

---

# Optimisation Continue

---

Jean-Philippe Préaux

---



# Table des matières

<b>Introduction</b>	<b>7</b>
1 Formulation . . . . .	9
1.1 Problème d'optimisation ; maximum et minimum . . . . .	9
1.2 Problème d'optimisation continue . . . . .	10
1.3 Extremum local . . . . .	11
2 Exemples de problèmes d'optimisation à une variable . . . . .	13
2.1 Minimisation des coûts dans la fabrication de boîtes cylindriques . .	13
2.2 Position d'équilibre d'un système de deux ressorts. . . . .	14
3 Problèmes d'optimisation sur plusieurs variables . . . . .	15
3.1 Production optimale d'une fonderie . . . . .	15
3.2 Problème de transport . . . . .	16
3.3 Régression linéaire . . . . .	17
3.4 Modélisation de données expérimentales . . . . .	17
<b>I Programmation linéaire</b>	<b>19</b>
I.1 Préliminaires . . . . .	19
I.1.1 Formulation . . . . .	19
I.1.2 Représentation matricielle . . . . .	20
I.1.3 Forme canonique . . . . .	20
I.1.4 Exemple de problème à deux variables - Résolution graphique . . . .	21
I.1.5 Généralisation . . . . .	21
I.2 Méthode du simplexe . . . . .	24
I.2.1 Problème de programmation linéaire sous forme normale . . . . .	24
I.2.2 Algorithme du simplexe I : préparation . . . . .	24
I.3 Résolution dans le cas général . . . . .	28
I.3.1 Ecrire un problème de maximisation sous forme normale . . . . .	28
I.3.2 Dualité minimum/maximum . . . . .	28
I.4 Programmation linéaire en nombres entiers . . . . .	30
Exercices. . . . .	32
<b>II Généralités sur l'optimisation</b>	<b>33</b>
II.1 Conditions suffisantes d'existence d'extrema globaux . . . . .	34
II.1.1 Compacité du domaine . . . . .	34

II.1.2 Applications coercives . . . . .	34
II.2 Recherche d'extrema locaux. . . . .	36
II.2.1 Condition nécessaire du 1 <sup>er</sup> ordre . . . . .	36
II.2.2 Conditions du second ordre . . . . .	37
II.3 Programmation convexe . . . . .	39
II.3.1 Applications convexes, strictement convexes . . . . .	39
II.3.2 Programmation convexe . . . . .	42
II.3.3 Applications elliptiques . . . . .	43
II.3.4 Programmation elliptique . . . . .	44
II.4 Programmation quadratique sans contraintes . . . . .	46
II.4.1 Applications quadratiques . . . . .	46
II.4.2 Programmation quadratique . . . . .	47
Exercices . . . . .	49
<b>III Programmation sous contraintes</b>	<b>51</b>
III.1 Optimisation sous contraintes égalitaires . . . . .	51
III.1.1 Enoncé du problème . . . . .	51
III.1.2 Exemples en dimension 2. . . . .	52
III.1.3 Principe de Lagrange . . . . .	54
III.1.4 Prise en compte de la convexité . . . . .	56
III.1.5 Conditions, nécessaire, suffisante, du second ordre . . . . .	57
III.1.6 Programmation quadratique sous contraintes égalitaires . . . . .	59
III.2 Optimisation sous contraintes : le cas général . . . . .	62
III.2.1 Conditions de Karush-Kuhn-Tucker . . . . .	62
III.2.2 Prise en compte de la convexité . . . . .	64
III.2.3 Qualification de contraintes affines et convexes . . . . .	65
III.2.4 Programmation quadratique sous contraintes . . . . .	66
III.2.5 Conditions nécessaire, suffisante, du second ordre . . . . .	67
III.2.6 Points-selles du Lagrangien : introduction à la dualité . . . . .	71
Exercices . . . . .	75
<b>IV Algorithmes itératifs</b>	<b>77</b>
IV.1 Méthodes itératives dans le cas sans contraintes . . . . .	79
IV.1.1 Méthode de Newton . . . . .	79
IV.1.2 Méthode de relaxation . . . . .	81
IV.1.3 Méthode de gradient à pas optimal . . . . .	83
IV.1.4 Méthode du gradient à pas fixe . . . . .	85
IV.1.5 Méthode du gradient conjugué . . . . .	86
IV.2 Méthodes itératives dans le cas sous contraintes . . . . .	89
IV.2.1 Méthode de relaxation sur un domaine produit d'intervalles . . . . .	91
IV.2.2 Méthode du gradient projeté . . . . .	92
IV.2.3 Méthode d'Uzawa . . . . .	93
Exercices . . . . .	95

<b>V Applications aux Maths numériques</b>	<b>97</b>
V.1 Résolution approchée d'un système d'équations . . . . .	98
V.1.1 Système d'équations linéaires de Cramer . . . . .	98
V.1.2 Système d'équations linéaires à matrice symétrique définie positive . . . . .	99
V.1.3 Inversion d'une matrice symétrique définie positive . . . . .	101
V.1.4 Résolution approchée d'un système d'équations non linéaires . . . . .	103
V.2 Approximation d'un nuage de points . . . . .	104
V.2.1 Approximation linéaire au sens des moindres carrés . . . . .	105
V.2.2 Exemple important : la droite de régression linéaire . . . . .	106
V.2.3 Exemple important : le polynôme d'interpolation de Lagrange . . . . .	107
V.2.4 Approximation minimax . . . . .	108
V.2.5 Approximation minimax linéaire . . . . .	108
Exercices . . . . .	110
<b>A Rappels de pré-requis Mathématiques</b>	<b>111</b>
A.1 Rappels d'analyse . . . . .	111
A.1.1 L'espace euclidien $\mathbb{R}^n$ . . . . .	111
A.1.2 Normes de $\mathbb{R}^n$ . . . . .	111
A.1.3 Topologie de $\mathbb{R}^n$ . . . . .	112
A.2 Rappels de calcul différentiel . . . . .	113
A.2.1 Applications différentiables . . . . .	113
A.2.2 Vecteur gradient . . . . .	113
A.2.3 Matrice hessienne . . . . .	113
A.2.4 Développements de Taylor . . . . .	114
A.2.5 Espace tangent . . . . .	114
A.3 Rappels sur les matrices . . . . .	116
A.3.1 Notations . . . . .	116
A.3.2 Norme matricielle . . . . .	116
A.3.3 Matrice (semi-)définie positive/négative . . . . .	117
<b>Correction des exercices</b>	<b>119</b>



# Introduction

L'optimisation est une discipline mathématique qui, bien qu'omniprésente depuis les origines, a pleinement pris son essor au cours du XX<sup>e</sup> siècle d'une part sous la stimulation du développement des sciences de l'industrie et de la planification, telles l'économie, la gestion, *etc.*, et des sciences appliquées aux technologies naissantes, comme l'automatique, le traitement du signal, *etc.*, et d'autre part grâce au développement de l'informatique qui a rendu efficiente ses méthodes algorithmiques jusque là impraticables.

Optimiser c'est choisir parmi plusieurs possibilités celle qui répond le mieux à certains critères. En ce sens il n'est pas de science ni même de domaine d'activité qui ne soit confronté à un problème d'optimisation. L'optimisation, et plus généralement la *Recherche opérationnelle*, intervient dès-lors pour appliquer l'outil mathématique à cette résolution, si tant est que le problème soit formalisable mathématiquement. De nos jours son champ d'application est on ne peut plus vaste : optimisation des ressources, des gains, des coûts dans l'industrie, optimisation du trafic aérien, ferroviaire, routier, dans le transport, optimisation de la couverture radar, de la réactivité d'intervention, de la gestion des stocks et des troupes dans le domaine militaire, *etc.*, sans parler des sciences dures, physique, chimie, informatique, automatique, traitement du signal, *etc.*, pour lesquels nombre de problèmes se ramènent et se résolvent par optimisation. C'est une discipline fondamentale dans les sciences de l'ingénieur, de l'économie et de la gestion, pour ne citer qu'elles.

Les premiers problèmes d'optimisation auraient été formulés par le mathématicien Euclide, au III<sup>e</sup> siècle av. J.C. dans *Les Eléments*. Trois siècles plus tard Héron d'Alexandrie énonce le principe du plus court chemin en optique. Au XVII<sup>e</sup> siècle l'apparition du calcul différentiel sous l'égide de Newton et de Leibnitz, et la théorie newtonienne de la mécanique entraînent l'invention des premières techniques d'otimisation, dont la méthode itérative de Newton pour chercher les extrema locaux d'une fonction. Durant le XVIII<sup>e</sup> siècle Euler et Lagrange développent le *calcul variationnel*, branche de l'analyse fonctionnelle dont le but est de trouver une application répondant au mieux à certains critères. Ce dernier invente une technique fondamentale en optimisation connue aujourd'hui sous le nom de *multiplicateurs de Lagrange*. Au XIX<sup>e</sup> siècle l'industrialisation en europe voit les économistes présenter un intérêt croissant pour les mathématiques et mettre en place des modèles économiques qu'il convient alors d'optimiser.

Au XX<sup>e</sup> siècle ce furent des aspects contrastés qui convergèrent vers le développement de l'optimisation, ou encore de la *programmation mathématique* et de la recherche opérationnelle. En Union Soviétique la planification fut une conséquence de la pensée commu-

niste et se concrétisa par des plan quinquénaux ou encore *gosplans*, tandis qu'aux Etats-Unis le développement du capitalisme accoucha de la recherche opérationnelle. Mais c'est avec l'apparition de l'informatique dans l'après-guerre que les techniques d'optimisation prirent toute leur ampleur et s'appliquèrent dans tous les champs d'activité.

L'un des premiers succès fût la *méthode du simplexe* s'appliquant en *programmation linéaire*, qui fut inventée en 1947 par le mathématicien américain Georges Dantzig. De par son efficacité pratique elle est devenue l'un des algorithmes les plus utilisés de l'histoire des mathématiques appliquées. Dantzig travaillait alors comme conseiller pour l'US air force sur la mécanisation des processus de planification, dans le but de les résoudre à l'aide de machines à cartes perforées. Notons d'ailleurs que le terme de *programmation* (mathématiques), synonyme d'optimisation, n'a rien à voir avec le sens qu'on lui donne en informatique, mais provient en fait du jargon militaire où il signifie *planification*.

C'est quelques années auparavant, peu avant la seconde guerre mondiale, que la programmation linéaire avait été développée par Leonid Kantorovich, professeur de mathématiques à l'université de Leningrad, qui avait été chargé par le gouvernement soviétique en 1938 d'optimiser la production industrielle de contreplaqué. Il y trouva des possibilités d'optimisation de la production économique soviétique. Il effectua par ailleurs de nombreux travaux en optimisation continue, dont des conditions de convergence pour la méthode de Newton. Ses théories ne furent publiées qu'après l'ère stalinienne ; il faillit être emprisonné deux fois et ne fut sauvé que pour son implication dans le programme nucléaire soviétique ; en effet ses travaux l'avaient conduit indirectement à réintroduire la théorie de l'utilité marginale qui s'oppose à la théorie économique marxiste ; ils ont trouvé leurs applications quelques années plus tard dans la libéralisation de l'économie soviétique. Conjointement avec T.Koopmans il obtint le prix nobel d'économie en 1975 "for their contributions to the theory of optimum allocation of resources"<sup>1</sup>.

De nos jours l'optimisation et plus généralement la recherche opérationnelle, reste un domaine fécond de la recherche en mathématiques qui bénéficie d'importants financements provenant aussi bien du domaine public que du domaine privé, et dont les retombées s'appliquent dans tous les domaines d'activité humaine se prêtant à la modélisation mathématique.

---

1. *Traduction* : "pour leurs contributions à la théorie de l'allocation optimale des ressources".



# 1 Formulation

## 1.1 Problème d'optimisation ; maximum et minimum

Soit  $n$  un entier strictement positif et soient :

$\mathcal{D} \subset \mathbb{R}^n$  un sous-ensemble non vide de  $\mathbb{R}^n$ , et  
 $f : \mathcal{D} \longrightarrow \mathbb{R}$  une application sur  $\mathcal{D}$  à valeurs réelles .

Un problème d'optimisation consiste à déterminer, lorsqu'il existe, un extremum, minimum ou maximum, de  $f$  sur  $\mathcal{D}$ . On note un tel problème :

$$\min_{\mathbf{x} \in \mathcal{D}} f(\mathbf{x}) \quad \text{ou} \quad \max_{\mathbf{x} \in \mathcal{D}} f(\mathbf{x}) .$$

Plus précisément :

- un minimum (ou minimum global)  $\mathbf{u}$  de  $f$  sur  $\mathcal{D}$  est un point  $\mathbf{u} \in \mathcal{D}$ , tel que  $\forall \mathbf{x} \in \mathcal{D}$ ,  $f(\mathbf{u}) \leq f(\mathbf{x})$ ,
- un maximum (ou maximum global)  $\mathbf{u}$  de  $f$  sur  $\mathcal{D}$  est un point  $\mathbf{u} \in \mathcal{D}$ , tel que  $\forall \mathbf{x} \in \mathcal{D}$ ,  $f(\mathbf{u}) \geq f(\mathbf{x})$ .

Lorsque l'inégalité est stricte  $\forall \mathbf{x} \in \mathcal{D} \setminus \{\mathbf{u}\}$  on parlera de minimum ou de maximum strict.

- La valeur  $f(\mathbf{u})$  prise par  $f$  en un minimum (resp. maximum) est sa valeur minimale (resp. maximale) et sera usuellement notée  $f_{\min}$  (resp.  $f_{\max}$ ).
- L'ensemble  $\mathcal{D}$  est appelé le domaine admissible, et la fonction  $f$  à minimiser la fonction coût, ou à maximiser la fonction objectif (ou fonction économique, *etc...*).

Un minimum (resp. maximum) de  $f$  est un maximum (resp. minimum) de  $-f$  et réciproquement, tandis la valeur minimale (resp. maximale) de  $f$  est l'opposé de la valeur maximale (resp. minimale) de  $-f$ . Pour cette raison on peut changer tout problème de minimisation en un problème de maximisation équivalent, et réciproquement.

L'optimisation se scinde essentiellement en deux disciplines dont les outils et méthodes sont très disparates :

Si  $\mathcal{D}$  est discret ( $\mathcal{D} \subset \mathbb{Z}^n$ , fini ou dénombrable), on parle d'**optimisation combinatoire**. Les outils proviennent essentiellement des mathématiques discrètes (théorie des graphes).

Si  $\mathcal{D}$  est continu, et  $f$  est continue, on parle d'**optimisation continue**. Les outils proviennent essentiellement de l'analyse (calcul différentiel, convexité) et de l'algèbre linéaire.

L'optimisation continue est des deux domaines probablement le plus "facile" car les outils d'analyse (comme les dérivées) sont des concepts puissants qui y sont fort utiles, tant du point de vue théorique que du point de vue algorithmique.

*Ce cours ne traitera que de l'optimisation continue.*

## 1.2 Problème d'optimisation continue

Sous la forme énoncée, la classe des problèmes d'optimisation continue est bien trop large pour espérer obtenir une méthode de résolution générale efficace. Aussi restreint-on cette classe de problèmes à des sous-classes, où des hypothèses restrictives permettent d'y établir des méthodes de résolution spécifiques. De telles hypothèses doivent être suffisamment fortes pour y établir des méthodes utilisables en pratique, et suffisamment faibles pour englober une large classe de problèmes.

En optimisation continue, dans la plupart des cas le domaine admissible  $\mathcal{D}$  est donné sous la forme (restrictive) suivante : soit  $\mathcal{U}$  un ouvert de  $\mathbb{R}^n$ ,

$$\mathcal{D} = \left\{ (x_1, x_2, \dots, x_n) \in \mathcal{U} \subset \mathbb{R}^n \mid \underbrace{\varphi_i(x_1, \dots, x_n) \leq 0, \quad i = 1, \dots, p}_{\text{contraintes inégalitaires}}, \underbrace{\psi_j(x_1, \dots, x_n) = 0, \quad j = 1, \dots, q}_{\text{contraintes égalitaires}} \right\}$$

Les applications  $\varphi_i, \psi_j$  sont appelées les applications contraintes et sont supposées non constantes ; les premières étant qualifiées d'inégalitaires et les dernières d'égalitaires.

On se restreint à des sous-classes de problèmes en posant des hypothèses sur les applications  $f, \varphi_i, \psi_j$ .

On parle de :

- **Programmation linéaire** : lorsque  $f, \varphi_1, \dots, \varphi_p, \psi_1, \dots, \psi_q$  sont des applications affines<sup>2</sup> et  $\mathcal{U} = \mathbb{R}^n$ .
- **Programmation quadratique** : lorsque  $f$  est une application quadratique,  $\varphi_1, \dots, \varphi_p, \psi_1, \dots, \psi_q$  sont des applications affines et  $\mathcal{U} = \mathbb{R}^n$ .
- **Programmation convexe** : problème de minimisation lorsque  $f$  et  $\varphi_1, \dots, \varphi_p$  sont des applications convexes,  $\psi_1, \dots, \psi_q$  sont des applications affines, et  $\mathcal{U}$  est convexe.

Dans ce cadre on verra comment établir des méthodes générales et des algorithmes pour les résoudre.

---

2. Rappelons qu'une application  $\varphi$  est affine s'il existe une application constante  $\psi$  telle que  $\phi - \psi$  soit linéaire.

### 1.3 Extremum local

Afin de rester général, on accordera une grande importance à la différentiabilité des applications considérées (à un ordre suffisant) qui procure des outils puissants et dans de nombreux cas des calculs efficaces pour donner des conditions nécessaires, suffisantes, d'existence d'*extrema locaux*. Cependant ces notions étant locales elles ne procurent une information que localement ; mais alliées à d'autres considérations (compacité, coercivité, convexité) elles peuvent se révéler fort utiles pour la recherche d'extrema.

- Un point  $\mathbf{u} \in \mathcal{D} \subset \mathbb{R}^n$  est un minimum local de  $f$  sur  $\mathcal{D}$  si il existe un voisinage  $\mathcal{V}(\mathbf{u})$  de  $\mathbf{u}$  dans  $\mathbb{R}^n$ , tel que  $\forall \mathbf{x} \in \mathcal{V}(\mathbf{u}) \cap \mathcal{D}, f(\mathbf{u}) \leq f(\mathbf{x})$ .
- Un point  $\mathbf{u} \in \mathcal{D} \subset \mathbb{R}^n$  est un maximum local de  $f$  sur  $\mathcal{D}$  si il existe un voisinage  $\mathcal{V}(\mathbf{u})$  de  $\mathbf{u}$  dans  $\mathbb{R}^n$ , tel que  $\forall \mathbf{x} \in \mathcal{V}(\mathbf{u}) \cap \mathcal{D}, f(\mathbf{u}) \geq f(\mathbf{x})$ .
- Lorsque les inégalités sont strictes  $\forall \mathbf{x} \in \mathcal{V}(\mathbf{u}) \cap \mathcal{D} \setminus \{\mathbf{u}\}$  on parle de minimum local ou de maximum local strict.

Clairement **tout extremum global est aussi un extremum local** (prendre  $\mathcal{V}(\mathbf{u}) = \mathbb{R}^n$ ). La réciproque est évidemment fausse, comme le montre l'exemple de la figure 1.

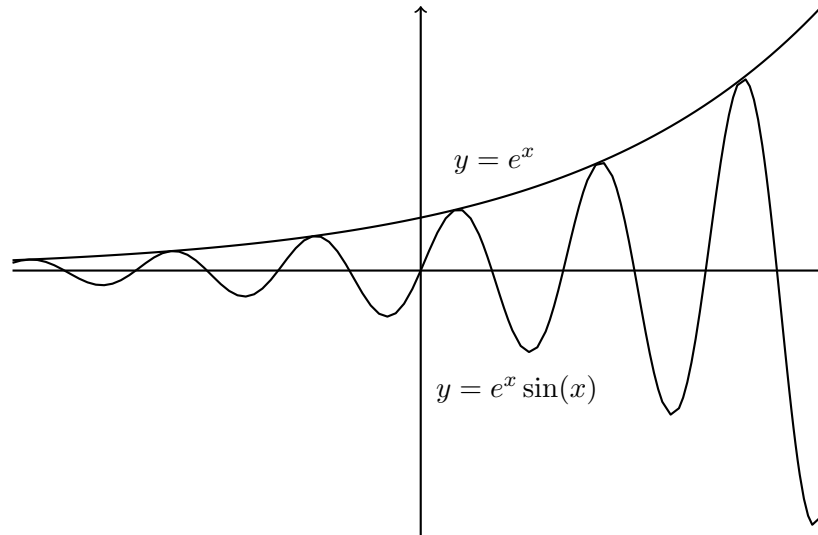


FIGURE 1 – L'application  $x \rightarrow e^x \sin(x)$  a une infinité de minima locaux (en  $-\frac{\pi}{4} [2\pi]$ ) et de maxima locaux (en  $\frac{3\pi}{4} [2\pi]$ ) mais aucun extremum global.

La recherche des extrema locaux d'une application  $f$  (suffisamment) différentiable sur un ouvert se fait usuellement via une étude locale. Sur un ouvert, en un extremum local les dérivées partielles s'annulent (condition d'Euler). C'est une condition nécessaire non suffisante ; il est utile de regarder les dérivées partielles secondes ; lorsque ces dernières

s'annulent il faut regarder les dérivées d'ordre 3, *etc...* (voir figure 2).

En fait pour une application définie sur un ouvert de  $\mathbb{R}$  et infiniment différentiable, on a le résultat suivant :

**Théorème .1 (Extrema locaux d'une application analytique réelle)** Soit  $\mathcal{U}$  un ouvert de  $\mathbb{R}$  et  $f : \mathcal{U} \rightarrow \mathbb{R}$  une application infiniment dérivable. Soit un point  $u \in \mathcal{U}$  en lequel au moins une des dérivées successives de  $f$  est non nulle.

Alors  $u$  est un extremum local de  $f$  si et seulement si il existe un entier  $n$  impair tel que :

$$\forall i = 1, \dots, n, f^{[i]}(u) = 0, \text{ et } f^{[n+1]}(u) \neq 0$$

De plus si  $f^{[n+1]}(u) > 0$  alors c'est un minimum local et sinon c'est un maximum local.

**Démonstration.** Soit  $n$  le plus grand entier tel que  $\forall k, 1 \leq k \leq n, f^{[k]}(u) = 0$ , s'il existe, et  $n = 0$  sinon. Considérons le développement de Taylor-Young de  $f$  au voisinage de  $u$  à l'ordre  $n + 1$ .

$$f(u+t) = f(u) + \underbrace{\frac{f^{[n+1]}(u)}{(n+1)!}}_{\neq 0} t^{n+1} + o(|t|^{n+1})$$

Il en découle que si  $n + 1$  est impair, on peut trouver  $t$  aussi proche que l'on veut de 0 tel que  $f(u+t) - f(u)$  et  $f(u-t) - f(u)$  soient de signes strictement opposés, et donc  $u$  n'est pas extremum local. Et si  $n + 1$  est pair alors pour tout  $t$  suffisamment proche de zéro,  $f(u+t) - f(u)$  garde un signe constant et donc  $f(u)$  est un extremum local, minimum si  $f^{[n+1]}(u) > 0$  et maximum si  $f^{[n+1]}(u) < 0$ .  $\square$

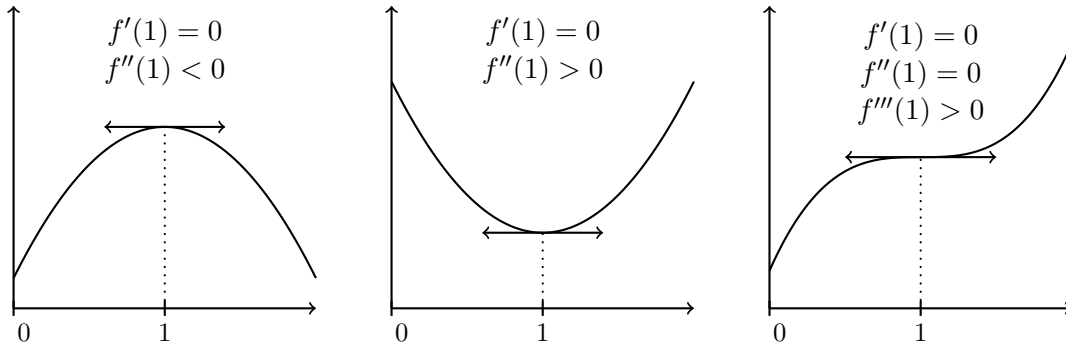


FIGURE 2 – Trois types de points critiques pour  $f : \mathbb{R} \rightarrow \mathbb{R}$  : un maximum local, un minimum local et un point d'inflexion.

Ce résultat se généralise en dimension supérieure, mais sa formulation y est bien plus technique et sans grande utilité (par cause de l'absence d'une théorie spectrale des  $p$ -formes lorsque  $p > 2$ ) ; nous ne l'aborderons pas. C'est pourquoi nous ne verrons des conditions, nécessaires, suffisantes en dimension supérieure, que jusqu'à l'ordre 2.

Attention, sur un domaine  $\mathcal{D}$  non ouvert, un extremum local n'est pas nécessairement un zéro de la dérivée (cf. figure 3). Nous verrons comment l'équation d'Euler se généralise en ce qu'on appelle les conditions de Lagrange (dans le cas où toutes les contraintes sont égalitaires) ainsi que les conditions de Karush-Kuhn-Tucker (dans le cas de contraintes égalitaires et inégalitaires).

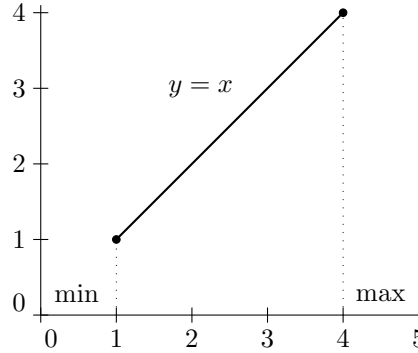


FIGURE 3 – Sur l'intervalle fermé  $[1, 4]$  l'application dérivable  $f(x) = x$  a un minimum en 1 et un maximum en 4, en lesquels la dérivée de  $f$  ne s'annule pas.

## 2 Exemples de problèmes d'optimisation à une variable

### 2.1 Minimisation des coûts dans la fabrication de boîtes cylindriques

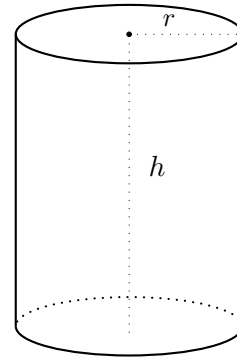
Dans la fabrication de boîtes de conserve cylindriques on minimise les coûts de matière première en cherchant le cylindre de surface minimale à volume constant. Considérons un cylindre, donné par sa hauteur  $h$  et le rayon  $r$  de sa base.

Le volume est :  $\pi r^2 h = K = \text{constante}$ .

L'aire est :  $2\pi r^2 + 2\pi r h$ .

Le problème d'optimisation s'écrit :

$$\begin{aligned} \min_{r, h} \quad & 2\pi r^2 + 2\pi r h \\ & \pi r^2 h = K \\ & r, h > 0 \end{aligned}$$



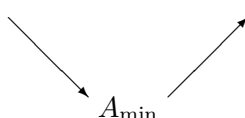
On utilise la contrainte égalitaire pour se ramener à un problème à une variable :

$$h = \frac{K}{\pi r^2} \implies \text{Aire}(r) = 2\pi r^2 + \frac{2K}{r}$$

$$\min_{r>0} A(r) = \pi r^2 + \frac{K}{r}$$

On étudie les variations de  $A(r)$  :

$$A'(r) = 2\pi r - \frac{K}{r^2} = \frac{2\pi r^3 - K}{r^2} \implies A'(r) \geq 0 \iff r \geq \sqrt[3]{\frac{K}{2\pi}}$$

$r$	0	$\sqrt[3]{\frac{K}{2\pi}}$	$+\infty$
$A'$	-	0	+
$A$			

Le minimum est :

$$r_{\min} = \sqrt[3]{\frac{K}{2\pi}}$$

$$\text{Alors } h_{\min} = \frac{K}{\pi^3 \frac{K^2}{4\pi^2}} = \sqrt[3]{\frac{4K}{\pi}} = 2r_{\min}$$

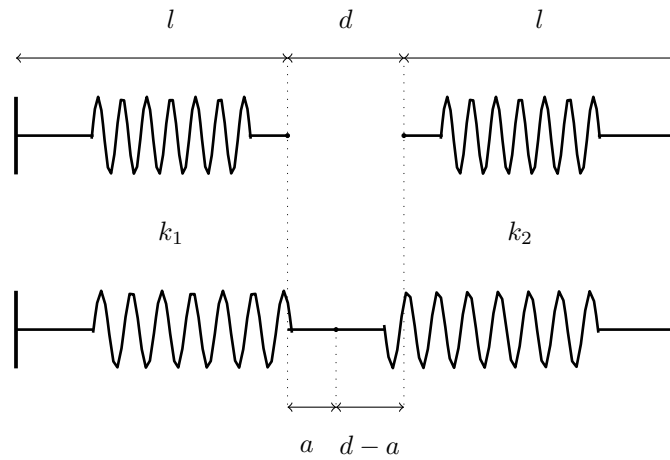
$$h_{\min} = 2r_{\min}$$

$$\text{Aire}_{\min} = 2\pi r_{\min}^2 + 2\pi r_{\min} h_{\min} = 2\pi \sqrt[3]{\frac{K^2}{4\pi^2}} + 4\pi \sqrt[3]{\frac{K^2}{4\pi^2}} = 3\sqrt[3]{\pi K^2}$$

$$\text{Aire}_{\min} = 3\sqrt[3]{\pi K^2}.$$

## 2.2 Position d'équilibre d'un système de deux ressorts.

Deux ressorts de coefficients de tension  $k_1, k_2$  et ayant même longueur à vide, ont chacun une extrémité fixe, et l'autre à distance mutuelle  $d$ . Lorsqu'on les attache par leur extrémité libre, comment s'exprime leur position d'équilibre (figure ci-dessous) ?



L'énergie potentielle à l'équilibre du premier ressort est :

$$E_1 = \frac{1}{2}k_1a^2$$

du deuxième ressort :

$$E_2 = \frac{1}{2}k_2(a-d)^2$$

L'énergie totale du système est :

$$E = E_1 + E_2 = \frac{1}{2}(k_1a^2 + k_2(a-d)^2)$$

La position d'équilibre est celle pour laquelle l'énergie potentielle du système est minimale. Il s'agit donc d'un problème d'optimisation qui s'exprime :

$$\min_{0 \leq a \leq d} E(a) = (k_1a^2 + k_2(a-d)^2)$$

$$E'(a) = 2(k_1 + k_2)a - 2k_2d; \text{ donc } E'(a) \geq 0 \iff a \geq \frac{k_2}{k_1 + k_2}d$$

$a$	0	$\frac{k_2d}{k_1 + k_2}$	$d$
$E'$	—	0	+
$E$	$k_2d^2$	$E_{\min}$	$k_1d^2$

La position d'équilibre est atteinte pour :

$$a = \frac{k_2}{k_1 + k_2}d$$

$$d - a = \frac{k_1}{k_1 + k_2}d$$

### 3 Problèmes d'optimisation sur plusieurs variables

#### 3.1 Production optimale d'une fonderie

Une fonderie fabrique 3 qualités de bronze à partir de cuivre et d'étain, en proportions variables. Elle dispose d'une quantité mensuelle de 65 tonnes de cuivre et de 5 tonnes d'étain.

Qualité	Bénéfice brut (K€/t)	% cuivre	% étain
A	2	90	10
B	1.6	93	7
C	1.8	95	5

Quelle production maximise le bénéfice mensuel ?

Notons :

- $x$  la quantité mensuelle produite (en tonne) de bronze de qualité  $A$ .
- $y$  la quantité mensuelle produite (en tonne) de bronze de qualité  $B$ .
- $z$  la quantité mensuelle produite (en tonne) de bronze de qualité  $C$ .

La fonction bénéfice brut mensuel à maximiser s'écrit :

$$f(x, y, z) = 2x + 1.6y + 1.8z$$

sous les contraintes inégalitaires :

$$90x + 93y + 95z \leq 6500$$

$$10x + 7y + 5z \leq 500$$

$$x, y, z \geq 0$$

Comment le résoudre ? Toutes les fonctions étant linéaires, on est dans le cadre de la programmation linéaire.

### 3.2 Problème de transport

On considère un problème de ravitaillement ; il fait partie d'une large classe de problème (amplement étudiée que ce soit en optimisation continue ou en optimisation combinatoire, et que l'on sait résoudre efficacement) très utile dans la pratique, plus généralement appelé *problème de transport*.

On souhaite ravitailler en carburant 3 sites à partir de 2 dépôts de capacité limitée. L'acheminement en carburant d'un dépôt à un site a un coût unitaire. Le tableau suivant résume chacun de ces coûts ainsi que la demande de chaque site, et le stock disponible dans chaque dépôt.

	site 1	site 2	site 3	diponibilité
dépôt 1	10	12	9	300
dépôt 2	11	11	10	450
demande	200	250	250	

Notons  $x_{ij}$ ,  $i = 1, 2$ ,  $j = 1, 2, 3$  la quantité de carburant (en unité de volume) acheminée du dépôt  $i$  au site  $j$ . Le coût d'acheminement est donné par la fonction coût :

$$f(x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23}) = 10x_{11} + 12x_{12} + 9x_{13} + 11x_{21} + 11x_{22} + 10x_{23}$$

qu'il s'agit de minimiser, sous les contraintes :

- Contraintes égalitaires provenant de la demande :

$$\begin{cases} x_{11} + x_{21} = 200 \\ x_{12} + x_{22} = 250 \\ x_{13} + x_{23} = 250 \end{cases}$$



– Contraintes inégalitaires provenant du stock disponible :

$$\begin{cases} x_{11} + x_{12} + x_{13} \leq 300 \\ x_{21} + x_{22} + x_{23} \leq 450 \end{cases}$$

– Contraintes de signe : (S)  $x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23} \geq 0$ .

Il s'agit encore d'un problème de programmation linéaire.

### 3.3 Régression linéaire

On considère un nuage de points  $(\mathbf{x}_n)_{n=1,\dots,N}$  dans  $\mathbb{R}^2$ . On cherche la droite affine qui approche le mieux ce nuage de points au sens des moindres carrés. Si l'on note  $\mathbf{x}_n = (x_n, y_n)$ , et  $\Delta : y = \alpha x + \beta$ , on cherche :

$$\min_{\alpha, \beta \in \mathbb{R}} \sum_{i=1}^N \|y_n - \alpha x_n - \beta\|^2$$

ce qui équivaut au problème de programmation quadratique :

$$\min_{\alpha, \beta \in \mathbb{R}} \sum_{i=1}^N (y_n - \alpha x_n - \beta)^2$$

On verra pourquoi ce problème admet toujours une solution, et l'on retrouvera les formules bien connues de la droite de régression linéaire.

### 3.4 Modélisation de données expérimentales

Plus généralement supposons que l'on ait effectué une série de  $p$  mesures dépendant d'un paramètre (évolution d'une concentration chimique, ou de toute autre grandeur physique, en fonction du temps, *etc...*).

paramètre	$t_1$	$t_2$	$\cdots$	$t_p$
valeur mesurée	$y_1$	$y_2$	$\cdots$	$y_p$

On souhaite modéliser ces données expérimentales par une certaine application mathématique  $F(\alpha_1, \dots, \alpha_n) : \mathbb{R} \rightarrow \mathbb{R}$  dépendant de paramètres réels  $(\alpha_1, \dots, \alpha_n) \in \mathcal{D} \subset \mathbb{R}^n$ . On cherche à déterminer les paramètres pour lesquels les valeurs prises par la fonction aux points  $t_1, \dots, t_p$  collent au mieux aux valeurs mesurées, dans un certain sens, disons par exemple au sens des moindres carrés.

Il s'agit alors du problème d'optimisation :

$$\min_{(\alpha_1, \dots, \alpha_p) \in \mathcal{D}} \sum_{i=1}^N (y_n - F(\alpha_1, \dots, \alpha_n)(x_n))^2$$

Ce problème est fondamental en sciences expérimentales.



# Chapitre I

## Programmation linéaire

Nous étudions dans ce chapitre la programmation linéaire, c'est à dire la classe des problèmes d'optimisation où la fonction objectif et les contraintes sont toutes affines. Il s'agit d'un domaine dont le champ d'application est énorme. Nous nous focalisons sur une méthode systématique de résolution, certainement la plus importante, la *méthode du simplexe* de Georges Dantzig. Ce n'est cependant pas la seule : pour des problèmes de grande taille on utilise en général plutôt la *méthode des points intérieurs*, que nous ne verrons pas (de toute façon, dans ce cas, les logiciels informatiques s'en chargent).

Contrairement aux autres chapitres, nous ne donnerons pas ici les preuves des résultats énoncés, nous en tenant aux idées directrices. Il aurait été autrement nécessaire de n'aborder ce chapitre que plus tard dans le déroulement du cours, bien qu'il s'agisse du domaine présentant le plus grand intérêt pratique. Par ailleurs la linéarité des applications considérées permet une résolution systématique, qui ne nécessite pas pour son application une compréhension fine, contrairement aux autres notions que nous aborderons par la suite.

### I.1 Préliminaires

#### I.1.1 Formulation

Dans tout ce chapitre  $n$  désigne un entier strictement positif et  $p$  désigne un entier. Soit  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  une application linéaire. Soient  $\varphi_1, \dots, \varphi_p : \mathbb{R}^n \rightarrow \mathbb{R}$ ,  $p$  applications linéaires, et  $(b_1, \dots, b_p) \in \mathbb{R}^p$ . Notons  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$  ; un problème de programmation linéaire s'exprime sous la forme :

trouver le minimum (respectivement le maximum) de  $f : \mathbb{R}^n \rightarrow \mathbb{R}$

$$\begin{array}{c} \min \\ \max \\ \mathbf{x} \end{array} f(\mathbf{x})$$

soumis aux contraintes :

$$\begin{cases} \varphi_1(x_1, \dots, x_n) \leq b_1 \\ \varphi_2(x_1, \dots, x_n) \leq b_2 \\ \vdots \\ \varphi_p(x_1, \dots, x_n) \leq b_p \end{cases}$$

$f, \varphi_1, \dots, \varphi_p$  étant des formes linéaires sur  $\mathbb{R}^n$ , on peut toujours noter pour certains vecteurs  $\mathbf{c} = (c_1, \dots, c_n) \in \mathbb{R}^n$ ,  $\mathbf{a}_i = (a_{i1}, \dots, a_{in}) \in \mathbb{R}^n$ , et  $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$  :

$$f(\mathbf{x}) = \sum_{i=1}^n c_i x_i = \langle \mathbf{c}, \mathbf{x} \rangle \quad ; \quad \varphi_i(\mathbf{x}) = \sum_{j=1}^n a_{ij} x_j = \langle \mathbf{a}_i, \mathbf{x} \rangle .$$

où  $\langle \cdot, \cdot \rangle$  désigne le produit scalaire usuel de  $\mathbb{R}^n$ .

### I.1.2 Représentation matricielle

Considérons, ce que nous appellerons *la matrice des contraintes*, et le vecteur des contraintes :

$$A = (a_{ij})_{\substack{i=1 \dots p \\ j=1 \dots n}} = \begin{pmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{p1} & \dots & a_{pn} \end{pmatrix}_{p \times n} \in \mathcal{M}_{p,n}(\mathbb{R}) \quad ; \quad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_p \end{pmatrix} \in \mathbb{R}^p .$$

Le problème d'optimisation s'écrit alors sous *forme matricielle* :

$$\begin{aligned} & \min \\ & \max \\ & \mathbf{x} \\ & A\mathbf{x} \leq \mathbf{b} \end{aligned}$$

Si de plus  $x_1, \dots, x_n \geq 0$ , on note  $\mathbf{x} \geq \mathbf{0}$ .

*Attention : L'inégalité entre vecteur  $\geq, \leq$  doit être comprise comme l'inégalité terme à terme, et ne définit pas un ordre sur les vecteurs !*

### I.1.3 Forme canonique

En programmation linéaire, un problème d'optimisation sous forme canonique est un problème sous la forme :

$$\begin{aligned} & \max_{\mathbf{x}} \langle \mathbf{c}, \mathbf{x} \rangle \\ & A\mathbf{x} \leq \mathbf{b} \\ & \mathbf{x} \geq \mathbf{0} \end{aligned}$$

Ce n'est pas restrictif : tout problème peut se mettre sous forme quadratique grâce à :

- $\min \langle \mathbf{c}, \mathbf{x} \rangle$  équivaut à  $\max \langle -\mathbf{c}, \mathbf{x} \rangle$ ,
- si  $x_i \not\geq 0$ , poser  $x_i = x_i^+ - x_i^-$  avec  $x_i^+, x_i^- \geq 0$ .

### I.1.4 Exemple de problème à deux variables - Résolution graphique

**Exemple.** Une société fabrique deux types de produits  $A$  et  $B$  (par exemple deux types de système audio), dont la vente lui rapporte un bénéfice brut respectif de 150 u.m. et de 450 u.m. ; sa production est limitée respectivement à 120 et 70 unités. Une même pièce  $P$  (par exemple un lecteur CD) rentre dans la fabrication d'une unité de  $A$ , ainsi que dans la fabrication d'une unité de  $B$ . Une même pièce  $Q$  (par exemple un haut-parleur) rentre dans la fabrication d'une unité de  $A$ , tandis que deux pièces  $Q$  sont nécessaires à la fabrication d'une pièce  $B$ . Elle dispose d'un stock de 140 pièces  $P$  et de 180 pièces  $Q$ . Comment gérer au mieux sa production en produits  $A$ ,  $B$  pour en retirer le bénéfice maximal ?

Notons :

$x$  la quantité de produits  $A$  fabriqués,

$y$  la quantité de produits  $B$  fabriqués,

$f(x, y) = 150x + 450y$  la fonction économique qui donne le bénéfice brut pour une production  $(x, y)$

Le problème d'optimisation se formalise alors :

$$\max_{(x,y)} f(x, y) = 150x + 450y$$

$$x \leq 120$$

$$y \leq 70$$

$$x + y \leq 140$$

$$x + 2y \leq 180$$

$$x, y \geq 0 \quad (\text{S})$$

Le domaine admissible est le polygone  $\mathcal{D} = \{(x, y) \in \mathbb{R}^2 \mid 0 \leq x \leq 120, 0 \leq y \leq 70, x + y \leq 140, x + 2y \leq 180\}$  de  $\mathbb{R}^2$ .

**Constatations :** On constate sur cet exemple (cf. figure I.1) :

- le domaine admissible est un hexagone convexe,
- la ligne de niveau  $k$  est la droite  $150x + 450y = k$  ; les lignes de niveau forment un faisceau de droites parallèles, de pente  $-1/3$ .

**On en déduit :** le maximum est atteint sur l'un des sommets de l'hexagone. Il suffit donc de calculer la valeur prise par  $f$  sur ses 6 sommets  $(0, 0)$ ,  $(0, 70)$ ,  $(120, 0)$ ,  $(40, 70)$ ,  $(100, 40)$ . Le calcul direct nous donne  $\underline{f_{\max} = 37500}$  est atteint au point  $\underline{\mathbf{u}_{\max} = (40, 70)}$ . Il faut produire 40 unités de  $A$  et 70 unités de  $B$ .

La figure I.3 représente la nappe représentative de  $f$  au dessus du domaine  $\mathcal{D}$ . C'est la portion (hexagonale) à la verticale de  $\mathcal{D} \subset \mathbb{R}^2$  du plan d'équation  $z = 150x + 450y$ .

### I.1.5 Généralisation

On tire ici des principes généraux, en dimension  $n$  quelconque, après les constatations faites sur l'exemple en dimension 2 de la section précédente.

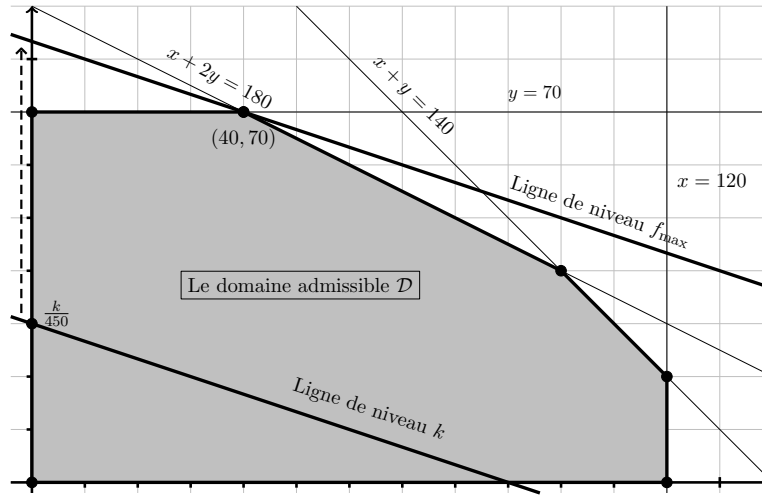


FIGURE I.1 – Le domaine admissible  $\mathcal{D} \subset \mathbb{R}^2$ , deux lignes de niveau, et le maximum  $(40, 70)$  de  $f$  sur  $\mathcal{C}$ .

• Chaque contrainte inégalitaire est l'équation d'un demi-espace. Sa frontière est un hyperplan affine (*i.e.* un sous-espace affine de codimension 1) dans l'espace  $\mathbb{R}^n$ . Ainsi le domaine admissible est une intersection d'un nombre fini de demi-espaces. C'est donc un polytope<sup>1</sup> convexe, ayant un nombre fini de sommets. Il peut être borné, ou non borné (cf. figure I.2).

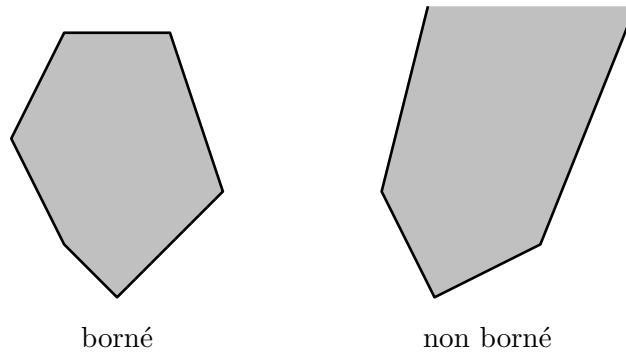


FIGURE I.2 – Deux polygones convexes de  $\mathbb{R}^2$ , l'un borné, l'autre non borné.

Le domaine est un fermé, aussi lorsque il est de plus borné c'est un compact de  $\mathbb{R}^n$ .

1. Un polytope généralise à toute dimension la notion de polygone dans  $\mathbb{R}^2$  et de polyèdre dans  $\mathbb{R}^3$ . Ici ce que l'on dénote par polytope, polygone, ou polyèdre est un peu plus général que la définition usuelle, puisqu'il peut être non borné. Une définition rigoureuse d'un *polytope convexe* est : *étant donné un nombre fini de segments et de demi-droites, c'est le plus petit convexe de  $\mathbb{R}^n$  les contenant.*

Or une application linéaire sur  $\mathbb{R}^n$  est continue. Ainsi lorsque le domaine est borné,  $f$  y prend un minimum ainsi qu'un maximum (cf. § 2.1.1).

- Lorsque  $f \neq 0$  les hyperplans de niveau (l'hyperplan de niveau  $k$  a pour équation  $\langle \mathbf{c}, \mathbf{x} \rangle = k$ ) sont tous parallèles (car de vecteur normal  $\mathbf{c}$ ). Avec ce qui précède, cela a pour conséquence que si un extremum existe il est atteint sur l'un des sommets du domaine polytope (éventuellement sur tous les points d'une de ses faces, et en particulier sur l'un des sommets aussi).

On résume toutes ces constatations dans le théorème suivant :

**Théorème I.1 (Programmation linéaire)** *En programmation linéaire, le domaine admissible, s'il est ni vide ni tout  $\mathbb{R}^n$ , est un polytope convexe ayant un nombre fini de sommets, qui peut être borné ou non borné. Si un extremum existe alors il est atteint sur l'un des sommets du polytope. Un point dans l'intérieur du domaine n'est jamais extremal si  $f \neq 0$ . Lorsque le polytope est borné,  $f$  y prend un minimum ainsi qu'un maximum.*

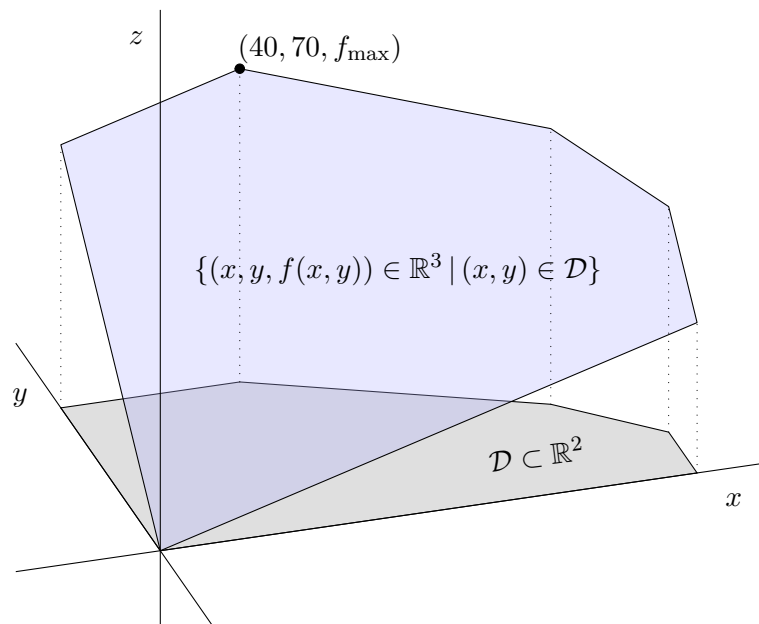


FIGURE I.3 – La nappe représentative de  $f$  au dessus du domaine  $\mathcal{D}$ .

Ce théorème fournit déjà une solution géométrique à un problème de programmation linéaire : il suffit de construire le polytope (on détermine ses sommets, arêtes, faces, etc...) en résolvant des systèmes d'équations linéaires. S'il est non borné certaines arêtes sont des demi-droites et en restreignant  $f$  à celles-ci on détermine si la fonction y tend vers  $+\infty$  ou  $-\infty$ . Ce faisant on sait alors si  $f$  admet un minimum ou un maximum. Si c'est le cas il suffit de calculer la valeur de  $f$  sur chacun des sommets pour déterminer un extremum.

Il faut cependant éviter cette méthode qui ne peut être utile qu'en petite dimension ( $\leq 3$ ) et lorsque le nombre de contraintes est faible. Nous allons voir dans la suite une méthode algébrique systématique pour résoudre un problème de programmation linéaire : *la méthode du simplexe*.

## I.2 Méthode du simplexe

La méthode du simplexe est une méthode algébrique, algorithmique, qui met à profit ces constatations géométriques. En partant d'un sommet elle se déplace successivement sur des sommets voisins qui accroît la valeur de la fonction, jusqu'à -si arrêt il y a- être parvenu sur un minimum local. La linéarité, et plus encore la convexité (cf. § 2.3), assure alors qu'il s'agit d'un minimum global.

### I.2.1 Problème de programmation linéaire sous forme normale

Un problème de programmation linéaire est *sous forme normale*, lorsqu'il s'écrit :

$$\left. \begin{array}{l} \max_{\mathbf{x}} \langle \mathbf{c}, \mathbf{x} \rangle \\ A\mathbf{x} \leq \mathbf{b} \\ \mathbf{x} \geq \mathbf{0} \quad (\text{S}) \end{array} \right\} \text{ forme canonique}$$

$$\left. \begin{array}{l} \mathbf{b} \geq \mathbf{0} \\ \mathbf{c} \geq \mathbf{0} \end{array} \right\} \text{ conditions de positivité}$$

c'est à dire lorsqu'en plus d'être sous forme canonique il vérifie les deux conditions de positivité. Une telle forme est restrictive, et l'on n'appliquera la méthode du simplexe qu'à des problèmes de programmation linéaire sous forme normale. On verra cependant par la suite comment ramener tout problème de programmation linéaire à un problème équivalent écrit sous forme normale.

### I.2.2 Algorithme du simplexe I : préparation

Afin d'appliquer la méthode du simplexe on commence par procéder de la façon suivante :

**a.** On change chacune des  $p$  contraintes inégalitaires en une contrainte égalitaire en introduisant une *variable d'écart* notée  $s_i$  ( $i = 1, \dots, p$ ) :

$$\langle \mathbf{a}_i, \mathbf{x} \rangle \leq b_i \iff \begin{cases} \langle \mathbf{a}_i, \mathbf{x} \rangle + s_i = b_i \\ s_i \geq 0 \end{cases}$$

(en présence de contraintes égalitaires, on les laisse inchangées).

**b.** On constitue la matrice suivante :



$x_1$	$\cdots$	$x_n$	$s_1$	$\cdots$	$s_p$	$\mathbf{b}$	
$a_{11}$	$\cdots$	$a_{1n}$	1		0	$b_1$	} partie centrale
$\vdots$	$A$	$\vdots$		$\ddots$		$\vdots$	
$a_{p1}$	$\cdots$	$a_{pn}$	0		1	$b_p$	
$c_1$	$\cdots$	$c_n$	0	$\cdots$	0	$f - 0$	} ligne résultat
$\underbrace{\hspace{1.5cm}}_{\mathbf{c}^\top}$							

La première ligne est optionnelle et purement nominative ; la dernière ligne s'appelle *ligne résultat* ; on travaillera sur cette dernière ainsi que sur la partie centrale. Le trait vertical symbolise l'égalité ; il sépare la partie gauche de la colonne droite.

### Algorithme du simplexe II : Implémentation

Pour implémenter l'algorithme du simplexe, on applique une suite de transformations sur cette matrice, insérées dans une boucle 'while'. Chacune consiste en un saut sur l'un des sommets voisins du polytope, qui maximise localement  $f$ , (du moins en l'absence d'un phénomène retors dit de *cyclage*, voir plus loin). A l'étape initiale il faut comprendre que l'on se trouve au sommet origine ; " $f - 0$ " dans la ligne résultat colonne droite signifie : en ce point, la valeur de  $f$  est 0.

#### Algorithme.

- Faire tant que la ligne résultat contient un terme  $> 0$  dans sa partie gauche.
- Le plus grand élément  $> 0$  de la ligne résultat partie gauche détermine la *colonne pivot*  $j$ .
- Choisir un *pivot* dans la colonne pivot  $j$ . C'est un élément  $(i, j)$ , que l'on note  $\alpha_{ij}$  dans la colonne pivot partie centrale choisi de façon à ce que  $b_i/\alpha_{ij}$  soit  $\geq 0$  et minimal.
- Si un tel élément pivot n'existe pas alors quitter : il n'y a pas de solution.
- Sinon le pivot choisi détermine la ligne pivot (L) (ou ligne de limitation). Ajouter autant de fois que nécessaire la ligne pivot aux autres lignes jusqu'à annuler tous les termes de la colonne pivot, autres que le pivot.
- Fin tant que.

(L)		$a_{1j}$		$b_j$	$-a_{1j}/a_{ij} \times (L)$
		$\vdots$		$\vdots$	
		$a_{ij}$		$b_i$	$b_i/a_{ij}$ minimal
		$\vdots$		$\vdots$	
		$a_{pj}$		$b_p$	$-a_{pj}/a_{ij} \times (L)$
		$c_{\max} > 0$		$f - R$	$-c/a_{ij} \times (L)$

Colonne  
pivot

	0		$b_j - a_{1j}b_i/a_{ij}$
	$\vdots$		$\vdots$
	$a_{ij}$		$b_i$
	$\vdots$		$\vdots$
	0		$b_p - a_{pj}b_i/a_{ij}$
	0		$f - R - c_jb_i/a_{ij}$

- Lorsque l'algorithme s'arrête en concluant à l'existence d'un maximum :

Barrer dans la matrice toutes les colonnes à la verticale d'éléments non nuls de la ligne résultat partie gauche. Poser que chacune des variables correspondantes est égale à 0. Puis déterminer la valeur des autres variables (on n'a en fait besoin que des  $x_1, \dots, x_n$ ) en résolvant le système linéaire dans la partie centrale du tableau.

