  et  $p^*$  est un point de maximum de la fonction  $p \in P \rightarrow \mathcal{L}(u^*, p) \in \mathbb{R}$ .)

**Exemple :** Prendre  $U = P = \mathbb{R}$  et  $\mathcal{L}(u, p) = u^2 - p^2$ . Il est facile de voir que  $(0, 0)$  est un point selle de  $\mathcal{L}$ .

On introduit les fonctions suivantes :  
 $G : U \rightarrow \overline{\mathbb{R}}$  et  $H : P \rightarrow \overline{\mathbb{R}}$  données par

$$G(u) = \sup_{q \in P} \mathcal{L}(u, q)$$

et respectivement par

$$H(p) = \inf_{v \in U} \mathcal{L}(v, p).$$

**Remarque :**  $(u^*, p^*)$  est un point selle de  $\mathcal{L}$  si et seulement si

$$G(u^*) = \mathcal{L}(u^*, p^*) = H(p^*).$$

On a le résultats suivant :

**Proposition 4.14.**

$$\sup_{p \in P} \inf_{u \in U} \mathcal{L}(u, p) \leq \inf_{u \in U} \sup_{p \in P} \mathcal{L}(u, p)$$

(autrement dit :  $\sup_{p \in P} H(p) \leq \inf_{u \in U} G(u)$ )

*Démonstration.* Soit  $(u, p) \in U \times P$  fixé arbitraire. Il est clair qu'on a

$$\inf_{v \in U} \mathcal{L}(v, p) \leq \mathcal{L}(u, p) \leq \sup_{q \in P} \mathcal{L}(u, q)$$

ce qui nous donne

$$H(p) \leq G(u), \quad \forall (u, p) \in U \times P.$$

En passant au "supremum" en  $p$  et au "infimum" en  $u$  on obtient le résultat. □

Nous avons :

**Théorème 4.15.** *Si  $(u^*, p^*)$  est un point selle de  $\mathcal{L}$  sur  $U \times P$  alors*

$$\sup_{p \in P} \inf_{u \in U} \mathcal{L}(u, p) = \mathcal{L}(u^*, p^*) = \inf_{u \in U} \sup_{p \in P} \mathcal{L}(u, p)$$

(autrement dit :  $\sup_{p \in P} H(p) = \mathcal{L}(u^*, p^*) = \inf_{u \in U} G(u)$ ).

*Démonstration.* La définition du point selle nous dit

$$H(p^*) = \mathcal{L}(u^*, p^*) = G(u^*)$$

ce qui nous permet d'écrire

$$\sup_{p \in P} H(p) \geq \mathcal{L}(u^*, p^*) \geq \inf_{u \in U} G(u).$$

A l'aide de la Proposition 4.14 on obtient le résultat attendu. □

**Exemple :** Reprennons l'exemple précédent :  $\mathcal{L} : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ ,  $\mathcal{L}(u, p) = u^2 - p^2$ .  
Nous avons

$$G(u) = \sup_{p \in \mathbb{R}} (u^2 - p^2) = u^2$$

et

$$H(p) = \inf_{u \in \mathbb{R}} (u^2 - p^2) = -p^2$$

On aussi :

$$\inf_{u \in \mathbb{R}} G(u) = 0, \quad \sup_{p \in \mathbb{R}} H(p) = 0, \quad \mathcal{L}(0, 0) = 0$$

(le point  $(0, 0)$  étant le point selle de  $\mathcal{L}$ ), ce qui vérifie l'égalité du Théorème 4.15.

### 4.2.3 Applications de la théorie du point selle à l'optimisation

On utilise ici de nouveau les notations du début de la Section 4.2, spécifiques aux problèmes d'optimisation.

Nous introduisons la fonction  $L : \mathbb{R}^n \times \mathbb{R}_+^m \rightarrow \mathbb{R}$  donnée par

$$L(x, p) = J(x) + \sum_{i=1}^m p_i \theta_i(x) \equiv J(x) + \langle p, \theta(x) \rangle$$

(où on voit  $p$  comme un vecteur colonne de composantes  $p_1, p_2, \dots, p_m$ ). On a le résultat suivant :

**Théorème 4.16.** *Si  $(x^*, p^*)$  est un point selle de  $L$  sur  $\mathbb{R}^n \times \mathbb{R}_+^m$  alors  $x^*$  est un point de minimum de  $J$  sur  $O$ . En plus  $(x^*, p^*)$  satisfait les conditions de Karush-Kuhn-Tucker, c'est à dire*

$$\begin{cases} \nabla J(x^*) + \sum_{i=1}^m p_i^* \nabla \theta_i(x^*) = 0 \\ p_i^* \theta_i(x^*) = 0, \quad \forall i = 1, \dots, m. \end{cases} \quad (4.16)$$

*Démonstration.* Nous avons

$$L(x^*, p^*) \geq L(x^*, p), \quad \forall p \in \mathbb{R}_+^m$$

ce qui donne

$$\sum_{i=1}^m p_i^* \theta_i(x^*) \geq \sum_{i=1}^m p_i \theta_i(x^*), \quad \forall p \in \mathbb{R}_+^m. \quad (4.17)$$

En prenant respectivement  $p = 0$  et  $p = 2p^*$  dans l'inégalité précédente on obtient

$$\langle p^*, \theta(x^*) \rangle = 0. \quad (4.18)$$

Maintenant pour un  $j \in \{1, 2, \dots, m\}$  fixé prenons dans (4.17)  $p = p^* + e_j$ . On déduit alors  $\theta_j(x^*) \leq 0$ . Comme ceci est vrai pour tout  $j$ , on déduit

$$x^* \in O. \quad (4.19)$$

D'autre part, on se sert de l'inégalité

$$L(x^*, p^*) \leq L(x, p^*), \quad \forall x \in \mathbb{R}^n$$

c'est à dire

$$J(x^*) + \langle p^*, \theta(x^*) \rangle \leq J(x) + \langle p^*, \theta(x) \rangle \quad \forall x \in \mathbb{R}^n. \quad (4.20)$$

En utilisant l'égalité (4.18) et le fait que  $\langle p^*, \theta(x) \rangle \leq 0$  pour  $x \in O$ , on déduit

$$J(x^*) \leq J(x), \quad \forall x \in O.$$

Ceci nous dit que  $x^*$  est un point de minimum de  $J$  sur  $O$ .

D'autre part, de (4.20) on déduit que  $x^*$  est un point de minimum de la fonction  $x \in \mathbb{R}^n \rightarrow J(x) + \sum_{i=1}^m p_i^* \theta_i(x)$ . Ceci nous donne, en utilisant le Théorème 2.15, la première égalité de (4.16). Finalement, de (4.18) et (4.19) on a

$$\sum_{j=1}^m p_j^* \theta_j(x^*) = 0 \quad \text{et} \quad p_i^* \theta_i(x^*) \leq 0 \quad \text{pour tout } i \in \{1, \dots, m\}$$

ce qui nous donne la deuxième partie de (4.16). Ceci finit la preuve du Théorème.  $\square$

**Remarque :** Ce résultat nous donne une condition suffisante pour avoir un minimum de  $J$  sur  $O$ . On verra aussi une réciproque, sous des hypothèses supplémentaires. On cherchera alors à résoudre le problème appelé **le problème dual** :

$$\sup_{p \in \mathbb{R}_+^m} \inf_{u \in \mathbb{R}^n} L(u, p)$$

(justifié par le Théorème 4.15) plus simple à résoudre que le problème de minimisation de  $J$  sur  $O$  (qu'on appellera **problème primal**).

