Université Paris Diderot L3 MIASHS – Année 2015-2016

Cours d'Optimisation

Matthieu Bonnivard (d'après le polycopié d'Olivier Bokanowski)

Références bibliographiques :

— Philippe G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation.*
— Jean-Baptiste Hiriart-Urruty, *Optimisation et analyse convexe (exercices corrigés).*
— Grégoire Allaire, *Analyse numérique et optimisation*, chap. 9 et 10.

Chapitre 1

Introduction à l'optimisation

1.1 Généralités. Exemple introductif

L'optimisation consiste en la recherche du minimum (ou du maximum) d'une certaine quantité, appelée coût ou objectif. Dans ce cours, on supposera que le coût dépend de N variables réelles, rassemblées en un vecteur $x = (x_1, \ldots, x_N) \in \mathbb{R}^N$, et qui fournissent une valeur $J(x)$ où J est une fonction de $\mathbb{R}^N$ dans $\mathbb{R}$. En général, les variables $x_1, \ldots, x_N$ ne seront pas autorisées à prendre n'importe quelle valeur, mais devront satisfaire des contraintes que l'on représentera par un sous ensemble $K \subset \mathbb{R}^N$. On écrira les problèmes d'optimisation sous la forme générale suivante :

$$(\mathcal{P}) \qquad \inf_{x \in K} J(x).$$

On dit que problème $(\mathcal{P})$ admet une solution s'il existe un choix de variables $x_0 \in K$ tel que

$$\forall x \in K \quad J(x_0) \leq J(x).$$

On dit alors que x_0 est un minimiseur (ou point de minimum) de J sur K, et que $J(x_0)$ est un minimum de J sur K.

Exemple 1.1. Un étudiant doit réviser pour ses examens. Il a 4 matières à passer et dispose d'une semaine de révisions, ce qui représente 42 heures de travail (en comptant 6 jours et 7 heures par jour). Pour $i = 1, \ldots, 4$, on note x_i le nombre d'heures de révisions

pour la matière numéro i. L'ensemble K est alors décrit par

$$K = \left\{ x \in \mathbb{R}^4, \quad \forall 1 \leq i \leq 4 \quad x_i \geq 0, \quad \sum_{i=1}^{4} x_i \leq 42 \right\}$$

On note $M(x)$ la moyenne des notes (sur 20) obtenues par l'étudiant après avoir révisé x_i heures la matière numéro i. L'objectif est de maximiser $M(x)$, ce qui revient à minimiser la différence $20 - M(x)$. On peut donc formuler le problème d'optimisation suivant :

$$\inf_{x \in K} (20 - M(x))$$

Remarque 1.1. Bien sûr, dans l'exemple précédent, on ne connaît pas la formule de $M(x)$ de manière explicite ! De plus, il est évident que la moyenne obtenue ne dépend pas seulement du nombre d'heures de révisions, mais de beaucoup d'autres paramètres (assiduité en TD, concentration lors des révisions, qualité du sommeil la veille des épreuves...). Le choix de la fonction coût découle d'un choix de *modélisation* du phénomène étudié. Cependant, dans ce cours, les fonctions coût seront considérées comme des données du problème.

Nous allons voir que la résolution d'un problème d'optimisation dépend en grande partie des propriétés mathématiques de la fonction J. Pour l'illustrer, plaçons-nous en dimension $N = 1$.

1.2 Quelques exemples en dimension $N = 1$

On considère un seul paramètre $x \in \mathbb{R}$, et une fonction coût $J : \mathbb{R} \to \mathbb{R}$. On choisit $K = \mathbb{R}$ ou $K = [c, d]$ un intervalle fermé non vide.

Exemple 1.2. Cas d'une fonction J générale (continue), qui n'a pas de min ni de max sur $\mathbb{R}$ (par exemple, affine), mais un min et un max sur tout intervalle fermé borné.

Exemple 1.3. Cas d'une fonction discontinue, qui possède un inf sur un intervalle fermé borné, mais n'atteint pas cet inf.

Exemple 1.4. Cas d'une fonction J convexe, mais pas strictement convexe (son graphe contient un segment) : existence d'un minimum mais pas unicité.

Exemple 1.5. Cas d'une fonction strictement convexe, dérivable : le minimum sur $\mathbb{R}$ est atteint au point x_0 qui satisfait $J'(x_0) = 0$. On dit que x_0 est un point critique de J.

Bilan. Face au problème $(\mathcal{P})$,

— l'existence d'un minimum est liée à la *continuité* de J,

— l'unicité du minimiseur x_0 est liée à la *convexité* (stricte) de J,

— l'équation satisfaite par x_0 est associée à la *dérivée* de J.

Toutes ces propriétés joueront des rôles analogues en dimension N ; la dérivée sera remplacée par les *dérivées partielles* ou *dérivées directionnelles* de J.

1.3 Rappels de calcul différentiel

On se place dans $\mathbb{R}^N$, muni de la norme euclidienne $\|\cdot\|$ et du produit scalaire euclidien $\langle\cdot,\cdot\rangle$. Pour $x \in \mathbb{R}^N$ et $R > 0$, on note $B_R(x)$ la boule ouverte de centre x et de rayon R et $\overline{B}_R(x)$ la boule fermée correspondante :

$$B_R(x) = \big\{y \in \mathbb{R}^N, \quad \|y - x\| < R\big\}$$
$$\overline{B}_R(x) = \big\{y \in \mathbb{R}^N, \quad \|y - x\| \leq R\big\}$$

Définition 1.1.

1. Un ensemble $U \subset \mathbb{R}^N$ est OUVERT si

$$\forall x \in U \quad \exists R > 0, \quad B(x, R) \subset U$$

2. Un ensemble $F \subset \mathbb{R}^N$ est FERMÉ si son complémentaire $\mathbb{R}^N \setminus F$ est ouvert.

Définition 1.2. Soit $U \subset \mathbb{R}^N$ un ouvert et $J : U \to \mathbb{R}$ une application. Soit $u \in U$.

1. On dit que J est DIFFÉRENTIABLE au point u s'il existe une application linéaire $L \in \mathcal{L}(\mathbb{R}^N, \mathbb{R})$ t.q. pour tout $h \in \mathbb{R}^N$ t.q. $u + h \in U$,

$$J(u + h) = J(u) + L(h) + o(\|h\|).$$

   La notation $o(\|h\|)$ signifie qu'il existe une fonction $\varepsilon : \mathbb{R}^N \to \mathbb{R}$ t.q. $\lim_{h\to 0} \varepsilon(h) = 0$, et qui permette d'écrire le reste sous la forme $o(\|h\|) = \|h\|\, \varepsilon(h)$.

   Si L existe, elle est unique ; on la note $L = DJ(u)$.

2. Soit $d \in \mathbb{R}^N \setminus \{0\}$. On dit que J admet une DÉRIVÉE DIRECTIONNELLE dans la direction d, au point u, si l'application $t \in \mathbb{R} \mapsto J(u + td)$ est dérivable en 0. Si c'est le cas, on note cette dérivée

$$\frac{\partial J}{\partial d}(u) := \lim_{t\to 0} \frac{J(u + td) - J(u)}{t}.$$

Si $d = e_i$ est l'un des vecteurs de base de $\mathbb{R}^N$, on appelle cette dérivée directionnelle la i-ème DÉRIVÉE PARTIELLE de J au point u, que l'on note

$$\frac{\partial J}{\partial x_i}(u) := \lim_{t\to 0} \frac{J(u+te_i) - J(u)}{t} = \lim_{t\to 0} \frac{J(u_1,\dots,u_{i-1},u_i+t,u_{i+1},\dots,u_N) - J(u_1,\dots,u_N)}{t}.$$

Proposition 1.1. *Si $J : U \to \mathbb{R}$ est différentiable au point $u \in U$, alors elle admet des dérivées directionnelles en toute direction au point u (et en particulier, des dérivées partielles). De plus, sa différentielle au point u s'écrit*

$$\forall h = (h_1,\dots,h_N) \in \mathbb{R}^N \quad DJ(u)(h) = \sum_{i=1}^{N} \frac{\partial J}{\partial x_i}(u)\ h_i.$$

En introduisant le gradient de J au point u, défini par

$$\nabla J(u) = \left[\frac{\partial J}{\partial x_1}(u) \dots \frac{\partial J}{\partial x_N}(u)\right]^T,$$

on peut écrire de manière condensée

$$DJ(u)(h) = \langle \nabla J(u), h\rangle. \tag{1.1}$$

On montre que si J est différentiable au point u, alors $\nabla J(u)$ est l'unique vecteur de $\mathbb{R}^N$ t.q. la relation (1.1) soit vérifiée.

Remarque 1.2 (Calcul du gradient)**.** Pour calculer le gradient de J au point u, il n'est pas toujours nécessaire de calculer explicitement toutes les dérivées partielles. Une autre méthode consiste à établir un développement limité de J sous la forme suivante :

$$J(u+h) = J(u) + \langle w, h\rangle + o(\|h\|)$$

où $w \in \mathbb{R}^N$ est un certain vecteur fixé. Alors, on peut affirmer que J est différentiable en u, et que

$$w = \nabla J(u).$$

Exercice 1.1. Montrer que si J est différentiable en u, alors pour tout $d \in \mathbb{R}^N \setminus \{0\}$, sa dérivée directionnelle dans la direction d s'écrit

$$\frac{\partial J}{\partial d}(u) = \langle \nabla J(u), d\rangle.$$

Chapitre 2

Existence de minimum, convexité, unicité

Cadre général. On considère un ouvert $U \subset \mathbb{R}^N$ et une fonction $J : U \to \mathbb{R}$ (fonction coût). On se donne un ensemble fermé non vide $K \subset U$ et on s'intéresse au problème d'optimisation suivant :

$$(\mathcal{P}) \qquad \inf_{x \in K} J(x)$$

On notera $I = \inf_{x \in K} J(x)$ avec la convention suivante. Soit $A := \{J(x),\ x \in K\}$; A est une partie de $\mathbb{R}$, non vide car K est non vide. On distingue deux cas de figure :

(i) si A n'est pas minorée, on pose $I = -\infty$;

(ii) si A est minorée, elle possède une borne inférieure et on pose $I = \inf A \in \mathbb{R}$.

La première question que l'on se pose est de savoir si I est atteint, c'est-à-dire s'il existe $x_0 \in K$ t.q. $I = J(x_0)$.

2.1 Existence de minimum

Définition 2.1 (Minimum global, minimum local)**.** Soit $x_0 \in K$. On dit que la fonction J admet

(i) un MINIMUM GLOBAL SUR K au point x_0, si

$$\forall x \in K,\ J(x_0) \leq J(x);$$

(ii) un MINIMUM LOCAL sur K au point x_0, si

$$\exists R > 0,\ \forall x \in B_R(x_0) \cap K,\ J(x_0) \leq J(x).$$

Pour établir l'existence de minimiseurs, la première étape consiste à approcher la borne inférieure I à l'aide d'une suite de points $x_n \in K$, qu'on appelle suite minimisante.

Définition 2.2. On appelle SUITE MINIMISANTE pour le problème $(\mathcal{P})$ toute suite $(x_n)_{n\in\mathbb{N}}$ à valeurs dans $\mathbb{R}^N$ t.q.

$$\forall n \in \mathbb{N},\ x_n \in K \qquad \text{et} \qquad \lim_{n\to\infty} J(x_n) = I.$$

Proposition 2.1. *Pour tout problème $(\mathcal{P})$, il existe au moins une suite minimisante.*

Preuve. On introduit l'ensemble $A := \{J(x),\ x \in K\}$. Distinguons deux cas de figure :

(i) Si A n'est pas minorée, alors par convention, $I = -\infty$. De plus, pour tout $m \in \mathbb{R}$ il existe un point $x \in K$ tel que $J(x) < m$. Pour tout $n \in \mathbb{N}$, en prenant $m = -n$ on en déduit l'existence d'un point $x_n \in K$ tel que $J(x_n) < -n$. En passant à la limite dans l'inégalité on obtient $\lim_{n\to\infty} J(x_n) = -\infty$, donc (x_n) est une suite minimisante.

(ii) Si A est minorée, alors elle possède une borne inférieure I (comme une partie non vide et minorée de $\mathbb{R}$). Alors par définition, pour tout $\varepsilon > 0$, il existe $y_\varepsilon \in A$ t.q. $I \leq y_\varepsilon < I + \varepsilon$ et par définition de A, il existe un $x_\varepsilon \in K$ t.q. $y_\varepsilon = J(x_\varepsilon)$. Ainsi pour tout $\varepsilon > 0$, il existe $x_\varepsilon \in K$ t.q.

$$I \leq J(x_\varepsilon) < I + \varepsilon.$$

On l'applique en remplaçant ε par $\frac{1}{n}$: pour tout $n \in \mathbb{N}^*$, il existe $x_n \in K$ t.q.

$$I \leq J(x_n) < I + \frac{1}{n}.$$

En passant à la limite quand $n \to \infty$, on obtient que $\lim_{n\to\infty} J(x_n) = I$.

□

Pour conclure à l'existence d'un minimum global pour le problème $(\mathcal{P})$, on aimerait montrer que $J(x_n)$ converge vers $J(x)$, où x est un certain point de K ; on aurait alors $J(x) = I$. En général, on n'aura pas la convergence de la suite (x_n) vers x, mais seulement la convergence d'une suite extraite $(x_{\varphi(n)})$, qui découlera de propriétés de **compacité**. Le passage à la limite dans la suite numérique $(J(x_{\varphi(n)}))$ nécessitera quant à lui la **continuité** de la fonction J au point x.

Théorème 2.1. *Si K est un compact non vide de $\mathbb{R}^N$ (i.e., un fermé borné), et si J est continue en tout point de K, alors J admet un minimum sur K.*

Preuve. Soit $(x_n)_{n\in\mathbb{N}}$ une suite minimisante. (x_n) est à valeurs dans K, qui est compact, donc il existe $x \in K$ et une suite extraite $(x_{\varphi(n)})$ t.q. $\lim_{n\to\infty} x_{\varphi(n)} = x$. Comme J est continue en x,

$$\lim_{n\to\infty} J(x_{\varphi(n)}) = J(x).$$

La suite $(J(x_{\varphi(n)}))$ étant une suite extraite de $(J(x_n))$, elle converge vers la même limite, d'où $I = J(x)$.

□

Lorsque l'ensemble des contraintes K n'est pas compact, on pourra pallier cette difficulté en considérant des fonctions J qui contraignent les suites minimisantes à rester dans un ensemble compact de $\mathbb{R}^N$. On introduit pour cela la notion de fonction coercive.

Définition 2.3. On suppose U non borné. On dit que J est COERCIVE (ou encore, infinie à l'infini), si

$$\lim_{x\in U, \|x\|\to\infty} J(x) = +\infty.$$

Théorème 2.2. *Soit $K \subset \mathbb{R}^N$ t.q. (i) K est fermé, non vide, (ii) J est continue en tout point de K, et (iii) J est coercive. Alors J admet un minimum global sur K.*

Preuve.

L'infimum I de J sur K étant soit un réel, soit égal à $-\infty$, on peut choisir un réel M t.q. $M > I$. Par définition de la coercivité, il existe alors $R > 0$ t.q.

$$\forall x \in U, \quad \|x\| > R \Rightarrow J(x) \geq M > I. \tag{2.1}$$

Soit $(x_n)_{n\in\mathbb{N}}$ une suite minimisante. Par définition, $\lim_{n\to\infty} J(x_n) = I$; comme $M > I$, il existe donc un entier N t.q.

$$\forall n \in \mathbb{N}, \quad n \geq N \Rightarrow J(x_n) < M.$$

D'après (2.1), on en déduit que pour $n \geq N$, $\|x_n\| \leq R$. Ainsi, la suite minimisante $(x_n)_{n\geq N}$ est à valeurs dans $K \cap \overline{B_R(0)}$, qui est compact ; on peut donc conclure en reprenant le raisonnement de la preuve du théorème 2.1.

□

2.2 Convexité et unicité du minimiseur

Soit $U \subset \mathbb{R}^N$ un ouvert, $J : U \to \mathbb{R}$ une application et $K \subset U$ un ensemble non vide.

Définition 2.4. On dit que l'ensemble K est CONVEXE si

$$\forall \theta \in [0,1],\ \forall (x,y) \in K^2, \quad (1-\theta)x + \theta y \in K.$$

Cela signifie que si deux points x, y sont dans K, alors le segment $[x, y]$, qui relie ces points, est contenu dans K.

Exemple 2.1. Les sous-ensembles convexes de $\mathbb{R}$ sont les intervalles.

Exercice 2.1. Soit $\mathcal{N} : \mathbb{R}^N \to \mathbb{R}_+$ une norme quelconque. Montrer que la boule unité (fermée) pour cette norme est nécessairement convexe.

Définition 2.5. On suppose que K est convexe. On dit que
• l'application J est CONVEXE sur K si

$$\forall \theta \in (0,1),\ \forall (x,y) \in K^2, \quad J((1-\theta)x + \theta y) \leq (1-\theta)J(x) + \theta J(y)$$

• J est STRICTEMENT CONVEXE sur K si

$$\forall \theta \in (0,1),\ \forall (x,y) \in K^2, \quad x \neq y \ \Rightarrow\ J((1-\theta)x + \theta y) < (1-\theta)J(x) + \theta J(y)$$

• J est α-CONVEXE sur K (pour un $\alpha \geq 0$), si

$$\forall \theta \in (0,1),\ \forall (x,y) \in K^2, \quad J((1-\theta)x + \theta y) + \frac{\alpha}{2}\theta(1-\theta)\|x-y\|^2 \leq (1-\theta)J(x) + \theta J(y)$$

En particulier si J est α-convexe avec $\alpha > 0$, elle est strictement convexe.

Exemple 2.2. $x \in \mathbb{R}^N \mapsto \|x\|$ est convexe mais pas strictement convexe.

Proposition 2.2. *Si J est strictement convexe sur K, alors J admet au plus un minimiseur sur K.*

Preuve. Par l'absurde : supposons que J, strictement convexe sur K, possède deux minimiseurs distints, x et y, sur K. On note m la valeur commune du minimum :

$m = J(x) = J(y)$. Soit $z = \frac{1}{2}(x+y)$ le milieu du segment $[x, y]$. K étant convexe, $z \in K$ et par la stricte convexité de J,

$$J(z) = J(\frac{1}{2}x + \frac{1}{2}y) < \frac{1}{2}J(x) + \frac{1}{2}J(y) = m.$$

Cela contredit la définition du minimum de J sur K.

□

Critères de convexité pour des fonctions différentiables.

Proposition 2.3. *On suppose que J est différentiable en tout point de K. On a équivalence entre :*
(i) J convexe sur K.
(ii) $\forall (u, v) \in K^2$, $J(v) \geq J(u) + \langle \nabla J(u),\ v - u \rangle$.
(iii) $\forall (u, v) \in K^2$, $\langle \nabla J(v) - \nabla J(u),\ v - u \rangle \geq 0$.

Remarque 2.1. L'équation de l'hyperplan tangent au graphe de J, au point $(u, J(u)) \in U \times \mathbb{R}$, s'écrit

$$y = J(u) + \langle \nabla J(u),\ x - u \rangle, \quad \text{pour } x \in \mathbb{R}^N,\ y \in \mathbb{R}.$$

La relation (ii) signifie géométriquement que le graphe de J est au-dessus de son hyperplan tangent en tout point.

Preuve. $(i) \Rightarrow (ii)$: Notons $u_\theta := (1-\theta)u + \theta v = u + \theta(v-u)$. Pour $\theta \in]0, 1]$, on a $J(u_\theta) \leq (1-\theta)J(u) + \theta J(v) = J(u) + \theta(J(v) - J(u))$, donc

$$J(v) - J(u) \geq \frac{1}{\theta}(J(u_\theta) - J(u)) \overset{\theta \to 0^+}{\to} \langle \nabla J(u), v - u \rangle.$$

$(ii) \Rightarrow (i)$: On a

$$J(u) \geq J(u_\theta) + \langle \nabla J(u_\theta), u - u_\theta \rangle \tag{2.2}$$
$$J(v) \geq J(u_\theta) + \langle \nabla J(u_\theta), v - u_\theta \rangle. \tag{2.3}$$

En sommant $(1-\theta)$ fois la relation (2.2) et θ fois la relation (2.3), et en utilisant le fait que $(1-\theta)(u - u_\theta) + \theta(v - u_\theta) = 0$, on obtient l'inégalité de convexité.
$(ii) \Rightarrow (iii)$: On écrit

$$J(v) \geq J(u) + \langle \nabla J(u), v - u \rangle \tag{2.4}$$
$$J(u) \geq J(v) + \langle \nabla J(v), u - v \rangle. \tag{2.5}$$

En sommant on obtient l'inégalité désirée.
$(iii) \Rightarrow (ii)$: Soit $g(t) := J(u + t(v-u))$ pour $t \in [0,1]$. On remarque que $g'(t) = \langle \nabla J(u_t),\ v-u\rangle$, et en particulier que $g'(0) = \langle \nabla J(u),\ v-u\rangle$. Ainsi, par hypothèse,

$$g'(t) - g'(0) = \langle \nabla J(u_t) - \nabla J(u),\ v-u\rangle = \frac{1}{t}\langle \nabla J(u_t) - \nabla J(u),\ u_t - u\rangle \geq 0.$$

D'autre part, comme $g \in C([0,1]) \cap \Delta(]0,1[)$, d'après le théorème des accroissements finis, il existe $c \in (0,1)$ tel que $\frac{g(1)-g(0)}{1} = g'(c) \geq g'(0)$. Ainsi $g(1) \geq g(0) + g'(0)$, ce qui donne l'inégalité désirée. □

Il existe des critères analogues permettant de caractériser les fonctions différentiables strictement convexes. C'est l'objet de la proposition suivante, dont la preuve est laissée en exercice.