						} (S)
0.....0	$r_i < 0$	0.....0	$r_j < 0$	0.....0	$f - f_{\max}$	
$x_i = 0$		$s_j = 0$				

$\Rightarrow$  On obtient le maximum  $(x_1, x_2, \dots, x_n)$ ; la valeur maximale  $f_{\max}$  de  $f$  se lit dans la ligne résultat partie droite.

**Remarque.** Même lorsqu'un maximum existe l'algorithme ne converge pas nécessairement et peut tourner indéfiniment dans certains cas exceptionnels ! En fait par construction, la suite  $(\mathbf{u}_k)_{k \in \mathbb{N}}$  construite vérifie  $f(\mathbf{u}_{k+1}) \geq f(\mathbf{u}_k)$ . Le plus souvent, pour tout  $k$ ,  $f(\mathbf{u}_{k+1}) > f(\mathbf{u}_k)$  et on a dans ce cas construit une suite de sommets du polytope, leur finitude impliquant alors la convergence. Seulement un phénomène de *cyclage* peut en théorie apparaître, *i.e.*  $f(\mathbf{u}_{k+1}) = f(\mathbf{u}_k)$  à partir d'un certain rang, et dans ce cas la méthode échoue à produire une solution. Il existe des façons de s'en prémunir, cependant dans la pratique ce phénomène exceptionnel n'arrive presque jamais.

**Exemple.** Reprenons l'exemple vu précédemment et résolu graphiquement au paragraphe §I.1.4.

$$\max_{(x,y)} f(x,y) = 150x + 450y$$

$$x \leq 120$$

$$y \leq 70$$

$$x + y \leq 140$$

$$x + 2y \leq 180$$

$$x, y \geq 0 \quad (\text{S})$$

$x$	$y$	$s_1$	$s_2$	$s_3$	$s_4$	
1	0	1	0	0	0	120
0	1	0	1	0	0	70
1	1	0	0	1	0	140
1	2	0	0	0	1	180
150	450	0	0	0	0	$f - 0$

$x$	$y$	$s_1$	$s_2$	$s_3$	$s_4$	
1	0	1	0	0	0	120
0	<u>1</u>	0	1	0	0	70 (L)
1	1	0	0	1	0	140 $-(L)$
1	2	0	0	0	1	180 $-2(L)$
150	<u>450</u>	0	0	0	0	$f - 0$ $-450(L)$

$x$	$y$	$s_1$	$s_2$	$s_3$	$s_4$	
1	0	1	0	0	0	120 $-(L)$
0	1	0	1	0	0	70
1	0	0	-1	1	0	70 $-(L)$
<u>1</u>	0	0	-1	0	1	40 (L)
<u>150</u>	0	0	-450	0	0	$f - 31500$ $-150(L)$

$x$	$y$	$s_1$	$s_2$	$s_3$	$s_4$	
0	0	1	1	0	-1	80
0	1	0	1	0	0	70
0	0	0	0	1	-1	30
1	0	0	-1	0	1	40
0	0	0	-300	0	-150	$f - 37500$

On obtient pour maximum  $x_1 = 40$ ,  $x_2 = 70$ , et  $f_{\max} = 37500$ . ( $s_1 = 80$ ,  $s_2 = 0$ ,  $s_3 = 30$ ,  $s_4 = 0$ ).

### I.3 Résolution dans le cas général

Ou comment ramener un problème de programmation linéaire à un problème équivalent mis sous forme normale.

#### I.3.1 Ecrire un problème de maximisation sous forme normale

Ou comment ramener un problème de maximisation linéaire à un problème équivalent mis sous forme normale.

- En l'absence de la contrainte de signe : (S) :  $\mathbf{x} \geq \mathbf{0}$ , c'est à dire si  $x_i \not\geq 0$ .

Poser  $x_i = x_i^+ - x_i^-$  avec  $x_i^+, x_i^- \geq 0$ .

- En présence de contraintes égalitaires.

Les insérer dans le tableau de la méthode du simplexe, sans ajouter de variable d'écart : cela revient à utiliser chacune de ces contraintes pour exprimer une variable en fonction des autres.

- Si  $\mathbf{c} \not\geq \mathbf{0}$ .

Par exemple si  $c_i < 0$  : poser  $x_i = 0$ . Dans la matrice de la méthode du simplexe cela revient à barrer (=supprimer) la colonne correspondante.

- Si  $\mathbf{b} \not\geq \mathbf{0}$ .

Par exemple  $b_i < 0$ . On change la contrainte inégalitaire en une contrainte égalitaire en insérant une nouvelle variable  $p_i \geq 0$ .

$$\begin{aligned} a_{i1}x_1 + \cdots + a_{in}x_n &\leq -|b_i| \\ \iff -a_{i1}x_1 - \cdots - a_{in}x_n - p_i &= |b_i| \quad \text{avec } p_i \geq 0. \end{aligned}$$

#### I.3.2 Dualité minimum/maximum

Ou comment ramener un problème de minimisation à un problème de maximisation équivalent en programmation linéaire.

Un problème de minimisation s'écrit sous forme canonique :

$$\begin{aligned} \min_{\mathbf{y}} \quad & \langle \mathbf{b}, \mathbf{y} \rangle \\ \mathbf{y} \quad & \geq \mathbf{0} \\ A^\top \mathbf{y} \quad & \geq \mathbf{c} \end{aligned}$$

il est sous forme normale si de plus :

$$\mathbf{b}, \mathbf{c} \geq \mathbf{0}$$

**Théorème I.2 (Dualité min/max)** *Tout problème de minimisation linéaire (resp. sous forme normale) est équivalent à un problème de maximisation linéaire (resp. sous forme normale) dans le sens suivant :*

$$\begin{array}{lll} \min_{\mathbf{y}} g(\mathbf{y}) = \langle \mathbf{b}, \mathbf{y} \rangle & & \max_{\mathbf{x}} f(\mathbf{x}) = \langle \mathbf{c}, \mathbf{x} \rangle \\ A^\top \mathbf{y} \geq \mathbf{c} & \iff & A\mathbf{x} \leq \mathbf{b} \\ \mathbf{y} \geq \mathbf{0} & & \mathbf{x} \geq \mathbf{0} \end{array}$$

- $g_{\min} = f_{\max}$ ,
- un minimum de  $g$  a pour coordonnées les opposés des valeurs dans la ligne résultat correspondant aux variables d'écart du problème de maximisation.

**Exemple.** On considère le problème de minimisation :

$$\begin{array}{c} \min_{x,y} 2x + 8y \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 1 & 3 \\ 3 & 2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} \geq \begin{pmatrix} 100 \\ 100 \\ 500 \\ 900 \\ 1200 \end{pmatrix} \\ x, y \geq 0 \end{array}$$

Il est équivalent au problème de maximisation (sous forme normale) :

$$\begin{array}{c} \max_{x_1, \dots, x_5} 100x_1 + 100x_2 + 500x_3 + 900x_4 + 1200x_5 \\ \begin{pmatrix} 1 & 0 & 1 & 1 & 3 \\ 0 & 1 & 1 & 3 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix} \leq \begin{pmatrix} 2 \\ 8 \end{pmatrix} \\ x_1, x_2, x_3, x_4, x_5 \geq 0 \end{array}$$

que l'on résout par la méthode du simplexe :

$$\begin{array}{cccccc|c} 1 & 0 & 1 & 1 & \boxed{3} & 1 & 0 & 2 \\ 0 & 1 & 1 & 3 & 2 & 0 & 1 & 8 \\ \hline 100 & 100 & 500 & 900 & 1200 & 0 & 0 & f - 0 \\ \\ 1 & 0 & 1 & \boxed{1} & 3 & 1 & 0 & 2 \\ -2/3 & 1 & 1/3 & 7/3 & 0 & -2/3 & 1 & 20/3 \\ \hline -300 & 100 & 100 & 500 & 0 & -400 & 0 & f - 800 \\ \\ 1 & 0 & 1 & 1 & 3 & 1 & 0 & 2 \\ -3 & \boxed{1} & -2 & 0 & -7 & -3 & 1 & 2 \\ \hline -800 & 100 & -400 & 0 & -1500 & -900 & 0 & f - 1800 \end{array}$$

$$\begin{array}{ccccccc|c}
1 & 0 & 1 & 1 & 3 & 1 & 0 & 2 \\
-3 & 1 & -2 & 0 & -7 & -3 & 1 & 2 \\
\hline
-900 & 0 & -200 & 0 & -800 & -\boxed{600} & -\boxed{100} & f - \boxed{2000}
\end{array}$$

On en déduit :

$$\begin{array}{l}
f_{\min} = 2000 \\
\min = (600, 100)
\end{array}$$

## I.4 Programmation linéaire en nombres entiers

Attention, lorsque un problème d'optimisation linéaire cherche une solution entière (c'est à dire à coordonnées entières), tout ce que l'on a vu jusqu'à présent ne s'applique pas ! En particulier la méthode du simplexe donne l'optimum sur les réels et non sur les entiers. Un tel problème s'appelle un problème de programmation linéaire en nombres entiers (PLNE), et c'est un domaine de recherche spécifique, utilisant ses outils propres, que nous ne traiterons pas ici.

**Prendre un arrondi entier d'un optimum ne fournit pas en général l'optimum en nombres entiers.**

**Exemple.** Considérons le problème d'optimisation linéaire suivant :

$$\begin{array}{l}
\max \quad x + 4y \\
y \geq 0, 4 \geq x \geq 0 \\
425x + 200y \leq 670
\end{array}$$

Le maximum est le point (4, 2.5) en lequel la fonction vaut 14. Le maximum sur les nombres entiers s'obtient en faisant 'descendre' la ligne de niveau max, jusqu'à passer par le premier point à coordonnées entières dans le domaine. On trouve le point (1, 3) maximum sur les entiers, en lequel la fonction vaut 13 (voir figure ci-dessous). En particulier l'optimum entier n'est pas l'arrondi entier de l'optimum.

Il est facile de construire des exemples sur le même modèle où l'optimum entier est aussi éloigné que l'on veut de l'optimum. Cependant, **si la méthode du simplexe nous retourne une solution en nombres entiers, c'est bien évidemment aussi l'optimum sur le problème en nombre entiers**, puisque c'est l'optimum sur les réels.

**Remarque.** Lorsque toutes les fonctions considérées ont des coefficients entiers, ou plus généralement rationnels, l'optimum, s'il existe, est toujours un nombre rationnel. En particulier si la solution optimale est recherchée uniquement sur les rationnels, les méthodes de ce chapitre s'appliquent dès lors que les coefficients des fonctions sont rationnels. Et sinon, puisque  $\mathbb{Q}$  est dense dans  $\mathbb{R}$ , en prenant une approximation rationnelle suffisamment proche de l'optimum réel trouvé, on peut se rapprocher autant que l'on veut d'un optimum rationnel ; d'ailleurs pour cette raison lorsque l'optimum est non rationnel, un

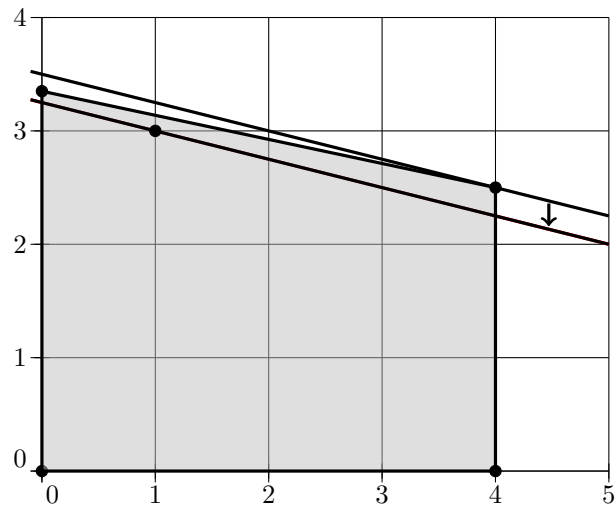


FIGURE I.4 – Exemple qui montre que le maximum sur les nombres entiers n'est pas l'arrondi entier de l'optimum réel.

optimum restreint aux rationnels n'existe pas, et on ne peut trouver qu'une approximation rationnelle d'un optimum réel.

## Exercices.

**Exercice 1.** Une usine produit deux types de produits finis  $x$  et  $y$  à partir d'une même matière première. Les produits  $x$  et  $y$  lui rapportent à la vente respectivement 8 et 4 euros le litre. La quantité de  $x$  et  $y$  produits est limitée par le stock de matière première disponible et par la durée du temps de travail. La fabrication d'un litre de produit  $x$  (resp.  $y$ ) nécessite  $1kg$  (resp.  $1kg$ ) de matière première. Il faut 15 heures de travail pour fabriquer 100l de  $x$  tandis qu'il faut 3 heures pour fabriquer 100l de  $y$ . On dispose de 1t de matière première et de 45 heures de travail chaque semaine.

Appliquer la méthode du simplexe pour maximiser le profit hebdomadaire.

**Exercice 2.** Résoudre par la méthode du simplexe le problème de production optimale d'une fonderie énoncé au § 3.1 de l'introduction.

**Exercice 3. a. Problème du consommateur.** On peut acheter 4 types d'aliments, dont la teneur en glucides et lipides est donnée dans le tableau suivant (par unité de poids et exprimée dans l'unité convenable) :

	type 1	type 2	type 3	type 4
glucides	2	1	0	1
lipides	1	2.5	2	4.5
prix	2	2	1	8

Le problème du consommateur consiste à obtenir au moindre coût au moins 12 unités de glucides et 7 unités de lipides.

Résoudre ce problème par la méthode du simplexe.

**b. Problème du concurrent.** Un vendeur concurrent souhaite s'approprier ce marché avec 2 nouveaux types d'aliment, dont les teneurs respectives en glucides et lipides sont données dans le tableau suivant (toujours exprimé par unité de volume dans l'unité convenable) :

	type 1	type 2
glucides	1	0
lipides	0	1

Il cherche à déterminer les prix de chacun de ces 2 produits lui permettant d'être le plus compétitif, tout en retirant le bénéfice maximal.

Déterminer les prix (par unité de poids) optimaux de ces 2 aliments.



## Chapitre II

# Généralités sur l'optimisation

**Notations.** On fixe les notations suivantes, et l'on renvoie au §A.2 et A.3 de l'annexe pour plus de précisions.

Dans tout ce qui suit  $n$  est un entier positif non nul. L'espace vectoriel réel de dimension  $n$  est muni de sa structure usuelle d'espace euclidien, c'est à dire du *produit scalaire usuel*  $\langle ., . \rangle$  et de la *norme associée*  $\|.\|_2$  (ou  $\|.\|$  lorsqu'il n'y a pas d'ambiguïté).

Si  $\mathcal{U}$  est un ouvert de  $\mathbb{R}^n$ ,  $\mathbf{x}_0 \in \mathbb{R}^n$  et  $f : \mathcal{U} \subset \mathbb{R}^n \longrightarrow \mathbb{R}$  est une application différentiable en  $\mathbf{x}_0$  on note :

$$\nabla f(\mathbf{x}_0) \triangleq \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}_0) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}_0) \end{pmatrix} \in \mathbb{R}^n$$

le *vecteur gradient de  $f$  en  $\mathbf{x}_0$*  (on prononce "nabla  $f$  de  $\mathbf{x}_0$ ").

On a alors le *développement de Taylor-Young de  $f$  à l'ordre 1 au voisinage de  $\mathbf{x}_0$*  :

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle + o(\|\mathbf{x} - \mathbf{x}_0\|)$$

Lorsque l'application  $f : \mathcal{U} \subset \mathbb{R}^n \longrightarrow \mathbb{R}$  est 2 fois différentiable en  $\mathbf{x}_0$  on note :

$$\nabla^2 f(\mathbf{x}_0) \triangleq \left( \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}_0) \right)_{\substack{i=1,2,\dots,n \\ j=1,2,\dots,n}} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\mathbf{x}_0) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}_0) \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}_0) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\mathbf{x}_0) \end{pmatrix}$$

la *matrice Hessienne de  $f$  en  $\mathbf{x}_0$* . C'est une matrice symétrique ; en particulier elle est diagonalisable.

On a alors au voisinage de  $\mathbf{x}_0$  le développement de Taylor-Young à l'ordre 2 :

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle + \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0) + o(\|\mathbf{x} - \mathbf{x}_0\|^2).$$

Puisque  $\nabla^2 f(\mathbf{x}_0)$  est symétrique,  $\mathbf{x}^\top \nabla^2 f(\mathbf{x}_0) \mathbf{x} = \langle \nabla^2 f(\mathbf{x}_0)^\top \mathbf{x}, \mathbf{x} \rangle = \langle \nabla^2 f(\mathbf{x}_0) \mathbf{x}, \mathbf{x} \rangle$ .

On s'intéressera à certaines propriétés des matrices Hessiennes, ou plus généralement des matrices carrées.

Une matrice  $A \in \mathcal{M}_n(\mathbb{R})$  est semi-définie positive si  $\forall \mathbf{x} \in \mathbb{R}^n$ ,  $\mathbf{x}^\top A \mathbf{x} \geq 0$ .

Une matrice  $A \in \mathcal{M}_n(\mathbb{R})$  est définie positive si  $\forall \mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ ,  $\mathbf{x}^\top A \mathbf{x} > 0$ .

On définit de façon analogue une matrice carrée semi-définie négative, définie négative.

## II.1 Conditions suffisantes d'existence d'extrema globaux

Nous voyons dans cette section deux conditions suffisantes d'existence d'extrema globaux : la compacité du domaine, et la coercivité de la fonction.

### II.1.1 Compacité du domaine

**Théorème II.1 (Existence d'extrema sur un domaine compact.)** *Si  $\mathcal{K}$  est un compact (i.e. est fermé et borné) de  $\mathbb{R}^n$ , et  $f : \mathcal{K} \rightarrow \mathbb{R}$  est continue, alors  $f$  admet un minimum ainsi qu'un maximum global sur  $\mathcal{K}$ .*

**Démonstration.** L'image d'un compact par une application continue est un compact. Ainsi  $f(\mathcal{K})$  est un compact de  $\mathbb{R}$ , c'est-à-dire un fermé borné. Puisque  $f(\mathcal{K})$  est borné il admet une borne inférieure  $m$  ainsi qu'une borne supérieure  $M$ . Par définition il existe une suite de points de  $f(\mathcal{K})$  convergeant vers  $M$ ; puisque  $f(\mathcal{K})$  est fermé,  $M \in f(\mathcal{K})$ . Le même raisonnement montre que  $m \in f(\mathcal{K})$ . Donc  $f^{-1}(\{m\})$  est non vide, et tous ses éléments sont des minima globaux de  $f$  sur  $\mathcal{K}$ , et de même  $f^{-1}(\{M\})$  est non vide et tous ses points sont des maxima globaux de  $f$  sur  $\mathcal{K}$ .  $\square$

Ce résultat n'est utile que face à un problème d'optimisation sous contraintes, car dans ce cas le domaine est toujours un fermé de  $\mathbb{R}^n$  et c'est le seul cas où il peut être borné, c'est-à-dire compact.

**Exemple.** soit  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  une application continue et soit  $\mathcal{C} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$  le cercle unité. Alors  $f$  admet (au moins) un maximum et un minimum sur  $\mathcal{C}$ . En effet  $\mathcal{C}$  est un compact de  $\mathbb{R}^2$  : d'une part c'est un fermé puisque c'est la préimage du fermé  $\{1\}$  de  $\mathbb{R}$  par l'application continue  $(x, y) \rightarrow x^2 + y^2$ ; d'autre part c'est un borné puisque la norme de  $(x, y) \in \mathcal{C}$  est uniformément majorée (égale à 1 pour la norme  $\|\cdot\|_2$ ).

### II.1.2 Applications coercives

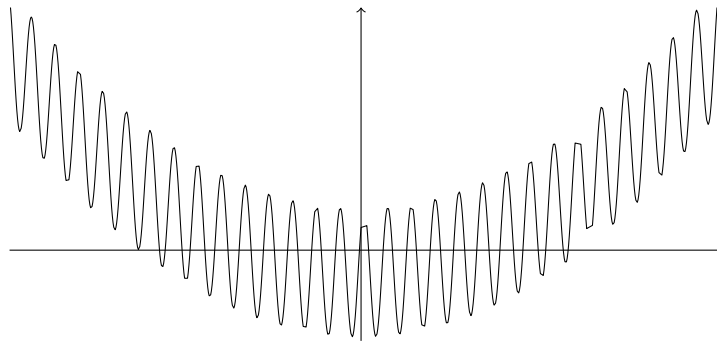
**Définition.** Une application  $f : \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}$  continue est coercive si  $\mathcal{D}$  est un fermé non borné et si :

$$\lim_{\|\mathbf{x}\| \rightarrow +\infty} f(\mathbf{x}) = +\infty$$

(souvent  $\mathcal{D} = \mathbb{R}^n$ ).

**Théorème II.2 (Une application coercive a un minimum.)** *Une application coercive admet un minimum global (et aucun maximum global). Si  $-f$  est coercive,  $f$  admet un maximum global (et aucun minimum global).*

**Démonstration.** Soit  $f : \mathcal{D} \rightarrow \mathbb{R}$  une application coercive. Soit  $a \in \mathbb{R}$ ; on choisit  $a$  suffisamment grand pour que  $\mathcal{K} = f^{-1}(]-\infty, a])$  soit non vide. Puisque  $f$  est continue et que  $] -\infty, a]$  est un fermé de  $\mathbb{R}$ ,  $\mathcal{K}$  est un fermé de  $\mathbb{R}^n$ . De plus  $\mathcal{K}$  est borné : autrement il contiendrait une suite de points  $(\mathbf{x}_n)_{n \in \mathbb{N}}$  avec  $\lim_{n \rightarrow \infty} \|\mathbf{x}_n\| = +\infty$  et  $\forall n \in \mathbb{N}, f(\mathbf{x}_n) \leq a$  ce qui contredirait le fait que  $f$  soit coercive. Ainsi  $\mathcal{K}$  est un compact de  $\mathbb{R}^n$ , et avec le théorème II.1  $f$  admet un minimum global  $\mathbf{u}$  sur  $\mathcal{K}$ , i.e.  $\forall \mathbf{x} \in \mathcal{K}, f(\mathbf{u}) \leq f(\mathbf{x}) \leq a$ . Or pour tout  $\mathbf{x} \in \mathcal{D} \setminus \mathcal{K}, f(\mathbf{x}) \geq a$ . Donc,  $\forall \mathbf{x} \in \mathcal{D}, f(\mathbf{u}) \leq f(\mathbf{x})$ , i.e.  $\mathbf{u}$  est un minimum global de  $f$  sur  $\mathcal{D}$ . Ceci montre la première assertion; la deuxième

FIGURE II.1 – Une application coercive  $f : \mathbb{R} \longrightarrow \mathbb{R}$ .

assertion est alors immédiate, puisqu'un minimum global de  $-f$  est un maximum global de  $f$ .  $\square$

**Exemple.** Une fonction polynomiale  $f : \mathbb{R} \longrightarrow \mathbb{R}$  de degré pair  $> 0$  est coercive sur  $\mathbb{R}$  si et seulement si le coefficient  $\alpha$  de son terme de plus haut degré est  $> 0$  : en effet  $\lim_{x \rightarrow \pm\infty} f(x) = +\infty$  si  $\alpha > 0$  et  $\lim_{x \rightarrow \pm\infty} f(x) = -\infty$  si  $\alpha < 0$ . Une fonction polynomiale de degré impair n'est jamais coercive sur  $\mathbb{R}$  mais, en notant  $\alpha$  le coefficient de son terme de plus haut degré, elle est coercive sur tout intervalle fermé  $[c, +\infty[$  si  $\alpha > 0$  et sur  $] -\infty, c]$  si  $\alpha < 0$ .

Une application polynomiale sur  $\mathbb{R}$  admet :

- Si son degré est pair et non nul :
  - Si le coefficient de son terme de plus haut degré est positif : un minimum global et aucun maximum global.
  - Si le coefficient de son terme de plus haut degré est négatif : un maximum global et aucun minimum global.
- Si son degré est impair : ni minimum ni maximum global.

## II.2 Recherche d'extrema locaux.

Nous voyons ici des conditions nécessaires, suffisantes, à l'ordre 1 et à l'ordre 2, pour qu'un point dans l'intérieur du domaine soit un extremum local d'une application différentiable (1 ou 2 fois). C'est l'outil principalement utilisé dans la recherche des extrema locaux en programmation sans contrainte. Attention tout ceci n'est valable que dans l'intérieur du domaine (ou autrement dit sur un domaine ouvert) ; nous généraliserons ces conditions sur tout le domaine dans le prochain chapitre.

**Rappels.** Soit  $\mathcal{D}$  un sous-ensemble de  $\mathbb{R}^n$ , et  $f : \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}$ .

Un point  $\mathbf{x}_0 \in \mathcal{D}$  est un extremum local de  $f$ , s'il existe un ouvert  $\mathcal{U} \subset \mathbb{R}^n$  contenant  $\mathbf{x}_0$ , tel que,  $\forall \mathbf{x} \in \mathcal{U} \cap \mathcal{D}$ ,  $f(\mathbf{x}) \geq f(\mathbf{x}_0)$  (respectivement  $f(\mathbf{x}) \leq f(\mathbf{x}_0)$ ). On dira alors que  $\mathbf{x}_0$  est un minimum local de  $f$  (respectivement maximum local).

Clairement tout extremum (resp. minimum, maximum) global de  $f$  est aussi un extremum (resp. minimum, maximum) local de  $f$ , tandis que la réciproque est évidemment fausse comme le montre l'exemple de la figure 1.

### II.2.1 Condition nécessaire du 1<sup>er</sup> ordre

**Rappel.** Si  $\mathcal{D}$  est un sous-ensemble de  $\mathbb{R}^n$ , l'intérieur de  $\mathcal{D}$ , noté  $\text{int}(\mathcal{D})$ , est le plus grand ouvert de  $\mathbb{R}^n$  inclus dans  $\mathcal{D}$ .

**Théorème II.3 (Equation d'Euler)** Soit  $f : \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}$ , différentiable en  $\mathbf{x}_0 \in \text{int}(\mathcal{D})$ . Si  $\mathbf{x}_0$  est un extremum local de  $f$ , alors  $\mathbf{x}_0$  est un point critique, i.e. :

$$\nabla f(\mathbf{x}_0) = \mathbf{0}.$$

**Démonstration.** Puisque  $f$  est différentiable en  $\mathbf{x}_0$ , les dérivées partielles de  $f$  en  $\mathbf{x}_0$  existent. Notons  $\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n\}$  la base canonique de  $\mathbb{R}^n$ . Pour tout  $i = 1, 2, \dots, n$  :

$$\begin{aligned} \frac{\partial f}{\partial x_i}(\mathbf{x}_0) &= \lim_{t \rightarrow 0} \frac{f(\mathbf{x}_0 + t\mathbf{e}_i) - f(\mathbf{x}_0)}{t} \\ &= \lim_{\substack{t \rightarrow 0 \\ <}} \underbrace{\frac{f(\mathbf{x}_0 + t\mathbf{e}_i) - f(\mathbf{x}_0)}{t}}_{=\Delta_g(t)} = \lim_{\substack{t \rightarrow 0 \\ >}} \underbrace{\frac{f(\mathbf{x}_0 + t\mathbf{e}_i) - f(\mathbf{x}_0)}{t}}_{=\Delta_d(t)}. \end{aligned}$$

Si  $\mathbf{x}_0$  est un extremum local de  $f$ , alors 0 est un extremum local de  $t \mapsto f(\mathbf{x}_0 + t\mathbf{e}_i)$  et donc  $\exists r > 0$  tel que lorsque  $t$  décrit  $] -r, r[$ ,  $f(\mathbf{x}_0 + t\mathbf{e}_i) - f(\mathbf{x}_0)$  garde un signe constant. Ainsi, lorsque  $t \in ] -r, r[$ , les taux d'accroissement de  $t \mapsto f(\mathbf{x}_0 + t\mathbf{e}_i)$  en 0, à droite et à gauche,  $\Delta_d(t)$  et  $\Delta_g(t)$ , sont de signes opposés. Donc, par passage à la limite,  $\frac{\partial f}{\partial x_i}(\mathbf{x}_0)$  est à la fois positif et négatif, et donc nécessairement nul. Ainsi si  $\mathbf{x}_0$  est un extremum local,  $\nabla f(\mathbf{x}_0) = \mathbf{0}$ .  $\square$

**Remarque 1 :** Le résultat est faux lorsque l'extremum local n'est pas dans l'intérieur de  $\mathcal{D}$  ; exemple : en programmation linéaire sur polyèdre convexe borné  $\mathcal{D}$  l'application linéaire  $f(\mathbf{x}) = \langle \mathbf{u}, \mathbf{x} \rangle$  admet toujours un minimum et un maximum global sur la frontière  $\partial\mathcal{D} = \mathcal{D} \setminus \text{int}(\mathcal{D})$  de  $\mathcal{D}$ , et en ces points  $\nabla f(\mathbf{x}) = \mathbf{u}$ .

**Remarque 2 :** C'est une condition nécessaire non suffisante, comme le montre la figure II.2.

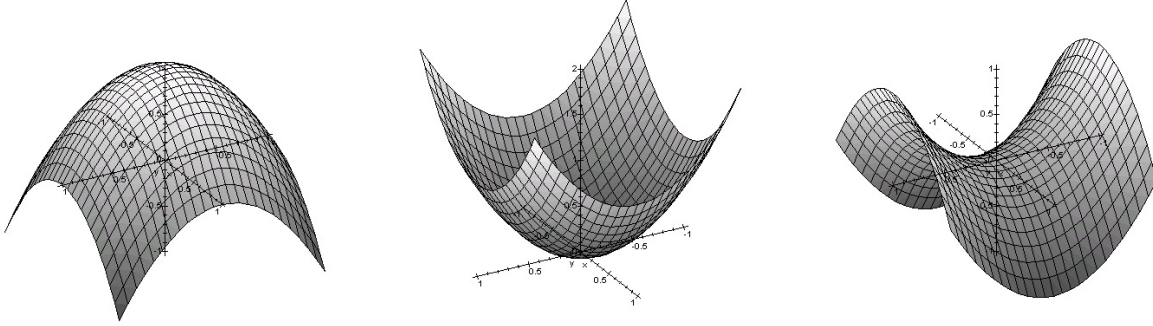


FIGURE II.2 – Trois types de points critiques pour  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  : un maximum local, un minimum local et un point-selle. Dans les deux premiers cas la matrice hessienne est respectivement semi-définie négative et semi-définie positive, dans le dernier cas elle a deux valeurs propres de signes opposés.

## II.2.2 Conditions du second ordre

Dans ce qui suit,  $\mathcal{D}$  désigne un sous-ensemble non vide de  $\mathbb{R}^n$ .

**Théorème II.4** Soit  $f : \mathcal{D} \subset \mathbb{R}^n \rightarrow \mathbb{R}$ , et  $\mathbf{u} \in \text{int}(\mathcal{D})$ , avec  $f$  2 fois différentiable en  $\mathbf{u}$ .

**1. (Condition nécessaire du 2<sup>e</sup> ordre.)**

Si  $\mathbf{u}$  est un minimum (resp. maximum) local de  $f$ , alors  $\nabla f(\mathbf{u}) = \mathbf{0}$  et  $\nabla^2 f(\mathbf{u})$  est semi-définie positive (resp. négative).

**2. (Condition suffisante du 2<sup>e</sup> ordre.)**

Si  $\nabla f(\mathbf{u}) = \mathbf{0}$  et  $\nabla^2 f(\mathbf{u})$  est définie positive (resp. négative), alors  $\mathbf{u}$  est un minimum (resp. maximum) local strict de  $f$ .

**Démonstration.** Nous montrons séparément les assertions 1 et 2.

**1.** Puisque  $\mathbf{u}$  est un extremum local alors  $\nabla f(\mathbf{u}) = \mathbf{0}$  (équation d'Euler, théorème II.3) et la formule de Taylor-Young à l'ordre 2 (cf. § A.2.4 page 114) s'écrit :

$$f(\mathbf{u} + \mathbf{x}) - f(\mathbf{u}) = \mathbf{x}^\top \nabla^2 f(\mathbf{u}) \mathbf{x} + o(\|\mathbf{x}\|^2)$$

Si  $\mathbf{u}$  est un minimum (resp. maximum) local de  $f$ , alors en appliquant la formule de Taylor-Young, il existe dans  $\mathcal{D}$  une boule ouverte  $B(\mathbf{0}, r)$  de  $\mathbb{R}^n$  centrée en  $\mathbf{0}$ , sur laquelle  $\mathbf{x}^\top \nabla^2 f(\mathbf{u}) \mathbf{x} \geq 0$  (resp.  $\leq 0$ ). Soit  $\mathbf{x} \in \mathbb{R}^n$ ; alors  $\mathbf{x}_0 = \frac{r}{2\|\mathbf{x}\|} \mathbf{x}$  est dans  $B(\mathbf{0}, r)$ , et donc  $\mathbf{x}_0^\top \nabla^2 f(\mathbf{u}) \mathbf{x}_0 \geq 0$  (resp.  $\leq 0$ ). Puisque  $\mathbf{x}^\top \nabla^2 f(\mathbf{u}) \mathbf{x} = \frac{r^2}{4\|\mathbf{x}\|^2} \mathbf{x}_0^\top \nabla^2 f(\mathbf{u}) \mathbf{x}_0 \geq 0$  (resp.  $\leq 0$ ),  $\nabla^2 f(\mathbf{u})$  est semi-définie positive (resp. négative).

**2.** Puisque  $\nabla f(\mathbf{u}) = \mathbf{0}$  la formule de Taylor-Young à l'ordre 2 (cf. proposition A.2.4) s'écrit :

$$f(\mathbf{u} + \mathbf{x}) - f(\mathbf{u}) = \mathbf{x}^\top \nabla^2 f(\mathbf{u}) \mathbf{x} + o(\|\mathbf{x}\|^2)$$

Puisque  $\nabla^2 f(\mathbf{u})$  est définie positive (resp. négative), alors  $\forall \mathbf{x} \in \mathbb{R}^n$ ,  $\mathbf{x}^\top \nabla^2 f(\mathbf{u}) \mathbf{x} > 0$  (resp.  $< 0$ ) et donc en appliquant la formule de Taylor-Young ci-dessus, il existe un ouvert contenant  $\mathbf{u}$  sur lequel  $f(\mathbf{x} + \mathbf{u}) - f(\mathbf{u}) \geq 0$  (resp.  $\leq 0$ ) :  $\mathbf{u}$  est donc un minimum (resp. maximum) local de  $f$ .  $\square$

**Exemple.** Soit l'application  $f$  de classe  $C^\infty$  (i.e. infiniment différentiable) :

$$\begin{aligned} f : \mathbb{R}^2 &\longrightarrow \mathbb{R} \\ (x, y) &\longrightarrow f(x, y) = x^3 + y^3 - 9xy \end{aligned}$$

Son vecteur gradient en un point  $(x, y)$  est :

$$\nabla f(x, y) = \begin{pmatrix} 3x^2 - 9y \\ 3y^2 - 9x \end{pmatrix},$$

et sa matrice Hessienne :

$$\nabla^2 f(x, y) = \begin{pmatrix} 6x & -9 \\ -9 & 6y \end{pmatrix}$$

Les points critiques, solutions de  $\nabla f(x, y) = 0$  sont les 2 points  $(0, 0)$  et  $(3, 3)$ . En ces points, les matrices Hessiennes sont :

$$\nabla^2 f(0, 0) = \begin{pmatrix} 0 & -9 \\ -9 & 0 \end{pmatrix} \quad \nabla^2 f(3, 3) = \begin{pmatrix} 18 & -9 \\ -9 & 18 \end{pmatrix}$$

- $\nabla^2 f(0, 0)$  a une trace nulle et un déterminant strictement négatif, elle n'est donc ni semi-définie positive, ni semi-définie négative :  $(0, 0)$  n'est pas un extremum local.
- $\nabla^2 f(3, 3)$  a une trace et un déterminant strictement positifs :  $(3, 3)$  est un minimum local.
- $f$  n'admet aucun extremum global puisque :

$$\lim_{x \rightarrow +\infty} f(x, 0) = +\infty \quad \lim_{x \rightarrow -\infty} f(x, 0) = -\infty.$$

**Remarques.** • La condition suffisante du 2<sup>e</sup> ordre s'utilise pour montrer qu'un point critique est un extremum local. La condition nécessaire du 2<sup>e</sup> ordre s'utilise pour montrer qu'un point critique n'est pas un extremum local. Lorsqu'en un point critique la matrice Hessienne est semi-définie positive ou négative, on ne sait à ce stade rien conclure ! Regarder à ce sujet l'exercice 2 page 49.

- Pour montrer qu'une matrice est/n'est pas définie positive/négative on utilise le théorème A.4, page 117. Pour montrer qu'une matrice est/n'est pas semi définie positive/négative on utilise le théorème A.5 page 117.
- L'étude locale ne suffit pas pour déterminer l'existence d'extrema globaux. En général on utilise des propriétés globales de l'application pour déterminer si parmi les extrema locaux, certains sont globaux. On montre le résultat suivant (peu utile en pratique, mais bon à savoir, voir exercice 5 page 49 pour une preuve.) :

**Proposition II.1** Soit  $\mathcal{D}$  un sous-ensemble connexe de  $\mathbb{R}^n$ ,  $f : \mathcal{D} \longrightarrow \mathbb{R}$  une application continue, et soit  $\mathbf{u} \in \mathcal{D}$  un minimum (resp. maximum) local de  $f$ . Alors  $\mathbf{u}$  est un minimum (resp. maximum) global de  $f$  si et seulement si,  $\forall \mathbf{x} \in \mathcal{D}$  tel que  $f(\mathbf{x}) = f(\mathbf{u})$ ,  $\mathbf{x}$  est un minimum (resp. maximum) local de  $f$ .

## II.3 Programmation convexe

Nous abordons ici les notions de convexité (large, stricte, forte) qui sont de première importance en optimisation :

- pour une application convexe un minimum local est aussi un minimum global,
- une application strictement convexe est convexe et un minimum, s'il existe, est unique (et donc strict),
- une application fortement convexe est strictement convexe et coercive, et donc admet un et un seul minimum global.

### II.3.1 Applications convexes, strictement convexes

**Définition.** Un sous-ensemble  $\mathcal{C}$  de  $\mathbb{R}^n$  est convexe si

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}, \forall t \in [0, 1], t\mathbf{x} + (1 - t)\mathbf{y} \in \mathcal{C}$$

(i.e. pour tout couple de points  $\mathbf{x}, \mathbf{y} \in \mathcal{C}$  le segment  $[\mathbf{x}, \mathbf{y}]$  est inclus dans  $\mathcal{C}$ , cf. figure II.3).

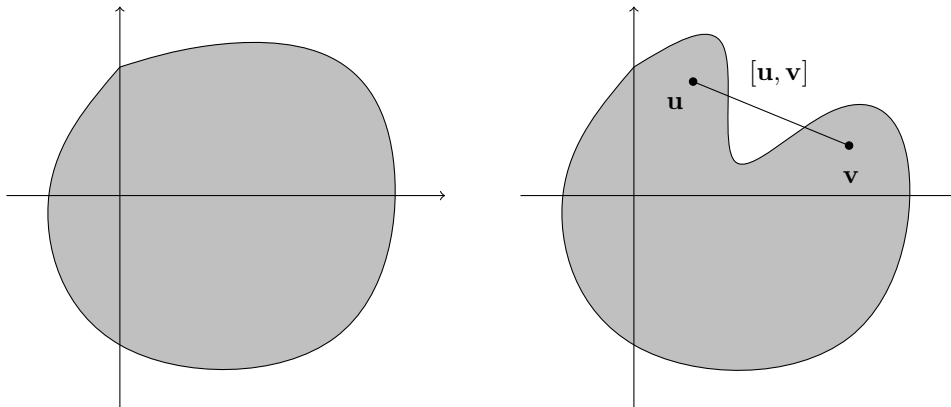


FIGURE II.3 – À gauche un sous-ensemble convexe de  $\mathbb{R}^2$ , à droite un sous-ensemble non convexe.

#### Propriétés.

- Tout sous-espace affine de  $\mathbb{R}^n$  (en particulier  $\mathbb{R}^n$ ) est convexe.
- Toute boule de  $\mathbb{R}^n$ , ouverte ou fermée, est un convexe de  $\mathbb{R}^n$ .
- L'intersection de convexes de  $\mathbb{R}^n$  est un convexe de  $\mathbb{R}^n$ .
- Si  $\mathcal{C}_1, \mathcal{C}_2$  sont deux convexes de  $\mathbb{R}^n$ , et  $\lambda \in \mathbb{R}$ , alors <sup>1</sup>  $\mathcal{C}_1 + \mathcal{C}_2$  et  $\lambda\mathcal{C}_1$  sont des convexes de  $\mathbb{R}^n$ .

1. En notant :  $\mathcal{C}_1 + \mathcal{C}_2 = \{\mathbf{x} + \mathbf{y} \in \mathbb{R}^n \mid \mathbf{x} \in \mathcal{C}_1, \mathbf{y} \in \mathcal{C}_2\}$ ;  $\lambda\mathcal{C}_1 = \{\lambda\mathbf{x} \in \mathbb{R}^n, \mid \mathbf{x} \in \mathcal{C}_1\}$ .

– Si  $\mathcal{C}_1$  est un convexe de  $\mathbb{R}^p$  et  $\mathcal{C}_2$  est un convexe de  $\mathbb{R}^q$ , leur produit cartésien  $\mathcal{C}_1 \times \mathcal{C}_2 = \{(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^p \times \mathbb{R}^q \mid \mathbf{x} \in \mathcal{C}_1, \mathbf{y} \in \mathcal{C}_2\}$  est un convexe de  $\mathbb{R}^p \times \mathbb{R}^q \approx \mathbb{R}^{p+q}$ .

**Définitions.** Soit  $\mathcal{C} \subset \mathbb{R}^n$  un ensemble convexe non vide et  $f : \mathcal{C} \rightarrow \mathbb{R}$ .

- L'application  $f$  est convexe si :

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}, \forall t \in [0, 1], \quad tf(\mathbf{x}) + (1-t)f(\mathbf{y}) \geq f(t\mathbf{x} + (1-t)\mathbf{y})$$

(i.e. dans  $\mathbb{R}^{n+1}$  le segment joignant  $(\mathbf{x}, f(\mathbf{x}))$  et  $(\mathbf{y}, f(\mathbf{y}))$  reste au-dessus de la nappe représentative de la fonction, cf. figure II.4.)

- L'application  $f$  est strictement convexe si :

$$\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}, \mathbf{x} \neq \mathbf{y}, \forall t \in ]0, 1[, \quad tf(\mathbf{x}) + (1-t)f(\mathbf{y}) > f(t\mathbf{x} + (1-t)\mathbf{y})$$

(i.e. dans  $\mathbb{R}^{n+1}$  le segment joignant  $(\mathbf{x}, f(\mathbf{x}))$  et  $(\mathbf{y}, f(\mathbf{y}))$  reste strictement au dessus de la nappe représentative de la fonction.)

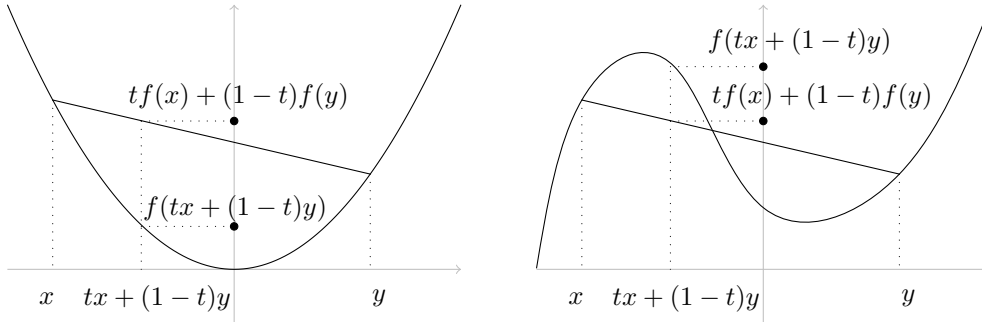


FIGURE II.4 – À gauche une application (strictement) convexe de  $\mathbb{R}$  dans  $\mathbb{R}$ , à droite une application non convexe.

### Propriétés.

- Toute application affine, définie sur un convexe, est convexe et non strictement convexe.
- La somme d'application (resp. strictement) convexes est (resp. strictement) convexe.
- Si  $f$  est (resp. strictement) convexe et  $\lambda \in \mathbb{R}_+$  (resp.  $\lambda \in \mathbb{R}_+^*$ ) alors  $\lambda f$  est (resp. strictement) convexe.
- Si  $f$  est (resp. strictement) convexe et  $a, b \in \mathbb{R}, a \neq 0$ , alors l'application  $\mathbf{x} \rightarrow f(a\mathbf{x} + b)$  est (resp. strictement) convexe.
- Si  $f_1, \dots, f_p$  sont (resp. strictement) convexes alors l'application  $\sup f_1, \dots, f_p$  est (resp. strictement) convexe.
- Une application convexe sur  $\mathcal{C}$  est continue en tout point de  $\text{int}(\mathcal{C})$ .

2. Définie par :  $\sup f_1, \dots, f_p : \mathbf{x} \rightarrow \sup\{f_1(\mathbf{x}), \dots, f_p(\mathbf{x})\}$ .



Les conditions formulées pour la définition d'une application convexe, strictement convexe, bien que significatives géométriquement, ne sont pas toujours pratiques, car il peut être difficile sous cette forme de vérifier si une application donnée les vérifie. Aussi donnons-nous des conditions du premier et du second ordre pour s'assurer de la convexité d'une fonction vérifiant des hypothèses adéquates de différentiabilité ; elles généralisent en dimension supérieure la caractérisation bien connue d'une application convexe  $f : \mathbb{R} \rightarrow \mathbb{R}$  : lorsque  $f$  est dérivable, sa dérivée  $f'$  est croissante, lorsque  $f$  est deux fois dérivable,  $\forall x, f'(x) \geq 0$ .

**Théorème II.5 (Caractérisations de la convexité.)** Soit  $\mathcal{U}$  un ouvert convexe de  $\mathbb{R}^n$  et  $f : \mathcal{U} \rightarrow \mathbb{R}$  une application.

1. (à l'ordre 1.) Si  $f$  est différentiable sur  $\mathcal{U}$ , alors
  - a.  $\forall \mathbf{x}, \mathbf{y} \in \mathcal{U}, f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \iff f$  est convexe sur  $\mathcal{U}$ ,
  - b.  $\forall \mathbf{x}, \mathbf{y} \in \mathcal{U}, \mathbf{x} \neq \mathbf{y}, f(\mathbf{y}) > f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle \iff f$  est strictement convexe sur  $\mathcal{U}$ .
2. (à l'ordre 2.) Si  $f$  est 2 fois différentiable sur  $\mathcal{U}$ , alors
  - a.  $\forall \mathbf{x} \in \mathcal{U}, \nabla^2 f(\mathbf{x})$  est semi-définie positive  $\iff f$  est convexe sur  $\mathcal{U}$ ,
  - b.  $\forall \mathbf{x} \in \mathcal{U}, \nabla^2 f(\mathbf{x})$  définie positive  $\implies f$  est strictement convexe.

**Remarques.** – Attention, l'implication de 2.b ne saurait admettre de réciproque en général comme le montre l'exemple de  $f(x) = x^4$  qui est strictement convexe sur  $\mathbb{R}^2$  tandis que  $f''(0) = 0$ . Par contre, comme nous le verrons, pour une fonction quadratique la réciproque est vraie.

– La première assertion exprime géométriquement que la nappe représentative d'une application (resp. strictement) convexe différentiable se situe au dessus de chacun de ses espaces tangents (resp. et ne l'intersecte qu'en un point).

**Démonstration.** Montrons séparément les assertions 1 et 2.

1. Soient  $\mathbf{x}, \mathbf{y}$  deux points distincts de  $\mathcal{U}$  et  $t \in ]0, 1[$ .

Si  $f$  est convexe,  $f(\mathbf{x} + t\mathbf{y}) \leq (1-t)f(\mathbf{x}) + tf(\mathbf{y})$ , ce qui s'écrit aussi  $\frac{1}{t}(f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})) \leq f(\mathbf{y}) - f(\mathbf{x})$ . Par passage à la limite :

$$\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle = \lim_{t \rightarrow 0} \frac{f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{t} \leq f(\mathbf{y}) - f(\mathbf{x}).$$

Si  $f$  est strictement convexe. Considérons un nombre  $\omega \in ]0, 1[$ , on vérifie :  $\mathbf{x} + t(\mathbf{y} - \mathbf{x}) = \frac{\omega-t}{\omega}\mathbf{x} + \frac{t}{\omega}(\mathbf{x} + \omega(\mathbf{y} - \mathbf{x}))$ . Alors en prenant  $0 < t \leq \omega$ , on déduit par convexité de  $f$  que :  $f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) \leq \frac{\omega-t}{\omega}f(\mathbf{x}) + \frac{t}{\omega}f(\mathbf{x} + \omega(\mathbf{y} - \mathbf{x}))$ . On en déduit alors :  $\frac{1}{t}(f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})) \leq \frac{1}{\omega}(f(\mathbf{x} + \omega(\mathbf{y} - \mathbf{x})) - f(\mathbf{x}))$ . Puisque  $f$  est strictement convexe on a d'autre part :  $\frac{1}{\omega}(f(\mathbf{x} + \omega(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})) < f(\mathbf{y}) - f(\mathbf{x})$ . On a donc établi la double inégalité suivante :  $\frac{1}{t}(f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})) \leq \frac{1}{\omega}(f(\mathbf{x} + \omega(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})) < f(\mathbf{y}) - f(\mathbf{x})$ . Par passage à la limite en faisant tendre  $t$  vers 0 et gardant  $\omega$  fixé, on obtient l'inégalité stricte recherchée :

$$\langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle = \lim_{t \rightarrow 0} \frac{f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{t} < f(\mathbf{y}) - f(\mathbf{x}).$$

Réciproquement, supposons que  $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$  pour tout  $\mathbf{x}, \mathbf{y} \in \mathcal{U}$ . Alors pour  $\mathbf{x} \neq \mathbf{y}$  dans  $\mathcal{U}$  et  $t \in ]0, 1[$ , on a en particulier  $f(\mathbf{y}) \geq f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) - t\langle \nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})), \mathbf{x} - \mathbf{y} \rangle$  ainsi que  $f(\mathbf{x}) \geq f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})) + (1-t)\langle \nabla f(\mathbf{y} + t(\mathbf{x} - \mathbf{y})), \mathbf{x} - \mathbf{y} \rangle$ . En multipliant la première inégalité par  $(1-t)$ , la deuxième par  $t$  puis en sommant on obtient :

$$f(t\mathbf{x} + (1-t)\mathbf{y}) \leq tf(\mathbf{x}) + (1-t)f(\mathbf{y})$$

ce qui montre la convexité de  $f$  ; lorsque les inégalités sont strictes on obtient une inégalité stricte et  $f$  est strictement convexe.

**2.** En appliquant la formule de Taylor-MacLaurin (cf. §A.2.4 page 114) au point  $\mathbf{x} \in \mathcal{U}$ , on obtient qu'il existe  $\theta > 0$  et  $\mathbf{z} \in \mathcal{U}$  tels que  $\mathbf{y} - \mathbf{x} = \theta(\mathbf{y} - \mathbf{z})$  et  $f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle = \frac{\theta^2}{2}(\mathbf{y} - \mathbf{z})^\top \nabla^2 f(\mathbf{z})(\mathbf{y} - \mathbf{z})$ . En appliquant **1** on déduit alors la convexité ou la stricte convexité de  $f$ .

Il reste à montrer l'implication réciproque dans **2.a**. Donnée un point  $\mathbf{x} \in \mathcal{U}$  considérons l'application  $g : \mathcal{U} \rightarrow \mathbb{R}$ , définie par  $g(\mathbf{y}) = f(\mathbf{y}) - \langle \nabla f(\mathbf{x}), \mathbf{y} \rangle$ . Alors  $g(\mathbf{y}) - g(\mathbf{x}) = f(\mathbf{y}) - f(\mathbf{x}) - \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$  et puisque  $f$  est convexe, avec **1.a**,  $\forall \mathbf{y} \in \mathcal{U}$ ,  $g(\mathbf{y}) - g(\mathbf{x}) \geq 0$ . Ainsi  $\mathbf{x}$  est un minimum de  $g$  sur  $\mathcal{U}$ . Or  $g$  est 2 fois différentiable et  $\nabla^2 g(\mathbf{x}) = \nabla^2 f(\mathbf{x})$ . En appliquant le théorème II.4.1 (condition nécessaire du second ordre), on en déduit que  $\nabla^2 f(\mathbf{x})$  est semi-définie positive.  $\square$

### II.3.2 Programmation convexe

On parle de programmation convexe lorsque :

–  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  est une application convexe, à optimiser sur

$$\mathcal{D} = \{\mathbf{x} \in \mathcal{U} \subset \mathbb{R}^n \mid \varphi_i(\mathbf{x}) = 0, \forall i = 1, \dots, p, \psi_j(\mathbf{x}) \leq 0, \forall j = 1, \dots, q\}$$

où :

- L'ensemble  $\mathcal{U}$  est un sous-ensemble convexe non vide de  $\mathbb{R}^n$ ,
- les applications  $\varphi_1, \dots, \varphi_p : \mathbb{R}^n \rightarrow \mathbb{R}$  sont affines,
- les applications  $\psi_1, \dots, \psi_q : \mathbb{R}^n \rightarrow \mathbb{R}$  sont convexes.

Dans ce cas  $\mathcal{D}$  est un sous-ensemble convexe de  $\mathbb{R}^n$ , comme le montre le résultat suivant :

**Proposition II.2 (Convexité du domaine.)** *Si les applications  $\varphi_1, \dots, \varphi_p : \mathbb{R}^n \rightarrow \mathbb{R}$  sont affines, si les applications  $\psi_1, \dots, \psi_q : \mathbb{R}^n \rightarrow \mathbb{R}$  sont convexes, et si  $\mathcal{U} \subset \mathbb{R}^n$  est un ensemble convexe, alors le domaine :*

$$\mathcal{D} = \{\mathbf{x} \in \mathcal{U} \subset \mathbb{R}^n \mid \varphi_i(\mathbf{x}) = 0, \forall i = 1, \dots, p, \psi_j(\mathbf{x}) \leq 0, \forall j = 1, \dots, q\}$$

*est un ensemble convexe de  $\mathbb{R}^n$*

**Démonstration.** Montrons tout d'abord que  $\psi_j$  étant convexe, l'ensemble  $\mathcal{C}_j = \{\mathbf{x} \in \mathbb{R}^n \mid \psi_j(\mathbf{x}) \leq 0\}$  est un convexe de  $\mathbb{R}^n$ . Soient  $\mathbf{x}, \mathbf{y} \in \mathcal{C}_j$ , on a  $\psi_j(\mathbf{x}) \leq 0$ , et  $\psi_j(\mathbf{y}) \leq 0$ . Alors par convexité de  $\psi_j$ , pour  $t \in [0, 1]$ ,  $\psi_j(t\mathbf{x} + (1-t)\mathbf{y}) \leq t\psi_j(\mathbf{x}) + (1-t)\psi_j(\mathbf{y}) \leq 0$ . Ainsi  $[\mathbf{x}, \mathbf{y}] \subset \mathcal{C}_j$  et donc  $\mathcal{C}_j$  est un convexe. D'autre part puisque  $\varphi_i$  est affine, l'ensemble  $\{\mathbf{x} \in \mathbb{R}^n \mid \varphi_i(\mathbf{x}) = 0\}$  est un sous-espace affine et donc un convexe de  $\mathbb{R}^n$ . Ainsi  $\mathcal{D}$  est une intersection de convexes de  $\mathbb{R}^n$  et est donc convexe.  $\square$

Si la convexité est une notion de première importance en optimisation, c'est d'abord parce qu'en programmation convexe un minimum local est aussi global. De plus une application strictement convexe admet au plus un minimum. C'est le résultat suivant.