#### 4.2.4 Le cas convexe

On verra ici des réciproques des Théorèmes 4.5 et 4.16 dans le cas où  $J$  et  $\theta_j$  sont convexes.

**Théorème 4.17.** (Réciproque du Théorème 4.5).

Supposons que toutes les fonctions  $J, \theta_1, \theta_2, \dots, \theta_m$  sont **convexes**. Supposons aussi qu'il existe  $p_1^*, p_2^*, \dots, p_m^* \in [0, +\infty[$  et  $x^* \in O$  tels que les conditions de Karush-Kuhn-Tucker (4.6) et (4.7) du Théorème 4.5 soient satisfaites. Alors

a)  $x^*$  est un point de minimum de  $J$  sur  $O$

b)  $(x^*, p^*)$  est un point selle de  $L$

(Rappel :  $L : \mathbb{R}^n \times \mathbb{R}_+^m \rightarrow \mathbb{R}$ ,  $L(x, p) = J(x) + \langle p, \theta(x) \rangle$ ).

*Démonstration.* a) Nous avons

$$J(x^*) \leq J(x^*) - \sum_{i=1}^m p_i^* \theta_i(y), \quad \forall y \in O \quad (\text{car } \theta_i(y) \leq 0, p_i^* \geq 0).$$

Ensuite de (4.7) et de la convexité de  $\theta_i$  on déduit

$$p_i^* \theta_i(y) = p_i^* [\theta_i(y) - \theta_i(x^*)] \geq p_i^* \langle \nabla \theta_i(x^*), y - x^* \rangle$$

On obtient alors des inégalités précédentes et de (4.6) :

$$J(x^*) \leq J(x^*) + \langle \nabla J(x^*), y - x^* \rangle, \quad \forall y \in O.$$

En utilisant cette fois-ci la convexité de  $J$  nous obtenons

$$J(x^*) \leq J(y), \quad \forall y \in O$$

ce qui finit la preuve de **a**).

**b**) Nous avons

$$L(x^*, p) = J(x^*) + \langle p, \theta(x^*) \rangle \leq J(x^*) + \langle p^*, \theta(x^*) \rangle = L(x^*, p^*), \quad \forall p \in \mathbb{R}_+^m$$

car  $\langle p, \theta(x^*) \rangle \leq 0$  et  $\langle p^*, \theta(x^*) \rangle = 0$ .

D'autre part, comme les fonctions  $J, \theta_1, \dots, \theta_m$  sont convexes alors l'application  $x \in \mathbb{R}^n \rightarrow L(x, p^*)$  est convexe. Mais l'égalité (4.6) nous dit que le gradient de cette application s'annule en  $x^*$ . On déduit alors (voir la remarque après Théorème 2.19) que  $x^*$  est un point de minimum de cette application et ceci finit la preuve.  $\square$

On finit cette section par les corollaires suivants :

**Corollaire 4.18.** (*Lien entre point selle et conditions de Karush-Kuhn-Tucker*)

Supposons que les fonctions  $J, \theta_1, \dots, \theta_m$  sont convexes. Alors  $(x^*, p^*)$  est point selle de  $L$  si et seulement si  $x^* \in O$  et les conditions (4.6) et (4.7) sont satisfaites.

*Démonstration.* C'est une conséquence immédiate des Théorèmes 4.16 et 4.17 partie **b**).  $\square$

**Corollaire 4.19.** (*Réciproque du Théorème 4.16*)

Supposons que les fonctions  $J, \theta_1, \dots, \theta_m$  sont convexes. Soit  $x^*$  un point de minimum de  $J$  sur  $O$  tel que les contraintes de  $O$  sont qualifiées en  $x^*$ . Alors il existe  $p^* \in \mathbb{R}_+^m$  tel que  $(x^*, p^*)$  est un point selle de  $L$ .

*Démonstration.* C'est une conséquence immédiate des Théorèmes 4.11 et 4.17 partie **b**).  $\square$

### 4.3 Algorithmes de minimisation avec contraintes

Le but de cette section est de donner des algorithmes numériques pour la résolution du problème de minimisation avec contraintes :

$$\min_{u \in U} J(u) \tag{4.21}$$

avec  $U \subset \mathbb{R}^n$  un ensemble de contraintes.

On rappelle les deux cas de contraintes que nous considérons dans ce cours :

1.  $U$  est un pavé, c'est à dire, il est de la forme  $U = [a_1, b_1] \times [a_2, b_2] \times \dots \times [a_n, b_n]$  (avec la convention qu'on peut avoir  $]-\infty, \infty[$  à la place de  $][a_i, \infty[$  ou  $]-\infty, \infty[$  à la place de  $][a_i, b_i]$ , donc le pavé peut être de volume infini).

2.  $U$  est de la forme

$$U = \{x \in \mathbb{R}^n, \theta_i(x) \leq 0, i = 1, 2, \dots, m\}$$

avec  $m \in \mathbb{N}^*$  et  $\theta_1, \theta_2, \dots, \theta_m$  des fonctions de  $\mathbb{R}^n$  dans  $\mathbb{R}$  données ( $U$  est donné par des contraintes inégalités larges).

Dans le premier cas on utilisera des extension naturelles des méthodes de minimisation sur  $\mathbb{R}^n$ , vues en Chapitre 3. Dans le deuxième cas on utilisera des méthodes du type dual.

### 4.3.1 Méthodes de relaxation

On suppose ici

$$U = [a_1, b_1] \times [a_2, b_2] \times \dots \times [a_n, b_n] \equiv \prod_{i=1}^n [a_i, b_i]. \quad (4.22)$$

L'algorithme dans ce cas ressemble à l'algorithme de relaxation pour minimisation sans contraintes.