Proposition 2.4. *On suppose que J est différentiable en tout point de K. Alors les propriétés suivantes sont équivalentes :*
(i) J strictement convexe sur K.
(ii) $\forall (u,v) \in K^2$, $u \neq v \Rightarrow J(v) > J(u) + \langle \nabla J(u),\ v-u\rangle$.
(iii) $\forall (u,v) \in K^2$, $u \neq v \Rightarrow \langle \nabla J(v) - \nabla J(u),\ v-u\rangle > 0$.

Exercice 2.2. On suppose J différentiable en tout point de K. Soit $\alpha \geq 0$. Montrer les équivalences suivantes :

$$\begin{aligned} J \text{ est } \alpha\text{-convexe sur } K &\iff \forall (u,v) \in K^2,\ J(v) \geq J(u) + \langle \nabla J(u), v-u\rangle + \frac{\alpha}{2}\|v-u\|^2 \\ &\iff \forall (u,v) \in K^2,\ \langle \nabla J(v) - \nabla J(u), v-u\rangle \geq \alpha \|v-u\|^2 \end{aligned}$$

Il existe aussi quelques critères de convexité portant sur la différentielle seconde de J.

Théorème 2.3. *On suppose que $K = \mathbb{R}^N$. Soit $J : \mathbb{R}^N \to \mathbb{R}$, deux fois différentiable sur $\mathbb{R}^N$. Alors*

$$J \text{ est convexe sur } \mathbb{R}^N \iff \Big(\forall u \in \mathbb{R}^N,\ \forall h \in \mathbb{R}^N,\ \langle D^2 J(u)h, h\rangle \geq 0\Big).$$

Preuve. Voir le TD.

Attention, l'énoncé de l'implication $(\Rightarrow)$ doit être adapté si on veut étudier la convexité sur un sous-ensemble K de $\mathbb{R}^N$.

Exercice 2.3. On suppose J deux fois différentiable en tout point de $K \subset \mathbb{R}^N$. Soit $\alpha \geq 0$. Montrer que

$$\left(\forall u \in K,\ \forall h \in \mathbb{R}^N,\ \langle D^2 J(u)h, h\rangle \geq \alpha \|h\|^2\right) \quad \Longrightarrow \quad J \text{ est } \alpha\text{-convexe sur } K,$$

et que l'equivalence est vraie pour $K = \mathbb{R}^N$.

Exemple 2.3. $x \in \mathbb{R}^N \mapsto \|x\|^2$ est $\alpha = 2$ convexe. En effet, sa matrice hessienne en tout x est 2 fois la matrice identité.

Exercice 2.4. On suppose J deux fois différentiable en tout point de K. Montrer que

$$\left(\forall u \in K,\ \forall h \in \mathbb{R}^N,\ \langle D^2 J(u)h, h\rangle > 0\right) \quad \Longrightarrow \quad J \text{ est strictement convexe sur } K.$$

Montrer que la réciproque est fausse (exemple : $J(x) = x^4$).

Premières applications.

Proposition 2.5. *Soit K un convexe fermé non vide de $\mathbb{R}^N$, contenu dans un ouvert U. Soit $J : U \to \mathbb{R}$, différentiable en tout point de K. On suppose que J est α-convexe sur K, avec $\alpha > 0$. Alors J possède un unique minimiseur sur K.*

Idée de la preuve : Soit u fixé dans K. De J α-convexe on déduit que pour tout $v \in K$,

$$\begin{aligned} J(v) &\geq J(u) + \langle \nabla J(u), v-u\rangle + \frac{\alpha}{2}\|v-u\|^2 \\ &\geq J(u) - \|\nabla J(u)\|\|v-u\| + \frac{\alpha}{2}\|v-u\|^2 \overset{\|v\|\to\infty}{\to} +\infty. \end{aligned}$$

Donc J est coercive, et admet un minimum sur K d'après le théorème 2.2. L'unicité du minimiseur provient de la stricte convexité de J sur K. □

Théorème 2.4 (projection sur un convexe fermé non vide)**.** *Soit K un convexe fermé non vide de $\mathbb{R}^N$. Alors*

$$\forall u \in \mathbb{R}^N,\ \exists! \bar{u} \in K, \quad \|\bar{u} - u\| = \min_{v \in K} \|v - u\|$$

On notera $\bar{u} = \Pi_K(u)$ la projection de u sur K.

Preuve. $\bar{u}$, s'il existe, est caractérisé de manière équivalente par

$$\bar{u} \in K, \quad \text{et} \quad \|\bar{u} - u\|^2 = \inf_{v \in K} \|v - u\|^2.$$

Il s'agit donc de la minimisation de la fonction $J(v) = \|v - u\|^2$ sur K. Cette fonction étant α-convexe (avec $\alpha = 2$), on peut appliquer la proposition précédente. □

Chapitre 3

Conditions d'optimalité : généralités

On considère un ouvert $U \subset \mathbb{R}^N$, une application $J : U \to \mathbb{R}^N$ et un sous-ensemble $K \subset U$. On note u un minimiseur local de J sur K (s'il existe), et $\overset{o}{K}$ l'ensemble des points intérieurs à K, c'est-à-dire

$$\overset{o}{K} = \{x \in K,\ \exists R > 0,\ B_R(x) \subset K\}.$$

3.1 Généralités dans le cas $u \in \overset{o}{K}$

Théorème 3.1. *On suppose que J admet un minimum local en u, sur K. Si J est différentiable en u et $u \in \overset{o}{K}$, alors $\nabla J(u) = 0$.*

Preuve. J admet un minimum local en u, donc il existe $R > 0$ t.q.

$$\forall v \in K, \quad \|v - u\| \leq R \Rightarrow J(v) \geq J(u).$$

Comme $u \in \overset{o}{K}$, quitte à réduire le rayon R, on peut supposer que $B_R(u) \subset K$. Pour montrer que $\nabla J(u) = 0$, nous allons montrer que pour tout $h \in \mathbb{R}^N \backslash \{0\}$, $\langle \nabla J(u), h \rangle = 0$. Par linéarité, il suffit de le montrer pour des vecteurs h de norme 1.

Soit $h \in \mathbb{R}^N$ t.q. $\|h\| = 1$. Soit $r \in (0, R]$. Alors le vecteur $u + rh \in B_R(u)$, et donc $J(u + rh) \geq J(u)$, que l'on peut écrire

$$\frac{J(u + rh) - J(u)}{r} \geq 0.$$

J étant différentiable en u, $\lim_{r\to 0} \frac{1}{r}(J(u+rh)-J(u)) = \langle \nabla J(u), h\rangle$, d'où en passant à la limite dans l'inégalité précédente :

$$\langle \nabla J(u), h\rangle \geq 0.$$

Enfin, on peut remplacer h par $-h$ et reprendre la démarche précédente pour obtenir l'inégalité dans l'autre sens, ce qui donne l'égalité. □

Définition 3.1. Soit $A \in \mathbb{R}^{N\times N}$ une matrice SYMÉTRIQUE. On dit que :
• A est POSITIVE (ou "$A \geq 0$"), si

$$\forall x \in \mathbb{R}^N,\ \langle Ax, x\rangle \geq 0.$$

• A est DÉFINIE POSITIVE (ou "$A > 0$") si

$$\forall x \in \mathbb{R}^N, \quad x \neq 0 \Rightarrow \langle Ax, x\rangle > 0.$$

Proposition 3.1. *Soit $A \in \mathbb{R}^{N\times N}$ une matrice symétrique. Alors les propriétés suivants sont équivalentes :*

(i) A est définie positive.

(ii) $\exists \alpha > 0, \quad \forall x \in \mathbb{R}^N \quad \langle Ax, x\rangle \geq \alpha \|x\|^2$.

Preuve. $(ii) \Rightarrow (i)$ est immédiat. Pour montrer $(i) \Rightarrow (ii)$, on considère l'application

$$f : \mathbb{R}^N \to \mathbb{R}, \quad x \mapsto \langle Ax, x\rangle.$$

Comme $f(x) = \sum_{i,j} A_{ij} x_j x_i$, f est une fonction polynômiale en les coordonnéees $x_1, \ldots, x_N$ de x, elle est donc continue sur $\mathbb{R}^N$. Notons S la sphère unité : $S = \left\{y \in \mathbb{R}^N,\ \|y\| = 1\right\}$. S est compacte donc d'après le théorème 2.1, f admet un minimum global sur S : il existe donc $y_0 \in S$ t.q.

$$\forall y \in S, \quad \langle Ay, y\rangle \geq \langle Ay_0, y_0\rangle.$$

On pose alors $\alpha = \langle Ay_0, y_0\rangle$; $\alpha > 0$ car A est définie positive, et pour tout $x \in \mathbb{R}^N \setminus \{0\}$, en notant $y = x/\|x\|$ (qui appartient à S),

$$\begin{aligned}
\langle Ax, x\rangle &= \langle A(y\|x\|), y\|x\|\rangle \\
&= \langle \|x\|(Ay), \|x\|y\rangle \\
&= \|x\|^2 \langle Ay, y\rangle \\
&\geq \|x\|^2 \langle Ay_0, y_0\rangle = \alpha \|x\|^2.
\end{aligned}$$

(ii) étant également vérifiée en $x = 0$, cela conclut la preuve. □

Remarque 3.1. D'après l'exercice 2.3, si J est deux fois différentiable en tout point de K et si pour tout $u \in K$, $D^2J(u)$ est définie positive, alors J est α-convexe pour un $\alpha > 0$.

Théorème 3.2 (Formules de Taylor à l'ordre 2)**.** *Soit $u \in U$ et $J : U \to \mathbb{R}$ une application deux fois différentiable en u. On note $D^2J(u)$ la matrice hessienne de J en u, définie par*

$$D^2J(u) = \left(\frac{\partial^2 J}{\partial x_i \partial x_j}\right)_{1\leq i,j\leq N}.$$

Alors on peut écrire la FORMULE DE TAYLOR-YOUNG *à l'ordre 2 au point u : il existe une fonction $\eta : \mathbb{R}^N \to \mathbb{R}$ t.q. $\lim_{h\to 0} \eta(h) = 0$ et*

$$\forall h \in \mathbb{R}^N \quad u+h \in U \Rightarrow J(u+h) = J(u)+\langle \nabla J(u), h\rangle + \frac{1}{2}\langle D^2J(u)h, h\rangle + \|h\|^2\, \eta(h). \quad (3.1)$$

On peut également donner une formulation plus précise appelée FORMULE DE TAYLOR-LAGRANGE *:*

$$\forall h \in \mathbb{R}^N \quad u+h \in U \Rightarrow \exists \theta \in]0,1[, \quad J(u+h) = J(u)+\langle \nabla J(u), h\rangle + \frac{1}{2}\langle D^2J(u+\theta h)h, h\rangle. \quad (3.2)$$

Théorème 3.3. *On suppose que J admet un minimum local en u, sur K. Si J est 2 fois différentiable en u et $u \in \overset{o}{K}$, alors $\nabla J(u) = 0$ et pour tout $h \in \mathbb{R}^N$, $\langle D^2J(u)h, h\rangle \geq 0$.*

Preuve. On sait d'après le théorème 3.1 que $\nabla J(u) = 0$. Il suffit de montrer que la seconde propriété. D'après la formule de Taylor (3.1), on a pour tout $h \in \mathbb{R}^N$ t.q. $u + h \in U$,

$$J(u+h) - J(u) = \frac{1}{2}\langle D^2J(u)h, h\rangle + \|h\|^2\, \eta(h) \quad (3.3)$$

avec $\lim_{h\to 0} \eta(h) = 0$. Par l'absurde, supposons qu'il existe un vecteur $x \in \mathbb{R}^N$ t.q. $\langle D^2J(u)x, x\rangle < 0$. Alors x est non nul et on peut définir le vecteur normalisé $\overline{x} = x/\|x\|$, qui vérifie également $\langle D^2J(u)\overline{x}, \overline{x}\rangle < 0$. Puisque $u \in \overset{o}{K}$, il existe $R > 0$ t.q. $B_R(u) \subset K$. Ainsi, pour tout $0 < \rho < R$, $u + \rho\overline{x} \in K$ donc d'après (3.3),

$$J(u+\rho\overline{x}) - J(u) = \frac{1}{2}\langle D^2J(u)(\rho\overline{x}), \rho\overline{x}\rangle + \|\rho\overline{x}\|^2\, \eta(\rho\overline{x}),$$

qui s'écrit encore, par bilinéarité du produit scalaire et par homogénéité de la norme,

$$\begin{aligned}\frac{J(u+\rho\overline{x})-J(u)}{\rho^2} &= \frac{1}{2}\langle D^2J(u)\overline{x},\overline{x}\rangle + \|\overline{x}\|^2\ \eta(\rho\overline{x}) \\ &= \frac{1}{2}\langle D^2J(u)\overline{x},\overline{x}\rangle + \eta(\rho\overline{x}).\end{aligned}$$

Enfin, $\lim_{h\to 0}\eta(h)=0$ donc il existe $\overline{R}>0$ t.q.

$$\forall h\in\mathbb{R}^N,\quad \|h\|<\overline{R}\Rightarrow \eta(h)<-\frac{1}{2}\langle D^2J(u)\overline{x},\overline{x}\rangle.$$

On en conclut que pour tout $\rho\in]0,\min(R,\overline{R})[$,

$$\frac{J(u+\rho\overline{x})-J(u)}{\rho^2}<0.$$

Cela contredit le fait que J possède un minimum local en u. □

Remarque 3.2. On dira typiquement que $\nabla J(u)=0$ est une condition d'optimalité d'ordre 1, et que la condition $\langle D^2J(u)h,h\rangle\geq 0$ pour tout $h\in\mathbb{R}^N$, est une condition d'optimalité d'ordre 2. L'équation $\nabla J(u)=0$ est parfois appeléé "équation d'Euler".

Théorème 3.4 (Réciproque)**.** *On suppose J deux fois différentiable en u, $u\in\overset{o}{K}$ et $\nabla J(u)=0$. Alors :*
(i) S'il existe $\alpha>0$ tel que pour tout $h\in\mathbb{R}^N$, $\langle D^2J(u)h,h\rangle\geq\alpha\|h\|^2$, alors J admet un minimum local en u.
(ii) S'il existe une boule $B_R(u)$ centrée en u, contenue dans K, telle que pour tout $v\in B_R(u)$ et tout $h\in\mathbb{R}^N$, $\langle D^2J(v)h,h\rangle\geq 0$, alors J admet un minimum local en u.

Preuve. Cas *(i)***.** D'après la formule de Taylor-Young (3.1) (sachant que $\nabla J(u)=0$), il existe une fonction $\eta:\mathbb{R}^N\to\mathbb{R}$ t.q. $\lim_{h\to 0}\eta(h)=0$ et

$$\forall v\in U,\quad J(v)-J(u)\geq(\alpha+\eta(v-u))\,\|v-u\|^2. \tag{3.4}$$

Comme $\eta(h)\to 0$ quand $h\to 0$, il existe $R>0$ t.q.

$$\forall h\in\mathbb{R}^N,\quad \|h\|\leq R\Rightarrow\eta(h)\geq-\alpha.$$

D'après (3.4), on en déduit que pour tout $v\in U$ t.q. $\|v-u\|\leq R$, $J(v)\geq J(u)$.
Cas *(ii)***.** On applique la formule (3.2) sur la boule $B_R(u)$: soit $v\in B_R(u)$, il existe donc $\theta\in]0,1[$ t.q.

$$J(v)=J(u)+\frac{1}{2}\langle D^2J(u+\theta(v-u))(v-u),v-u\rangle.$$

Comme le point $u+\theta(v-u)$ appartient encore à la boule $B_R(u)$, l'hypothèse *(ii)* implique $\langle D^2J(u+\theta(v-u))(v-u), v-u\rangle \geq 0$, donc $J(v) \geq J(u)$.

3.2 Conditions d'optimalité dans le cas où K est convexe

Les résultats suivants sont fondamentaux dans ce cours.

Théorème 3.5 (CO1)**.** *Soit $J : U \to \mathbb{R}$ et $K \subset U$ un sous-ensemble convexe. On suppose que J admet un minimum local sur K, au point $u \in K$, et que J est différentiable en u. Alors*

$$\forall v \in K, \quad \langle \nabla J(u), v-u\rangle \geq 0.$$

C'est une condition d'optimalité d'ordre 1.

Preuve. Soit $v \in K \setminus \{u\}$. Pour $\theta \in]0,1[$, on note $u_\theta := (1-\theta)u + \theta v = u + \theta(v-u)$. J admet un minimum local en u sur K donc il existe $R > 0$ t.q.

$$\forall w \in K \cap B_R(u) \quad J(w) - J(u) \geq 0.$$

Par convexité de K, le point u_θ appartient à K quel que soit $\theta \in]0,1[$. De plus, si $\theta < R/\|v-u\|$, $u_\theta \in B_R(u)$, d'où :

$$\forall \theta \in]0, \min(1, \frac{R}{\|v-u\|})[\quad J(u_\theta) - J(u) \geq 0.$$

Le résultat s'en déduit par passage à la limite quand $\delta \to 0$, puisque

$$\langle \nabla J(u), v-u\rangle = \lim_{\theta \to 0^+} \frac{J(u_\theta) - J(u)}{\theta}.$$

□

Théorème 3.6. *Soit $J : U \to \mathbb{R}$, $K \subset U$ un sous-ensemble convexe et $u \in U$. On suppose J convexe sur K, et différentiable en u. Alors les conditions suivantes sont équivalentes :*

(i) J admet un minimum local sur K au point u ;

(ii)

$$\begin{cases} u \in K \\ \forall v \in K, \ \langle \nabla J(u), v-u\rangle \geq 0 \end{cases} \tag{3.5}$$

(iii) J admet un minimum global sur K au point u.
On dira que la condition d'optimalité d'ordre 1 (3.5) *caractérise la minimalité de u.*

Preuve. On a déjà vu que si $u \in K$ est un point de minimum local, alors on a (3.5). Réciproquement, si l'on a (3.5), alors par convexité de J, pour tout v dans K :

$$J(v) \geq J(u) + \langle \nabla J(u), v - u \rangle \geq J(u)$$

Donc J possède un minimum global en u sur K. □

Remarque 3.3. On peut montrer que pour une fonction J convexe, l'équivalence entre minimum local et minimum global reste valable sans hypothèse de différentiabilité de J (voir le TD).

En application directe nous avons le résultat suivant.

Théorème 3.7 (Projection sur un convexe fermé)**.** *Soit $u \in \mathbb{R}^N$, et $K \subset \mathbb{R}^N$ un convexe fermé non vide.*
(i) $\exists! \bar{u} = \Pi_K(u)$ dans K, t.q. $\|\bar{u} - u\| = \inf_{v \in K} \|v - u\|$.
(ii) $\bar{u}$ est caractérisé par

$$\langle u - \bar{u},\ v - \bar{u} \rangle \leq 0, \quad \forall v \in K. \tag{3.6}$$

(iii) De plus l'application $u \to \Pi_K(u)$ est 1-Lipschitzienne :

$$\|\Pi_K(u_2) - \Pi_K(u_1)\| \leq \|u_2 - u_1\|, \quad \forall u_1, u_2 \in \mathbb{R}^N.$$

Preuve. Voir le TD.

3.3 Le cône tangent $T_K(u)$

On se place maintenant dans un cadre plus général, où $K \subset \mathbb{R}^N$ est supposé seulement fermé et non vide. Pour généraliser la propriété (3.5) lorsque K n'est plus nécessairement convexe, on a besoin d'introduire la notion de cône tangent en un point $u \in K$.