**Théorème II.6 (Programmation convexe.)** Soient  $\mathcal{C}$  un sous-ensemble convexe de  $\mathbb{R}^n$ ,  $f : \mathcal{C} \rightarrow \mathbb{R}$  une application convexe et  $\mathbf{x}_0 \in \mathcal{C}$ .

1. Les conditions suivantes sont équivalentes :

(i)  $\mathbf{x}_0$  est un minimum local de  $f$ ,

(ii)  $\mathbf{x}_0$  est un minimum global de  $f$ .

Si de plus  $f$  est différentiable en  $\mathbf{x}_0 \in \mathcal{C}$ , (i) et (ii) sont équivalents à :

(iii) si  $\mathbf{x}_0 \in \text{int}(\mathcal{C})$ ,  $\nabla f(\mathbf{x}_0) = \mathbf{0}$ .

(iv)  $\forall \mathbf{x} \in \mathcal{C}$ ,  $\langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle \geq 0$

2. Si  $f$  est strictement convexe,  $f$  admet au plus un minimum, et un minimum de  $f$  est toujours strict.

**Démonstration.** L'implication (ii)  $\Rightarrow$  (i) est évidente ; montrons la réciproque. Soit  $\mathbf{x}_0$  un minimum local de  $f$  sur  $\mathcal{C}$  soit  $\mathbf{y} = \mathbf{x}_0 + \mathbf{z}$  un point quelconque de  $\mathcal{C}$  et  $t \in [0, 1]$ . La convexité de  $f$  implique que  $f(\mathbf{x}_0 + t\mathbf{z}) \leq (1-t)f(\mathbf{x}_0) + tf(\mathbf{y})$ , ce qui s'écrit aussi  $f(\mathbf{x}_0 + t\mathbf{z}) - f(\mathbf{x}_0) \leq t(f(\mathbf{y}) - f(\mathbf{x}_0))$ . Le point  $\mathbf{x}_0$  étant un minimum relatif, il existe  $t_0 \in ]0, 1[$  tel que  $0 \leq f(\mathbf{x}_0 + t_0\mathbf{z}) - f(\mathbf{x}_0)$ . Cela montre que  $0 \leq f(\mathbf{y}) - f(\mathbf{x}_0)$  et donc  $\mathbf{x}_0$  est un minimum global de  $f$  sur  $\mathcal{C}$ .

Si  $f$  est strictement convexe le même raisonnement conduit aux inégalités  $0 \leq f(\mathbf{x}_0 + t\mathbf{z}) < (1-t)f(\mathbf{x}_0) + tf(\mathbf{y})$  qui montrent que  $\mathbf{x}_0$  est un minimum strict, et en particulier est unique.

L'équivalence entre (ii) et (iv) est une conséquence immédiate du théorème II.5.1.a.

Pour finir montrons l'équivalence entre (i) et (iii). Si  $f$  est différentiable en  $\mathbf{x}_0 \in \text{int}(\mathcal{C})$  et si  $\mathbf{x}_0$  est un minimum local de  $f$  sur  $\mathcal{C}$ , la équation d'Euler implique que  $\nabla f(\mathbf{x}_0) = \mathbf{0}$ . Réciproquement considérons une boule ouverte  $\mathcal{B}$  centrée en  $\mathbf{x}_0$  dans  $\text{int}(\mathcal{C})$  ;  $\mathcal{B}$  est un ouvert convexe et  $f$  est clairement encore convexe sur  $\mathcal{B}$ . Si  $\nabla f(\mathbf{x}_0) = \mathbf{0}$ , la condition 1 du théorème II.5 implique que  $\mathbf{x}_0$  est un minimum de  $f$  sur  $\mathcal{B}$ , et donc un minimum local de  $f$  sur  $\mathcal{C}$ .  $\square$

**Remarque.** Une application convexe ou strictement convexe n'admet pas forcément un minimum. Exemple :  $f(x) = e^x$  est une application strictement convexe ( $f''(x) = e^x > 0$ ) n'admettant aucun minimum, puisque  $f'(x) = e^x \neq 0$ .

### II.3.3 Applications elliptiques

Une application convexe, ou même strictement convexe n'admet pas en général de minimum global. Il existe une condition plus forte, la *forte convexité*<sup>3</sup> qui assure l'existence d'un unique minimum global. Nous voyons cette notion ici dans un cadre plus restreint où l'application considérée est de plus de classe  $C^1$  sur  $\mathbb{R}^n$  ; on parle alors plutôt d'application  $\alpha$ -elliptique, ou encore *elliptique*.

**Définition.** Soit  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  une application de classe  $C^1$ . L'application  $f$  est elliptique ou encore  $\alpha$ -elliptique, s'il existe un réel  $\alpha > 0$ , tel que :

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

On peut pour une application deux fois différentiable en donner une caractérisation à l'ordre 2.

3. En toute généralité une application  $f$  définie sur un domaine convexe  $\mathcal{C}$  de  $\mathbb{R}^n$  est dite *fortement convexe* ou encore  $\alpha$ -convexe, si il existe un réel  $\alpha > 0$ , tel que  $\forall \mathbf{x}, \mathbf{y} \in \mathcal{C}, \forall t \in [0, 1], tf(\mathbf{x}) + (1-t)f(\mathbf{y}) \geq f(t\mathbf{x} + (1-t)\mathbf{y}) + \alpha \frac{t(1-t)}{2} \|\mathbf{x} - \mathbf{y}\|^2$ . Lorsque  $f$  est de classe  $C^1$  et  $\mathcal{C} = \mathbb{R}^n$  cette définition est équivalente à l' $\alpha$ -ellipticité de  $f$ .

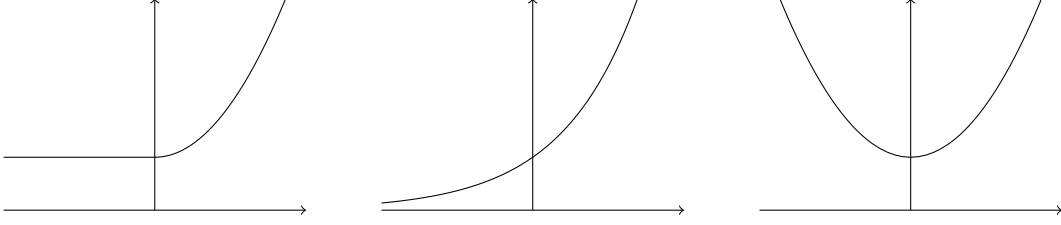


FIGURE II.5 – De gauche à droite, les graphes d'une application convexe, strictement convexe, elliptique (ou fortement convexe).

**Proposition II.3 (Caractérisation de l'ellipticité à l'ordre 2.)** Soit  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  une application deux fois différentiable. Alors  $f$  est  $\alpha$ -elliptique si et seulement si :

$$\forall \mathbf{x}, \mathbf{u} \in \mathbb{R}^n, \quad \mathbf{x}^\top \nabla^2 f(\mathbf{u}) \mathbf{x} \geq \alpha \|\mathbf{x}\|^2$$

**Démonstration.** Si  $f$  est  $\alpha$ -elliptique et deux fois différentiable, on a

$$\mathbf{x}^\top \nabla^2 f(\mathbf{u}) \mathbf{x} = \lim_{t \rightarrow 0} \frac{\langle \nabla f(\mathbf{u} + t\mathbf{x}) - \nabla f(\mathbf{u}), \mathbf{x} \rangle}{t} = \lim_{t \rightarrow 0} \frac{\langle \nabla f(\mathbf{u} + t\mathbf{x}) - \nabla f(\mathbf{u}), t\mathbf{x} \rangle}{t^2} \geq \alpha \|\mathbf{x}\|^2.$$

Réciproquement, on considère l'application  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  définie par  $g(\mathbf{z}) = \langle \nabla f(\mathbf{z}), \mathbf{y} - \mathbf{x} \rangle$ , avec  $\mathbf{x}, \mathbf{y}$  fixés dans  $\mathbb{R}^n$ , et on lui applique la formule de Taylor-McLaurin à l'ordre 1 au voisinage de  $\mathbf{x}$ . Il existe  $\theta \in [0, 1]$  tel que :

$$\begin{aligned} \langle \nabla f(\mathbf{y}) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle &= g(\mathbf{y}) - g(\mathbf{x}) = \langle \nabla g(\mathbf{x} + \theta(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle \\ &= (\mathbf{y} - \mathbf{x})^\top \nabla^2 f(\mathbf{x} + \theta(\mathbf{y} - \mathbf{x})) (\mathbf{y} - \mathbf{x}) \geq \alpha \|\mathbf{y} - \mathbf{x}\|^2, \end{aligned}$$

par hypothèse, et donc  $f$  est  $\alpha$ -elliptique.  $\square$

**Remarque.** Ce résultat peut s'interpréter par : une application  $f$  deux fois différentiable est  $\alpha$ -elliptique si et seulement si pour tout  $\mathbf{x} \in \mathbb{R}^n$  les valeurs propres de  $\nabla^2 f(\mathbf{x})$  sont minorées par  $\alpha$ .

### II.3.4 Programmation elliptique

Le résultat suivant assure de l'existence d'un minimum pour une application elliptique sur un domaine convexe fermé.

**Théorème II.7 (Programmation elliptique.)** Soit  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  une application  $\alpha$ -elliptique. Alors :

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

De plus  $f$  est coercive et strictement convexe. Sur un domaine convexe fermé et non vide de  $\mathbb{R}^n$ , elle admet un unique minimum.

**Démonstration.** Appliquons la formule de Taylor avec reste intégral à l'ordre 1 :

$$\begin{aligned}
 f(\mathbf{y}) - f(\mathbf{x}) &= \int_0^1 \langle \nabla f(\mathbf{x} + \theta(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle d\theta \\
 &= \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \langle \nabla f(\mathbf{x} + \theta(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle d\theta \\
 &\geq \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \alpha \theta \|\mathbf{y} - \mathbf{x}\|^2 d\theta \\
 &= \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\alpha}{2} \|\mathbf{y} - \mathbf{x}\|^2.
 \end{aligned}$$

On déduit alors de cette inégalité : d'une part,

$$\forall \mathbf{x} \neq \mathbf{y} \in \mathbb{R}^n, \quad f(\mathbf{y}) > f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle$$

et donc avec le théorème II.5.1.b  $f$  est strictement convexe. D'autre part  $f$  est coercive puisque :

$$f(\mathbf{x}) \geq f(\mathbf{0}) + \langle \nabla f(\mathbf{0}), \mathbf{x} \rangle + \frac{\alpha}{2} \|\mathbf{x}\|^2 \geq f(\mathbf{0}) - \|\nabla f(\mathbf{0})\| \|\mathbf{x}\| + \frac{\alpha}{2} \|\mathbf{x}\|^2.$$

Soit  $\mathcal{C}$  un domaine convexe fermé et non vide de  $\mathbb{R}^n$ . Si  $\mathcal{D}$  est non borné, alors  $f$  y admet un minimum par le théorème II.2. Si  $\mathcal{D}$  est borné, alors  $f$  y admet un minimum par le théorème II.1 et le fait qu'une application convexe est continue. De plus ce minimum est unique puisque  $f$  est strictement convexe sur  $\mathcal{D}$ .  $\square$

## II.4 Programmation quadratique sans contraintes

Nous appliquons ici les résultats obtenus dans les sections précédentes au cas particulier important de la programmation quadratique sans contraintes.

### II.4.1 Applications quadratiques

**Définition.** Une application  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  est quadratique lorsque c'est un polynôme de degré 2.

Une application quadratique est de la forme :

$$f(x_1, x_2, \dots, x_n) = \underbrace{\frac{1}{2} \sum_{i=1}^n a_{ii} x_i^2 + \sum_{i < j} a_{ij} x_i x_j}_{\text{forme quadratique}} - \underbrace{\sum_{i=1}^n b_i x_i}_{\text{forme linéaire}} + \underbrace{c}_{\text{constante}}$$

En posant :

$$A = (a_{ij})_{\substack{i=1, \dots, n \\ j=1, \dots, n}} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \in \mathcal{M}_n(\mathbb{R}) \quad ; \quad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \in \mathbb{R}^n$$

et  $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n$  on écrit l'application quadratique  $f$  sous *forme matricielle* :

$$f(\mathbf{x}) = \frac{1}{2} \langle A\mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x} \rangle + c$$

ou encore :

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top A \mathbf{x} - \mathbf{b}^\top \mathbf{x} + c$$

L'intérêt ne réside pas que dans la concision de l'écriture : on obtient immédiatement le vecteur gradient et la matrice Hessienne :

**Théorème II.8 (Gradient, matrice hessienne, d'une application quadratique.)**

Soit  $f(\mathbf{x}) = \frac{1}{2} \langle A\mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x} \rangle + c$  une application quadratique. Alors  $f$  est infiniment différentiable et,

$$\begin{aligned} \nabla f(\mathbf{x}) &= A\mathbf{x} - \mathbf{b}, \\ \nabla^2 f(\mathbf{x}) &= A. \end{aligned}$$

**Démonstration.**  $f$  est polynomiale, et donc infiniment différentiable. Pour tout  $i = 1, 2, \dots, n$ , le calcul donne  $\frac{\partial f}{\partial x_i}(\mathbf{x}) = \sum_{j=1}^n a_{ij} x_j - b_i$ , donc  $\nabla f(\mathbf{x}) = A\mathbf{x} - \mathbf{b}$ . Pour tout  $i, j = 1, 2, \dots, n$ ,  $\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = a_{ij}$ , et l'on obtient  $\nabla^2 f(\mathbf{x}) = A$ . □

## II.4.2 Programmation quadratique

La convexité d'une application quadratique est totalement caractérisée par sa matrice hessienne (contrairement au cas d'une application quelconque, comparer avec le théorème II.5.2). De plus dans ce cas strictement convexe = fortement convexe.

**Théorème II.9 (Convexité d'une application quadratique.)** *Soit l'application quadratique  $f(\mathbf{x}) = \frac{1}{2}\langle A\mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x} \rangle + c$ . Alors :*

- $f$  convexe  $\iff A$  semi-définie positive
- $f$  strictement convexe  $\iff f$  fortement convexe  $\iff A$  définie positive.

**Démonstration.** Puisque  $\nabla^2 f(\mathbf{x}) = A$ , par définition  $A$  est définie positive si et seulement si  $f$  est  $\lambda_1$ -convexe (i.e. fortement convexe) où  $\lambda_1 > 0$  désigne la plus petite valeur propre de  $A$ . La forte convexité impliquant la convexité stricte, avec le théorème II.5 il ne reste plus qu'à montrer que si  $f$  est strictement convexe alors  $A$  est définie positive. Supposons donc que  $f$  est strictement convexe; en particulier  $f$  est convexe et donc  $A$  est semi-définie positive. Procédons par l'absurde en supposant que  $A$  n'est pas définie positive : ainsi  $A$  admet 0 pour valeur propre, et soit  $E_0 = \ker A$  le sous-espace propre associé; il est de dimension au moins 1 et pour tout  $\mathbf{x} \in E_0$ ,  $\langle A\mathbf{x}, \mathbf{x} \rangle = 0$ . Puisque  $f$  est quadratique le développement de Taylor-Young à l'ordre 2 s'écrit ici :

$$\forall \mathbf{x} \in \mathbb{R}^n, f(\mathbf{u} + \mathbf{x}) - f(\mathbf{u}) = \langle \nabla f(\mathbf{u}), \mathbf{x} \rangle + \frac{1}{2}\langle A\mathbf{x}, \mathbf{x} \rangle$$

Ainsi,  $\forall \mathbf{x} \in E_0$ ,

$$f(\mathbf{u} + \mathbf{x}) = f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{x} \rangle$$

et donc la restriction de  $f$  à  $E_0$  est une application affine, et donc convexe mais non strictement convexe. Cela contredit le fait que  $f$  est strictement convexe. Ainsi  $A$  est définie positive.  $\square$

Nous pouvons maintenant caractériser les extrema d'une application quadratique.

**Théorème II.10 (Programmation quadratique.)** *Soit  $f(\mathbf{x}) = \frac{1}{2}\langle A\mathbf{x}, \mathbf{x} \rangle - \langle \mathbf{b}, \mathbf{x} \rangle + c$  une application quadratique sur  $\mathbb{R}^n$  et  $\mathbf{u} \in \mathbb{R}^n$ . Si  $A$  est semi-définie positive (resp. négative), alors les propositions suivantes sont équivalentes :*

- $\mathbf{u}$  est un minimum (resp. maximum) local de  $f$ ,
- $\mathbf{u}$  est un minimum (resp. maximum) global de  $f$ ,
- $A\mathbf{u} = \mathbf{b}$ , i.e.  $\mathbf{u}$  est solution du système d'équations linéaires  $A\mathbf{x} = \mathbf{b}$ .

*Si  $A$  est définie positive  $f$  admet un unique minimum (resp. maximum) global.*

*Si  $A$  n'est pas semi-définie positive (resp. négative)  $f$  n'admet aucun minimum (resp.) maximum local ou global.*

**Démonstration.** Puisque  $A$  est semi-définie positive (resp. négative),  $f$  (resp.  $-f$ ) est convexe et l'équivalence des 3 assertions découle des théorèmes II.6 et II.8. Si  $A$  est définie positive,  $\det(A) \neq 0$ , le système  $A\mathbf{u} = \mathbf{b}$  est de Cramer et  $f$  admet donc un unique extremum. Si  $A$  n'est pas semi-définie positive (resp. négative) la condition nécessaire du second ordre (théorème II.4.1) avec le théorème II.8 impliquent que  $f$  n'a aucun minimum (resp. maximum) local ni donc aussi global.  $\square$

**Exemple.** Soit  $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{b}^\top \mathbf{x}$  avec  $A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$  et  $\mathbf{b} = \begin{pmatrix} -3 \\ 1 \\ -2 \end{pmatrix}$ . Le polynôme caractéristique de  $A$  est

$$p_A(\lambda) = 8 - 12\lambda + 6\lambda^2 - \lambda^3 = \prod_{i=1}^3 \lambda_i - \lambda \sum_{i \neq j} \lambda_i \lambda_j + \lambda^2 \sum_{i=1}^3 \lambda_i - \lambda^3,$$

où  $\lambda_1, \lambda_2, \lambda_3$  désignent les valeurs propres de  $A$  ( $A$  est diagonalisable puisque symétrique réelle). Ainsi (cf. théorème A.4),

$$\prod_{i=1}^3 \lambda_i = 8 > 0, \quad \sum_{i \neq j} \lambda_i \lambda_j = 12 > 0, \quad \sum_{i=1}^3 \lambda_i = 6 > 0 \quad \implies \quad \lambda_1, \lambda_2, \lambda_3 > 0$$

et donc  $A$  est définie positive  $\implies f$  a un unique minimum global qui est l'unique solution de  $A\mathbf{x} = \mathbf{b}$ .

$$A\mathbf{x} = \mathbf{b} \iff \begin{cases} 2x & -y & & = & -3 \\ -x & +2y & -z & = & 1 \\ & -y & +2z & = & -2 \end{cases} \iff \mathbf{x} = \begin{pmatrix} -9/4 \\ -3/2 \\ -7/4 \end{pmatrix} \underline{\text{minimum de } f}.$$



## Exercices.

**Exercice 1.** Déterminer les extrema locaux et globaux de l'application  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  définie par :

$$f(x, y) = x^3 + y^3 + x^2 + y^2 - 1 .$$

**Exercice 2.** On considère l'application  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  définie par :

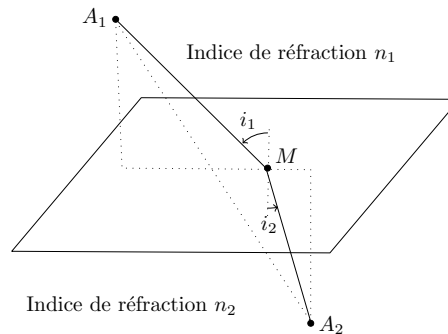
$$f(x, y) = x^4 + y^4 - x^3 - y^3 .$$

- Que peut-on dire de l'existence d'extrema globaux pour  $f$  ?
- Déterminer tous les extrema globaux de  $f$ .
- Montrer le résultat :

*Soit  $g : \mathbb{R}^n \rightarrow \mathbb{R}$  une application différentiable et  $\mathbf{u}$  un point critique de  $g$ , alors  $\mathbf{u}$  est un minimum local de  $g$  si et seulement si  $g$  est convexe sur une boule ouverte centrée en  $\mathbf{u}$ .*

- En déduire tous les extrema locaux de  $f$ .

**Exercice 3.** Un rayon lumineux effectue un trajet spatial d'un point  $A_1$  situé dans un milieu ayant pour indice de réfraction  $n_1$  à un point  $A_2$  situé dans un milieu ayant pour indice de réfraction  $n_2$  ; les deux milieux étant séparés par un plan.



En appliquant le principe que la lumière parcourt le trajet le plus rapide, retrouver la loi de Descartes de réfraction de la lumière :  $n_1 \sin i_1 = n_2 \sin i_2$ .

**Exercice 4.** Le but de l'exercice est de prouver le *théorème de projection convexe* :

*Soit  $\mathcal{C}$  un sous-ensemble convexe fermé non vide de  $\mathbb{R}^n$ . Donnée  $\mathbf{u} \in \mathbb{R}^n$  il existe un unique point  $P_{\mathcal{C}}(\mathbf{u}) \in \mathcal{C}$ , tel que :*

$$\|P_{\mathcal{C}}(\mathbf{u}) - \mathbf{u}\| = \min_{\mathbf{v} \in \mathcal{C}} \|\mathbf{v} - \mathbf{u}\| .$$

*On l'appelle le projeté de  $\mathbf{u}$  sur  $\mathcal{C}$ . Il est caractérisé par :*

$$\forall \mathbf{v} \in \mathcal{C}, \langle P_{\mathcal{C}}(\mathbf{u}) - \mathbf{u}, \mathbf{v} - P_{\mathcal{C}}(\mathbf{u}) \rangle \geqslant 0 .$$

De plus l'application  $P_C$  est contractante, i.e. :

$$\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \|P_C(\mathbf{x}) - P_C(\mathbf{y})\| \leq \|\mathbf{x} - \mathbf{y}\| .$$

- a. Prouver l'existence et l'unicité de  $P_C(\mathbf{u})$ .
- b. Prouver la caractérisation donnée de  $P_C(\mathbf{u})$ .
- c. Utiliser cette caractérisation pour prouver que  $P_C$  est une application contractante.

**Exercice 5.** Le but de l'exercice est de prouver la proposition II.1 :

Soient  $\mathcal{D} \subset \mathbb{R}^n$  connexe,  $f : \mathcal{D} \rightarrow \mathbb{R}$  continue, et  $\mathbf{u} \in \mathcal{D}$  un min (resp. max) local de  $f$ . Alors  $\mathbf{u}$  est un min (resp. max) global de  $f$  ssi  $\forall \mathbf{x}$  tel que  $f(\mathbf{x}) = f(\mathbf{u})$ ,  $\mathbf{x}$  est un min (resp. max) local de  $f$ .

Sans perte de généralité, quitte à changer  $f$  en  $-f$ , on la montrera pour  $\mathbf{u}$  un min local.

Soit  $u = f(\mathbf{u}) \in \mathbb{R}$ .

1. Montrer que  $f^{-1}(]-\infty, u])$  est un ouvert de  $\mathcal{D}$ .
2. Montrer que  $\mathcal{C}_D f^{-1}(]-\infty, u])$  est un voisinage de tout point de  $f^{-1}(\{u\})$  pour  $v > u$ .
3. Soit  $\mathbf{x} \in f^{-1}(\{u\})$ ; appliquer l'hypothèse que  $\mathbf{x}$  est un min local pour montrer que  $\mathcal{C}_D f^{-1}(]-\infty, u])$  est un voisinage de  $\mathbf{x}$ .
4. Dédire de 2 et 3 que  $f^{-1}(]-\infty, u])$  est un fermé de  $\mathcal{D}$ .
5. Appliquer la connexité de  $\mathcal{D}$  avec 1 et 4 pour montrer que  $f^{-1}(]-\infty, u]) = \emptyset$ . Conclure.

## Chapitre III

# Programmation sous contraintes

**Problème :** Soit  $\mathcal{D}$  un sous-ensemble propre (*i.e.*  $\neq \mathbb{R}^n$ ) et non vide de  $\mathbb{R}^n$ . Soit l'application  $f : \mathcal{D} \rightarrow \mathbb{R}$  dont on cherche les extrema.

- Lorsque  $\mathcal{D}$  est un ouvert de  $\mathbb{R}^n$  et  $f$  est différentiable (1 ou 2 fois) sur  $\mathcal{D}$  les notions vues au chapitre II s'appliquent pour étudier les extrema locaux, et on peut dans certains cas en déduire les extrema globaux de  $f$ .
- Lorsque  $\mathcal{D}$  n'est pas un ouvert, les notions du chapitre II s'avèrent insuffisantes pour étudier les extrema locaux et globaux de  $f$ .

Comment généraliser les conditions du 1<sup>er</sup> et 2<sup>e</sup> ordre vues au chapitre II dans le cas sous contraintes ? C'est l'objet de ce chapitre.

Nous procédons en deux étapes. Nous considérons dans une première partie le cas plus restrictif où toutes les contraintes sont égalitaires ; l'équation d'Euler se généralise par les conditions de Lagrange. Nous voyons ensuite dans une deuxième partie le cas général sous contraintes égalitaires et inégalitaires ; les conditions de Lagrange s'y généralisent par les conditions de Karush-Kuhn-Tucker.

### III.1 Optimisation sous contraintes égalitaires

#### III.1.1 Enoncé du problème

Soit  $\mathcal{U}$  un ouvert non vide de  $\mathbb{R}^n$  (le plus souvent  $\mathcal{U} = \mathbb{R}^n$ ). Soit  $f : \mathcal{U} \rightarrow \mathbb{R}$  une application différentiable sur  $\mathcal{U}$ , et  $\varphi_1, \varphi_2, \dots, \varphi_p : \mathcal{U} \rightarrow \mathbb{R}$  des applications de classe  $C^1$  sur  $\mathcal{U}$  (*i.e.*  $\varphi_i$  est différentiable et  $\nabla \varphi_i : x \mapsto \nabla \varphi_i(x)$  est continue ; c'est le cas en particulier lorsque  $\varphi_i$  est 2 fois différentiable). Soit le domaine  $\mathcal{D}$  :

$$\mathcal{D} = \{\mathbf{x} \in \mathcal{U} \mid \varphi_i(\mathbf{x}) = 0, \forall i = 1, 2, \dots, p\}.$$

Le problème :

$$\min_{\mathbf{x} \in \mathcal{D}} f(\mathbf{x}) \quad (\text{respectivement } \max_{\mathbf{x} \in \mathcal{D}} f(\mathbf{x}))$$

est un problème de minimisation (respectivement de maximisation) sous contraintes égalitaires.

**Remarque.** Lorsque  $\mathcal{U} = \mathbb{R}^n$ ,  $\mathcal{D} = \bigcap_{i=1}^p \varphi_i^{-1}(\{0\})$  est un fermé de  $\mathbb{R}^n$ , et très souvent d'intérieur vide. Or l'équation d'Euler ne s'applique que dans l'intérieur de  $\mathcal{D}$ .

### III.1.2 Exemples en dimension 2.

Nous voyons ici deux exemples, en dimension 2, qui vont nous permettre par une approche géométrique de dégager les idées directrices pour nous conduire aux conditions de Lagrange.

• **Exemple A.** Soit  $\mathcal{D} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$  le cercle unité de centre  $(0, 0)$ . Soit  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $f(x, y) = x$ . L'application  $f$  n'a aucun point critique sur  $\mathcal{D}$  :  $\nabla f(x, y) = (1, 0) \neq \mathbf{0}$ . Or  $\mathcal{D}$  est un compact (car fermé et borné) de  $\mathbb{R}^n \implies \exists$  un minimum et un maximum de  $f$  sur  $\mathcal{D}$ . (On constate que l'équation d'Euler ne s'applique pas ici.) On a deux solutions évidentes, un maximum  $(1, 0)$  et un minimum  $(-1, 0)$ .

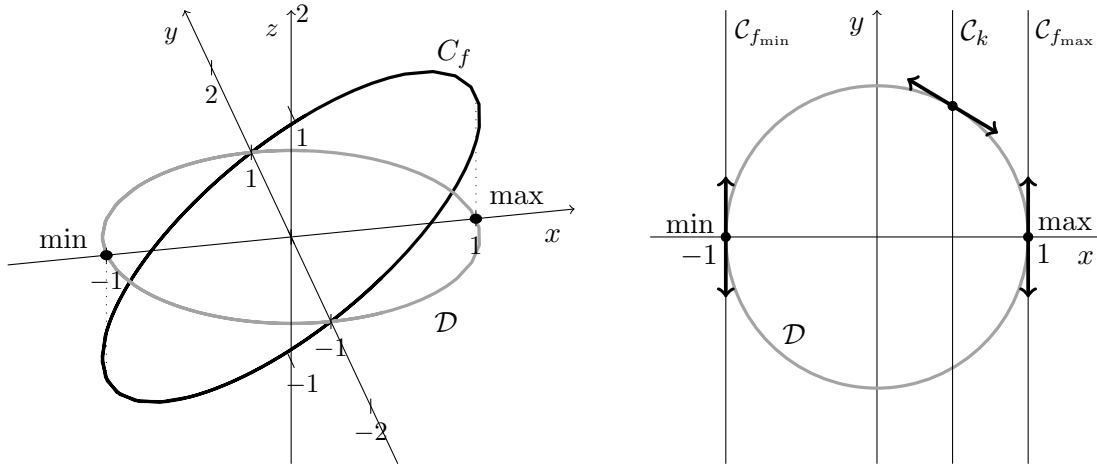


FIGURE III.1 – Sur la figure de gauche : la courbe représentative  $C_f$  dans  $\mathbb{R}^3$  de  $f$  au-dessus du domaine  $\mathcal{D}$ , ainsi que le minimum et le maximum. Sur la figure de droite : le domaine  $\mathcal{D}$  et les lignes de niveau dans  $\mathbb{R}^2$  ; aux extrema les lignes de niveau sont tangentes au domaine.

Graphiquement (cf. figure III.1) on constate qu'aux extrema trouvés les courbes de niveau sont tangentes au domaine.

• **Exemple B.** Soit  $\mathcal{D} = \{(x, y) \in \mathbb{R}^2 \mid xy = 1\}$  et  $f(x, y) = x^2 + y^2$ . L'application  $f$  est coercive sur le fermé (non borné)  $\mathcal{D} \implies \exists$  un minimum et  $\nexists$  de maximum de  $f$  sur  $\mathcal{D}$ .

On se ramène à un problème sans contrainte à une seule variable, en utilisant la contrainte pour supprimer une variable. Soit  $y = \frac{1}{x}$ ,  $f_x : \mathbb{R}_* \rightarrow \mathbb{R}$ ,  $f_x(x) = f(x, \frac{1}{x}) = \frac{x^4+1}{x^2}$ . On étudie les variations de  $f_x$ , sa dérivée est  $f'_x(x) = 2\frac{x^4-1}{x^3}$ .

$x$	$-\infty$	$-1$	$0$	$1$	$+\infty$	
$f'_x$	$-$	$0$	$+$	$-$	$0$	$+$
$f_x$	$+\infty$	$\searrow$	$2$	$\nearrow$	$+\infty$	$+\infty$

Ainsi  $f_x$  a deux minima globaux :  $x = \pm 1 \Rightarrow f$  a deux minima globaux sur  $\mathcal{D}$  : **a**(1,1) et **b**(-1,-1).

On trace dans  $\mathbb{R}^2$  le domaine  $\mathcal{D}$  ainsi que des courbes de niveau de  $f$ . Une fois de plus on constate qu'aux extrema trouvés les courbes de niveau de  $f$  sont tangentes au domaine  $\mathcal{D}$  (cf. figure III.2).

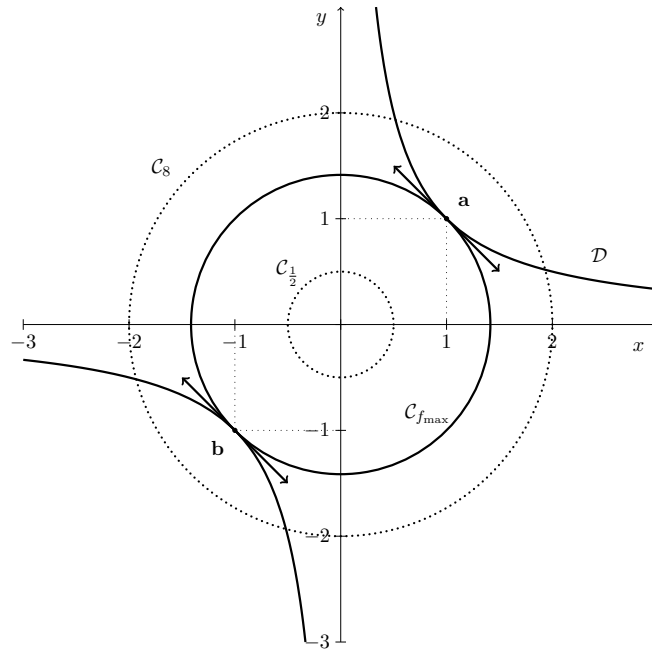


FIGURE III.2 – Le domaine admissible  $\mathcal{D} \subset \mathbb{R}^2$  et 3 courbes de niveau.

- Que dire de cette constatation ? Est-ce une coïncidence ou un principe général ? Dans ce dernier cas est-ce une condition nécessaire, suffisante, à l'existence d'extrema ?

Soit  $\mathcal{D} = \{\mathbf{x} \in \mathbb{R}^2 \mid \varphi(\mathbf{x}) = 0\}$ ; puisque  $\varphi$  est de classe  $C^1$  et que  $\nabla\varphi(\mathbf{a}) \neq 0$ , l'espace tangent à  $\mathcal{D}$  en  $\mathbf{a}$  est engendré par un vecteur  $\mathbf{d}_\mathbf{a} \neq \mathbf{0}$  (cf. théorème A.2). Si  $\mathbf{x}$  est dans un voisinage de  $\mathbf{a}$  dans  $\mathcal{D}$ , alors  $\mathbf{x} = \mathbf{a} + \lambda\mathbf{d}_\mathbf{a} + o(|\lambda|)$  pour  $\lambda$  dans un voisinage de 0. Or puisque  $f$  est différentiable on a la formule de Taylor à l'ordre 1,  $f(\mathbf{x}) = f(\mathbf{a}) + \langle \nabla f(\mathbf{a}), \mathbf{x} - \mathbf{a} \rangle + o(\|\mathbf{x} - \mathbf{a}\|)$ . En prenant  $\mathbf{x}$  dans  $\mathcal{D}$ ,  $\mathbf{x} = \mathbf{a} + \lambda\mathbf{d}_\mathbf{a} + o(|\lambda|)$ , ainsi :

$$\begin{aligned} f(\mathbf{x}) &= f(\mathbf{a}) + \langle \nabla f(\mathbf{a}), \lambda\mathbf{d}_\mathbf{a} \rangle + o(|\lambda|) \\ f(\mathbf{x}) &\approx f(\mathbf{a}) + \lambda \langle \nabla f(\mathbf{a}), \mathbf{d}_\mathbf{a} \rangle \end{aligned}$$

$\implies$  si  $\langle \nabla f(\mathbf{a}), \mathbf{d}_\mathbf{a} \rangle \neq 0$  alors  $f(\mathbf{x}) - f(\mathbf{a})$  ne garde pas un signe constant lorsque  $\mathbf{x}$  est dans un voisinage de  $\mathbf{a}$  dans  $\mathcal{D}$ .

$\implies$  Si  $\langle \nabla f(\mathbf{a}), \mathbf{d}_\mathbf{a} \rangle \neq 0$  alors  $\mathbf{a}$  n'est pas un extremum local.

Or en tout point  $\mathbf{x}$  le vecteur gradient a la propriété d'être orthogonal à la courbe de niveau (cela se vérifie aisément à l'aide du développement de Taylor-Young à l'ordre 1).

$\implies$  Une condition nécessaire pour que  $\mathbf{a}$  soit un extremum local est bien que la courbe de niveau passant par  $\mathbf{a}$  soit tangente à  $\mathcal{D}$ . Notre constatation s'avère être une condition nécessaire à l'existence d'extrema.

On peut l'exprimer par une équation. L'équation de la tangente à la courbe de niveau passant par  $\mathbf{a}$  est :

$$\langle \nabla f(\mathbf{a}), \mathbf{x} - \mathbf{a} \rangle = 0.$$

L'équation de la tangente au domaine  $\mathcal{D}$  en  $\mathbf{a}$  est :

$$\langle \nabla\varphi(\mathbf{a}), \mathbf{x} - \mathbf{a} \rangle = 0.$$

Donc la condition nécessaire s'écrit : " $\nabla f(\mathbf{a})$  et  $\nabla\varphi(\mathbf{a})$  sont colinéaires". On aboutit à : Si  $\mathbf{a}$  un extremum local alors  $\exists \lambda_1, \lambda_2$  non tous deux nuls, tels que :

$$\lambda_1 \nabla f(\mathbf{a}) + \lambda_2 \nabla\varphi(\mathbf{a}) = \mathbf{0}$$

Nous allons généraliser cette relation (Conditions de Lagrange) pour énoncer un principe plus général : le principe de Lagrange.

### III.1.3 Principe de Lagrange

Il s'agit d'une condition nécessaire à l'existence d'un extremum local pour un problème sous contrainte égalitaire, généralisant l'équation d'Euler. Plus généralement l'étude des extrema de  $f$  sur  $\mathcal{D}$  se ramène à l'étude d'extrema sans contrainte d'une fonction appelée le *Lagrangien du problème*.

**Théorème III.1 (Condition de Lagrange.)** Soient  $f$  une application différentiable sur un ouvert  $\mathcal{U}$  non vide de  $\mathbb{R}^n$  et  $\varphi_1, \dots, \varphi_p$  des applications de classe  $C^1$  sur  $\mathcal{U}$ .

$$\mathcal{D} = \{\mathbf{x} \in \mathcal{U} \mid \varphi_i(\mathbf{x}) = 0, \forall i = 1, 2, \dots, p\}$$

Si  $\mathbf{u} \in \mathcal{D}$  est un extremum local de  $f$  sur  $\mathcal{D}$ , et si les vecteurs  $\nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u})$  forment une famille linéairement indépendante, alors  $\exists! \lambda_1, \lambda_2, \dots, \lambda_p \in \mathbb{R}$ , appelés multiplicateurs de Lagrange, tels que :

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

**Démonstration.** Soit  $\mathbf{u} \in \mathcal{D}$  un point vérifiant les hypothèses du théorème. Le fait que  $\varphi_1, \dots, \varphi_p$  soient de classe  $C^1$  et que la famille  $\nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u})$  soit linéairement indépendante a pour conséquence l'existence d'un espace tangent à  $\mathcal{D}$  en  $\mathbf{u}$  de codimension  $p$  (cf. théorème A.2), qui est :

$$T_{\mathbf{u}}\mathcal{D} = \langle \nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u}) \rangle^\perp = \{ \mathbf{x} \in \mathbb{R}^n \mid \langle \nabla\varphi_i(\mathbf{u}), \mathbf{x} \rangle = 0, \forall i = 1, \dots, p \}$$

Soit  $\mathbf{e}_1, \dots, \mathbf{e}_{n-p}$  une base orthogonale de  $T_{\mathbf{u}}\mathcal{D}$ . Lorsque  $\mathbf{x}$  est dans un voisinage de  $\mathbf{u}$  dans  $\mathcal{D}$ , il existe  $\mathbf{a} = (a_1, \dots, a_{n-p}) \in \mathbb{R}^{n-p}$  tel que  $\mathbf{x} = \mathbf{u} + \sum_{i=1}^{n-p} a_i \mathbf{e}_i + o(\|\mathbf{a}\|)$  et  $\mathbf{a}$  décrit un voisinage de  $\mathbf{0}$  dans  $\mathbb{R}^{n-p}$  lorsque  $\mathbf{x}$  décrit un voisinage de  $\mathbf{u}$ . Or on a au voisinage de  $\mathbf{u}$  dans  $\mathcal{U}$  le développement de Taylor-Young à l'ordre 1 :  $f(\mathbf{x}) = f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{x} - \mathbf{u} \rangle + o(\|\mathbf{x} - \mathbf{u}\|)$ . Pour  $i = 1, \dots, p$ , notons  $\mathbf{x}(a_i)$  le projeté orthogonal de  $\mathbf{x}$  sur la droite affine  $\mathbf{u} + \langle \mathbf{e}_i \rangle$ . On a alors  $\mathbf{x}(a_i) = \mathbf{u} + a_i \mathbf{e}_i + o(|a_i|)$  et  $f(\mathbf{x}(a_i)) - f(\mathbf{u}) = a_i \langle \nabla f(\mathbf{u}), \mathbf{e}_i \rangle + o(|a_i|)$  pour  $a_i$  dans un voisinage de 0. Ainsi, si  $\langle \nabla f(\mathbf{u}), \mathbf{e}_i \rangle \neq 0$ ,  $f(\mathbf{x}(a_i)) - f(\mathbf{u})$  ne garde pas un signe constant lorsque  $a_i$  décrit un voisinage de 0, ce qui contredit le fait que  $\mathbf{u}$  soit un extremum de  $f$  sur  $\mathcal{D}$ . Ainsi, nécessairement,  $\forall i = 1, \dots, n-p$ ,  $\langle \nabla f(\mathbf{u}), \mathbf{e}_i \rangle = 0$ . Donc  $\nabla f(\mathbf{u}) \in \langle \mathbf{e}_1, \dots, \mathbf{e}_{n-p} \rangle^\perp = \langle \nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u}) \rangle$ ; ainsi  $\exists \lambda_1, \dots, \lambda_p \in \mathbb{R}$  tels que  $\nabla f(\mathbf{u}) = \sum_{i=1}^p \lambda_i \nabla\varphi_i(\mathbf{u})$ . Ils sont uniques puisque la famille  $\nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u})$  est libre.  $\square$

**Interprétation géométrique.** Si les vecteurs  $\nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u})$  sont linéairement indépendants l'espace tangent  $T_{\mathbf{u}}\mathcal{D}$  à  $\mathcal{D}$  en  $\mathbf{u}$  existe et (cf. théorème A.2 page 116) :

$$T_{\mathbf{u}}\mathcal{D} = \{ \mathbf{x} \in \mathbb{R}^n \mid \langle \nabla\varphi_i(\mathbf{u}), \mathbf{x} \rangle = 0, \forall i = 1, \dots, p \} .$$

Les conditions de Lagrange expriment qu'en un point extremum, le vecteur gradient est orthogonal à l'espace tangent. Noter que  $\nabla f(\mathbf{u})$  (resp.  $-\nabla f(\mathbf{u})$ ) est la direction locale de plus grand accroissement (resp. de plus grande décroissance) de  $f$ , et qu'elle est orthogonale aux hypersurfaces de niveau. Ainsi, on retrouve la constatation déjà faite, qu'en un extremum local l'hypersurface de niveau est tangente au domaine.

**Remarque.** Plus généralement, sans l'hypothèse (*de qualification des contraintes*) " $\nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u})$  sont linéairement indépendants", on a, tout du moins, le résultat plus faible suivant :

$\mathbf{u} \in \mathcal{D}$  est un extremum local  $\implies \exists \lambda_0, \lambda_1, \dots, \lambda_p$  non tous nuls tels que :

$$\lambda_0 \nabla f(\mathbf{u}) + \sum_{i=1}^p \lambda_i \nabla\varphi_i(\mathbf{u}) = \mathbf{0}$$

Seulement lorsque  $\lambda_0 = 0$  l'équation n'est pas informative pour  $f$ ... Aussi on suppose que les  $\nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u})$  sont linéairement indépendants : cela assure que  $\lambda_0 \neq 0$ .

**Formulation lagrangienne.** Le Lagrangien du problème est l'application :

$$\begin{aligned} \mathcal{L} : \mathbb{R}^n \times \mathbb{R}^p &\longmapsto \mathbb{R} \\ (\mathbf{x}, \lambda) &\longrightarrow \mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_{i=1}^p \lambda_i \varphi_i(\mathbf{x}) \end{aligned}$$

Lorsque  $f, \varphi_1, \dots, \varphi_p$  sont différentiables, le vecteur gradient du Lagrangien en  $\mathbf{u} \in \mathbb{R}^n$  est :

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{u}, \lambda) = \begin{pmatrix} \frac{\partial \mathcal{L}}{\partial x_1}(\mathbf{u}, \lambda) \\ \vdots \\ \frac{\partial \mathcal{L}}{\partial x_n}(\mathbf{u}, \lambda) \end{pmatrix} = \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{u}) + \sum_{i=1}^p \lambda_i \frac{\partial \varphi_i}{\partial x_1}(\mathbf{u}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{u}) + \sum_{i=1}^p \lambda_i \frac{\partial \varphi_i}{\partial x_n}(\mathbf{u}) \end{pmatrix} = \nabla f(\mathbf{u}) + \sum_{i=1}^p \lambda_i \nabla \varphi_i(\mathbf{u}) .$$

(on ne dérive que par rapport aux variables primaires!).

La condition de Lagrange s'écrit alors, sous les hypothèses adéquates :

$$\boxed{\mathbf{u} \text{ est un extremum local} \implies \exists ! \lambda \in \mathbb{R}^p, \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{u}, \lambda) = \mathbf{0}}$$

On définit aussi, lorsque  $f, \varphi_1, \dots, \varphi_p$  sont deux fois différentiables, la matrice hessienne du Lagrangien en  $\mathbf{u} \in \mathbb{R}^n$ , par :

$$\nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda) = \left( \frac{\partial^2 \mathcal{L}}{\partial x_i \partial x_j}(\mathbf{u}, \lambda) \right)_{\substack{i=1 \dots n \\ j=1 \dots n}} = \nabla^2 f(\mathbf{u}) + \sum_{i=1}^p \lambda_i \nabla^2 \varphi_i(\mathbf{u}) .$$

La condition de lagrange n'est qu'une condition nécessaire; elle généralise dans le cas de contraintes égalitaires l'équation d'Euler (théorème II.3). Elle ne permet pas de déterminer si une solution trouvée est bien un extremum local. Nous allons améliorer ce critère en tenant compte, comme nous l'avons fait dans le cas sans contrainte, d'une part de la convexité et d'autre part de conditions nécessaires, suffisantes du second ordre.

### III.1.4 Prise en compte de la convexité

Dans le cas de la programmation convexe, nous allons établir que les conditions de Lagrange sont nécessaires et suffisantes à l'existence d'un minimum, qui plus est, global. De plus elles ne nécessitent plus aucune hypothèse de qualification des contraintes; cela est vrai plus généralement dès lors que les contraintes sont affines.

**Proposition III.1 (Simplification de l'énoncé sous contraintes affines.)** *Si toutes les contraintes  $\varphi_1, \dots, \varphi_p$  sont affines, les conditions de Lagrange restent vraies, à l'exception de l'unicité des multiplicateurs de lagrange, sans l'hypothèse de qualification des contraintes :  $\nabla \varphi_1(\mathbf{u}), \dots, \nabla \varphi_p(\mathbf{u})$  est libre.*



**Démonstration.** Lorsque  $\varphi_1, \dots, \varphi_p$  sont affines, les vecteurs  $\nabla\varphi_1(\mathbf{x}), \dots, \nabla\varphi_p(\mathbf{x})$  ne dépendent pas de  $\mathbf{x} \in \mathcal{U}$ . Si la famille  $\nabla\varphi_1(\mathbf{x}), \dots, \nabla\varphi_p(\mathbf{x})$  n'est pas libre alors, sans perte de généralité,  $\nabla\varphi_p(\mathbf{x})$  est une combinaison linéaire de  $\nabla\varphi_1(\mathbf{x}), \dots, \nabla\varphi_{p-1}(\mathbf{x}) : \forall \mathbf{x} \in \mathcal{U}, \nabla\varphi_p(\mathbf{x}) = \sum_{i=1}^{p-1} \rho_i \nabla\varphi_i(\mathbf{x})$ . Or puisque les  $\varphi_i$  sont affines  $\varphi_i(\mathbf{x}) = \langle \nabla\varphi_i(\mathbf{x}), \mathbf{x} \rangle + k_i$ ; ainsi on obtient,  $\forall \mathbf{x} \in \mathcal{U}, \varphi_p(\mathbf{x}) = k + \sum_{i=1}^{p-1} \rho_i \varphi_i(\mathbf{x})$ . En particulier en prenant  $\mathbf{x} \in \mathcal{D}, k = 0$  et la contrainte  $\varphi_p(\mathbf{x}) = 0$  est redondante de sorte qu'on peut la supprimer. On procède de même tant que c'est possible pour aboutir à une sous-famille de contraintes égalitaires  $\varphi_1, \dots, \varphi_r$ , pour  $1 \leq r \leq p$ , vérifiant  $\nabla\varphi_1(\mathbf{x}), \dots, \nabla\varphi_r(\mathbf{x})$  est libre. Pour cette sous-famille, en un extremum local  $\mathbf{u} \in \mathcal{D}$  les conditions de Lagrange s'appliquent et donc  $\exists \lambda_1, \dots, \lambda_r \in \mathbb{R}$  tel que  $\nabla f(\mathbf{u}) + \sum_{i=1}^r \lambda_i \nabla\varphi_i(\mathbf{u}) = 0$ . En posant  $\lambda_{r+1} = \dots = \lambda_p = 0$ , on a aussi  $\nabla f(\mathbf{u}) + \sum_{i=1}^p \lambda_i \nabla\varphi_i(\mathbf{u}) = 0$ .  $\square$

**Théorème III.2 (CNS en programmation convexe.)** *Si  $\mathcal{U}$  est un ouvert convexe, si  $f$  est différentiable et convexe et si  $\varphi_1, \dots, \varphi_p$  sont affines, alors  $\mathbf{u}$  est un minimum global de  $f$  sur  $\mathcal{D} = \{\mathbf{x} \in \mathcal{U} \mid \varphi_i(\mathbf{x}) = 0, \forall i = 1, \dots, p\}$ , si et seulement si  $\exists \lambda_1, \dots, \lambda_p \in \mathbb{R}$  tels que :*

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

**Démonstration.** Soit  $\mathbf{u} \in \mathcal{D}$  un minimum de  $f$  sur  $\mathcal{D}$ . Avec la proposition III.1 on peut appliquer le théorème III.1 et on obtient les conditions nécessaires de Lagrange en  $\mathbf{u}$ . Pour la réciproque on renvoie à la preuve du théorème III.6 où elle est démontrée dans un cadre plus général.  $\square$

### III.1.5 Conditions, nécessaire, suffisante, du second ordre

En l'absence de l'hypothèse de convexité les conditions de Lagrange n'offrent qu'une condition nécessaire à l'existence d'extrema (qui plus est sous des conditions suffisantes de qualification des contraintes), et il est utile d'établir des conditions nécessaire, suffisante à l'ordre 2, comme nous l'avons fait dans le cas sans contrainte. C'est ce que nous faisons ici.

Soit  $\mathcal{U}$  un ouvert de  $\mathbb{R}^n$ , soit  $\varphi_1, \dots, \varphi_p : \mathcal{U} \rightarrow \mathbb{R}$  des applications de classe  $C^1$  et soit  $\mathcal{D} = \{\mathbf{x} \in \mathcal{U} \subset \mathbb{R}^n \mid \varphi_i(\mathbf{x}) = 0, \forall i = 1, \dots, p\} \subset \mathbb{R}^n$ . Si en  $\mathbf{u}$  les vecteurs  $\nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u})$  sont linéairement indépendants, alors l'espace tangent  $T_{\mathbf{u}}\mathcal{D}$  à  $\mathcal{D}$  en  $\mathbf{u}$  existe (voir page 116) et

$$T_{\mathbf{u}}\mathcal{D} = \{\mathbf{x} \in \mathbb{R}^n \mid \langle \nabla\varphi_i(\mathbf{u}), \mathbf{x} \rangle = 0, \forall i = 1, \dots, p\} .$$

C'est un sous-espace vectoriel de  $\mathbb{R}^n$  de dimension  $n - p$ .

**Théorème III.3 (Conditions, nécessaire, suffisante, du 2<sup>e</sup> ordre.)** *Soit  $\mathcal{U}$  un ouvert non vide de  $\mathbb{R}^n$  et soit  $f, \varphi_1, \dots, \varphi_p : \mathcal{U} \rightarrow \mathbb{R}$  des applications deux fois différentiables. Soit  $\mathbf{u} \in \mathcal{D}$  tel que  $\nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u})$  soient linéairement indépendants. Alors :*

(CN) *Si  $\mathbf{u}$  est un minimum (resp. maximum) local de  $f$  sur  $\mathcal{D}$ , alors :*

$$\begin{aligned} & \exists ! \lambda \in \mathbb{R}^p, \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{u}, \lambda) = 0, \text{ et} \\ & \forall \mathbf{x} \in T_{\mathbf{u}}\mathcal{D}, \langle \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda) \mathbf{x}, \mathbf{x} \rangle \geq 0 \text{ (resp. } \leq 0), \end{aligned}$$

(CS) Si

$$\begin{aligned} & \exists \lambda \in \mathbb{R}^p, \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{u}, \lambda) = 0, \text{ et} \\ & \forall \mathbf{x} \in T_{\mathbf{u}}\mathcal{D} \setminus \{0\}, \langle \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda) \mathbf{x}, \mathbf{x} \rangle > 0 \text{ (resp. } < 0), \end{aligned}$$

alors  $\mathbf{u}$  est un minimum (resp. maximum) local strict de  $f$  sur  $\mathcal{D}$ .

**Démonstration.** Soit  $\mathbf{x} \in \mathcal{D}$  un point dans un voisinage de  $\mathbf{u}$  dans  $\mathcal{D}$ , En écrivant le développement de Taylor-Young de  $f$  à l'ordre 2 au voisinage de  $\mathbf{u}$ ,

$$f(\mathbf{x}) - f(\mathbf{u}) = \langle \nabla f(\mathbf{u}), \mathbf{x} - \mathbf{u} \rangle + \frac{1}{2}(\mathbf{x} - \mathbf{u})^\top \nabla^2 f(\mathbf{u})(\mathbf{x} - \mathbf{u}) + o(\|\mathbf{x} - \mathbf{u}\|^2) \quad (E)$$

ainsi que le développement de Taylor-Young de  $\varphi_i$  à l'ordre 2 au voisinage de  $\mathbf{u}$  :

$$0 = \varphi_i(\mathbf{x}) - \varphi_i(\mathbf{u}) = \langle \nabla \varphi_i(\mathbf{u}), \mathbf{x} - \mathbf{u} \rangle + \frac{1}{2}(\mathbf{x} - \mathbf{u})^\top \nabla^2 \varphi_i(\mathbf{u})(\mathbf{x} - \mathbf{u}) + o(\|\mathbf{x} - \mathbf{u}\|^2) \quad (E_i)$$

Ainsi en formant l'équation  $(E) - \sum_{i=1}^p \lambda_i (E_i)$  on obtient en posant  $\lambda = (\lambda_1, \dots, \lambda_p)$  :

$$f(\mathbf{x}) - f(\mathbf{u}) = \mathcal{L}(\mathbf{x}, \lambda) - \mathcal{L}(\mathbf{u}, \lambda) = \langle \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{u}, \lambda), \mathbf{x} - \mathbf{u} \rangle + \frac{1}{2}(\mathbf{x} - \mathbf{u})^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda)(\mathbf{x} - \mathbf{u}) + o(\|\mathbf{x} - \mathbf{u}\|^2) \quad (*)$$

De plus, par définition de l'espace tangent,  $\mathbf{x} - \mathbf{u} = \mathbf{d} + o(\|\mathbf{x} - \mathbf{u}\|)$  où  $\mathbf{d} \in T_{\mathbf{u}}\mathcal{D}$ . Après avoir remarqué tout cela le même argument que celui utilisé dans la preuve du théorème II.4, en remplaçant la formule de Taylor-Young à l'ordre 2 par l'équation  $(*)$  et l'équation d'Euler (théorème II.3) par la condition de Lagrange (théorème III.1) prouve les deux conditions.  $\square$

**Remarques.** – La condition  $\forall \mathbf{x} \in T_{\mathbf{u}}\mathcal{D}, \langle \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda) \mathbf{x}, \mathbf{x} \rangle \geq 0$  est vérifiée notamment lorsque  $\nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda)$  est semi-définie positive (et similairement pour semi-définie négative).