**Algorithme :**

On se donne  $u^{(k)} = (u_1^{(k)}, u_2^{(k)}, \dots, u_n^{(k)})^T$  et on va calculer  $u^{(k+1)} = (u_1^{(k+1)}, u_2^{(k+1)}, \dots, u_n^{(k+1)})^T$  en  $n$  pas successifs. On a :

$$\begin{aligned} J(u_1^{(k+1)}, u_2^{(k)}, \dots, u_n^{(k)}) &= \min_{y \in [a_1, b_1]} J(y, u_2^{(k)}, \dots, u_n^{(k)}) \\ J(u_1^{(k+1)}, u_2^{(k+1)}, u_3^{(k)}, \dots, u_n^{(k)}) &= \min_{y \in [a_2, b_2]} J(u_1^{(k+1)}, y, u_3^{(k)}, \dots, u_n^{(k)}) \\ &\dots \\ J(u_1^{(k+1)}, \dots, u_{n-1}^{(k+1)}) &= \min_{y \in [a_n, b_n]} J(u_1^{(k+1)}, \dots, u_{n-1}^{(k+1)}, y) \end{aligned}$$

Comme dans le cas sans contraintes, on a le théorème de convergence suivant :

**Théorème 4.20.** *Si  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  est une fonction elliptique et l'ensemble  $U$  est donné par (4.22) alors la méthode de relaxation pour la minimisation de  $J$  sur  $U$  est bien définie et converge.*

*Démonstration.* La preuve du fait que la méthode est bien définie se fait exactement comme dans le Théorème 3.3. La convergence est admise !  $\square$

### 4.3.2 Méthodes de projection

Rappelons d'abord le résultat fondamental suivant :

**Théorème 4.21.** (Théorème de projection sur un convexe fermé)

Soit  $U \subset \mathbb{R}^n$  un ensemble convexe, fermé, non-vidé et soit  $v \in \mathbb{R}^n$ . Alors il existe un unique élément  $Pv \in U$  (on peut le noter pour plus de précision par  $P_U(v)$ ) tel que

$$\|v - P_U(v)\| = \inf_{u \in U} \|v - u\| = \min_{u \in U} \|v - u\|.$$

L'élément  $P_U(v) \in U$  s'appelle **la projection de  $v$  sur  $U$  en  $\mathbb{R}^n$** . En plus on a

a) L'élément  $P_U(v) \in U$  satisfait aussi

$$\langle v - P_U(v), u - P_U(v) \rangle \leq 0, \quad \forall u \in U \quad (4.23)$$

b) Si  $w \in U$  est tel que

$$\langle v - w, u - w \rangle \leq 0, \quad \forall u \in U \quad (4.24)$$

alors  $w = P_U(v)$  (donc  $w = P_U(v)$  est l'unique élément de  $U$  satisfaisant (4.24)).

c) On a

$$\|P_U(v_1) - P_U(v_2)\| \leq \|v_1 - v_2\|, \quad \forall v_1, v_2 \in \mathbb{R}^n$$

(la fonction  $P_U$  n'augmente pas les distances). Ceci implique que  $P_U$  est une fonction lipschitzienne, donc **continue**.

d) On a

$$v = P_U(v) \quad \text{si et seulement si} \quad v \in U$$

e) Si  $U$  est le sous-espace affine de  $\mathbb{R}^n$  donné par  $U = a + U_0$  avec  $a \in \mathbb{R}^n$  et  $U_0$  un sous-espace vectoriel de  $\mathbb{R}^n$  alors l'inégalité (4.23) devient l'égalité

$$\langle v - P_U(v), u \rangle = 0, \quad \forall u \in U_0$$

(c'est à dire  $v - P_U(v) \perp U_0$ ).

*Idée de la preuve*

Tout revient à minimiser sur  $U$  la fonction  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  définie par

$$g(u) = \|u - v\|^2.$$

Montrons d'abord que  $g$  est une fonction elliptique. Pour cela, on utilise les calculs et les résultats du Chapitre 2. On peut écrire

$$g(u) = \langle v - u, v - u \rangle = \langle u, u \rangle - 2 \langle u, v \rangle + \langle v, v \rangle.$$

ce qui donne

$$\nabla g(u) = 2u - 2v$$

$$\nabla^2 g(u) = 2I_n \quad (\text{donc constante indépendante de } u).$$

Comme  $2I_n$  est une matrice symétrique et définie positive (matrice SDP) alors  $g$  est elliptique.

Comme  $U$  est fermé convexe alors il existe un unique élément  $Pv = P_U(v) \in U$  tel que  $g(P_U(v)) \leq g(u)$ ,  $\forall u \in U$  c'est à dire

$$\|v - P_U(v)\| \leq \|v - u\|, \quad \forall u \in U.$$

ce qui donne le résultat principal du théorème. Nous avons aussi (voir Théorème 2.15) :

$$\langle \nabla g(P_U(v)), u - P_U(v) \rangle \geq 0, \quad \forall u \in U$$

ce qui nous donne facilement

$$\langle v - P_U(v), u - P_U(v) \rangle \leq 0, \quad \forall u \in U.$$

ce qui prouve **a)**. La partie **b)** est une conséquence de l'ellipticité et de l'unicité. D'autre part, soient  $v_1, v_2 \in \mathbb{R}^n$ . On a alors

$$\langle v_1 - Pv_1, u - Pv_1 \rangle \leq 0, \quad \forall u \in U$$

$$\langle v_2 - Pv_2, u - Pv_2 \rangle \leq 0, \quad \forall u \in U$$

En prenant  $u = Pv_2$  dans la première inégalité et  $u = Pv_1$  dans la deuxième et en faisant la somme, on trouve

$$\langle Pv_2 - Pv_1, v_1 - Pv_1 - v_2 + Pv_2 \rangle \leq 0$$

ce qui donne

$$\|Pv_2 - Pv_1\|^2 \leq \langle Pv_2 - Pv_1, v_2 - v_1 \rangle \leq \|Pv_2 - Pv_1\| \|v_2 - v_1\|$$

(on a utilisé l'inégalité de Cauchy-Schwarz). Ceci nous donne **c)** par simplification. La partie **d)** est évidente.

Pour la partie **e)**, soit  $U_0$  un sous-espace vectoriel de  $\mathbb{R}^n$ ,  $a \in \mathbb{R}^n$ ,  $U = a + U_0$  et  $h \in U_0$  arbitraire et fixé. On va prendre en (4.23) :  $u = Pv \pm h$  qui est un élément de  $U$  car  $Pv \in U$  et  $h \in U_0$ . Ceci donne

$$\pm \langle v - Pv, h \rangle \leq 0$$

d'où on déduit

$$\langle v - Pv, h \rangle = 0$$

ce qui finit la preuve.

**Exemple 1.**

Si  $n = 1$  et  $U = [a, b]$  (avec  $-\infty \leq a < b \leq +\infty$ ) alors il est facile de voir que pour tout  $v \in \mathbb{R}$  on a

$$P_U(v) = \begin{cases} a & \text{si } v < a & \text{(partie inexistente si } a = -\infty) \\ v & \text{si } v \in U \\ b & \text{si } v > b & \text{(partie inexistente si } b = +\infty) \end{cases}$$



On a évidemment le cas particulier : si  $U = [0, +\infty[$  alors

$$P_U(v) = \begin{cases} v & \text{si } v \geq 0 \\ 0 & \text{si } v < 0 \end{cases}$$

On peut aussi noter dans ce cas

$$P_U(v) = v^+, \quad \forall v \in \mathbb{R}.$$

### Exemple 2.

Supposons que  $U$  est un pavé, donc

$$U = \prod_{i=1}^n [a_i, b_i] \in \mathbb{R}^n$$

avec  $-\infty \leq a_i < b_i \leq +\infty$ ,  $i = 1, \dots, n$ . Alors  $\forall v = (v_1, v_2, \dots, v_n)^T \in \mathbb{R}^n$  on a

$$P_U(v) = (P_1 v_1, P_2 v_2, \dots, P_n v_n)^T$$

avec  $P_k v_k =$  la projection de  $v_k$  en  $\mathbb{R}$  sur l'intervalle  $[a_k, b_k]$  (voir TD pour la preuve).

