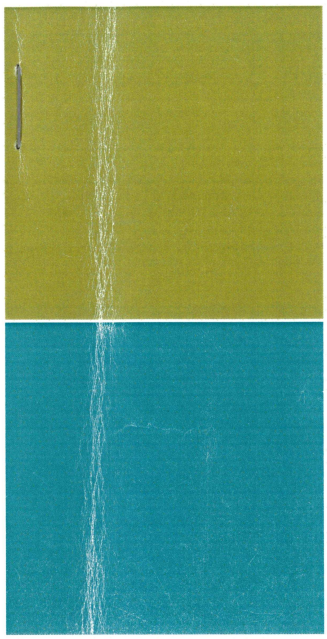


Cours AO101

Optimisation quadratique

École Nationale Supérieure
de **Techniques Avancées**

ENSTA

IP PARIS

Cours AO101
Optimisation quadratique

Wim van Ackooij

(1er mars 2021)

Table des matières

1 Introduction **7**

1.1 Quelques exemples 8

1.2 Plan et objectifs de ce cours 11

2 Existence, Unicité d'un minimum **15**

2.1 Cadre du problème 15

2.2 Existence d'un minimum : résultats généraux 15

2.2.1 Exemple de résolution d'un problème de minimisation 17

2.3 Convexité et unicité 18

2.4 Propriétés des fonctions convexes. 22

3 Conditions nécessaires et suffisantes **25**

3.1 Introduction 25

3.2 La géométrie 26

3.3 Conditions d'optimalité 27

3.3.1 Cas général 27

3.3.2 Cas convexe 30

3.4 Moindres carrés linéaires 33

3.4.1 Problèmatique d'une application 33

3.4.2 Le formalisme abstrait et son étude : pourquoi des carrés ? . . . 35

3.4.3 L'approche directe 36

3.4.4 Une astuce de calcul 38

3.4.5 Existence du point de minimum 38

4 Algorithmes pour problèmes sans contraintes : Fonctionnelle quadratique **41**

4.1 Précisions Numériques. Critères associés à la convergence 42

4.1.1 Test d'arrêt. 43

4.1.2 Evaluer le coût calcul d'une méthode itérative. 43
4.1.3 Stockage mémoire. 45
4.1.4 Comparaison d'algorithmes 45
4.1.5 Observations finales 46
4.2 Conditionnement d'un problème 46
4.2.1 Taux et vitesse de convergence. 48
4.3 Méthodes de descente 48
4.3.1 Relaxation 49
4.3.2 Gradient à pas fixe, à pas optimal 53
4.3.3 Gradient conjugué 57
4.3.4 Extensions 62

5 Conditions nécessaires d'optimalité II 63
5.1 Introduction 63
5.2 Contraintes d'égalité affines 64
5.2.1 Application : fonctionnelle quadratique 66
5.3 Contraintes d'inégalité affines 68
5.4 Contraintes d'égalité et d'inégalité affines 71
5.5 Cas général : conditions de Karush-Kuhn-Tucker* 73
5.5.1 Le Lagrangien - in a nutshell 73

6 Algorithmes pour problèmes contraints 77
6.1 Méthode du gradient projeté. 77
6.2 Méthode d'Uzawa 80
6.3 Techniques de pénalisation 81
6.4 Optimisation par résolution directe des conditions du premier ordre . . 83
6.4.1 Elimination des contraintes - première approche 84
6.4.2 Elimination des contraintes - deuxième approche 85

A Quelques rappels de calcul différentiel 89
A.1 Différentiabilité 89
A.2 Propriétés de la différentielle 95
A.3 Différentielles d'ordre supérieur et formules de Taylor 98
A.3.1 Différentielles d'ordre supérieur 99
A.3.2 Formules de Taylor 101

B Quelques rappels de l'algèbre linéaire 103

B.1 Normes matricielles 103
B.2 Décomposition en valeurs singulières 106
B.3 Méthodes itératives de résolution de systèmes 111
B.3.1 Méthode de Jacobi 113
B.3.2 Méthode de Gauss-Seidel 114

C Chemins et cônes derivables **117**

Chapitre 1

Introduction

L'optimisation est un concept qui fait partie intégrante de la vie courante. Citons quelques exemples tout à fait banals, mais représentatifs :

Quel est le meilleur itinéraire pour aller d'un point A à un point B en voiture ?
Au tennis, comment maximiser l'effet, la vitesse d'une balle de service ?
Peut-on gagner contre la banque à la roulette au casino ?
A la bourse, comment maximiser les profits tout en minimisant les risques ?
Pourquoi tel composant chimique réagit-il avec tel autre ?
etc.

Une stratégie raisonnable est d'essayer de modéliser chacun de ces problèmes, c'est-à-dire de les reformuler sous une forme mathématique, puis de résoudre/optimiser les modèles mathématiques ainsi obtenus, et enfin de tester les résultats sur les situations pratiques... La modélisation, la mise en équations, ne sera que très marginalement étudiée dans ce cours (voir le chapitre 3.4). La modélisation est un travail à part. On peut penser que cette activité est du ressort du physicien, du chimiste, de l'économiste, du joueur, et elle l'est en partie. Toutefois afin d'aboutir à un modèle bien posé, resoluble, potentiellement même efficacement, une interaction entre l'ingénieur-optimiseur et le modeleur (si les deux ne sont pas confondus) est nécessaire. Modéliser, et en fin de compte résoudre, est souvent un aller-retour "incessant" entre modélisation, théorie, algorithmique et implementation numérique. Ce déroulement nécessite méthode, rigueur, et avant tout un besoin de toujours expliciter les raisons qui poussent à faire un certain choix de modélisation plutôt qu'un autre.

L'ingénieur, à qui revient la charge de résoudre ces modèles, se doit de les bien connaître, notamment en ce qui concerne les hypothèses sous lesquelles le modèle est valide, avant d'envisager leur résolution. Dans cette optique, le thème de ce cours est la construction, et la justification mathématique, de méthodes de résolution de ces modèles. Nous

considérerons principalement des modèles simplifiés, que nous nous attacherons à analyser (mathématiquement) en détail. Nous proposerons également des méthodes de résolution approchées, c'est-à-dire leur résolution numérique sur ordinateur. En particulier, nous ferons appel à des outils d'analyse (topologie, calcul différentiel, convexité), mais aussi à de nombreuses branches d'algèbre linéaire. En ce sens, la distinction algèbre/analyse, classique en classes préparatoires, s'estompera.