– La condition  $\forall \mathbf{x} \in T_{\mathbf{u}}\mathcal{D} \setminus \{0\}, \langle \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda) \mathbf{x}, \mathbf{x} \rangle > 0$  est vérifiée notamment lorsque  $\nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda)$  est définie positive (et similairement pour définie négative).

– Comme dans le cas sans contraintes la condition suffisante s'utilise pour prouver qu'un point est extremum, tandis que la condition nécessaire s'utilise pour prouver qu'un point n'est pas un extremum. En un point critique  $\mathbf{u}$  du lagrangien, si l'on n'a que l'inégalité large  $\forall \mathbf{x} \in T_{\mathbf{u}}\mathcal{D}, (\mathbf{x} - \mathbf{u})^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda)(\mathbf{x} - \mathbf{u}) \geq 0$  (ou  $\leq 0$ ) on ne peut rien en déduire.

– A part dans le cas où  $\nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda)$  est définie positive ou négative, ces conditions sont bien moins pratiques à manier que les conditions d'ordre 2 dans le cas sans contrainte (théorème II.4) puisqu'on ne possède pas de critère simple ou général pour vérifier les inégalités sur  $T_{\mathbf{u}}\mathcal{D}$ . Aussi leur usage en est-il bien moins systématique. Cependant en l'absence d'informations supplémentaires (convexité,...) c'est ce que l'on peut faire de mieux, et elles peuvent s'avérer parfois très utiles.

**Exemple.** Soit  $f(x, y) = 3x^2 + 4y^2$  sur  $\mathcal{D} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$ . Puisque  $f$  est continue sur le compact  $\mathcal{D}$ , il existe un minimum et un maximum global de  $f$  sur  $\mathcal{D}$ .

$$\nabla f(x, y) = \begin{pmatrix} 6x \\ 8y \end{pmatrix} \quad ; \quad \nabla \varphi(x, y) = \begin{pmatrix} 2x \\ 2y \end{pmatrix} \quad ; \quad \nabla_{\mathbf{x}} \mathcal{L}(x, y, \lambda) = \begin{pmatrix} 6x + 2\lambda x \\ 8y + 2\lambda y \end{pmatrix}$$

Puisque  $\nabla\varphi(x, y) \neq 0$  sur  $\mathcal{D}$  on peut appliquer les conditions de Lagrange :

$$\begin{cases} 6x + 2\lambda x = 0 & (1) \\ 8y + 2\lambda y = 0 & (2) \\ x^2 + y^2 = 1 & (3) \end{cases}$$

$$\lambda \neq -4 \text{ ou } -3 \implies x = y = 0 \text{ impossible avec (3)}$$

$$\lambda = -4 \xrightarrow{(1)} x = 0 \xrightarrow{(3)} y = \pm 1$$

$$\lambda = -3 \xrightarrow{(2)} y = 0 \xrightarrow{(3)} x = \pm 1$$

On obtient 4 solutions :

$$\mathbf{a} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} ; \mathbf{b} = \begin{pmatrix} 0 \\ -1 \end{pmatrix} \text{ (avec } \lambda = -4)$$

$$\mathbf{c} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} ; \mathbf{d} = \begin{pmatrix} -1 \\ 0 \end{pmatrix} \text{ (avec } \lambda = -3)$$

En  $\mathbf{u} = \mathbf{a}$  ou  $\mathbf{b}$ ,  $\nabla\varphi(\mathbf{u}) = \begin{pmatrix} 0 \\ \pm 2 \end{pmatrix}$ , et donc  $T_{\mathbf{u}}\mathcal{D} = \left\langle \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right\rangle$ .

$$\nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, -4) = \begin{pmatrix} -2 & 0 \\ 0 & 0 \end{pmatrix}$$

Pour tout  $\mathbf{x} = (x, 0) \neq 0 \in T_{\mathbf{u}}\mathcal{D}$ ,  $\mathbf{x}^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, -4) \mathbf{x} = -2x^2 < 0$ . Donc  $\mathbf{a}$  et  $\mathbf{b}$  sont deux maxima locaux.

En  $\mathbf{u} = \mathbf{c}$  ou  $\mathbf{d}$ ,  $\nabla\varphi(\mathbf{u}) = \begin{pmatrix} \pm 2 \\ 0 \end{pmatrix}$ , et donc  $T_{\mathbf{u}}\mathcal{D} = \left\langle \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\rangle$ .

$$\nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, -3) = \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix}$$

Pour tout  $\mathbf{y} = (0, y) \neq 0 \in T_{\mathbf{u}}\mathcal{D}$ ,  $\mathbf{y}^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, -3) \mathbf{y} = 2y^2 > 0$ . Donc  $\mathbf{c}$  et  $\mathbf{d}$  sont deux minima locaux.

Tous les extrema sont globaux, par compacité, et car  $f(\mathbf{a}) = f(\mathbf{b})$  et  $f(\mathbf{c}) = f(\mathbf{d})$ .

### III.1.6 Programmation quadratique sous contraintes égalitaires

Nous appliquons maintenant à la programmation quadratique sous contraintes égalitaires les notions vues ci-dessus, plus précisément les conditions de Lagrange et la prise en compte de la convexité.

Soit  $f(\mathbf{u}) = \frac{1}{2}\mathbf{u}^\top A\mathbf{u} - \mathbf{b}^\top \mathbf{u}$  où  $A \in \mathcal{M}_n(\mathbb{R})$  est symétrique et  $\mathbf{b} \in \mathbb{R}^n$ , ainsi que les contraintes affines :

$$\begin{cases} \varphi_1(\mathbf{u}) = \sum_{j=1}^n m_{1j}x_j = c_1 \\ \vdots \\ \varphi_i(\mathbf{u}) = \sum_{j=1}^n m_{ij}x_j = c_i \\ \vdots \\ \varphi_p(\mathbf{u}) = \sum_{j=1}^n m_{pj}x_j = c_p \end{cases}$$

On note  $M = (m_{ij})_{\substack{i=1\dots p \\ j=1\dots n}} \in \mathcal{M}_{p,n}(\mathbb{R})$  et  $\mathbf{c} = (c_1, \dots, c_p) \in \mathbb{R}^p$ .

Les contraintes étant affines on pourra se passer de l'hypothèse de qualification des contraintes pour appliquer les conditions de Lagrange (cf. proposition III.1 plus haut). Elles s'écrivent ici :

$$\begin{aligned} \nabla_X \mathcal{L}(\mathbf{u}, \lambda) &= \nabla f(\mathbf{u}) + \sum_{i=1}^p \lambda_i \nabla \varphi_i(\mathbf{u}) = \mathbf{0} \\ \Leftrightarrow \begin{cases} A\mathbf{u} - \mathbf{b} + M^\top \lambda &= \mathbf{0} \\ M\mathbf{u} &= \mathbf{c} \end{cases} \\ \Leftrightarrow \left( \begin{array}{c|c} A & M^\top \\ \hline M & 0 \end{array} \right) \begin{pmatrix} \mathbf{u} \\ \lambda \end{pmatrix} &= \begin{pmatrix} \mathbf{b} \\ \mathbf{c} \end{pmatrix} \quad (S) \end{aligned}$$

**Théorème III.4 (Programmation quadratique sous contraintes égalitaires.)**

Soit  $f(\mathbf{u}) = \frac{1}{2}\mathbf{u}^\top A\mathbf{u} - \mathbf{b}^\top \mathbf{u}$  sous la contrainte  $M\mathbf{u} = \mathbf{c}$ . Supposons  $p < n$  et le domaine non vide.

- Si  $A$  est semi-définie positive (resp. négative), alors un extremum, s'il existe, est global et caractérisé par le système (S).
- Si  $A$  est définie positive (resp. négative) alors  $\exists!$  minimum (resp. maximum) global et il est caractérisé par le système (S).

**Démonstration.** Quitte à changer  $f$  en  $-f$  on se ramène au cas où  $A$  est (semi-)définie positive. La première assertion est une conséquence immédiate des théorèmes II.9 et III.2. Quand à la deuxième assertion, le domaine étant non vide c'est un sous-espace affine de  $\mathbb{R}^n$  de dimension  $> 0$  et par suite un fermé non borné ; le théorème II.9 montre que  $f$  est fortement convexe et donc admet un unique minimum global.  $\square$

**Exemple.** Considérons dans  $\mathbb{R}^3$  la droite  $\Delta$  d'équation :

$$\begin{cases} \varphi_1(x, y, z) = 10x + 15y + 20z - 60 = 0 \\ \varphi_2(x, y, z) = -6x + 5y + 10z - 20 = 0 \end{cases}$$

Quelle est la distance de  $\Delta$  à l'origine ?

La distance de  $\Delta$  à l'origine est par définition la distance minimale de l'origine à un point de  $\Delta$ . Le problème se ramène donc au problème d'optimisation :

$$\min_{(x,y,z) \in \Delta} f(x, y, z) = x^2 + y^2 + z^2$$

La fonction  $f$  est quadratique,  $f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top A \mathbf{x}$  avec :

$$\mathbf{x} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} ; \quad A = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}$$

Les contraintes sont affines, la matrice et le vecteur des contraintes sont :

$$M = \begin{pmatrix} 10 & 15 & 20 \\ -6 & 5 & 10 \end{pmatrix} ; \quad \mathbf{c} = \begin{pmatrix} 60 \\ 20 \end{pmatrix}$$

Puisque  $A$  est définie positive, il existe un unique minimum, global, de  $f$ , caractérisé par le système :

$$\left( \begin{array}{c|c} A & M^\top \\ \hline M & 0 \end{array} \right) \begin{pmatrix} \mathbf{u} \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ \mathbf{c} \end{pmatrix}$$

$$\left( \begin{array}{ccc|cc} 2 & 0 & 0 & 10 & -6 \\ 0 & 2 & 0 & 15 & 5 \\ 0 & 0 & 2 & 20 & 10 \\ \hline 10 & 15 & 20 & 0 & 0 \\ -6 & 5 & 10 & 0 & 0 \end{array} \right) \begin{pmatrix} x \\ y \\ z \\ \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 60 \\ 20 \end{pmatrix}$$

Le système a pour solution exacte :

$$\mathbf{u}_{\min} = \left( \frac{88}{141}, \frac{884}{705}, \frac{1232}{705} \right), \quad \lambda_1 = -\frac{536}{3525}, \quad \lambda_2 = -\frac{32}{705}$$

donc,

$$f_{\min} = \left( \frac{88}{141} \right)^2 + \left( \frac{884}{705} \right)^2 + \left( \frac{1232}{705} \right)^2 = \frac{3536}{705} \approx 5,1055$$

et la distance  $\sqrt{f_{\min}}$  de  $\Delta$  à l'origine est proche de  $\sqrt{5}$ .

## III.2 Optimisation sous contraintes : le cas général

On généralise ici les conditions vues précédemment dans le cas où toutes les contraintes étaient égalitaires au cas général où les contraintes sont égalitaires ou inégalitaires. Les conditions de Lagrange se généralisent aux conditions de Karush-Kuhn-Tucker.

### III.2.1 Conditions de Karush-Kuhn-Tucker

Le théorème suivant généralise les conditions de Lagrange (théorème III.1) au cas de contraintes égalitaires et inégalitaires.

Soit  $\mathcal{U}$  un ouvert non vide de  $\mathbb{R}^n$  (souvent,  $\mathcal{U} = \mathbb{R}^n$ ) ; soit  $f : \mathcal{U} \rightarrow \mathbb{R}$  une application différentiable. Soient  $\varphi_1, \dots, \varphi_p, \psi_1, \dots, \psi_q : \mathcal{U} \rightarrow \mathbb{R}$  des applications de classe  $C^1$  ( $p, q \geq 0$ ). Et soit :

$$\mathcal{D} = \{\mathbf{x} \in \mathcal{U} \mid \underbrace{\varphi_i(\mathbf{x}) = 0, \forall i = 1, \dots, p}_{\text{contraintes égalitaires}} ; \underbrace{\psi_j(\mathbf{x}) \leq 0, \forall j = 1, \dots, q}_{\text{contraintes inégalitaires}}\}$$

Pour énoncer le théorème nous avons encore besoin d'une hypothèse de qualification des contraintes en  $\mathbf{u} \in \mathcal{D}$  :

$(QC)_{\mathbf{u}} : \{\nabla \varphi_1(\mathbf{u}), \dots, \nabla \varphi_p(\mathbf{u})\} \cup \{\nabla \psi_j(\mathbf{u}) \mid \psi_j(\mathbf{u}) = 0\}$  est une famille de vecteurs linéairement indépendante.

**Théorème III.5 (Conditions nécessaires de Karush-Kuhn-Tucker.)** *Sous les hypothèses énoncées ci-dessus, si  $\mathbf{u} \in \mathcal{D}$  est un minimum local de  $f$  sur  $\mathcal{D}$ , et si l'hypothèse  $(QC)_{\mathbf{u}}$  est vérifiée, alors  $\exists ! \lambda_1, \dots, \lambda_p, \mu_1, \dots, \mu_q$  tels que :*

- (i)  $\nabla f(\mathbf{u}) + \sum_{i=1}^p \lambda_i \nabla \varphi_i(\mathbf{u}) + \sum_{j=1}^q \mu_j \nabla \psi_j(\mathbf{u}) = \mathbf{0}$
- (ii)  $\forall j = 1, \dots, q, \mu_j \geq 0$
- (iii)  $\forall j = 1, \dots, q, \mu_j = 0$  si  $\psi_j(\mathbf{u}) < 0$ .

**Démonstration.** On ne montrera la condition (ii) que sous l'hypothèse plus forte que toutes les applications sont deux fois différentiables et l'on admettra qu'elle reste vraie sans cette hypothèse. C'est un moindre coût face au gain de simplicité apportée par notre preuve.

On se ramène à un problème d'optimisation à contrainte égalitaire en ajoutant  $q$  variables  $\mathbf{y} = (y_1, \dots, y_q)$  :

$$\begin{cases} \min_{\mathbf{x}, \mathbf{y}} f(\mathbf{x}) \\ \varphi_i(\mathbf{x}) = 0 & \forall i = 1, \dots, p \\ \psi_j(\mathbf{x}) + y_j^2 = 0 & \forall j = 1, \dots, q \end{cases}$$

son domaine  $\mathcal{D}'$  est inclus dans  $\mathcal{D} \times \mathbb{R}^q \subset \mathbb{R}^{n+q}$ , et il a pour minimum local  $(\mathbf{u}, \mathbf{y})$  si et seulement si  $\mathbf{u}$  est un minimum local de  $f$  sur  $\mathcal{D}$ . On va lui appliquer les conditions de Lagrange (théorème III.1). En  $\mathbf{u} \in \mathcal{D}$ , l'hypothèse  $(QC)_{\mathbf{u}}$  nous assure que l'hypothèse de qualification des contraintes en  $(\mathbf{u}, \mathbf{y})$  nécessaire à son application est satisfaite. Notons  $\mathcal{L}(\mathbf{u}, \mathbf{y}, \lambda, \mu)$  le lagrangien de ce problème et  $\mathcal{L}(\mathbf{u}, \lambda, \mu)$  le lagrangien du problème obtenu en prenant  $y_1 = \dots = y_q = 0$ . On obtient  $\exists \lambda_1, \dots, \lambda_p, \mu_1, \dots, \mu_q$ , tels que :

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{u}, \mathbf{y}, \lambda, \mu) = \begin{pmatrix} \nabla f(\mathbf{u}) \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \sum_{i=1}^p \lambda_i \begin{pmatrix} \nabla \varphi_i(\mathbf{u}) \\ 0 \\ \vdots \\ 0 \end{pmatrix} + \sum_{j=1}^q \mu_j \begin{pmatrix} \nabla \psi_j(\mathbf{u}) \\ 0 \\ 2y_j \\ 0 \end{pmatrix} = \mathbf{0}$$

En particulier on obtient la condition (i), et de plus  $\forall j = 1, \dots, q, \mu_j y_j = 0$ , c'est à dire  $\mu_j = 0$  dès que  $\psi_j(\mathbf{u}) < 0$ , c'est la condition (iii).

Pour montrer la condition (ii), comme dit plus haut, on suppose en outre que les applications sont deux fois différentiables. On applique la condition nécessaire à l'ordre 2 (théorème III.3). On a la matrice :

$$\nabla_x^2 \mathcal{L}(\mathbf{u}, \mathbf{y}, \lambda, \mu) = \begin{pmatrix} \boxed{\nabla_x^2 \mathcal{L}(\mathbf{u}, \lambda, \mu)} & 0 & \cdots & 0 \\ & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 \end{pmatrix} \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 0 \\ 2\mu_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 2\mu_q \end{pmatrix}$$

Or puisque  $\mu_j y_j = 0, y_j = 0$  dès que  $\mu_j \neq 0$ . En particulier le vecteur  $\mathbf{e}_j$  de  $\mathbb{R}^{n+q}$  dont la  $(n+j)$ -ème coordonnée est 1 et toutes les autres sont nulles est dans l'espace tangent à  $\mathcal{D}'$  en  $(\mathbf{u}, \mathbf{y})$ . Puisque  $\mathbf{e}_j^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \mathbf{y}, \lambda, \mu) \mathbf{e}_j = 2\mu_j$ , la condition nécessaire du second ordre implique que  $\mu_j \geq 0$ . On obtient donc la condition (ii).

L'unicité des multiplicateurs est une conséquence immédiate de l'hypothèse de qualification des contraintes.  $\square$

**Remarques.** – Les  $\lambda_i, \mu_j$  sont appelés les *multiplicateurs de Lagrange-KKT* ou *multiplicateurs de Lagrange généralisés*.

– Pour un problème de maximum il suffit de changer la condition (ii) en (ii') :  $\forall j = 1, \dots, q, \mu_j \leq 0$ .

– Une contrainte inégalitaire  $\psi_j$  est dite *insaturée* ou *inactive* en  $\mathbf{u}$  si  $\psi_j(\mathbf{u}) < 0$ , et sinon elle est dite *saturée* ou *active*.

– Dans le cas d'une contrainte insaturée  $\psi_j(\mathbf{u}) \neq 0$ , le coefficient de Lagrange-KKT correspondant est nul :  $\mu_j = 0$ , c'est-à-dire que cette contrainte ne compte pas. Lorsque toutes les contraintes inégalitaires sont insaturées en  $\mathbf{u}$  on retrouve les conditions de Lagrange. C'était prévisible, l'ensemble des points de  $\mathbb{R}^n$  où toutes les contraintes inégalitaires sont insaturées est un ouvert  $\mathcal{U}$  et l'on est dans le cadre d'application du théorème de Lagrange (théorème III.1), son domaine n'étant plus défini sur  $\mathcal{U}$  que par les contraintes égalitaires.

– En l'absence de l'hypothèse de qualification des contraintes, on a tout de même l'existence de  $\lambda_1, \dots, \lambda_p$  et  $\mu_0, \mu_1, \dots, \mu_q$ , vérifiant les conditions (ii) et (iii), non nécessairement uniques, tels que  $\mu_0 \geq 0$  et :

$$\mu_0 \nabla f(\mathbf{u}) + \sum_{i=1}^p \lambda_i \nabla \varphi_i(\mathbf{u}) + \sum_{j=1}^q \mu_j \nabla \psi_j(\mathbf{u}) = \mathbf{0}$$

Seulement lorsque  $\mu_0 = 0$  cette équation n'est pas informative sur  $f$ . Les hypothèses de qualification des contraintes sont là pour pallier à cette éventualité.

**Notations Lagrangiennes.** Comme pour un problème d'optimisation sous contraintes égalitaires, on parle du lagrangien d'un problème d'optimisation sous contraintes égalitaires et inégalitaires. C'est l'application :

$$\begin{aligned} \mathcal{L} : \mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}_+)^q &\longmapsto \mathbb{R} \\ (\mathbf{x}, \lambda, \mu) &\longrightarrow \mathcal{L}(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \sum_{i=1}^p \lambda_i \varphi_i(\mathbf{x}) + \sum_{j=1}^q \mu_j \psi_j(\mathbf{x}) . \end{aligned}$$

Lorsque toutes les applications sont différentiables, le vecteur gradient du lagrangien en  $\mathbf{u} \in \mathbb{R}^n$  est :

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{u}, \lambda, \mu) = \nabla f(\mathbf{u}) + \sum_{i=1}^p \lambda_i \nabla \varphi_i(\mathbf{u}) + \sum_{j=1}^q \mu_j \nabla \psi_j(\mathbf{u}) .$$

La condition (i) de KKT s'écrit alors, sous les hypothèses adéquates :

$$\exists ! \lambda \in \mathbb{R}^p, \mu \in (\mathbb{R}_+)^q, \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{u}, \lambda, \mu) = \mathbf{0} .$$

On définit aussi, lorsque les applications sont deux fois différentiables, la matrice Hessienne du lagrangien en  $\mathbf{u} \in \mathbb{R}^n$ , par :

$$\nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda, \mu) = \nabla^2 f(\mathbf{u}) + \sum_{i=1}^p \lambda_i \nabla^2 \varphi_i(\mathbf{u}) + \sum_{j=1}^q \mu_j \nabla^2 \psi_j(\mathbf{u}) .$$

### III.2.2 Prise en compte de la convexité

Dans le cas de la programmation convexe, c'est-à-dire si de plus :

- $\mathcal{U}$  est un ouvert convexe de  $\mathbb{R}^n$  (notamment lorsque  $\mathcal{U} = \mathbb{R}^n$ ),
- $f$  est convexe,
- $\varphi_i, i = 1, \dots, p$ , est linéaire,
- $\psi_j, j = 1, \dots, q$  est convexe,

d'une part, les conditions de KKT sont aussi suffisantes, comme montré ci-dessous, d'autre part avec le théorème II.6, le minimum  $\mathbf{u}$  n'est pas seulement local mais aussi global. Ainsi, en programmation convexe (différentiable), nous arrivons à une caractérisation quasi-complète d'un minimum, exprimée par des conditions du 1er ordre, que ce soit dans le cas sans contraintes ou sous contraintes, avec pour seul bémol, lors de la présence de contraintes inégalitaires, le fait que les conditions de KKT ne sont nécessaires qu'en un minimum  $\mathbf{u}$  vérifiant une hypothèse de qualification des contraintes.

#### **Théorème III.6 (Suffisance des conditions KKT en programmation convexe.)**

*En programmation convexe, les conditions (i), (ii) et (iii) de KKT en  $\mathbf{u} \in \mathcal{D}$  sont aussi suffisantes pour que  $\mathbf{u}$  soit un minimum global de  $f$ .*

**Démonstration.** Soit  $\mathbf{u} \in \mathcal{D}$  en lequel les conditions (i), (ii) et (iii) de KKT sont satisfaites pour certains  $\lambda_1, \dots, \lambda_p, \mu_1, \dots, \mu_q$ . Soit  $\mathbf{v}$  un point quelconque de  $\mathcal{D}$ . Puisque  $\mathbf{v} \in \mathcal{D}$  et  $\mu_j \geq 0$ ,  $f(\mathbf{u}) \leq f(\mathbf{u}) - \sum_{i=1}^p \lambda_i \varphi_i(\mathbf{v}) - \sum_{j=1}^q \mu_j \psi_j(\mathbf{v})$ . Puisque  $\mathbf{u}$  vérifie les conditions (ii) et (iii) de KKT,  $f(\mathbf{u}) \leq f(\mathbf{u}) - \sum_{i=1}^p \lambda_i (\varphi_i(\mathbf{v}) - \varphi_i(\mathbf{u})) - \sum_{j=1}^q \mu_j (\psi_j(\mathbf{v}) - \psi_j(\mathbf{u}))$ . Puisque les  $\varphi_i, \psi_j$  sont convexes, en appliquant le théorème II.5.1 on obtient  $f(\mathbf{u}) \leq f(\mathbf{v}) - \sum_{i=1}^p \lambda_i \nabla \varphi_i(\mathbf{u})(\mathbf{v} - \mathbf{u}) - \sum_{j=1}^q \mu_j \nabla \psi_j(\mathbf{u})(\mathbf{v} - \mathbf{u})$ . Alors avec la condition (i) de KKT en  $\mathbf{u}$ ,  $f(\mathbf{u}) \leq f(\mathbf{v}) + \nabla f(\mathbf{u})(\mathbf{v} - \mathbf{u})$ . Puisque  $f$  est convexe, on utilise le théorème II.5.1 avec cette dernière inégalité pour en déduire que  $f(\mathbf{u}) \leq f(\mathbf{v})$ ; donc  $\mathbf{u}$  est un minimum global de  $f$  sur  $\mathcal{D}$ .  $\square$



**Remarque.** Nul besoin d'hypothèse de qualification des contraintes pour la suffisance des conditions KKT ; elles sont cependant nécessaires pour la nécessité des conditions. Elles se simplifient cependant considérablement, comme nous le voyons ci-après.

### III.2.3 Qualification de contraintes affines et convexes

Les nouvelles conditions de qualification des contraintes que nous allons énoncer tirent parti de la convexité, ou de l'affinité, des applications contraintes ; pour s'appliquer il n'est nullement besoin que la fonction  $f$  vérifie une quelconque hypothèse de convexité.

Comme dans la section précédente, on peut simplifier l'énoncé du théorème III.5 en se passant de l'hypothèse de qualification des contraintes lorsque les contraintes égalitaires ainsi que les contraintes inégalitaires actives sont affines. On perd ce faisant l'unicité des multiplicateurs de Lagrange-KKT.

**Proposition III.2 (Qualification de contraintes affines.)** *Si en  $\mathbf{u} \in \mathcal{D}$ , toutes les contraintes égalitaires et inégalitaires actives sont affines alors on peut se passer de l'hypothèse  $(QC)_{\mathbf{u}}$  dans le théorème III.5, à ceci près que les multiplicateurs de Lagrange-KKT ne sont plus nécessairement uniques.*

**Démonstration.** La preuve procède de la même façon que pour la proposition III.1. □

Nous noterons cette nouvelle hypothèse de qualification des contraintes :

$(\overline{QC}_{\mathbf{u}})$  : Toutes les contraintes égalitaires ainsi que toutes les contraintes inégalitaires actives en  $\mathbf{u}$  sont affines.

En fait lorsque toutes les contraintes sont convexes, il suffit même que cette condition s'applique en un point  $\omega \in \mathcal{D}$  arbitraire, et pas nécessairement en  $\mathbf{u}$ . C'est le résultat que nous énonçons ci-dessous.

Lorsque les contraintes égalitaires sont affines et que les contraintes inégalitaires sont convexes, on peut affaiblir l'hypothèse de qualification des contraintes dans le théorème de Karush-Kuhn-Tucker, en une condition qui ne dépend plus du point considéré.

$(\overline{\overline{QC}})$  : Les contraintes égalitaires sont affines, les contraintes inégalitaires sont convexes, et  $\exists \omega \in \mathcal{U}$ , tel que pour  $i = 1, \dots, q$  soit  $\psi_i$  est affine soit  $\psi_i(\omega) < 0$ .

**Proposition III.3 (Qualification de contraintes convexes.)** *Le théorème III.5 reste vrai, à l'exception de l'unicité des multiplicateurs de Lagrange-KKT, sous l'hypothèse de qualification des contraintes  $(\overline{\overline{QC}})$ .*

On en admettra la preuve, qui aurait nécessité de prouver le théorème III.5 sous une hypothèse de qualification des contraintes plus faibles, l'hypothèse Mangasarian-Fromovitz :

$$\begin{aligned}
(\overline{QC}_{\mathbf{u}}) : \quad & \{\nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u})\} \text{ forment une famille libre, et} \\
& \exists \mathbf{d}_0 \in \mathbb{R}^n, \quad \text{tq } \langle \nabla\varphi_i(\mathbf{u}), \mathbf{d}_0 \rangle = 0, \quad \forall i = 1, \dots, p, \\
& \text{et } \langle \nabla\psi_j(\mathbf{u}), \mathbf{d}_0 \rangle < 0, \quad \forall j \text{ tel que } \psi_j(\mathbf{u}) = 0.
\end{aligned}$$

Le théorème III.5 reste vrai sous cette hypothèse (plus faible) de qualification des contraintes, hormis cependant l'unicité des multiplicateurs de Lagrange-KKT ; c'est d'ailleurs sous cette hypothèse qu'il est en général énoncé. Il n'est pas difficile de vérifier que  $(QC_{\mathbf{u}}) \implies (\overline{QC}_{\mathbf{u}})$  et que la réciproque en est fausse.

### III.2.4 Programmation quadratique sous contraintes

Nous pouvons d'ores et déjà appliquer tout ce que nous avons vu à la programmation quadratique. Le théorème qui suit est une conséquence immédiate des résultats précédents.

**Théorème III.7 (Programmation quadratique sous contraintes.)** *Soit  $f$  une application quadratique,  $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{b}^\top \mathbf{x}$  où  $A$  est une matrice symétrique, et soit  $\mathcal{D}$  un domaine défini par des contraintes égalitaires et inégalitaires affines.*

*Alors, si  $A$  est semi-définie positive (resp. négative) un minimum (resp. maximum) global de  $f$  sur  $\mathcal{D}$ , s'il existe, est caractérisé par les conditions (i), (ii) (resp. (ii')), (iii) de KKT.*

*Si de plus  $A$  est définie positive (resp. négative),  $f$  admet un unique minimum (resp. maximum) global sur  $\mathcal{D}$ .*

Nous en laissons la preuve en guise d'exercice d'application.

**Exemple.** Résoudre le problème de programmation quadratique suivant.

$$\min x^2 + y^2 + z^2 \quad \text{sous les contraintes} \quad \begin{cases} x + y + z = 3 \\ 2x - y + z \leq 5 \end{cases}$$

C'est équivalent au problème consistant à déterminer dans  $\mathbb{R}^3$  la distance d'un demi-plan (défini par les contraintes) à l'origine.

Les contraintes étant affines, on peut appliquer telles quelles les conditions (KKT) : en un minimum  $\mathbf{u}$ ,  $\exists \lambda \in \mathbb{R}, \mu \in \mathbb{R}_+$  tels que  $\nabla f(\mathbf{u}) + \lambda \nabla \varphi(\mathbf{u}) + \mu \nabla \psi(\mathbf{u}) = 0$  :

$$\begin{aligned}
(i) \quad & \begin{cases} 2x + \lambda + 2\mu = 0 \\ 2y + \lambda - \mu = 0 \\ 2z + \lambda + \mu = 0 \end{cases} \\
(ii) \quad & \mu(2x - y + z - 5) = 0 \\
(iii) \quad & \mu \geq 0
\end{aligned}$$

Supposons que  $\mu \neq 0 \implies 2x - y + z = 5$

$$(i) \implies \begin{cases} x = (-\lambda - 2\mu)/2 \\ y = (-\lambda + \mu)/2 \\ z = (-\lambda - \mu)/2 \end{cases}$$

$$x + y + z = 3 \implies -\lambda - 2\mu - \lambda + \mu - \lambda - \mu = 6 \implies 3\lambda + 2\mu = -6 \quad (\text{a})$$

$$2x - y + z = 5 \implies 2(-\lambda - 2\mu) - (-\lambda + \mu) + (-\lambda - \mu) = 10 \implies \lambda + 3\mu = -5 \quad (\text{b})$$

En formant (a)−3(b) on obtient  $2\mu - 9\mu = -6 + 15$ , soit  $\mu = -9/7 < 0$  ce qui contredit (iii)!

Ainsi  $\mu = 0$ . Donc (i) devient :

$$\begin{cases} 2x + \lambda = 0 \\ 2y + \lambda = 0 \\ 2z + \lambda = 0 \end{cases} \implies 2(x + y + z) + 3\lambda = 0 \implies 2 \times 3 + 3\lambda = 0 \implies \lambda = -2$$

$\implies x = y = z = 1$ . On obtient pour solution  $\mathbf{u} = (1, 1, 1)$ .

Pour conclure que  $\mathbf{u}$  est le minimum global de  $f$  sur  $\mathcal{D}$ , on a plusieurs possibilités :

- $f$  est quadratique sous contraintes affines, de matrice  $A = \text{Id}$  définie positive.
- $f$  est (fortement) convexe sous contraintes affines.
- $f$  est coercive.

### III.2.5 Conditions nécessaire, suffisante, du second ordre

Nous établissons ici une condition suffisante et une condition nécessaire du second ordre pour qu'un point soit extremum. La difficulté est qu'en présence de contraintes inégalitaires il n'existe plus en un point  $\mathbf{u} \in \mathcal{D}$  d'espace tangent  $T_{\mathbf{u}}\mathcal{D}$  à  $\mathcal{D}$ , hormis lorsque  $\mathbf{u}$  est dans l'intérieur de  $\mathcal{D}$ , c'est à dire, plus géométriquement, que  $\mathbf{u}$  ne ressemble plus localement à un espace affine. Pour pallier à cette absence nous devons introduire la notion de *cône tangent*. Elle nous permettra par ailleurs d'interpréter géométriquement les conditions de KKT.

#### Définitions.

- En  $\mathbf{u} \in \mathcal{D}$ , l'ensemble des indices de contraintes actives,  $J(\mathbf{u}) \subset \{1, \dots, q\}$ , est l'ensemble des indices  $j$  pour lesquels la  $j$ -ème contrainte inégalitaire est active en  $\mathbf{u}$  :

$$J(\mathbf{u}) = \left\{ j \in \{1, \dots, q\} \mid \psi_j(\mathbf{u}) = 0 \right\} .$$

- Le cône tangent à  $\mathcal{D}$  en  $\mathbf{u}$ , noté  $\mathcal{C}_{\mathbf{u}}\mathcal{D}$  est le sous-ensemble de  $\mathbb{R}^n$  :

$$\mathcal{C}_{\mathbf{u}}\mathcal{D} = \left\{ \mathbf{d} \in \mathbb{R}^n \mid \langle \nabla \varphi_i(\mathbf{u}), \mathbf{d} \rangle = 0, i = 1, \dots, q, \langle \nabla \psi_j(\mathbf{u}), \mathbf{d} \rangle \leq 0, j \in J(\mathbf{u}) \right\}$$

Le cône tangent généralise en cas de contraintes inégalitaires la notion d'espace tangent, au sens du résultat qui suit. Ce n'est plus un espace vectoriel, mais une intersection de sous-espaces vectoriels et de demi-espaces.

**Proposition III.4 (Le cône tangent en tant qu'espace tangent.)** *Si en  $\mathbf{u} \in \mathcal{D}$  une des hypothèses de qualification des contraintes  $(QC)_{\mathbf{u}}$ ,  $(\overline{QC})_{\mathbf{u}}$ ,  $(\overline{QC})$  ou  $(\overline{QC})_{\mathbf{u}}$  est vérifiée, alors le cône tangent  $\mathcal{C}_{\mathbf{u}}\mathcal{D}$  est l'ensemble des directions  $\mathbf{d} \in \mathbb{R}^n$  pour lesquelles soit  $\mathbf{d} = \mathbf{0}$  soit il existe une suite  $(\mathbf{u}_k)_{k \in \mathbb{N}}$  dans  $\mathcal{D}$ , non stationnaire, tendant vers  $\mathbf{u}$ , avec :*

$$\mathbf{u}_k = \mathbf{u} + \frac{\|\mathbf{u}_k - \mathbf{u}\|}{\|\mathbf{d}\|} \mathbf{d} + o(\|\mathbf{u}_k - \mathbf{u}\|)$$

**Démonstration.** Notons  $\mathcal{D}_0 = \{\mathbf{x} \in \mathbb{R}^n \mid \varphi_i(\mathbf{x}) = 0, \forall i = 1, \dots, p\}$ . Sous l'une quelconque des hypothèses de qualification des contraintes, l'espace tangent en  $\mathbf{u}$  à  $\mathcal{D}_0$  est  $T_{\mathbf{u}}\mathcal{D}_0 = \{\mathbf{d} \in \mathbb{R}^n \mid \langle \nabla \varphi_i(\mathbf{u}), \mathbf{d} \rangle = 0, \forall i = 1, \dots, p\}$ .

Soit  $(\mathbf{u}_k)_{k \in \mathbb{N}}$  une suite de  $\mathcal{D}$  non stationnaire qui tend vers  $\mathbf{u}$ . C'est aussi en particulier une suite de  $\mathcal{D}_0$  tendant vers  $\mathbf{u}$ , et par définition de l'espace tangent  $T_{\mathbf{u}}\mathcal{D}_0$ , il existe  $\mathbf{d} \in T_{\mathbf{u}}\mathcal{D}_0$ , tel que :  $\mathbf{u}_k = \mathbf{u} + \frac{\|\mathbf{u}_k - \mathbf{u}\|}{\|\mathbf{d}\|} \mathbf{d} + o(\|\mathbf{u}_k - \mathbf{u}\|)$ . Montrons que  $\mathbf{d} \in \mathcal{C}_{\mathbf{u}}\mathcal{D}$ . Soit  $j \in J(\mathbf{u})$ , i.e.  $\psi_j(\mathbf{u}) = 0$ . En utilisant un développement de Taylor-Young de  $\psi_j$  au voisinage de  $\mathbf{u}$ , on obtient :

$$\psi_j(\mathbf{u}_k) - \psi_j(\mathbf{u}) = \psi_j(\mathbf{u}_k) = \frac{\|\mathbf{u}_k - \mathbf{u}\|}{\|\mathbf{d}\|} \langle \nabla \psi_j(\mathbf{u}), \mathbf{d} \rangle + o(\|\mathbf{u}_k - \mathbf{u}\|) \leq 0$$

puisque  $\mathbf{u}_k \in \mathcal{D}$ . Il en découle que  $\langle \nabla \psi_j(\mathbf{u}), \mathbf{d} \rangle \leq 0$  ce qui montre que  $\mathbf{d} \in \mathcal{C}_{\mathbf{u}}\mathcal{D}$ .

Montrons la réciproque. Soit  $\mathbf{d} \neq \mathbf{0} \in \mathcal{C}_{\mathbf{u}}\mathcal{D} \subset T_{\mathbf{u}}\mathcal{D}_0$ . Par définition de l'espace tangent  $T_{\mathbf{u}}\mathcal{D}_0$ , il existe une suite  $(\mathbf{u}_k)_{k \in \mathbb{N}}$  de  $\mathcal{D}_0$ , non stationnaire, tendant vers  $\mathbf{u}$ , avec  $\mathbf{u}_k = \frac{\|\mathbf{u}_k - \mathbf{u}\|}{\|\mathbf{d}\|} \mathbf{d} + o(\|\mathbf{u}_k - \mathbf{u}\|)$ . Il nous suffit de montrer qu'à partir d'un certain rang  $k$ ,  $\mathbf{u}_k \in \mathcal{D}$ . Soit  $j \in J(\mathbf{u})$ , c'est à dire tel que  $\psi_j(\mathbf{u}) = 0$ . Alors en considérant un développement de Taylor-Young à l'ordre 1 de  $\psi_j$  au voisinage de  $\mathbf{u}$ , on obtient comme ci-dessus

$$\psi_j(\mathbf{u}_k) = \frac{\|\mathbf{u}_k - \mathbf{u}\|}{\|\mathbf{d}\|} \underbrace{\langle \nabla \psi_j(\mathbf{u}), \mathbf{d} \rangle}_{\leq 0} + o(\|\mathbf{u}_k - \mathbf{u}\|).$$

Cela montre que pour  $k$  suffisamment grand, pour tout  $j \in J(\mathbf{u})$ ,  $\psi_j(\mathbf{u}_k) \leq 0$ . Par ailleurs, pour  $j \notin J(\mathbf{u})$ ,  $\psi_j(\mathbf{u}) < 0$  et par continuité de  $\psi_j$ , à partir d'un certain rang  $\psi_j(\mathbf{u}_k) < 0$ . Tout cela montre que  $\mathbf{u}_k$  est dans  $\mathcal{D}$  pour  $k$  assez grand, ce qui achève la preuve.  $\square$

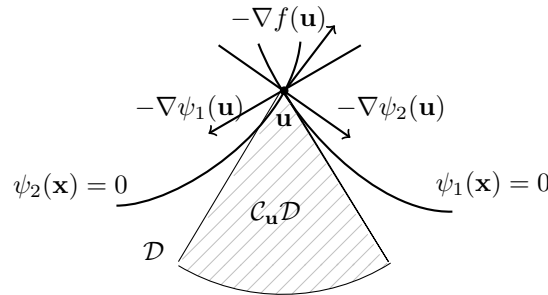


FIGURE III.3 – Le cône tangent  $\mathcal{C}_{\mathbf{u}}\mathcal{D}$  à  $\mathcal{D}$  en  $\mathbf{u}$ . C'est une intersection de demi-espaces (pour chaque contrainte inégalitaire active) et de sous-espaces vectoriels (pour chaque contrainte égalitaire). On a représenté aussi la direction de plus grande pente  $-\nabla f(\mathbf{u})$ . Si  $\mathbf{u}$  est un minimum local elle se trouve dans le cône polaire de  $\mathcal{C}_{\mathbf{u}}\mathcal{D}$  (délimité ici par  $\mathbf{u}$ ,  $\nabla \psi_1(\mathbf{u})$  et  $\nabla \psi_2(\mathbf{u})$ ).

**Interprétation géométrique des conditions KKT.** Les conditions nécessaires de KKT, s'expriment géométriquement par : en un minimum (resp. maximum) local  $\mathbf{u}$  de  $f$  sur  $\mathcal{D}$ , la direction de plus grande décroissance (resp. accroissement) de  $f$ ,  $-\nabla f(\mathbf{u})$  (resp.  $\nabla f(\mathbf{u})$ ) est dans le cône polaire de  $\mathcal{C}_{\mathbf{u}}\mathcal{D}$ , c'est à dire  $\{\mathbf{c} \in \mathbb{R}^n \mid \forall \mathbf{d} \in \mathcal{C}_{\mathbf{u}}\mathcal{D}, \langle \mathbf{d}, \mathbf{c} \rangle \leq 0\}$  (voir figure III.3). En présence de contraintes uniquement égalitaires, le cône polaire n'est rien d'autre que l'orthogonal de l'espace tangent.

**Théorème III.8 (Condition suffisante du 2<sup>e</sup> ordre.)** Soit  $\mathcal{U} \subset \mathbb{R}^n$  un ouvert, on suppose que  $f, \varphi_1, \dots, \varphi_p, \psi_1, \dots, \psi_q$  sont deux fois différentiables sur  $\mathcal{U}$ , et que  $\mathbf{u} \in \mathcal{D}$  vérifie une des hypothèses de qualification des contraintes.

Si  $\mathbf{u} \in \mathcal{D}$  vérifie les conditions (i), (ii), et (iii) de KKT, en particulier si il existe  $\lambda \in \mathbb{R}^p$ ,  $\mu \in (\mathbb{R}_+)^q$  tels que :

$$\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{u}, \lambda, \mu) = 0$$

et si de plus  $\forall \mathbf{d} \neq \mathbf{0} \in \mathcal{C}_{\mathbf{u}}\mathcal{D}$ , :

$$\langle \nabla_{\mathbf{x}}^2\mathcal{L}(\mathbf{u}, \lambda, \mu) \mathbf{d}, \mathbf{d} \rangle > 0 .$$

(i.e.  $\nabla_{\mathbf{x}}^2\mathcal{L}(\mathbf{u}, \lambda, \mu)$  est définie positive sur  $\mathcal{C}_{\mathbf{u}}\mathcal{D}$ ), alors  $\mathbf{u}$  est un minimum local strict de  $f$  sur  $\mathcal{D}$ .

**Démonstration** On note au point  $\mathbf{u} \in \mathcal{D}$ ,  $\mathcal{L}(\mathbf{u}, \lambda, \mu)$  le laplacien du problème avec contraintes égalitaires et inégalitaires et  $\mathcal{L}(\mathbf{u}, \lambda)$  le laplacien du problème ne comportant que les contraintes égalitaires. Soit  $\mathbf{x}$  dans un voisinage de  $\mathbf{u}$  dans  $\mathcal{D}$ . Comme dans la preuve du théorème III.3, on établit :

$$f(\mathbf{x}) - f(\mathbf{u}) = \mathcal{L}(\mathbf{x}, \lambda) - \mathcal{L}(\mathbf{u}, \lambda) = \langle \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{u}, \lambda), \mathbf{x} - \mathbf{u} \rangle + \frac{1}{2}(\mathbf{x} - \mathbf{u})^\top \nabla_{\mathbf{x}}^2\mathcal{L}(\mathbf{u}, \lambda)(\mathbf{x} - \mathbf{u}) + o(\|\mathbf{x} - \mathbf{u}\|^2) \quad (E)$$

Par ailleurs, pour  $j = 1, \dots, q$  le développement de Taylor-Young à l'ordre 2 de  $\psi_j$  dans un voisinage de  $\mathbf{u}$  s'écrit :

$$\psi_j(\mathbf{x}) - \psi_j(\mathbf{u}) = \langle \nabla \psi_j(\mathbf{u}), \mathbf{x} - \mathbf{u} \rangle + \frac{1}{2}(\mathbf{x} - \mathbf{u})^\top \nabla^2 \psi_j(\mathbf{u})(\mathbf{x} - \mathbf{u}) + o(\|\mathbf{x} - \mathbf{u}\|^2) \quad (E_j)$$

On forme alors l'équation (E) +  $\sum_{j=1}^q \mu_j(E_j)$ . On obtient :

$$f(\mathbf{x}) - f(\mathbf{u}) + \sum_{j=1}^q \mu_j(\psi_j(\mathbf{x}) - \psi_j(\mathbf{u})) = \langle \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{u}, \lambda, \mu), \mathbf{x} - \mathbf{u} \rangle + \frac{1}{2}(\mathbf{x} - \mathbf{u})^\top \nabla_{\mathbf{x}}^2\mathcal{L}(\mathbf{u}, \lambda, \mu)(\mathbf{x} - \mathbf{u}) + o(\|\mathbf{x} - \mathbf{u}\|^2) \quad (*)$$

Les conditions (ii) et (iii) de KKT impliquent que  $\sum_{j=1}^q \mu_j(\psi_j(\mathbf{x}) - \psi_j(\mathbf{u})) \leq 0$ , et la condition (i) que  $\langle \nabla_{\mathbf{x}}\mathcal{L}(\mathbf{u}, \lambda, \mu), \mathbf{x} - \mathbf{u} \rangle = 0$ , ainsi pour  $\mathbf{d} \in \mathcal{C}_{\mathbf{u}}\mathcal{D}$  :

$$f(\mathbf{x}) - f(\mathbf{u}) \geq \frac{1}{2}(\mathbf{x} - \mathbf{u})^\top \nabla_{\mathbf{x}}^2\mathcal{L}(\mathbf{u}, \lambda, \mu)(\mathbf{x} - \mathbf{u}) + o(\|\mathbf{x} - \mathbf{u}\|^2) = \frac{\|\mathbf{x} - \mathbf{u}\|^2}{2\|\mathbf{d}\|^2} \mathbf{d}^\top \nabla_{\mathbf{x}}^2\mathcal{L}(\mathbf{u}, \lambda, \mu)\mathbf{d} + o(\|\mathbf{x} - \mathbf{u}\|^2)$$

Puisque  $\mathbf{d}^\top \nabla_{\mathbf{x}}^2\mathcal{L}(\mathbf{u}, \lambda, \mu)\mathbf{d} > 0$  on obtient  $f(\mathbf{x}) - f(\mathbf{u}) > 0$  pour  $\mathbf{x}$  suffisamment proche de  $\mathbf{u}$  :  $\mathbf{u}$  est un minimum local de  $f$  sur  $\mathcal{D}$ .  $\square$

Afin d'énoncer une condition nécessaire du second ordre, nous devons nous restreindre à un sous-ensemble du cône tangent. On se place pour cela sous les hypothèses du théorème III.5.

**Définition.** Soit  $\mathbf{u} \in \mathcal{D}$ , un point vérifiant une hypothèse de qualification des contraintes ainsi que les conditions (i), (ii) et (iii) de KKT.

• Notons :

$$J^+(\mathbf{u}) = \left\{ j \in \{1, \dots, q\} \mid \psi_j(\mathbf{u}) = 0 \text{ et } \mu_j > 0 \right\}$$

C'est l'ensemble des indices des contraintes *fortement actives* en  $\mathbf{u}$ . Notons que, plus simplement,  $J^+(\mathbf{u}) = \left\{ j \in \{1, \dots, q\} \mid \mu_j \neq 0 \right\}$ .

• Notons

$$\mathcal{C}_{\mathbf{u}}^+ \mathcal{D} = \left\{ \mathbf{d} \in \mathcal{C}_{\mathbf{u}} \mathcal{D} \mid \langle \nabla \psi_j(\mathbf{u}), \mathbf{d} \rangle = 0, j \in J^+(\mathbf{u}), \right\} .$$

**Théorème III.9 (Condition nécessaire du second ordre)** Soit  $\mathcal{U} \subset \mathbb{R}^n$  un ouvert, on suppose que  $f, \varphi_1, \dots, \varphi_p, \psi_1, \dots, \psi_q$  sont deux fois différentiables sur  $\mathcal{U}$ , et que  $\mathbf{u} \in \mathcal{D}$  vérifie une des hypothèses de qualification des contraintes.

Si  $\mathbf{u}$  est un minimum local de  $f$  sur  $\mathcal{D}$ , les conditions (i), (ii), (iii) de KKT sont satisfaites, en particulier  $\exists \lambda \in \mathbb{R}^p, \mu \in (\mathbb{R}_+)^q$  tels que :

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{u}, \lambda, \mu) = 0$$

et de plus,  $\forall \mathbf{d} \in \mathcal{C}_{\mathbf{u}}^+ \mathcal{D}$ , :

$$\langle \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda, \mu) \mathbf{d}, \mathbf{d} \rangle \geq 0 .$$

(i.e.  $\nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda, \mu)$  est semi-définie positive sur  $\mathcal{C}_{\mathbf{u}} \mathcal{D}$ ).

**Démonstration.** Il suffit de montrer que sous ces hypothèses,  $\forall \mathbf{d} \in \mathcal{C}_{\mathbf{u}}^+ \mathcal{D}$ ,  $\langle \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda, \mu) \mathbf{d}, \mathbf{d} \rangle \geq 0$ . Comme dans la preuve du théorème III.5, on retranscrit le problème d'optimisation comme le problème d'optimisation (Q) sous contraintes égalitaires :

$$\begin{cases} \min_{\mathbf{x}, \mathbf{y}} f(\mathbf{x}) \\ \varphi_i(\mathbf{x}) = 0 & \forall i = 1, \dots, p \\ \psi_j(\mathbf{x}) + y_j^2 = 0 & \forall j = 1, \dots, q \end{cases}$$

Notons  $\mathcal{D}'$  son domaine, il intersecte  $\mathbb{R}^n \times \mathbf{0}$  en  $\mathcal{D} \times \mathbf{0}$ , et notons  $\mathbf{v} = (\sqrt{-\psi_1(\mathbf{u})}, \dots, \sqrt{-\psi_q(\mathbf{u})}) \in \mathbb{R}^q$  de sorte que  $\mathbf{u}' = (\mathbf{u}, \mathbf{v}) \in \mathcal{D}'$ . Comme conséquence de la proposition III.4, l'espace tangent  $T_{\mathbf{u}'} \mathcal{D}'$  intersecte  $\mathbb{R}^n \times \mathbf{0}$  en  $\mathcal{C}_{\mathbf{u}} \mathcal{D} \times \mathbf{0}$ . Le point  $\mathbf{u}'$  est un minimum local de (Q), et vérifie l'hypothèse de qualification des contraintes, par construction et car  $\mathbf{u}$  vérifie une telle hypothèse. La condition nécessaire du second ordre (théorème III.3) implique alors que  $\forall \mathbf{x} \in \mathbb{R}^n, \mathbf{y} \in \mathbb{R}^q$ , tels que  $(\mathbf{x}, \mathbf{y}) \in T_{\mathbf{u}'} \mathcal{D}'$ , on a  $(\mathbf{x}, \mathbf{y})^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \mathbf{v}, \lambda, \mu)(\mathbf{x}, \mathbf{y}) \geq 0$ . Or,

$$\nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \mathbf{v}, \lambda, \mu) = \begin{pmatrix} \boxed{\nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda, \mu)} & 0 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & \dots & \dots & 0 & 2\mu_1 & \dots & 0 \\ \vdots & & & \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & \dots & 0 & 0 & \dots & 2\mu_q \end{pmatrix}$$

et donc, en notant  $\mathbf{y} = (y_1, \dots, y_q)$ ,

$$(\mathbf{x}, \mathbf{y})^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \mathbf{v}, \lambda, \mu)(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda, \mu) \mathbf{x} + 2 \sum_{i=1}^q \mu_i y_i .$$

Notons  $T_{\mathbf{u}'}^+ \mathcal{D}' = \{\mathbf{d} \in T_{\mathbf{u}'} \mathcal{D}' \mid \forall j \in J^+(\mathbf{u}), \langle \nabla \psi_j(\mathbf{u}), \mathbf{d} \rangle = 0\}$ . Or si  $\mathbf{d} \in T_{\mathbf{u}'} \mathcal{D}'$ , en notant  $\mathbf{v}_j = \sqrt{-\psi_j(\mathbf{u})}$ ,

$$\forall j = 1, \dots, q, \quad (\nabla \psi_j(\mathbf{u})^\top \mid 0 \dots 0 \ 2\mathbf{v}_j \ 0 \dots 0) \mathbf{d} = 0$$

donc pour tout  $\mathbf{d} \in T_{\mathbf{u}'}^+ \mathcal{D}'$ , si  $j \in J^+(\mathbf{u})$ , la  $j^e$  coordonnée de  $\mathbf{d}$  est nulle. D'autre part si  $j \notin J^+(\mathbf{u})$ , on a  $\mu_j = 0$ . Ainsi :  $\forall (\mathbf{x}, \mathbf{y}) \in T_{\mathbf{u}'}^+ \mathcal{D}'$ ,

$$0 = (\mathbf{x}, \mathbf{y})^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \mathbf{v}, \lambda, \mu) (\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda, \mu) \mathbf{x}.$$

Or,  $T_{\mathbf{u}'}^+ \mathcal{D}' \cap (\mathbb{R}^n \times \mathbf{0}) = \mathcal{C}_{\mathbf{u}}^+ \mathcal{D} \times \mathbf{0}$ . On a donc montré que  $\forall \mathbf{x} \in \mathcal{C}_{\mathbf{u}}^+ \mathcal{D}$ ,  $\mathbf{x}^\top \nabla_{\mathbf{x}}^2 \mathcal{L}(\mathbf{u}, \lambda, \mu) \mathbf{x} = 0$ .  $\square$

### III.2.6 Points-selles du Lagrangien : introduction à la dualité

On se rappelle qu'au problème d'optimisation sous contrainte :

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \varphi_i(\mathbf{x}) &= 0, \forall i = 1, \dots, p, \\ \psi_j(\mathbf{x}) &\leq 0, \forall j = 1, \dots, q, \end{aligned} \tag{P}$$

on associe le lagrangien du problème (cf. p.63) :

$$\begin{aligned} \mathcal{L} : \mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}_+)^q &\longmapsto \mathbb{R} \\ (\mathbf{x}, \lambda, \mu) &\longrightarrow \mathcal{L}(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \sum_{i=1}^p \lambda_i \varphi_i(\mathbf{x}) + \sum_{j=1}^q \mu_j \psi_j(\mathbf{x}) \end{aligned}$$

(en notant  $\lambda = (\lambda_1, \dots, \lambda_p)$  et  $\mu = (\mu_1, \dots, \mu_q)$ ).