*Cas particulier :* Si  $U = [0, +\infty[^n$  alors  $P_U(v) = (v_1^+, v_2^+, \dots, v_n^+)^T$ .

**Remarque :** Dans le cas où  $U$  est donné par des contraintes inégalité générales, alors il peut être difficile en pratique de calculer la projection en  $\mathbb{R}^n$  sur  $U$ .

Revenons au problème de minimisation (4.21). On supposera dans la suite que l'ensemble  $U$  des contraintes est **convexe** et **fermé**.

Supposons que  $x^* \in U$  est tel que

$$J(x^*) = \min_{x \in U} J(x).$$

Alors

$$\langle \nabla J(x^*), x - x^* \rangle \geq 0, \quad \forall x \in U \quad (4.25)$$

qui est équivalente pour tout  $\rho > 0$  à l'inégalité

$$\rho \langle -\nabla J(x^*), x - x^* \rangle \leq 0, \quad \forall x \in U.$$

Ceci donne

$$\langle [x^* - \rho \nabla J(x^*)] - x^*, x - x^* \rangle \leq 0, \quad \forall x \in U.$$

En utilisant le Théorème de projection cette dernière inégalité est équivalente à l'égalité

$$x^* = P_U(x^* - \rho \nabla J(x^*)).$$

Alors la solution du problème de minimisation (4.21) est un point fixe de l'application  $g : U \mapsto U$  donnée par

$$g(x) = P_U(x - \rho \nabla J(x)).$$

Une idée naturelle pour approcher numériquement ce point fixe est d'utiliser une méthode itérative, qui consiste à construire une suite  $x^{(k)} \in U$  par récurrence :

$$x^{(k+1)} = P_U(x^{(k)} - \rho \nabla J(x^{(k)})), \quad k \in \mathbb{N}$$

avec  $\rho > 0$  donné. Il est aussi possible de faire varier  $\rho$  à chaque pas. Donc l'algorithme général sera dans ce cas :

$$\begin{cases} x^{(k+1)} = P_U(x^{(k)} - \rho_k \nabla J(x^{(k)})), & k \in \mathbb{N} \\ x^{(0)} \text{ donné} \in U \end{cases} \quad (4.26)$$

avec  $\rho_k > 0$  données. C'est la **méthode de gradient avec projection à pas variable**. On l'appelle **méthode de gradient avec projection à pas fixe** si  $\rho_k$  est indépendant de  $k$ .

On a le résultat suivant :

**Théorème 4.22.** *Soit  $U \subset \mathbb{R}^n$  un ensemble fermé, convexe et non vide et  $J : \mathbb{R}^n \mapsto \mathbb{R}$  une fonction elliptique telle que  $\nabla J$  soit lipschitzienne (Donc il existe  $M > 0$  et  $\alpha > 0$  tels que  $\forall x, y \in \mathbb{R}^n$  on a*

$$\langle \nabla J(x) - \nabla J(y), x - y \rangle \geq \alpha \|x - y\|^2$$

et

$$\|\nabla J(x) - \nabla J(y)\| \leq M \|x - y\|. \quad )$$

Supposons qu'il existe des nombres réels  $a_1$  et  $a_2$  satisfaisant

$$0 < a_1 \leq a_2 < \frac{2\alpha}{M^2}$$

tels que

$$a_1 \leq \rho_k \leq a_2 \quad \forall k \in \mathbb{N}.$$

Alors la méthode (4.26) converge et la convergence est géométrique (c'est à dire, il existe  $C \geq 0$  et  $\beta \in [0, 1[$  tels que

$$\|x^{(k)} - x^*\| \leq C\beta^k, \quad \forall k \in \mathbb{N}. \quad )$$

*Démonstration.* Par définition nous avons

$$x^{(k+1)} = P_U(x^{(k)} - \rho_k \nabla J(x^{(k)}))$$

et aussi

$$x^* = P_U(x^* - \rho_k \nabla J(x^*))$$

En faisant la différence et en utilisant la Partie **c)** du Théorème de projection, on déduit

$$\|x^{(k+1)} - x^*\| \leq \|x^{(k)} - x^* - \rho_k [\nabla J(x^{(k)}) - \nabla J(x^*)]\|$$

c'est à dire

$$\|x^{(k+1)} - x^*\|^2 \leq \|x^{(k)} - x^*\|^2 + \rho_k^2 \|\nabla J(x^{(k)}) - \nabla J(x^*)\|^2 - 2\rho_k \langle x^{(k)} - x^*, \nabla J(x^{(k)}) - \nabla J(x^*) \rangle.$$

En utilisant les hypothèses sur  $J$  on déduit

$$\|x^{(k+1)} - x^*\|^2 \leq (1 - 2\alpha\rho_k + M^2\rho_k^2)\|x^{(k)} - x^*\|^2.$$

La suite de la preuve est exactement comme dans la preuve du Théorème 3.6 (à partir de l'inégalité (3.14)).  $\square$

**Remarque :** Cette méthode est applicable uniquement si on peut calculer facilement la projection  $P_U$ , par exemple si  $U$  est un pavé en  $\mathbb{R}^n$ . Pour un ensemble  $U$  donné par des contraintes inégalités larges, il n'est pas facile en général d'utiliser cette méthode.

**Exemple :** Supposons que  $J$  est une fonction quadratique associée à une matrice SDP, donc  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  est donnée par

$$J(x) = \frac{1}{2} \langle Ax, x \rangle - \langle b, x \rangle + c$$

avec  $A \in \mathcal{M}_n(\mathbb{R})$  une matrice carrée réelle et SDP,  $b \in \mathbb{R}^n$ ,  $c \in \mathbb{R}$ .

Supposons aussi que  $U = [0, +\infty[^n$ .

Rappelons qu'on a dans ce cas :

$$\nabla J(x) = Ax - b, \quad \forall x \in \mathbb{R}^n$$

$$P_U(x_1, x_2, \dots, x_n) = (x_1^+, x_2^+, \dots, x_n^+)$$

Alors (4.26) devient

$$x_i^{(k+1)} = \max\{x_i^{(k)} - \rho_k (Ax^{(k)} - b)_i, 0\}, \quad i = 1, \dots, n \text{ et } k \in \mathbb{N}.$$

### 4.3.3 Méthodes de pénalisation extérieure

On considère toujours le problème de minimisation (4.21).

On introduira une fonction  $\psi : \mathbb{R}^n \mapsto \mathbb{R}$  ayant les propriétés suivantes :

$$\begin{cases} \psi & \text{fonction continue et convexe} \\ \psi(x) \geq 0, & \forall x \in \mathbb{R}^n \\ \psi(x) = 0 & \text{si et seulement si } x \in U. \end{cases} \quad (4.27)$$

**Remarque :** Une fonction  $\psi$  avec les propriétés de (4.27) est appelée **fonction de pénalisation extérieure pour  $U$** . On introduit ensuite pour tout  $\epsilon > 0$  "assez petit" une fonction  $J_\epsilon : \mathbb{R}^n \mapsto \mathbb{R}$  définie par

$$J_\epsilon(x) = J(x) + \frac{1}{\epsilon}\psi(x). \quad (4.28)$$

La fonction  $J_\epsilon$  s'appelle **fonction pénalisée de  $J$** .