1.1 Quelques exemples

Exemple 1.1.1 (Télephonie mobile)**.** *Un réseau de téléphonie cellulaire est un ensemble de cellules (géographiques) couvrant un territoire. On cherche à regrouper les cellules en un nombre K de zones, de manière à réaliser un compromis entre la puissance (de calcul) nécessaire pour gérer tous les appels à l'intérieur d'une zone donnée (*paging*), et celle nécessaire pour informer le système central qu'un utilisateur change de zone (*location updating*) . Clairement, si on utilise autant de zones que de cellules (une zone contient une seule cellule), tout l'effort est porté sur le* location updating *; le cas extrême contraire (une seule zone contenant toutes les cellules), requiert une très grande puissance de calcul pour le paging. Mathématiquement, ce problème revient à partitionner un graphe ; il intervient dans de nombreux autres problèmes d'optimisation Pour résoudre le problème, on a les données statistiques suivantes : f_{ij} désigne le coût (en temps de calcul) correspondant au flux moyen d'utilisateurs observés entre la cellule i et la cellule j ; c_i est le coût moyen de paging dans la cellule i (en gros proportionnel au nombre moyen d'appels dans cette cellule). On cherche alors des valeurs X_{ij} valant* 1 *si les cellules i et j sont dans la même zone et* 0 *sinon. Le problème de minimisation du coût s'écrit alors :*

$$\text{Trouver } (X_{ij})_{i,j} \text{ qui minimise } \sum_i \left(\sum_j f_{i,j}(1 - X_{ij}) + X_{ij} c_i \right)$$
$$\text{avec } X_{ij} \in \{0,1\}, \quad 1 \leq \sum_j X_{ij} < K, \quad \text{pour tout } i.$$

Exemple 1.1.2 (Optimisation de portefeuille)**.** *On considère un problème d'optimisation de portefeuille. On suppose qu'on a N actions représentées par des variables aléatoires $R_1, \ldots, R_N$. Chaque action rapporte en moyenne $e_i = E[R_i]$ (E désigne "l'espérance") au bout d'un an. On suppose qu'on investit une somme donnée, et on note $x_i \in \mathbb{R}$ la proportion de la somme investie dans l'action i. Ainsi $1 = \sum_{i=1}^N x_i$. Le portefeuille total est représenté par $R = \sum_{i=1}^N x_i R_i$ et rapporte donc en moyenne le rendement $E[R] = \sum_{i=1}^N x_i e_i$. Le risque du portefeuille est lui modélisé par $\sigma^2(x) =$*

$E[(R - E[R])^2]$, *c'est la mesure de la fluctuation autour du rendement moyen. En notant* $A_{ij} = E(R_i - E(R_i))(R_j - E(R_j))$ *la matrice de covariance des* (R_i), *on trouve l'expression*

$$\sigma^2(x) = \langle x, Ax \rangle .$$

Le but est de trouver la répartition $x = (x_1, \ldots, x_N)^\mathsf{T}$ *minimisant le risque global du portefeuille* $\sigma(x)$ *à rendement au moins égal à* r_0 *(*$r_0 > 0$ *fixé).*

Ainsi le problème mathématique peut s'écrire

$$\min_{x \in K} J(x)$$

avec $J(x) = \frac{1}{2} \langle x, Ax \rangle$ *et* $K = \{x \in \mathbb{R}^N,\ \sum_{i=1}^N x_i = 1 \text{ et } \sum_{i=1}^N x_i e_i \geq r_0\}$.

Exemple 1.1.3 (Identification des paramètres de modèles démographiques)**.** *On dispose de données statistiques contenant en particulier les mesures e la taille de population d'un pays sur une certaine période de temps. On souhaite déterminer une loi d'évolution de la population en fonction du temps. Cela permet en particulier de faire des prévisions sur une évolution future en fonction des données historiques.*

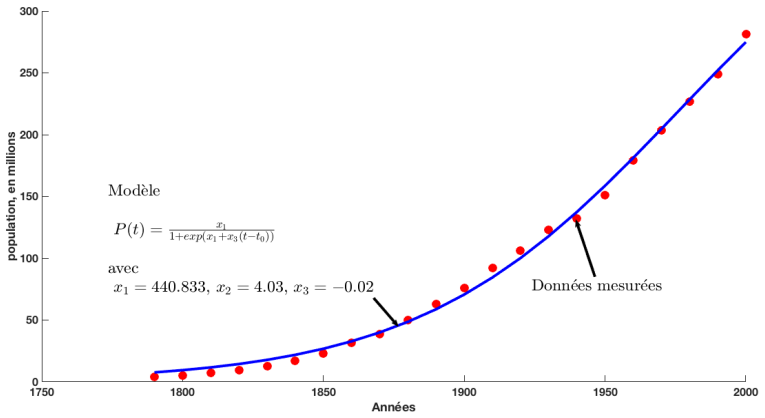


FIGURE 1.1 – Identification du modèle de croissance démographique au sens des moindres carrés

Supposons que les données disponibles sont de la forme (t_i, P_i), $i = 1, \ldots, N$ *où* t_i *sont les années et* P_i *les mesures correspondantes de la taille de la population. Pour identifier la loi d'évolution, on cherche le plus souvent à identifier les paramètres d'une loi-type. Par exemple, une des lois souvent utilisées en sciences démographiques est le modèle logistique*

$$P(t) = \frac{x_1}{1 + e^{x_2 + x_3(t - t_0)}}$$

où

— *t est le temps (en années) (t_0 est la première date de mesure disponible) et $P(t)$ est la taille de la population*
— *x_1 représente l'asymptote à l'infini*
— *x_2 est un paramètre représentant la population initiale, relativement à la taille asymptotique de la population*
— *x_3 est un paramètre de taux de croissance*

Pour déterminer le modèle "le plus fidèle" aux données historiques, on cherche à déterminer le vecteur des paramètres $x^* = (x_1^*, x_2^*, x_3^*) \in \mathbb{R}^3$ qui minimise la la fonctionnelle

$$\inf_{x\in\mathbb{R}^3} J(x)$$

qui représente l'erreur quadratique de représentation de l'ensemble des points $(t_i, P_i),\ i = 1, \ldots, N$ par la fonction $P(t)$:

$$J(x) = \sum_{i=1}^{N} \|P(t_i, x) - P_i\|^2.$$

Sur la figure 1.1 sont représentées (en rouge) les mesures de la taille de la population des Etats Unis de 1790 à 2000 (d'après http://www.census-charts.com/Population/pop-us-1790-2000.html). La courbe bleue représente le modèle avec les paramètres optimaux déterminés à partir de ces données.