**Définition.** On appelle point-selle du lagrangien tout triplet  $(x^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}_+)^q$  vérifiant :

$$\forall (\mathbf{x}, \lambda, \mu) \in \mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}_+)^q, \quad \mathcal{L}(\mathbf{x}^*, \lambda, \mu) \leq \mathcal{L}(\mathbf{x}^*, \lambda^*, \mu^*) \leq \mathcal{L}(\mathbf{x}, \lambda^*, \mu^*).$$

c'est-à-dire que :  
 –  $\mathbf{x}^*$  est un minimum de  $\mathbf{x} \mapsto \mathcal{L}(\mathbf{x}, \lambda^*, \mu^*)$ , et  
 –  $(\lambda^*, \mu^*)$  est un maximum de  $(\lambda, \mu) \mapsto \mathcal{L}(\mathbf{x}^*, \lambda, \mu)$ .

**Proposition III.5 (Caractérisation d'un point-selle.)** Si  $(\mathbf{x}^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}_+)^q$  est un point-selle de  $\mathcal{L}(\mathbf{x}, \lambda, \mu)$ , alors :

$$\sup_{\lambda, \mu} \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \mu) = \mathcal{L}(\mathbf{x}^*, \lambda^*, \mu^*) = \inf_{\mathbf{x}} \sup_{\lambda, \mu} \mathcal{L}(\mathbf{x}, \lambda, \mu).$$

**Démonstration.** On a toujours :  $\inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda^*, \mu^*) \leq \mathcal{L}(\mathbf{x}^*, \lambda^*, \mu^*) \leq \sup_{\lambda, \mu} \mathcal{L}(\mathbf{x}^*, \lambda, \mu)$ , ce qui implique :

$$\sup_{\lambda, \mu} \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \mu) \leq \inf_{\mathbf{x}} \sup_{\lambda, \mu} \mathcal{L}(\mathbf{x}, \lambda, \mu).$$

D'autre part, puisque  $(\mathbf{u}^*, \lambda^*, \mu^*)$  est un point-selle, on a :

$$\inf_{\mathbf{x}} \sup_{\lambda, \mu} \mathcal{L}(\mathbf{x}, \lambda, \mu) \leq \sup_{\lambda, \mu} \mathcal{L}(\mathbf{x}^*, \lambda, \mu) = \mathcal{L}(\mathbf{x}^*, \lambda^*, \mu^*) = \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda^*, \mu^*) \leq \sup_{\lambda, \mu} \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \mu).$$

et on a donc la conclusion recherchée.  $\square$

**Théorème III.10 (un point-selle du lagrangien fournit une solution à  $(P)$ .)** *Si  $(\mathbf{x}^*, \lambda^*, \mu^*) \in \mathbb{R}^n \times \mathbb{R}^p \times (\mathbb{R}_+)^q$  est un point-selle du lagrangien du problème  $(P)$ , alors  $\mathbf{x}^*$  est solution du problème  $(P)$ .*

**Démonstration.** L'inégalité  $\mathcal{L}(\mathbf{x}^*, \lambda, \mu) \leq \mathcal{L}(\mathbf{x}^*, \lambda^*, \mu^*)$ ,  $\forall (\lambda, \mu) \in \mathbb{R}^p \times (\mathbb{R}_+)^q$  montre que :

$$\sum_{i=1}^p (\lambda_i - \lambda_i^*) \varphi_i(\mathbf{x}^*) + \sum_{j=1}^q (\mu_j - \mu_j^*) \psi_j(\mathbf{x}^*) \leq 0 .$$

En faisant tendre  $\mu_j$  vers  $+\infty$  cela montre que  $\psi_j(\mathbf{x}^*) \leq 0$ . En faisant tendre  $\lambda_i$  vers  $+\infty$  cela montre que  $\varphi_i(\mathbf{x}^*) \leq 0$ , et en le faisant tendre vers  $-\infty$ , que  $\varphi_i(\mathbf{x}^*) \geq 0$ , et donc  $\varphi_i(\mathbf{x}^*) = 0$ . Ainsi  $\mathbf{x}^*$  est dans le domaine admissible  $\mathcal{D}$ . En particulier,  $\sum_{i=1}^p \lambda_i^* \varphi_i(\mathbf{x}^*) + \sum_{j=1}^q \mu_j^* \psi_j(\mathbf{x}^*) \leq 0$ . Mais avec l'inégalité ci-dessus, en prenant  $\mu_j = 0$  pour  $j = 1, \dots, q$ , on obtient aussi  $\sum_{i=1}^p \lambda_i^* \varphi_i(\mathbf{x}^*) + \sum_{j=1}^q \mu_j^* \psi_j(\mathbf{x}^*) \geq 0$ ; on obtient donc :

$$\sum_{i=1}^p \lambda_i^* \varphi_i(\mathbf{x}^*) + \sum_{j=1}^q \mu_j^* \psi_j(\mathbf{x}^*) = 0 .$$

En combinant cette égalité avec  $\mathcal{L}(\mathbf{x}^*, \lambda^*, \mu^*) \leq \mathcal{L}(\mathbf{x}, \lambda^*, \mu^*)$  pour tout  $\mathbf{x} \in \mathbb{R}^n$ , on obtient donc :  $\forall \mathbf{x} \in \mathbb{R}^n$

$$f(\mathbf{x}^*) \leq f(\mathbf{x}) + \sum_{i=1}^p \lambda_i^* \varphi_i(\mathbf{x}) + \sum_{j=1}^q \mu_j^* \psi_j(\mathbf{x})$$

en particulier pour tout  $\mathbf{x} \in \mathcal{D}$ ,

$$f(\mathbf{x}^*) \leq f(\mathbf{x})$$

ce qui montre que  $\mathbf{x}^*$  est solution de  $(P)$ . □

Remarquer que ce résultat ne nécessite aucune hypothèse que ce soit sur  $f$  ou sur les contraintes. Par contre la réciproque ne peut s'établir que sous des hypothèses relativement fortes.

**Théorème III.11 (Cas où la réciproque est vraie.)** *Supposons que  $f$  est dérivable et convexe,  $\varphi_1, \dots, \varphi_p$  sont  $C^1$  et affines, et  $\psi_1, \dots, \psi_q$  sont  $C^1$  et convexes. Soit  $\mathbf{x}^*$  un point du domaine admissible vérifiant une hypothèse de qualification des contraintes.*

*Si  $\mathbf{x}^*$  est solution du problème  $(P)$ , alors il existe  $(\lambda^*, \mu^*) \in \mathbb{R}^p \times (\mathbb{R}_+)^q$  tel que  $(\mathbf{x}^*, \lambda^*, \mu^*)$  soit un point-selle du lagrangien.*

**Démonstration.** On est dans le cadre d'application des conditions KKT (théorème III.5). Il existe donc  $\lambda^* \in \mathbb{R}^p$ ,  $\mu^* \in (\mathbb{R}_+)^q$ , vérifiant les conditions de Karush-Kuhn-Tucker :

$$\sum_{i=1}^p \lambda_i^* \varphi_i(\mathbf{x}^*) + \sum_{j=1}^q \mu_j^* \psi_j(\mathbf{x}^*) = 0, \quad \text{et} \quad \nabla f(\mathbf{x}^*) + \sum_{i=1}^p \lambda_i^* \nabla \varphi_i(\mathbf{x}^*) + \sum_{j=1}^q \mu_j^* \nabla \psi_j(\mathbf{x}^*) = \mathbf{0} .$$

La première condition de KKT, avec le fait que  $\mathbf{x}^* \in \mathcal{D}$ , montre que pour tout  $(\lambda, \mu) \in \mathbb{R}^p \times (\mathbb{R}_+)^q$  :

$$\mathcal{L}(\mathbf{x}^*, \lambda, \mu) = f(\mathbf{x}^*) + \sum_{i=1}^p \lambda_i^* \varphi_i(\mathbf{x}^*) + \sum_{j=1}^q \mu_j^* \psi_j(\mathbf{x}^*) \leq f(\mathbf{x}^*) = \mathcal{L}(\mathbf{x}^*, \lambda^*, \mu^*) .$$

Pour ce couple  $(\lambda^*, \mu^*)$ , l'application  $\mathbf{x} \mapsto \mathcal{L}(\mathbf{x}, \lambda^*, \mu^*)$  est convexe car somme d'une application affine, et donc convexe,  $\mathbf{x} \mapsto \sum_{i=1}^p \lambda_i^* \varphi_i(\mathbf{x})$  et de  $q$  applications convexes  $\mathbf{x} \mapsto \mu_j^* \psi_j(\mathbf{x})$ . Ainsi la deuxième condition de KKT montre que  $\mathbf{x}^*$  en est un minimum global (théorème III.6), et donc  $\forall \mathbf{x} \in \mathbb{R}^n$ ,  $\mathcal{L}(\mathbf{x}^*, \lambda^*, \mu^*) \leq \mathcal{L}(\mathbf{x}, \lambda^*, \mu^*)$ . Ceci montre que  $(\mathbf{x}^*, \lambda^*, \mu^*)$  est un point-selle du lagrangien. □



**Exemple.** Le problème de minimisation suivant :

$$\min_{x \geq 0} x$$

est un problème de programmation convexe qui a pour solution évidente  $x = 0$ . Son lagrangien est  $\mathcal{L}(x, \mu) = x - \mu x$  qui admet un unique point-selle sur  $\mathbb{R} \times \mathbb{R}_+$  en  $x = 0$ ,  $\mu = 1$ , (voir figure III.4) qui fournit le minimum et le multiplicateur de Lagrange associé.

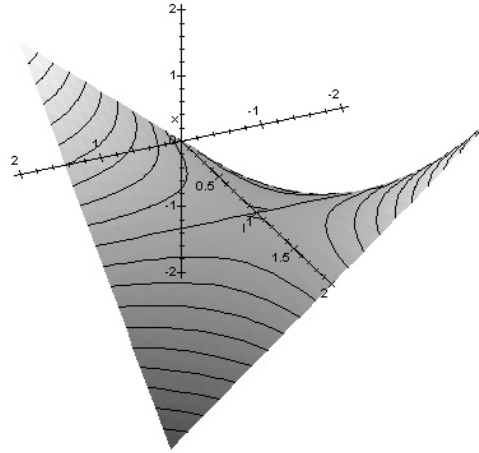


FIGURE III.4 – Le graphe du lagrangien  $\mathcal{L}(x, \mu) = x - \mu x$  admet un unique point selle en  $x = 0$ ,  $\mu = 1$ .

Ainsi, en programmation convexe, sous des hypothèses adéquates (de différentiabilité et de qualification des contraintes), une solution  $\mathbf{u}^*$  du problème  $(P)$  correspond exactement avec le premier argument d'un point-selle  $(\mathbf{u}^*, \lambda^*, \mu^*)$  du lagrangien. La connaissance des arguments  $(\lambda^*, \mu^*)$  d'un point-selle permettrait donc de ramener le problème  $(P)$  à un problème sans contrainte :

$$\inf_{\mathbf{x} \in \mathbb{R}^n} \mathcal{L}(\mathbf{x}, \lambda^*, \mu^*) .$$

Comment trouver un tel couple  $(\lambda^*, \mu^*)$  ?

Avec la proposition III.5, on a

$$\mathcal{L}(\mathbf{x}^*, \lambda^*, \mu^*) = \inf_{\mathbf{x} \in \mathbb{R}^n} \mathcal{L}(\mathbf{x}, \lambda^*, \mu^*) = \sup_{\lambda, \mu} \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda, \mu) .$$

On se ramène donc à chercher  $(\lambda^*, \mu^*) \in \mathbb{R}^p \times (\mathbb{R}_+)^q$  comme solution du problème :

$$F(\lambda^*, \mu^*) = \sup_{\mu \geq 0, \lambda} F(\lambda, \mu) \quad (Q)$$

où :

$$\begin{aligned} F : \mathbb{R}^p \times (\mathbb{R}_+)^q &\longrightarrow \overline{\mathbb{R}} \\ (\lambda, \mu) &\longrightarrow F(\lambda, \mu) = \inf_{\mathbf{x} \in \mathbb{R}^n} \mathcal{L}(\mathbf{x}, \lambda, \mu) . \end{aligned}$$

Le problème  $(Q)$  est appelé le problème dual de  $(P)$ , qui est alors appelé *problème primal*. C'est un problème d'optimisation sous contraintes, mais avec des contraintes particulièrement simples, puisqu'il ne s'agit que de contraintes de signe sur les coefficients de  $\mu$ .

Par construction, sous les hypothèses du théorème III.11, si  $\mathbf{x}^*$  est un minimum du problème  $(P)$  alors le problème  $(Q)$  admet une solution  $(\lambda^*, \mu^*) \in \mathbb{R}^p \times (\mathbb{R}_+)^q$  telle que  $(\mathbf{x}^*, \lambda^*, \mu^*)$  soit un point-selle du lagrangien.

**Exemple.** Si l'on reprend l'exemple ci-dessus,  $F(\mu) = -\infty$  si  $\mu \neq 1$  et  $F(1) = 1$ . Le problème dual a donc pour solution  $\mu = 1$ .

## Exercices.

**Exercice 1.** Retrouver les résultats obtenus aux exemples A et B du § III.1.2 en appliquant les conditions de Lagrange.

**Exercice 2.** Considérons l'application  $f : \mathbb{R}^2 \rightarrow \mathbb{R}$  définie par :

$$f(x, y) = x^3 + y^3 + x^2 + y^2 - 1$$

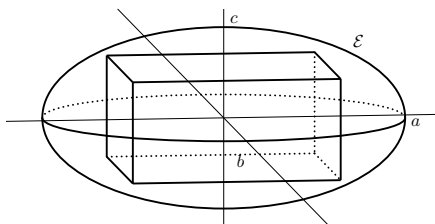
déjà étudiée dans l'exercice 1 du chapitre 2.

Soit  $\mathcal{C} = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}$  le cercle unité de  $\mathbb{R}^2$ . Justifier de l'existence d'extrema globaux de  $f$  sur  $\mathcal{D}$ , et les déterminer.

**Exercice 3.** On évalue que le volume de vente d'un produit est fonction du nombre de publicités dans les magazines  $x$  et du nombre de minutes de temps de télévision  $y$  :  $f(x, y) = 3xy - x^2 - 3y^2$ . Chaque publicité dans les magazines et chaque minute de télévision coûtent 100 u.m.. On dispose de 2800 u.m. de budget de publicité. Comment l'allouer de façon optimale pour maximiser la vente de ce produit ?

**Exercice 4.** (Problème de Kepler.)

Inscrire dans l'ellipsoïde  $\mathcal{E} = \{(x, y, z) \in \mathbb{R}^3 \mid x^2/a^2 + y^2/b^2 + z^2/c^2 = 1\}$  le parallélépipède de volume maximal dont les arêtes sont parallèles aux axes.



**Exercice 5.** (Problème de Tartaglia)

Décomposer le nombre 8 en deux parties positives  $p_1, p_2$  de sorte que le produit de leur produit par leur différence soit maximal.

**Exercice 6.** Soit  $A$  une matrice symétrique réelle. Justifier que le problème :

$$\max_{\|\mathbf{x}\| \leq 1} \mathbf{x}^\top A \mathbf{x}$$

admet une solution  $\mathbf{u}$ . Que représente  $\mathbf{u}$  pour la matrice  $A$  ?



# Chapitre IV

## Algorithmes itératifs

Soit le problème d'optimisation :

$$\min_{\mathbf{x} \in \mathcal{D}} f(\mathbf{x})$$

pour  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  et  $\mathcal{D} \subset \mathbb{R}^n$  défini à l'aide de fonctions contraintes égalitaires et inégalitaires. On cherche à se rapprocher algorithmiquement d'une solution en construisant une suite  $(\mathbf{u}_n)_{n \in \mathbb{N}}$  de  $\mathbb{R}^n$  qui converge vers une solution  $\mathbf{u}$  du problème, c'est à dire un minimum global de  $f$  sur  $\mathcal{D}$ . On établit pour cela plusieurs méthodes, et on s'intéresse à :

- des conditions suffisantes sur  $f$  et  $\mathcal{D}$  pour que  $\mathbf{u}_n \rightarrow \mathbf{u}$ ,
- sa vitesse de convergence.

Nous établirons d'abord des algorithmes dans le cas sans contrainte, *i.e.*  $\mathcal{D} = \mathbb{R}^n$ , puis dans le cas sous contraintes. Nous verrons plusieurs types de méthodes ; on les classe parfois en :

- *Les méthodes directes.* Elles n'utilisent pas les dérivées de l'application. Nous n'en verrons aucune, mais on peut citer comme exemple la *méthode de Hooke-Jeeves* qui consiste à fixer un pas  $\rho > 0$  puis à construire  $\mathbf{u}_{k+1}$  à partir de  $\mathbf{u}_k$ , en choisissant parmi tous les points  $\mathbf{u}_k \pm \rho \mathbf{e}_i$  (où  $\mathbf{e}_i, i = 1, \dots, n$  désignent les vecteurs de la base canonique) celui dont la valeur prise par  $f$  est minimale. Si  $\mathbf{u}_{k+1} = \mathbf{u}_k$  on diminue le pas. Cette méthode est employée pour minimiser une application non différentiable ; sa convergence, lorsqu'elle a lieu, est très lente.
- *Les méthodes de descente.* Elles utilisent les dérivées d'ordre 1. Ici  $\mathbf{u}_{k+1}$  est construit à partir de  $\mathbf{u}_k$  en choisissant une *direction de descente*  $\mathbf{d}_k$  et un *pas de descente*  $\rho_k > 0$ , tels que  $\mathbf{u}_{k+1} = \mathbf{u}_k + \rho_k \mathbf{d}_k$ . Nous en verrons plusieurs : la *méthode de relaxation* qui prend pour direction de descente la direction des axes de façon cyclique et calcule le pas de descente en se ramenant à un problème de minimisation à une variable ; la *méthode du gradient à pas optimal* qui prend comme direction de descente la direction locale de plus grande pente, l'opposé du vecteur gradient, et détermine le pas optimal en se ramenant à un problème à une variable ; ces deux dernières méthodes ont pour désavantage la résolution à chaque pas

d'un problème à une variable ; la *méthode du gradient à pas fixe* s'en démarque en fixant le pas ; enfin nous verrons la *méthode du gradient conjugué*, très ingénieuse, qui pour une fonction quadratique elliptique trouve le minimum en au plus  $n$ -itérations : on parle d'une *méthode exacte* par opposition aux *méthodes approchées* qui ne peuvent qu'approcher une solution.

- *Les méthodes utilisant les dérivées secondes.* Il s'agit essentiellement de la *méthode de Newton* (et de ses variantes) qui plus généralement détermine les zéros d'une application. Sous des hypothèses suffisantes sa convergence est très rapide ; elle a cependant pour désavantage de n'être que locale : il faut choisir le point initial suffisamment proche du minimum.

Dans le cas sous contraintes, les méthodes que nous verrons sont déduites des méthodes ci-dessus en utilisant le théorème de projection convexe : *méthode de relaxation sous contraintes*, *méthode de gradient projeté*. Elles nécessitent cependant d'exprimer l'opérateur de projection convexe, ce qui n'est possible que dans des cas très simples. La *méthode d'Uzawa* quant à elle contourne cette difficulté en mettant à profit la théorie de la dualité convexe (§ III.2.6) pour résoudre le problème dual qui n'invoque quant à lui que des contraintes de signe.

La convergence de toutes ces méthodes ne peut s'établir que sous de fortes hypothèses, tout au moins (hormis pour la méthode de Newton) l'ellipticité de  $f$  et la convexité du domaine  $\mathcal{D}$ . En fait ce sont les méthodes algorithmiques à utiliser en programmation convexe, et uniquement dans ce cadre. En programmation non convexe on utilise d'autres méthodes, stochastiques, programmation dynamique *etc...*, de la recherche opérationnelle qui sortent du cadre de ce cours (étudiées dans le cours de Recherche Opérationnelle EA3).

Pour aborder, concept essentiel, la vitesse de convergence de ces algorithmes, nous avons besoin comme préliminaire de donner quelques définitions (à rapprocher du concept d'ordre d'une application).

**Définitions.** Soit  $(\mathbf{u}_n)_{n \in \mathbb{N}}$  une suite de  $\mathbb{R}^n$  qui converge vers  $\mathbf{u} \in \mathbb{R}^n$ . On note  $\mathbf{r}_n = \mathbf{u} - \mathbf{u}_n$ .

- Si  $\exists k < 1$  et  $n_0 \in \mathbb{N}$  tel que  $\forall n > n_0$ ,  $\|\mathbf{r}_{n+1}\| \leq k \|\mathbf{r}_n\|$  la convergence est dite *géométrique*.
- Si  $\exists p \in \mathbb{N}^*$  et  $n_0 \in \mathbb{N}$  tel que  $\forall n > n_0$ ,  $\|\mathbf{r}_{n+1}\| \leq \|\mathbf{r}_n\|^p$  la convergence est dite *d'ordre  $p$*  ; si  $p = 1$  elle est dite *linéaire* et si  $p = 2$  elle est dite *quadratique*.
- Si  $\exists N_0$  tel que  $\forall n \geq N_0$ ,  $\|\mathbf{r}_n\| = 0$ , la convergence est dite *finie*. C'est ce qui caractérise les *méthodes exactes*.

## IV.1 Méthodes itératives dans le cas sans contraintes

### IV.1.1 Méthode de Newton

Pour chercher un extremum  $\mathbf{u}$  d'une fonction différentiable, on peut se ramener à chercher ses points critiques,  $\nabla f(\mathbf{u}) = 0$ . Résoudre cette équation n'est pas toujours facile, ni même faisable. Il est utile de considérer une méthode de calcul approché. Nous voyons ici la (célèbre) méthode de Newton, qui d'une façon plus générale permet d'approcher les zéros d'une fonction (sous certaines hypothèses).

Méthode de Newton pour une application dérivable  $f : \mathbb{R} \rightarrow \mathbb{R}$ . On cherche  $u \in \mathbb{R}$ , tel que  $f(u) = 0$ . Au voisinage de  $x_0$ ,  $f(x) \approx f(x_0) + f'(x_0)(x - x_0)$ .  
 $\Rightarrow$  Si  $f'(x_0) \neq 0$ , on considère :

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

et on construit une suite par récurrence, par (figure IV.1) :

$$\text{si } f'(x_n) \neq 0, \quad x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

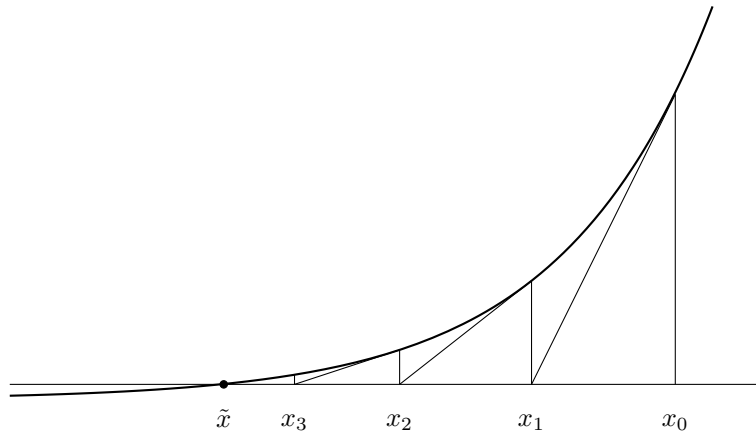


FIGURE IV.1 – La méthode de descente de Newton pour la recherche du zéro d'une application  $f : \mathbb{R} \rightarrow \mathbb{R}$ .

Sous certaines hypothèses, la suite  $(x_n)_{n \in \mathbb{N}}$  converge vers un zéro de  $f$ . Plus précisément et plus généralement :

**Théorème IV.1 (Convergence de la méthode de Newton.)** Soit  $F : \mathbb{R}^n \rightarrow \mathbb{R}^n$  de classe  $C^1$ , et  $\mathbf{u}$  un zéro isolé de  $F$ . Si la matrice jacobienne  $DF(\mathbf{u})$  de  $F$  en  $\mathbf{u}$  est inversible, alors il existe une boule  $B(\mathbf{u})$  centrée en  $\mathbf{u}$ , telle que  $\forall \mathbf{u}_0 \in B(\mathbf{u})$ , la suite :

$$\mathbf{u}_{n+1} = \mathbf{u}_n - DF(\mathbf{u}_n)^{-1}F(\mathbf{u}_n)$$

soit contenue dans  $B(\mathbf{u})$  et converge vers  $\mathbf{u}$  seul zéro de  $F$  dans  $B(\mathbf{u})$ . De plus la convergence est géométrique.

**Démonstration.** Par continuité de  $DF$  d'une part et d'autre part parce que  $\mathbf{u}$  est un zéro isolé, il existe un nombre  $\epsilon > 0$  tel que  $DF(\mathbf{v})$  soit inversible pour tout  $\mathbf{v} \in B(\mathbf{u}) = \overline{B(\mathbf{u}, \epsilon)}$ , et tel que  $\mathbf{u}$  soit le seul zéro de  $f$  dans  $B(\mathbf{u})$ . Supposons que  $\mathbf{u}_n$  soit dans  $B(\mathbf{u})$ , de sorte que  $DF(\mathbf{u}_n)$  soit inversible. Puisque  $F(\mathbf{u}) = 0$ , on a  $\mathbf{u}_{n+1} - \mathbf{u} = \mathbf{u}_n - \mathbf{u} - (DF(\mathbf{u}_n))^{-1}(F(\mathbf{u}_n) - F(\mathbf{u}))$ . En effectuant un développement de Taylor-Young à l'ordre 1 de  $F$  au voisinage de  $\mathbf{u}_n$ , cette formule devient  $\mathbf{u}_{n+1} - \mathbf{u} = (DF(\mathbf{u}_n))^{-1}(\|\mathbf{u} - \mathbf{u}_n\| + o(\|\mathbf{u} - \mathbf{u}_n\|))$ . Puisque  $DF$  est continue sur le compact  $B(\mathbf{u})$ , elle y est uniformément continue et donc il existe une constante  $C$  ne dépendant pas de  $n$  telle que  $\|\mathbf{u}_{n+1} - \mathbf{u}\| \leq C\|\mathbf{u}_n - \mathbf{u}\|$ . En choisissant  $\epsilon$  suffisamment petit pour que  $\epsilon C < 1$ ,  $\mathbf{u}_{n+1}$  reste dans la boule  $B(\mathbf{u})$ . Ainsi par récurrence, avec ce choix de  $\epsilon$  et en prenant  $\mathbf{u}_0$  dans  $B(\mathbf{u})$  la suite  $(\mathbf{u}_n)_{n \in \mathbb{N}}$  reste dans  $B(\mathbf{u})$ , il existe  $K = \epsilon C < 1$  tel que  $\|\mathbf{u}_n - \mathbf{u}\| \leq K^n \|\mathbf{u}_1 - \mathbf{u}_0\|$  et donc  $(\mathbf{u}_n)_{n \in \mathbb{N}}$  converge géométriquement vers  $\mathbf{u}$ .  $\square$

**Avantage :** La convergence est rapide. Si  $F$  est supposée de classe  $C^2$  la convergence est même quadratique (la preuve procède de la même façon ; il suffit de poursuivre le développement de Taylor-Young jusqu'à l'ordre 2, et d'utiliser la continuité uniforme de la différentielle seconde ; on la laisse en guise d'exercice.)

**Désavantages :** – Il faut prendre  $\mathbf{u}_0$  suffisamment proche du zéro  $\mathbf{u}$ .

– Le calcul de la matrice jacobienne et son inversion sont coûteux en temps de calcul. Pour cette raison ont été développées des méthodes dites 'quasi-Newton' où  $DF(\mathbf{u}_n)$  est remplacée par une matrice moins coûteuse à inverser ; par exemple la matrice identité : c'est la *méthode des approximations successives*.

**Application à la recherche de minimum.** Pour résoudre

$$\min f(\mathbf{x})$$

avec  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ , on applique la méthode de Newton à  $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ .

**Théorème IV.2 (Application de la méthode de Newton à l'optimisation.)** Soit  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  de classe  $C^2$  et  $\mathbf{u}$  un minimum local de  $f$  isolé. Si  $\nabla^2 f(\mathbf{u})$  est définie positive, alors  $\exists B(\mathbf{u})$  une boule centrée en  $\mathbf{u}$ , tel que  $\forall \mathbf{u}_0 \in B(\mathbf{u})$ , la suite  $\mathbf{u}_n$  définie par :

$$\mathbf{u}_{n+1} = \mathbf{u}_n - \nabla^2 f(\mathbf{u}_n)^{-1} \nabla f(\mathbf{u}_n)$$

converge géométriquement vers le minimum  $\mathbf{u}$ .

**Remarques.** – Si  $f$  est de classe  $C^3$  la convergence est même quadratique !

– Il s'agit d'une *méthode locale* dans le sens où il faut être suffisamment proche d'un extremum (que justement l'on cherche) pour converger. On peut raffiner le résultat pour



majorer cette distance.

- Notons que si la méthode de Newton a pour désavantage de ne converger que localement, elle a par contre l'avantage d'être la seule méthode de ce chapitre qui ne nécessite aucune hypothèse de convexité ; elle a donc un champ d'applications très large.
- Cette méthode est amplement employée en informatique, du fait de sa rapidité de convergence, par exemple pour le calcul approché de  $\sqrt{\alpha}$  (avec  $f(x) = x^2 - \alpha$ ) ou de  $1/\alpha$  (avec  $f(x) = \alpha x - 1$ ). Voir l'exercice 1.

### IV.1.2 Méthode de relaxation

Dans une méthode de descente, on construit une suite  $(\mathbf{u}_n)_n$  en choisissant en chaque point  $\mathbf{u}_n$  une direction de descente  $\mathbf{d}_n$  et un pas de descente  $\rho_n$ . Une méthode pour construire  $\rho_n$  peut consister à se ramener à un problème d'optimisation en dimension 1 par :

$$f(\mathbf{u}_n + \rho_n \mathbf{d}_n) = \inf_{\rho \in \mathbb{R}} f(\mathbf{u}_n + \rho \mathbf{d}_n)$$

La méthode de relaxation consiste en une telle méthode, où l'on prend pour direction de descente successivement chacun des axes. Plus formellement la suite  $(\mathbf{u}_k)_{k \in \mathbb{N}}$  est construite ainsi :  $\mathbf{u}_0$  est choisi arbitrairement, en pratique, si possible, proche d'un minimum, et si

$$\mathbf{u}_k = (x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)})$$

$$\mathbf{u}_{k+1} = (x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_n^{(k+1)}) \text{ est construit par :}$$

$$f(x_1^{(k+1)}, x_2^{(k)}, \dots, x_n^{(k)}) = \inf_{x \in \mathbb{R}} f(x, x_2^{(k)}, \dots, x_n^{(k)})$$

$$f(x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_n^{(k)}) = \inf_{x \in \mathbb{R}} f(x_1^{(k+1)}, x, \dots, x_n^{(k)})$$

⋮

$$f(x_1^{(k+1)}, \dots, x_{n-1}^{(k+1)}, x_n^{(k+1)}) = \inf_{x \in \mathbb{R}} f(x_1^{(k+1)}, \dots, x_{n-1}^{(k+1)}, x)$$

**Théorème IV.3 (Convergence de la méthode de relaxation.)** *Si  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  est elliptique la méthode de relaxation converge vers son unique minimum.*

**Démonstration.** Notons  $u_{k:l} = (x_1^{k+1}, \dots, x_l^{k+1}, x_l^k, \dots, x_n^k)$ , de sorte que  $u_k = u_{k:0}$  et  $u_{k+1} = u_{k:n}$ . Notons  $\mathbf{e}_1, \dots, \mathbf{e}_n$  les vecteurs de la base canonique. Puisque  $f$  est elliptique, il en est de même de chacune des applications  $\varphi_{k:l} : \rho \in \mathbb{R} \rightarrow f(\mathbf{u}_{k:l-1} + \rho \mathbf{e}_l)$ , qui admet donc un unique minimum global (cf. théorème II.7) caractérisé par l'équation d'Euler  $\varphi'_{k:l}(\mathbf{u}_{k:l}) = 0$ . Le point  $\mathbf{u}_{k,l}$  est donc bien défini, et donc la suite  $(\mathbf{u}_n)_{n \in \mathbb{N}}$  aussi. Ecrivons :

$$f(\mathbf{u}_k) - f(\mathbf{u}_{k+1}) = f(\mathbf{u}_{k:0}) - f(\mathbf{u}_{k:n}) = \sum_{l=1}^n f(\mathbf{u}_{k:l-1}) - f(\mathbf{u}_{k:l}) .$$

Puisque  $f$  est  $\alpha$ -elliptique,

$$f(\mathbf{u}_{k:l-1}) - f(\mathbf{u}_{k:l}) \geq \langle \nabla f(\mathbf{u}_{k:l}), \mathbf{u}_{k:l-1} - \mathbf{u}_{k:l} \rangle + \frac{\alpha}{2} \|\mathbf{u}_{k:l-1} - \mathbf{u}_{k:l}\|^2 .$$

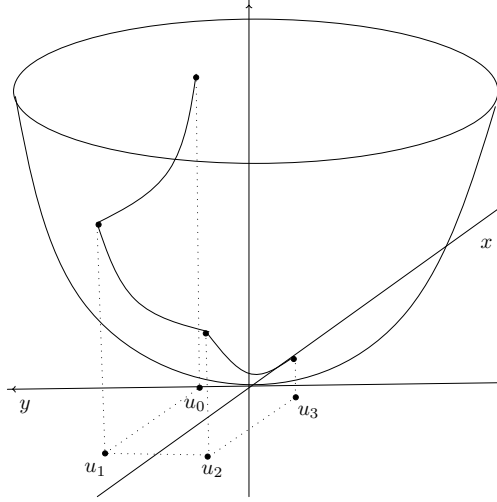


FIGURE IV.2 – Dans la méthode de relaxation on prend comme direction de descente successivement les directions des axes et on détermine le pas en résolvant un problème de minimisation à une variable.

Or par construction pour  $1 \leq l \leq n$ ,  $\langle \nabla f(\mathbf{u}_{k:l}), \mathbf{u}_{k:l-1} - \mathbf{u}_{k:l} \rangle = \frac{\partial f}{\partial x_l}(\mathbf{u}_{k:l})(x_l^k - x_l^{k+1}) = 0$ . Et comme  $\|\mathbf{u}_{k:l-1} - \mathbf{u}_{k:l}\|^2 = |x_l^k - x_l^{k+1}|^2$  pour  $1 \leq l \leq n$ , on obtient finalement :

$$f(\mathbf{u}_k) - f(\mathbf{u}_{k+1}) \geq \frac{\alpha}{2} \sum_{l=1}^n |x_l^k - x_l^{k+1}|^2 = \frac{\alpha}{2} \|\mathbf{u}_k - \mathbf{u}_{k+1}\|^2. \quad (*)$$

La suite  $(f(\mathbf{u}_k))_{k \in \mathbb{N}}$  est décroissante, par construction, et minorée, puisque  $f$  admet un minimum (car elliptique), et donc convergente. Avec (\*) on en déduit que  $\lim_{k \rightarrow +\infty} \|\mathbf{u}_k - \mathbf{u}_{k+1}\| = 0$ , et donc aussi  $\lim_{k \rightarrow +\infty} \|\mathbf{u}_{k:l} - \mathbf{u}_{k+1}\| = 0$ , pour  $0 \leq l \leq n-1$ .

Notons  $\mathbf{u} = (x_1, \dots, x_n)$  le minimum de  $f$ . Puisque  $f$  est  $\alpha$ -elliptique et que  $\nabla f(\mathbf{u}) = \mathbf{0}$  (condition d'Euler) :

$$\alpha \|\mathbf{u}_{k+1} - \mathbf{u}\|^2 \leq \langle \nabla f(\mathbf{u}_{k+1}) - \nabla f(\mathbf{u}), \mathbf{u}_{k+1} - \mathbf{u} \rangle = \langle \nabla f(\mathbf{u}_{k+1}), \mathbf{u}_{k+1} - \mathbf{u} \rangle = \sum_{l=1}^n \frac{\partial f}{\partial x_l}(\mathbf{u}_{k+1})(x_l^{k+1} - x_l),$$

avec l'inégalité de Cauchy-Schwartz,

$$\sum_{l=1}^n \frac{\partial f}{\partial x_l}(\mathbf{u}_{k+1})(x_l^{k+1} - x_l) \leq \left( \sum_{l=1}^n \left( \frac{\partial f}{\partial x_l}(\mathbf{u}_{k+1}) \right)^2 \right)^{\frac{1}{2}} \|\mathbf{u}_{k+1} - \mathbf{u}\|$$

et il découle alors de ces deux dernières inégalités et du fait que par construction  $\frac{\partial f}{\partial x_l}(\mathbf{u}_{k:l}) = 0$  (condition d'Euler), que

$$\|\mathbf{u}_{k+1} - \mathbf{u}\| \leq \frac{1}{\alpha} \left( \sum_{l=1}^n \left( \frac{\partial f}{\partial x_l}(\mathbf{u}_{k+1}) - \frac{\partial f}{\partial x_l}(\mathbf{u}_{k:l}) \right)^2 \right)^{\frac{1}{2}} \quad (**)$$

Or puisque  $\lim_{k \rightarrow +\infty} \|\mathbf{u}_{k:l} - \mathbf{u}_{k+1}\| = 0$ , et que  $x \mapsto \frac{\partial f}{\partial x_l}(x)$  est continue et donc uniformément continue sur tout compact, on en déduit :  $\lim_{k \rightarrow +\infty} (\frac{\partial f}{\partial x_l}(\mathbf{u}_{k+1}) - \frac{\partial f}{\partial x_l}(\mathbf{u}_{k:l})) = 0$ . On déduit alors de (\*\*) que  $\mathbf{u}_k$  tend vers  $\mathbf{u}$ .  $\square$

**Remarques.** – L’inégalité (\*\*) obtenue dans la preuve donne une majoration de l’erreur à l’étape  $k + 1$ .

– Le théorème reste vrai sous les hypothèses plus générales où  $f$  est  $C^1$ , strictement convexe et coercive. La preuve en est cependant plus délicate.

### IV.1.3 Méthode de gradient à pas optimal

Une méthode de type gradient est une méthode de descente où la direction choisie en chaque point  $\mathbf{x}$  est celle de plus grande pente, c’est à dire<sup>1</sup> :  $-\nabla f(\mathbf{x})$ .

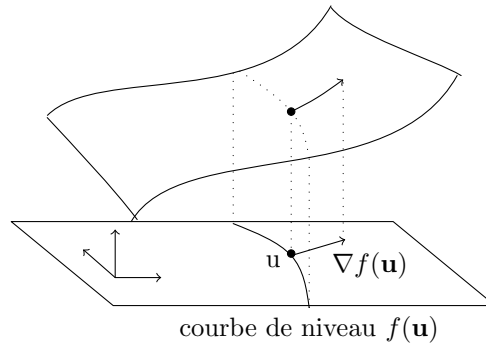


FIGURE IV.3 – La direction locale en  $\mathbf{u}$  de plus grand accroissement de  $f$  est  $\nabla f(\mathbf{u})$ , la direction de plus grande descente  $-\nabla f(\mathbf{u})$ .

On construit par récurrence une suite de points  $(\mathbf{u}_k)_{k \in \mathbb{N}}$ , par la formule :

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \rho_k \nabla f(\mathbf{u}_k)$$

La *méthode du gradient à pas optimal* détermine à chaque itération le pas  $\rho_k$  par :

$$f(\mathbf{u}_k - \rho_k \nabla f(\mathbf{u}_k)) = \inf_{\rho \in \mathbb{R}} f(\mathbf{u}_k - \rho \nabla f(\mathbf{u}_k))$$

c’est à dire en se ramenant à un problème à une seule variable.

#### **Théorème IV.4 (Convergence de la méthode du gradient à pas optimal.)**

Si  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  est  $\alpha$ -elliptique, la méthode du gradient à pas optimal converge vers l’unique minimum de  $f$ . L’erreur à l’étape  $k$  est majorée par :

$$\|\mathbf{r}_k\| = \|\mathbf{u}_k - \mathbf{u}\| \leq \frac{1}{\alpha} \|\nabla f(\mathbf{u}_k)\|.$$

1. cela découle immédiatement de la formule de Taylor-Young à l’ordre 1.

**Démonstration.** L'ellipticité de  $f$  implique l'existence d'un unique minimum  $\mathbf{u}$  caractérisé par l'équation d'Euler  $\nabla f(\mathbf{u}) = \mathbf{0}$  (cf. théorèmes II.7 et II.6). Sans perte de généralité on suppose que  $\forall k \geq 0, \nabla f(\mathbf{u}_k) \neq \mathbf{0}$ , car autrement la méthode est convergente en un nombre fini d'itérations. Chacune des applications  $\varphi_k : \rho \in \mathbb{R} \mapsto f(\mathbf{u}_k - \rho \nabla f(\mathbf{u}_k))$  est aussi elliptique et admet donc un unique minimum  $\rho_k$  caractérisé par l'équation d'Euler  $\varphi'_k(\rho_k) = 0$ . La formule de dérivation d'une application composée donne :

$$\varphi'_k(\rho) = -\langle \nabla f(\mathbf{u}_k - \rho \nabla f(\mathbf{u}_k)), \nabla f(\mathbf{u}_k) \rangle, \quad (i)$$

d'où on déduit la relation :

$$\langle \nabla f(\mathbf{u}_{k+1}), \nabla f(\mathbf{u}_k) \rangle = 0 \quad (ii)$$

qui montre que **deux directions de descente successives sont orthogonales**. Puisque  $\mathbf{u}_{k+1} = \mathbf{u}_k - \rho_k \nabla f(\mathbf{u}_k)$ , on déduit de (i) que  $\langle \nabla f(\mathbf{u}_{k+1}), \mathbf{u}_{k+1} - \mathbf{u}_k \rangle = 0$ . Donc par ellipticité de  $f$  (théorème II.3)  $f(\mathbf{u}_k) - f(\mathbf{u}_{k+1}) \geq \frac{\alpha}{2} \|\mathbf{u}_k - \mathbf{u}_{k+1}\|^2$ . Or la suite  $(f(\mathbf{u}_k))_{k \in \mathbb{N}}$  est décroissante par construction et minorée par sa valeur minimale  $f(\mathbf{u})$ , d'où on déduit que  $\lim_{k \rightarrow \infty} (f(\mathbf{u}_k) - f(\mathbf{u}_{k+1})) = 0$ , et avec la dernière équation on en déduit que :

$$\lim_{k \rightarrow \infty} \|\mathbf{u}_k - \mathbf{u}_{k+1}\| = 0. \quad (iii)$$

En utilisant (ii) d'une part  $\|\nabla f(\mathbf{u}_k)\|^2 \leq \langle \nabla f(\mathbf{u}_k), \nabla f(\mathbf{u}_k) - \nabla f(\mathbf{u}_{k+1}) \rangle$  et d'autre part avec l'inégalité de Cauchy-Schwartz,  $\langle \nabla f(\mathbf{u}_k), \nabla f(\mathbf{u}_k) - \nabla f(\mathbf{u}_{k+1}) \rangle \leq \|\nabla f(\mathbf{u}_k)\| \|\nabla f(\mathbf{u}_k) - \nabla f(\mathbf{u}_{k+1})\|$  et donc

$$\|\nabla f(\mathbf{u}_k)\| \leq \|\nabla f(\mathbf{u}_k) - \nabla f(\mathbf{u}_{k+1})\|. \quad (iv)$$

Puisque la suite  $(f(\mathbf{u}_k))_{k \in \mathbb{N}}$  est décroissante elle est bornée, et  $f$  étant coercive (cf. théorème II.7) la suite  $(\mathbf{u}_k)_{k \in \mathbb{N}}$  est aussi nécessairement bornée. Puisque  $\nabla f$  est continue par hypothèse, elle est uniformément continue sur les compacts. On déduit alors de (iii) que  $\lim_{k \rightarrow \infty} \|\nabla f(\mathbf{u}_k) - \nabla f(\mathbf{u}_{k+1})\| = 0$ , et avec (iv) que

$$\lim_{k \rightarrow \infty} \nabla f(\mathbf{u}_k) = \mathbf{0}. \quad (v)$$

En utilisant successivement, l' $\alpha$ -ellipticité de  $f$ , la condition  $\nabla f(\mathbf{u}) = \mathbf{0}$  et l'inégalité de Cauchy-Schwartz, on obtient :

$$\alpha \|\mathbf{u}_k - \mathbf{u}\|^2 \leq \langle \nabla f(\mathbf{u}_k) - \nabla f(\mathbf{u}), \mathbf{u}_k - \mathbf{u} \rangle = \langle \nabla f(\mathbf{u}_k), \mathbf{u}_k - \mathbf{u} \rangle \leq \|\nabla f(\mathbf{u}_k)\| \|\mathbf{u}_k - \mathbf{u}\|$$

dont on déduit :

$$\|\mathbf{u}_k - \mathbf{u}\| \leq \frac{1}{\alpha} \|\nabla f(\mathbf{u}_k)\|$$

et il découle alors de (v) que la suite  $\mathbf{u}_k$  converge vers  $\mathbf{u}$ .  $\square$

**Remarque.** Un point essentiel de la preuve réside dans le fait que  $\nabla f(\mathbf{u}_k)$  et  $\nabla f(\mathbf{u}_{k+1})$  sont orthogonaux. On peut mettre à profit cela pour s'abstenir de résoudre à chaque étape un problème d'optimisation à une variable dans le cas d'une fonction quadratique elliptique. C'est ce que nous faisons ci-dessous.

### Le cas d'une fonction quadratique elliptique.

Soit  $f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top A \mathbf{x} - \mathbf{b}^\top \mathbf{x} + c$  une fonction quadratique elliptique (*i.e.*  $A$  est définie positive). Le théorème précédent s'applique, mais de plus on peut ici donner une formule explicite pour le pas optimal  $\rho_k$ .

**Théorème IV.5 (Pas optimal en programmation quadratique elliptique.)** *Dans le cas de la fonction quadratique elliptique  $f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top A \mathbf{x} - \mathbf{b}^\top \mathbf{x} + c$ , le pas optimal  $\rho_k$  est donné par :*

$$\rho_k = \frac{\|A\mathbf{u}_k - \mathbf{b}\|^2}{\langle A(\mathbf{u}_k - \mathbf{u}), A(\mathbf{u}_k - \mathbf{u}) \rangle} = \frac{\|\nabla f(\mathbf{u}_k)\|^2}{\langle A\nabla f(\mathbf{u}_k), \nabla f(\mathbf{u}_k) \rangle}.$$

**Démonstration.** Mettons à profit le fait établi dans la preuve du théorème IV.4 que  $\nabla f(\mathbf{u}_k)$  et  $\nabla f(\mathbf{u}_{k+1})$  sont orthogonaux. Puisque :

$$\begin{aligned}\langle \nabla f(\mathbf{u}_{k+1}), \nabla f(\mathbf{u}_k) \rangle &= 0 \\ &= \langle A(\mathbf{u}_k - \rho_k(A\mathbf{u}_k - b)) - b, A\mathbf{u}_k - b \rangle \quad (\text{cf. théorème II.8}).\end{aligned}$$

Par bilinéarité du produit scalaire :

$$\begin{aligned}\Rightarrow \quad \rho_k \langle A(A\mathbf{u}_k - b), A\mathbf{u}_k - b \rangle &= \langle A\mathbf{u}_k - b, A\mathbf{u}_k - b \rangle \\ \Rightarrow \quad \rho_k &= \frac{\|A\mathbf{u}_k - b\|^2}{\langle A(A\mathbf{u}_k - b), A\mathbf{u}_k - b \rangle} = \frac{\|\mathbf{d}_k\|^2}{\langle A\mathbf{d}_k, \mathbf{d}_k \rangle}.\end{aligned}$$

□

#### IV.1.4 Méthode du gradient à pas fixe

Les méthodes de relaxation et de gradient à pas optimal ont en commun la recherche à chaque pas d'un pas de descente optimal, en se ramenant à un problème uni-dimensionnel. C'est pour s'abstraire de cette recherche du pas qu'on développe la méthode du gradient à pas fixe. Il s'agit d'une méthode de gradient où le pas de descente est fixé à  $\rho > 0$  :

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \rho \nabla f(\mathbf{u}_k).$$

Sous des hypothèses suffisantes, on peut choisir le pas pour s'assurer de la convergence.

#### Théorème IV.6 (Convergence de la méthode du gradient à pas fixe.)

Soit  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  une application  $\alpha$ -elliptique dont la différentielle est lipschitzienne, c'est à dire qu'il existe  $M > 0$  telle que  $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ ,

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq M \|\mathbf{x} - \mathbf{y}\|.$$

Si le pas  $\rho$  est choisi tel que :

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

alors la méthode du gradient à pas fixe converge géométriquement vers l'unique minimum global de  $f$ .

**Démonstration.** Par ellipticité de  $f$  le minimum  $\mathbf{u}$  existe, est unique, et est caractérisé par l'équation d'Euler  $\nabla f(\mathbf{u}) = \mathbf{0}$ . On peut donc écrire  $\mathbf{u}_{k+1} - \mathbf{u} = (\mathbf{u}_k - \mathbf{u}) - \rho(\nabla f(\mathbf{u}_k) - \nabla f(\mathbf{u}))$ . Ainsi, et en utilisant l' $\alpha$ -ellipticité de  $f$  et le fait que  $\nabla f$  est  $M$ -lipschitzienne :

$$\|\mathbf{u}_{k+1} - \mathbf{u}\|^2 = \|\mathbf{u}_k - \mathbf{u}\|^2 - 2\rho \langle \nabla f(\mathbf{u}_k) - \nabla f(\mathbf{u}), \mathbf{u}_k - \mathbf{u} \rangle + \rho^2 \|\nabla f(\mathbf{u}_k) - \nabla f(\mathbf{u})\|^2 \leq (1 - 2\alpha\rho + M^2\rho^2) \|\mathbf{u}_k - \mathbf{u}\|^2.$$

On vérifie facilement que le trinôme  $t(\rho) = 1 - 2\alpha\rho + M^2\rho^2$  est convexe et a une valeur comprise dans  $]0, 1[$  si et seulement si  $\rho \in ]0, \frac{2\alpha}{M^2}[$ . Alors si  $0 < a \leq \rho \leq b < \frac{2\alpha}{M^2}$ ,

$$\sqrt{1 - 2\alpha\rho + M^2\rho^2} \leq \beta \triangleq \sqrt{\max\{t(a), t(b)\}} < 1.$$

On a alors  $\|\mathbf{u}_{k+1} - \mathbf{u}\| \leq \beta \|\mathbf{u}_k - \mathbf{u}\| \leq \beta^{k+1} \|\mathbf{u}_0 - \mathbf{u}\|$ , et la suite  $(\mathbf{u}_k)_{k \in \mathbb{N}}$  converge donc géométriquement vers  $\mathbf{u}$ . □

**Remarques.** – Lorsque  $f$  est deux fois différentiable les hypothèses du théorème reviennent à l'existence de deux réels strictement positifs  $\alpha \leq M$  tels que pour tout  $\mathbf{x} \in \mathbb{R}^n$ , toutes les valeurs propres de  $\nabla^2 f(\mathbf{x})$  sont dans  $[\alpha, M]$ .

– En général le meilleur pas de descente donné par la preuve est  $\rho = \alpha/M^2 = \min_{\rho} (1 - 2\alpha\rho + M^2\rho^2)$ .

– Pour  $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{b}^\top \mathbf{x}$  une fonction quadratique elliptique,  $\alpha$  et  $M$  sont respectivement donnés par la plus petite (la plus grande) valeur propre de  $A$ . On peut vérifier que dans ce cas le meilleur pas de descente est  $\frac{2}{\lambda_1 + \lambda_n}$  où  $\lambda_1, \lambda_n$  désignent la plus petite et la plus grande valeur propre de  $A$ .

– Il faut noter que contrairement à la méthode de relaxation ou du gradient à pas optimal on peut avoir  $f(\mathbf{u}_{k+1}) > f(\mathbf{u}_k)$ .

#### IV.1.5 Méthode du gradient conjugué

Même si la direction opposée au gradient est localement la direction de plus grande descente locale, ce n'est pas en appliquant une méthode de descente du type gradient que l'on converge le plus rapidement vers un minimum. Et ce n'est pas ce que l'on peut faire de mieux à l'ordre 1. L'idée de la méthode du gradient conjugué est de construire  $\mathbf{u}_{k+1}$  comme le minimum de la fonction sur l'espace affine  $\mathbf{u}_k + \langle \mathbf{d}_0, \dots, \mathbf{d}_k \rangle$ . Dans le cas d'une fonction quadratique elliptique la méthode converge en au plus  $n$ -itérations : c'est une méthode exacte particulièrement rapide. Le point essentiel réside dans le fait que dans ce cas la famille des directions successives  $\mathbf{d}_0, \dots, \mathbf{d}_k$  est orthogonale pour le produit scalaire associé à la matrice  $A$ , et en particulier est une famille libre dans un espace vectoriel de dimension finie. Cette méthode ingénieuse prend en compte la géométrie globale de la nappe représentative de la fonction. Cette propriété n'est plus vérifiée pour une fonction quelconque. Elle se généralise cependant à des fonctions non quadratiques par des méthodes telles que *Fletcher-Reeves*<sup>2</sup> ou *Polak-Ribière*<sup>3</sup>.

Soit  $f$  une fonction quadratique elliptique,  $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top A\mathbf{x} - \mathbf{b}^\top \mathbf{x} + c$ . La méthode du gradient conjugué est la méthode de descente définie par :

• **Etape 1 :**

$$\mathbf{d}_0 = \nabla f(\mathbf{u}_0) = A\mathbf{u}_0 - \mathbf{b}$$

$$\rho_0 = \frac{\|\mathbf{d}_0\|^2}{\langle A\mathbf{d}_0, \mathbf{d}_0 \rangle}$$

$$\mathbf{u}_1 = \mathbf{u}_0 - \rho_0 \mathbf{d}_0$$

2. R.FLETCHER, C.M.REEVES, *Function minimization by conjugate gradients*, Computer Journal, **7** (1964), pp.149–154.

3. E.POLAK, G.RIBIÈRE, *Sur la convergence de la méthode des gradients conjugués*, Revue Française d'Informatique et de Recherche Opérationnelle, **16**(1) (1969).

• **Etape  $k + 1$  :**

$$\begin{aligned} \mathbf{d}_k &= \nabla f(\mathbf{u}_k) + \frac{\|\nabla f(\mathbf{u}_k)\|^2}{\|\nabla f(\mathbf{u}_{k-1})\|^2} \mathbf{d}_{k-1} \quad \text{soit :} \quad \mathbf{d}_k = A\mathbf{u}_k - \mathbf{b} + \frac{\|A\mathbf{u}_k - \mathbf{b}\|^2}{\|A\mathbf{u}_{k-1} - \mathbf{b}\|^2} \mathbf{d}_{k-1} \\ \rho_k &= \frac{\langle \nabla f(\mathbf{u}_k), \mathbf{d}_k \rangle}{\langle A\mathbf{d}_k, \mathbf{d}_k \rangle} \quad \rho_k = \frac{\langle A\mathbf{u}_k - \mathbf{b}, \mathbf{d}_k \rangle}{\langle A\mathbf{d}_k, \mathbf{d}_k \rangle} \\ \mathbf{u}_{k+1} &= \mathbf{u}_k - \rho_k \mathbf{d}_k \end{aligned}$$

Et ce tant que  $\nabla f(\mathbf{u}_k) = A\mathbf{u}_k - \mathbf{b} \neq \mathbf{0}$ .

**Théorème IV.7 (Convergence de la méthode du gradient conjugué.)** *La méthode du gradient conjugué appliquée à une fonction quadratique elliptique de  $\mathbb{R}^n$  converge en au plus  $n$  itérations.*

**Démonstration.** Puisque  $f$  est elliptique, il existe un unique minimum  $u$  caractérisé par l'équation d'Euler  $\nabla f(\mathbf{u}) = A\mathbf{u} - \mathbf{b} = \mathbf{0}$ . Aussi lorsque  $\nabla f(\mathbf{u}_k) = \mathbf{0}$  l'algorithme s'arrête (devient stationnaire) en  $\mathbf{u}_k$  minimum de  $f$ .