La **méthode de pénalisation extérieure** du problème (4.21) consiste à minimiser sur  $\mathbb{R}^n$  la fonction  $J_\epsilon$ , avec un  $\epsilon > 0$  qui "tend vers 0". La partie  $\frac{1}{\epsilon}\psi(x)$  pénalise l'éloignement de  $x$  par rapport à  $U$ .

On réduit donc le problème de minimisation avec contraintes (de  $J$  sur  $U$ ) à un problème de minimisation sans contraintes (de  $J_\epsilon$  sur  $\mathbb{R}^n$ ). On applique pour la minimisation sans contraintes de  $J_\epsilon$  tout algorithme vu en Chapitre 3.

Le résultat suivant de convergence justifie cette méthode :

**Théorème 4.23.** *Soit  $J : \mathbb{R}^n \mapsto \mathbb{R}$  une fonction continue, coercive et strictement convexe,  $U \subset \mathbb{R}^n$  un ensemble convexe, fermé et non-vide et  $\psi : \mathbb{R}^n \mapsto \mathbb{R}$  une fonction de pénalisation extérieure de  $U$  (donc satisfaisant les hypothèses de (4.27)).*

*Alors pour tout  $\epsilon > 0$  il existe un unique élément  $x_\epsilon \in \mathbb{R}^n$  vérifiant*

$$J_\epsilon(x_\epsilon) = \min_{x \in \mathbb{R}^n} J_\epsilon(x).$$

En plus on a

$$x_\epsilon \rightarrow x^* \quad \text{pour } \epsilon \mapsto 0$$

où  $x^*$  est l'unique point de  $U$  satisfaisant

$$J(x^*) \leq J(x), \quad \forall x \in U.$$

*Démonstration.* Il est facile de voir que pour tout  $\epsilon > 0$  fixé,  $J_\epsilon$  est une fonction continue, coercive et strictement convexe sur  $\mathbb{R}^n$ ; donc il existe un unique point de minimum de  $J_\epsilon$  sur  $\mathbb{R}^n$  qu'on va noter par  $x_\epsilon$ . Il est facile de voir que

$$J_\epsilon \geq J \quad \text{et} \quad J_\epsilon = J \quad \text{sur } U.$$

On déduit alors

$$J(x_\epsilon) \leq J_\epsilon(x_\epsilon) \leq J_\epsilon(x) = J(x), \quad \forall x \in U. \quad (4.29)$$

Ceci nous dit que la suite  $J(x_\epsilon)$  est bornée, ce qui implique que la suite  $x_\epsilon$  est bornée (car sinon, il y aurait une sous-suite de  $x_\epsilon$  dont la norme tend vers  $+\infty$ , ce qui impliquerait, par coercivité de  $J$ , que  $J(x_\epsilon) \mapsto +\infty$ ; ceci est impossible à cause de (4.29)).

Comme  $x_\epsilon$  est borné en  $\mathbb{R}^n$ , le Théorème de Bolzano-Weierstrass nous dit qu'il existe une sous-suite  $x_{\epsilon'}$  de  $x_\epsilon$  tel que  $x_{\epsilon'}$  converge vers un élément  $x^*$  de  $\mathbb{R}^n$ . En passant à la limite  $\epsilon' \mapsto 0$  en (4.29) on déduit, grâce à la continuité de  $J$  que

$$J(x^*) \leq J(x), \quad \forall x \in U. \quad (4.30)$$

Pour montrer que  $x^*$  est un point de minimum de  $J$  sur  $U$  il reste à montrer que  $x^* \in U$ ; grâce à la troisième propriété de (4.27) ceci revient à montrer que  $\psi(x^*) = 0$ . En effet nous avons :

$$0 \leq \psi(x_{\epsilon'}) = \epsilon' [J_{\epsilon'}(x_{\epsilon'}) - J(x_{\epsilon'})] \leq \epsilon' [J(x) - J(x_{\epsilon'})]$$

pour un  $x \in U$  fixé. Par continuité de  $J$  nous avons que  $J(x_{\epsilon'})$  converge, donc le terme  $J(x) - J(x_{\epsilon'})$  est borné. Ceci implique que  $\epsilon' [J(x) - J(x_{\epsilon'})] \rightarrow 0$  pour  $\epsilon' \rightarrow 0$ . On obtient alors  $\psi(x_{\epsilon'}) \mapsto 0$ . Par continuité de  $\psi$  on déduit  $\psi(x^*) = 0$ , c'est à dire  $x^* \in U$ .

Comme  $J$  est strictement convexe,  $x^*$  est l'unique point de minimum de  $J$  sur  $U$ .

Il nous reste à démontrer que toute la séquence  $x_\epsilon$  converge vers  $x^*$ . Montrons ce résultat par absurd : supposons que  $x_\epsilon$  ne converge pas vers  $x^*$  ; ceci implique qu'il existe une sous-suite  $x_{\epsilon''}$  de  $x_\epsilon$  et un  $\delta > 0$  tels que

$$\|x_{\epsilon''} - x^*\| \geq \delta. \quad (4.31)$$

En reproduisant pour  $x_{\epsilon''}$  le même argument que pour  $x_\epsilon$ , on peut extraire une sous-suite  $x_{\epsilon'''}$  de  $x_{\epsilon''}$  qui converge vers  $x^*$  (l'unicité de  $x^*$  est ici essentielle) ; ceci contredit (4.31) et finit la preuve du Théorème. □

**Remarque :** Cette méthode est utile s'il est facile de construire une fonction  $\psi$  avec les propriétés (4.27).

**Exemple :** On appelle **distance** d'un point  $x \in \mathbb{R}^n$  à l'ensemble  $U$ , notée  $\text{dist}(x, U)$  la quantité donnée par

$$\text{dist}(x, U) = \|x - P_U(x)\|.$$

(rappelons que  $U$  est fermé et convexe). On définit alors la fonction  $\psi : \mathbb{R}^n \mapsto \mathbb{R}$  par

$$\psi(x) = \text{dist}^2(x, U) = \|x - P_U(x)\|^2.$$

Il est très facile de montrer toutes les propriétés de (4.27) sauf la convexité de  $\psi$ . Pour la convexité de  $\psi$  dans un cas particulier de  $U$  voir TD.

#### 4.3.4 Méthode d'Uzawa

La méthode d'Uzawa est basée sur la formulation duale et la recherche des points selle et elle est bien adaptée pour des ensembles de contraintes définies par des inégalités larges.

On va considérer ici des ensembles de contraintes comme dans la Section 4.2, c'est à dire

$$U = O = \{x \in \mathbb{R}^n, \theta_i(x) \leq 0, \quad \forall i = 1, 2, \dots, m\}$$

avec  $\theta_1, \theta_2, \dots, \theta_m : \mathbb{R}^n \mapsto \mathbb{R}$ ,  $m \in \mathbb{N}^*$ . On peut encore écrire

$$U = O = \{x \in \mathbb{R}^n, \theta(x) \leq 0\}.$$

Rappelons que nous considérons le problème de minimisation (4.21).