On rappelle que $\Gamma \subset \mathbb{R}^N$ est un cône si

$$\forall \lambda \geq 0,\ \forall x \in \Gamma, \quad \lambda x \in \Gamma.$$

Définition 3.2. On appelle CÔNE TANGENT en u, et on note $T_K(u)$, l'ensemble déterminé par l'une des définitions équivalentes suivantes :

$$T_K(u) := \left\{ d \in \mathbb{R}^N \;\middle|\; \exists t_n > 0,\ t_n \overset{n\to\infty}{\to} 0,\ \exists u_n \in K,\ \lim_{n\to\infty} \frac{u_n - u}{t_n} = d \right\} \tag{3.7}$$

$$= \left\{ d \in \mathbb{R}^N \;\middle|\; \exists t_n > 0,\ t_n \overset{n\to\infty}{\to} 0,\ \exists d_n \in \mathbb{R}^N,\ u + t_n d_n \in K \text{ et } \lim_{n\to\infty} d_n = d \right\} \tag{3.8}$$

$$= \left\{ d \in \mathbb{R}^N \;\middle|\; \exists t_n > 0,\ t_n \overset{n\to\infty}{\to} 0,\ \exists u_n \in K,\ u_n = u + t_n d + o(t_n) \right\} \tag{3.9}$$

Une direction $d \in T_K(u)$ est appelée DIRECTION ADMISSIBLE.

Pour vérifier que les définitions (3.7)-(3.8)-(3.9) sont équivalentes, il suffit de poser $d_n = \frac{u_n - u}{t_n}$ et d'écrire $u + t_n d_n = u_n \in K$, ou encore $u_n = u + t_n d + t_n(d_n - d) = u + t_n d + o(t_n)$.

Interprétation. L'ensemble $T_K(u)$ contient en particulier les tangentes en u aux courbes issues du point u et contenues dans K. Si une telle courbe est paramétrée par une application régulière $\gamma : \mathbb{R}_+ \to \mathbb{R}^N$, à valeurs dans K et t.q. $\gamma(0) = u$, en notant $d = \gamma'(0)$, on obtient par développement limité en $t = 0$:

$$\gamma(t) = u + td + o(t) \quad \text{qd } t \to 0.$$

Étant donnée une suite (t_n) de réels t.q. $t_n > 0$ et $\lim_{n\to\infty} t_n = 0$, en définissant $u_n = \gamma(t_n)$, la suite (u_n) est à valeurs dans k et vérifie

$$u_n = u + t_n d + o(t_n) \quad \text{qd } n \to \infty.$$

Proposition 3.2. *$T_K(u)$ est un cône fermé, non vide.*

Preuve. $T_K(u)$ est non vide car il contient 0 (prendre $u_n = u$ et (t_n) une suite quelconque). Si $d \in T_K(u)$, on considère des suites $(t_n), (u_n)$ comme dans la définition (3.7). Pour tout $\lambda > 0$, il suffit alors de remplacer t_n par $t'_n = t_n/\lambda$ pour obtenir $\lim_{n\to\infty} \frac{u_n - u}{t'_n} = \lambda d$, d'où $\lambda d \in K$. $T_K(u)$ est donc un cône.
Montrons que $T_K(u)$ est fermé. Soit $(d_k)_{k\in\mathbb{N}}$ une suite de vecteurs de $T_K(u)$ convergeant vers un vecteur $d \in \mathbb{R}^N$. Montrons que $d \in T_K(u)$. Comme $d = \lim_{k\to\infty} d_k$ et que chaque d_k s'écrit également comme une limite de suite, on va utiliser un argument diagonal.

Pour tout $k \in \mathbb{N}$, on note $(t_{k,n})_{n\in\mathbb{N}}$ une suite de réels strictement positifs convergeant vers 0 et $(u_{k,n})_{n\in\mathbb{N}}$ une suite de points de K t.q.

$$\lim_{n\to\infty} \frac{u_{k,n} - u}{t_{k,n}} = d_k.$$

Le principe de l'argument diagonal est de construire à partir des valeurs $u_{k,n}, t_{k,n}$, dépendant de deux indices k, n, des suites $u_{k,n(k)}, t_{k,n(k)}$ dépendant uniquement de k et t.q.

$$\lim_{k\to\infty} t_{k,n(k)} = 0 \quad \text{et} \quad \lim_{k\to\infty} \frac{u_{k,n(k)} - u}{t_{k,n(k)}} = d. \tag{3.10}$$

Pour cela, on considère une suite $(\varepsilon_k)_{k\in\mathbb{N}}$ de réels strictement positifs et qui converge vers 0. Par définition des limites, pour tout $k \in \mathbb{N}$ fixé, il existe un entier $n(k)$ t.q.

$$0 < t_{k,n(k)} \leq \varepsilon_k \quad \text{et} \quad \left\| \frac{u_{k,n(k)} - u}{t_{k,n(k)}} - d_k \right\| \leq \varepsilon_k.$$

Par inégalité triangulaire, on obtient alors

$$\forall k \in \mathbb{N} \quad \left\| \frac{u_{k,n(k)} - u}{t_{k,n(k)}} - d \right\| \leq \left\| \frac{u_{k,n(k)} - u}{t_{k,n(k)}} - d_k \right\| + \|d - d_k\| \leq \varepsilon_k + \|d - d_k\|.$$

On en déduit (3.10) par passage à la limite quand $k \to \infty$ dans les inégalités précédentes. □

Exercice 3.1. Soit $u \in K$. Montrer que :
• si $u \in \overset{o}{K}$, alors $T_K(u) = \mathbb{R}^N$;
• si K est convexe, alors $T_K(u)$ contient $\big\{v - u,\ v \in K\big\}$.

Définition 3.3. On appelle CÔNE ENGENDRÉ par des vecteurs $a_1, \ldots, a_p$ de $\mathbb{R}^N$, l'ensemble noté $\Gamma(a_1, \ldots, a_p)$ et défini par :

$$\Gamma(a_1, \ldots, a_p) := \left\{ \sum_{i=1}^{p} \lambda_i a_i,\ \lambda_i \geq 0 \right\}.$$

Exercice 3.2. Montrer que $\Gamma(a_1, \ldots, a_p)$ est un cône convexe fermé non vide. (Indication : le caractère fermé pourra être démontré en raisonnant par récurrence sur p.)

Définition 3.4. Étant donné un ensemble $A \subset \mathbb{R}^N$, on appelle CÔNE POLAIRE de A (noté A^o) l'ensemble

$$A^o \ := \ \big\{v \in \mathbb{R}^N,\ \forall a \in A,\ \langle v, a\rangle \leq 0\big\}. \tag{3.11}$$

Théorème 3.8 (CO1)**.** *Soit u un point de minimum local de J sur K. On suppose J différentiable en u. Alors*

$$\forall d \in T_K(u), \quad \langle \nabla J(u), d \rangle \geq 0.$$

Au vu de la définition 3.4, cette propriété s'énonce également comme suit :

$$-\nabla J(u) \in T_K(u)^o. \tag{3.12}$$

Preuve. Soit u un point de minimum local de J sur K et $d \in T_K(u)$. En reprenant les notations des définitions (3.7)–(3.9), on écrit $u_n - u = t_n d_n$ avec $t_n \to 0$ et $d_n \to d$. Par définition du gradient de J en u, il existe une suite $\varepsilon_n \in \mathbb{R}^N$ t.q. $\varepsilon_n \to 0$ et

$$J(u_n) - J(u) = \langle \nabla J(u), u_n - u \rangle + \|u_n - u\|\varepsilon_n = t_n \langle \nabla J(u), d_n \rangle + t_n \|d_n\| \varepsilon_n.$$

Or $u_n \in K$ et u_n converge vers u, donc par définition du minimum local, il existe $N \in \mathbb{N}$ t.q.

$$\forall n \in \mathbb{N},\ n \geq N \ \Rightarrow\ J(u_n) - J(u) \geq 0,$$

d'où l'on déduit, après division par $t_n > 0$ dans l'égalité qui précède :

$$\forall n \in \mathbb{N},\ n \geq N \ \Rightarrow\ \langle \nabla J(u), d_n \rangle + \|d_n\| \varepsilon_n \geq 0.$$

Le résultat s'en déduit par passage à la limite car $\langle \nabla J(u), d_n \rangle \to \langle \nabla J(u), d \rangle$ et $\|d_n\| \varepsilon_n \to 0$. □

Théorème 3.9 (CO2)**.** *Soit u un point de minimum local de J sur K. On suppose J deux fois différentiable en u. Alors*

$$\forall d \in T_K(u),\ \begin{cases} \text{soit } \langle \nabla J(u), d \rangle > 0, \\ \text{soit } \langle \nabla J(u), d \rangle = 0, \ \text{ et } \langle D^2 J(u) d, d \rangle \geq 0 \end{cases}$$

Un vecteur d vérifiant $\langle \nabla J(u), d \rangle = 0$ est appelé "direction critique".

Preuve. Soit u un point de minimum local de J sur K et $u \in T_K(u)$. D'après la condition d'optimalité d'ordre 1, seul le cas $\langle \nabla J(u), d \rangle = 0$ est à considérer. On procède comme dans la preuve précédente en écrivant $u_n - u = t_n d_n$, avec $d_n \to d$, $t_n \to 0$.

D'après la formule de Taylor-Young à l'ordre 2, il existe une suite $\varepsilon_n \in \mathbb{R}^N$ t.q. $\varepsilon_n \to 0$ et

$$J(u)-J(u) = \frac{1}{2}\langle D^2J(u)(u_n-u), u_n-u\rangle + \|u_n-u\|^2\varepsilon_n = \frac{1}{2}t_n^2\langle D^2J(u)d_n, d_n\rangle + t_n^2\|d_n\|^2\varepsilon_n.$$

Par définition du minimum local, puisque $u_n \to u$, on a pour n assez grand : $J(u_n) \geq J(u)$. On en déduit après division par t_n^2 dans l'égalité précédente : pour n assez grand,

$$\frac{1}{2}\langle D^2J(u)d_n, d_n\rangle + \|d_n\|^2\varepsilon_n \geq 0.$$

Puisque $\langle D^2J(u)d_n, d_n\rangle \to \langle D^2J(u)d, d\rangle$ et $\|d_n\|^2\varepsilon_n \to 0$, on en déduit le résultat par passage à la limite. □

Chapitre 4

Algorithmes de minimisation sans contrainte

Dans tout ce chapitre, nous allons considérer une fonction $J : \mathbb{R}^N \to \mathbb{R}$, que l'on supposera α-convexe (pour un $\alpha > 0$) et différentiable pour garantir l'existence d'un unique $u \in \mathbb{R}^N$ t.q.

$$\forall v \in \mathbb{R}^N \quad J(u) \leq J(v)$$

(voir les résultats du chapitre 2). Rappelons que dans ce cas, le point u est caractérisé par l'équation d'Euler :

$$\nabla J(u) = 0. \tag{4.1}$$

Trouver u est donc équivalent à résoudre (4.1). Cependant, en général, il n'est pas possible de déterminer une formule explicite pour u à partir du système d'équations (4.1) (car ces équations peuvent être non linéaires par rapport aux coordonnées $u_1, \ldots, u_N$ du vecteur inconnu u). C'est pourquoi on est amené à chercher une valeur approchée de u. Pour construire cette approximation, nous allons utiliser des algorithmes itératifs, qui se présentent sous la forme d'algorithmes de descente.

4.1 Algorithmes itératifs et algorithmes de descente

Définition 4.1. Un ALGORITHME ITÉRATIF est défini par une application vectorielle $\mathbb{A} : \mathbb{R}^N \to \mathbb{R}^N$ qui génère une suite de vecteurs $(u^{(n)})_{n\in\mathbb{N}}$, à l'aide d'une construction de

la forme :

$$\text{Choisir } u^{(0)} \in \mathbb{R}^N \text{ (phase d'initialisation de l'algorithme)}$$
$$\text{Calculer } u^{(n+1)} = \mathbb{A}(u^{(n)}) \text{ (n-ième itération)}$$

Ce que l'on espère, c'est que la suite $(u^{(n)})_{n\in\mathbb{N}}$ converge vers le minimiseur u cherché ; on dit alors que l'algorithme converge vers la solution du problème de minimisation. Si un algorithme converge, on pourra mesurer son efficacité suivant deux critères :

— sa VITESSE DE CONVERGENCE, qui mesure la « rapidité » avec laquelle la suite $(u^{(n)})_{n\in\mathbb{N}}$ converge vers le point u ;

— sa COMPLEXITÉ CALCULATOIRE, qui mesure le coût des opérations nécessaires pour obtenir une itération. Le coût global est alors donné par le coût d'une itération multiplié par le nombre d'itérations nécessaires pour obtenir la solution escomptée avec une certaine précision ε fixée *a priori*.

La précision ε est associée à un CRITÈRE D'ARRÊT, permettant à l'algorithme de s'arrêter et de fournir une valeur approchée $u^{(n)}$ du minimiseur, que l'on jugera « acceptable ». Sachant que la solution exacte satisfait l'équation d'Euler (4.1), ce critère d'arrêt pourra prendre, par exemple, la forme suivante :

$$\|\nabla u^{(n)}\| \leq \varepsilon. \tag{4.2}$$

Ainsi, l'algorithme fournira comme résultat le premier vecteur $u^{(n)}$ obtenu, satisfaisant la condition (4.2).

Dans ce chapitre, nous nous intéressons plus particulièrement à la vitesse de convergence des algorithmes. Pour comparer ces vitesses de convergence, on introduit les définitions suivantes.

Définition 4.2. Supposons connue une suite $(u^{(n)})_{n\in\mathbb{N}}$, obtenue à l'aide d'un algorithme itératif, et telle que $\lim_{n\to\infty} u^{(n)} = u$. Pour tout $n \in \mathbb{N}$, on définit l'erreur $e^{(n)}$ à l'itération n par

$$e^{(n)} = \|u - u^{(n)}\|.$$

- On dira que la vitesse de convergence de l'algorithme est LINÉAIRE si

$$\exists C \in [0,1[,\ \forall u^{(0)} \in \mathbb{R}^N, \quad e^{(n+1)} \leq C e^{(n)}. \tag{4.3}$$

Cette propriété s'écrit de manière équivalente :

$$\exists C \in [0,1[,\ \forall u^{(0)} \in \mathbb{R}^N, \quad e^{(n)} \leq C^n\ e^{(0)}.$$

Pour cette raison, on dira également que si une méthode satisfait (4.3), la convergence est GÉOMÉTRIQUE (l'erreur se comporte comme une suite géométrique de raison inférieure strictement à 1).

- La méthode sera dite D'ORDRE p si elle satisfait une relation du type

$$\exists C \in [0,1[,\ \forall u^{(0)} \in \mathbb{R}^N, \quad e^{(n+1)} \leq C(e^{(n)})^p. \tag{4.4}$$

Si $p = 2$, on dira que la vitesse de convergence est QUADRATIQUE.

Algorithmes de descente. Les algorithmes que nous allons considérer pour les problèmes d'optimisation ont la forme générale suivante :

$$u^{(0)} \text{ étant donné, calculer } u^{(n+1)} = u^{(n)} + \rho^{(n)} d^{(n)}. \tag{4.5}$$

Le vecteur $d^{(n)}$ s'appelle la direction de descente, et le réel $\rho^{(n)} > 0$ le pas de la méthode à la n-ième itération. On pratique, on choisira la direction et le pas afin que l'inégalité suivante soit satisfaite :

$$J(u^{(n+1)}) \leq J(u^{(n)}).$$

De tels algorithmes sont appelés ALGORITHMES DE DESCENTE.

4.2 Algorithmes de gradient

Supposons que l'on cherche à définir un algorithme de descente suivant le procédé (4.5). Partant d'une valeur $u^{(n)}$, écrivons la formule de Taylor à l'ordre 1 pour J au point $u^{(n)}$:

$$J(u^{(n+1)}) = J(u^{(n)} + \rho^{(n)} d^{(n)}) = J(u^{(n)}) + \langle \nabla J(u^{(n)}), \rho^{(n)} d^{(n)} \rangle + \rho^{(n)} \|d^{(n)}\|\ \eta(\rho^{(n)} d^{(n)}),$$

où $\eta : \mathbb{R}^N \to \mathbb{R}$ est une fonction vérifiant $\lim_{h \to 0} \eta(h) = 0$. Par linéarité du produit scalaire, on peut donc écrire :

$$J(u^{(n+1)}) - J(u^{(n)}) = \rho^{(n)} \left[\langle \nabla J(u^{(n)}), d^{(n)} \rangle + \|d^{(n)}\|\ \eta(\rho^{(n)} d^{(n)}) \right].$$

Étant donné que η tend vers 0 lorsque son argument tend vers 0, on peut supposer que, pour $\rho^{(n)}$ suffisamment petit, le signe du second membre va être le même que le signe de $\langle \nabla J(u^{(n)}), d^{(n)} \rangle$ (rappelons que $\rho^{(n)} > 0$). Pour s'assurer que $J(u^{(n+1)}) - J(u^{(n)}) \leq 0$, un choix possible pour $d^{(n)}$ est donc

$$d^{(n)} = -\nabla J(u^{(n)}). \tag{4.6}$$

On obtient alors :

$$\begin{aligned} J(u^{(n+1)}) - J(u^{(n)}) &= \rho^{(n)} \left[-\|\nabla J(u^{(n)})\|^2 + \|\nabla J(u^{(n)})\| \, \eta(-\rho^{(n)} \nabla J(u^{(n)})) \right] \\ &= -\rho^{(n)} \|\nabla J(u^{(n)})\| \left[\|\nabla J(u^{(n)})\| + \eta(-\rho^{(n)} \nabla J(u^{(n)})) \right]. \end{aligned}$$

Cette quantité sera négative si l'on choisit un pas $\rho^{(n)}$ suffisamment petit. En effet, on peut supposer $\|\nabla J(u^{(n)})\| > 0$ (sinon $u^{(n)} = u$ et l'algorithme s'arrêterait), et alors il existe un $\rho_{max} > 0$ t.q. pour tout choix de $\rho^{(n)}$ t.q. $0 < \rho^{(n)} < \rho_{max}$, $\eta(-\rho^{(n)} \nabla J(u^{(n)})) > -\|\nabla J(u^{(n)})\|$, ce qui donne le résultat.

Les algorithmes de descente utilisant la direction (4.6) à chaque itération s'appellent des ALGORITHMES DE GRADIENT. Dans le raisonnement précédent, la borne supérieure ρ_{max} sur le pas dépend *a priori* de l'itération n, puisqu'elle dépend de la fonction η (qui dépend elle-même de $u^{(n)}$) et de la norme de $\nabla J(u^{(n)})$. Sous une hypothèse supplémentaire sur le gradient de J, on peut en fait établir des bornes uniformes sur $\rho^{(n)}$ permettant de garantir la convergence des algorithmes de gradient ; c'est l'objet du théorème suivant.

Théorème 4.1. *Soit $J : \mathbb{R}^N \to \mathbb{R}$ une application différentiable, α-convexe pour un $\alpha > 0$. On suppose que $\nabla J : \mathbb{R}^N \to \mathbb{R}^N$ est une application M-lipschitzienne, pour une constante $M > 0$, c'est-à-dire :*

$$\forall u, v \in \mathbb{R}^N \quad \|\nabla J(u) - \nabla J(v)\| \leq M \|u - v\|. \tag{4.7}$$

On considère deux réels a, b t.q.

$$0 < a < b < \frac{2\alpha}{M^2},$$

et l'on se donne une suite de pas $\rho^{(n)}$ t.q.

$$\forall n \in \mathbb{N} \quad \rho^{(n)} \in [a, b].$$

Alors, pour toute valeur initiale $u^{(0)} \in \mathbb{R}^N$, la méthode de gradient définie par l'itération

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

converge ; de plus, la convergence est géométrique : il existe une constante $0 < C < 1$, dépendant uniquement de α, M, a et b, t.q.

$$\forall n \in \mathbb{N}, \quad \|u^{(n)} - u\| \leq C^n \|u^{(0)} - u\|$$

Preuve. En utilisant la condition (4.1), on peut écrire

$$\begin{aligned} u^{(n+1)} - u &= (u^{(n)} - u) - \rho^{(n)} \nabla J(u^{(n)}) \\ &= (u^{(n)} - u) - \rho^{(n)} \left[\nabla J(u^{(n)}) - \nabla J(u) \right]. \end{aligned}$$

Puisque pour tout vecteur $v \in \mathbb{R}^N$, $\|v\|^2 = \langle v, v \rangle$, on obtient en développant les produits scalaires :

$$\|u^{(n+1)} - u\|^2 = \|u^{(n)} - u\|^2 - 2\rho^{(n)} \langle \nabla J(u^{(n)}) - \nabla J(u), u^{(n)} - u \rangle + (\rho^{(n)})^2 \|\nabla J(u^{(n)}) - \nabla J(u)\|^2$$

J étant α-convexe et différentiable, on a, d'après l'exercice 2.2,

$$\langle \nabla J(u^{(n)}) - \nabla J(u), u^{(n)} - u \rangle \geq \alpha \|u^{(n)} - u\|^2,$$

et d'après la condition de Lipschitz sur ∇J,

$$|\nabla J(u^{(n)}) - \nabla J(u)\|^2 \leq M^2 \|u^{(n)} - u\|^2.$$

Puisque $\rho^{(n)} > 0$, on en déduit l'estimation suivante :

$$\|u^{(n+1)} - u\|^2 \leq \left(1 - 2\alpha\rho^{(n)} + M^2 (\rho^{(n)})^2 \right) \|u^{(n)} - u\|^2.$$

On note $\tau : \mathbb{R} \to \mathbb{R}$ la fonction trinôme définie par $\tau(\rho) = 1 - 2\alpha\rho + M^2\rho^2$. Remarquons tout d'abord que pour tout $\rho > 0$, $\tau(\rho) \geq 0$. En effet, son discriminant est égal à $4(\alpha^2 - M^2)$, il est donc négatif puisque d'après les hypothèses faites sur J, on a nécessairement $\alpha \leq M$. Pour s'en convaincre, on remarque que, en appliquant l'exercice 2.2 et l'inégalité de Cauchy-Schwarz,

$$\begin{aligned} \forall u, v \in \mathbb{R}^N \quad \alpha \|u - v\|^2 &\leq \langle \nabla J(u) - \nabla J(v), u - v \rangle \\ &\leq \|\nabla J(u) - \nabla J(v)\| \, \|u - v\|, \end{aligned}$$

ce qui implique, d'après la condition (4.7) :

$$\forall u, v \in \mathbb{R}^N \quad \alpha\|u - v\| \leq \|\nabla J(u) - \nabla J(v)\| \leq M\|u - v\|.$$

Le minimum de τ est atteint au point $\rho_{min} = \frac{\alpha}{M^2}$; de plus, $\tau(0) = 1$ et par symétrie, $\tau(\frac{2\alpha}{M^2}) = 1$. Ainsi, si l'on fixe $0 < a < b < \frac{2\alpha}{M^2})$, on obtient pour tout $\rho \in [a, b]$,

$$\tau(\rho) \leq \max(\tau(a), \tau(b)) < 1.$$

En notant $C = [\max(\tau(a), \tau(b))]^{1/2}$, on vérifie que pour toute suite $(\rho^{(n)})_{n\in\mathbb{N}}$ à valeurs dans $[a, b]$,

$$\forall n \in \mathbb{N} \qquad \|u^{(n+1)} - u\| \leq C\|u^{(n)} - u\|.$$

□

Corollaire 4.1 (convergence de l'algorithme de gradient à pas fixe)**.** *Soit $J : \mathbb{R}^N \to \mathbb{R}$ une application différentiable, α-convexe pour un $\alpha > 0$. On suppose que $\nabla J : \mathbb{R}^N \to \mathbb{R}^N$ est M-lipschitzienne. On fixe un pas constant $\rho \in]0, \frac{2\alpha}{M^2}[$. Alors, pour toute valeur initiale $u^{(0)} \in \mathbb{R}^N$, la méthode de gradient à pas fixe, définie par l'itération*

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

converge ; de plus, la convergence est géométrique.