Nous procédons par récurrence pour montrer l'hypothèse suivante :  $\forall k \in \mathbb{N}$  tel que  $\nabla f(\mathbf{u}_l) \neq \mathbf{0}$  et  $\rho_l \neq 0$  pour  $l < k$  :

$$(H_k) : \quad \begin{cases} \langle \nabla f(\mathbf{u}_k), \nabla f(\mathbf{u}_j) \rangle = 0 & \forall 0 \leq j < k & (H_k^1) \\ \langle \nabla f(\mathbf{u}_k), \mathbf{d}_j \rangle = 0 & \forall 0 \leq j < k & (H_k^2) \\ \langle \mathbf{d}_k, A\mathbf{d}_j \rangle = 0 & \forall 0 \leq j < k & (H_k^3) \end{cases}$$

Preuve de  $H_k^1$ . Pour  $k \geq 0$  :

$$\langle \nabla f(\mathbf{u}_{k+1}), \mathbf{d}_k \rangle = \langle A(\mathbf{u}_k - \rho_k \mathbf{d}_k) - \mathbf{b}, \mathbf{d}_k \rangle = \langle A\mathbf{u}_k - \mathbf{b}, \mathbf{d}_k \rangle - \rho_k \langle A\mathbf{d}_k, \mathbf{d}_k \rangle = \underbrace{\langle \nabla f(\mathbf{u}_k), \mathbf{d}_k \rangle - \rho_k \langle A\mathbf{d}_k, \mathbf{d}_k \rangle}_{\text{par définition de } \rho_k} = 0,$$

En particulier cela montre l'étape initiale ( $H_1^1$ ) de la première hypothèse de récurrence ( $H_k^1$ ).

Montrons que  $(H_k) \implies (H_{k+1}^1)$ . On vient de voir que  $\langle \nabla f(\mathbf{u}_{k+1}), \mathbf{d}_k \rangle = 0$ , montrons que  $\langle \nabla f(\mathbf{u}_{k+1}), \mathbf{d}_j \rangle = 0$  pour  $0 \leq j < k$  :

$$\langle \nabla f(\mathbf{u}_{k+1}), \mathbf{d}_j \rangle = \langle \nabla f(\mathbf{u}_{k+1}), \mathbf{d}_j \rangle - \underbrace{\langle \nabla f(\mathbf{u}_k), \mathbf{d}_j \rangle}_{=0} = \underbrace{\langle \nabla f(\mathbf{u}_{k+1}) - \nabla f(\mathbf{u}_k), \mathbf{d}_j \rangle}_{=A(\mathbf{u}_k - \rho_k \mathbf{d}_k) - A\mathbf{u}_k} = \underbrace{-\rho_k \langle A\mathbf{d}_k, \mathbf{d}_j \rangle}_{\text{puisque } A = A^\top} = -\rho_k \langle \mathbf{d}_k, A\mathbf{d}_j \rangle = 0$$

en utilisant l'hypothèse ( $H_k^3$ ).

Preuve de ( $H_k^2$ ). Pour  $0 \leq j \leq k$ ,

$$\langle \nabla f(\mathbf{u}_{k+1}), \nabla f(\mathbf{u}_j) \rangle = \begin{cases} \langle \nabla f(\mathbf{u}_{k+1}), \mathbf{d}_0 \rangle = 0 & \text{si } j = 0, \\ \langle \nabla f(\mathbf{u}_{k+1}), \mathbf{d}_j \rangle - \alpha_j \langle \nabla f(\mathbf{u}_{k+1}), \mathbf{d}_{j-1} \rangle = 0 & \text{si } j > 0 \end{cases}$$

en utilisant  $H_{k+1}^1$  : pour  $j = 0$  puisque  $\mathbf{d}_0 = \nabla f(\mathbf{u}_0)$ , et pour  $j > 0$  car par définition  $\nabla f(\mathbf{u}_j) = \mathbf{d}_j - \alpha_j \mathbf{d}_{j-1}$  (où  $\alpha_j$  est défini ci-dessous).

Il ne reste qu'à montrer  $(H_k) \implies (H_{k+1}^3)$ . Avant cela montrons que pour  $k \geq 1$  :

$$\alpha_k \triangleq \frac{\|\nabla f(\mathbf{u}_k)\|^2}{\|\nabla f(\mathbf{u}_{k-1})\|^2} = -\frac{\langle \nabla f(\mathbf{u}_k), A\mathbf{d}_{k-1} \rangle}{\langle A\mathbf{d}_{k-1}, \mathbf{d}_{k-1} \rangle} \quad \text{de sorte que :} \quad \mathbf{d}_k = \nabla f(\mathbf{u}_k) + \alpha_k \mathbf{d}_{k-1} \quad (*)$$

Puisque  $\nabla f(\mathbf{u}_{k-1}) - \nabla f(\mathbf{u}_k) = A\mathbf{u}_{k-1} - \mathbf{b} - A(\mathbf{u}_{k-1} - \rho_{k-1} \mathbf{d}_{k-1}) - \mathbf{b} = \rho_{k-1} A\mathbf{d}_{k-1}$ , on a :

$$A\mathbf{d}_{k-1} = \frac{\nabla f(\mathbf{u}_{k-1}) - \nabla f(\mathbf{u}_k)}{\rho_{k-1}} \quad \text{et } (H_k^1) \implies \langle \nabla f(\mathbf{u}_k), A\mathbf{d}_{k-1} \rangle = -\frac{\|\nabla f(\mathbf{u}_k)\|^2}{\rho_{k-1}}$$

D'autre part (par construction de  $\rho_{k-1}$  puis en appliquant  $(H_{k-1}^2)$ ) :

$$\langle \text{Ad}_{k-1}, \mathbf{d}_{k-1} \rangle = \frac{\langle \mathbf{d}_{k-1}, \nabla f(\mathbf{u}_{k-1}) \rangle}{\rho_{k-1}} = \frac{\langle \nabla f(\mathbf{u}_{k-1}) + \alpha_{k-1} \mathbf{d}_{k-2}, \nabla f(\mathbf{u}_{k-1}) \rangle}{\rho_{k-1}} = \frac{\|\nabla f(\mathbf{u}_{k-1})\|^2}{\rho_{k-1}}$$

ce qui montre (\*).

Revenons à la preuve de  $(H_k^3)$ . Avec (\*),

$$\langle \mathbf{d}_1, \text{Ad}_0 \rangle = \langle \nabla f(\mathbf{u}_1), \text{Ad}_0 \rangle + \alpha_1 \langle \mathbf{d}_0, \text{Ad}_0 \rangle = 0$$

qui prouve  $(H_1^3)$ . Pour  $0 \leq j < k$ , en utilisant  $\mathbf{d}_{k+1} = \nabla f(\mathbf{u}_{k+1}) + \alpha_{k+1} \mathbf{d}_k$ , on obtient :

$$\langle \mathbf{d}_{k+1}, \text{Ad}_j \rangle = \langle \nabla f(\mathbf{u}_{k+1}), \text{Ad}_j \rangle + \alpha_{k+1} \underbrace{\langle \mathbf{d}_k, \text{Ad}_j \rangle}_{=0} = \underbrace{\langle \nabla f(\mathbf{u}_{k+1}), \text{Ad}_j \rangle}_{\text{puisque } \nabla f(\mathbf{u}_{j+1}) = \nabla f(\mathbf{u}_j) - \rho_j \text{Ad}_j} = \frac{1}{\rho_j} \langle \nabla f(\mathbf{u}_{k+1}), \nabla f(\mathbf{u}_j) - \nabla f(\mathbf{u}_{j+1}) \rangle = 0$$

en appliquant  $(H_{k+1}^1)$ . Par ailleurs, avec (\*) :

$$\langle \mathbf{d}_{k+1}, \text{Ad}_k \rangle = \langle \nabla f(\mathbf{u}_{k+1}), \text{Ad}_k \rangle - \frac{\langle \nabla f(\mathbf{u}_{k+1}), \text{Ad}_k \rangle}{\langle \text{Ad}_k, \mathbf{d}_k \rangle} \langle \mathbf{d}_k, \text{Ad}_k \rangle = \langle \nabla f(\mathbf{u}_{k+1}), \text{Ad}_k \rangle - \langle \nabla f(\mathbf{u}_{k+1}), \text{Ad}_k \rangle = 0$$

ce qui achève la preuve de  $(H_k^3)$  et donc de  $(H_k)$  pour  $k \in \mathbb{N}$  avec  $\nabla f(\mathbf{u}_l) \neq \mathbf{0}$  et  $\rho_l \neq 0$  pour tout  $0 \leq l < k$ .

Or, en appliquant  $(H_k^2)$ , on obtient :

$$\|\mathbf{d}_k\|^2 = \|\nabla f(\mathbf{u}_k)\|^2 + \alpha_k^2 \|\mathbf{d}_{k-1}\|^2 \quad \text{et} \quad \langle \nabla f(\mathbf{u}_k), \mathbf{d}_k \rangle = \langle \nabla f(\mathbf{u}_k), \nabla f(\mathbf{u}_k) \rangle + \alpha_k \langle \nabla f(\mathbf{u}_k), \mathbf{d}_{k-1} \rangle = \|\nabla f(\mathbf{u}_k)\|^2$$

et donc  $\rho_k \neq 0$  et  $\mathbf{d}_k \neq \mathbf{0}$  tant que  $\nabla f(\mathbf{u}_k) \neq \mathbf{0}$ . Ainsi l'algorithme se poursuit tant que  $\nabla f(\mathbf{u}_k) \neq \mathbf{0}$ .

Puisque  $A$  est définie positive,  $(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{x}^\top A \mathbf{y}$  est un produit scalaire. Or avec  $(H_k^3)$  les directions  $\mathbf{d}_0, \dots, \mathbf{d}_k$  sont orthogonales pour ce produit scalaire. En particulier ils forment une famille libre tant qu'ils sont non nuls. Ainsi après au plus  $n$  itérations l'algorithme s'arrête en un point critique, et donc en un minimum.  $\square$

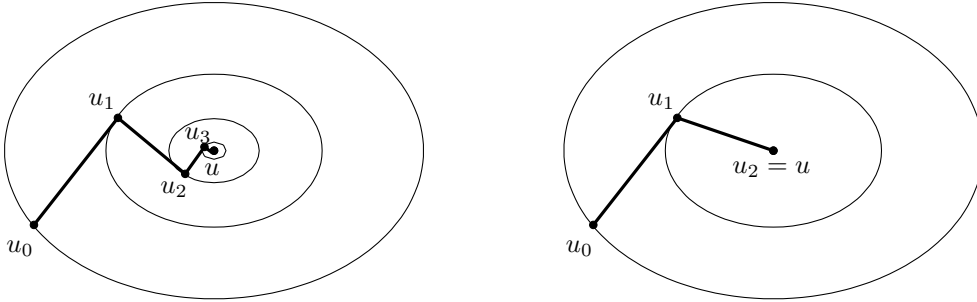


FIGURE IV.4 – Comparaison de la méthode du gradient à pas optimal (à gauche) et de la méthode du gradient conjugué (à droite) pour minimiser  $f(x, y) = \frac{x^2}{a^2} + \frac{y^2}{b^2}$ . Lorsque  $a^2 \neq b^2$  la méthode du gradient conjugué converge en deux étapes, tandis que la méthode du gradient à pas optimal ne converge pas en un nombre fini d'étapes.

**Remarque.** La méthode du gradient conjugué est en fait apparue initialement <sup>4</sup> comme méthode de résolution d'un système d'équations, comme nous le verrons dans le prochain chapitre.

4. M.R.HESTENES, E.STIEFEL, *Methods of conjugate gradients for solving linear systems*, National Bureau of Standards Journal of Research, **49** (1952), pp.409–436.



**Exemple.**

Minimisation de la fonction quadratique :

$$f(x, y, z) = x^2 + y^2 + z^2 + xy + xz + yz + x - y + 3z .$$

Le tableau suivant permet de comparer quatre méthodes de minimisation de l'application quadratique  $f$ . Son minimum est  $(0.25, -1.75, 2.25)$  en lequel la fonction vaut  $-4.375$ .

Itérations	Gradient conjugué	Gradient optimal	Gradient fixe	Relaxation
point initial	$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$
1	$\begin{pmatrix} -0.5454 \\ -0.3181 \\ 1.4545 \end{pmatrix}$	$\begin{pmatrix} -0.5454 \\ -0.3181 \\ 1.4545 \end{pmatrix}$	$\begin{pmatrix} -1.4000 \\ -1.6000 \\ 0.6000 \end{pmatrix}$	$\begin{pmatrix} -2.0000 \\ -1.0000 \\ 3.0000 \end{pmatrix}$
2	$\begin{pmatrix} 0.2500 \\ -1.7500 \\ 2.2500 \end{pmatrix}$	$\begin{pmatrix} 0.3086 \\ -1.4569 \\ 2.3086 \end{pmatrix}$	$\begin{pmatrix} 0.5200 \\ -0.4000 \\ 2.5200 \end{pmatrix}$	$\begin{pmatrix} -0.5000 \\ -1.7500 \\ 2.6250 \end{pmatrix}$
3		$\begin{pmatrix} 0.1878 \\ -1.6381 \\ 2.1878 \end{pmatrix}$	$\begin{pmatrix} -0.3440 \\ -1.6960 \\ 1.6560 \end{pmatrix}$	$\begin{pmatrix} 0.0625 \\ -1.8437 \\ 2.3906 \end{pmatrix}$
5		$\begin{pmatrix} 0.2451 \\ -1.7412 \\ 2.2451 \end{pmatrix}$	$\begin{pmatrix} 0.0361 \\ -1.7305 \\ 2.0361 \end{pmatrix}$	$\begin{pmatrix} 0.2587 \\ -1.7749 \\ 2.2580 \end{pmatrix}$
10		$\begin{pmatrix} 0.2500 \\ -1.7499 \\ 2.2500 \end{pmatrix}$	$\begin{pmatrix} 0.2545 \\ -1.7273 \\ 2.2545 \end{pmatrix}$	$\begin{pmatrix} 0.2500 \\ -1.7499 \\ 2.2499 \end{pmatrix}$
20		$\begin{pmatrix} 0.2500 \\ -1.7499 \\ 2.2500 \end{pmatrix}$	$\begin{pmatrix} 0.2500 \\ -1.7498 \\ 2.2500 \end{pmatrix}$	$\begin{pmatrix} 0.2500 \\ -1.7500 \\ 2.2499 \end{pmatrix}$
30		$\begin{pmatrix} 0.2500 \\ -1.7500 \\ 2.2500 \end{pmatrix}$	$\begin{pmatrix} 0.2500 \\ -1.7499 \\ 2.2500 \end{pmatrix}$	$\begin{pmatrix} 0.2499 \\ -1.7500 \\ 2.2500 \end{pmatrix}$
50			$\begin{pmatrix} 0.2500 \\ -1.7499 \\ 2.2500 \end{pmatrix}$	$\begin{pmatrix} 0.2500 \\ -1.7500 \\ 2.2500 \end{pmatrix}$
80			$\begin{pmatrix} 0.2500 \\ -1.7500 \\ 2.2500 \end{pmatrix}$	

(Voir la figure IV.5 qui donne le code utilisé pour l'implémentation sous `matlab`.)

## IV.2 Méthodes itératives dans le cas sous contraintes

Dans le cas sous contraintes de domaine convexe fermé, on établit des méthodes itératives en appliquant les méthodes sans contraintes vues précédemment tout en projetant à chaque itération le point obtenu sur le domaine. On utilise pour cela le *théorème de projection convexe* que nous rappelons ci-dessous. C'est un grand classique dont on peut trouver une preuve dans l'exercice 4 du chapitre II, page 49.

<pre> %% Méthode du gradient conjugué %% A=[2 1 1;1 2 1;1 1 2]; % matrice A b=[1;-1;3];           % vecteur b % etape 1 u=[1;2;3];            % point initial u0 d=A*u-b;              % d0 r=(d'*d)/(d'*A*d);    % r0 v=u;                  % u0 u=u-r*d;              % u1 % etape k &gt; 1 for i=1:2     Gu=A*u-b; Gv=A*v-b;     d=Gu+(Gu'*Gu)/(Gv'*Gv)*d;     if d==0         break;     end     r=((A*u-b)*d)/(d'*A*d);     v=u;     u=u-r*d;          % uk+1 end umin=u                % minimum de f fmin=0.5*u'*A*u-b'*u % valeur min de f grad=A*u-b           % gradient de f au min </pre>	<pre> %% Méthode du gradient à pas optimal %% A=[2 1 1;1 2 1;1 1 2]; % matrice A b=[1;-1;3];           % vecteur b % Initialisation N=30;                  % Nbre itérations u=[1;2;3];            % point initial u0 % Implémentation for i=1:N     d=A*u-b;          % dk     if d==0         break;     end     r=(d'*d)/(d'*A*d); % rk     u=u-r*d;          % uk+1 end umin=u                % minimum de f fmin=0.5*u'*A*u-b'*u % valeur min de f grad=A*u-b           % gradient de f au min </pre>
<pre> %% Méthode du gradient à pas fixe %% A=[2 1 1;1 2 1;1 1 2]; % matrice A b=[1;-1;3];           % vecteur b % Initialisation N=80;                  % Nbre itérations u=[1;2;3];            % point u0 r=2/(1+4);            % pas de descente % Implémentation for i=1:N     d=A*u-b;          % dk     if d==0         break;     end     u=u-r*d;          % uk+1 end N umin=u                % minimum de f fmin=0.5*u'*A*u-b'*u % valeur min de f grad=A*u-b           % gradient de f au min </pre>	<pre> %% Méthode de relaxation %% A=[2 1 1;1 2 1;1 1 2]; % matrice A b=[1;-1;3];           % vecteur b % Initialisation n=3;                   % dimension N=50;                  % Nbre itérations u=[1;2;3];            % point initial u0 % Implémentation for i=1:N     for j=1:n          % calcul de uk         a=[A(j,1:j-1) A(j,j+1:n)];         v=[u(1:j-1); u(j+1:n)];         u(j)=(b(j)-a*v)/A(j,j);     end end umin=u                % minimum de f fmin=0.5*u'*A*u-b'*u % valeur min de f grad=A*u-b           % gradient de f au min </pre>

FIGURE IV.5 – Le code sous matlab des implémentations des méthodes du gradient à pas conjugué, du gradient à pas variable, du gradient à pas fixe et de la méthode de relaxation.

**Théorème IV.8 (Théorème de projection convexe.)** Soit  $\mathcal{C}$  un sous-ensemble non vide fermé, convexe de  $\mathbb{R}^n$ . Donné  $\mathbf{u} \in \mathbb{R}^n$  il existe un unique  $P_{\mathcal{C}}(\mathbf{u}) \in \mathcal{C}$  tel que

$$\|\mathbf{u} - P_{\mathcal{C}}(\mathbf{u})\| = \inf_{\mathbf{v} \in \mathcal{C}} \|\mathbf{u} - \mathbf{v}\| ,$$

et  $P_{\mathcal{C}}(\mathbf{u})$  est caractérisé par l'inégalité :

$$\forall \mathbf{v} \in \mathcal{C}, \quad \langle P_{\mathcal{C}}(\mathbf{u}) - \mathbf{u}, \mathbf{v} - P_{\mathcal{C}}(\mathbf{u}) \rangle \geq 0 .$$

L'application  $P_{\mathcal{C}} : \mathbb{R}^n \longrightarrow \mathcal{C}$  ainsi définie est appelée l'opérateur de projection sur  $\mathcal{C}$ . C'est

une application contractante, i.e. :

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

### IV.2.1 Méthode de relaxation sur un domaine produit d'intervalles

Considérons une application  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  elliptique que l'on souhaite minimiser sur un domaine de la forme :

$$\mathcal{D} = \prod_{i=1}^n [a_i, b_i] \quad \text{où } a_i, b_i \in \overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}.$$

Dans ce cas  $\mathcal{D}$  est un fermé convexe de  $\mathbb{R}^n$  (car produit direct de fermés convexes de  $\mathbb{R}$ ), et puisque  $f$  est elliptique elle y admet un unique minimum. On adapte alors la méthode de relaxation naturellement par :

$$\begin{aligned} \mathbf{u}_k &= (x_1^{(k)}, x_2^{(k)}, \dots, x_n^{(k)}) \in \mathcal{D} \\ \mathbf{u}_{k+1} &= (x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_n^{(k+1)}) \in \mathcal{D} \text{ est construit par :} \\ f(x_1^{(k+1)}, x_2^{(k)}, \dots, x_n^{(k)}) &= \inf_{a_1 \leq x \leq b_1} f(x, x_2^{(k)}, \dots, x_n^{(k)}) \\ f(x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_n^{(k)}) &= \inf_{a_2 \leq x \leq b_2} f(x_1^{(k+1)}, x, \dots, x_n^{(k)}) \\ &\vdots \\ f(x_1^{(k+1)}, \dots, x_{n-1}^{(k+1)}, x_n^{(k+1)}) &= \inf_{a_n \leq x \leq b_n} f(x_1^{(k+1)}, \dots, x_{n-1}^{(k+1)}, x) \end{aligned}$$

(les inégalités ayant lieu dans  $\overline{\mathbb{R}}$  et étant bien évidemment strictes lorsque  $a_i, b_i = \pm\infty$ .)

#### Théorème IV.9 (Convergence de la méthode de relaxation sous contraintes.)

Soit  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  une application elliptique sur  $\mathcal{D} \subset \mathbb{R}^n$  qui est un produit d'intervalles :

$$\mathcal{D} = \prod_{i=1}^n [a_i, b_i] \quad \text{où } a_i, b_i \in \overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}, \quad a_i \leq b_i.$$

La méthode de relaxation converge vers le minimum de  $f$  sur  $\mathcal{D}$ .

**Démonstration.** La preuve est identique à celle dans le cas sans contraintes, à l'exception près que l'on remplace les conditions nécessaires et suffisantes de minimum :

$$\begin{aligned} \nabla f(\mathbf{u}) &= \mathbf{0}, \quad \text{par } \forall \mathbf{v} \in \mathcal{D}, \quad \nabla f(\mathbf{u})(\mathbf{v} - \mathbf{u}) \geq 0, \text{ et} \\ \frac{\partial f}{\partial x_l}(\mathbf{u}_{k:l}) &= 0, \quad \text{par } \forall v_l \in [a_l, b_l], \quad \frac{\partial f}{\partial x_l}(\mathbf{u}_{k:l})(v_l \mathbf{e}_l - \mathbf{u}_{k:l}) \geq 0, \quad \forall 1 \leq l \leq n \end{aligned}$$

qui sont encore nécessaires et suffisantes pour un minimum  $\mathbf{u} \in \mathcal{D}$  (théorème II.6.1.(iv)). □

### IV.2.2 Méthode du gradient projeté

La méthode du gradient projeté consiste à projeter sur le domaine  $\mathcal{C}$  (convexe, fermé, non vide) les points obtenus à chaque itération par la méthode du gradient à pas fixe. C'est-à-dire, soit  $\rho > 0$  un pas de descente :

$$\mathbf{u}_{k+1} = P_{\mathcal{C}}(\mathbf{u}_k - \rho \nabla f(\mathbf{u}_k)) .$$

Sa convergence est assurée sous les mêmes hypothèses que pour la méthode du gradient à pas fixe par le théorème suivant :

#### Théorème IV.10 (Convergence de la méthode du gradient projeté.)

Soit  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  une application  $\alpha$ -elliptique et un domaine  $\mathcal{C}$  non vide fermé et convexe. On suppose de plus que  $\nabla f : \mathbb{R}^n \rightarrow \mathbb{R}^n$  est  $M$ -lipschitzienne (c'est à dire  $\exists M > 0$ ,  $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ ,  $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq M\|\mathbf{x} - \mathbf{y}\|$ ).

Si le pas de descente  $\rho$  est choisi tel que :

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

alors la méthode du gradient projeté converge géométriquement vers le minimum de  $f$  sur  $\mathcal{C}$ .

**Démonstration.** Puisque  $f$  est elliptique et  $\mathcal{C}$  est fermé convexe non vide,  $f$  admet un unique minimum global  $\mathbf{u}$  sur  $\mathcal{C}$ . Définissons l'application  $g : \mathbb{R}^n \rightarrow \mathcal{C}$  par  $g(\mathbf{x}) = P_{\mathcal{C}}(\mathbf{x} - \rho \nabla f(\mathbf{x}))$ , en prenant  $\rho > 0$ . L'opérateur de projection étant une application contractante, on a :

$$\begin{aligned} \|g(\mathbf{x}) - g(\mathbf{y})\|^2 &= \|P_{\mathcal{C}}(\mathbf{x} - \rho \nabla f(\mathbf{x})) - P_{\mathcal{C}}(\mathbf{y} - \rho \nabla f(\mathbf{y}))\|^2 \\ &\leq \|(\mathbf{x} - \mathbf{y}) - \rho(\nabla f(\mathbf{x}) - \nabla f(\mathbf{y}))\|^2 , \\ &= \|\mathbf{x} - \mathbf{y}\|^2 - 2\rho \langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle + \rho^2 \|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \end{aligned}$$

et puisque  $f$  est  $\alpha$ -elliptique et  $\nabla f$  est  $M$ -lipschitzienne :

$$\leq (1 - 2\rho + M^2\rho^2) \|\mathbf{x} - \mathbf{y}\|^2 .$$

Comme dans la preuve du théorème IV.6, on établit l'existence de  $\beta \geq 0$ , tels que :  $\sqrt{1 - 2\alpha\rho + M^2\rho^2} \leq \beta < 1$ .

Le point  $\mathbf{u}$  est un point fixe de l'application  $g$ . En effet, avec le théorème II.5.1.a,  $\forall \mathbf{x} \in \mathcal{C}$ ,  $\langle \nabla f(\mathbf{u}), \mathbf{x} - \mathbf{u} \rangle \geq 0$ , et donc pour tout  $\rho > 0$  et  $\mathbf{x} \in \mathcal{C}$ ,  $\langle \mathbf{u} - (\mathbf{u} - \rho \nabla f(\mathbf{u})), \mathbf{x} - \mathbf{u} \rangle \geq 0$  qui implique avec la caractérisation de  $P_{\mathcal{C}}(\mathbf{u})$  donnée dans le théorème IV.8 que  $\mathbf{u} = P_{\mathcal{C}}(\mathbf{u} - \rho \nabla f(\mathbf{u})) = g(\mathbf{u})$ . D'autre part chaque élément de la suite  $\mathbf{u}_{k+1} = \mathbf{u}_k - \rho \nabla f(\mathbf{u}_k)$  vérifie  $g(\mathbf{u}_k) = \mathbf{u}_{k+1}$ . Ainsi on a :

$$\|\mathbf{u}_{k+1} - \mathbf{u}\| = \|g(\mathbf{u}_k) - g(\mathbf{u})\| \leq \beta \|\mathbf{u}_k - \mathbf{u}\|$$

qui montre la conclusion. □

Ainsi la méthode du gradient conjugué permet en théorie de déterminer le minimum d'une fonction elliptique à dérivée lipschitzienne sur un convexe fermé quelconque. C'est cependant illusoire : on ne sait pas en général construire l'opérateur de projection sur un convexe. Les seuls exemples notables étant lorsque  $\mathcal{C} = \prod_{i=1}^n [a_i, b_i]$  est un produit d'intervalles, ou lorsque  $\mathcal{C}$  est une boule fermée  $\mathcal{C} = \overline{B(\mathbf{x}_0, r)}$ . Aussi emploie-t-on plutôt, du moins lorsque les contraintes sont affines, la méthode d'Uzawa, que nous allons voir, qui met à profit la notion de dualité et résout le problème dual, où l'opérateur de projection est alors on ne peut plus simple, les contraintes n'étant plus alors que des contraintes de signe.

### IV.2.3 Méthode d'Uzawa

Cette méthode applique la théorie de la dualité convexe (vue au § III.2.6), et recherche dans un problème de programmation convexe un point-selle du lagrangien. Il s'agit en fait de la méthode du gradient projeté appliquée au problème dual. Mais dans ce cas l'opérateur de projection de  $\mathbb{R}^q$  sur  $(\mathbb{R}_+)^q$  est particulièrement simple à écrire ; c'est là qu'en réside tout l'intérêt.

#### Algorithme d'Uzawa.

On construit une suite  $(\mathbf{x}_k)_{k \in \mathbb{N}}$  de  $\mathbb{R}^n$  et deux suites  $(\lambda_k)_{k \in \mathbb{N}}$  de  $\mathbb{R}^p$  et  $(\mu_k)_{k \in \mathbb{N}}$  de  $(\mathbb{R}_+)^q$  de la façon suivante.

- **Initialement.** On fixe  $\rho > 0$  et on choisit arbitrairement  $(\lambda_0, \mu_0) \in \mathbb{R}^p \times (\mathbb{R}_+)^q$ .
- **Itération  $k$ .** On détermine  $\mathbf{x}_k \in \mathbb{R}^n$  par :

$$\mathbf{x}_k \text{ est solution de } \min_{\mathbf{x} \in \mathbb{R}^n} \mathcal{L}(\mathbf{x}, \lambda_k, \mu_k)$$

$$\text{soit encore : } \nabla_x \mathcal{L}(\mathbf{x}_k, \lambda_k, \mu_k) = \mathbf{0} .$$

On détermine  $\lambda_{k+1}$  et  $\mu_{k+1}$  par :

$$\begin{aligned} \lambda_{k+1} &= \lambda_k + \rho \cdot (\varphi_1(\mathbf{x}_k), \dots, \varphi_p(\mathbf{x}_k)) , \\ \mu_{k+1} &= P_{(\mathbb{R}_+)^q}(\mu_k + \rho \cdot (\psi_1(\mathbf{x}_k), \dots, \psi_q(\mathbf{x}_k))) . \end{aligned}$$

Sa convergence est assurée sous certaines hypothèses par le résultat suivant.

**Théorème IV.11 (Convergence de la méthode d'Uzawa.)** *On suppose que  $f$  est  $\alpha$ -elliptique, que  $\varphi_1, \dots, \varphi_p$  et  $\psi_1, \dots, \psi_q$  sont affines, i.e.,*

$$\mathcal{D} = \left\{ \mathbf{x} \in \mathbb{R}^n \mid A\mathbf{x} = \mathbf{b} ; C\mathbf{x} \leq \mathbf{d} \right\}$$

avec  $A \in \mathcal{M}_{p,n}(\mathbb{R})$ ,  $\mathbf{b} \in \mathbb{R}^p$ ,  $C \in \mathcal{M}_{q,n}(\mathbb{R})$  et  $\mathbf{d} \in \mathbb{R}^q$ . Alors en choisissant  $\rho$  tel que

$$0 < \rho < \frac{2\alpha}{\|A\|^2 + \|C\|^2}$$

la suite  $(\mathbf{u}_k)_{k \in \mathbb{N}}$  converge vers l'unique minimum de  $f$  sur  $\mathcal{D}$ .

**Démonstration.** Sous ces hypothèses le domaine  $\mathcal{D}$  est un convexe fermé, et l'application  $f$  étant elliptique, y admet un unique minimum  $\mathbf{u}$ , solution du problème que nous appellerons  $(P)$ . Il en est de même pour chacun des problèmes de minimisation permettant de déterminer  $\mathbf{u}_k$  dans la méthode d'Uzawa. De plus le problème  $(Q)$  dual de  $(P)$  admet aussi une solution.

Pour éviter les confusions, notons pour  $m > 0$ ,  $\langle \cdot, \cdot \rangle_m$ , le produit scalaire usuel de  $\mathbb{R}^m$  et  $\|\cdot\|_m$  la norme associée, tandis que  $\langle \cdot, \cdot \rangle$  et  $\|\cdot\|$  désigneront le produit scalaire usuel de  $\mathbb{R}^n$  et sa norme associée. Avec ces notations :

$$\mathcal{L}(\mathbf{x}, \lambda, \mu) = f(\mathbf{x}) + \langle A\mathbf{x} - \mathbf{b}, \lambda \rangle_p + \langle C\mathbf{x} - \mathbf{d}, \mu \rangle_q = f(\mathbf{x}) + \langle A^\top \lambda, \mathbf{x} \rangle_p - \langle \mathbf{b}, \lambda \rangle_p + \langle C^\top \mu, \mathbf{x} \rangle_q - \langle \mathbf{d}, \mu \rangle_q .$$

On notera encore, pour plus de concision,  $\varphi(\mathbf{x}) = A\mathbf{x} - \mathbf{b} = (\varphi_1(\mathbf{x}), \dots, \varphi_p(\mathbf{x}))$  et  $\psi(\mathbf{x}) = C\mathbf{x} - \mathbf{d} = (\psi_1(\mathbf{x}), \dots, \psi_q(\mathbf{x}))$ .

Soit  $(\lambda^*, \mu^*) \in \mathbb{R}^p \times (\mathbb{R}_+)^q$  une solution du problème dual  $(Q)$ , de sorte que  $(\mathbf{u}, \lambda^*, \mu^*)$  soit un point-selle du lagrangien ; en particulier on vérifie les conditions de KKT :  $\langle \psi(\mathbf{u}), \mu^* \rangle_q = 0$ , et  $\nabla f(\mathbf{u}) + A^\top \lambda^* + C^\top \mu^* = \mathbf{0}$ . Puisque la solution  $\mathbf{u}$  est dans le domaine admissible  $\mathcal{D}$ , d'une part  $\psi(\mathbf{u}) = \mathbf{0}$  et d'autre part il découle de la première

condition de KKT ci-dessus que  $\forall \mu \in (\mathbb{R}_+)^q$ ,  $\langle \psi(\mathbf{u}), \mu - \mu^* \rangle_q \leq 0$ . Cette dernière relation s'écrit aussi pour  $\rho > 0$ ,  $\langle \mu^* - (\mu^* + \rho \psi(\mathbf{u})), \mu - \mu^* \rangle_q \geq 0$  pour tout  $\mu \in (\mathbb{R}_+)^q$ . Ceci montre (cf. théorème IV.8) que  $\mu^*$  est la projection sur  $(\mathbb{R}_+)^q$  de  $\mu^* + \rho \psi(\mathbf{u})$ . On a donc établi :

$$\begin{cases} \nabla f(\mathbf{u}) + A^\top \lambda^* + C^\top \mu^* = \mathbf{0} \\ \lambda^* = \lambda^* + \rho \varphi(\mathbf{u}) \\ \mu^* = P_{(\mathbb{R}_+)^q}(\mu^* + \rho \psi(\mathbf{u})) \end{cases}$$

Par construction de la méthode d'Uzawa, on a pour tout  $k \geq 0$  :

$$\begin{cases} \nabla f(\mathbf{u}_k) + A^\top \lambda_k + C^\top \mu_k = \mathbf{0} \\ \lambda_{k+1} = \lambda_k + \rho \varphi(\mathbf{u}_k) \\ \mu_{k+1} = P_{(\mathbb{R}_+)^q}(\mu_k + \rho \psi(\mathbf{u}_k)) \end{cases}$$

dont on déduit, puisque la projection convexe est contractante :

$$\begin{cases} \nabla f(\mathbf{u}_k) - \nabla f(\mathbf{u}) + A^\top (\lambda_k - \lambda^*) + C^\top (\mu_k - \mu^*) = \mathbf{0} & (1) \\ \|\lambda_{k+1} - \lambda^*\|_p = \|\lambda_k - \lambda^* + \rho A(\mathbf{u}_k - \mathbf{u})\|_p & (2) \\ \|\mu_{k+1} - \mu^*\|_q \leq \|\mu_k - \mu^* + \rho C(\mathbf{u}_k - \mathbf{u})\|_q & (3) \end{cases}$$

Montrons maintenant que  $(\mathbf{u}_k)_{k \in \mathbb{N}}$  converge vers  $\mathbf{u}$ ; nous n'utiliserons que ces trois dernières relations (1), (2), (3). En élevant au carré (2) et (3) on obtient :

$$\begin{aligned} \|\lambda_{k+1} - \lambda^*\|_p^2 &= \|\lambda_k - \lambda^*\|_p^2 + 2\rho \langle A^\top (\lambda_k - \lambda^*), \mathbf{u}_k - \mathbf{u} \rangle_p + \rho^2 \|A(\mathbf{u}_k - \mathbf{u})\|_p^2 \\ \|\mu_{k+1} - \mu^*\|_q^2 &\leq \|\mu_k - \mu^*\|_q^2 + 2\rho \langle C^\top (\mu_k - \mu^*), \mathbf{u}_k - \mathbf{u} \rangle_q + \rho^2 \|C(\mathbf{u}_k - \mathbf{u})\|_q^2 \end{aligned}$$

ce qui donne en les additionnant puis en tenant compte de (1),

$$\begin{aligned} \|(\lambda_{k+1}, \mu_{k+1}) - (\lambda^*, \mu^*)\|_{p+q}^2 &\leq \|(\lambda_k, \mu_k) - (\lambda^*, \mu^*)\|_{p+q}^2 - 2\rho \langle \nabla f(\mathbf{u}_k) - \nabla f(\mathbf{u}), \mathbf{u}_k - \mathbf{u} \rangle \\ &\quad + \rho^2 \|A(\mathbf{u}_k - \mathbf{u})\|_p^2 + \rho^2 \|C(\mathbf{u}_k - \mathbf{u})\|_q^2 \end{aligned}$$

puisque  $f$  est  $\alpha$ -elliptique et par propriété de compatibilité de la norme matricielle,

$$\leq \|(\lambda_k, \mu_k) - (\lambda^*, \mu^*)\|_{p+q}^2 - \rho(2\alpha - \rho(\|A\|^2 + \|C\|^2)) \|\mathbf{u}_k - \mathbf{u}\|_q^2.$$

En particulier, en prenant  $0 \leq \rho \leq \frac{2\alpha}{\|A\|^2 + \|C\|^2}$ ,  $\implies \|(\lambda_{k+1}, \mu_{k+1}) - (\lambda^*, \mu^*)\|_{p+q} \leq \|(\lambda_k, \mu_k) - (\lambda^*, \mu^*)\|_{p+q}$  pour tout  $k \geq 0$ . Ainsi la suite  $(\|(\lambda_k, \mu_k) - (\lambda^*, \mu^*)\|_{p+q})_{k \in \mathbb{N}}$  est décroissante et minorée par 0 et donc convergente. Cela entraîne :

$$\lim_{k \rightarrow \infty} (\|(\lambda_{k+1}, \mu_{k+1}) - (\lambda^*, \mu^*)\|_{p+q}^2 - \|(\lambda_k, \mu_k) - (\lambda^*, \mu^*)\|_{p+q}^2) = 0,$$

et alors :

$$0 < \rho < \frac{2\alpha}{\|A\|^2 + \|C\|^2} \implies \lim_{k \rightarrow \infty} \|\mathbf{u}_k - \mathbf{u}\| = 0$$

Ce qui montre la conclusion.  $\square$

**Remarques.** – Il s'agit en fait essentiellement de la méthode du gradient projeté appliquée au problème dual. Ce qui explique la similarité des conditions de convergence.

– Le résultat reste essentiellement vrai sous des hypothèses plus générales :  $\psi_1, \dots, \psi_q$  sont différentiables et lipschitziennes, et le lagrangien admet un point-selle. La preuve suit les mêmes lignes; on pourra l'adapter en guise d'exercice (par quoi est remplacé  $\|C\|$  dans la conclusion?). Ces conditions sont immédiatement vérifiées lorsque les contraintes inégalitaires sont affines.

– On peut utiliser dans la conclusion n'importe quelle norme matricielle compatible avec la norme vectorielle usuelle ( $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ ), c'est à dire telle que  $\|A\mathbf{x}\| \leq \|A\| \|\mathbf{x}\|$ . On a un large choix, voir §A.3.2 de l'annexe.

## Exercices.

**Exercice 1.** Appliquer la méthode de Newton pour donner une valeur approchée de  $\sqrt{2}$ ,  $\sqrt[3]{2}$ , que l'on comparera avec une valeur donnée par une calculatrice. Essayer plusieurs points initiaux.

**Exercice 2.** Considérons le cas d'un problème de programmation quadratique elliptique sans contraintes :

$$\min_{\mathbf{x}} \frac{1}{2} \mathbf{x}^\top A \mathbf{x} - \mathbf{b}^\top \mathbf{x} .$$

Comment s'exprime dans ce cas la méthode de Newton ? Sa convergence dépend-elle du point initial ? En combien d'itérations converge-t-elle ? Réinterpréter la méthode de Newton.

**Exercice 3.** Considérons le cas d'un problème de programmation quadratique elliptique sans contraintes :

$$\min_{\mathbf{x}} \frac{1}{2} \mathbf{x}^\top A \mathbf{x} - \mathbf{b}^\top \mathbf{x} .$$

Comment s'exprime dans ce cas la méthode de relaxation ?

**Exercice 4.** Considérons le cas d'un problème de programmation quadratique elliptique sous contraintes :

$$\begin{aligned} \min_{\mathbf{x}} \quad & \frac{1}{2} \mathbf{x}^\top A \mathbf{x} - \mathbf{b}^\top \mathbf{x} \\ & C \mathbf{x} = \mathbf{c} \\ & D \mathbf{x} \leq \mathbf{d} \end{aligned}$$

où  $A \in \mathcal{M}_n(\mathbb{R})$  définie positive,  $\mathbf{b} \in \mathbb{R}^n$ ,  $C \in \mathcal{M}_{p,n}(\mathbb{R})$ ,  $\mathbf{c} \in \mathbb{R}^p$ ,  $D \in \mathcal{M}_{q,n}(\mathbb{R})$ ,  $\mathbf{d} \in \mathbb{R}^q$ .

- Comment s'exprime le vecteur gradient du lagrangien  $\mathcal{L}(\mathbf{x}, \lambda, \mu)$  de ce problème ?
- Exprimer les conditions nécessaires de KKT pour ce problème sous forme matricielle.
- Comment s'exprime ici la méthode d'Uzawa ?





## Chapitre V

# Applications aux Maths numériques

Les domaines d'application de l'optimisation sont innombrables. Nous passons ici en revue quelques exemples en mathématiques numériques. Nous avons fait le choix de la simplicité et de la concision ; les développements que nous faisons découlent presque immédiatement des notions abordées dans les chapitres précédents, et auraient tout aussi bien pu être présentés sous forme d'exercices d'application ; par ailleurs la liste que nous donnons est loin d'être exhaustive. Ils revêtent cependant un grand intérêt et sont très largement utilisés.

Nous passons sous silence certains aspects présentant pourtant une grande importance. Notamment le domaine du *calcul variationnel* : il s'agit de l'optimisation d'applications définies non plus sur  $\mathbb{R}^n$  mais sur un espace fonctionnel réel ; un minimum n'est plus un point de  $\mathbb{R}^n$  mais une application définie sur  $\mathbb{R}^n$ . Cela aurait nécessité d'énoncer toute cette théorie sur des espaces vectoriels réels de dimension pouvant être infinie, et plus précisément sur des *espaces de Hilbert*<sup>1</sup>. La plupart des résultats que nous avons vus y restent vrais, sans apporter de difficulté supplémentaire, tandis que le champ d'application de la théorie s'élargit considérablement. Une de ses applications en est la théorie du *contrôle optimal* fondamentale en automatique.

**Exemple de problème variationnel.** *Problème de la brachistochrone.* Quelle forme doit avoir un toboggan pour que la durée de descente (sans frottements) soit minimale. Ce problème revient à déterminer l'application  $f : \mathbb{R} \rightarrow \mathbb{R}$  qui minimise le critère :

$$\int_{x_A}^{x_B} \frac{\sqrt{1 + f'(x)^2}}{\sqrt{f(x)}} dx ?$$

Réponse : c'est une *cycloïde*. Elle peut se décrire comme la courbe décrite par le point d'une roue roulant sur une surface plane.

---

1. Un espace vectoriel réel muni d'un produit scalaire (*i.e.* d'une forme bilinéaire symétrique définie positive) qui est complet pour la norme induite (*i.e.* toute suite de Cauchy est convergente) est appelé un *espace de Hilbert*. En dimension finie il s'agit des *espaces euclidiens*.

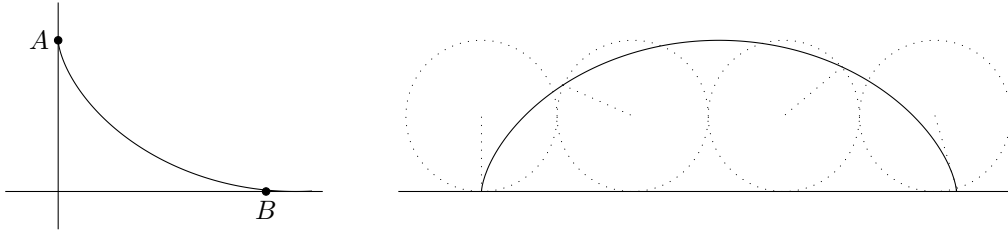


FIGURE V.1 – Une cycloïde décrit le trajet sans frottements du point  $A$  au point  $B$  de durée minimale d'un corps soumis à un champ de pesanteur (à gauche). On peut la voir (à un signe près) comme le trajet que suit la valve d'une roue de vélo (à droite).

## V.1 Résolution approchée d'un système d'équations

### V.1.1 Système d'équations linéaires de Cramer

Soit  $M \in \mathcal{M}_n(\mathbb{R})$  une matrice inversible, et soit  $\mathbf{c} \in \mathbb{R}^n$ . On considère le système d'équations linéaires de Cramer :

$$M\mathbf{x} = \mathbf{c} \quad (*)$$

La résolution d'un tel système intervient très fréquemment dans tous les domaines d'applications mathématiques. Lorsque  $n$  est grand la résolution directe de ce système –par la méthode du pivot de Gauss, ou par les formules de Cramer, par exemple– est fastidieuse et prend un temps de calcul pouvant être pénalisant. Aussi est-il très utile dans la pratique de disposer d'algorithmes de résolution approchée de système d'équations linéaires, plus rapides ou moins gourmands en ressources. Nous allons appliquer les résultats établis dans les précédents chapitres pour y parvenir.

Pour ce faire commençons par nous ramener au cas d'une matrice symétrique définie positive. Il suffit de multiplier à gauche par la matrice transposée  $M^\top$  ; c'est une opération peu coûteuse en ressources.

Poser  $A = M^\top M$  et  $\mathbf{b} = M^\top \mathbf{c}$  ;  $A$  est symétrique définie positive (cf. théorème A.4 page 117) et le système linéaire

$$A\mathbf{x} = \mathbf{b}$$

est équivalent au système linéaire (\*).

Dans la suite nous ne considérerons plus que des systèmes linéaires à matrice symétrique définie positive.

### V.1.2 Système d'équations linéaires à matrice symétrique définie positive

Soit  $A \in \mathcal{M}_n(\mathbb{R})$  une matrice symétrique définie positive ; soit  $\mathbf{b} \in \mathbb{R}^n$ . On souhaite appliquer un algorithme pour déterminer une valeur approchée de la solution du système de Cramer :

$$A\mathbf{x} = \mathbf{b} \quad (*)$$

Résoudre ce système, on l'a vu, équivaut au problème d'optimisation quadratique elliptique :

$$\min_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top A \mathbf{x} - \mathbf{b}^\top \mathbf{x}.$$

En particulier une méthode itérative de recherche de minimum fournit une méthode approchée de résolution du système linéaire.

**Méthode du gradient à pas fixe.** En notant  $\lambda_1, \lambda_n$  la plus petite et la plus grande valeur propre de  $A$ , elle s'écrit ici (voir § IV.1.4) :

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \frac{2}{\lambda_1 + \lambda_n} (A\mathbf{u}_k - \mathbf{b}) .$$

**Méthode du gradient à pas optimal.** Elle s'écrit ici (voir § IV.1.3) :

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \frac{\|A\mathbf{u}_k - \mathbf{b}\|^2}{\langle A(A\mathbf{u}_k - \mathbf{b}), A\mathbf{u}_k - \mathbf{b} \rangle} (A\mathbf{u}_k - \mathbf{b}) .$$

**Méthode du gradient conjugué.** Il s'agit cette fois-ci d'une méthode exacte (si l'on exclut les erreurs d'approximation) qui converge en au plus  $n$  itérations (voir § IV.1.5) :

• **Etape 1 :**

$$\mathbf{d}_0 = A\mathbf{u}_0 - \mathbf{b}$$

$$\rho_0 = \frac{\|\mathbf{d}_0\|^2}{\langle A\mathbf{d}_0, \mathbf{d}_0 \rangle}$$

$$\mathbf{u}_1 = \mathbf{u}_0 - \rho_0 \mathbf{d}_0$$

• **Etape  $k + 1$  :**

$$\mathbf{d}_k = A\mathbf{u}_k - \mathbf{b} + \frac{\|A\mathbf{u}_k - \mathbf{b}\|^2}{\|A\mathbf{u}_{k-1} - \mathbf{b}\|^2} \mathbf{d}_{k-1}$$

$$\rho_k = \frac{\langle A\mathbf{u}_k - \mathbf{b}, \mathbf{d}_k \rangle}{\langle A\mathbf{d}_k, \mathbf{d}_k \rangle}$$

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \rho_k \mathbf{d}_k$$

Tant que  $\nabla f(\mathbf{u}_k) = A\mathbf{u}_k - \mathbf{b} \neq \mathbf{0}$ .

Le nombre d'opérations à effectuer est de l'ordre de  $O(n^3)$  et ne présente pas de grand avantage par rapport à d'autres méthodes directes telles la *méthode de Cholesky*; en outre c'est dans la pratique un leurre de considérer cette méthode comme directe, les erreurs d'approximations dans les calculs successifs nécessitant de poursuivre la méthode au-delà des  $n$  itérations théoriques, en ajoutant un critère d'arrêt (tel  $\|A\mathbf{u}_k - \mathbf{b}\| < \epsilon$ .) Par contre elle présente une très bonne stabilité par rapport aux erreurs d'arrondi. D'autre part pour des matrices creuses (*i.e.* comportant beaucoup de zéros), le calcul des  $A\mathbf{d}_k$ , les plus coûteux numériquement, peut dans ce cas se faire à l'aide de relations de récurrence et améliore considérablement la rapidité de calcul; c'est le cas par exemple dans le cas de discrétisation de problèmes aux limites par des méthodes de différences finies. Cela permet une réduction spectaculaire de la quantité de calculs nécessaires à son application; c'est alors une méthode de résolution approchée des plus efficaces.

### Méthode de Gauss-Seidel.

La méthode de relaxation appliquée à  $f$  fournit une méthode approchée de résolution de (\*) connue sous le nom de méthode de Gauss-Seidel. Notons  $A = (a_{ij})_{i=1..n, j=1..n}$ , elle devient dans ce cas (voir l'exercice 3 du chapitre IV) :

$$(*) \quad \Longleftrightarrow \quad \begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ \vdots & \vdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n &= b_n \end{cases}$$

Choisir arbitrairement un point initial  $\mathbf{u}_0$  et construire le point  $\mathbf{u}_{k+1} = (x_1^{k+1}, \dots, x_n^{k+1})$  à partir du point  $\mathbf{u}_k = (x_1^k, \dots, x_n^k)$  de la façon suivante :

$$\begin{cases} \boxed{a_{11}x_1^{k+1}} + a_{12}x_2^k + \cdots + a_{1n}x_n^k &= b_1 &\Longrightarrow x_1^{k+1} \\ a_{11}x_1^{k+1} + \boxed{a_{12}x_2^{k+1}} + \cdots + a_{1n}x_n^k &= b_2 &\Longrightarrow x_2^{k+1} \\ \vdots & \vdots & \vdots \\ a_{11}x_1^{k+1} + a_{12}x_2^{k+1} + \cdots + \boxed{a_{1n}x_n^{k+1}} &= b_n &\Longrightarrow x_n^{k+1} \end{cases}$$

**Remarque.** Clairement, pour appliquer la méthode, les coefficients diagonaux de  $A$  doivent être tous non nuls. C'est bien le cas puisque  $A$  est définie positive (cf. théorème A.4).

En collectant tous les résultats établis dans le chapitre précédent concernant la convergence de chacune de ces méthodes, on peut énoncer ici :

**Théorème V.1 (Convergence des méthodes de résolution approchée.)** *Lorsque  $A$  est une matrice symétrique définie positive, chacune des suites  $(\mathbf{u}_k)_{k \in \mathbb{N}}$  construites selon les méthodes du gradient à pas variable, à pas fixe, du gradient conjugué ou de Gauss-Seidel, convergent vers la solution  $\tilde{\mathbf{u}}$  de (\*).*

### V.1.3 Inversion d'une matrice symétrique définie positive

Soit  $A \in \mathcal{M}_n(\mathbb{R})$  une matrice symétrique définie positive, et donc en particulier inversible. Inverser la matrice  $A$  permet bien évidemment une résolution directe du système (\*) vu ci-dessus, de sorte que ce que nous allons voir présente encore une méthode de résolution de systèmes linéaires. Cependant l'inversion d'une matrice intervient très fréquemment dans de nombreux problèmes (la méthode de Newton par exemple) et présente un intérêt qui ne se réduit pas seulement à la résolution de systèmes linéaires.

L'inversion d'une matrice est une opération coûteuse numériquement, pouvant naïvement se ramener à la résolution d'un système de Cramer de  $n$  équations qui se résout en  $O(n^3)$ . C'est cependant un problème incontournable dans bon nombre de problèmes de mathématiques appliquées. Aussi plusieurs méthodes ont-elles été développées pour l'effectuer au mieux, citons par exemple la méthode du pivot de Gauss, ou la factorisation  $LU$ , qui toutes deux se ramènent à l'inversion de matrices triangulaires. Nous allons développer une technique d'inversion basée sur la méthode du gradient conjugué ; ce que nous voyons ici ne s'adapte qu'à une matrice symétrique définie positive. Il s'agit d'une méthode exacte, pour peu que l'on oublie les erreurs d'arrondi.

**Définition.** Soit  $A \in \mathcal{M}_n(\mathbb{R})$  une matrice symétrique définie positive, et  $\omega_1, \dots, \omega_p$  une famille de  $p$  vecteurs de  $\mathbb{R}^n$ . La famille est dite  $A$ -orthogonale si pour tout  $1 \leq i, j \leq p$ ,  $i \neq j$ ,  $\omega_i^\top A \omega_j = 0$ .

Clairement, puisque lorsque  $A$  est symétrique définie positive, la forme bilinéaire  $(\mathbf{x}, \mathbf{y}) \mapsto \mathbf{x}^\top A \mathbf{y}$  définit un produit scalaire, une famille de vecteurs non nuls  $A$ -orthogonale est une famille libre.

Soit  $\omega_1, \dots, \omega_p$  une famille de  $p$  vecteurs non nuls de  $\mathbb{R}^n$   $A$ -orthogonale ( $p \leq n$ ). On construit une suite finie  $C_1, \dots, C_p$  de matrices dans  $\mathcal{M}_n(\mathbb{R})$  de la façon suivante :

$$C_k = \sum_{i=1}^k \frac{\omega_i \omega_i^\top}{\omega_i^\top A \omega_i}, \quad k = 1, \dots, p.$$

**Théorème V.2 (Calcul itératif de l'inverse de la matrice  $A$ .)** Si  $A \in \mathcal{M}_n(\mathbb{R})$  est symétrique définie positive et si  $\omega_1, \dots, \omega_n$  est une famille  $A$ -orthogonale de vecteurs non nuls de  $\mathbb{R}^n$ , alors :

$$C_n = A^{-1}.$$

**Démonstration.** Par construction, pour  $j = 1, \dots, k$ ,

$$C_k A \omega_j = \sum_{i=1}^k \frac{\omega_i \omega_i^\top A \omega_j}{\omega_i^\top A \omega_i} = \frac{\omega_j \omega_j^\top A \omega_j}{\omega_j^\top A \omega_j} = \omega_j.$$

Posons  $D_k = \text{Id} - C_k A$  ; avec ce qui précède, pour  $j = 1, \dots, k$ ,

$$D_k \omega_j = \omega_j - C_k A \omega_j = \omega_j - \omega_j = \mathbf{0}.$$

En particulier  $D_n \omega_j = \mathbf{0}$  pour tout  $j = 1, \dots, n$ . Or  $\omega_1, \dots, \omega_n$  est une famille libre et donc une base de  $\mathbb{R}^n$ . Ainsi,  $D_n$  est la matrice nulle  $\mathbf{0}$ . Donc  $D_n = \text{Id} - C_n A = \mathbf{0} \implies C_n = A^{-1}$ .  $\square$

Pour appliquer cette méthode il suffit donc de construire une famille  $A$ -orthogonale de  $n$  vecteurs non nuls. On peut appliquer la méthode du gradient conjugué à la forme quadratique  $f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^\top A \mathbf{x}$  qui permet de construire une famille de  $p$  vecteurs non nuls  $A$ -orthogonale. Si  $p < n$  on complète cette famille (que l'on peut aussi construire directement) par :

**Théorème V.3 (Construction d'une famille  $A$ -orthogonale de  $n$  vecteurs.)** Soit  $\omega_1, \dots, \omega_p$  une famille de  $p < n$  vecteurs non nuls  $A$ -orthogonale. Posons pour  $k = 1, \dots, p$ ,

$$D_k = \text{Id} - C_k A .$$

Si  $D_p$  est la matrice nulle alors,  $C_p = A^{-1}$ , et sinon soit  $\mathbf{u} \notin \ker D_p$  et  $\omega_{p+1} = D_p \mathbf{u}$ . Alors  $\omega_1, \dots, \omega_{p+1}$  est une famille  $A$ -orthogonale de  $p+1$  vecteurs non nuls.