Nous faisons dans cette partie les hypothèses suivantes :

1.  $J$  est elliptique, donc  $J$  est de classe  $C^1$  et il existe  $\alpha > 0$  tel que

$$\langle J(x) - J(y), x - y \rangle \geq \alpha \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n$$

2. Pour tout  $k = 1, \dots, m$  on a que  $\theta_k$  est une fonction convexe et de classe  $C^1$ . En plus  $\theta$  est lipschitzienne, donc il existe  $M > 0$  tel que

$$\|\theta(x) - \theta(y)\| \leq M\|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

3. Les contraintes de  $O$  sont qualifiées en tout point  $x \in O$ .

Nous introduisons l'application  $L : \mathbb{R}^n \times \mathbb{R}_+^m \mapsto \mathbb{R}$  donnée par

$$L(x, p) = J(x) + \langle p, \theta(x) \rangle = J(x) + \sum_{k=1}^m p_k \theta_k(x), \quad \forall x \in \mathbb{R}^n, \quad \forall p \in \mathbb{R}_+^m.$$

Grâce à la convexité de toutes les fonctions  $\theta_k$  l'ensemble  $U$  est convexe (Proposition 4.1). Comme l'ensemble  $U$  est fermé et convexe et  $J$  elliptique, il existe un unique point de minimum  $x^* \in U$  de  $J$  sur  $U$ . Le Corollaire 4.19 nous dit alors qu'il existe  $p^* = (p_1^*, p_2^*, \dots, p_m^*)^T \in \mathbb{R}_+^m$  tel que  $(x^*, p^*)$  est un point selle de  $L$ . Du Théorème 4.15 on déduit alors que

$$\sup_{p \in \mathbb{R}_+^m} \inf_{x \in \mathbb{R}^n} L(x, p) = L(x^*, p^*) = \inf_{x \in \mathbb{R}^n} \sup_{p \in \mathbb{R}_+^m} L(x, p).$$

Introduisons la fonction  $H : \mathbb{R}_+^m \mapsto \overline{\mathbb{R}}$  par

$$H(p) = \inf_{x \in \mathbb{R}^n} L(x, p).$$

La Proposition 2.5 et les hypothèses faites sur  $J$  et les  $\theta_k$  nous disent que pour tout  $p \in \mathbb{R}_+^m$  fixé l'application

$$x \in \mathbb{R}^n \mapsto L(x, p) \in \mathbb{R}$$

est elliptique. Alors il existe un unique point de minimum de cette application sur  $\mathbb{R}^n$ , que nous allons noter  $\bar{x}^p$ . Donc  $\bar{x}^p \in \mathbb{R}^n$  est tel que

$$L(\bar{x}^p, p) \leq L(x, p) \quad \forall x \in \mathbb{R}^n.$$

On déduit alors que  $H$  prend ses valeurs en  $\mathbb{R}$  et que

$$H(p) = \min_{x \in \mathbb{R}^n} L(x, p) = L(\bar{x}^p, p), \quad \forall p \in \mathbb{R}_+^m.$$

Le problème **dual** de (4.21) sera

$$\text{Trouver } \tilde{p}^* \in \mathbb{R}_+^m \text{ tel que } H(\tilde{p}^*) = \sup_{p \in \mathbb{R}_+^m} H(p) \quad (4.32)$$

(le problème (4.21) sera appelé problème **primal**).

**Remarque :** Le problème de minimisation (4.32) peut s'écrire

$$-H(\tilde{p}^*) = \inf_{p \in \mathbb{R}_+^m} [-H(p)].$$

Supposons que  $\tilde{p}^*$  est bien une solution du problème de maximisation (4.32). Alors on a

$$\sup_{p \in \mathbb{R}_+^m} \inf_{x \in \mathbb{R}^n} L(x, p) = \sup_{p \in \mathbb{R}_+^m} H(p) = H(\tilde{p}^*) = L(\tilde{x}^*, \tilde{p}^*)$$

où on a noté  $\tilde{x}^* = \bar{x}^{\tilde{p}^*}$ .

Alors le point selle  $(x^*, p^*)$  que nous cherchons, se trouve parmi les solutions  $(\tilde{x}^*, \tilde{p}^*)$  provenant de (4.32).

Comme l'ensemble des contraintes du problème dual (4.32) est un pavé (c'est l'ensemble  $\mathbb{R}_+^m$ ) on peut alors utiliser un algorithme de gradient avec projection à pas variable :

$$p^{(k+1)} = P_{\mathbb{R}_+^m} [p_i^{(k)} + \rho_k \nabla H(p^{(k)})]$$

ce qui donne, en tenant compte de l'expression de la projection sur  $\mathbb{R}_+^m$  :

$$p^{(k+1)} = \max \left\{ p_i^{(k)} + \rho_k \frac{\partial H}{\partial p_i}(p^{(k)}), 0 \right\} \quad i = 1, \dots, m \quad (4.33)$$

avec  $\rho_k > 0$ .

Il nous reste à calculer  $\frac{\partial H}{\partial p_i}$ .

Nous avons le **calcul formel** suivant, en supposant que l'application  $p \mapsto \bar{x}^p$  est de classe  $C^1$  : comme  $H(p) = J(\bar{x}^p) + \sum_{j=1}^m p_j \theta_j(\bar{x}^p)$  alors

$$\frac{\partial H}{\partial p_i} = \langle \nabla J(\bar{x}^p), \frac{\partial \bar{x}^p}{\partial p_i} \rangle + \theta_i(\bar{x}^p) + \sum_{j=1}^m p_j \langle \nabla \theta_j(\bar{x}^p), \frac{\partial \bar{x}^p}{\partial p_i} \rangle$$

Mais comme  $\nabla J(\bar{x}^p) + \sum_{j=1}^m p_j \nabla \theta_j(\bar{x}^p) = 0$  (conséquence immédiate du fait que  $\bar{x}^p$  est le point de minimum de la fonction  $x \mapsto L(x, p)$ ) alors il reste

$$\frac{\partial H}{\partial p_i} = \theta_i(\bar{x}^p).$$

Le résultat précis avec démonstration rigoureuse est donné dans le lemme suivant :

**Lemme 4.24.** *La fonction  $H$  est de classe  $C^1$  et on a  $\forall p \in \mathbb{R}_+^m$  :*

$$\frac{\partial H}{\partial p_i}(p) = \theta_i(\bar{x}^p), \quad i = 1, 2, \dots, m.$$

*Démonstration.* Soit  $p \in \mathbb{R}_+^m$  et  $t \in \mathbb{R}$  tel que  $p + te_i \in \mathbb{R}_+^m$ . Le but principal est de montrer que la limite pour  $t \mapsto 0$  de  $\frac{1}{t}[H(p + te_i) - H(p)]$  existe et de calculer cette limite.