Une autre stratégie consiste à déterminer, s'il existe, un pas optimal à chaque itération. En notant $d^{(n)} = -\nabla J(u^{(n)})$ la direction de descente à l'itération n, cela revient à déterminer un point sur la droite passant par $u^{(n)}$ et dirigée par $d^{(n)}$, qui minimise la valeur de J sur cette droite. Autrement dit, on cherche à chaque itération n un pas $\rho^{(n)} \in \mathbb{R}$ t.q.

$$J(u^{(n)} + \rho^{(n)}d^{(n)}) = \min_{\rho\in\mathbb{R}} J(u^{(n)} + \rho d^{(n)}). \tag{4.8}$$

L'algorithme de gradient à pas optimal consiste alors, à partir d'une valeur initiale $u^{(0)}$, à réaliser l'itération

$$u^{(n+1)} = u^{(n)} + \rho^{(n)}d^{(n)}, \quad \text{avec } d^{(n)} := -\nabla J(u^{(n)}),$$

et où $\rho^{(n)}$ est (si possible) défini par (4.8).

Théorème 4.2 (convergence de l'algorithme de gradient à pas optimal)**.** *Soit $J : \mathbb{R}^N \to \mathbb{R}$, α-convexe, différentiable, t.q. ∇J soit M-lipschitzien. Pour tout point de départ $u^0 \in \mathbb{R}^N$, l'algorithme de gradient à pas optimal est bien défini, et converge vers l'unique minimiseur u :*

$$\lim_{n\to\infty} u^{(n)} = u.$$

Preuve. Notons que le minimiseur u est bien défini et caractérisé par $\nabla J(u) = 0$ (cas J est différentiable et α-convexe). On peut supposer que pour tout $n \in \mathbb{N}$, $d^{(n)} \neq 0$, sinon l'algorithme converge en un nombre fini d'itérations. Pour tout $n \in \mathbb{N}$, on définit l'application $g_n : \mathbb{R} \to \mathbb{R}$, par

$$\forall \rho \in \mathbb{R}, \quad g_n(\rho) := J(u^{(n)} + \rho d^{(n)}).$$

Alors g_n possède un unique minimiseur $\rho^{(n)}$ sur $\mathbb{R}$. En effet,

- g_n est continue sur $\mathbb{R}$ comme composée de la fonction affine $\rho \in \mathbb{R} \mapsto u^{(n)} + \rho d^{(n)} \in \mathbb{R}^N$ par la fonction continue J ;
- g_n est coercive. Pour le voir, il suffit de remarquer que, puisque $d^{(n)} \neq 0$,

  $$\lim_{|\rho|\to\infty} \|u^{(n)} + \rho d^{(n)}\| = +\infty,$$

  et de combiner cette propriété avec la coercivité de J ;
- g_n est strictement convexe sur $\mathbb{R}$. En effet, soit ρ_1, ρ_2 deux réels distincts et $\theta \in]0, 1[$. Alors, en utilisant la stricte convexité de J,

$$\begin{aligned} g_n((1-\theta)\rho_1 + \theta\rho_2) &= J(u^{(n)} + [(1-\theta)\rho_1 + \theta\rho_2]d^{(n)}) \\ &= J((1-\theta)[u^{(n)} + \rho_1 d^{(n)}] + \theta[u^{(n)} + \rho_2 d^{(n)}]) \\ &< (1-\theta)J(u^{(n)} + \rho_1 d^{(n)}) + \theta J(u^{(n)} + \rho_2 d^{(n)}) = (1-\theta)g_n(\rho_1) + \theta g_n(\rho_2). \end{aligned}$$

L'unicité du minimiseur $\rho^{(n)}$ découle alors du théorème 2.2 et de la proposition 2.2. De plus, d'après les théorèmes 3.1 et 3.6, $\rho^{(n)}$ est caractérisé par la propriété :

$$g_n'(\rho^{(n)}) = 0.$$

Pour calculer g_n', on fixe $\rho \in \mathbb{R}$ et $h \in \mathbb{R} \setminus \{0\}$, et on écrit le quotient différentiel

$$\frac{g_n(\rho+h) - g_n(\rho)}{h} = \frac{J(u^{(n)} + (\rho+h)d^{(n)}) - J(u^{(n)} + \rho d^{(n)})}{h} = \frac{J((u^{(n)} + \rho d^{(n)}) + hd^{(n)}) - J(u^{(n)} + \rho d^{(n)})}{h},$$

d'où en passant à la limite quand $h \to 0$:

$$g_n'(\rho) = \langle \nabla J(u^{(n)} + \rho d^{(n)}), d^{(n)} \rangle.$$

En appliquant cette formule avec $\rho = \rho^{(n)}$, puisque $d^{(n+1)} = \nabla J(u^{(n)} + \rho^{(n)} d^{(n)})$, on en déduit la propriété suivante :

$$\langle d^{(n+1)}, d^{(n)} \rangle = 0. \tag{4.9}$$

Ainsi, dans l'algorithme de gradient à pas optimal, deux directions de descente successives sont orthogonales.

Notons que $J(u^{n+1}) \leq J(u^{(n)})$, par définition. Donc la suite $(J(u^{(n)}))_{n\in\mathbb{N}}$ est décroissante minorée (J est minorée par $J(u)$), donc convergente vers une limite notée ℓ. D'autre part, par α-convexité de J on a

$$J(u^{(n)}) \geq J(u^{(n+1)}) + \langle \nabla J(u^{(n+1)}), u^{(n)} - u^{(n+1)} \rangle + \frac{\alpha}{2} \|u^{(n)} - u^{(n+1)}\|^2$$

De plus on remarque que $\langle \nabla J(u^{(n+1)}), u^{(n)} - u^{(n+1)} \rangle = -\rho_n \langle d^{n+1}, d^n \rangle = 0$, et donc

$$\frac{\alpha}{2} \|u^{(n)} - u^{(n+1)}\|^2 \leq J(u^{(n)}) - J(u^{(n+1)}) \overset{n\to\infty}{\to} \ell - \ell = 0.$$

On a donc déja montré que $\|u^{(n)} - u^{(n+1)}\| \to 0$.

Par ailleurs, comme $\langle \nabla J(u^{(n)}), \nabla J(u^{(n+1)}) \rangle = 0$, en développant la norme au carré et en utilisant la condition de Lipschitz sur ∇J,

$$M^2 \|u^{(n)} - u^{(n+1)}\|^2 \geq \|\nabla J(u^{(n)}) - \nabla J(u^{(n+1)})\|^2 = \|\nabla J(u^{(n)})\|^2 + \|\nabla J(u^{(n+1)})\|^2,$$

et donc

$$\|\nabla J(u^{(n)})\| \leq M \|u^{(n)} - u^{(n+1)}\|.$$

Enfin, on l'α-convexité de J permet également d'écrire :

$$\begin{aligned} \alpha \|u^{(n)} - u\|^2 &\leq \langle u^{(n)} - u, \nabla J(u^{(n)}) - \nabla J(u) \rangle = \langle u^{(n)} - u, \nabla J(u^{(n)}) \rangle \\ &\leq \|u^{(n)} - u\| \|\nabla J(u^{(n)})\| \end{aligned}$$

On en déduit finalement

$$\|u^{(n)} - u\| \leq \frac{1}{\alpha} \|\nabla J(u^{(n)})\| \leq \frac{M}{\alpha} \|u^{(n)} - u^{(n+1)}\| \overset{n\to\infty}{\to} 0.$$

□

4.3 Cas particulier : fonctionnelles quadratiques

Définition 4.3. On appelle FONCTIONNELLE QUADRATIQUE une application $J : \mathbb{R}^N \to \mathbb{R}$ de la forme

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

où $A \in \mathbb{R}^{N\times N}$ est une matrice symétrique et $b \in \mathbb{R}^N$.

Proposition 4.1. *Soit $A \in \mathbb{R}^{N\times N}$ une matrice symétrique, $b \in \mathbb{R}^N$ et $J : \mathbb{R}^N \to \mathbb{R}$ la fonctionnelle quadratique associée. Alors J est indéfiniment dérivable et on a les formules suivantes pour son gradient et sa matrice hessienne en tout point :*

$$\forall x \in \mathbb{R}^N, \qquad \nabla J(x) = Ax - b, \quad D^2J(x) = A.$$

Corollaire 4.2. *Soit $A \in \mathbb{R}^{N\times N}$ une matrice symétrique, $b \in \mathbb{R}^N$ et $J : \mathbb{R}^N \to \mathbb{R}$ la fonctionnelle quadratique associée. On suppose de plus que A est* DÉFINIE POSITIVE. *On note λ_1 (resp. λ_N) la plus petite (resp., la plus grande) valeur propre de A. Alors J est α-convexe pour la constante $\alpha = \lambda_1$ et ∇J est $M-$lipschitzien pour la constante $M = \lambda_N$.*

Preuve. Voir le TD.

Remarque 4.1. Une fonctionnelle quadratique associée à une matrice A symétrique définie positive est parfois appelée FONCTIONNELLE QUADRATIQUE ELLIPTIQUE.

Remarque 4.2. En vertu du corollaire 4.2, on peut donc appliquer les algorithmes de gradient à pas fixe ou à pas optimal pour la minimisation d'une fonctionnelle quadratique elliptique. Cependant, en utilisant la linéarité du gradient, on peut montrer que, dans ce cas, l'algorithme de gradient à pas fixe converge pour une plage plus large de choix possibles de ρ que celle obtenue au théorème 4.1 (voir le TD).

4.3.1 Calcul du pas optimal pour l'algorithme de gradient

Soit $A \in \mathbb{R}^{N\times N}$ une matrice symétrique définie positive, $b \in \mathbb{R}^N$ et $J : \mathbb{R}^N \to \mathbb{R}$ définie par $J(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$ pour tout $x \in \mathbb{R}^N$. On se donne une valeur initiale $u^{(0)} \in \mathbb{R}^N$ et on applique l'algorithme de gradient à pas optimal à la minimisation de la fonctionnelle J sur $\mathbb{R}^N$. Pour chaque $n \in \mathbb{N}$, on réalise donc l'itération

$$u^{(n+1)} = u^{(n)} + \rho^{(n)}d^{(n)}, \quad \text{avec } d^{(n)} := -(Au^{(n)} - b),$$

où $\rho^{(n)}$ est défini par (4.8). Montrons que $\rho^{(n)}$ peut, dans ce cas, se calculer par une formule explicite. En effet, on a vu dans la preuve du théorème 4.2, que deux directions de descente successives étaient orthogonales (relation (4.9)), ce qui s'écrit également

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

En utilisant l'expression de $u^{(n+1)}$ et la formule du gradient de J, on obtient donc

$$\begin{aligned}
0 &= \left\langle A\left(u^{(n)} - \rho^{(n)}(Au^{(n)} - b)\right) - b, Au^{(n)} - b \right\rangle \\
&= \left\langle Au^{(n)} - b - \rho^{(n)} A\left(Au^{(n)} - b\right), Au^{(n)} - b \right\rangle \\
&= \|Au^{(n)} - b\|^2 - \rho^{(n)} \left\langle A(Au^{(n)} - b), Au^{(n)} - b \right\rangle \\
&= \|d^{(n)}\|^2 - \rho^{(n)} \langle Ad^{(n)}, d^{(n)} \rangle.
\end{aligned}$$

On peut supposer $d^{(n)} \neq 0$ (sinon cela signifie que $\nabla J(u^{(n)}) = 0$, c'est-à-dire que $u^{(n)} = u$ et il est inutile de calculer le pas $\rho^{(n)}$ suivant), et donc $\langle Ad^{(n)}, d^{(n)} \rangle \neq 0$ puisque A est définie positive. Ainsi, $\rho^{(n)}$ s'exprime par la formule

$$\rho^{(n)} = \frac{\|d^{(n)}\|^2}{\langle Ad^{(n)}, d^{(n)} \rangle}. \tag{4.10}$$

4.3.2 Choix de la direction de descente $d^{(n)}$. La notion de direction conjuguée

Nous allons voir sur un exemple que le choix de la direction de descente $d^{(n)} = -\nabla J(u^{(n)})$ n'est pas forcément le plus efficace. On se place en dimension $N = 2$ et on définit la fonctionnelle $J : (x, y) \in \mathbb{R}^2 \mapsto \frac{1}{2}(x^2 + 2y^2)$, associée à la matrice $A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$. Le gradient de J est défini en tout point $(x, y) \in \mathbb{R}^2$ par

$$\nabla J(x, y) = \begin{pmatrix} x \\ 2y \end{pmatrix} = A \begin{pmatrix} x \\ y \end{pmatrix}.$$

Le minimum de J sur $\mathbb{R}^2$ est clairement atteint au point $(0, 0)$. Examinons les premières étapes de l'algorithme de gradient à pas optimal. On se donne un premier point $u^{(0)} = (x^{(0)}, y^{(0)})^T$ et on définit la direction de descente

$$d^{(0)} = -\nabla J(u^{(0)}) = -Au^{(0)} = \begin{pmatrix} -x^{(0)} \\ -2y^{(0)} \end{pmatrix}.$$

Le point suivant $u^{(1)}$ est sur la droite passant par $u^{(0)}$ et dirigée par $d^{(0)}$, c'est-à-dire l'ensemble des points de la forme $(x^{(0)} - \rho x^{(0)}, y^{(0)} - 2\rho y^{(0)})$ pour un $\rho \in \mathbb{R}$. Cette droite passera par la solution exacte $u = (0, 0)$, si et seulement si il existe $\rho \in \mathbb{R}$ t.q.

$$x^{(0)} - \rho x^{(0)} = 0 \quad \text{et} \quad y^{(0)} - 2\rho y^{(0)} = 0.$$

On voit que c'est possible uniquement si l'une des valeurs $x^{(0)}$ ou $y^{(0)}$ est nulle. Ainsi, si l'on part d'un point $u^{(0)}$ qui n'est pas sur l'un des axes Ox ou Oy, le point suivant $u^{(1)}$ ne coïncidera pas avec la solution exacte. En utilisant la formule (4.10), on peut montrer que ce phénomène se reproduit à chaque itération ; plus précisément, si l'on part d'un $u^{(0)}$ t.q. $x^{(0)} \neq 0$ et $y^{(0)} \neq 0$, alors à chaque itération n, le point $u^{(n)} = (x^{(n)}, y^{(n)})$ vérifie également $x^{(n)} \neq 0$ et $y^{(n)} \neq 0$. Par conséquent, l'algorithme ne pourra pas atteindre la valeur exacte du minimiseur $u = (0, 0)$ en un nombre fini d'itérations.

Pourtant, en partant par exemple de la valeur $u^{(1)}$, la meilleure direction à choisir pour poursuivre la descente est la direction du vecteur $u^{(1)}$ lui-même (puisque la droite correspondasnte passe par l'origine). Nous allons voir que cette direction vérifie une propriété remarquable. Reprenons pour cela la formule (4.10). Elle nous permet d'écrire l'expression suivante :

$$u^{(1)} = u^{(0)} - \frac{\|d^{(0)}\|^2}{\langle Ad^{(0)}, d^{(0)} \rangle} d^{(0)}.$$

Puisque $d^{(0)} = -Au^{(0)}$, on peut alors calculer $\langle Au^{(1)}, d^{(0)} \rangle$:

$$\begin{aligned} \langle Au^{(1)}, d^{(0)} \rangle &= \langle Au^{(0)}, d^{(0)} \rangle - \frac{\|d^{(0)}\|^2}{\langle Ad^{(0)}, d^{(0)} \rangle} \langle Ad^{(0)}, d^{(0)} \rangle \\ &= \|d^{(0)}\|^2 - \|d^{(0)}\|^2 \\ &= 0. \end{aligned}$$

Ainsi, on pourra utiliser une autre direction de descente que celle du gradient au point $u^{(1)}$; cette direction $d^{(1)}$, colinéaire à $u^{(1)}$, pourra être définie (au signe près) par

$$d^{(1)} \neq 0, \quad \langle Ad^{(1)}, d^{(0)} \rangle = 0.$$

Une telle direction $d^{(1)}$ vérifiant $\langle Ad^{(1)}, d^{(0)} \rangle = 0$ s'appelle une DIRECTION CONJUGUÉE à la direction $d^{(0)}$, relativement à la matrice A.

4.4 Méthode du gradient conjugué

Dans toute cette section, on considère une fonctionnelle quadratique elliptique

$$J : x \in \mathbb{R}^N \mapsto \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$$

où $A \in \mathbb{R}^{N\times N}$ est une matrice symétrique définie positive et $b \in \mathbb{R}^N$ est un vecteur fixé.

Nous avons vu que les méthodes de gradient consistent, à partir d'un point $u^{(n)}$ calculé à l'itération n, à chercher le point suivant $u^{(n+1)}$ sur la droite passant par $u^{(n)}$ et dirigée par $d^{(n)} = -\nabla J(u^{(n)})$. Ainsi, seule l'information fournie par le gradient de J à l'étape n est utilisée pour déterminer la prochaine direction de descente. Le principe de la méthode du gradient conjugué consiste, au contraire, à utiliser tous les gradients calculés précédemment par l'algorithme $\nabla J(u^{(0)}), \nabla J(u^{(1)}), \ldots, \nabla J(u^{(n)})$ pour déterminer la prochaine direction de descente.

Principe de l'algorithme de gradient conjugué. Un vecteur initial $u^{(0)}$ ayant été donné, supposons les vecteurs $u^{(1)}, u^{(2)}, \ldots, u^{(n)}$ déjà calculés. On peut faire l'hypothèse

$$\nabla J(u^{(k)}) \neq 0, \quad 0 \leq k \leq n,$$

sinon la valeur exacte du minimiseur aurait déjà été atteinte. Pour $k = 0, 1, \ldots, n$, appelons G_n le sous-espace de $\mathbb{R}^N$ engendré par les gradients $\nabla J(u^{(0)}), \nabla J(u^{(1)}), \ldots, \nabla J(u^{(n)})$; c'est donc un sous-espace de dimension au plus $n+1$. L'idée de la méthode consiste à définir le vecteur suivant $u^{(n+1)}$ comme le minimiseur de J sur le plan affine passant par $u^{(n)}$ et dirigé par G_n. Ainsi, en notant $u^{(n)} + G_n$ ce plan affine,

$$u^{(n)} + G_n := \{u^{(n)} + w;\ w \in G_n\} = \left\{ u^{(n)} + \sum_{k=0}^{n} \delta_k \nabla J(u^{(k)})\ ;\ \delta_k \in \mathbb{R},\ 0 \leq k \leq n \right\},$$

le point $u^{(n+1)}$ vérifie :

$$u^{(n+1)} \in (u^{(n)} + G_n) \quad \text{et} \quad J(u^{(n+1)}) = \min_{v \in u^{(n)}+G_n} J(v). \tag{4.11}$$

L'ensemble $u^{(n)} + G_n$ étant fermé et convexe, et J étant coercive et strictement convexe, le problème de minimisation ci-dessus admet une solution unique.