Exemple 1.1.4 (Planification de production d'énergie)**.** *On considère un réseau de production énergétique formé de K centrales thermiques. Le réseau doit fournir une puissance commune supérieure à une demande donnée D en MW. On notera la puissance fournie par chaque station i, P_i. Chaque station peut produire une puissance en respectant les limites qui lui sont propres :*

$$P_i^{min} \leq P_i \leq P_i^{max},\ i = 1, \ldots, K.$$

Le coût, en Euros, de production de chaque station associé à une puissance P est défini par le fonction

$$C_i(P) = a_i P^2 + b_i P + c_i,\ i = 1, \ldots, K.$$

où $(a_i, b_i, c_i),\ i = 1, \ldots, K.$ sont les coefficients donnés, propres à chaque station. La production d'une puissance P provoque l'émission de polluants, mesurée en kg, par la fonction

$$E_i(P) = d_i P^2 + e_i P + f_i,\ i = 1, \ldots, K.$$

où $(d_i, e_i, f_i)),\ i = 1, \ldots, K.$ sont les coefficients donnés, propres à chaque station. Le gestionnaire du réseau doit répartir les objectifs de production $P = (p_1, \ldots, P_K)$ entre

les stations du réseau de façon à ce que le réseau puisse satisfaire la demande D. Deux problèmes d'optimisation peuvent être considérés pour déterminer le vecteur P.

Optimisation des coûts de production

$$\inf_{P\in\mathbb{R}^K} \sum_{i=1}^{K} C_i(P_i)$$

$$P_i^{min} \leq P_i \leq P_i^{max},\ i = 1, \ldots, K.$$
$$\sum_{i=1}^{K} P_i \geq D$$

Optimisation des émissions polluantes

$$\inf_{P\in\mathbb{R}^K} \sum_{i=1}^{K} E_i(P_i)$$

$$P_i^{min} \leq P_i \leq P_i^{max},\ i = 1, \ldots, K.$$
$$\sum_{i=1}^{K} P_i \geq D$$

Exemple 1.1.5 (Problème de spectroscopie)**.** *On cherche la concentration respective de n produits dans un mélange gazeux (ce problème intervient par exemple, lorsqu'on cherche à mesurer la "qualité" de l'air Parisien). Pour cela, on éclaire le mélange à diverses longueurs d'ondes λ_i, $i = 1, \ldots, m$, et on mesure (par spectroscopie) l'intensité correspondante, que l'on collecte dans un vecteur $b \in \mathbb{R}^m$. En laboratoire, on a mesuré les coefficients d'absorption a_{ij} correspondant à chaque gaz j et à chaque longueur d'onde i, d'où une matrice $A = (a_{ij}) \in \mathbb{R}^{n\times m}$.*

En notant X_i la concentration du i^e produit, le problème mathématique s'écrit :

$$\textit{Trouver } X \in [0,1]^n,\ \|AX - b\| = \min_{y\in[0,1]^n} \|Ay - b\|.$$

1.2 Plan et objectifs de ce cours

L'objectif de ce cours est de donner un aperçu à la fois théorique et pratique d'une partie du domaine de l'optimisation. Les exemples cités ci-dessus montrent l'implication de cette branche des mathématiques dans différents domaines (physique, finance, économie, ... etc). Si le travail de l'ingénieur commence par la modélisation et la compréhension du problème posé, il se prolonge naturellement par l'étude mathématique

du cadre permettant d'analyser les modèles ; on s'intéresse alors à l'existence et à l'unicité ou la multiplicité des solutions, à leur caractérisation et à toutes autres propriétés qualitatives.

En pratique, lorsque l'on résout un problème d'optimisation, on utilise des algorithmes permettant d'approcher numériquement la solution d'un problème du type

$$\text{Trouver } u \in K, \text{ tel que } J(u) = \inf_{v \in K} J(v)$$

ou bien

$$\text{Trouver } u \in K, \text{ tel que } J(u) = \sup_{v \in K} J(v),$$

où J est une fonctionnelle définie sur un ensemble K non vide, à valeurs dans $\mathbb{R}$. Avant d'envisager l'utilisation d'un algorithme, il est naturel [1] de répondre aux questions ci-dessous :

(i) Que veut dire une solution ?
(ii) **Existe**-t-il une solution u ? Est-elle **unique** ?
(iii) Comment la **caractériser** ?
(iv) Quel(s) **algorithme(s)** permet(tent) de calculer la solution ?
(v) Quel est alors l'algorithme le plus **efficace** ?

Le plan de ce cours est le suivant. Le chapitre 2 est consacré aux questions d'existence et d'unicité du ou des minima. Nous nous attachons en particulier à l'étude de problèmes posés non pas sur l'espace entier, mais plutôt sur une partie de celui-ci ; on parle alors de problème de minimisation avec contraintes.

Dans le chapitre 3, nous analysons les conditions d'optimalité. Là encore, nous nous interesserons aux cas de problèmes avec contraintes, et en particulier, aux cas de contraintes d'égalités ou inégalités affines.

Dans le chapitre suivant, nous étudions un problème classique de minimisation, appelé moindres carrés linéaires, qui peut être perçu comme une généralisation de la résolution d'un système linéaire.