**Étape 1.** On montre d'abord que l'application  $p \mapsto \bar{x}^p$  est continue.

Soit  $p \geq 0$  fixé et  $p' \geq 0$  tel que  $p'$  converge vers  $p$ . Les vecteurs  $\bar{x}^p$  et  $\bar{x}^{p'}$  satisfont respectivement

$$\nabla J(\bar{x}^p) + \sum_{j=1}^m p_j \nabla \theta_j(\bar{x}^p) = 0$$

et

$$\nabla J(\bar{x}^{p'}) + \sum_{j=1}^m p'_j \nabla \theta_j(\bar{x}^{p'}) = 0$$

En faisant la différence entre ces deux égalités et en faisant le produit scalaire avec  $\bar{x}^{p'} - \bar{x}^p$  on trouve

$$\begin{aligned} \langle \nabla J(\bar{x}^{p'}) - \nabla J(\bar{x}^p), \bar{x}^{p'} - \bar{x}^p \rangle + \sum_j p'_j \langle \nabla \theta_j(\bar{x}^{p'}) - \nabla \theta_j(\bar{x}^p), \bar{x}^{p'} - \bar{x}^p \rangle \\ + \sum_j (p'_j - p_j) \langle \nabla \theta_j(\bar{x}^p), \bar{x}^{p'} - \bar{x}^p \rangle = 0 \end{aligned}$$

On utilise ensuite l'inégalité d'ellipticité de  $J$  et le fait que les fonctions  $\theta_j$  sont convexes et on déduit

$$\alpha \|\bar{x}^{p'} - \bar{x}^p\|^2 \leq - \sum_j (p'_j - p_j) \langle \nabla \theta_j(\bar{x}^p), \bar{x}^{p'} - \bar{x}^p \rangle$$

En utilisant l'inégalité de Cauchy-Schwarz, on trouve

$$\alpha \|\bar{x}^{p'} - \bar{x}^p\|^2 \leq \sum_j |p'_j - p_j| \|\nabla \theta_j(\bar{x}^p)\| \cdot \|\bar{x}^{p'} - \bar{x}^p\|$$

ce qui nous donne immédiatement

$$\bar{x}^{p'} \mapsto \bar{x}^p \quad \text{si} \quad p' \mapsto p.$$

On a donc la continuité souhaitée.

**Etape 2.** Nous avons d'une part

$$H(p) = L(\bar{x}^p, p) \leq L(\bar{x}^{p+te_i}, p)$$

et d'autre part

$$H(p + te_i) = L(\bar{x}^{p+te_i}, p + te_i) \leq L(\bar{x}^p, p + te_i).$$

Ceci nous donne

$$H(p+te_i) - H(p) \leq L(\bar{x}^p, p+te_i) - L(\bar{x}^p, p) = J(\bar{x}^p) + \sum_j p_j \theta_j(\bar{x}^p) + t\theta_i(\bar{x}^p) - J(\bar{x}^p) - \sum_j p_j \theta_j(\bar{x}^p)$$

c'est à dire

$$H(p + te_i) - H(p) \leq t\theta_i(\bar{x}^p). \tag{4.34}$$

Nous avons aussi

$$\begin{aligned} H(p + te_i) - H(p) \geq L(\bar{x}^{p+te_i}, p + te_i) - L(\bar{x}^{p+te_i}, p) = J(\bar{x}^{p+te_i}) + \\ \sum_j p_j \theta_j(\bar{x}^{p+te_i}) + t\theta_i(\bar{x}^{p+te_i}) - J(\bar{x}^{p+te_i}) - \sum_j p_j \theta_j(\bar{x}^{p+te_i}) \end{aligned}$$



c'est à dire

$$H(p + te_i) - H(p) \geq t\theta_i(\bar{x}^{p+te_i}). \quad (4.35)$$

En divisant (4.34) et (4.35) par  $t \neq 0$  on trouve

$$\theta_i(\bar{x}^{p+te_i}) \leq \frac{H(p + te_i) - H(p)}{t} \leq \theta_i(\bar{x}^p), \quad \text{si } t > 0$$

et

$$\theta_i(\bar{x}^p) \leq \frac{H(p + te_i) - H(p)}{t} \leq \theta_i(\bar{x}^{p+te_i}), \quad \text{si } t < 0.$$

En passant à la limite  $t \mapsto 0$  et en utilisant le fait que  $\bar{x}^{p+te_i} \mapsto \bar{x}^p$  (conséquence de l'Etape 1) on déduit

$$\frac{H(p + te_i) - H(p)}{t} \rightarrow \theta_i(\bar{x}^p) \quad \text{si } t \mapsto 0.$$

En utilisant de nouveau le résultat de l'Etape 1 on obtient le résultat attendu.  $\square$

On peut alors remplacer l'expression de  $\frac{\partial H}{\partial p_i}$  dans (4.33) et on trouve

$$p^{(k+1)} = \max \left\{ p_i^{(k)} + \rho_k \theta_i(\bar{x}^{p^{(k)}}), 0 \right\} \quad i = 1, \dots, m$$

On va noter dans la suite  $x^{(k)} = \bar{x}^{p^{(k)}}$ . Alors **l'algorithme d'Uzawa** est le suivant :

**pas 1.** On pose  $k = 0$ , on choisit  $p^{(0)} \in \mathbb{R}_+^m$  (par exemple  $p^{(0)} = 0$ ) et on choisit  $k_{max} \in \mathbb{N}$  assez grand (par exemple  $k_{max} = 1000$ ) et  $\epsilon > 0$  assez petit (par exemple  $\epsilon = 10^{-6}$ ).

**pas 2.** Calculer  $x^{(k)} \in \mathbb{R}^n$  qui minimise sur  $\mathbb{R}^n$  la fonction  $x \in \mathbb{R}^n \rightarrow J(x) + \sum_{j=1}^m p_j^{(k)} \theta_j(x)$  (appliquer l'un des algorithmes de minimisation sans contrainte vus en Chapitre 3).

Si ( $k = 2$ ) alors

Si  $\|x^{(k)} - x^{(k-1)}\| \leq \epsilon$  alors  $x^{(k)}$  est le point de minimum recherché (l'algorithme a convergé)

Sinon, va au **pas 3**.

**pas 3.**

Si ( $k = k_{max}$ ) alors l'algorithme diverge.

Sinon, faire :

$$p^{(k+1)} = \max \left\{ p_i^{(k)} + \rho_k \theta_i(x^{(k)}), 0 \right\} \quad i = 1, \dots, m$$

ensuite  $k = k + 1$ , retour au **pas 2**.

Nous finissons par le résultat de convergence suivant :

**Théorème 4.25.** (convergence de l'algorithme d'Uzawa)

Soient  $a_1, a_2 \in \mathbb{R}$  tels que

$$0 < a_1 \leq a_2 < \frac{2\alpha}{M^2}.$$

Supposons que les facteurs  $\rho_k$  sont tels que

$$a_1 \leq \rho_k \leq a_2 \quad \forall k \in \mathbb{N}.$$

Alors la suite  $x^{(k)}$  générée par l'algorithme d'Uzawa converge vers  $x^*$  l'unique solution du problème de minimisation (4.21).