**Démonstration.** Il est clair par construction que si  $D_p$  est la matrice nulle alors  $C_p = A^{-1}$ . Supposons que ce n'est pas le cas et soit  $\mathbf{u} \notin \ker D_p$ . Pour  $j = 1, \dots, p$

$$\begin{aligned} (D_p \mathbf{u})^\top A \omega_j &= \mathbf{u}^\top D_p^\top A \omega_j = \mathbf{u}^\top (\text{Id} - A C_p) A \omega_j \\ &= \mathbf{u}^\top A \omega_j - \mathbf{u}^\top A (C_p A \omega_j) \end{aligned}$$

or  $C_p A \omega_j = \omega_j$ , voir la preuve du précédent théorème,

$$= \mathbf{u}^\top A \omega_j - \mathbf{u}^\top A \omega_j = 0$$

Ainsi  $\omega_1, \dots, \omega_{p+1}$  est une famille  $A$ -orthogonale, non nulle puisque  $\omega_{k+1} = D_p \mathbf{u} \neq \mathbf{0}$ .  $\square$

**Exemple.** Soit la matrice symétrique :

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

qui est définie positive car à trace et déterminant  $> 0$ . Prenons  $\omega_1 = (1, 0)$ , alors :

$$C_1 = \frac{\omega_1 \omega_1^\top}{\omega_1^\top A \omega_1} = \frac{1}{2} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1/2 & 0 \\ 0 & 0 \end{pmatrix} \quad ; \quad D_1 = \text{Id} - C_1 A = \begin{pmatrix} 0 & -1/2 \\ 0 & 1 \end{pmatrix} .$$

Posons  $\mathbf{u} = (0, 1) \notin \ker D_1$  et  $\omega_2 = D_1 \mathbf{u} = (-\frac{1}{2}, 1)$ . Alors :

$$\frac{\omega_2 \omega_2^\top}{\omega_2^\top A \omega_2} = \begin{pmatrix} 1/6 & -1/3 \\ -1/3 & 2/3 \end{pmatrix} \quad ; \quad C_2 = C_1 + \frac{\omega_2 \omega_2^\top}{\omega_2^\top A \omega_2} = \begin{pmatrix} 2/3 & -1/3 \\ -1/3 & 2/3 \end{pmatrix} = A^{-1} .$$

### V.1.4 Résolution approchée d'un système d'équations non linéaires

Soit  $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$  une application de classe  $C^1$ ; résoudre  $F(\mathbf{x}) = \mathbf{0}$  équivaut à résoudre un système de  $n$  équations à  $n$  inconnues :

$$\begin{cases} g_1(x_1, \dots, x_n) &= 0 \\ \vdots & \vdots \\ g_n(x_1, \dots, x_n) &= 0 \end{cases}$$

et les  $g_1, \dots, g_n$  sont des applications de classe  $C^1$ . On peut le résoudre en appliquant la méthode de Newton (voir § IV.1.1) :

$$\mathbf{x}_{n+1} = \mathbf{x}_n - DF(\mathbf{x}_n)^{-1}F(\mathbf{x}_n)$$

Chaque itération revient à résoudre le système d'équations linéaires d'inconnue  $\delta \mathbf{x}_n$  :

$$DF(\mathbf{x}_n)\delta \mathbf{x}_n = F(\mathbf{x}_n)$$

puis à poser :  $\mathbf{x}_{n+1} = \mathbf{x}_n - \delta \mathbf{x}_n$ .

Seulement chaque itération est coûteuse en temps de calcul. Aussi peut-on lui préférer en pratique une **méthode quasi-Newton** qui consiste en chaque itération à remplacer  $DF(\mathbf{x}_n)$  par une matrice  $A_n(\mathbf{x}_n)$  pour laquelle la résolution du système linéaire est moins coûteuse. Il y a beaucoup de telles méthodes, en voici deux :

- Fixer un entier  $k$  et poser  $\forall n = p, p+1, \dots, p+k$ ,  $A_n(\mathbf{x}_n) = DF(\mathbf{x}_p)$  et  $A_{p+k+1}(\mathbf{x}_{p+k+1}) = DF(\mathbf{x}_{p+k+1})$  (c'est-à-dire conserve la matrice  $DF(\mathbf{x}_p)$  sur  $k$  itérations). Lorsque  $k$  est suffisamment petit, cette méthode quasi-Newton converge.
- Poser  $\forall n \in \mathbb{N}$ ,  $A_n(\mathbf{x}_n) = Id$ , i.e.  $\mathbf{x}_{n+1} = \mathbf{x}_n - F(\mathbf{x}_n)$ . C'est la **méthode des approximations successives**.

Pour établir la convergence d'une méthode quasi-Newton on peut utiliser le résultat technique suivant, que nous admettons :

**Théorème V.4 (Convergence des méthodes quasi-Newton.)** Soit  $\mathbf{x}_0 \in \mathbb{R}^n$ ; s'il existe 3 constantes  $r, M, \beta$  telles que :  $r > 0$ ,  $B = B(\mathbf{x}_0, r)$ ,

$$\begin{aligned} \sup_{k \in \mathbb{N}} \sup_{\mathbf{x} \in B} \|A_k^{-1}(\mathbf{x})\|_2 &\leq M \\ \sup_{k \in \mathbb{N}} \sup_{\mathbf{x}, \mathbf{x}' \in B} \|DF(\mathbf{x}) - A_k(\mathbf{x}')\|_2 &\leq \frac{\beta}{M} \\ \|F(\mathbf{x}_0)\| &\leq \frac{r}{M}(1 - \beta) \end{aligned}$$

alors la suite définie par  $\mathbf{x}_{k+1} = \mathbf{x}_k - A_k^{-1}(\mathbf{x}_k)$  converge vers l'unique zéro  $\tilde{\mathbf{x}}$  de  $F$  dans  $B$ , et la convergence est géométrique :

$$\forall k \in \mathbb{N}, \quad \|\mathbf{x}_k - \tilde{\mathbf{x}}\| \leq \frac{\|\mathbf{x}_1 - \mathbf{x}_0\|}{1 - \beta} \beta^k$$

Nous restons succinct et ne développons pas plus loin ces techniques quasi-Newton. Elles pourraient mériter un chapitre à elles seules. Ce qui d'ailleurs ne serait pas cher payer au vu de leur puissance et de leur vaste champ d'applications.

## V.2 Approximation d'un nuage de points

Soit  $p$  un entier strictement positif, et un nuage (=suite finie) de  $p$  points de  $\mathbb{R}^2$  :

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_p, y_p)\} \quad .$$

Soit

$$\{f_{\mathbf{u}} : \mathbb{R} \longrightarrow \mathbb{R} \mid \mathbf{u} \in \mathbb{R}^n\}$$

une famille d'applications dépendant continûment de  $n$  paramètres réels. Soit  $\|\cdot\|_{\mathbb{R}^p}$  une norme de  $\mathbb{R}^p$ , posons :

$$Z(\mathbf{u}) = \begin{pmatrix} y_1 - f_{\mathbf{u}}(x_1) \\ y_2 - f_{\mathbf{u}}(x_2) \\ \vdots \\ y_p - f_{\mathbf{u}}(x_p) \end{pmatrix}$$

et considérons le problème d'optimisation :

$$\min_{\mathbf{u} \in \mathbb{R}^n} \|Z(\mathbf{u})\|_{\mathbb{R}^p}$$

C'est un **problème d'approximation** des  $p$  points  $(x_1, y_1), \dots, (x_p, y_p)$  de  $\mathbb{R}^2$  par une application de la classe  $\{f_{\mathbf{u}} \mid \mathbf{u} \in \mathbb{R}^n\}$ .

- Lorsque  $\|Z(\mathbf{u})\|$  a pour valeur minimale 0, il s'agit d'un **problème d'interpolation**. Dans ce cas le résultat est indépendant de la norme  $\|\cdot\|_{\mathbb{R}^p}$  considérée.
- Lorsque  $\|\cdot\|_{\mathbb{R}^p} = \|\cdot\|_2$ , c'est-à-dire,  $\|\mathbf{x}\|_2 = (\sum_{i=1}^n x_i^2)^{\frac{1}{2}}$  il s'agit d'un **problème d'approximation au sens des moindres carrés**.
- Lorsque  $\|\cdot\|_{\mathbb{R}^p} = \|\cdot\|_{\infty}$ , c'est-à-dire,  $\|\mathbf{x}\|_{\infty} = \sup \{|x_1|, \dots, |x_n|\}$ , il s'agit d'un **problème d'approximation au sens de Tchebychev** ou encore d'un **problème d'approximation minimax**.
- Lorsque  $f_{\mathbf{u}}$  dépend linéairement des paramètres  $\mathbf{u} = (u_1, u_2, \dots, u_n)$  on parle d'**approximation linéaire**; c'est-à-dire,  $\forall x \in \mathbb{R}, \mathbf{u} \longrightarrow f_{\mathbf{u}}(x)$  est linéaire. Dans ce cas, on a la notation matricielle :

$$\forall i = 1, \dots, p, \quad f_{\mathbf{u}}(x_i) = \sum_{j=1}^n z_{ij} u_j$$

$$M = \begin{pmatrix} z_{11} & \cdots & z_{1n} \\ \vdots & & \vdots \\ z_{p1} & \cdots & z_{pn} \end{pmatrix} \quad ; \quad \mathbf{u} = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \quad ; \quad \mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix} \quad ,$$

et le problème s'écrit :

$$\min_{\mathbf{u} \in \mathbb{R}^n} \|M\mathbf{u} - \mathbf{y}\|$$



### V.2.1 Approximation linéaire au sens des moindres carrés

Soit  $M \in \mathcal{M}_{p,n}(\mathbb{R})$  et  $\mathbf{y} \in \mathbb{R}^p$  donnés. On s'intéresse à :

$$\min_{\mathbf{u} \in \mathbb{R}^n} \|M\mathbf{u} - \mathbf{y}\|_2 = \min_{\mathbf{u} \in \mathbb{R}^n} (\|M\mathbf{u} - \mathbf{y}\|_2)^2 = \min_{\mathbf{u} \in \mathbb{R}^n} \sum_{i=1}^p ((M\mathbf{u})_i - \mathbf{y}_i)^2 . \quad (*)$$

• **Existence d'une solution.**

$Im(M)$  est un sous-espace vectoriel de  $\mathbb{R}^p$  ; c'est donc un fermé convexe non vide. Avec le théorème de projection convexe (cf. théorème IV.8) :

$$\exists ! \tilde{\mathbf{u}} = M\mathbf{u} \text{ tel que } \|\tilde{\mathbf{u}} - \mathbf{y}\| = \min_{\mathbf{v} \in Im(M)} \|\mathbf{v} - \mathbf{y}\| .$$

et un élément  $\mathbf{u}^*$  tel que  $\tilde{\mathbf{u}} = M\mathbf{u}^*$  est une solution du problème d'approximation.

• **Si  $M$  est inversible ( $rg M = n = p$ ).**

Le problème (\*) admet une unique solution  $\mathbf{u}^* = M^{-1}\tilde{\mathbf{u}}$ , où  $\tilde{\mathbf{u}}$  est le projeté de  $\mathbf{y}$  sur  $Im M$ .

• **Comment déterminer la solution ?**

Posons :

$$\begin{aligned} J(\mathbf{u}) &= \frac{1}{2} (\|M\mathbf{u} - \mathbf{y}\|_2)^2 - \frac{1}{2} (\|\mathbf{y}\|)^2 \\ &= \frac{1}{2} \langle M\mathbf{u} - \mathbf{y}, M\mathbf{u} - \mathbf{y} \rangle - \frac{1}{2} \langle \mathbf{y}, \mathbf{y} \rangle \\ &= \frac{1}{2} \langle M\mathbf{u}, M\mathbf{u} \rangle - \langle \mathbf{y}, M\mathbf{u} \rangle \\ J(\mathbf{u}) &= \frac{1}{2} \langle M^\top M\mathbf{u}, \mathbf{u} \rangle - \langle M^\top \mathbf{y}, \mathbf{u} \rangle \end{aligned}$$

$J : \mathbb{R}^n \longrightarrow \mathbb{R}$  est une fonction quadratique de matrice Hessienne  $M^\top M$ , et :

$$\min_{\mathbf{u} \in \mathbb{R}^n} J(\mathbf{u}) = \min_{\mathbf{u} \in \mathbb{R}^n} \|M\mathbf{u} - \mathbf{y}\|_2$$

Or  $\forall M \in \mathcal{M}_{p,n}(\mathbb{R})$ , la matrice carrée  $M^\top M \in \mathcal{M}_{n,n}(\mathbb{R})$  est semi-définie positive (cf. théorème A.5 page 117) et même définie positive lorsque  $M$  est de rang maximal. Ainsi un minimum  $\mathbf{u}^*$  de (\*) est caractérisé comme solution du système linéaire :

$$M^\top M\mathbf{u} = M^\top \mathbf{y}$$

d'inconnue  $\mathbf{u} \in \mathbb{R}^p$ .

Si en outre  $p = n$  et  $M$  est inversible, alors  $M^\top M$  est définie positive. Dans ce cas (cf. théorème II.10) il existe une unique solution  $\mathbf{u}^* \in \mathbb{R}^p$  de (\*) caractérisée par le système de Cramer :

$$M\mathbf{u} = \mathbf{y} .$$

Il s'agit alors d'un problème d'interpolation linéaire.

On peut résumer tous ces faits dans le théorème suivant :

**Théorème V.5 (Approximation linéaire au sens des moindres carrés.)**

Un problème d'approximation linéaire d'un nuage de points au sens des moindres carrés admet toujours une solution. En notant  $M \in \mathcal{M}_{p,n}(\mathbb{R})$  la matrice associée, une solution est caractérisée par le système linéaire d'inconnue  $\mathbf{u}$  :

$$M^\top M \mathbf{u} = M^\top \mathbf{y}$$

De plus la solution est unique si et seulement si  $M$  est de rang maximal. Lorsque  $M$  est inversible la solution est aussi caractérisée par le système de Cramer  $M \mathbf{u} = \mathbf{y}$ , et il s'agit alors d'une interpolation (la valeur minimale est 0).

**V.2.2 Exemple important : la droite de régression linéaire**

On cherche la droite  $y = ax + b$  qui approche le mieux le nuage de points  $(x_1, y_1), \dots, (x_p, y_p)$  de  $\mathbb{R}^2$  au sens des moindres carrés. Soit  $f_{a,b}(x) = ax + b$ ; on cherche :

$$\min_{a,b \in \mathbb{R}} \sum_{i=1}^p (y_i - ax_i - b)^2$$

Avec les notations précédentes, posons  $u = (a, b)$  et :

$$M = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_p & 1 \end{pmatrix} \quad M^\top M = \begin{pmatrix} \sum_{i=1}^p x_i^2 & \sum_{i=1}^p x_i \\ \sum_{i=1}^p x_i & p \end{pmatrix} \quad M^\top \mathbf{y} = \begin{pmatrix} \sum_{i=1}^p x_i y_i \\ \sum_{i=1}^p y_i \end{pmatrix}$$

Or  $\det(M^\top M) = p \sum_{i=1}^p x_i^2 - (\sum_{i=1}^p x_i)^2 \neq 0$ . Ainsi existe-t-il une unique solution  $(a, b)$ , caractérisée par le système :

$$M^\top M \mathbf{u} = M^\top \mathbf{y} \quad \Longleftrightarrow \quad \begin{cases} a \sum_{i=1}^p x_i^2 + b \sum_{i=1}^p x_i = \sum_{i=1}^p x_i y_i \\ a \sum_{i=1}^p x_i + bp = \sum_{i=1}^p y_i \end{cases}$$

$$\Rightarrow \begin{cases} a = \frac{p \sum_{i=1}^p x_i y_i - \sum_{i=1}^p x_i \sum_{i=1}^p y_i}{p \sum_{i=1}^p x_i^2 - (\sum_{i=1}^p x_i)^2} \\ b = \frac{\sum_{i=1}^p x_i^2 \sum_{i=1}^p y_i - \sum_{i=1}^p x_i \sum_{i=1}^p x_i y_i}{p \sum_{i=1}^p x_i^2 - (\sum_{i=1}^p x_i)^2} \end{cases}$$

### V.2.3 Exemple important : le polynôme d'interpolation de Lagrange

Soit  $\{(x_1, y_1), \dots, (x_p, y_p)\}$  un nuage de points, soit  $\mathbf{u} = (u_0, u_1, \dots, u_{n-1}) \in \mathbb{R}^n$ , et soit l'application  $f_{\mathbf{u}} : \mathbb{R} \rightarrow \mathbb{R}$  polynômiale à coefficients réels de degré au plus  $n-1$ ,  $f_{\mathbf{u}}(x) = u_0 + u_1x + u_2x^2 + \dots + u_{n-1}x^{n-1}$ .

Le problème d'approximation du nuage de points par une application polynomiale de degré au plus  $n-1$  au sens des moindres carrés est un problème d'approximation linéaire, et sa matrice associée est :

$$M = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{n-1} \\ \vdots & & & & \vdots \\ 1 & x_p & x_p^2 & \cdots & x_p^{n-1} \end{pmatrix} \in \mathcal{M}_{p,n}(\mathbb{R})$$

Elle est de rang maximal lorsque  $n \leq p$  (voir plus bas le déterminant d'une matrice carrée de Vandermonde), et donc

$$M^T M = \begin{pmatrix} 1 & \sum_{i=1}^p x_i & \sum_{i=1}^p x_i^2 & \cdots & \sum_{i=1}^p x_i^{n-1} \\ \sum_{i=1}^p x_i & \sum_{i=1}^p x_i^2 & \sum_{i=1}^p x_i^3 & & \vdots \\ \sum_{i=1}^p x_i^2 & \sum_{i=1}^p x_i^3 & \sum_{i=1}^p x_i^4 & & \vdots \\ \vdots & & & \ddots & \vdots \\ \sum_{i=1}^p x_i^{n-1} & \cdots & \cdots & \cdots & \sum_{i=1}^p x_i^{2n-2} \end{pmatrix}$$

est lorsque  $n \leq p$  inversible, et même symétrique définie positive. Il existe donc lorsque  $n \leq p$  une unique solution  $\mathbf{u}^* \in \mathbb{R}^n$ , pour lequel  $f_{\mathbf{u}^*}$  est la meilleure approximation du nuage de points au sens des moindres carrés. Une solution est caractérisée par le système d'inconnue  $\mathbf{u}$ ,  $M^T M \mathbf{u} = M^T \mathbf{y}$ , qui s'écrit ici :

$$\begin{cases} u_0 & + u_1 \sum_{i=1}^p x_i & + \cdots & + u_{n-1} \sum_{i=1}^p x_i^{n-1} & = \sum_{i=1}^p y_i \\ u_0 \sum_{i=1}^p x_i & + u_1 \sum_{i=1}^p x_i^2 & + \cdots & + u_{n-1} \sum_{i=1}^p x_i^n & = \sum_{i=1}^p x_i y_i \\ \vdots & & & & \vdots \\ u_0 \sum_{i=1}^p x_i^{n-1} & + u_1 \sum_{i=1}^p x_i^n & + \cdots & + u_{n-1} \sum_{i=1}^p x_i^{2n-2} & = \sum_{i=1}^p x_i^{n-1} y_i \end{cases}$$

- Lorsque  $n < p$ . La solution optimale  $f_{\mathbf{u}^*}$  est unique et approxime au mieux le nuage de points au sens des moindres carrés. On peut en déterminer une valeur approchée en

implémentant les méthodes de résolution approchées vues au chapitre 4.

- Lorsque  $n = p$ . La matrice  $M$  est connue sous le nom de *matrice de Vandermonde*, et son déterminant est non nul égal à  $\prod_{1 \leq j < i \leq n} (x_i - x_j)$ . La matrice  $M$  est inversible, et il existe donc une unique solution  $\mathbf{u}^*$  et  $f_{\mathbf{u}^*}$  est le polynôme de degré minimal interpolant le nuage de points. C'est le **polynôme d'interpolation de Lagrange**. On peut vérifier qu'il s'écrit explicitement :

$$f_{\mathbf{u}^*}(x) = \sum_{i=1}^p y_i \prod_{j=1, j \neq i}^p \frac{x - x_j}{x_i - x_j}$$

- Lorsque  $n > p$ . L'ensemble des solutions est isomorphe à un sous-espace affine de l'espace vectoriel  $\mathbb{R}_{n-1}[x]$  des polynômes à coefficients réels de degré au plus  $n - 1$ . Une solution particulière est donnée par le polynôme d'interpolation de Lagrange (de degré  $p - 1$ ), et une base du sous-espace vectoriel sous-jacent est donnée par :

$$\left\{ \prod_{i=1}^p (x - x_i); x \prod_{i=1}^p (x - x_i); \dots; x^{n-1-p} \prod_{i=1}^p (x - x_i) \right\}$$

C'est l'ensemble des polynômes de degré au plus  $n - 1 - p$  ayant  $x_1, x_2, \dots, x_p$  pour racines, c'est-à-dire les solutions du problème homogène associé.

#### V.2.4 Approximation minimax

Donnés un nuage de points  $\{(x_1, y_1), (x_2, y_2), \dots, (x_p, y_p)\}$  et une classe d'applications  $f_{\mathbf{u}} : \mathbb{R} \rightarrow \mathbb{R}$  dépendant d'un paramètre  $\mathbf{u} \in \mathbb{R}^n$ , on cherche, une fois posé  $\mathbf{x} = (x_1, \dots, x_p)$  et  $\mathbf{y} = (y_1, \dots, y_p)$  à résoudre le problème d'optimisation :

$$\min_{\mathbf{u} \in \mathbb{R}^n} \|\mathbf{y} - f_{\mathbf{u}}(\mathbf{x})\|_{\infty} = \min_{\mathbf{u} \in \mathbb{R}^n} \left( \max_{k=1..p} |y_k - f_{\mathbf{u}}(x_k)| \right)$$

Il s'écrit aussi comme le problème d'optimisation avec  $2p$  contraintes inégalitaires suivant :

$$\begin{aligned} & \min_{r \in \mathbb{R}, \mathbf{u} \in \mathbb{R}^n} r \\ & \begin{cases} y_k - f_{\mathbf{u}}(x_k) - r \leq 0 \\ -y_k + f_{\mathbf{u}}(x_k) - r \leq 0 \\ k = 1, 2, \dots, p \end{cases} \end{aligned}$$

(implicitement  $r \geq 0$  puisque  $|y_k - f_{\mathbf{u}}(x_k)| \leq r$ .)

#### V.2.5 Approximation minimax linéaire

Lorsque  $f_{\mathbf{u}}$  dépend linéairement de  $\mathbf{u}$ , c'est à dire lorsque  $\forall x \in \mathbb{R}, \mathbf{u} \rightarrow f_{\mathbf{u}}(x)$  est linéaire :  $f_{\mathbf{u}}(x_i) = \sum_{j=1}^p z_{ij} u_j$ , notons  $M = (z_{ij})_{\substack{i=1..p \\ j=1..n}} \in \mathcal{M}_{pn}(\mathbb{R})$ , le problème s'écrit matriciellement :

$$\min_{\mathbf{u} \in \mathbb{R}^n} \|M\mathbf{u} - \mathbf{y}\|_{\infty}$$

qui est équivalent à :

$$\min_{r \in \mathbb{R}, \mathbf{u} \in \mathbb{R}^n} r$$

$$\begin{cases} y_k - \sum_{i=1}^p z_{ik} u_i - r \leq 0 \\ -y_k + \sum_{i=1}^p z_{ik} u_i - r \leq 0 \\ k = 1, 2, \dots, p \end{cases}$$

C'est un problème de programmation linéaire. Puisque  $r$  est minoré,  $r \geq 0$ , il existe une solution au problème que l'on détermine avec la méthode du simplexe. On a donc montré :

**Théorème V.6** *Un problème d'approximation minimax linéaire s'exprime comme un problème de programmation linéaire et admet toujours au moins une solution.*

### Interprétation d'un problème minimax linéaire.

**Interprétation algébrique.** Notons pour  $i = 1, 2, \dots, n$ ,  $\mathbf{z}_i = (z_{1i}, z_{2i}, \dots, z_{pi}) \in \mathbb{R}^p$ , c'est-à-dire les  $n$  vecteurs colonnes de la matrice  $M$ .

- Si la famille  $\mathbf{z}_1, \dots, \mathbf{z}_n$  n'est pas linéairement indépendante alors un des paramètres  $u_1, \dots, u_n$  peut être supprimé : en effet, si par exemple  $\mathbf{z}_1 = \sum_{i=2}^n \lambda_i \mathbf{z}_i$  alors remplacer dans les équations  $u_1$  par  $\sum_{i=2}^n \lambda_i u_i$ . On réduit le problème à un problème minimax linéaire équivalent et de dimension inférieure.

Aussi suppose-t-on dans la suite que les  $\mathbf{z}_1, \dots, \mathbf{z}_n$  forment une famille libre.

- Lorsque  $p \leq n$  la valeur minimale de  $\|\mathbf{M}\mathbf{u} - \mathbf{y}\|$  est 0 ; le problème consiste en la résolution du système linéaire  $\mathbf{M}\mathbf{u} = \mathbf{y}$  d'inconnue  $\mathbf{u}$  ; appliquer ici la méthode du simplexe au problème linéaire équivalent consiste en fait à le résoudre par une variante de la méthode du pivot de Gauss ; cela ne présente pas un grand intérêt.

- Par contre lorsque  $n < p$  et que le système linéaire  $\mathbf{M}\mathbf{u} = \mathbf{y}$  n'admet pas de solution : la solution  $\tilde{\mathbf{u}}$  au problème minimax est optimale dans le sens où c'est l'élément le plus proche (pour la norme  $\|\cdot\|_\infty$ ) d'être une solution.

**Interprétation géométrique.** Notons pour  $i = 1, 2, \dots, p$ ,  $\bar{\mathbf{z}}_i = (z_{i1} \ z_{i2} \ \dots \ z_{in}) \in \mathcal{M}_{1,n}(\mathbb{R})$ , c'est-à-dire les  $p$  matrices lignes de la matrice  $M$ . On considère les  $p$  hyperplans,  $\mathcal{H}_1, \dots, \mathcal{H}_p$ , définis par  $\mathcal{H}_i = \{\mathbf{u} \in \mathbb{R}^n \mid \bar{\mathbf{z}}_i \mathbf{u} = y_i\}$ .

- Une solution  $\tilde{\mathbf{u}}$  du problème minimax est un point de  $\mathbb{R}^n$  dont la distance maximale à la famille d'hyperplans  $\mathcal{H}_1, \dots, \mathcal{H}_p$  est minimale.

- Exemple en dimension 2 : les  $p$  hyperplans sont des droites.

- Si  $p = 1$  une solution  $\tilde{\mathbf{u}}$  est n'importe quel point de la droite  $\mathcal{H}_1$ .

- Si  $p = 2$  ; si les 2 droites  $\mathcal{H}_1, \mathcal{H}_2$  sont non parallèles, l'unique solution  $\tilde{\mathbf{u}}$  est leur point d'intersection ; c'est la solution d'un système de 2 équations linéaires. Si  $\mathcal{H}_1, \mathcal{H}_2$  sont parallèles ; soit elles sont confondues, et dans ce cas tout point de  $\mathcal{H}_1 = \mathcal{H}_2$  est solution ; soit

elles sont disjointes, et l'ensemble des solutions est une droite parallèle et équidistante à  $\mathcal{H}_1, \mathcal{H}_2$ .

– Si  $p = 3$ ; si les 3 droites sont deux à deux non parallèles elles découpent un triangle et la solution du problème minimax est le centre du cercle inscrit à ce triangle (cf. fig. V.2; si  $\mathcal{H}_1, \mathcal{H}_2$  sont parallèles et  $\mathcal{H}_3$  ne leur est pas parallèle, la solution est le point de  $\mathcal{H}_3$  équidistant de  $\mathcal{H}_1, \mathcal{H}_2$ ; etc...

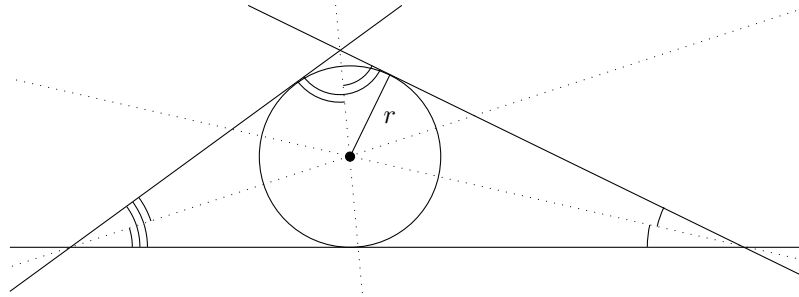


FIGURE V.2 – Le centre du cercle inscrit est la solution d'un problème minimax d'approximation linéaire à deux paramètres d'un nuage de 3 points.

## Exercices.

**Exercice 1.** Justifier que la matrice symétrique  $A$  ci-dessous est définie positive.

$$\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$$

Déterminer son inverse.

**Exercice 2.** Déterminer l'espace des polynômes  $P$  de degré au plus 5 interpolant les points  $(0, 0)$ ,  $(1, 1)$  et  $(2, 2)$ .

## Annexe A

# Rappels de pré-requis Mathématiques

### A.1 Rappels d'analyse

#### A.1.1 L'espace euclidien $\mathbb{R}^n$

Soit  $\mathbb{R}^n$  L'espace vectoriel réel de dimension  $n \in \mathbb{N}_*$ . On notera sa base canonique  $\mathbf{e}_1, \dots, \mathbf{e}_n$ . On munit  $\mathbb{R}^n$  du *produit scalaire usuel*  $\langle \cdot, \cdot \rangle$  et de la *norme associée*  $\|\cdot\|_2$  (ou  $\|\cdot\|$  lorsqu'il n'y a pas d'ambiguïté). C'est à dire que si dans la base canonique de  $\mathbb{R}^n$  les vecteurs  $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$  s'écrivent  $\mathbf{u} = (u_1, u_2, \dots, u_n)$  et  $\mathbf{v} = (v_1, v_2, \dots, v_n)$ , alors :

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v} = u_1 v_1 + u_2 v_2 + \dots + u_n v_n$$
$$\|\mathbf{u}\| \triangleq \langle \mathbf{u}, \mathbf{u} \rangle^{\frac{1}{2}} = \sqrt{u_1^2 + u_2^2 + \dots + u_n^2}$$

On parle alors de l'*espace euclidien*  $(\mathbb{R}^n, \langle \cdot, \cdot \rangle)$  de dimension  $n$ . On y vérifie l'*inégalité de Cauchy-Schwartz* :

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\| \quad \text{soit} \quad \left| \sum_{i=1}^n x_i y_i \right| \leq \left( \sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}} \left( \sum_{i=1}^n y_i^2 \right)^{\frac{1}{2}}.$$

#### A.1.2 Normes de $\mathbb{R}^n$

Une norme  $\|\cdot\|$  sur  $\mathbb{R}^n$  est une application de  $\mathbb{R}^n$  dans  $\mathbb{R}_+$  vérifiant :  $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \forall \lambda \in \mathbb{R}$  :

$$\begin{aligned} (\text{séparation}) \quad & \|\mathbf{x}\| = 0 \implies \mathbf{x} = \mathbf{0}, \\ (\text{homogénéité}) \quad & \|\lambda \mathbf{x}\| = |\lambda| \|\mathbf{x}\|, \\ (\text{sous-additivité}) \quad & \|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|. \end{aligned}$$

Lorsque l'on munit  $\mathbb{R}^n$  d'une norme  $\|\cdot\|$ , on parle de l'*espace normé*  $(\mathbb{R}^n, \|\cdot\|)$ .

Voici plusieurs exemples de normes sur  $\mathbb{R}^n$  :

- La norme 1 :  $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$ ,
- La norme 2 :  $\|\mathbf{x}\|_2 = \left( \sum_{i=1}^n x_i^2 \right)^{\frac{1}{2}}$ ,

- La norme  $p$  :  $\|\mathbf{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{\frac{1}{p}}$ ,
- La norme sup :  $\|\mathbf{x}\|_\infty = \max \{|x_1|, \dots, |x_n|\}$ .

Sur  $\mathbb{R}^n$  toutes les normes sont équivalentes, c'est à dire si  $\|\cdot\|_a$  et  $\|\cdot\|_b$  désignent deux normes de  $\mathbb{R}^n$ , il existe  $c_1, c_2 > 0$  tels que  $\forall \mathbf{x} \in \mathbb{R}^n$  :

$$\|\mathbf{x}\|_a \leq c_1 \|\mathbf{x}\|_b \quad \text{et} \quad \|\mathbf{x}\|_b \leq c_2 \|\mathbf{x}\|_a .$$

L'espace normé  $(\mathbb{R}^n, \|\cdot\|)$  est un espace *complet* : toute suite de Cauchy y est convergente.

### A.1.3 Topologie de $\mathbb{R}^n$

Soit  $\mathbf{u} \in \mathbb{R}^n$  et  $r > 0$ . Une *boule ouverte* de l'espace normé  $(\mathbb{R}^n, \|\cdot\|)$  centrée en  $\mathbf{u}$  et de rayon  $r$  est :

$$B(\mathbf{u}, r) \triangleq \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{u} - \mathbf{x}\| < r\}$$

On munit  $\mathbb{R}^n$  d'une *topologie* naturelle : un sous-ensemble  $\mathcal{U} \subset \mathbb{R}^n$  est un *ouvert* de  $\mathbb{R}^n$  si pour tout  $\mathbf{u} \in \mathcal{U}$ ,  $\mathcal{U}$  contient une boule ouverte centrée en  $\mathbf{u}$ . C'est la topologie engendrée par les boules ouvertes. Elle ne dépend pas de la norme considérée.

#### Propriétés :

- $\emptyset$  et  $\mathbb{R}^n$  sont des ouverts de  $\mathbb{R}^n$ ,
- une réunion d'ouverts est un ouvert,
- une intersection finie d'ouverts est un ouvert.
- Un sous-ensemble  $\mathcal{E}$  de  $\mathbb{R}^n$  contient un unique ouvert maximal pour l'inclusion ; on le note  $\text{int}(\mathcal{E})$  et on l'appelle l'*intérieur* de  $\mathcal{E}$ .

Un sous-ensemble  $\mathcal{V}$  de  $\mathbb{R}^n$  est un *fermé* pour cette topologie si son complémentaire est un ouvert. Toute boule fermée  $\overline{B(\mathbf{u}, r)} = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{u} - \mathbf{x}\| \leq r\}$  est un fermé de  $\mathbb{R}^n$ .

#### Propriétés :

- $\emptyset$  et  $\mathbb{R}^n$  sont des fermés,
- une intersection de fermés est un fermé,
- une réunion finie de fermés est un fermés.

Dans cette topologie,  $\emptyset$  et  $\mathbb{R}^n$  sont les seuls sous-ensembles de  $\mathbb{R}^n$  à la fois ouverts et fermés : on dit que  $\mathbb{R}^n$  est *connexe*.

Une application de  $\mathbb{R}^n$  dans  $\mathbb{R}^m$  est *continue* si pour tout ouvert (resp. fermé)  $\mathcal{U}$  de  $(\mathbb{R}^m, \|\cdot\|)$ ,  $f^{-1}(\mathcal{U}) \triangleq \{\mathbf{x} \in \mathbb{R}^n, f(\mathbf{x}) \in \mathcal{U}\}$  est un ouvert (resp. fermé) de  $(\mathbb{R}^n, \|\cdot\|)$ .

Un sous-ensemble  $\mathcal{K}$  de  $\mathbb{R}^n$  est un *compact* si il est fermé et borné (*i.e.*  $\exists C > 0, \forall \mathbf{x} \in \mathcal{K}, \|\mathbf{x}\| \leq C$ ).

Si  $f$  est une application continue de  $\mathbb{R}^n$  dans  $\mathbb{R}^m$  et si  $\mathcal{K}$  est un compact de  $(\mathbb{R}^n, \|\cdot\|)$ , alors  $f(\mathcal{K})$  est un compact de  $(\mathbb{R}^m, \|\cdot\|)$ .



## A.2 Rappels de calcul différentiel

### A.2.1 Applications différentiables

Soient  $\mathcal{U}$  un ouvert non-vidé de  $\mathbb{R}^n$ ,  $\mathbf{x}_0 \in \mathcal{U}$ , et  $f : \mathcal{U} \rightarrow \mathbb{R}^p$ . L'application  $f$  est *différentiable en  $\mathbf{x}_0$*  si il existe une application linéaire  $Df(\mathbf{x}_0) : \mathbb{R}^n \rightarrow \mathbb{R}^p$  (la *différentielle de  $f$  en  $\mathbf{x}_0$* ), tel que pour tout  $\mathbf{x} \in \mathcal{U}$ ,  $f(\mathbf{x}) = f(\mathbf{x}_0) + Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + o(\|\mathbf{x} - \mathbf{x}_0\|)$ .

Ce qu'on peut aussi écrire :

$$\forall \varepsilon > 0, \exists r > 0, \text{ tel que } \forall \mathbf{x} \in \mathcal{U}, \|\mathbf{x} - \mathbf{x}_0\| < r \implies \|f(\mathbf{x}) - f(\mathbf{x}_0) - Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)\| < \varepsilon.$$

La notation de Landau :  $o(\|\mathbf{x} - \mathbf{x}_0\|^p)$ , ( $p \in \mathbb{N}$ ), signifie  $\|\mathbf{x} - \mathbf{x}_0\|^p \theta(\mathbf{x})$  où  $\lim_{\mathbf{x} \rightarrow \mathbf{x}_0} \theta(\mathbf{x}) = 0$ .

### A.2.2 Vecteur gradient

Dans ce qui suit on considère le cas particulier d'une application  $f : \mathcal{U} \subset \mathbb{R}^n \rightarrow \mathbb{R}$ , c'est à dire à valeur réelle.

Lorsque  $f$  est différentiable en  $\mathbf{x}_0$ , les dérivées partielles de  $f$  en  $\mathbf{x}_0$  existent. Soit :

$$\nabla f(\mathbf{x}_0) \triangleq \begin{pmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}_0) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}_0) \end{pmatrix} \in \mathbb{R}^n$$

C'est le vecteur gradient de  $f$  en  $\mathbf{x}_0$  (on prononce "nabla f de  $\mathbf{x}_0$ "). On a alors :

$$Df(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) = \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle$$

Lorsque  $f : \mathcal{U} \rightarrow \mathbb{R}$  est différentiable sur  $\mathcal{U}$  (i.e. en tout point de  $\mathcal{U}$ ), on définit sur  $\mathcal{U}$  l'application gradient :

$$\begin{aligned} \nabla f : \mathcal{U} &\rightarrow \mathbb{R}^n \\ \mathbf{x} &\rightarrow \nabla f(\mathbf{x}) \end{aligned}$$

(Remarque : lorsque  $f : \mathcal{U} \subset \mathbb{R} \rightarrow \mathbb{R}$ ,  $\nabla f$  n'est rien d'autre que l'application dérivée  $f'$ .)

Une application  $f : \mathcal{U} \subset \mathbb{R}^n \rightarrow \mathbb{R}$  est de classe  $C^1$  lorsqu'elle est différentiable sur  $\mathcal{U}$  et que  $\nabla f : \mathcal{U} \rightarrow \mathbb{R}^n$  est continue.

### A.2.3 Matrice hessienne

L'application  $f : \mathcal{U} \subset \mathbb{R}^n \rightarrow \mathbb{R}$  est *2 fois différentiable en  $\mathbf{x}_0 \in \mathcal{U}$* , si  $f$  est différentiable sur un ouvert  $\mathcal{V}$  contenant  $\mathbf{x}_0$ , et si  $\nabla f : \mathcal{V} \rightarrow \mathbb{R}^n$  est différentiable en  $\mathbf{x}_0$ . Dans ce cas les dérivées partielles secondes de  $f$  en  $\mathbf{x}_0$  existent et de plus on a  $\forall i, j, \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}_0) = \frac{\partial^2 f}{\partial x_j \partial x_i}(\mathbf{x}_0)$  (*formule de Schwartz*). On note :

$$\nabla^2 f(\mathbf{x}_0) \triangleq \left( \frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}_0) \right)_{\substack{i=1,2,\dots,n \\ j=1,2,\dots,n}} = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1}(\mathbf{x}_0) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(\mathbf{x}_0) \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(\mathbf{x}_0) & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_n}(\mathbf{x}_0) \end{pmatrix}_{n \times n}$$

la matrice Hessienne de  $f$  en  $\mathbf{x}_0$ . C'est une matrice symétrique de  $\mathcal{M}_n(\mathbb{R})$ .

*Remarque.* Puisque  $\nabla^2 f(\mathbf{x}_0)$  est symétrique :

$$\mathbf{x}^\top \nabla^2 f(\mathbf{x}_0) \mathbf{x} = \langle \nabla^2 f(\mathbf{x}_0)^\top \mathbf{x}, \mathbf{x} \rangle = \langle \nabla^2 f(\mathbf{x}_0) \mathbf{x}, \mathbf{x} \rangle .$$

### A.2.4 Développements de Taylor

#### Formule de Taylor-Young (à l'ordre 1 et 2)

Lorsque  $f : \mathcal{U} \subset \mathbb{R}^n \longrightarrow \mathbb{R}$  est différentiable en  $\mathbf{x}_0$  on a le *développement de Taylor-Young* de  $f$  à l'ordre 1 au voisinage de  $\mathbf{x}_0$  :

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle + o(\|\mathbf{x} - \mathbf{x}_0\|)$$

(Cette condition est équivalente à la définition de la différentiabilité de  $f$  en  $\mathbf{x}_0$ .)

Lorsque  $f : \mathcal{U} \subset \mathbb{R}^n \longrightarrow \mathbb{R}$  est 2 fois différentiable en  $\mathbf{x}_0$ , on a le développement de Taylor-Young à l'ordre 2 au voisinage de  $\mathbf{x}_0$  :

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} - \mathbf{x}_0 \rangle + \frac{1}{2} (\mathbf{x} - \mathbf{x}_0)^\top \nabla^2 f(\mathbf{x}_0) (\mathbf{x} - \mathbf{x}_0) + o(\|\mathbf{x} - \mathbf{x}_0\|^2).$$

Ces deux formules de Taylor-Young à l'ordre 1 et 2 sont fondamentales pour établir des conditions nécessaires, suffisantes à l'existence d'extrema locaux.

Les formules de Taylor-MacLaurin et de Taylor avec reste intégral qui suivent donnent plus de précision sur le reste. Elles nous sont bien moins essentielles, n'apparaissant que sporadiquement dans certaines preuves du § 2.3.

#### Formule de Taylor-MacLaurin (à l'ordre 2)

Lorsque  $f : \mathcal{U} \subset \mathbb{R}^n \longrightarrow \mathbb{R}$  est deux fois différentiable sur  $\mathcal{U}$ ,  $\exists \theta \in ]0, 1[$  tel que :

$$f(\mathbf{x}_0 + \mathbf{x}) = f(\mathbf{x}_0) + \langle \nabla f(\mathbf{x}_0), \mathbf{x} \rangle + \frac{1}{2} \mathbf{x}^\top \nabla^2 f(\mathbf{x}_0 + \theta \mathbf{x}) \mathbf{x} .$$

#### Formule de Taylor avec reste intégral (à l'ordre 1)

Lorsque  $f : \mathcal{U} \subset \mathbb{R}^n \longrightarrow \mathbb{R}$  est de classe  $C^1$ ,  $\exists \theta \in ]0, 1[$  tel que :

$$f(\mathbf{x}_0 + \mathbf{x}) = f(\mathbf{x}_0) + \int_0^1 (1-t) \langle \nabla f(\mathbf{x}_0 + t\mathbf{x}), \mathbf{x} \rangle dt .$$

### A.2.5 Espace tangent

Soit  $f : \mathbb{R}^n \longrightarrow \mathbb{R}$  une application. La nappe représentative de  $f$  ou *graphe* de  $f$  est définie comme le sous-ensemble de  $\mathbb{R}^{n+1}$  :

$$C_f \triangleq \left\{ (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R} \mid \mathbf{y} = f(\mathbf{x}) \right\} .$$

Lorsque  $f$  est différentiable sur un ouvert  $\mathcal{U}$  de  $\mathbb{R}^n$ ,  $C_f$  admet en chaque point  $(\mathbf{u}, f(\mathbf{u}))$  où  $\mathbf{u} \in \mathcal{U}$  un *espace tangent*, noté  $\mathcal{T}_{\mathbf{u}} C_f$  et donné par :

$$\mathcal{T}_{\mathbf{u}} C_f \triangleq \left\{ (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R} \mid \mathbf{y} = \langle \nabla f(\mathbf{u}), \mathbf{x} \rangle \right\} .$$

C'est un sous-espace vectoriel de  $\mathbb{R}^{n+1}$  de dimension  $n$ .

L'hyperplan tangent à  $C_f$  en  $\mathbf{u}$  a pour équation  $\mathbf{y} = f(\mathbf{u}) + \langle \nabla f(\mathbf{u}), \mathbf{x} - \mathbf{u} \rangle$ ; c'est un espace affine dont le sous-espace vectoriel sous-jacent  $\mathbf{y} = \langle \nabla f(\mathbf{u}), \mathbf{x} \rangle$  n'est autre que l'espace tangent  $T_{\mathbf{u}}C_f$ .

Plus généralement soit  $f : \mathbb{R}^n \rightarrow \mathbb{R}^p$  une application différentiable sur un ouvert  $\mathcal{U}$  de  $\mathbb{R}^n$ . Le graphe de  $f$  est le sous-ensemble de  $\mathbb{R}^{n+p}$  :

$$C_f \triangleq \left\{ (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^p \mid \mathbf{y} = f(\mathbf{x}) \right\}$$

et  $C_f$  admet en chaque point  $(\mathbf{u}, f(\mathbf{u}))$  où  $\mathbf{u} \in \mathcal{U}$  un espace tangent, noté  $T_{\mathbf{u}}C_f$  donné par :

$$T_{\mathbf{u}}C_f \triangleq \left\{ (\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n \times \mathbb{R}^p \mid \mathbf{y} = Df(\mathbf{u})(\mathbf{x}) \right\}$$

où  $Df(\mathbf{u}) : \mathbb{R}^n \rightarrow \mathbb{R}^p$  est sa différentielle en  $\mathbf{u}$ . C'est un sous-espace vectoriel de  $\mathbb{R}^{n+p}$  de dimension  $n$ .

Lorsque  $\mathcal{D} \subset \mathbb{R}^n$  admet en  $\mathbf{u} \in \mathcal{D}$  un espace tangent  $T_{\mathbf{u}}\mathcal{D}$ , ce dernier est l'ensemble des directions  $\mathbf{d} \in \mathbb{R}^n$  pour lesquelles soit  $\mathbf{d} = \mathbf{0}$  soit il existe une suite  $(\mathbf{u}_k)_{k \in \mathbb{N}}$  dans  $\mathcal{D}$ , non stationnaire, tendant vers  $\mathbf{u}$ , tel que :

$$\mathbf{u}_k = \mathbf{u} + \frac{\|\mathbf{u}_k - \mathbf{u}\|}{\|\mathbf{d}\|} \mathbf{d} + o(\|\mathbf{u}_k - \mathbf{u}\|) .$$

L'intervention de la notion d'espace tangent est essentielle en optimisation sous contrainte lorsqu'appliquée au domaine admissible. Pour le caractériser par une expression explicite nous utilisons le théorème des fonctions implicites (ou plutôt d'un cas particulier de ce théorème).

**Théorème A.1 (Théorème des fonctions implicites.)** Soient  $\mathcal{U}$  un ouvert de  $\mathbb{R}^p \times \mathbb{R}^{n-p}$  et

$$\begin{aligned} \varphi : \mathbb{R}^p \times \mathbb{R}^{n-p} &\longrightarrow \mathbb{R}^p \\ (\mathbf{y}, \mathbf{x}) &\longrightarrow \begin{pmatrix} \varphi_1(\mathbf{y}, \mathbf{x}) \\ \vdots \\ \varphi_p(\mathbf{y}, \mathbf{x}) \end{pmatrix} \end{aligned}$$

une application de classe  $C^1$ . Soient  $\mathbf{v} \in \mathbb{R}^p$  et  $\mathbf{u} \in \mathbb{R}^{n-p}$  tels que  $(\mathbf{v}, \mathbf{u}) \in \mathcal{U}$  et tels que :

$$\varphi(\mathbf{v}, \mathbf{u}) = \mathbf{0} \quad \text{et la matrice} \quad \left( \frac{\partial \varphi_i}{\partial \mathbf{e}_j}(\mathbf{v}, \mathbf{u}) \right)_{\substack{i=1..p \\ j=1..p}} \quad \text{soit inversible.}$$

Alors il existe un ouvert  $\mathcal{U}_1$  de  $\mathbb{R}^p$ , un ouvert  $\mathcal{U}_2$  de  $\mathbb{R}^{n-p}$ , tels que  $(\mathbf{v}, \mathbf{u}) \in \mathcal{U}_1 \times \mathcal{U}_2 \subset \mathcal{U}$ , et une application  $f : \mathcal{U}_2 \rightarrow \mathbb{R}^p$  continue telle que

$$\left\{ (\mathbf{y}, \mathbf{x}) \in \mathcal{U}_1 \times \mathcal{U}_2 \mid \varphi(\mathbf{y}, \mathbf{x}) = \mathbf{0} \right\} = \left\{ (\mathbf{y}, \mathbf{x}) \in \mathbb{R}^p \times \mathcal{U}_2 \mid \mathbf{y} = f(\mathbf{x}) \right\} .$$

De plus  $f$  est différentiable en  $\mathbf{u}$ .

Comme conséquence, le résultat suivant est essentiel en optimisation sous contrainte égalitaire :

**Théorème A.2 (Espace tangent d'un domaine égalitaire.)** Soit  $\mathcal{U}$  un ouvert de  $\mathbb{R}^n$  et  $\varphi_1, \dots, \varphi_p : \mathcal{U} \rightarrow \mathbb{R}$  des applications de classe  $C^1$ . Soit le domaine

$$\mathcal{D} = \left\{ \mathbf{x} \in \mathcal{U} \mid \varphi_1(\mathbf{x}) = \dots = \varphi_p(\mathbf{x}) = 0 \right\} .$$

Si  $\mathbf{u} \in \mathcal{D}$  et si  $\nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u})$  forment une famille libre, alors l'espace tangent  $T_{\mathbf{u}}\mathcal{D}$  en  $\mathbf{u}$  à  $\mathcal{D}$  existe et est donné par :

$$T_{\mathbf{u}}\mathcal{D} = \left\{ \mathbf{d} \in \mathbb{R}^n \mid \forall i = 1, \dots, p, \langle \nabla\varphi_i(\mathbf{u}), \mathbf{d} \rangle = 0 \right\} .$$

**Démonstration.** Notons  $\varphi : \mathcal{U} \rightarrow \mathbb{R}^p$  l'application de classe  $C^1$  qui est définie par  $\varphi(\mathbf{x}) = (\varphi_1(\mathbf{x}), \dots, \varphi_p(\mathbf{x}))$ . Le fait que la famille  $\nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u})$  soit libre revient à dire que la matrice  $\left( \frac{\partial\varphi_i}{\partial\mathbf{e}_j}(\mathbf{u}) \right)_{\substack{i=1..p \\ j=1..n}}$  est de rang  $p$ . Alors quitte à permuter ses colonnes on peut supposer que sa sous-matrice carrée constituée des colonnes 1 à  $p$  est inversible. En notant  $\mathbf{u}_{\mathbf{y}}$  et  $\mathbf{u}_{\mathbf{x}}$  les projetés de  $\mathbf{u}$  sur  $\mathbb{R}^p \times \mathbf{0}$  et sur  $\mathbf{0} \times \mathbb{R}^{n-p}$ , le théorème des fonctions implicites fournit  $f : \mathbb{R}^{n-p} \rightarrow \mathbb{R}^p$  tel que  $\mathbf{u}_{\mathbf{y}} = f(\mathbf{u}_{\mathbf{x}})$ . De plus  $f$  est différentiable en  $\mathbf{u}_{\mathbf{x}}$ . En particulier  $\mathcal{D}$  admet en  $\mathbf{u}$  un espace tangent de dimension  $n - p$ .

Soit  $(\mathbf{u}_k)_{k \in \mathbb{N}}$  une suite de  $\mathcal{D}$  non stationnaire qui tend vers  $\mathbf{u} \in \mathcal{D}$ , avec  $\mathbf{u}_k = \mathbf{u} + \frac{\|\mathbf{u}_k - \mathbf{u}\|}{\|\mathbf{d}\|} \mathbf{d} + o(\|\mathbf{u}_k - \mathbf{u}\|)$ . Alors si  $i \in \{1, \dots, p\}$ ,  $\varphi_i(\mathbf{u}_k) = \varphi_i(\mathbf{u}) = 0$ . En utilisant le développement de Taylor-Young au rang 1 au voisinage de  $\mathbf{u}$ , pour  $k$  suffisamment grand,

$$\underbrace{\varphi_i(\mathbf{u}_k)}_{=0} = \underbrace{\varphi_i(\mathbf{u})}_{=0} + \frac{\|\mathbf{u}_k - \mathbf{u}\|}{\|\mathbf{d}\|} \langle \nabla\varphi_i(\mathbf{u}), \mathbf{d} \rangle + o(\|\mathbf{u}_k - \mathbf{u}\|) .$$

On a donc nécessairement  $\langle \nabla\varphi_i(\mathbf{u}), \mathbf{d} \rangle = 0$ . Ainsi  $T_{\mathbf{u}}\mathcal{D}$  est un sous-espace vectoriel de dimension  $n - p$  de l'orthogonal :  $\langle \nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u}) \rangle^\perp$ . Or puisque  $\nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u})$  forment une famille libre de dimension  $p$  dans  $\mathbb{R}^n$ , ce dernier a pour dimension  $n - p$ . Ainsi  $T_{\mathbf{u}}\mathcal{D}$  coïncide avec  $\langle \nabla\varphi_1(\mathbf{u}), \dots, \nabla\varphi_p(\mathbf{u}) \rangle^\perp$ .  $\square$

## A.3 Rappels sur les matrices

### A.3.1 Notations

On note  $\mathcal{M}_{p,n}(\mathbb{R})$  l'espace vectoriel des matrices  $p \times n$  à coefficient réel.

Si  $A \in \mathcal{M}_{p,n}(\mathbb{R})$  on note  $A^\top$  sa matrice transposée.

On note  $\mathcal{M}_n(\mathbb{R})$  l'algèbre des matrices carrées  $n \times n$  à coefficient réel.

Pour  $A \in \mathcal{M}_n(\mathbb{R})$  on note  $\det(A)$  son *déterminant* et  $tr(A)$  sa trace.