Enfin, dans le dernier chapitre, nous construirons des algorithmes permettant de calculer numériquement une approximation du minimum. Deux points-clefs sont à noter dès à présent au sujet de ces algorithmes :

> Ils sont basés sur les caractérisations obtenues dans les chapitres théoriques.
> Ils sont itératifs : à partir d'une initialisation u^0, on calcule u^1, puis u^2, etc. jusqu'à arriver à une solution numérique correcte.

1. Même si cette procédure n'est pas toujours respectée en pratique...

Dans l'Annexe, nous rappelons les notions élémentaires, ainsi que les théorèmes fondamentaux, associés à la différentiabilité d'une fonctionnelle définie sur un espace vectoriel normé, à valeurs dans un espace vectoriel normé.

Bien sûr, la recherche en optimisation est très dynamique, et la théorie en constante évolution. Aussi, les résultats présentés ci-après ne représentent qu'une petite introduction à l'art de l'optimisation. Des généralisations et approfondissements seront proposés lors d'autres cours de l'ENSTA, en deuxième (cf. [8]) et troisième années.
Nous renvoyons également le lecteur aux ouvrages [1, 4], qui proposent de nombreuses extensions, tout en restant tout à fait abordable pour le (futur) ingénieur.

L'approfondissement peut aussi se penser en consultant par exemple [2].

Chapitre 2

Existence, Unicité d'un minimum

2.1 Cadre du problème

Dans toute la suite, et sauf indication contraire, $\mathbb{V}$ désignera toujours l'espace vectoriel normé $\mathbb{R}^n$, K un sous-ensemble non vide de $\mathbb{V}$ et J une fonctionnelle continue définie sur K, à valeurs dans $\mathbb{R}$. Considérons le problème d'optimisation suivant :

$$\text{Trouver } u \in K, \text{ tel que } J(u) = \inf_{v \in K} J(v).$$

Dans ce problème, il ne s'agit pas seulement de vérifier que $\inf_{v\in K} J(v) \in \mathbb{R}$, mais aussi que cette valeur inférieure est atteinte par un (voire plusieurs) point u de K.

2.2 Existence d'un minimum : résultats généraux

Commençons d'abord par quelques définitions.

Définition 2.2.1. *$u \in K$ est un* **point de minimum local** *de J sur K si, et seulement si*

$$\exists \eta > 0, \quad \forall v \in K, \quad \|v - u\| < \eta \Longrightarrow J(u) \leq J(v).$$

$u \in K$ est un **point de minimum global** *de J sur K si, et seulement si*

$$\forall v \in K, \quad J(u) \leq J(v).$$

Définition 2.2.2. *On dit qu'une suite $(u_k)_{k\in\mathbb{N}}$ d'éléments de K est une* **suite minimisante** *si, et seulement si,*

$$\lim_{k\to+\infty} J(u_k) = \inf_{v\in K} J(v).$$

Remarque 2.2.1. *Par définition de la notion d'infimum, il existe toujours des suites minimisantes !*

Intéressons nous maintenant à la question d'existence de minima, et rappelons d'abord le théorème suivant, bien connu.

Théorème 2.2.1. *Si K est compact, non-vide et J est continue sur K, alors J atteint ses extréma :*

$$\exists (u_{min}, u_{max}) \in K \times K, \text{ tels que } J(u_{min}) = \inf_{v \in K} J(v), \ J(u_{max}) = \sup_{v \in K} J(v).$$

On peut établir une variante du théorème 2.2.1, valable lorsque $\mathbb{V}$ est de dimension *finie*. Ce résultat est fort utile si K est non compact.

Définition 2.2.3. *On dit qu'une fonctionnelle J est infinie à l'infini dans K si, et seulement si,*

$$\text{pour toute suite } (v_n)_n \subset K, \quad \lim_{n \to +\infty} \|v_n\| = +\infty \Longrightarrow \lim_{n \to +\infty} J(v_n) = +\infty. \tag{2.1}$$

Théorème 2.2.2. *Dans le cas où $\mathbb{V} = \mathbb{R}^n$, si K est un fermé (et non-vide), et si J est continue et infinie à l'infini dans K, alors elle admet un minimum global sur K. De plus, de toute suite minimisante, on peut extraire une sous-suite qui converge vers un point de minimum.*

Démonstration. Soit $(u_k)_k$ une suite minimisante.

- $(u_k)_k$ est bornée : en effet, supposons qu'il existe une sous-suite extraite, $(u_{k'})_{k'}$, telle que $\|u_{k'}\| \to +\infty$; comme J est infinie à l'infini, on infère que $J(u_{k'}) \to +\infty$, ce qui contredit le fait que (u_k) est une suite minimisante (en particulier, $\lim_k J(u_k) < +\infty$.)
- Comme $(u_k)_k$ est bornée, on peut en extraire une sous-suite, toujours notée $(u_{k'})_{k'}$, qui converge vers un point u. C'est une suite d'éléments de K qui est fermé, donc $u \in K$.
- Par ailleurs, comme J est continue, $\lim_{k'} J(u_{k'}) = J(\lim_{k'} u_{k'}) = J(u)$. Enfin, la sous-suite $(u_{k'})_{k'}$ est minimisante ; ainsi $J(u) = \inf_{v \in K} J(v)$, et u est un minimum global de J sur K.

□

Remarque 2.2.2. *La propriété "infinie à l'infini dans K" assure que toute suite minimisante de J est bornée. Il est important de noter que cette propriété est automatiquement vérifiée si K est borné, on retrouve ainsi le résultat classique du théorème 2.2.1. Il est aussi évident que* (2.1) *est vraie, si, et seulement si,*

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

Remarque 2.2.3 (*)**.** *Lorsque la dimension de $\mathbb{V}$ est infinie, la proposition précédente est fausse ! On peut en effet construire des contre-exemples, lorsque la dimension est infinie.*

Nous allons expliquer pourquoi la démonstration de la proposition ne s'applique pas dans un espace de dimension infinie. Pour cela, rappelons un théorème dû à Riesz.