On peut prévoir d'ores et déjà que la définition (4.11) fournira une valeur $J(u^{(n+1)})$ inférieure à celle que l'on aurait obtenue en appliquant une itération de la méthode

de gradient à pas optimal. En effet, la droite $\{u^{(n)} - \rho\nabla J(u^{(n)}),\ \rho \in \mathbb{R}\} =: u^{(n)} + \mathrm{Vect}(\nabla J(u^{(n)}))$ sur laquelle on minimise J pour mettre à jour le point $u^{(n)}$ dans la méthode de gradient à pas optimal, est contenue dans le plan $u^{(n)} + G_n$; par conséquent,

$$\min\Big\{J(v),\ v \in u^{(n)} + G_n\Big\} \leq \min\Big\{J(v),\ v \in u^{(n)} + \mathrm{Vect}(\nabla J(u^{(n)}))\Big\}.$$

Cependant, pour que la définition (4.11) soit applicable en pratique, il faut s'assurer que le problème de minimisation associé, qui porte sur $n+1$ variables $\delta_0, \delta_1, \ldots, \delta_n$, soit facile à résoudre. Nous allons voir que c'est le cas, et que sa résolution repose sur l'utilisation des DIRECTIONS CONJUGUÉES associées à la matrice A.

Remarquons tout d'abord que les solutions des problèmes successifs de minimisation

$$u^{(k+1)} \in (u^{(k)} + G_k) \quad \text{et} \quad J(u^{(k+1)}) = \min_{v \in u^{(k)} + G_k} J(v) = \min_{w \in G_k} J(u^{(k)} + w), \quad 0 \leq k \leq n, \tag{4.12}$$

vérifient

$$\langle \nabla J(u^{(k+1)}), w\rangle = 0 \quad \text{pour tout } w \in G_k. \tag{4.13}$$

En effet, pour $w \in G_k$, puisque G_k est un espace vectoriel et que $u^{(k+1)} \in (u^{(k)} + G_k)$, on a également, pour tout $t > 0$, $u^{(k+1)} + tw \in (u^{(k)} + G_k)$. Par définition du minimum, on peut donc écrire

$$0 \leq \frac{J(u^{(k+1)} + tw) - J(u^{(k+1)})}{t}$$

et passer à la limite quand $t \to 0$, ce qui donne $\langle \nabla J(u^{(k+1)}), w\rangle \geq 0$. En remplaçant w par $-w$ (ce qui est autorisé car G_k est un espace vectoriel), on obtient l'inégalité contraire, d'où (4.13). En particulier, on peut donc écrire :

$$\langle \nabla J(u^{(k+1)}), \nabla J(u^{(l)})\rangle = 0, \quad 0 \leq l \leq k \leq n,$$

c'est-à-dire que les gradients $\nabla J(u^{(k)})$, $0 \leq k \leq n+1$ sont deux à deux orthogonaux.

Remarque 4.3. Cette propriété est plus forte que la propriété (4.9) établie pour l'algorithme de gradient à pas optimal, où seulement deux gradients consécutifs sont orthogonaux.

Cette orthogonalité montre, en particulier, que les gradients $\nabla J(u^{(0)}), \nabla J(u^{(1)}), \ldots, \nabla J(u^{(n)})$ forment une *famille libre* (on a supposé qu'ils étaient tous non nuls). Cela implique que l'algorithme converge en au plus N itérations : en effet, si les N premiers vecteurs

$\nabla J(u_k)$, $0 \leq k \leq N-1$ sont différents de zéro, alors nécessairement le vecteur suivant $\nabla J(u^{(N)})$ est nul, sinon le sous-espace $G_N \subset \mathbb{R}^N$ contiendrait $N+1$ vecteurs indépendants. Par conséquent, $\nabla J(u_N) = 0$ et donc $u_N = u$.

Supposons que, à partir d'un vecteur initial $u^{(0)}$, on ait construit les vecteurs $u^{(1)}, \ldots, u^{(n)}$ en résolvant les problèmes de minimisation successifs définis par (4.12). Pour tout $0 \leq k \leq n$, on note $\Delta^{(k)} := u^{(k+1)} - u^{(k)}$ la différence entre deux approximations successives. Par construction, chaque vecteur $\Delta^{(k)}$ appartient au sous-espace G_k, dont il existe $k+1$ paramètres réels $\delta_0^k, \delta_1^k, \ldots, \delta_k^k$ t.q.

$$\Delta^{(k)} = \sum_{l=0}^{k} \delta_l^k \nabla J(u^{(l)}).$$

Nous allons montrer que ces vecteurs sont conjugués par rapport à la matrice A.

Définition 4.4. Soit $A \in \mathbb{R}^{N\times N}$ une matrice symétrique. On dit que des vecteurs $w^{(0)}, w^{(1)}, \ldots, w^{(n)}$ de $\mathbb{R}^N$ sont CONJUGUÉS par rapport à la matrice A s'ils vérifient :

$$w^{(k)} \neq 0, \quad 0 \leq k \leq n, \quad \text{et} \quad \langle Aw^{(l)}, w^{(m)} \rangle = 0, \quad 0 \leq m < l \leq n.$$

Remarque 4.4. Si l'on suppose de plus que A est définie positive, alors l'application

$$(x, y) \in \mathbb{R}^N \mapsto \langle Ax, y \rangle \in \mathbb{R}$$

définit un produit scalaire sur $\mathbb{R}^N$ (le produit scalaire usuel correspondant au choix $A = I_N$). Une famille de vecteurs non nuls est donc conjuguée par rapport à A si elle est orthogonale pour ce produit scalaire ; en particulier, elle forme donc une famille libre.

Montrons que les vecteurs $\Delta^{(k)}$ introduits plus haut sont conjugués par rapport à A. En utilisant l'expression de ∇J, on remarque que

$$\forall v, w \in \mathbb{R}^N \quad \nabla J(v+w) = A(v+w) - b = \nabla J(v) + Aw,$$

ce qui permet d'écrire

$$\nabla J(u^{(k+1)}) = \nabla J(u^{(k)} + \Delta^{(k)}) = \nabla J(u^{(k)}) + A\Delta^{(k)}, \quad 0 \leq k \leq n.$$

En utilisant l'orthogonalité des gradients $\nabla J(u^{(k)})$, $0 \leq k \leq n$, et en développant le produit scalaire, on obtient

$$0 = \langle \nabla J(u^{(k+1)}), \nabla J(u^{(k)}) \rangle = \|\nabla J(u^{(k)})\|^2 + \langle A\Delta^{(k)}, \nabla J(u^{(k)}) \rangle, \quad 0 \leq k \leq n.$$

Comme on a supposé $\nabla J(u^{(k)}) \neq 0$, on en déduit que $\langle A\Delta^{(k)}, \nabla J(u^{(k)})\rangle \neq 0$ et donc $\Delta^{(k)} \neq 0$ pour tout $0 \leq k \leq n$.

D'autre part, en écrivant $u^{(k+1)} = u^{(k)} + \Delta^{(k)}$, on calcule de la même manière

$$0 = \langle \nabla J(u^{(k+1)}), \nabla J(u^{(l)})\rangle = \langle \nabla J(u^{(k)}), \nabla J(u^{(l)})\rangle + \langle A\Delta^{(k)}, \nabla J(u^{(l)})\rangle, \quad 0 \leq l < k \leq n,$$

ce qui donne

$$\langle A\Delta^{(k)}, \nabla J(u^{(l)})\rangle = 0, \quad 0 \leq l < k \leq n.$$

Pour un entier m t.q. $0 \leq m < k \leq n$, chaque vecteur $\Delta^{(m)} \in G_m$ est une combinaison linéaire des vecteurs $\nabla J(u^{(l)})$, pour $0 \leq l \leq m$. Par conséquent, l'égalité précédente entraîne

$$\langle A\Delta^{(k)}, \Delta^{(m)}\rangle = 0, \quad 0 \leq m < k \leq n.$$

On peut montrer que la conjugaison par rapport à A des directions de descente $\Delta^{(k)}$, permet, à chaque itération n, de déterminer la direction de descente suivante et de résoudre le problème de minimisation (4.11) par des formules explicites. On aboutit aux expressions suivantes.

Définition 4.5 (Algorithme de gradient conjugué)**.** Soit $A \in \mathbb{R}^{N\times N}$ une matrice symétrique définie positive, $b \in \mathbb{R}^N$ et J la fonctionnelle quadratique associée. L'algorithme de GRADIENT CONJUGUÉ est le suivant. On se donne un point initial $u^{(0)}$. Si $\nabla J(u^{(0)}) \neq 0$, on définit $d^{(0)} = -\nabla J(u^{(0)})$, et tant que $\nabla J(u^{(n)}) \neq 0$, on réalise l'itération :

$$\begin{aligned} d^{(n)} &= -\nabla J(u^{(n)}) + \frac{\|\nabla J(u^{(n)})\|^2}{\|\nabla J(u^{(n-1)})\|^2} d^{(n-1)}, \\ \rho^{(n)} &= \frac{\|\nabla J(u^{(n)})\|^2}{\langle Ad^{(n)}, d^{(n)}\rangle}, \\ u^{(n+1)} &= u^{(n)} + \rho^{(n)} d^{(n)}. \end{aligned}$$

Chapitre 5

Contraintes d'égalité

Les contraintes d'égalité considérées sont du type

$$K := \left\{ x \in \mathbb{R}^N,\ \varphi_1(x) = 0,\ \ldots,\ \varphi_p(x) = 0 \right\} \tag{5.1}$$

où les fonctions $\varphi_i : \mathbb{R}^N \to \mathbb{R}$ sont données, et où p est un entier ≥ 1 qui représente le nombre de contraintes. On définira aussi la fonction à valeurs vectorielles $\varphi : \mathbb{R}^N \to \mathbb{R}^p$, par

$$\varphi(x) := \begin{pmatrix} \varphi_1(x) \\ \vdots \\ \varphi_p(x) \end{pmatrix}$$

Contraintes d'égalité affines. Il s'agit du cas particulier où chaque φ_i est affine : il existe des coefficients (c_{ij}) et (f_i) t.q.

$$\varphi_i(x) = \sum_{j=1}^{N} c_{ij} x_j - f_i.$$

En particulier en notant $x = (x_1 \ldots x_N)^T$ et

$$C := \begin{bmatrix} c_{11} & \cdots & c_{1N} \\ \vdots & & \vdots \\ c_{p1} & \cdots & c_{pN} \end{bmatrix} \quad \text{et} \quad f := \begin{pmatrix} f_1 \\ \vdots \\ f_p \end{pmatrix}, \tag{5.2}$$

on a l'égalité dans $\mathbb{R}^p$:

$$\varphi(x) = Cx - f.$$

K est donc un sous-espace affine de $\mathbb{R}^N$ dirigé par le sous-espace vectoriel $\{w \in \mathbb{R}^N,\ Cw = 0\}$. En effet, si $x, y \in K$, leur différence $x - y$ vérifie $C(x - y) = 0$. En particulier, on peut vérifier facilement que pour le cas de contraintes d'égalité affines, K est convexe.

5.1 Cas des contraintes d'égalité affines

Étant donné un ensemble $A \subset \mathbb{R}^N$, on notera $A^\perp$ son orthogonal, défini par

$$A^\perp := \{x \in \mathbb{R}^N, \quad \forall a \in A,\ \langle x, a\rangle = 0\}.$$

Rappelons que pour tout ensemble A, son orthogonal $A^\perp$ est un sous-espace vectoriel de $\mathbb{R}^N$.

Lemme 5.1. *Lorsque K est affine, le cône tangent en tout point u de K est un espace vectoriel donné par*

$$T_K(u) = \{d \in \mathbb{R}^N, Cd = 0\} = \{\nabla\varphi_1, \ldots, \nabla\varphi_p\}^\perp.$$

Preuve. On commence par établir que $T_K(u) = \{d \in \mathbb{R}^N, Cd = 0\}$. Pour cela, on procède par double inclusion.

— Soit $d \in T_K(u)$; d'après la définition 3.2, il existe des suites $u_n \in K$ et $t_n > 0$ t.q.

$$\lim_{n\to\infty} t_n = 0 \quad \text{et} \quad \lim_{n\to\infty} \frac{u_n - u}{t_n} = d.$$

Puisque $u_n, u \in K$, $C(u_n - u) = 0$ donc par linéarité,

$$C\left(\frac{u_n - u}{t_n}\right) = 0 \quad \text{pour tout } n.$$

En passant à la limite quand $n \to \infty$, on obtient $Cd = 0$.

— Réciproquement, soit $d \in \mathbb{R}^N$ t.q. $Cd = 0$. Montrons que $d \in T_K(u)$. Pour tout $n \in \mathbb{N}^*$, on pose $t_n = \frac{1}{n}$ et on définit $u_n = u + \frac{1}{n}d$. Puisque $Cd = 0$, on vérifie que $Cu_n = Cu = f$, ce qui montre que la suite $(u_n)_{n\in\mathbb{N}^*}$ est à valeurs dans K. Par conséquent, $d \in T_K(u)$.

Pour conclure la preuve, on observe enfin que $\nabla\varphi_i(x) = (c_{i1} \dots c_{iN})^T$, donc $(Cd)_i = \langle \nabla\varphi_i, d\rangle$. Ainsi

$$Cd = 0 \;\Leftrightarrow\; \forall i = 1, \dots, p,\ \langle \nabla\varphi_i, d\rangle = 0.$$

L'ensemble des vecteurs $d \in \mathbb{R}^N$ tels que $Cd = 0$ est donc exactement l'ensemble des vecteurs orthogonaux à tous les vecteurs $\nabla\varphi_1, \dots, \nabla\varphi_p$, d'où le résultat. □

Notons au passage les identités suivantes :

$$Cd = \begin{pmatrix} \langle \nabla\varphi_1, d\rangle \\ \vdots \\ \langle \nabla\varphi_p, d\rangle \end{pmatrix}, \qquad C \equiv \left[\nabla\varphi_1,\ \dots,\ \nabla\varphi_p\right]^T,$$

et encore

$$C^T = \left[\nabla\varphi_1,\ \dots,\ \nabla\varphi_p\right].$$

Théorème 5.1 (**multiplicateurs de Lagrange, contraintes d'égalité affines).** *Soit K un ensemble d'égalités affines. Si J est différentiable en $u \in K$ et si u est un minimum local de J sur K, alors*

$$\exists \lambda \in \mathbb{R}^p, \quad \nabla J(u) + C^T\lambda = 0 \tag{5.3a}$$

$$Cu - f = 0. \tag{5.3b}$$

On dit que les composantes du vecteur $\lambda = (\lambda_1 \dots \lambda_p)^T$ sont les multiplicateurs de Lagrange associés aux contraintes $\varphi_i(x) = 0$. Les conditions (5.3) *constituent les conditions d'optimalité d'ordre* 1 *du problème de minimisation. Elles s'écrivent de manière équivalente*

$$\exists \lambda \in \mathbb{R}^p, \quad \nabla J(u) + \sum_{i=1}^{p} \lambda_i \nabla\varphi_i(u) = 0 \tag{5.4a}$$

$$\varphi(u) = 0. \tag{5.4b}$$

Preuve. Nous avons établi au chapitre 3 que pour un problème général d'optimisation sous contraintes, les conditions d'optimalité d'ordre 1 s'écrivaient : pour tout $d \in T_K(u)$,

$$\langle -\nabla J(u), d\rangle \leq 0.$$

Mais comme ici $T_K(u)$ est un espace vectoriel, on a aussi $-d \in T_K(u)$ et donc

$$\langle -\nabla J(u), -d\rangle \leq 0.$$

En combinant les deux inégalités, on obtient

$$\forall d \in T_K(u), \quad \langle -\nabla J(u), d\rangle = 0$$

Ainsi,

$$-\nabla J(u) \in T_K(u)^\perp \equiv \{\nabla\varphi_1, \ldots, \nabla\varphi_p\}^{\perp\perp}$$

Rappelons alors le resultat suivant :

Lemme 5.2 (Théoreme du « bi-orthogonal »). *Pour tout sous-ensemble A non vide de $\mathbb{R}^N$,*

$$A^{\perp\perp} = \mathrm{V}ect(A)$$

Preuve du lemme 5.2. En exercice. On pourra montrer les propriétés suivantes :

- $A \subset A^{\perp\perp}$;
- $A^\perp = (\mathrm{V}ect(A))^\perp$;
- $\mathrm{V}ect(A)$ et $A^\perp$ sont en somme directe (pour cela, utiliser une base orthonormale de $\mathrm{V}ect(A)$ et la projection orthogonale de $\mathbb{R}^N$ sur $\mathrm{V}ect(A)$) ;

pour en déduire que $\mathrm{V}ect(A) \subset A^{\perp\perp}$ et conclure par un argument de dimension.

Fin de la preuve du théorème 5.1. Ainsi $-\nabla J(u) \in \mathrm{V}ect\{\nabla\varphi_1, \ldots, \nabla\varphi_p\}$, donc il existe des réels $\lambda_1 \ldots, \lambda_p$ t.q.

$$-\nabla J(u) = \sum_{i=1}^{p} \lambda_i \nabla\varphi_i(u).$$

Par ailleurs, en notant $\lambda = (\lambda_1, \ldots, \lambda_p)^T \in \mathbb{R}^p$, on peut écrire

$$\sum_{i=1}^{p} \nabla\varphi_i(u)\lambda_i = [\nabla\varphi_1(u), \ldots, \nabla\varphi_p(u)]\ \lambda \equiv C^T\lambda.$$

Cela conclut la preuve pour les deux versions (5.3) et (5.4). □

Exercice 2.

Soit $J(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$, où A est une matrice symétrique définie positive. Écrire les conditions d'optimalité du théorème des multiplicateurs de Lagrange dans ce cadre ; montrer qu'on obtient un système linéaire en les inconnues (u, λ).

5.3 Cas de contraintes d'égalité quelconques

Le cas général est plus difficile mais va conduire, sous des hypothèses adéquates (dites de « qualification des contraintes »), à des conditions d'optimalité similaires. On considère dans cette section que les fonctions $\varphi_1, \ldots, \varphi_p : \mathbb{R}^N \to \mathbb{R}$, sont de classe C^1.

Définition 5.1. On dira que les contraintes d'égalité (5.1) sont QUALIFIÉES en $u \in K$ si l'une des conditions suivantes est satisfaite :

- soit les contraintes sont LINÉAIRES : chaque fonction φ_i est affine ;
- soit la famille $\{\nabla\varphi_1(u), \ldots, \nabla\varphi_p(u)\}$ est LIBRE.

Théorème 5.2. *Soit u un point de K, un minimiseur local de J sur K. On suppose que J est différentiable en u et que les contraintes sont qualifiées en u. Alors le théorème des multiplicateurs de Lagrange est encore valable :*

$$\exists \lambda \in \mathbb{R}^p, \quad \nabla J(u) + \sum_{i=1}^{p} \lambda_i \nabla\varphi_i(u) = 0 \tag{5.5a}$$

$$\varphi(u) = 0. \tag{5.5b}$$

Preuve. Le coeur de la démonstration repose sur la caractérisation

$$T_K(u) = \{\nabla\varphi_1(u), \ldots, \nabla\varphi_p(u)\}^{\perp}.$$

L'inclusion $\subset$ est facile à montrer ; la réciproque nécessite l'usage du théorème des fonctions implicites. Le détail de la preuve est fait dans le Complément 5.4. Le reste de la preuve est similaire au cas des contraintes affines.

5.4 Complément : preuve du Théorème 5.2

On suppose que $d \in \{\nabla\varphi_1(u), \ldots, \nabla\varphi_p(u)\}^{\perp}$ (c'est-à-dire $\langle \nabla\varphi_i(u), d\rangle = 0$ pour tout i), et on désire montrer que $d \in T_K(u)$. Notons que la matrice jacobienne $D\varphi$ est une matrices $p \times N$, qui s'écrit

$$D\varphi = \left(\frac{\partial\varphi_i}{\partial x_j}\right)_{ij} = \begin{pmatrix} \nabla\varphi_i^T \\ \vdots \\ \nabla\varphi_p^T \end{pmatrix}.$$

Donc par hypothèse on a

$$D\varphi(u)\cdot d = \begin{pmatrix} \langle \nabla\varphi_i(u), d\rangle \\ \vdots \\ \langle \nabla\varphi_p(u), d\rangle \end{pmatrix} = 0$$

Le cas où les p contraintes sont toutes affines ayant déjà été traité, on suppose donc que les gradients $\nabla\varphi_1(u), \ldots, \nabla\varphi_p(u)$ sont linéairement indépendants (ce qui impose en particulier que $N \geq p$, puisque les gradients forment une famille llibre de p vecteurs de $\mathbb{R}^N$). Cela signifie que la matrice $D\varphi$, dont les p lignes sont formées de vecteurs linéairement indépendants, est de rang p. Par conséquent, $D\varphi$ contient également p vecteurs colonnes indépendants. Quitte à réordonner les coordonnées, on peut donc supposer que les p premiers vecteurs colonnes de $D\varphi(u)$ forment une famille libre dans $\mathbb{R}^p$.