### A.3.2 Norme matricielle

On peut munir  $\mathcal{M}_{p,n}(\mathbb{R})$  d'une norme de plusieurs façons. Une norme matricielle  $\|\cdot\|$  est *compatible* avec une norme vectorielle  $\|\cdot\|$  si  $\forall A \in \mathcal{M}_{p,n}(\mathbb{R})$  et  $\forall \mathbf{x} \in \mathbb{R}^n$ ,  $\|A\mathbf{x}\| \leq \|A\| \|\mathbf{x}\|$ . Voici quelques exemples de normes matricielles compatibles avec la norme euclidienne  $\|\cdot\|_2$  :

– La *norme de Frobenius* :

$$\|A\|_f = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2} = tr(AA^\top) .$$

– La norme induite par  $\|\cdot\|_2$  :

$$\|A\|_{\text{sup}} = \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|_2}{\|\mathbf{x}\|_2}.$$

– Si  $A$  est une matrice carrée diagonalisable, elle coïncide avec la norme spectrale :

$$\|A\|_s = \max \{|\lambda| \mid \lambda \text{ est une valeur propre de } A\}.$$

### A.3.3 Matrice (semi-)définie positive/négative

Une notion d'importance en optimisation est la propriété de la matrice Hessienne d'une application d'être (semi)-définie positive ou négative.

#### Définitions.

Une matrice  $A \in \mathcal{M}_n(\mathbb{R})$  est semi-définie positive si  $\forall \mathbf{x} \in \mathbb{R}^n, \mathbf{x}^\top A \mathbf{x} \geq 0$ .

Une matrice  $A \in \mathcal{M}_n(\mathbb{R})$  est définie positive si  $\forall \mathbf{x} \in \mathbb{R}^n \setminus \{0\}, \mathbf{x}^\top A \mathbf{x} > 0$ .

On définit de façon analogue une matrice carrée semi-définie négative, définie négative.

Dans les cas nous intéressant, les matrices considérées sont symétriques (*i.e.*  $A^\top = A$ ) réelles. On dispose du résultat important suivant :

**Théorème A.3 (Symétrique réelle  $\implies$  diagonalisable.)** *Toute matrice symétrique réelle est diagonalisable.*

Pour déterminer si une matrice symétrique est définie positive on utilisera le résultat suivant qui en donne plusieurs caractérisations.

#### Théorème A.4 (Caractérisation des matrices symétriques définies positives)

Soit  $A = (a_{ij})_{i,j=1..n}$  une matrice symétrique. Les assertions suivantes sont équivalentes :

- (i)  $A$  est définie positive.
- (ii) Toutes les valeurs propres de  $A$  sont  $> 0$ .
- (iii)  $\forall i = 0, \dots, n, c_i > 0$ , où  $p_A(\lambda) = \sum_{i=0}^n (-1)^i c_i \lambda_i^n$  est le polynôme caractéristique de  $A$ .
- (iv) Les déterminants  $\det(A_k)$  où  $A_k$  désigne  $A_k = (a_{ij})_{i,j=1..k}$  sont tous  $> 0$ .
- (v) Il existe une matrice  $M$  inversible tel que  $M^\top M = A$ .

De plus :

- (a) Si  $A$  est définie positive alors  $\forall i = 1, \dots, n, a_{ii} > 0$ .
- (b) Si  $A \in \mathcal{M}_2(\mathbb{R})$ ,  $A$  est définie positive si et seulement si  $\det(A) > 0$  et  $\text{tr}(A) > 0$ .

Pour déterminer qu'une matrice symétrique est semi-définie positive on utilisera le résultat suivant qui en donne plusieurs caractérisations.

#### Théorème A.5 (Caractérisation des matrices symétriques semi-définies positives)

Soit  $A = (a_{ij})_{i,j=1..n}$  une matrice symétrique. Les assertions suivantes sont équivalentes :

- (i)  $A$  est semi-définie positive.
- (ii) Toutes les valeurs propres de  $A$  sont  $\geq 0$ .
- (iii)  $\forall i = 0, \dots, n, c_i \geq 0$ , où  $p_A(\lambda) = \sum_{i=0}^n (-1)^i c_i \lambda_i^n$  est le polynôme caractéristique de  $A$ .
- (iv) Les mineurs principaux de  $A$  sont tous  $\geq 0$ .
- (v) Il existe une matrice  $M$  tel que  $M^\top M = A$ .

De plus :

- (a) Si  $A$  est semi-définie positive alors  $\forall i = 1, \dots, n, a_{ii} \geq 0$ .
- (b) Si  $A \in \mathcal{M}_2(\mathbb{R})$ ,  $A$  est définie positive si et seulement si  $\det(A) \geq 0$  et  $\text{tr}(A) \geq 0$ .



# Correction des exercices

## Chapitre I.

**Exercice 1.** On note  $x, y$  les quantités en litre de produits finis.

La fonction économique à maximiser -qui représente le bénéfice brut- est :

$$f(x, y) = 8x + 4y$$

sous les contraintes :

$$\begin{cases} x + y \leq 1000 \\ 15x + 3y \leq 4500 \\ x, y \geq 0 \end{cases}$$

Méthode du simplexe :

1	1	1	0	1000	$-\frac{1}{15}l$	0	4/5	1	-1/15	700	$l$
15	3	0	1	4500	$l$	15	3	0	1	4500	$-\frac{15}{4}l$
8	4	0	0	$f - 0$	$-\frac{8}{15}l$	0	12/5	0	-8/15	$f - 2400$	$-3l$

0	4/5	1	-1/15	700	$\Rightarrow y = 875$
15	0	-15/4	5/4	1875	$\Rightarrow x = 125$
0	0	-3	-1/3	$f - 4500$	$f_{max} = 4500$

**Exercice 2.** Le problème s'écrit (voir § 3.1) :

$$\begin{aligned} \max_{x, y, z} \quad & 2x + 1.6y + 1.8z \\ & 90x + 93y + 95z \leq 6500 \\ & 10x + 7y + 5z \leq 500 \\ & x, y, z \geq 0 \end{aligned}$$

Le problème est écrit sous forme normale, on lui applique la méthode du simplexe :

90	93	95	1	0	6500	0	30	50	1	-9	2000
10	7	5	0	1	500	10	7	5	0	1	500
2	1.6	1.8	0	0	$f - 0$	0	0.2	0.8	0	-0.2	$f - 100$

0	30	50	1	-9	2000	$\Rightarrow x_3 = 40$
10	4	0	-0.1	1.9	300	$\Rightarrow x_1 = 30$
0	-0.28	0	-0.016	-1.856	$f - 132$	$\Rightarrow f_{max} = 132$

On obtient  $x_1 = 30$ ,  $x_2 = 0$ ,  $x_3 = 40$ , pour  $f_{max} = 132$ . Il faut produire 30t, 0t et 40t, respectivement de bronze de qualités  $A$ ,  $B$ ,  $C$ , pour un bénéfice maximal de 132000 €. Les stocks de cuivre et d'étain sont

épuisés ( $s_1 = s_2 = 0$ ).

**Exercice 3. a.** En appelant  $x_1, x_2, x_3, x_4$  les quantités exprimées en unité de poids de chacun des 4 types d'aliment, le problème s'écrit :

$$\begin{aligned} \min_{x_1, \dots, x_4} \quad & 2x_1 + 2x_2 + x_3 + 8x_4 \\ \begin{cases} 2x_1 + x_2 + x_4 \geq 12 \\ x_1 + 2.5x_2 + 2x_3 + 4.5x_4 \geq 7 \end{cases} \\ x_1, x_2, x_3, x_4 \geq 0 \end{aligned}$$

Par dualité min/max, il est équivalent au problème :

$$\begin{aligned} \max_{y_1, y_2} \quad & 12y_1 + 7y_2 \\ \begin{cases} 2y_1 + y_2 \leq 2 \\ y_1 + 2.5y_2 \leq 2 \\ 2y_2 \leq 1 \\ y_1 + 4.5y_2 \leq 8 \end{cases} \\ y_1, y_2 \leq 0 \end{aligned}$$

Il est écrit sous forme normale ; on applique la méthode du simplexe.

$$\begin{array}{cccccc|cccc|c} \boxed{2} & 1 & 1 & 0 & 0 & 0 & 2 & 2 & 1 & 1 & 0 & 0 & 0 & 2 \\ 1 & 2.5 & 0 & 1 & 0 & 0 & 2 & 0 & 2 & -0.5 & 1 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 & 1 & 0 & 1 & 0 & \boxed{2} & 0 & 0 & 1 & 0 & 1 \\ 1 & 4.5 & 0 & 0 & 0 & 1 & 8 & 0 & 4 & -0.5 & 0 & 0 & 1 & 7 \\ \hline 12 & 7 & 0 & 0 & 0 & 0 & f-0 & 0 & 1 & -6 & 0 & 0 & 0 & f-12 \end{array}$$
  

$$\begin{array}{cccccc|cccc|c} 2 & 0 & 1 & 0 & -0.5 & 0 & 1.5 \\ 0 & 0 & -0.5 & 1 & -1 & 0 & 0 \\ 0 & 2 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & -0.5 & 0 & -2 & 1 & 5 \\ \hline 0 & 0 & \boxed{-6} & \boxed{0} & \boxed{-0.5} & \boxed{0} & f-12.5 \end{array}$$

Il faut acheter 6 u.p. d'aliment de type 1, 0.5u.p. d'aliment de type 3, et aucun aliment de types 2,4. Pour un coût de 12.5 u.m. on obtient 12 u. de glucides et 7 u. de lipides.

**b.** En appelant  $y_1, y_2$  le prix par unité de volume des aliments 1 et 2, il s'agit de maximiser la fonction (c'est la gain obtenu pour l'achat permettant d'obtenir 12u. de glucides et 7u. de lipides) :

$$g(y_1, y_2) = 12y_1 + 7y_2$$

Pour être compétitif le coût de ses produits pour obtenir la même quantité de glucides et lipides que dans les produits concurrents doit leur être inférieur ou égal. Cela s'exprime :

$$\begin{cases} 2y_1 + y_2 \leq 2 \\ y_1 + 2.5y_2 \leq 2 \\ 2y_2 \leq 1 \\ y_1 + 4.5y_2 \leq 8 \end{cases}$$

Ce problème n'est rien d'autre que le problème max dual du problème du consommateur. On lui a déjà appliqué la méthode du simplexe. On obtient en résolvant le système restant :  $y_1 = 0.75$ u.m. et  $y_2 = 0.5$ u.m..



## Chapitre II.

**Exercice 1.** L'application  $f$  est infiniment différentiable, car polynomiale. Pour mener une étude locale on détermine en chaque point son vecteur gradient et sa matrice Hessienne.

$$\nabla f(x, y) = \begin{pmatrix} 3x^2 + 2x \\ 3y^2 + 2y \end{pmatrix} \quad \nabla^2 f(x, y) = \begin{pmatrix} 6x + 2 & 0 \\ 0 & 6y + 2 \end{pmatrix}$$

On recherche ses points critiques,

$$\nabla f(x, y) = \begin{pmatrix} 3x^2 + 2x \\ 3y^2 + 2y \end{pmatrix} = 0 \iff \begin{cases} 3x^2 + 2x = 0 \\ 3y^2 + 2y = 0 \end{cases} \iff \begin{cases} x = 0 \text{ ou } x = -\frac{2}{3} \\ y = 0 \text{ ou } y = -\frac{2}{3} \end{cases}$$

Les points critiques sont donc :  $(0, 0)$ ,  $(0, -2/3)$ ,  $(-2/3, 0)$ ,  $(-2/3, -2/3)$ .

On évalue en chaque point critique la matrice Hessienne.

$$\nabla^2 f(0, 0) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix} \text{ est définie positive : } (0, 0) \text{ est un minimum local.}$$

$$\nabla^2 f(0, -\frac{2}{3}) = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix} \text{ n'est pas semi-définie : } (0, -\frac{2}{3}) \text{ n'est pas un extremum.}$$

$$\nabla^2 f(-\frac{2}{3}, 0) = \begin{pmatrix} -2 & 0 \\ 0 & 2 \end{pmatrix} \text{ n'est pas semi-définie : } (-\frac{2}{3}, 0) \text{ n'est pas un extremum.}$$

$$\nabla^2 f(-\frac{2}{3}, -\frac{2}{3}) = \begin{pmatrix} -2 & 0 \\ 0 & -2 \end{pmatrix} \text{ est définie négative : } (-\frac{2}{3}, -\frac{2}{3}) \text{ est un maximum local.}$$

L'application  $f$  n'admet pas d'extremum global, car elle est surjective sur  $\mathbb{R}$  :

$$\lim_{x \rightarrow +\infty} f(x, 0) = +\infty \quad \lim_{x \rightarrow -\infty} f(x, 0) = -\infty.$$

**Exercice 2.** Soit l'application  $f(x, y) = x^4 + y^4 - x^3 - y^3$ .

a. Montrons que  $f$  est coercive. Formons :

$$f(x, y) - x^2 - y^2 = x^2(x^2 - x - 1) + y^2(y^2 - y - 1).$$

Le trinôme  $t^2 - t - 1$  est positif lorsque  $t \notin ]\frac{-1-\sqrt{5}}{2}, \frac{-1+\sqrt{5}}{2}[$  et a pour minimum  $t = -\frac{1}{2}$  en lequel il vaut  $-\frac{1}{4}$ . Ainsi lorsque  $x$  ou  $y$  est suffisamment grand,  $f(x, y) \geq \|(x, y)\|^2$ . Toutes les normes étant équivalentes sur  $\mathbb{R}^2$ ,  $\exists C > 0$  tel que :

$$\|(x, y)\|_\infty = \sup\{|x|, |y|\} \geq C\|(x, y)\|_2 = \sqrt{x^2 + y^2}.$$

Ainsi lorsque  $\|(x, y)\|_2$  tend vers  $+\infty$ ,  $\sup\{|x|, |y|\}$  tend aussi vers  $+\infty$ , de sorte que  $f(x, y) \geq \|(x, y)\|_2$  et tend aussi vers  $+\infty$ . Donc  $f$  est coercive.

On en déduit (théorème II.2) l'existence d'un minimum global et d'aucun maximum global pour  $f$  sur  $\mathbb{R}^2$ .

b. Afin de déterminer le(s) minimum(s) de  $f$  on cherche ses extrema locaux en poursuivant une étude locale. L'application  $f$  est infiniment différentiable. Son vecteur gradient et sa matrice hessienne s'expriment :

$$\nabla f(x, y) = \begin{pmatrix} 4x^3 - 3x^2 \\ 4y^3 - 3y^2 \end{pmatrix} \quad \nabla^2 f(x, y) = \begin{pmatrix} 12x^2 - 6x & 0 \\ 0 & 12y^2 - 6y \end{pmatrix}.$$

Ainsi  $f$  a 4 points critiques  $A = (0, 0)$ ,  $B = (0, \frac{3}{4})$ ,  $C = (\frac{3}{4}, 0)$  et  $D = (\frac{3}{4}, \frac{3}{4})$ . En  $D$  la matrice hessienne est définie positive :  $D$  est un minimum local de  $f$ . En  $A$ ,  $B$ ,  $C$  la matrice hessienne est semi-définie positive : on ne peut rien déduire sur la nature des points critiques  $A$ ,  $B$ ,  $C$ .

Pour déterminer le(s) minimum(s) de  $f$  il suffit d'évaluer  $f$  en ses 4 points critiques :

$$f(0, 0) = 0 > f(0, \frac{3}{4}) = f(\frac{3}{4}, 0) = -\frac{3^3}{4^4} > f(\frac{3}{4}, \frac{3}{4}) = -2\frac{3^3}{4^4}.$$

Ainsi le minimum de  $f$  est le point  $D = (\frac{3}{4}, \frac{3}{4})$ .

c. Soit  $\mathbf{u}$  un point critique de  $g$ , i.e.  $\nabla g(\mathbf{u}) = \mathbf{0}$ .

Si  $\mathbf{u}$  est un minimum local de  $g$ , il existe une boule ouverte  $\mathcal{B}$  centrée en  $\mathbf{u}$  tel que,  $\forall \mathbf{x} \in \mathcal{B}$ ,  $g(\mathbf{x}) \geq g(\mathbf{u}) = g(\mathbf{u}) + \langle \nabla g(\mathbf{u}), \mathbf{x} - \mathbf{u} \rangle$ . Avec le théorème II.5.1 on en déduit que la restriction de  $g$  à  $\mathcal{B}$  est une application convexe.

Réciproquement, avec le théorème II.5.1, si  $g$  est convexe sur une boule  $\mathcal{B}$  centrée en  $\mathbf{u}$ , alors  $\forall \mathbf{x} \in \mathcal{B}$ ,  $g(\mathbf{x}) \geq g(\mathbf{u}) + \langle \nabla g(\mathbf{u}), \mathbf{x} - \mathbf{u} \rangle = g(\mathbf{u})$  et en particulier  $\mathbf{u}$  est un minimum local de  $g$ .

d. Revenons en à l'étude des extrema locaux de  $f$ . Nous avons déterminé la matrice Hessienne de  $f$ . Le binôme  $12t^2 - 6t$  ne garde pas un signe constant sur un voisinage de 0. Ainsi sur aucun voisinage de  $A$ ,  $B$  et  $C$ , la matrice hessienne ne reste semi-définie positive ou négative. Avec le théorème II.5.2, l'application  $f$  n'est ni localement convexe ni localement concave sur un voisinage convexe de  $A$ ,  $B$  ou  $C$ . Ainsi en appliquant le résultat établi en c), ni  $A$ , ni  $B$ , ni  $C$  n'est un extremum local de  $f$ .

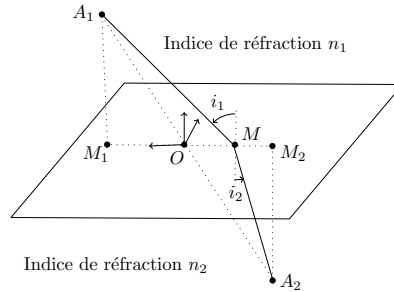
**Exercice 3.** On rappelle que l'indice de réfraction est  $n_i = \frac{c}{v_i}$  où  $c$  désigne la vitesse de propagation de la lumière dans le vide et  $v_i$  sa vitesse de propagation dans le milieu.

La lumière parcourt le trajet qui minimise le temps de parcours. Ce dernier est :

$$\frac{A_1 M}{v_1} + \frac{A_2 M}{v_2} = \frac{n_1 A_1 M}{c} + \frac{n_2 A_2 M}{c}.$$

Il s'agit donc de déterminer le point  $M$  de façon à minimiser le *chemin optique*  $n_1 A_1 M + n_2 A_2 M$ .

On se donne un repère orthonormé construit de la façon suivante (voir la figure ci-après) : soit  $O$  le point d'intersection de la droite  $(A_1 A_2)$  avec le plan de séparation que nous appellerons  $(P)$ . Soient  $M_1$  et  $M_2$  les projetés orthogonaux respectifs de  $A_1$  et  $A_2$  sur  $(P)$ . Le segment  $[M_1 M_2]$  passe par  $O$ . On choisit un repère orthonormal d'origine  $O$ , tel que  $(O_i)$  est confondu avec  $(M_1 M_2)$  et  $(O_j)$  est dans  $(P)$ ; alors  $k$  est orthogonal à  $(P)$ .



Les coordonnées de  $M, A_1, A_2$  dans ce repère sont respectivement  $(x, y, 0)$ ,  $(x_1, 0, z_1)$  et  $(x_2, 0, z_2)$ . Le chemin optique s'exprime alors :

$$f(x, y) = n_1 \sqrt{(x - x_1)^2 + y^2 + z_1^2} + n_2 \sqrt{(x - x_2)^2 + y^2 + z_2^2}$$

et il s'agit de le minimiser. L'application  $f$  est clairement coercive et admet donc un minimum. Etudions ses points critiques :

$$\begin{aligned} \frac{\partial f}{\partial x}(x, y) &= n_1 \frac{x - x_1}{\sqrt{(x - x_1)^2 + y^2 + z_1^2}} + n_2 \frac{x - x_2}{\sqrt{(x - x_2)^2 + y^2 + z_2^2}} \\ \frac{\partial f}{\partial y}(x, y) &= n_1 \frac{y}{\sqrt{(x - x_1)^2 + y^2 + z_1^2}} + n_2 \frac{y}{\sqrt{(x - x_2)^2 + y^2 + z_2^2}} \end{aligned}$$

Puisque  $\frac{\partial f}{\partial y}(x, y) = 0$ , on a  $y = 0$ . Ainsi  $M$  est situé sur la droite  $(M_1 M_2)$ .

Puisque  $\frac{\partial f}{\partial x}(x, y) = 0$ ,  $x - x_1$  et  $x - x_2$  sont de signes opposés, ainsi (par exemple)  $x_1 \leq x \leq x_2$  :  $M$  est situé sur le segment  $[M_1 M_2]$ . Alors au point  $M(x, 0)$ ,

$$\frac{\partial f}{\partial x}(x, 0) = n_1 \frac{x - x_1}{\sqrt{(x - x_1)^2 + z_1^2}} + n_2 \frac{x - x_2}{\sqrt{(x - x_2)^2 + z_2^2}} = n_1 \frac{M_1 M}{A_1 M} - n_2 \frac{M_2 M}{A_2 M} = 0$$

ce qui implique  $n_1 \frac{M_1 M}{A_1 M} = n_2 \frac{M_2 M}{A_2 M}$  ; or  $\frac{M_k M}{A_k M} = \cos(\frac{\pi}{2} - i_k) = \sin i_k$ ,  $k = 1, 2$ . On trouve donc qu'au minimum on a :

$$n_1 \sin i_1 = n_2 \sin i_2 .$$

**Exercice 4. a.** Le problème de minimisation  $\min_{\mathbf{v} \in \mathcal{C}} \|\mathbf{u} - \mathbf{v}\|$  est équivalent au problème  $\min_{\mathbf{v} \in \mathcal{C}} \|\mathbf{u} - \mathbf{v}\|^2$ . Or l'application

$$f : \mathbf{x} \mapsto \|\mathbf{u} - \mathbf{x}\|^2 = \sum_{i=1}^n (u_i - x_i)^2 = \mathbf{x}^\top \text{Id } \mathbf{x} - 2\mathbf{u}^\top \mathbf{x} + \|\mathbf{u}\|^2$$

est une application quadratique de matrice hessienne  $2\text{Id}$ . Avec le théorème II.9,  $f$  est une application elliptique, et donc strictement convexe et coercive. Le domaine  $\mathcal{C}$  étant convexe fermé et non vide elle y admet un unique minimum,  $P_{\mathcal{C}}(\mathbf{u})$ .

**b.** On est dans le cadre de la programmation convexe, et  $f$  est différentiable. La caractérisation de  $P_{\mathcal{C}}(\mathbf{u})$  est donnée par le théorème II.6.(iv) :  $\forall \mathbf{v} \in \mathcal{C}$ ,  $\langle \nabla f(P_{\mathcal{C}}(\mathbf{u})), \mathbf{v} - P_{\mathcal{C}}(\mathbf{u}) \rangle \geq 0$ . Or  $\nabla f(P_{\mathcal{C}}(\mathbf{u})) = 2(P_{\mathcal{C}}(\mathbf{u}) - \mathbf{u})$ . On obtient donc :

$$\forall \mathbf{v} \in \mathcal{C}, \quad 2\langle P_{\mathcal{C}}(\mathbf{u}) - \mathbf{u}, \mathbf{v} - P_{\mathcal{C}}(\mathbf{u}) \rangle \geq 0$$

et la caractérisation donnée en découle immédiatement.

**c.** En appliquant la caractérisation des points  $P_{\mathcal{C}}(\mathbf{x})$  et  $P_{\mathcal{C}}(\mathbf{y})$  :

$$\langle P_{\mathcal{C}}(\mathbf{x}) - \mathbf{x}, P_{\mathcal{C}}(\mathbf{y}) - P_{\mathcal{C}}(\mathbf{x}) \rangle \geq 0$$

$$\langle P_{\mathcal{C}}(\mathbf{y}) - \mathbf{y}, P_{\mathcal{C}}(\mathbf{x}) - P_{\mathcal{C}}(\mathbf{y}) \rangle \geq 0$$

En additionnant ces deux inégalités :

$$\langle P_{\mathcal{C}}(\mathbf{x}) - P_{\mathcal{C}}(\mathbf{y}) - \mathbf{x} + \mathbf{y}, P_{\mathcal{C}}(\mathbf{y}) - P_{\mathcal{C}}(\mathbf{x}) \rangle \geq 0 ,$$

soit

$$\langle \mathbf{y} - \mathbf{x}, P_{\mathcal{C}}(\mathbf{y}) - P_{\mathcal{C}}(\mathbf{x}) \rangle \geq \|P_{\mathcal{C}}(\mathbf{y}) - P_{\mathcal{C}}(\mathbf{x})\|^2$$

et en appliquant l'inégalité de Cauchy-Schwartz au membre de gauche :

$$\|\mathbf{y} - \mathbf{x}\| \|P_{\mathcal{C}}(\mathbf{y}) - P_{\mathcal{C}}(\mathbf{x})\| \geq \langle \mathbf{y} - \mathbf{x}, P_{\mathcal{C}}(\mathbf{y}) - P_{\mathcal{C}}(\mathbf{x}) \rangle \geq \|P_{\mathcal{C}}(\mathbf{y}) - P_{\mathcal{C}}(\mathbf{x})\|^2$$

dont on déduit l'inégalité recherchée.

#### Exercice 5.

**1.** Puisque  $] -\infty, u[$  est un ouvert de  $\mathbb{R}$  et que  $f : \mathcal{D} \rightarrow \mathbb{R}$  est continue,  $f^{-1}(] -\infty, u[)$  est un ouvert de  $\mathcal{D}$ .

**2.** Soit  $r = v - u$ , alors tout point  $x$  de la boule ouverte  $B$  de  $\mathbb{R}$  centrée en  $v$  et de rayon  $r$  vérifie  $x \geq u$ , en particulier  $f^{-1}(B)$  est contenu dans  $\mathcal{L}_{\mathcal{D}} f^{-1}(] -\infty, u[)$  et contient  $f^{-1}(\{v\})$ . Puisque  $B$  est un ouvert de  $\mathbb{R}$  et  $f$  est continue,  $f^{-1}(B)$  est un ouvert de  $\mathcal{D}$ . Donc  $\mathcal{L}_{\mathcal{D}} f^{-1}(] -\infty, u[)$  est un voisinage de tout point de  $f^{-1}(\{v\})$ .

**3.** Soit  $\mathbf{x} \in f^{-1}(\{u\})$  ; puisque  $\mathbf{x}$  est un min local il existe un ouvert  $\mathcal{U}$  de  $\mathbb{R}^n$  contenant  $\mathbf{x}$  tel que  $\forall \mathbf{y} \in \mathcal{U} \cap \mathcal{D}$ ,  $f(\mathbf{y}) \geq f(\mathbf{x}) = u$ . Ainsi  $(\mathcal{U} \cap \mathcal{D}) \subset \mathcal{L}_{\mathcal{D}} f^{-1}(] -\infty, u[)$ , et par définition c'est un ouvert de  $\mathcal{D}$  ;  $\mathcal{L}_{\mathcal{D}} f^{-1}(] -\infty, u[)$  est donc un voisinage de  $\mathbf{x}$  dans  $\mathcal{D}$ .

**4.** On déduit de 2 et 3 que  $\mathcal{L}_{\mathcal{D}} f^{-1}(] -\infty, u[)$  est un voisinage de tous ses points. C'est donc un ouvert de  $\mathcal{D}$  et donc son complément  $f^{-1}(] -\infty, u[)$  est un fermé de  $\mathcal{D}$ .

**5.** On a montré en 1 et 4 que  $f^{-1}(] -\infty, u[)$  est à la fois fermé et ouvert dans  $\mathcal{D}$ . Puisque  $\mathcal{D}$  est connexe,  $f^{-1}(] -\infty, u[)$  est soit  $\emptyset$  soit  $\mathcal{D}$ . Puisque  $\mathbf{u} \in \mathcal{D}$  n'est pas dans  $f^{-1}(] -\infty, u[)$ , c'est l'ensemble vide. Ainsi,  $\forall \mathbf{x} \in \mathcal{D}$ ,  $f(\mathbf{x}) \geq f(\mathbf{u})$  ;  $\mathbf{u}$  est donc un minimum global de  $f$  sur  $\mathcal{D}$ .

### Chapitre III.

**Exercice 1.** On traite séparément les exemples A et B.

*Exemple A.*  $f(x, y) = x$  et  $\mathcal{D} = \{x^2 + y^2 = 1\}$ ; on retrouve les solutions évidentes trouvées au § III.1.2 en appliquant ici les conditions de Lagrange.

On l'a vu, puisque  $f$  est continue et  $\mathcal{D}$  est compact, il existe un minimum et un maximum global.

$$\nabla f(x, y) = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ et } \nabla \varphi(x, y) = \begin{pmatrix} 2x \\ 2y \end{pmatrix}.$$

Puisque  $\nabla \varphi(x, y) \neq 0$  sur  $\mathcal{D}$ ,  $\nabla \varphi(x, y)$  forme une famille linéairement indépendante. On applique la condition nécessaire de Lagrange :

$$(x, y) \text{ extremum local} \implies \nabla_x \mathcal{L}(x, y, \lambda) = \mathbf{0} \implies \begin{cases} 1 + 2\lambda x = 0 & (1) \\ 2\lambda y = 0 & (2) \\ x^2 + y^2 = 1 & (3) \end{cases}$$

On résout ce système d'inconnues  $x, y$  :

$$\left. \begin{array}{l} (2) \implies y = 0 \text{ ou } \lambda = 0 \\ (1) \implies \lambda \neq 0 \end{array} \right\} \implies y = 0 \xrightarrow{(3)} x = \pm 1.$$

On obtient deux solutions :

$$\mathbf{a} = (1, 0) \text{ (avec } \lambda = -1/2) \quad ; \quad \mathbf{b} = (-1, 0) \text{ (avec } \lambda = 1/2).$$

Puisque  $f(\mathbf{a}) = 1$  et  $f(\mathbf{b}) = -1$ , et que l'on connaît déjà l'existence d'un minimum et d'un maximum global, on peut d'ores et déjà conclure que  $\mathbf{b}$  est le minimum et  $\mathbf{a}$  est le maximum (globaux). On retrouve cependant qu'il sont minimum et maximum (local) en appliquant les conditions du second ordre.

$$\nabla_x^2 \mathcal{L}(x, y, \lambda) = \begin{pmatrix} 2\lambda & 0 \\ 0 & 2\lambda \end{pmatrix}$$

L'espace tangent à  $\mathcal{D}$  en  $\mathbf{u} = \mathbf{a}$  ou  $\mathbf{b}$  est ici le même,  $T_{\mathbf{u}}\mathcal{D} = \text{Vect}((0, 1))$ ; ça n'a ici peu d'importance, puisque :

$$\text{en } \mathbf{a}, \nabla_x^2 \mathcal{L}(\mathbf{a}, -1/2) = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \text{ est définie négative} \implies \mathbf{a} \text{ est un max,}$$

$$\text{en } \mathbf{b}, \nabla_x^2 \mathcal{L}(\mathbf{b}, 1/2) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \text{ est définie positive} \implies \mathbf{b} \text{ est un min,}$$

et puisque ce sont les seuls extrema locaux de  $f$  sur  $\mathcal{D}$ , ce sont des extrema globaux par compacité.

*Exemple B.* On reprend l'exemple B du § III.1.2 :

$$f(x, y) = x^2 + y^2 \text{ et } \mathcal{D} = \{(x, y) \in \mathbb{R}^2 \mid xy = 1\}.$$

$$\nabla f(x, y) = \begin{pmatrix} 2x \\ 2y \end{pmatrix} \quad ; \quad \nabla \varphi(x, y) = \begin{pmatrix} y \\ x \end{pmatrix} \quad ; \quad \nabla_x \mathcal{L}(x, y, \lambda) = \begin{pmatrix} 2x + \lambda y \\ 2y + \lambda x \end{pmatrix}$$

Sur  $\mathcal{D}$ ,  $\nabla \varphi(x, y) \neq 0$ , aussi on peut appliquer les conditions de Lagrange, ce qui nous amène à résoudre le système :

$$\begin{cases} 2x + \lambda y = 0 & (1) \\ 2y + \lambda x = 0 & (2) \\ y = 1/x & (3) \end{cases}$$

En formant l'équation (1) - (2) on obtient :

$$(x - y)(2 - \lambda) = 0 \implies \begin{cases} x = y \xrightarrow{(3)} x^2 = 1 \implies x = \pm 1 \\ \text{ou} \\ \lambda = 2 \xrightarrow{(1)} y = -x \xrightarrow{(3)} x^2 = -1 \text{ impossible.} \end{cases}$$

On obtient deux solutions :

$$\mathbf{a} = (1, 1) \quad \text{et} \quad \mathbf{b} = (-1, -1) \quad (\text{avec } \lambda = -2).$$

On détermine en ces deux points la matrice Hessienne du Lagrangien :

$$\nabla_x^2 \mathcal{L}(x, y, \lambda) = \begin{pmatrix} 2 & \lambda \\ \lambda & 2 \end{pmatrix} \quad ; \quad \nabla_x^2 \mathcal{L}(\mathbf{a}, -2) = \nabla_x^2 \mathcal{L}(\mathbf{b}, -2) = \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix}$$

En  $\mathbf{u} = \mathbf{a}$  ou  $\mathbf{b}$ ,  $T_{\mathbf{u}}\mathcal{D} = \text{Vect}((1, -1)) = \{(t, -t) \mid t \in \mathbb{R}\}$  (orthogonal de  $\nabla\varphi(\mathbf{u}) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$  et  $\begin{pmatrix} -1 \\ -1 \end{pmatrix}$ ). Or, si  $t \neq 0$  :

$$(t, -t) \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} \begin{pmatrix} t \\ -t \end{pmatrix} = 8t^2 > 0$$

Ainsi,  $\forall \mathbf{x} \neq 0 \in T_{\mathbf{u}}\mathcal{D}$ ,  $\mathbf{x}^\top \nabla_x^2 \mathcal{L}(\mathbf{u}, -2) \mathbf{x} > 0$ , et donc  $\mathbf{a}$  et  $\mathbf{b}$  sont deux minima locaux stricts. Ils sont en fait globaux car  $f$  est coercive sur le fermé  $\mathcal{D}$  et  $f(\mathbf{a}) = f(\mathbf{b})$ .

**Exercice 2.** L'application  $f$  a déjà été étudiée dans l'exercice 1 du chapitre 2.

On se souvient qu'elle n'admet pas d'extremum global sur  $\mathbb{R}^2$  car elle est surjective.

Le cercle  $\mathcal{C}$  est un compact (fermé borné) de  $\mathbb{R}^2$ , donc  $f$  étant continue elle admet un minimum et un maximum global sur  $\mathcal{C}$ . Il s'agit d'un problème d'optimisation sous contrainte égalitaire.

Soit la contrainte égalitaire  $\varphi(x, y) = x^2 + y^2 - 1 = 0$ . On a  $\nabla\varphi(x, y) = \begin{pmatrix} 2x \\ 2y \end{pmatrix} \neq 0$  sur  $\mathcal{C}$ . Donc en tout point de  $\mathcal{C}$  les contraintes sont qualifiées. On a donc en tout extrémum  $(x, y) \in \mathcal{C}$  de  $f$ , les conditions de Lagrange :

$$\begin{aligned} \exists \lambda \in \mathbb{R}, \nabla f(x, y) + \lambda \nabla\varphi(x, y) &= 0 \\ \implies \begin{cases} 3x^2 + 2x + 2\lambda x &= 0 \\ 3y^2 + 2y + 2\lambda y &= 0 \end{cases} &\implies \begin{cases} x(3x + 2 + 2\lambda) &= 0 \\ y(3y + 2 + 2\lambda) &= 0 \end{cases} \implies \begin{cases} x = 0 \text{ ou } x = -\frac{2+2\lambda}{3} \\ y = 0 \text{ ou } y = -\frac{2+2\lambda}{3} \end{cases} \end{aligned}$$

Puisque  $(x, y) \in \mathcal{C}$ , on a  $x^2 + y^2 = 1$ , et donc :

$$\begin{aligned} x = 0 &\implies y = \pm 1 \\ y = 0 &\implies x = \pm 1 \\ x = y = -\frac{2+2\lambda}{3} &\implies 2 \left( \frac{2+2\lambda}{3} \right)^2 = 1 \implies \lambda = -1 \pm \frac{3}{4}\sqrt{2} \implies x = y = \pm \frac{1}{\sqrt{2}} \end{aligned}$$

On obtient donc 6 points vérifiant les conditions nécessaires de Lagrange :

Point	(0, 1)	(0, -1)	(1, 0)	(-1, 0)	$(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$	$(-\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$
Valeur de $f$	1	-1	1	-1	$\frac{1}{\sqrt{2}}$	$-\frac{1}{\sqrt{2}}$

Ainsi  $f$  admet sur  $\mathcal{C}$  :

- pour minima les 2 points  $(-1, 0)$  et  $(0, -1)$ ,
- pour maxima les 2 points  $(1, 0)$  et  $(0, 1)$ .

**Exercice 3.** Le problème se formule :

$$\begin{aligned} \max_{x, y} \quad & 3xy - x^2 - 3y^2 \\ & x + y = 28 \\ & x, y > 0 \end{aligned}$$

Il s'agit d'un problème de programmation quadratique sous contrainte égalitaire, sur l'ouvert  $\mathcal{U} = (\mathbb{R}_+^*)^2$ . La matrice hessienne de  $f$  est

$$A = \begin{pmatrix} -2 & 3 \\ 3 & -6 \end{pmatrix}.$$

Puisque  $\det A = 3 > 0$  et  $\text{tr}(A) = -8 < 0$ ,  $A$  est définie négative *i.e.*  $f$  est strictement concave et admet donc au plus un maximum  $\mathbf{u}$ , et s'il existe  $(x, y)$  est solution sur  $(\mathbb{R}_+)^2$  du système :

$$\begin{cases} -2x + 3y + \lambda = 0 \\ 3x - 6y + \lambda = 0 \\ x + y = 28 \end{cases}$$

Que l'on résout pour obtenir  $x = 18$  et  $y = 10$ . L'allocation optimale est 18 publicités en magazine et 10mn de télévision.

**Exercice 4.** (Problème de Kepler.)

Le problème se formule :

$$\begin{cases} \max xyz \\ x^2/a^2 + y^2/b^2 + z^2/c^2 = 1 \\ x, y, z \geq 0 \end{cases}$$

On calcule :

$$\nabla f(\mathbf{x}) = \begin{pmatrix} yz \\ xz \\ xy \end{pmatrix} ; \quad \nabla \varphi(\mathbf{x}) = \begin{pmatrix} 2x/a^2 \\ 2y/b^2 \\ 2z/c^2 \end{pmatrix} \neq 0 \text{ sur } \mathcal{E} \text{ si } xyz \neq 0.$$

Les contraintes inégalitaires étant toutes insaturées car  $x, y, z > 0 \implies \mu_1 = \mu_2 = \mu_3 = 0$ . En appliquant (KKT) :

$$\begin{aligned} \nabla f(\mathbf{x}) + \lambda \nabla \varphi(\mathbf{x}) &= 0 \\ \begin{cases} yz + \lambda 2x/a^2 = 0 \\ xz + \lambda 2y/b^2 = 0 \\ xy + \lambda 2z/c^2 = 0 \end{cases} \end{aligned}$$

On multiplie la première ligne par  $x$ , la deuxième par  $y$  et la dernière par  $z$ , puis on somme : on obtient  $3xyz + 2\lambda = 0 \implies yz = -2\lambda/(3x) \implies -2\lambda/(3x) + \lambda 2x/a^2 = 0 \implies 2\lambda(3x^2 - a^2)/(3a^2x) = 0$ . Or  $\lambda = 0$  est impossible car autrement  $xyz = 0$ . Donc  $3x^2 = a^2$ . De la même façon  $3y^2 = b^2$  et  $3z^2 = c^2$ . Donc :

$$x = \frac{a}{\sqrt{3}} ; \quad y = \frac{b}{\sqrt{3}} ; \quad z = \frac{c}{\sqrt{3}}$$

Et par suite, par compacité :

$$\text{Vol}_{\max} = \frac{abc}{3\sqrt{3}}$$

Le volume maximal du parallélépipède rectangle inscrit dans une ellipsoïde est  $1/(3\sqrt{3})$  fois le volume du parallélépipède dans lequel  $\mathcal{E}$  est inscrit.

**Exercice 5.** (Problème de Tartaglia)

Le problème se formule :

$$\begin{aligned} &\begin{cases} \max p_1 p_2 \times (p_2 - p_1) = p_1 p_2^2 - p_1^2 p_2 \\ p_1 + p_2 = 8 \\ p_1, p_2 \geq 0 \end{cases} \\ &\mathbf{u} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix} ; \quad \nabla f(\mathbf{u}) = \begin{pmatrix} p_2^2 - 2p_1 p_2 \\ 2p_1 p_2 - p_1^2 \end{pmatrix} \\ &\nabla \varphi(\mathbf{u}) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} ; \quad \nabla \psi_1(\mathbf{u}) = \begin{pmatrix} -1 \\ 0 \end{pmatrix} ; \quad \nabla \psi_2(\mathbf{u}) = \begin{pmatrix} 0 \\ -1 \end{pmatrix} \end{aligned}$$

Les contraintes étant affines on peut appliquer les conditions (KKT). Clairement les contraintes inégalitaires sont insaturées :  $p_1, p_2 > 0 \implies \mu_1 = \mu_2 = 0$ . On obtient :

$$\begin{cases} p_2^2 - 2p_1 p_2 + \lambda = 0 & (l_1) \\ 2p_1 p_2 - p_1^2 + \lambda = 0 & (l_2) \end{cases}$$

En formant  $(l_1) - (l_2)$  on obtient  $p_1^2 + p_2^2 - 4p_1p_2 = (p_1 + p_2)^2 - 6p_1p_2 = 0$ . Puisque  $p_1 + p_2 = 8$ , on a  $p_1p_2 = \frac{32}{3}$ . Donc  $p_1, p_2$  sont les racines du polynôme  $x^2 - 8x + \frac{32}{3}$ . On trouve  $(\Delta = \frac{8^2}{3})$  :

$$p_1 = 4 - \frac{4}{\sqrt{3}} \quad ; \quad p_2 = 4 + \frac{4}{\sqrt{3}} .$$

*Question. Par compacité il existe aussi un minimum global que les conditions de Lagrange doivent déterminer ! Réponse : c'est  $(p_2, p_1)$ .*

**Exercice 6.** Le domaine  $\mathcal{D} = \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\| \leq 1\}$  est un compact. L'application  $f(\mathbf{x}) = \mathbf{x}^\top A \mathbf{x}$  est quadratique et donc continue. Ainsi  $f$  admet (au moins) un maximum sur  $\mathcal{D}$ .

Appliquons les conditions de KKT. La contrainte est  $\psi(\mathbf{x}) = \sum_{i=1}^n x_i^2 - 1$ . En un maximum  $\mathbf{u}$ , il existe  $\mu \leq 0$ , tel que :

$$\nabla f(\mathbf{u}) + \mu \nabla \psi(\mathbf{u}) = 2A\mathbf{u} + 2\mu \text{Id } \mathbf{u} = \mathbf{0} \quad \implies \quad A\mathbf{u} = -\mu \mathbf{u} .$$

Ainsi :

- si  $\mu \neq 0$ ,  $\mathbf{u}$  est un vecteur propre de  $A$  associé à la valeur propre  $-\mu > 0$ . Dans ce cas la contrainte est saturée,  $\|\mathbf{u}\| = 1$ , et  $f(\mathbf{u}) = -\mu \mathbf{u}^\top \mathbf{u} = -\mu \|\mathbf{u}\|^2 = -\mu$ .
- si  $\mu = 0$ ,  $\mathbf{u}$  est un vecteur de  $\ker A$ .

On peut conclure : si  $A$  a une valeur propre  $> 0$ ,  $\mathbf{u}$  est un vecteur propre unitaire associé à la plus grande valeur propre de  $A$ . Sinon  $\mathbf{u}$  est n'importe quel élément du noyau de  $A$ .

## Chapitre IV.

**Exercice 1.** On applique la méthode de Newton pour la recherche de zéro de l'application  $f(x) = x^2 - 2$ . Elle s'écrit :

$$\mathbf{u}_{k+1} = \mathbf{u}_k - f'(\mathbf{u}_k)^{-1} f(\mathbf{u}_k) = \mathbf{u}_k - (2\mathbf{u}_k)^{-1} (\mathbf{u}_k^2 - 2) = \frac{\mathbf{u}_k}{2} + \frac{1}{\mathbf{u}_k} .$$

En prenant  $\mathbf{u}_0 = 1$ , on obtient :

Avec 10 chiffres significatifs :  $\sqrt{2} = 1.41421356237309$ .

$\mathbf{u}_1 = 1.5000000000000000$ ,

$\mathbf{u}_2 = 1.4166666666666667$ ,

$\mathbf{u}_3 = 1.41421568627451$ ,

$\mathbf{u}_4 = 1.41421356237469$ ,

$\mathbf{u}_5 = 1.41421356237309$ ,

La convergence est particulièrement rapide ! On trouve une valeur approchée à  $10^{-10}$  près en 5 itérations. Elle dépend ici peu du point base. Avec un point base négatif elle converge néanmoins vers  $-\sqrt{2}$ , c'est à dire vers l'autre zéro de  $f$ . Pour des valeurs initiales s'éloignant de la solution le nombre d'itération nécessaire est plus important. Voir en guise d'exemple le code matlab pour l'implémentation :

```

%% Méthode de Newton pour sqrt(2) %%
format long;
u=1;
N=5;
for i=1:N
    u=u/2+1/u
end
```

Voir aussi à ce sujet l'exercice 1 du TP n°4.

**Exercice 2.** Dans ce cas  $\forall \mathbf{x} \in \mathbb{R}^n$ ,  $\nabla f(\mathbf{x}) = A\mathbf{x} - \mathbf{b}$  et  $\nabla^2 f(\mathbf{x}) = A$ . La méthode de Newton s'exprime :

$$\mathbf{u}_{k+1} = \mathbf{u}_k - \nabla^2 f(\mathbf{u}_k)^{-1} \nabla f(\mathbf{u}_k) = \mathbf{u}_k - A^{-1}(A\mathbf{u}_k - \mathbf{b}) = A^{-1}\mathbf{b} .$$

Puisque  $A$  est définie positive,  $A^{-1}\mathbf{b}$  est l'unique minimum de  $f$  sur  $\mathbb{R}^n$  (cf. théorème II.10). Ainsi la méthode de Newton équivaut à la résolution directe de ce problème. Sa convergence se fait en une itération

et ne dépend pas du point base. Elle n'apporte donc rien ici. En général, la méthode de Newton peut se réinterpréter de la façon suivante : elle revient à approcher au voisinage de  $\mathbf{u}_k$  l'application à minimiser par une application quadratique.

**Exercice 3.** Soit  $t \in \{1, \dots, n\}$ .

$$\begin{aligned} f(x_1, x_2, \dots, x_n) &= \frac{1}{2} \sum_{i=1}^n a_{ii} x_i^2 + \sum_{i < j} a_{ij} x_i x_j - \sum_{i=1}^n b_i x_i \\ &= \underbrace{\frac{1}{2} a_{tt} x_t^2 + x_t \sum_{\substack{j=1 \\ j \neq t}}^n a_{tj} x_j - b_t x_t}_{\text{dépend de } x_t} + \underbrace{\frac{1}{2} \sum_{\substack{i=1 \\ i \neq t}}^n a_{ii} x_i^2 + \sum_{\substack{i < j \\ i, j \neq t}} a_{ij} x_i x_j - \sum_{\substack{i=1 \\ i \neq t}}^n b_i x_i}_{\text{ne dépend pas de } x_t} \end{aligned}$$

Alors :

$$\frac{\partial f}{\partial x_t}(x_1, \dots, x_n) = a_{tt} x_t + \sum_{\substack{j=1 \\ j \neq t}}^n a_{tj} x_j - \mathbf{b} = \sum_{j=1}^n a_{tj} x_j - \mathbf{b}$$

Puisque  $A$  est définie positive,  $f$  est elliptique. Cela implique qu'en tout point  $\mathbf{u} \in \mathbb{R}^n$ , et pour tout  $t \in \{1, \dots, n\}$  chacune des applications  $x \mapsto f(\mathbf{u} + x \mathbf{e}_t)$  est strictement convexe et coercive. Chacune admet donc un minimum global caractérisé par la condition d'Euler :

$$\frac{\partial f}{\partial x_t}(\mathbf{u} + x \mathbf{e}_t) = 0 .$$

Ainsi, si :

$$\begin{aligned} f(x_1^{(k+1)}, x_2^{(k)}, \dots, x_n^{(k)}) &= \inf_{x \in \mathbb{R}} f(x, x_2^{(k)}, \dots, x_n^{(k)}) \\ f(x_1^{(k+1)}, x_2^{(k+1)}, \dots, x_n^{(k)}) &= \inf_{x \in \mathbb{R}} f(x_1^{(k+1)}, x, \dots, x_n^{(k)}) \\ &\vdots \\ f(x_1^{(k+1)}, \dots, x_{n-1}^{(k+1)}, x_n^{(k+1)}) &= \inf_{x \in \mathbb{R}} f(x_1^{(k+1)}, \dots, x_{n-1}^{(k+1)}, x) \end{aligned}$$

Alors donné  $\mathbf{u}_k = (x_1^{(k)}, \dots, x_n^{(k)})$ , le point  $\mathbf{u}_{k+1} = (x_1^{(k+1)}, \dots, x_n^{(k+1)})$  construit par la méthode de relaxation est caractérisé par :

$$\begin{aligned} a_{11} x_1^{(k+1)} + a_{12} x_2^{(k)} + \dots + a_{1n} x_n^{(k)} &= b_1 \\ &\vdots \\ a_{21} x_1^{(k+1)} + a_{22} x_2^{(k+1)} + \dots + a_{2n} x_n^{(k)} &= b_2 \\ a_{n1} x_1^{(k+1)} + a_{n2} x_2^{(k+1)} + \dots + a_{nn} x_n^{(k+1)} &= b_n . \end{aligned}$$

On le construit donc grâce à :

$$\begin{aligned} x_1^{(k+1)} &= \frac{1}{a_{11}} (b_1 - a_{12} x_2^{(k)} - \dots - a_{1n} x_n^{(k)}) \\ x_2^{(k+1)} &= \frac{1}{a_{22}} (b_2 - a_{21} x_1^{(k+1)} - \dots - a_{2n} x_n^{(k)}) \\ &\vdots \\ x_n^{(k+1)} &= \frac{1}{a_{nn}} (b_n - a_{n1} x_1^{(k+1)} - \dots - a_{nn-1} x_{n-1}^{(k+1)}) . \end{aligned}$$

en effet la matrice  $A$  étant définie positive,  $a_{11}, a_{22}, \dots, a_{nn} \neq 0$ .

**Exercice 4. a.** Le vecteur gradient du lagrangien du problème est :

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{u}, \lambda, \mu) = \nabla f(\mathbf{u}) + \sum_{i=1}^p \lambda_i \nabla \varphi_i(\mathbf{u}) + \sum_{j=1}^q \mu_j \nabla \psi_j(\mathbf{u}) .$$



En notant  $\mathbf{c}_i$  et  $\mathbf{d}_j$  la  $i^e$  ligne de la matrice  $C$  et la  $j^e$  ligne de la matrice  $D$ ,  $\nabla\varphi_i(\mathbf{u}) = \mathbf{c}_i^\top$  et  $\nabla\psi_j(\mathbf{u}) = \mathbf{d}_j^\top$ . Aussi :

$$\begin{aligned}\nabla_{\mathbf{x}}\mathcal{L}(\mathbf{u}, \lambda, \mu) &= \nabla f(\mathbf{u}) + \sum_{i=1}^p \lambda_i \mathbf{c}_i^\top + \sum_{j=1}^q \mu_j \mathbf{d}_j^\top \\ &= A\mathbf{u} - \mathbf{b} + C^\top \lambda + D^\top \mu .\end{aligned}$$

b. Les conditions de KKT s'écrivent ici :

$$\begin{aligned}A\mathbf{u} - \mathbf{b} + C^\top \lambda + D^\top \mu &= \mathbf{0} , \\ \mu &\geq \mathbf{0} , \\ \mu^\top (D\mathbf{u} - \mathbf{d}) &= 0 .\end{aligned}$$

c. En utilisant l'expression du vecteur gradient du lagrangien obtenue en a. :

$$\begin{aligned}\mathbf{x}_k &= A^{-1}(\mathbf{b} - C^\top \lambda - D^\top \mu) , \\ \lambda_{k+1} &= \lambda_k + \rho(C\mathbf{x}_k - \mathbf{c}) , \\ \mu_{k+1} &= P_{(\mathbb{R}_+)^q}(\mu_k + \rho(D\mathbf{x}_k - \mathbf{d})) .\end{aligned}$$

## Chapitre V.

**Exercice 1.** On applique le théorème A.4 :

$$\det(A) = 2 \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} - \begin{vmatrix} 1 & 1 \\ 1 & 2 \end{vmatrix} + \begin{vmatrix} 1 & 1 \\ 2 & 1 \end{vmatrix} = 2 \times 3 - 1 + (-1) = 4, \quad \det(A_2) = \begin{vmatrix} 2 & 1 \\ 1 & 2 \end{vmatrix} = 3, \quad \det(A_1) = 2 .$$

Ainsi  $A$  est définie positive.

Prenons  $\omega = (1, 0, 0)$ , alors :

$$C_1 = \frac{\omega_1 \omega_1^\top}{\omega_1^\top A \omega_1} = \frac{1}{2} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} ; \quad D_1 = \text{Id} - C_1 A = \begin{pmatrix} 0 & -1/2 & -1/2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} .$$

Posons  $\mathbf{u} = (0, 2, 0) \notin \ker D_1$  et  $\omega_2 = D_1 \mathbf{u} = (-1, 2, 0)$ . Alors :

$$\frac{\omega_2 \omega_2^\top}{\omega_2^\top A \omega_2} = \frac{1}{6} \begin{pmatrix} 1 & -2 & 0 \\ -2 & 4 & 0 \\ 0 & 0 & 0 \end{pmatrix} ; \quad C_2 = C_1 + \frac{\omega_2 \omega_2^\top}{\omega_2^\top A \omega_2} = \begin{pmatrix} 2/3 & -1/3 & 0 \\ -1/3 & 2/3 & 0 \\ 0 & 0 & 0 \end{pmatrix} .$$

Posons  $v = (0, 0, 3) \notin \ker D_2$  et  $\omega_3 = D_2 v = (-1, -1, 3)$ . Alors,

$$\frac{\omega_3 \omega_3^\top}{\omega_3^\top A \omega_3} = \frac{1}{12} \begin{pmatrix} 1 & 1 & -3 \\ 1 & 1 & -3 \\ -3 & -3 & 9 \end{pmatrix} ; \quad C_3 = C_2 + \frac{\omega_3 \omega_3^\top}{\omega_3^\top A \omega_3} = \begin{pmatrix} 3/4 & -1/4 & -1/4 \\ -1/4 & 3/4 & -1/4 \\ -1/4 & -1/4 & 3/4 \end{pmatrix} = A^{-1} .$$

**Exercice 2.** On commence par déterminer le polynôme  $p(x)$  d'interpolation de Lagrange des points  $(0, 0)$ ,  $(1, 1)$ ,  $(2, 2)$ . On pourrait remarquer que  $x \mapsto x$  interpole ces points et est de degré 1 et donc minimal ; ainsi  $p(x) = x$ . On applique cependant naïvement la formule :

$$\begin{aligned}p(x) &= \sum_{i=1}^3 y_i \prod_{j=1, j \neq i}^3 \frac{x - x_j}{x_i - x_j} \\ &= 1 \times \frac{x - 0}{1 - 0} \times \frac{x - 2}{1 - 2} + 2 \times \frac{x - 0}{2 - 0} \times \frac{x - 1}{2 - 1} \\ &= x(2 - x) + x(x - 1) \\ &= x\end{aligned}$$

L'ensemble des polynômes de degré au plus 5 interpolant ces 3 points est l'ensemble des polynômes :

$$P_{a,b,c}(x) = x + x(x-1)(x-2)(a+bx+cx^2)$$

où  $a, b, c$  décrivent  $\mathbb{R}$ .