Théorème 2.2.3. *Soit $\mathbb{V}$ un espace vectoriel normé et $B(0,1) = \{v \in \mathbb{V} \ : \ \|v\| \leq 1\}$ sa boule unité fermée. Alors, $\mathbb{V}$ est de dimension finie si, et seulement si, $B(0,1)$ est compacte.*

A partir de ce résultat, on voit qu'il ne sert à rien de se ramener à une suite bornée, si l'on reprend la démonstration dans le cas de la dimension infinie. En effet, les éléments de la suite appartiennent bien à une boule fermée et bornée, mais celle-ci n'est plus compacte. On ne peut alors plus considérer une sous-suite qui converge...

— Il est également indispensable que l'ensemble K soit *fermé*. Si on considère par exemple la fonction $x \mapsto x^2$ sur $K = \mathbb{R}_\star^+$, on a bien une fonction continue, infinie à l'infini, définie sur K non vide, mais K n'est pas fermé... Elle n'admet pas de point de minimum sur K.

— Notons aussi que la condition (2.1) n'assure pas l'existence d'un maximum. Cependant, il n'est pas difficile maintenant d'énoncer un résultat d'existence du maximum sous une hypothèse semblable à (2.1). Ce point est laissé en exercice au lecteur.

2.2.1 Exemple de résolution d'un problème de minimisation

Dans cette section, nous considérons le problème de minimisation "classique" du polynôme $P(x) = \alpha\, x^2 - \beta\, x + \gamma$ sur $\mathbb{R}$. Ce problème est simple mais instructif.

Par définition, si x_0 est un minimum local de P, il existe $\eta > 0$ tel que, pour tout h vérifiant $|h| < \eta$, on ait $P(x_0 + h) \geq P(x_0)$. Par différence, on obtient $h(2\alpha\, x_0 + \alpha\, h - \beta) \geq 0$.

Si on choisit h dans $]0, \eta[$, on a alors $2\alpha\, x_0 + \alpha\, h - \beta \geq 0$; on fait tendre h vers 0, pour arriver à $2\alpha\, x_0 - \beta \geq 0$.

En prenant h négatif, on obtient cette fois $2\alpha\, x_0 - \beta \leq 0$.

Ainsi, une condition *nécessaire* d'existence de minimum est que

$$2\alpha\, x_0 = \beta. \tag{2.2}$$

Réciproquement, si x_0 est tel que $2\alpha\, x_0 = \beta$, on trouve $P(x_0 + h) = P(x_0) + \alpha\, h^2$. Pour garantir l'existence d'un minimum (qui sera d'ailleurs global), α doit être positif ou nul. Notons enfin que pour que (2.2) possède une solution, il faut soit $\alpha \neq 0$, soit $\alpha = \beta = 0$. Dans le premier cas, il existe une solution et une seule, et dans le second cas, x_0 est quelconque.

En conclusion, nous sommes arrivés au résultat suivant :

(I) $\alpha < 0$: la condition (2.2) n'est pas *suffisante* pour déterminer le minimum et d'ailleurs il n'existe pas de minimum.

(II) $\alpha \geq 0$: la condition (2.2) permet de calculer le minimum lorsqu'il existe.

Si $\alpha > 0$, il existe un minimum x_0 unique, égal à $x_0 = \beta/2\,\alpha$.

Si $\alpha = 0$ et $\beta = 0$, tout élément de $\mathbb{R}$ réalise le minimum.

Si $\alpha = 0$ et $\beta \neq 0$, il n'existe pas de minimum.

Le cas $\alpha \geq 0$ correspond à une fonction $P(x)$ convexe, notion que nous abordons à la section suivante.

2.3 Convexité et unicité

Une catégorie très importante parmi les fonctionnelles est celle des fonctionnelles convexes, pour lesquelles on peut obtenir des informations sur l'ensemble des minima. Nous le verrons au chapitre suivant, que nous pouvons aussi obtenir, une caractérisation de ces minima. En effet, lorsque la fonctionnelle J est convexe, le minimum, qui *à priori* peut-être *local*, devient *global* et même dans certains cas unique.

Définition 2.3.1. *On dit qu'un sous-ensemble K de $\mathbb{V}$ est* **convexe** *si, et seulement si, pour tout couple d'éléments (u, v), le segment $[u, v]$ est inclus dans K : $\forall u, v \in K$, $\forall t \in [0, 1]$, $u + t(v - u) = (1 - t)u + t\, v \in K$.*

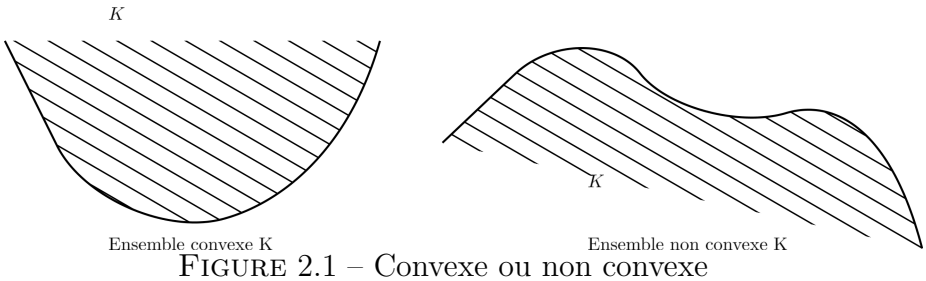


FIGURE 2.1 – Convexe ou non convexe

On définit aussi une fonction convexe de la manière suivante :

Définition 2.3.2. *Soit J une fonctionnelle définie sur un sous-ensemble convexe non vide K de $\mathbb{V}$, à valeurs dans $\mathbb{R}$. On dit que J est* **convexe** *si et seulement si*

$$\forall u, v \in K, \quad u \neq v, \quad \forall \theta \in]0,1[\qquad J(\theta u + (1-\theta)v) \leq \theta J(u) + (1-\theta)J(v).$$

Dans le cas d'une inégalité stricte, on dit que la fonctionnelle J est **strictement convexe**.
Enfin, s'il existe $\alpha > 0$ tel que

$$\forall u, v \in K, \quad u \neq v, \quad \forall \theta \in]0,1[\qquad J(\theta u+(1-\theta)v) \leq \theta J(u)+(1-\theta)J(v)-\frac{\alpha}{2}\theta(1-\theta)\|u-v\|^2,$$

nous dirons que J est **α-convexe**.