Notons pour un vecteur x de $\mathbb{R}^N$, $x = (x^1, x^2)$ où x^1 contient les p premières composantes de x et x^2 les $N-p$ autres ($N-p \geq 0$). En particulier, $\varphi(x) = \varphi(x^1, x^2)$, et on peut noter $D_1\varphi$ et $D_2\varphi$ les dérivées par rapport aux coordonnées x^1 et x^2 respectivement. L'hypothese d'indépendance des gradients revient donc à dire que $D_1\varphi(u)$ est une matrice inversible. D'après le théorème des fonctions implicites il existe des voisinages de u^1 et de u^2 et une fonction Ψ de classe C^1 t.q., dans ces voisinages,

$$\varphi(x^1, x^2) = 0 \Leftrightarrow x^1 = \Psi(x^2).$$

En particulier, $u^1 = \Psi(u^2)$.

Afin de construire une suite x_n dans K, on procède en posant d'abord

$$x_n^2 = u^2 + \frac{1}{n}d^2,$$

et

$$x_n^1 = \Psi(x_n^2).$$

Par construction, on a donc $x_n \in K$ pour n assez grand. Ensuite, par développement limité,

$$\begin{aligned} x_n^1 &= \Psi(u^2) + \frac{1}{n}D_2\Psi(u^2)d^2 + o(\frac{1}{n}). \\ &= u^1 + \frac{1}{n}D_2\Psi(u^2)d^2 + o(\frac{1}{n}). \end{aligned}$$

Montrons que

$$D_2\Psi(u^2)d^2 = d^1. \tag{5.6}$$

On aura ainsi $x_n = u + \frac{1}{n}d + o(\frac{1}{n})$, avec $x_n \in K$, et donc $d \in T_K(u)$.

D'abord, on a

$$0 = D\varphi(u)d = [D_1\varphi D_2\varphi] \begin{pmatrix} d^1 \\ d^2 \end{pmatrix} = D_1\varphi(u)d^1 + D_2\varphi(u)d^2. \tag{5.7}$$

De l'identité $\varphi(\Psi(x^2), x^2)) = 0$, en différentiant en u^2, on obtient

$$D_1\varphi(u)D_2\Psi(u^2) + D_2\varphi(u) = 0. \tag{5.8}$$

On applique cette dernière identité à d^2, et en identifiant avec (5.7), et en simplifiant par $D_1\varphi(u)$ qui est inversible, on obtient l'identité désirée (5.6). □

Chapitre 6

Contraintes d'inégalité, contraintes mixtes

On considère des contraintes d'inégalité du type

$$K := \Big\{ x \in \mathbb{R}^N,\ \varphi_1(x) \leq 0,\ \ldots,\ \varphi_p(x) \leq 0 \Big\} \tag{6.1}$$

où $\varphi_i : \mathbb{R}^N \to \mathbb{R}$, et où p est un entier ≥ 1 qui représente le nombre de contraintes. On notera aussi la fonction $\varphi : \mathbb{R}^N \to \mathbb{R}^p$, définie par $\varphi(x) := (\varphi_1(x) \ldots \varphi_p(x))^T$.

On utilisera la notation $X \leq Y$, pour deux vecteurs $X = (x_i)$ et $Y = (y_i)$, lorsque $x_i \leq y_i$, $\forall i$, ainsi que la notation $X \leq 0$ pour dire que $x_i \leq 0$, $\forall i$. Ainsi, l'ensemble K s'écrira $K \equiv \{x \in \mathbb{R}^N,\ \varphi(x) \leq 0\}$.

Contraintes d'inégalité affines. Il s'agit du cas particulier où chaque φ_i est affine : il existe des coefficients (c_{ij}) et (f_i) t.q. $\varphi_i(x) = \sum_{j=1}^N c_{ij} x_j - f_i$. En particulier en notant $x = (x_1, \ldots, x_N)^T$ et $C = (c_{ij})$, $f = (f_i)$, on a

$$K \equiv \{x \in \mathbb{R}^N,\ Cx - f \leq 0\}.$$

Définition 6.1. Pour $u \in K$, on note $A(u) := \{i \in \{1, \ldots, p\},\ \varphi_i(u) = 0\}$ l'ensemble des **contraintes actives**, ou « **saturées** ».

Il sera important de distinguer
— les contraintes **actives** ($\varphi_i(u) = 0$),
— les contraintes **inactives** ($\varphi_i(u) < 0$).

6.1 Cas des contraintes d'inégalité affines

On note que si K est un ensemble de contraintes d'inégalité affines, alors K est un convexe.

Lemme 6.1. *(i) De manière générale (K quelconque), on a*

$$T_K(u) \subset \{\nabla\varphi_i(u),\ i \in A(u)\}^o$$

(ii) Lorsque K est affine, on a

$$T_K(u) = \{\nabla\varphi_i(u),\ i \in A(u)\}^o$$

On peut dire que le cône des directions admissibles, en tout point u de K, est « le polaire des gradients des contraintes actives ».

Preuve du lemme 6.1.
Cas (i) : soit $d \in T_K(u)$ et $i \in A(u)$, il faut montrer que $\langle d, \nabla\varphi_i(u)\rangle \leq 0$. $d \in T_K(u)$ donc il existe des suites $t_n \searrow 0$, $u_n \in K$ t.q. $\lim_{n\to\infty} \frac{u_n-u}{t_n} = d$. En notant $d_n = \frac{u_n-u}{t_n}$ et en utilisant le fait que $u_n \in K$ et un développement de Taylor de φ_i en u, dans la direction d_n, on obtient :

$$\begin{aligned} 0 \geq \varphi_i(u_n) &= \varphi_i(u + t_n d_n) \\ &= \varphi_i(u) + t_n\langle\nabla\varphi_i(u), d_n\rangle + o(t_n). \end{aligned}$$

Or $i \in A(u)$ donc $\varphi_i(u) = 0$, d'où la relation

$$\langle\nabla\varphi_i(u), d_n\rangle = \frac{\varphi_i(u_n)}{t_n} + o(1).$$

Comme $\varphi_i(u_n) \leq 0$ et que $\lim_{n\to\infty}\langle\nabla\varphi_i(u), d_n\rangle = \langle\nabla\varphi_i(u), d\rangle$, le résultat s'en déduit par passage à la limite dans l'égalité précédente.
Cas (ii) : il s'agit de montrer que l'inclusion réciproque est vraie, dans le cas d'inégalités affines. Soit donc $d \in \{\nabla\varphi_i(u),\ i \in A(u)\}^o$. On définit $u_n = u + \frac{1}{n}d$ (ce qui correspond au choix $t_n = \frac{1}{n}, d_n = d$ pour vérifier la définition de $T_K(u)$). Vérifions que pour n assez grand, $u_n \in K$. Considérons tout d'abord le cas des contraintes inactives. Soit $j \in \{1, \ldots, p\}\backslash A(u)\,;\, \varphi_j(u) < 0$ et $u_n \to u$ donc par continuité, il existe un entier $N_j \in \mathbb{N}$

t.q. pour tout $n \geq N_j$, $\varphi_j(u_n) < 0$. En prenant $N = \max\{N_j,\ j \in \{1,\dots,p\} \setminus A(u)\}$, on a donc :

$$\forall n \in \mathbb{N},\ n \geq N\ \Rightarrow \forall j \in \{1,\dots,p\} \setminus A(u),\ \phi_j(u_n) < 0.$$

Pour le cas des contraintes actives, on remarque que pour des contraintes affines, la formule de développement de Taylor utilisée dans le cas précédent devient exacte (c'est-à-dire sans terme de reste) : si $i \in A(u)$, $\varphi_i(u) = 0$ et on peut écrire le développement suivant :

$$\varphi_i(u_n) = \frac{1}{n}\langle \nabla\varphi_i(u), d\rangle \leq 0$$

puisque d appartient au cône polaire des directions admissibles. On en déduit que pour tout $n \geq N$, $u_n \in K$.

□

Pour exprimer la condition d'optimalité $-\nabla J(u) \in T_K(u)^o$, on aura donc besoin de décrire un « bipolaire », c'est-à-dire le cône polaire d'un cône polaire. Rappelons la notation :

$$\Gamma(a_1,\dots,a_p) := \{\sum_{i=1}^{p} \lambda_i a_i,\ \lambda_i \geq 0\}.$$

Lemme 6.2 (de Farkas - ou « Théorème du bipolaire »)**.** *Pour tout $a_1,\dots,a_p$ dans $\mathbb{R}^N$, on a*

$$(a_1,\dots,a_p)^{oo} = \Gamma(a_1,\dots,a_p)$$

Pour la preuve de ce résultat, nous avons besoin de deux résultats préliminaires.

Lemme 6.3. *(Théorème de séparation.) Soit K un convexe fermé non vide de $\mathbb{R}^N$, et $u \notin K$. Alors il existe $d \in \mathbb{R}^N$, $\exists b \in \mathbb{R}$,*

$$\langle d, u\rangle < b < \langle d, x\rangle \quad \forall x \in K.$$

(On dit que l'hyperplan $\langle d, x\rangle = b$ sépare u de K.)

Preuve. Soit $p = \Pi_K(u)$, la projection de u sur K. $u \notin K$ donc $p \neq u$. Soit $d := p - u$ (donc $d \neq 0$). On a

$$\forall x \in K, \quad 0 \geq (x - p, u - p) = -(d, x - p).$$

Donc

$$(x, d) \geq (d, p) = (d, d) + (u, d).$$

Au final, on choisit $b = \frac{1}{2}(d, d) + (u, d)$: on vérifie que $(x, d) > b$, et $b > (u, d)$. □

Lemme 6.4. $\Gamma(a_1, \ldots, a_p)$ *est fermé*

Preuve. Peut se faire par récurrence sur p. □

Preuve du Lemme 6.2 : D'abord on vérifie facilement l'inclusion $\Gamma(a_1, \ldots, a_p) \subset \{a_1, \ldots, a_p\}^{oo}$. Réciproquement, on considère $K := \Gamma(a_1, \ldots, a_p)$. C'est un convexe, fermé, non vide ($0 \in K$). Supposons (par l'absurde) l'existence d'un élément u, $u \notin K$ et $u \in \{a_1, \ldots, a_p\}^{oo}$. D'après le Théorème de séparation, $\exists d \in \mathbb{R}^N$, $b \in \mathbb{R}$,

$$\langle d, v\rangle > b > \langle d, u\rangle, \quad \forall v \in K. \tag{6.2}$$

Notons que
(i) $\langle d, u\rangle < 0$ car on peut prendre $v = 0 \in K$ dans (6.2).
(ii) Aussi, $\forall \lambda \geq 0$, $\lambda a_i \in K$ donc

$$b < \langle d, \lambda a_i\rangle = \lambda \langle d, a_i\rangle.$$

A la limite $\lambda \to +\infty$, cela implique que $\langle d, a_i\rangle \geq 0$. Donc

$$-d \in \{a_1, \ldots, a_p\}^{o}.$$

(iii) Par définition, puisque $u \in \{a_1, \ldots, a_p\}^{oo}$, on a donc $\langle -d, u\rangle \leq 0$, soit $0 \leq \langle d, u\rangle$. C'est en contradiction avec (i). □

Voici un premier énoncé du Théorème de Karush, Kuhn et Tucker ou "KKT", dans le cas simplifié de contraintes d'inégalité affines. (Le théorème général de KKT concerne en fait les contraintes mixtes et sera vu plus loin.)

Théorème 6.1 (**Karush, Kuhn et Tucker, cas d'inégalités affines**). *Soit K un ensemble d'inégalités affines. On suppose que u est un minimiseur de J sur K, et que J est différentiable en $u \in K$. Alors*

$$\exists \lambda = (\lambda_1, \ldots, \lambda_p)^T \in \mathbb{R}^p, \quad \nabla J(u) + C^T \lambda = 0 \tag{6.3a}$$

$$\lambda \geq 0, \quad Cu - f \leq 0, \tag{6.3b}$$

$$\forall i = 1, \ldots, p, \ (Cu - f)_i = 0 \ \textit{ou} \ \lambda_i = 0. \tag{6.3c}$$

On dira encore que $\lambda = (\lambda_1, \ldots, \lambda_p)^T$ sont des multiplicateurs. L'ensemble des conditions (6.3) *représentent les conditions d'optimalité d'ordre* 1 *du problème de minimisation.*

Elles s'écrivent de manière équivalente

$$\exists \lambda = (\lambda_1, \dots, \lambda_p)^T \in \mathbb{R}^p, \quad \nabla J(u) + \sum_{i=1}^{p} \lambda_i \nabla \varphi_i(u) = 0 \tag{6.4a}$$

$$\lambda \geq 0, \ \varphi(u) \leq 0, \tag{6.4b}$$

$$\langle \varphi(u), \lambda \rangle = 0. \tag{6.4c}$$

Preuve. On a vu que pour u, point de minimum local de J sur K : $(\nabla J(u), d) \geq 0$ $\forall d \in T_K(u)$, soit, d'après le Lemme 6.1 et le Lemme de Farkas 6.2 :

$$\begin{aligned} -\nabla J(u) &\in T_K(u)^o = \left\{ \nabla \varphi_i(u), \ i \in A(u) \right\}^{oo} \\ &\in \Gamma\{\varphi_i(u), \ i \in A(u)\}. \end{aligned}$$

Donc il existe $(\lambda_1, \dots, \lambda_p) \in (\mathbb{R}_+)^p$ t.q.

$$-\nabla J(u) = \sum_{i=1}^{p} \lambda_i \nabla \varphi_i(u),$$

où l'on a choisi simplement $\lambda_i = 0$ si $i \notin A(u)$. En particulier, on a donc soit $\varphi_i(u) = 0$, soit $\varphi_i(u) < 0$ et dans ce cas $i \notin A(u)$ et donc $\lambda_i = 0$. Cela implique aussi que

$$\langle \lambda, \varphi(u) \rangle = \sum_i \lambda_i \varphi_i(u) = 0,$$

ce qui conclut la preuve des relations (6.4). Pour obtenir l'écriture vectorielle (6.3), on utilise le fait que $\sum_i \lambda_i \nabla \varphi_i(u) = C^T \lambda$ pour $\lambda = (\lambda_1, \dots, \lambda_p)^T$.

6.2 Cas général - contraintes d'inégalité

Définition 6.2 (qualification des contraintes)**.** Soit $u \in K$. On suppose ici que les fonctions φ_i sont différentiables en u. On dira que les contraintes sont QUALIFIÉES en u si

$$\exists w \in \mathbb{R}^N, \quad \forall i \in A(u), \quad \begin{cases} \text{soit } \langle \nabla \varphi_i(u), w \rangle < 0, \\ \text{soit } \langle \nabla \varphi_i(u), w \rangle = 0, \text{ et } \varphi_i \text{ affine.} \end{cases} \tag{6.5}$$

Géométriquement, $\nabla\varphi_i(u)$ représente la normale sortante à la courbe $\varphi_i(v) = 0$, en $v = u$ (normale dirigée suivant la région ou $\varphi_i > 0$). Donc cela revient à supposer qu'on a un vecteur w qui est rentrant pour les contraintes (actives) d'inégalité, et strictement rentrant par rapport aux contraintes d'inégalité (actives) non affines.

Théorème 6.2 (CO1, contraintes générales d'inégalité)**.** *On suppose que u est un point de minimum local de J sur K, que $J, \varphi_1, \ldots, \varphi_p$ sont différentiables en u et que les contraintes sont qualifiées en u. Alors les conclusions du théorème KKT,* (6.4)*, restent valables.*

Preuve. Toute la preuve repose sur la caractérisation suivante.

$$T_K(u) = \{\nabla\varphi_i(u),\ i \in A(u)\}^o. \tag{6.6}$$

Pour le vérifier, notons W l'ensemble de droite. On a déjà vu que $T_K(u) \subset W$ (Lemme 6.1). Réciproquement, prenons $d \in W$: on a

$$\langle d, \nabla\varphi_i(u)\rangle \leq 0, \quad \forall i \text{ t.q. } \varphi_i(u) = 0. \tag{6.7}$$

Pour montrer que $d \in T_K(u)$, nous allons procéder indirectement : en considérant un vecteur w vérifiant les propriétés (6.5), nous allons montrer que pour $\lambda > 0$ fixé, $d+\lambda w \in T_K(u)$. Comme $T_K(u)$ est fermé, en passant à la limite quand $\lambda \to 0$, on conclura alors que la direction limite d appartient également à $T_K(u)$.

Ainsi, soit w un vecteur vérifiant (6.5) et soit $\lambda > 0$; on introduit la suite $u_n = u + \frac{1}{n}(d + \lambda w)$. Montrons que pour n assez grand, $u_n \in K$.

Si $i \notin A(u)$ alors $\varphi_i(u) < 0$ et donc $\varphi_i(u + \frac{1}{n}(d + \lambda w)) < 0$ pour n assez grand.

Si $i \in A(u) : \varphi_i(u) = 0$. Premier sous-cas : $\langle\nabla\varphi_i(u), w\rangle < 0$: alors, en utilisant (6.7),

$$\begin{aligned}\varphi_i(u_n) &= \varphi_i(u) + \frac{1}{n}\langle\nabla\varphi_i(u), d + \lambda w\rangle + o(\frac{1}{n}),\\ &\leq \frac{1}{n}\Big(\lambda\langle\nabla\varphi_i(u), w\rangle + o(1)\Big).\end{aligned}$$

Ansi $\varphi_i(u_n) < 0$ pour n assez grand. Deuxième sous-cas : $\langle\nabla\varphi_i(u), w\rangle = 0$ avec φ_i affine. Alors

$$\begin{aligned}\varphi_i(u_n) &= \varphi_i(u) + \frac{1}{n}\langle\nabla\varphi_i(u), d + \lambda w\rangle \quad (\text{car } \varphi_i \text{ est affine})\\ &= \frac{1}{n}\langle\nabla\varphi_i(u), d\rangle \ \leq\ 0.\end{aligned}$$

On en déduit que pour n assez grand, $\forall i,\ \varphi_i(u_n) \leq 0$. Cela montre que $d + \lambda w \in T_K(u)$. On conclut alors que $d \in T_K(u)$, ce qui conclut la preuve de (6.6), et du Théorème 6.2. □

6.3 Contraintes mixtes

On considère maintenant le cas le plus général des contraintes mixtes :

$$K := \left\{ x \in \mathbb{R}^N,\ \varphi_i(x) = 0,\ 1 \leq i \leq p, \quad \psi_j(x) \leq 0,\ 1 \leq j \leq q \right\} \tag{6.8}$$

où $\varphi_i, \psi_j : \mathbb{R}^N \to \mathbb{R}$, avec $p, q \geq 0$ entiers qui représentent le nombre de contraintes d'égalité ou d'inégalité, respectivement. On notera aussi les fonctions $\varphi : \mathbb{R}^N \to \mathbb{R}^p$ et $\psi : \mathbb{R}^N \to \mathbb{R}^q$:

$$\varphi(x) := (\varphi_1(x), \ldots, \varphi_p(x))^T, \quad \psi(x) := (\psi_1(x), \ldots, \psi_p(x))^T,$$

de sorte que $K = \{x,\ \varphi(x) = 0 \text{ et } \psi(x) \leq 0\}$.

Le cas particulier des contraintes affines (mixtes) s'écrit alors

$$K = \{x,\ Cx - f = 0,\ Dx - g \leq 0\},$$

pour des matrices $C \in \mathbb{R}^{p \times N}$, $f \in \mathbb{R}^p$ et $D \in \mathbb{R}^{q \times N}$, $g \in \mathbb{R}^q$.

Nous allons pouvoir écrire les conditions d'optimalité pour un minimum sous contraintes mixtes, dans deux cas : soit dans le cas des contraintes affines, soit dans un cadre général sous une hypothèse de qualification des contraintes. Le résultat final sera le Théorème de Karush, Kuhn et Tucker.

Définition 6.3 (qualification des contraintes)**.** Soit $u \in K$. On suppose que les fonctions φ_i sont de classe C^1 au voisinage de u, et les ψ_j sont différentiables en u. On dira que les contraintes sont QUALIFIÉES en u si : soit toutes les contraintes sont affines, soit il existe un vecteur $w \in \mathbb{R}^N$,

- $\{\nabla\varphi_1(u), \ldots, \nabla\varphi_p(u)\}$ libre, (6.9)

  et $\langle \nabla\varphi_i(u), w \rangle = 0,\ \forall 1 \leq i \leq p,$ (6.10)

- $\forall i \in A(u)$, $\begin{cases} \text{soit } \langle \nabla\psi_i(u), w \rangle < 0, \\ \text{soit } \langle \nabla\psi_i(u), w \rangle = 0, \text{ et } \psi_i \text{ affine.} \end{cases}$ (6.11)

Géométriquement cela revient à supposer qu'on a un vecteur w qui est tangent aux contraintes d'égalité, et rentrant pour les contraintes d'inégalité (strictement rentrant si ψ_i n'est pas affine).