*Démonstration.* Il est clair que  $x^{(k)}$  et  $x^*$  existent et sont uniques, grâce aux hypothèses du début du paragraph 4.3.4. On rappelle aussi qu'il existe  $p^* \in \mathbb{R}_+^m$  tel que  $(x^*, p^*)$  soit un point selle de  $L$ .

**Étape 1.** Nous montrons les inégalités suivantes :

$$\langle \nabla J(x^{(k)}), x - x^{(k)} \rangle + \sum_{j=1}^m p_j^{(k)} [\theta_j(x) - \theta_j(x^{(k)})] \geq 0, \quad \forall x \in \mathbb{R}^n \quad (4.36)$$

et

$$\langle \nabla J(x^*), x - x^* \rangle + \sum_{j=1}^m p_j^* [\theta_j(x) - \theta_j(x^*)] \geq 0, \quad \forall x \in \mathbb{R}^n \quad (4.37)$$

*Preuve de (4.36) :* Dans l'inégalité

$$J(z) + \sum_{j=1}^m p_j^{(k)} \theta_j(z) \geq J(x^{(k)}) + \sum_{j=1}^m p_j^{(k)} \theta_j(x^{(k)}), \quad \forall z \in \mathbb{R}^n$$

prenons  $z = x^{(k)} + t(x - x^{(k)})$  avec  $x \in \mathbb{R}^n$  arbitraire et  $t \in [0, 1]$ . Grâce à la convexité des  $\theta_j$  nous avons

$$\theta_j(x^{(k)} + t(x - x^{(k)})) - \theta_j(x^{(k)}) \leq t[\theta_j(x) - \theta_j(x^{(k)})]$$

ce qui donne avec l'inégalité précédente :

$$J(x^{(k)} + t(x - x^{(k)})) - J(x^{(k)}) + t \sum_{j=1}^m p_j^{(k)} [\theta_j(x) - \theta_j(x^{(k)})] \geq 0, \quad \forall t \in ]0, 1].$$

En divisant par  $t$  et en passant à la limite  $t \mapsto 0$  on obtient (4.36).

La preuve de (4.37) est analogue.

**Étape 2.** En posant  $x = x^*$  en (4.36) et  $x = x^{(k)}$  en (4.37) et en faisant la somme des deux inégalités, on obtient :

$$\langle \nabla J(x^*) - \nabla J(x^{(k)}), x^{(k)} - x^* \rangle + \sum_{j=1}^m (p_j^* - p_j^{(k)}) [\theta_j(x^{(k)}) - \theta_j(x^*)] \geq 0$$

ce qui donne, en utilisant l'hypothèse d'ellipticité de  $J$  :

$$\langle p^{(k)} - p^*, \theta(x^{(k)}) - \theta(x^*) \rangle \leq -\alpha \|x^{(k)} - x^*\|^2. \quad (4.38)$$

D'autre part, nous avons par hypothèse

$$p^{(k+1)} = P_{\mathbb{R}_+^m} (p^{(k)} + \rho_k \theta(x^{(k)}).) \quad (4.39)$$

Comme  $p^*$  réalise le minimum sur  $\mathbb{R}_+^m$  de la fonction  $p \rightarrow -J(x^*) - \langle p^*, \theta(x^*) \rangle$  nous avons

$$p^* = P_{\mathbb{R}_+^m} (p^* + \rho_k \theta(x^*)) \quad (4.40)$$

En faisant la différence entre (4.39) et (4.40) et en utilisant la partie c) du Théorème de projection, on déduit

$$\|p^{(k+1)} - p^*\| \leq \|p^{(k)} - p^* + \rho_k [\theta(x^{(k)}) - \theta(x^*)]\|$$

c'est à dire,

$$\|p^{(k+1)} - p^*\|^2 \leq \|p^{(k)} - p^*\|^2 + \rho_k^2 \|\theta(x^{(k)}) - \theta(x^*)\|^2 + 2\rho_k \langle p^{(k)} - p^*, \theta(x^{(k)}) - \theta(x^*) \rangle$$

En utilisant le fait que  $\theta$  est lipschitzienne, on déduit

$$\|p^{(k+1)} - p^*\|^2 \leq \|p^{(k)} - p^*\|^2 + \rho_k^2 M^2 \|x^{(k)} - x^*\|^2 + 2\rho_k \langle p^{(k)} - p^*, \theta(x^{(k)}) - \theta(x^*) \rangle \quad (4.41)$$

On obtient ensuite facilement des inégalités (4.38) et (4.41) :

$$\|p^{(k+1)} - p^*\|^2 \leq \|p^{(k)} - p^*\|^2 + \rho_k^2 M^2 \|x^{(k)} - x^*\|^2 - 2\rho_k \alpha \|x^{(k)} - x^*\|^2$$

c'est à dire

$$\|p^{(k+1)} - p^*\|^2 \leq \|p^{(k)} - p^*\|^2 + (\rho_k^2 M^2 - 2\rho_k \alpha) \|x^{(k)} - x^*\|^2 \quad (4.42)$$

Introduisons la fonction  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  donnée par

$$\psi(\rho) = 2\alpha\rho - M^2\rho^2$$

On observe que  $\psi(\rho) > 0, \forall \rho \in ]0, \frac{2\alpha}{M^2}[$  et que

$$\psi(\rho) \geq \min\{\psi(a_1), \psi(a_2)\} \quad \forall \rho \in [a_1, a_2].$$

En choisissant alors  $\beta = \min\{\psi(a_1), \psi(a_2)\} > 0$  on déduit

$$2\alpha\rho_k - M^2\rho_k^2 \geq \beta \quad \forall k \in \mathbb{N}.$$

Ceci nous permet de déduire de (4.42) :

$$\|p^{(k+1)} - p^*\|^2 \leq \|p^{(k)} - p^*\|^2 - \beta \|x^{(k)} - x^*\|^2 \quad (4.43)$$

On obtient alors

$$\|p^{(k+1)} - p^*\|^2 \leq \|p^{(k)} - p^*\|^2$$

ce qui nous dit que la suite réelle  $\|p^{(k)} - p^*\|^2$  est décroissante. Comme elle est aussi positive, elle est convergente, et soit  $l \geq 0$  sa limite. On obtient de (4.43) :

$$\|x^{(k)} - x^*\|^2 \leq \frac{1}{\beta} [\|p^{(k)} - p^*\|^2 - \|p^{(k+1)} - p^*\|^2]$$

Comme le membre de droite de cette inégalité converge vers 0, on déduit que  $x^{(k)}$  converge vers  $x^*$ . Ceci finit la preuve.  $\square$

FIN

# Bibliographie

- [1] D. Azé et J.B. Hiriart-Urruty, *Analyse variationnelle et optimisation*, Cépadues, 2010.
- [2] M. Bergounioux, *Optimisation et contrôle des systèmes linéaires*, Dunod, Paris, 2001.
- [3] J. Cea, *Optimisation : théorie et algorithmes*, Dunod, 1971.
- [4] P. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson, Paris, 1982.
- [5] P. Ciarlet, B. Miara, J.M. Thomas, *Exercices l'analyse numérique matricielle et d'optimisation*, Masson, Paris, 1987.