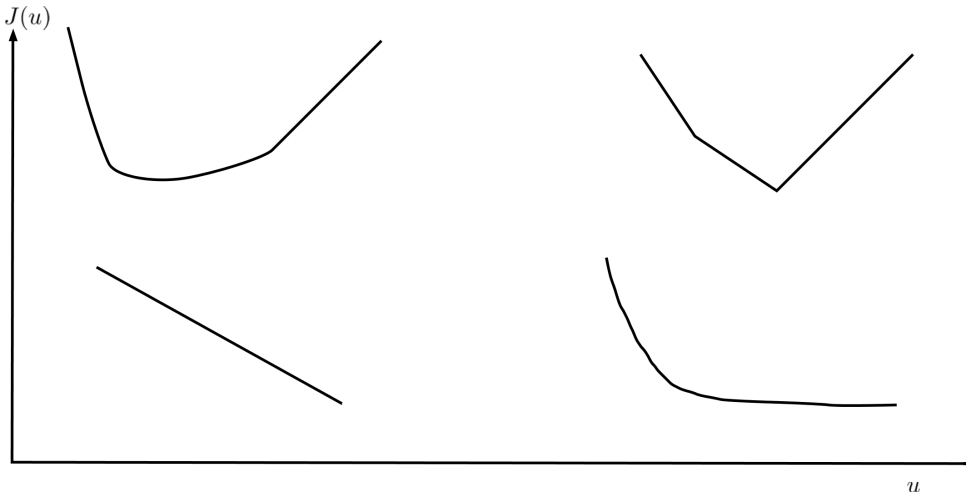


FIGURE 2.2 – Exemples de fonctions convexes

Remarque 2.3.1. *(***géométrique***) La convexité de J signifie que le graphe de J est en-dessous de toutes ses cordes. (Voir Fig.2.3).*

Exercice 2.3.1. *Montrer que si J est α-convexe et différentiable en un point, alors elle est infinie à l'infini (NB. On ne fait aucune hypothèse de continuité sur J.)*

Nous allons maintenant établir un premier résultat sur les fonctionnelles convexes, à savoir que tout point de minimum local est en fait un point de minimum global. Commençons par la proposition suivante.

Proposition 2.3.1. *Soit J une fonctionnelle convexe définie sur un convexe non vide K :*

1. *si u et v sont deux points de minimum locaux, alors $J(u) = J(v)$;*

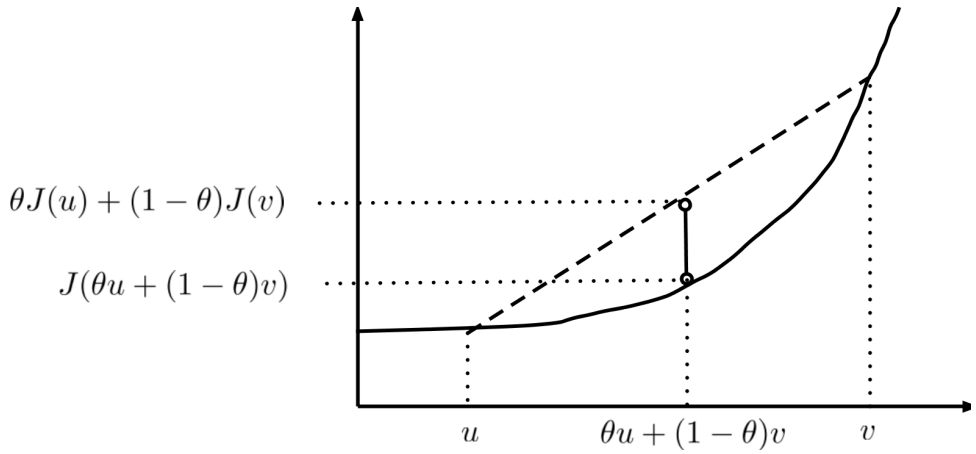


FIGURE 2.3 – Remarque 2.3.1

2. *si de plus J est strictement convexe, alors $u = v$.*

Démonstration. 1. Soient u et v deux minima. Comme K est convexe, $u + \theta(v - u) \in K$ pour tout $\theta \in]0, 1[$. De plus $\exists \theta_0$, t.q. pour tout $\theta \in]0, \theta_0]$, $J(u) \leq J(u + \theta(v - u))$. Comme J est convexe, on obtient, pour tout $\theta \in]0, \theta_0]$,

$$J(u) \leq J(u + \theta(v - u)) = J((1 - \theta)u + \theta\, v) \leq (1 - \theta)J(u) + \theta\, J(v),$$

et on en déduit facilement que $J(u) \leq J(v)$. De même on montre que $J(u) \geq J(v)$. On a donc bien $J(u) = J(v)$.

2. Supposons que J soit strictement convexe. Si u et v sont deux points de minimum, on a vu que $J(u) = J(v)$. Si u et v sont distincts,

$$J(u + \theta(v - u)) < (1 - \theta)J(u) + \theta\, J(v) = J(u), \quad \forall \theta \in]0, 1[,$$

ce qui contredit le fait que u est un minimum local.
On a donc bien $u = v$.

□

Nous en déduisons

Théorème 2.3.1. *Soit J une fonctionnelle convexe définie sur un convexe non vide K :*

1. *tout point de minimum local est un point de minimum* global *;*
2. *si de plus J est strictement convexe, le point de minimum, s'il existe, est* unique.

Démonstration. 1. On reprend le raisonnement ci-dessus avec u minimum local et $v \in K$ quelconque. On obtient que $J(u) \leq J(v)$. Ainsi u est un minimum global.

2. Immédiate.

□

Bien entendu, le théorème précédent donne un cadre (stricte convexité) où le minimum, lorsqu'il existe, est unique. Cependant il existe des fonctions qui ne sont pas strictement convexe et qui pourtant n'admette qu'un seul minimum (voir ci-dessous).