On remarque que si toutes les contraintes sont affines, alors tout point u de K est qualifié.

Lemme 6.5. *Si les contraintes sont qualifiées en un point $u \in K$ (et en particulier pour des contraintes affines), on a*

$$T_K(u) = \{\nabla\varphi_i(u), 1 \leq i \leq p\}^\perp \cap \{\nabla\psi_j(u), j \in A(u)\}^o. \tag{6.12}$$

Preuve. On commence par vérifier l'inclusion $\subset$. Ensuite, pour l'inclusion réciproque : la preuve est simple dans le cas affine ; dans le cas général, on procède comme dans la preuve du Théorème 5.2, et de celle du Théorème 6.2. En supposant que d est dans l'ensemble de droite de (6.12), on pose $d' = d + \lambda w$ pour un $\lambda > 0$. On construit une suite u_n vérifiant $\varphi_i(u_n) = 0$, $u_n = u + \frac{1}{n}d' + o(\frac{1}{n})$. Enfin on vérifie que cette suite vérifie aussi $\psi_j(u_n) \leq 0$, pour n assez grand. Ainsi $d + \lambda w \in T_K(u)$, pour tout $\lambda > 0$, et on conclut à $d \in T_K(u)$.

Théorème 6.3 (**Karush, Kuhn et Tucker, cas mixte).** *Soit K un ensemble de contraintes mixtes comme défini par* (6.8)*. On suppose que u est un point de minimum de J sur K, J est différentiable en $u \in K$, et les contraintes sont qualifiées en u. Alors*

$$\exists \lambda = (\lambda_1, \ldots, \lambda_p)^T \in \mathbb{R}^p,\ \mu = (\mu_1, \ldots, \mu_q)^T \in \mathbb{R}^q, \tag{6.13a}$$
$$\nabla J(u) + \sum_{i=1}^{p} \lambda_i \nabla\varphi_i(u) + \sum_{j=1}^{q} \mu_j \nabla\psi_j(u) = 0,$$
$$\varphi_i(u) = 0, \quad \forall 1 \leq i \leq p, \tag{6.13b}$$
$$\mu_j \geq 0, \quad \psi_j(u) \leq 0, \ \textit{et} \ \ \mu_j\psi_j(u) = 0, \quad \forall 1 \leq j \leq q. \tag{6.13c}$$

On dira encore que les $\lambda = (\lambda_1, \ldots, \lambda_p)^T$ et $\mu = (\mu_1, \ldots, \mu_q)^T$ sont des multiplicateurs. L'ensemble des conditions (6.3) *représentent les conditions d'optimalité d'ordre* 1 *du problème de minimisation, ou* **conditions (KKT)**.

Preuve. On remarque que

$$\begin{aligned} T_K(u) &= \{\nabla\varphi_i(u), 1 \leq i \leq p\}^\perp \cap \{\nabla\psi_j(u), j \in A(u)\}^o \\ &= \{(\nabla\varphi_i(u), -\nabla\varphi_i(u))_{1\leq i\leq p}, (\nabla\psi_j(u))_{j\in A(u)}\}^o, \end{aligned}$$

et donc, d'après le lemme de Farkas 6.2,

$$-\nabla J(u) \quad \in \quad T_K(u)^o = \Gamma\Big(\pm \nabla\varphi_i(u)_{1\leq i\leq p}, (\nabla\psi_j(u))_{j\in A(u)} \Big)$$

En particulier il existe des coefficients $\lambda_i^1, \lambda_i^2 \geq 0$ et $\mu_j \geq 0$ t.q.

$$\begin{aligned} -\nabla J(u) &= \sum_i \lambda_i^1 \nabla\varphi_i(u) + \lambda_i^2(-\nabla\varphi_i(u)) + \sum_{j\in A(u)} \mu_j \nabla\psi_j(u) \\ &= \sum_i \lambda_i \nabla\varphi_i(u) + \sum_{j\in A(u)} \mu_j \nabla\psi_j(u). \end{aligned}$$

On conclut comme dans le cas des contraintes d'inégalité.

Théorème 6.4. *Réciproquement, si les conditions (KKT) sont satisfaites, si J est convexe sur K, si les φ_i sont affines et les ψ_j sont convexes (avec $J, \psi_1, \ldots, \psi_q$ différentiables en $u \in K$), alors u est un minimiseur global de J sur K.*

Preuve. Posons

$$\mathcal{L}(v, \alpha, \beta) := J(v) + \sum_i \alpha_i \varphi_i(v) + \sum_j \beta_j \psi_j(v),$$

aussi appelé Lagrangien du problème. On a $v \to \mathcal{L}(v, \lambda, \mu)$ convexe, puisque J convexe, $v \to \sum_i \alpha_i \varphi_i(v)$ est affine donc convexe, et les $\mu_j \geq 0$ donc $\sum_j \beta_j \psi_j(v)$ est convexe. Enfin la somme de fonctions convexes est convexe. De plus, $\nabla_v \mathcal{L}(u, \lambda, \mu) = \nabla J(u) + \sum_{i=1}^p \lambda_i \nabla\varphi_i(u) + \sum_{j=1}^q \mu_j \nabla\psi_j(u) = 0$ d'après KKT. Ainsi, u est un minimiseur global de $\mathcal{L}$ sur $\mathbb{R}^N$:

$$\mathcal{L}(u, \lambda, \mu) \leq \mathcal{L}(v, \lambda, \mu), \qquad \forall v \in \mathbb{R}^N.$$

Mais par ailleurs, au vu des conditions (KKT), on a $J(u) = \mathcal{L}(u, \lambda, \mu)$, et pour v dans K on voit que $\mathcal{L}(v, \lambda, \mu) \leq J(v)$ (en utilisant que $\mu \geq 0$). Ainsi $J(u) \leq J(v)$ pour tout $v \in K$. □

Chapitre 7

Algorithmes de minimisation pour les problèmes avec contraintes

7.1 Algorithme de gradient projeté

On suppose que $K \subset \mathbb{R}^N$ est un convexe fermé non vide, et $J : \mathbb{R}^N \to \mathbb{R}$. On cherche à minimiser la fonctionnelle J sur K. On suppose J différentiable.

Algorithme de Gradient Projeté (GP)
On prend un point de départ $u^0 \in \mathbb{R}^N$. On se donne un pas fixe $\rho > 0$.
On itère sur $n \geq 0$:

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

Théorème 7.1. *Soit $J : \mathbb{R}^N \to \mathbb{R}$, α-convexe, différentiable, avec ∇J : M-lipschitzien pour un $M > 0$.*
(i) Il existe un unique minimiseur u de J sur K, et, pour tout $\rho > 0$, ce minimiseur est caractérisé par

$$u = \Pi_K(u - \rho \nabla J(u)). \tag{7.1}$$

(ii) Si $\rho \in]0, \frac{2\alpha}{M^2}[$*, alors pour tout* $u^0 \in \mathbb{R}^N$*, l'algorithme (GP) converge vers* u *:*

$$\lim_{n\to\infty} u^n = u.$$

(iii) Enfin, la convergence est linéaire : $\exists 0 \leq R < 1$, $\exists C \geq 0$, $\|u^n - u\| \leq CR^n$ *(*$\forall n \geq 0$*).*

On pourrait aussi décider de faire varier le pas à chaque itération, et proposer une méthode de gradient à pas optimal projeté.

Preuve. (i) : Comme J est α-convexe et différentiable, le minimiseur u de J sur K est bien défini et unique. De plus il vérifie $u \in K$ et la condition d'optimalité

$$\langle \nabla J(u),\ v - u\rangle \geq 0, \quad \forall v \in K. \tag{7.2}$$

On en déduit que $u \in K$ et

$$\langle u - \rho\nabla J(u) - u,\ v - u\rangle \leq 0, \quad \forall v \in K.$$

Ces deux propriétés caractérisent le fait que

$$u = \Pi_K(u - \rho\nabla J(u)). \tag{7.3}$$

Réciproquement, cette relation est équivalente à (7.2). Comme J est convexe, cette condition implique que u est un minimiseur global de J sur K.

(ii)-(iii) : On peut faire la différence entre le schéma et (7.3). En utilisant le fait que Π_K est 1-lipschitzienne,

$$\begin{aligned} \|u^{n+1} - u\|^2 &= \|\Pi_K(u^n - \rho\nabla J(u^n)) - \Pi_K(u - \rho\nabla J(u))\|^2 && (7.4)\\ &\leq \|(u^n - \rho\nabla J(u^n)) - (u - \rho\nabla J(u))\|^2 && (7.5)\end{aligned}$$

La fin de la preuve est la même que pour la méthode de gradient à pas fixe.

□

Le problème du schéma est qu'il faut pouvoir calculer Π_K. Si Π_K est facilement calculable, on peut utiliser l'algorithme de gradient projeté (voir section 7.2). Sinon, on verra d'autres algorithmes à la section 7.3.

7.2 Cas particuliers de projections

Lemme 7.1. *Si $A \subset \mathbb{R}^p$ et $B \subset \mathbb{R}^q$ sont deux ensembles convexes fermés non vides, et $(x, y) \in \mathbb{R}^p \times \mathbb{R}^q$:*

$$\Pi_{A\times B}((x, y)) = (\Pi_A(x), \Pi_B(y)).$$

Ceci se généralise facilement à un produit :

$$\Pi_{A^1\times\cdots\times A^k}((x_1, \ldots, x_k)) = (\Pi_{A_1}(x_1), \ldots, \Pi_{A_k}(x_k)).$$

Lemme 7.2. *Si $-\infty \leq a \leq b \leq +\infty$,*

$$\Pi_{[a,b]}(x) \quad = \quad \begin{cases} a & si\ x \leq a \\ x & si\ x \in [a, b] \\ b & si\ x > b \end{cases} \quad = \quad \min(\max(x, a), b).$$

Corollaire : Projection sur un parallelépipède. Si $K = \prod_{i=1}^N [a_i, b_i]$ et $x = (x_i)_{1\leq i\leq N}$, alors

$$\Pi_K(x) = \Big(\Pi_{[a_i,b_i]}(x_i)\Big)_{1\leq i\leq N} \equiv \Big(\min(\max(x_i, a_i), b_i)\Big)_{1\leq i\leq N}$$

Dans le cas particulier où $K = (\mathbb{R}_+)^p$, on obtient $\Pi_K(x) = \max(x, 0) = \Big(\max(x_i, 0)\Big)_{1\leq i\leq N}$.

En conclusion, si K est particulier (un parallèlépipède, une boule), on peut savoir calculer $\Pi_K(x)$ et l'algorithme de gradient projeté est envisageable. Dans le cas général, on ne sait pas calculer $\Pi_K(x)$ et il faut recourir à d'autres méthodes.

7.3 Algorithme d'Uzawa : contraintes d'égalité

On considère le cas de p contraintes d'égalité **affines**

$$K := \{x,\ Cx - f = 0\}.$$

Dans ce cas, les conditions d'optimalité s'écrivent

$$\exists \lambda \in \mathbb{R}^p, \qquad \nabla J(u) + C^T\lambda = 0 \tag{7.6a}$$

$$Cu - f = 0 \tag{7.6b}$$

On réécrit ces équations, pour un $\rho > 0$, sous la forme

$$\exists \lambda \in \mathbb{R}^p, \qquad \nabla J(u) + C^T \lambda = 0 \tag{7.7a}$$

$$\lambda = \lambda + \rho(Cu - f). \tag{7.7b}$$

Cette dernière forme suggère alors l'algorithme suivant.

Algorithme d'Uzawa (U1), contraintes d'égalité affines

On prend un multiplicateur de départ $\lambda^0 \in \mathbb{R}^p$. On fixe un pas $\rho > 0$.
Puis on itère sur $n \geq 0$:

(i) Calculer u^n t.q. $\nabla J(u^n) + C^T \lambda^n = 0$.
(ii) Calculer $\lambda^{n+1} = \lambda^n + \rho(Cu^n - f)$.

Pour que l'algorithme soit bien défini il faudra montrer l'existence d'un vecteur u^n solution de (i).

Théorème 7.2 (Cas de contraintes d'égalité affines)**.** *Soit $J : \mathbb{R}^N \to \mathbb{R}$, α-convexe, différentiable, et un ensemble de contraintes $K := \{x,\ Cx - f = 0\}$, supposé non vide.*
(i) Il existe un unique minimiseur u de J sur K.
(ii) Pour tout $\rho \in]0, \frac{2\alpha}{\|C\|^2}[$, pour tout λ^0 de départ, l'algorithme d'Uzawa (U1) converge : $\lim_{n\to\infty} u^n = u$.
(iii) Si, de plus, C est surjective et si ∇J est continue, alors on a aussi la convergence de λ^n vers un unique λ solution de (7.7a).

Preuve. (i) est classique. (ii). Commençons par vérifier l'existence de u^n. On introduit pour cela une fonction $\mathcal{L} : \mathbb{R}^N \times \mathbb{R}^p \to \mathbb{R}$, appelée lagrangien du problème, et définie par :

$$\forall (v, \mu) \in \mathbb{R}^N \times \mathbb{R}^p \quad \mathcal{L}(v, \mu) = J(v) + \langle \mu, Cv - f \rangle.$$

Supposons λ^n connu, et considérons l'application

$$v \in \mathbb{R}^N \mapsto \mathcal{L}(v, \lambda^n) = J(v) + \langle \lambda^n, Cv - f \rangle.$$

C'est une fonction strictement convexe de v ; en effet, J est strictement convexe et les contraintes étant affines, le terme $\langle \lambda^n, Cv - f \rangle$ est également une fonction affine de v et en particulier c'est une fonction convexe de v. De plus, λ^n étant fixé, la fonction $v \mapsto$

$\mathcal{L}(v, \lambda^n)$ est coercice car J est coercive, avec une croissance quadratique à l'infini (car α-convexe), et les contraintes sont des fonctions affines donc à croissance linéaire à l'infini. Ainsi $\mathcal{L}(\cdot, \lambda^n)$ possède un unique minimiseur u^n sur $\mathbb{R}^N$, caractérisé par $\nabla_v \mathcal{L}(u^n, \lambda^n) = 0$, c'est-à-dire

$$\nabla J(u^n) + C^T \lambda^n = 0.$$

Cela prouve l'existence et l'unicité de u^n.

Travaillons ensuite sur la convergence des λ^n :

$$\begin{aligned} \|\lambda^{n+1} - \lambda\|^2 &= \|(\lambda^n + \rho(Cu^n - f)) - (\lambda + \rho(Cu - f))\|^2 \\ &= \|\lambda^n - \lambda + \rho C(u^n - u)\|^2 \\ &= \|\lambda^n - \lambda\|^2 + 2\rho\langle \lambda^n - \lambda,\ C(u^n - u)\rangle + \rho^2 \|C(u^n - u)\|^2. \end{aligned}$$

On a d'une part $\|C(u^n - u)\|^2 \leq \|C\|^2\ \|u^n - u\|^2$, d'autre part, en utilisant les relations sur les gradients et l'α-convexité de J :

$$\begin{aligned} \langle \lambda^n - \lambda,\ C(u^n - u)\rangle &= \langle C^T(\lambda^n - \lambda),\ u^n - u\rangle \\ &= -\langle \nabla J(u^n) - \nabla J(u),\ u^n - u\rangle \leq -\alpha \|u^n - u\|^2. \end{aligned}$$

Ainsi,

$$\|\lambda^{n+1} - \lambda\|^2 \leq \|\lambda^n - \lambda\|^2 - \gamma \|u^n - u\|^2, \tag{7.8}$$

avec

$$\gamma := \rho(2\alpha - \rho\|C\|^2).$$

En particuliern si $0 < \rho < \frac{2\alpha}{\|C\|^2}$, on a $\gamma > 0$.

La suite $n \to \|\lambda^n - \lambda\|^2$ est alors décroissante, minorée (par 0), donc convergente vers une limite notée ℓ. Ensuite on renverse l'inégalité (7.8) pour écrire

$$\gamma \|u^n - u\|^2 \leq \|\lambda^n - \lambda\|^2 - \|\lambda^{n+1} - \lambda\|^2 \overset{n\to\infty}{\to} \ell - \ell = 0$$

Cela démontre la convergence de la suite u^n vers u, mais pas nécessairement la convergence de la suite λ^n.

(iii) C^T est alors injective ; en effet, C est surjective signifie que l'application $X \in \mathbb{R}^N \mapsto CX \in \mathbb{R}^p$ est surjective, ou encore que $rg(C) = p$; ainsi $rg(C^T) = rg(C) = p$ et donc, d'après le théorème du rang, $dim(KerC^T) + rg(C^T) = p$ d'où $dim(KerC^T) = 0$.

Par conséquent CC^T est inversible ; en effet, c'est une matrice carrée dont le noyau est réduit à 0 (si un vecteur $X \in \mathbb{R}^N$ vérifie $CC^T X = 0$, alors $\|C^T X\|^2 = \langle C^T X, C^T X\rangle = \langle X,\ CC^T X\rangle = 0$, donc $C^T X = 0$ et $X = 0$ puisque C^T est injective). En utilisant la relation $C^T\lambda^n = -\nabla J(u^n)$, on en déduit $CC^T\lambda^n = -C\nabla J(u^n)$ et donc

$$\lambda^n = -(CC^T)^{-1}C\ \nabla J(u^n).$$

Comme $u^n \to u$, par continuité de ∇J, on obtient la convergence des λ^n vers un vecteur $\lambda \in \mathbb{R}^p$. Enfin par passage à la limite dans la relation $\nabla J(u^n) + C^T\lambda^n = 0$, on en déduit que λ satisfait (7.7a) (l'unicité d'un tel λ s'obtient en écrivant comme ci-dessus, $\lambda = -(CC^T)^{-1}C\ \nabla J(u)$, ce qui définit λ de manière unique puisque u est également défini de manière unique). □

7.4 Algorithme d'Uzawa : contraintes d'inégalité

7.4.1 Contraintes d'inégalité affines

On considère le cas de p contraintes d'inégalités **affines**

$$K := \{x,\ Cx - f \leq 0\}.$$

Rappelons que si u est un point de minimum local de J sur K, et si J est différentiable en u, alors on peut écrire les conditions (KKT) sous la forme suivante :

$$\exists \lambda \in \mathbb{R}^p, \qquad \nabla J(u) + C^T\lambda = 0 \tag{7.9a}$$

$$\lambda \geq 0,\ Cu - f \leq 0,\ \langle \lambda, Cu - f\rangle = 0. \tag{7.9b}$$

Le lemme suivant permet de réécrire le deuxième jeu d'équations sur λ de manière plus compacte :

Lemme 7.3. *Soit $F := (\mathbb{R}_+)^p$, et $\rho > 0$. Pour tout $\lambda \in \mathbb{R}^p$, $C \in \mathbb{R}^{p\times N}$ et $f \in \mathbb{R}^p$, on a*

$$\Big(\lambda \geq 0,\ \ Cu - f \leq 0,\ \langle \lambda, Cu - f\rangle = 0\Big) \quad \Longleftrightarrow \quad \lambda = \Pi_F(\lambda + \rho(Cu - f)).$$

Preuve. On procède par double implication. Supposons que $\lambda \geq 0$, $Cu - f \leq 0$ et $\langle \lambda, Cu - f\rangle = 0$, et montrons que $\lambda = \Pi_F(\lambda + \rho(Cu - f))$. Comme $\lambda \geq 0$, $\lambda \in F$; il s'agit donc de montrer que pour tout $\mu \in F$,

$$\langle \lambda - \big(\lambda + \rho(Cu - f)\big), \lambda - \mu\rangle \leq 0.$$

Or,

$$\begin{aligned}\langle \lambda - \big(\lambda + \rho(Cu-f)\big), \lambda - \mu\rangle &= -\langle \rho(Cu-f), \lambda-\mu\rangle \\ &= -\rho\langle Cu-f, \lambda\rangle + \rho\langle Cu-f, \mu\rangle \leq 0\end{aligned}$$

puisque $\langle Cu-f, \lambda\rangle = 0$, $\rho > 0$, $Cu - f \leq 0$ et $\mu \geq 0$.