Exemple 2.3.1 (Projection sur un convexe fermé)**.** *Soit K une partie convexe non vide et fermée de $\mathbb{R}^n$ et soit $w \in \mathbb{R}^n$. Considérons le problème*

$$\text{Trouver } u \in K, \quad \|w-u\| = \min_{v\in K} \|w-v\|. \tag{2.3}$$

($\|\cdot\|$ désigne la norme Euclidienne.) Posons $J : v \in K \longmapsto \|v-w\|$. Il est clair que l'application J est continue. De plus $J(v) \geq |\|v\| - \|w\||$, donc

$$\lim_{v\in K, \|v\|\to+\infty} J(v) = +\infty,$$

*ce qui prouve que J est "*infinie à l'infini*". L'ensemble K étant fermé et non vide, on conclut que J atteint le minimum sur K. Il nous reste à prouver que la solution est unique. Remarquons d'abord que la fonction J est convexe sans être strictement convexe, on ne peut donc pas utiliser directement le théorème 2.3.1 pour prouver l'unicité.*

Soient u_1 et u_2 deux solutions de (2.3)*. Puisque K est convexe et J est convexe, alors $\frac{u_1+u_2}{2}$ est aussi une solution de* (2.3)*, et on a :*

$$\min_{v\in K} \|v-w\| = J(u_1) = J(u_2) = J(\frac{u_1+u_2}{2}).$$

D'autre part, on a :

$$\begin{aligned} \| u_1-u_2 \|^2 &= 2(\| u_1-w \|^2 + \| u_2-w \|^2) - 4 \| \frac{u_1+u_2}{2} - w \|^2 \\ &= 2([J(u_1)]^2 + [J(u_2)]^2) - 4[J(\frac{u_1+u_2}{2})]^2 = 0. \end{aligned}$$

On a donc $u_1 = u_2$, ce qui prouve bien l'unicité de la solution qu'on notera u.

Ainsi on a prouvé que pour tout $w \in \mathbb{V}$, il existe $u \in K$ tel que $\|w-u\| = \min_{v\in K} \|w-v\|$. L'élément u est appelé projection de w sur le convexe fermé K, on le notera $P_K(w) := u$. De plus on a :

Proposition 2.3.2. *Si K est un convexe fermé, alors tout élément w de $\mathbb{V}$ admet une projection unique $P_K(w)$ sur K. De plus l'application $w \longmapsto P_K(w)$ est contractante.*

Démonstration. Il reste à prouver que P_K est contractante. □

2.4 Propriétés des fonctions convexes.

Théorème 2.4.1. *Soit J une fonctionnelle différentiable sur un sous-ensemble K convexe non vide.*
Les assertions suivantes sont équivalentes.

(i) J est convexe sur K.

(ii) $\forall u, v \in K,\ u \neq v,\ J(v) \geq J(u) + \langle \nabla J(u), v - u \rangle$.

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

De même, les assertions suivantes sont équivalentes.

(iv) J est strictement convexe sur K.

(v) $\forall u, v \in K,\ u \neq v,\ J(v) > J(u) + \langle \nabla J(u), v - u \rangle$.

(vi) $\forall u, v \in K,\ u \neq v,\ \langle \nabla J(u) - \nabla J(v), u - v \rangle > 0$.

Démonstration. Montrons d'abord que $(i) \Longrightarrow (ii)$. Pour cela supposons que J est convexe sur K, et prenons u et v deux éléments distincts de K. Pour tout $\theta \in]0,1[$, on a :

$$J(u + \theta(v-u)) = J(\theta v + (1-\theta)u) \leq \theta J(v) + (1-\theta)J(u).$$

Ce qui implique, pour $\theta \in]0,1[$, l'inégalité suivante :

$$\frac{J(u + \theta(v-u)) - J(u)}{\theta} \leq J(v) - J(u).$$

Par passage à la limite lorsque $\theta \to 0^+$, on obtient : $\langle \nabla J(u), v - u \rangle \leq J(v) - J(u)$, et donc (ii).

Supposons maintenant que (ii) est satisfaite est prouvons que (iii) est alors aussi vérifiée. Pour cela, prenons $u, v \in K$. De (ii), on a :

$$\begin{aligned} J(v) &\geq J(u) + \langle \nabla J(u), v - u \rangle , \\ J(u) &\geq J(v) + \langle \nabla J(v), u - v \rangle . \end{aligned}$$

En additionnant les deux inégalités, on obtient bien le résultat (iii).

Reste maintenant à prouver $(iii) \Rightarrow (i)$. Soient $u, v \in K$, et $\theta \in]0,1[$. Considérons la fonction $\mu : t \in]0,1[\longmapsto J(v + t\theta(u-v))$. Comme J est différentiable, la fonction μ est de classe C^1 et $\mu'(t) = \theta \langle \nabla J(v + t\theta(u-v)), u - v \rangle$. La formule de Taylor-Mac-Laurin autour de 0 donne

$$\exists \lambda_1 \in]0,1[, \quad \mu(1) = \mu(0) + \mu'(\lambda_1),$$

ou encore,

$$\begin{aligned} \exists \lambda_1 \in]0,1[, \quad J(\theta u + (1-\theta)v) &= J(v + \theta(u-v)) \\ &= J(v) + \lambda_1 \theta \langle \nabla J(v + \lambda_1 \theta(u-v)), u - v \rangle . \end{aligned} \quad (2.4)$$

Avec le même raisonnement, on a : $\exists \lambda_2 \in]0,1[$ tel que

$$\begin{aligned} J(\theta u + (1-\theta)v) =& J(u + (1-\theta)(v-u)) \\ =& J(u) + \lambda_2(1-\theta)\left\langle \nabla J(u + \lambda_2(1-\theta)(v-u)), v-u \right\rangle . \end{aligned} \tag{2.5}$$

En multipliant (2.4) par $(1-\theta)$ et (2.5) par θ, et en additionnant les égalités obtenues, on arrive à :

$$\begin{aligned} J(\theta u + (1-\theta)v) = \theta J(u) + (1-\theta)J(v) + \\ \theta(1-\theta)\left\langle \nabla J(v + \lambda_1\theta(u-v)) - \nabla J(u + \lambda_2(1-\theta)(v-u)), u-v \right\rangle . \end{aligned} \tag{2.6}$$