Réciproquement, supposons que $\lambda = \Pi_F(\lambda + \rho(Cu-f))$; alors $\lambda \geq 0$ et pour tout $\mu \geq 0$,

$$-\langle \rho(Cu-f), \lambda - \mu\rangle \leq 0 \quad \text{donc} \quad \langle Cu-f, \lambda-\mu\rangle \geq 0.$$

En prenant $\mu = 0 \in \mathbb{R}^p$ (resp. $\mu = 2\lambda$), on obtient $\langle Cu-f, \lambda\rangle \geq 0$ (resp. $\langle Cu-f, -\lambda\rangle \geq 0$), d'où $\langle Cu-f, \lambda\rangle = 0$. Enfin, d'après la formule de projection sur $F = (\mathbb{R}_+)^p$, on peut écrire pour chaque $i \in \{1, \dots, p\}$,

$$\lambda_i = \max(\lambda_i + \rho(Cu-f)_i, 0).$$

En particulier, $\lambda_i \geq \lambda_i + \rho(Cu-f)_i$ donc $(Cu-f)_i \leq 0$. Cela montre que $Cu - f \leq 0$. □

Ainsi on peut réécrire les conditions d'optimalité sous la forme suivante, pour tout $\rho > 0$:

$$\exists \lambda \in \mathbb{R}^p, \qquad \nabla J(u) + C^T\lambda = 0 \tag{7.10a}$$

$$\lambda = \Pi_F(\lambda + \rho(Cu-f)). \tag{7.10b}$$

Cela suggère alors l'algorithme suivant.

Algorithme d'Uzawa (U2), contraintes d'inégalité affines

On prend un multiplicateur de départ $\lambda^0 \in (\mathbb{R}_+)^p$. On fixe un pas $\rho > 0$.
Puis on itère sur $n \geq 0$:

(i) Calculer u^n t.q. $\nabla J(u^n) + C^T\lambda^n = 0$
(ii) Calculer $\lambda^{n+1} = \Pi_F\Big(\lambda^n + \rho(Cu^n - f)\Big)$.

Théorème 7.3 (Cas de contraintes d'inégalité affines)**.** *Soit $J : \mathbb{R}^N \to \mathbb{R}$, α-convexe, différentiable, et un ensemble de contraintes $K := \{x \in \mathbb{R}^N,\ Cx - f \leq 0\}$, supposé non vide.*
(i) Il existe un unique minimiseur u de J sur K.

(ii) Pour tout $\rho \in]0, \frac{2\alpha}{\|C\|^2}[$*, pour tout* λ^0 *de départ, l'algorithme d'Uzawa (U2) est bien défini et converge :* $\lim_{n\to\infty} u^n = u$.
(iii) Si, de plus, C *est surjective et* ∇J *est continue, alors on a aussi la convergence de la suite* λ^n *vers un unique* λ *solution de* (7.10a).

Preuve. La preuve est pratiquement identique à celle du théorème 7.2 ; la seule différence provient de la projection sur F, qui cependant n'a pas d'influence sur la convergence de u^n, comme on le remarque en écrivant :

$$\begin{aligned} \|\lambda^{n+1} - \lambda\|^2 &= \|\Pi_F(\lambda^n + \rho(Cu^n - f)) - \Pi_F(\lambda + \rho(Cu - f))\|^2 \\ &\leq \|(\lambda^n + \rho(Cu^n - f)) - (\lambda + \rho(Cu - f))\|^2 \end{aligned}$$

(puisque la projection Π_F est une application 1-lipschitzienne). □

7.4.2 Contraintes d'inégalité convexes

On considère maintenant le cas de contraintes de la forme

$$K := \{x \in \mathbb{R}^N,\ \varphi_i(x) \leq 0, \quad 1 \leq i \leq p\},$$

où chaque contrainte $\varphi_i : \mathbb{R}^N \to \mathbb{R}$ est supposée **convexe**. On note $\varphi(x) = (\varphi_1(x), \ldots, \varphi_p(x))^T$, de sorte que K s'écrive aussi $\{x, \varphi(x) \leq 0\}$.

Rappelons que si u est un point de minimum local de J sur K, avec J, φ_i différentiables en u, et si les contraintes sont qualifiées en u, alors on peut écrire les conditions (KKT) sous la forme suivante :

$$\exists \lambda \in \mathbb{R}^p, \qquad \nabla J(u) + \sum_{i=1}^{p} \lambda_i \nabla \varphi_i(u) = 0 \tag{7.11a}$$

$$\lambda \geq 0,\ \varphi(u) \leq 0,\ \langle \lambda, \varphi(u) \rangle = 0. \tag{7.11b}$$

Le lemme suivant permet de réécrire le deuxième jeu d'équations sur λ de manière plus compacte :

Lemme 7.4. *Soit* $F := (\mathbb{R}_+)^p$*, et* $\rho > 0$*. Pour tout* $\lambda \in \mathbb{R}^p$ *et* $\varphi(u) \in \mathbb{R}^p$*, on a*

$$\Big(\lambda \geq 0,\ \varphi(u) \leq 0,\ \langle \lambda, \varphi(u) \rangle = 0\Big) \quad \Longleftrightarrow \quad \lambda = \Pi_F(\lambda + \rho\varphi(u)).$$

Preuve. La preuve est identique à celle du lemme 7.3, où $Cu - f$ est remplacé par $\varphi(u)$. □

Ainsi on peut réécrire les conditions d'optimalité sous la forme suivante, pour tout $\rho > 0$:

$$\exists \lambda \in \mathbb{R}^p, \qquad \nabla J(u) + \sum_{i=1}^{p} \lambda_i \nabla \varphi_i(u) = 0 \tag{7.12a}$$

$$\lambda = \Pi_F(\lambda + \rho\varphi(u)). \tag{7.12b}$$

Ceci suggère alors l'algorithme suivant.

Algorithme d'Uzawa (U2), contraintes d'inégalités convexes

On prend un multiplicateur de départ $\lambda^0 \in (\mathbb{R}_+)^p$. On fixe un pas $\rho > 0$.
Puis on itère sur $n \geq 0$:

(i) Calculer u^n t.q. $\nabla J(u^n) + \sum_{i=1}^{p} \lambda_i^n \nabla \varphi_i(u^n) = 0$
(ii) Calculer $\lambda^{n+1} = \Pi_F\Big(\lambda^n + \rho\varphi(u^n)\Big)$.

Pour que l'algorithme soit bien défini il faudra montrer l'existence d'un vecteur u^n solution de (i).

Théorème 7.4 (Cas de contraintes d'inégalité convexes)**.** *Soit $J : \mathbb{R}^N \to \mathbb{R}$, α-convexe, différentiable, et un ensemble de contraintes $K := \{x \in \mathbb{R}^N,\ \varphi_i(x) \leq 0,\ i = 1, \ldots, p\}$ avec φ_i **convexes**, différentiables. On suppose de plus que l'application $x \in \mathbb{R}^N \mapsto \varphi(x) := (\varphi_1(x), \ldots, \varphi_p(x))^T$ est M-lipschitzienne. On suppose enfin que les contraintes sont qualifiées au point u.*
(i) Il existe un unique minimiseur u de J sur K.
(ii) Pour tout $\rho \in]0, \frac{2\alpha}{M^2}[$, pour tout λ^0 de départ, l'algorithme d'Uzawa (U2) est bien défini et converge : $\lim u^n = u$.
(iii) Si, de plus, la matrice $C(u) = [\nabla\varphi_1(u), \ldots, \nabla\varphi_p(u)]^T$ est surjective, et si J et φ sont de classe C^1, alors on a aussi la convergence de la suite λ^n.

Preuve. (i) est classique.
(ii). Commençons par vérifier l'existence de u^n. On note que $\mathcal{L}(v, \lambda^n) = J(v) + \sum_i \lambda_i^n \varphi_i(v)$ est une fonction strictement convexe de v, car J est strictement convexe, les contraintes φ_i sont convexes et les λ_i^n sont positifs. Elle est coercice car J l'est, avec une croissance

quadratique à l'infini (car α-convexe), et les φ_i sont à croissance au plus linéaire à l'infini (car Lipschitz). Ainsi $\mathcal{L}(.,\lambda^n)$ possède un unique minimiseur u^n sur $\mathbb{R}^N$, caractérisé par $\nabla_v \mathcal{L}(u^n,\lambda^n)=0$, soit

$$\nabla J(u^n)+\sum_{i=1}^{p}\lambda_i^n\nabla\varphi_i(u^n)=0.$$

(Ce qui prouve donc l'existence et l'unicité de u^n.)

Remarquons que par définition du minimiseur u^n,

$$\forall v\in\mathbb{R}^N \quad \mathcal{L}(u^n,\lambda^n)\leq\mathcal{L}(v,\lambda^n),$$

c'est-à-dire

$$\forall v\in\mathbb{R}^N \quad J(u^n)+\langle\lambda^n,\varphi(u^n)\rangle\leq J(v)+\langle\lambda^n,\varphi(v)\rangle. \tag{7.13}$$

Soit $w\in\mathbb{R}^N$ et $t\in]0,1[$; en appliquant (7.13) avec $v=u^n+t(w-u^n)$, on obtient

$$J(u^n+t(w-u^n))-J(u^n)+\sum_{i=1}^{p}\lambda_i^n\left(\varphi_i(u^n+t(w-u^n))-\varphi_i(u^n)\right)\geq 0. \tag{7.14}$$

Mais par convexité des φ_i,

$$\varphi_i(u^n+t(w-u^n))-\varphi_i(u^n)\leq t(\varphi_i(w)-\varphi_i(u^n)).$$

D'après (7.14), on en déduit :

$$J(u^n+t(w-u^n))-J(u^n)+t\sum_{i=1}^{p}\lambda_i^n\left(\varphi_i(w)-\varphi_i(u^n)\right)\geq 0.$$

En divisant par t et en passant à la limite quand $t\to 0$, on obtient :

$$\forall w\in\mathbb{R}^N \quad \langle\nabla J(u^n),w-u^n\rangle+\sum_{i=1}^{p}\lambda_i^n\left(\varphi_i(w)-\varphi_i(u^n)\right)\geq 0. \tag{7.15}$$

Considérons à présent le point u et le vecteur λ ; d'après les conditions (KKT), ils vérifient la relation

$$\nabla J(u)+\sum_{i=1}^{p}\lambda_i\nabla\varphi_i(u)=0. \tag{7.16}$$

En définissant comme plus haut le lagrangien $\mathcal{L}(v,\lambda) = J(v) + \sum_i \lambda_i \varphi_i(v)$, on constate que la relation (7.16) s'écrit $\nabla_v \mathcal{L}(u,\lambda) = 0$, ce qui montre que u est le minimiseur unique de l'application $v \mapsto \mathcal{L}(v,\lambda)$, sur $\mathbb{R}^N$ (l'existence et l'unicité d'un tel minimiseur s'obtiennent par les mêmes arguments que pour l'application $v \mapsto \mathcal{L}(v,\lambda^n)$). On a donc :

$$\forall v \in \mathbb{R}^N \quad \mathcal{L}(u,\lambda) \leq \mathcal{L}(v,\lambda).$$

En appliquant le même raisonnement que précédemment, on en déduit

$$\forall w \in \mathbb{R}^N \quad \langle \nabla J(u), w-u\rangle + \sum_{i=1}^{p} \lambda_i \left(\varphi_i(w) - \varphi_i(u)\right) \geq 0. \tag{7.17}$$

En prenant $w = u$ dans (7.15), $w = u^n$ dans (7.17), on obtient

$$\forall w \in \mathbb{R}^N \quad \langle \nabla J(u^n), u-u^n\rangle + \sum_{i=1}^{p} \lambda_i^n \left(\varphi_i(u) - \varphi_i(u^n)\right) \geq 0.$$

$$\forall w \in \mathbb{R}^N \quad \langle \nabla J(u), u^n-u\rangle + \sum_{i=1}^{p} \lambda_i \left(\varphi_i(u^n) - \varphi_i(u)\right) \geq 0.$$

En sommant, on en déduit

$$\langle \nabla J(u) - \nabla J(u^n), u^n-u\rangle + \sum_{i=1}^{p} (\lambda_i - \lambda_i^n)(\varphi_i(u^n) - \varphi_i(u)) \geq 0.$$

En utilisant l'α-convexité de J, on en déduit finalement

$$\begin{aligned} \langle \lambda^n - \lambda,\ \varphi(u^n) - \varphi(u)\rangle &\leq -\langle \nabla J(u^n) - \nabla J(u), u^n - u\rangle \\ &\leq -\alpha \|u^n - u\|^2. \end{aligned} \tag{7.18}$$

Nous allons utiliser l'estimation (7.18) pour démontrer la convergence de la suite $\|\lambda^n - \lambda\|$. Pour cela, on écrit :

$$\begin{aligned} \|\lambda^{n+1} - \lambda\|^2 &= \left\|\Pi_F\left(\lambda^n + \rho\varphi(u^n)\right) - \Pi_F\left(\lambda + \rho\varphi(u)\right)\right\|^2 \\ &\leq \|(\lambda^n + \rho\varphi(u^n)) - (\lambda + \rho\varphi(u))\|^2 \\ &\leq \|(\lambda^n - \lambda) + \rho(\varphi(u^n) - \varphi(u))\|^2 \\ &\leq \|\lambda^n - \lambda\|^2 + 2\rho\langle \lambda^n - \lambda, \varphi(u^n) - \varphi(u)\rangle + \rho^2 M^2 \|u^n - u\|^2 \\ &\leq \|\lambda^n - \lambda\|^2 - 2\alpha\rho\|u^n - u\|^2 + \rho^2 M^2 \|u^n - u\|^2 \\ &\leq \|\lambda^n - \lambda\|^2 - \gamma\|u^n - u\|^2 \end{aligned}$$

en utilisant le caractère Lipschitz de φ et l'inégalité (7.18) avec

$$\gamma := \rho(2\alpha - \rho M^2).$$

On conclut alors exactement comme pour la convergence de l'algorithme (U1). On obtient donc la convergence de la suite u^n, mais pas nécessairement celle de la suite λ^n.

(*iii*) Si les fonctions φ_i sont de classe C^1, alors l'application $v \in \mathbb{R}^N \mapsto C(v)$ est continue (rappelons que la matrice $C(v)$ est définie par $C(v) = [\nabla\varphi_1(v), \ldots, \nabla\varphi_p(v)]^T$). De plus (voir la preuve de la convergence de l'algorithme d'Usawa pour le cas des contraintes d'égalité affines), $C(u)$ étant surjective, la matrice $C(u)^T$ est injective, et dans ce cas, la matrice $C(u)C(u)^T$ est inversible. Par conséquent, son déterminant est non nul. Par continuité du déterminant et de l'application $v \in \mathbb{R}^N \mapsto C(v)C(v)^T$, puisque $\det(C(u)C(u)^T) \neq 0$, la convergence de u^n vers u entraîne que pour n assez grand, on a également $\det(C(u^n)C(u^n)^T) \neq 0$, c'est-à-dire que la matrice $C(u^n)C(u^n)^T$ est inversible.

Or, la relation

$$\nabla J(u^n) + \sum_{i=1}^{p} \lambda_i^n \nabla\varphi_i(u^n) = 0$$

s'écrit

$$\nabla J(u^n) + C(u^n)^T\lambda^n = 0,$$

d'où

$$C(u^n)\nabla J(u^n) + C(u^n)C(u^n)^T\lambda^n = 0.$$

Pour n assez grand, la matrice $C(u^n)C(u^n)^T$ est inversible, ce qui permet d'exprimer λ^n sous la forme suivante :

$$\lambda^n = -(C(u^n)C(u^n)^T)^{-1}C(u^n)\nabla J(u^n).$$

Comme $u^n \to u$, on en déduit par continuité la convergence des λ^n vers une limite λ^* qui s'écrit

$$\lambda^* = -(C(u)C(u)^T)^{-1}C(u)\nabla J(u).$$

En passant à la limite dans la relation

$$\nabla J(u^n) + C(u^n)^T\lambda^n = 0,$$

on obtient (par continuité des dérivées partielles de J)

$$\nabla J(u) + C(u)^T \lambda^* = 0.$$

Ainsi λ^* est solution de (7.12a) ; c'est même l'unique solution du système, puisque par injectivité de $C(u)^T$, si un vecteur μ^* satisfait $\nabla J(u) + C(u)^T \mu^* = 0$, alors $C(u)^T \mu^* = C(u)^T \lambda^*$ et donc $\mu^* = \lambda^*$.

□

7.5 Méthode de pénalisation

On considère une fonctionnelle J, continue, coercive sur un ensemble de contraintes K, où $K \subset \mathbb{R}^N$ est supposé fermé et non vide. On considère le problème d'optimisation sous contrainte suivant :

$$\inf_{v \in K} J(v) \tag{7.19}$$

On introduit une fonction $\varphi : \mathbb{R}^N \to \mathbb{R}$, continue, telle que, pour tout v dans $\mathbb{R}^N$:

$$\varphi(v) \geq 0,$$

et

$$\varphi(v) = 0 \quad \Leftrightarrow \quad v \in K.$$

On considère enfin pour tout $n \in \mathbb{N}^*$ et $v \in \mathbb{R}^N$, la fonctionnelle

$$J_n(v) := J(v) + n\varphi(v), \tag{7.20}$$

et le problème pénalisé correspondant, sur $\mathbb{R}^N$:

$$\inf_{v \in \mathbb{R}^N} J_n(v). \tag{7.21}$$

L'avantage du problème pénalisé (7.21) est qu'il s'agit d'un problème de minimisation sans contrainte (posé sur $\mathbb{R}^N$).

Exemples type de pénalisations :

1. Pour $K := \{v,\ Cv - f = 0\}$, $\varphi(v) := \|Cv - f\|^2$.
2. Pour $K := \{v,\ Cv - f \leq 0\}$, $\varphi(v) := \|\max(Cv - f, 0)\|^2$.

On a alors le résultat suivant.

Théorème 7.5. *On suppose que* (7.19) *admet un unique minimiseur noté* u*. On suppose que pour tout* $n \in \mathbb{N}^*$*,* u_n *est un point de minimum du problème* (7.21) *sur* $\mathbb{R}^N$*. Alors*

$$\lim_{n\to\infty} u_n = u.$$

Preuve. Comme u_n est un minimiseur pour (7.21) sur $\mathbb{R}^N$, on a en particulier $J_n(u_n) \leq J_n(u)$, c'est-à-dire

$$J(u_n) + n\varphi(u_n) \leq J(u) \tag{7.22}$$

(puisque $u \in K$ on a $\varphi(u) = 0$). Comme $\varphi \geq 0$, on en déduit que $J(u_n) \leq J(u)$ et donc que $J(u_n)$ est une suite bornée. Comme J est coercive, cela implique que u_n est également une suite bornée.

Pour démontrer la convergence de la suite (u_n), on va procéder indirectement en considérant des sous-suites convergentes. On rappelle pour cela un lemme de topologie :

Lemme 7.5. *Soit* (v_n) *une suite à valeurs dans* $\mathbb{R}^N$*, telle que de toute suite extraite, on puisse extraire une sous-suite convergente vers une même limite* $v \in \mathbb{R}^N$*. Alors toute la suite* (v_n) *est convergente, et de limite* v*.*

Soit (u_{n_k}) une sous-suite de (u_n). Comme u_{n_k} est bornée dans $\mathbb{R}^N$, on peut à nouveau en extraire une sous-suite convergente, vers un $v \in \mathbb{R}^N$. (On note encore u_{n_k} cette sous-suite.) Tout d'abord

$$J(u_{n_k}) \leq J(u),$$

donc par continuité de J et en passant à la limite,

$$J(v) \leq J(u).$$

D'autre part, on a aussi d'après (7.22),

$$\varphi(u_{n_k}) \leq \frac{1}{n_k}\ (J(u) - J(u_{n_k})) \leq \frac{C}{n_k}.$$

Ainsi, à la limite,

$$\varphi(v) \leq 0.$$

On en déduit donc que $\varphi(v) = 0$, c'est-à-dire $v \in K$. Comme $J(v) \leq J(u)$, par unicité du minimiseur pour le problème (7.19), cela implique que $v = u$. D'apres le lemme précédent, on conclut donc que tout la suite (u_n) converge vers u. □

Estimation d'erreur. On peut dans certains cas estimer l'erreur faite sur le problème pénalisé, en fonction de n ; par exemple, dans le cas de contraintes d'égalité $K := \{v,\ Cv-f = 0\}$. On considère $\varphi(v) = \frac{1}{2}\|Cv-f\|^2$; un calcul donne $\nabla\varphi(v) = C^T(Cv-f)$. Supposons J différentiable. La condition d'optimalité pour u_n s'écrit alors

$$\nabla J(u_n) + nC^T(Cu_n - f) = 0.$$

De plus $Cu - f = 0$ donc on a également $C^T(Cu - f) = 0$. Ainsi, on obtient

$$C^TC(u_n - u) = -\frac{1}{n}\nabla J(u_n).$$

Si C^TC est inversible (ce qui est équivalent à supposer C injective), alors après multiplication par $(C^TC)^{-1}$ on obtient

$$\|u_n - u\| \leq \frac{1}{n}\ \|(C^TC)^{-1}\|\ \|\nabla J(u_n)\| \leq \frac{c_0}{n},$$

où c_0 est une constante ($\|\nabla J(u_n)\|$ est bornée car u_n est une suite bornée).