Posant $w_1 := v + \lambda_1\theta(u-v)$, $w_2 := u + \lambda_2(1-\theta)(v-u)$, on trouve $w_2 - w_1 = (1 - \lambda_1\theta - \lambda_2(1-\theta))(u-v)$. Notons d'abord que du fait que $\lambda_1, \lambda_2 \in]0,1[$, le coefficient $\gamma = 1 - \lambda_1\theta - \lambda_2(1-\theta) > 0$. D'autre part, de (iii), on a :

$$\left\langle \nabla J(w_1) - \nabla J(w_2), u-v \right\rangle = \frac{1}{\gamma}\left\langle \nabla J(w_1) - \nabla J(w_2), w_2 - w_1 \right\rangle \leq 0.$$

Avec (2.6) on conclut que :

$$J(\theta u + (1-\theta)v) \leq \theta J(u) + (1-\theta)J(v).$$

c.a.d. la convexité de J. □

Remarque 2.4.1. *(**géométrique**) La convexité de J (point (ii)) signifie que le graphe de J est au-dessus du graphe de l'application affine tangente à J en u (c'est-à-dire l'application $v \mapsto J(u) + \langle \nabla J(u), v-u \rangle$), en tout point u de K (voir Fig. 2.4).*

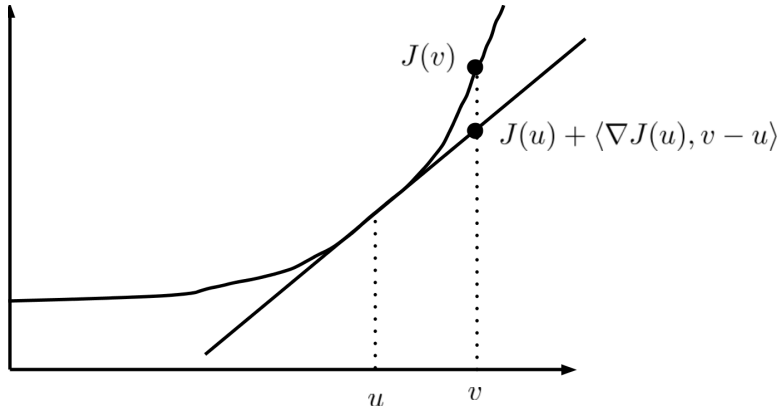


FIGURE 2.4 – Remarque 2.4.1

Remarque 2.4.2. *En ce qui concerne l'α-convexité, on peut prouver les équivalences ci-dessous. Soit J une fonctionnelle différentiable sur un sous-ensemble K.*
Les assertions suivantes sont équivalentes.

(i) *J est α-convexe sur K.*
(ii) $\forall u, v \in K,\ u \neq v,\ J(v) \geq J(u) + \langle \nabla J(u), v-u \rangle + \frac{\alpha}{2}\|u-v\|^2$.
(iii) $\forall u, v \in K,\ u \neq v,\ \langle \nabla J(u) - \nabla J(v), u-v \rangle \geq \alpha \|u-v\|^2$.

Notons aussi le résultat suivant :

Théorème 2.4.2. *Soit J une fonctionnelle de $\mathbb{V}$ dans $\mathbb{R}$, de classe $\mathcal{C}^2$. Alors J est convexe si, et seulement si,*

$$\forall u, d \in \mathbb{V}, \quad \left\langle \nabla^2 J(u)d, d \right\rangle \geq 0. \tag{2.7}$$

Démonstration. Supposons que J est convexe. Soient $u, d \in \mathbb{V}$, et $\theta \in]0,1[$. Du théorème 2.4.1, on a :

$$0 \leq \frac{\langle \nabla J(u+\theta d) - \nabla J(u), d \rangle}{\theta}.$$

En faisant tendre $\theta \to 0$, il vient :

$$0 \leq \left\langle \nabla^2 J(u)d, d \right\rangle.$$

Réciproquement, supposons que (2.7) est satisfaite. De la formule de Taylor-Mac-Laurin, on a :

$$\exists \lambda \in]0,1[, \quad J(v) = J(u) + \langle \nabla J(u), v-u \rangle + \frac{1}{2}\left\langle \nabla^2 J(u+\lambda(v-u))(v-u), v-u \right\rangle.$$

Or de (2.7), on obtient :

$$\frac{1}{2}\left\langle \nabla^2 J(u+\lambda(v-u))(v-u), v-u \right\rangle \geq 0,$$

et par suite,

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

Ce qui, d'après le théorème 2.4.1, conclut que J est convexe. □

Pour finir, notons aussi que pour une fonctionnelle convexe, on a des résultats généraux concernant sa *régularité* (nous renvoyons le lecteur à [8]).

Chapitre 3

Conditions nécessaires et suffisantes

3.1 Introduction

Dans ce chapitre on considère les fonctionnelles dérivables au sens de Gateaux (cf. l'Annexe), sauf mention explicite du contraire. Dans l'ensemble de notre chapitre, la fonctionnelle J est considérée comme une fonction $J : \mathbb{R}^n \to \mathbb{R}$, c.à.d., $\mathbb{V} = \mathbb{R}^n$ (Pour des extensions vers des espaces d'Hilbert, ou Banach, on se réfère à [5, 10], mais également [12].

Nous nous intéressons ici aux **conditions d'optimalité** d'une fonctionnelle J sur un ensemble non vide K. Nous allons étudier d'abord le cas général, poursuivi par celui d'un sous-ensemble K convexe.

Notre intérêt se porte tout d'abord sur la caractérisation d'un minimum ou plus généralement une solution du problème :

$$\min_{u \in \mathbb{V}} J(u),$$

où l'on suppose par exemple que ce problème admet une solution $\bar{u}$ locale. Nous pouvons donc trouver un voisinage U de $\bar{u}$, de sorte à ce que $J(v) \geq J(\bar{u})$ pour tout $v \in U$.

Prenons h, assez petit pour que $u = \bar{u} + h \in U$ et considérons le développement limité :

$$J(\bar{u}) \leq J(u) = J(\bar{u}) + \langle \nabla J(\bar{u}), h \rangle + o(\|h\|),$$

où $o(\|h\|) \to 0$ lorsque $h \to 0$. Nous en déduisons $\langle \nabla J(\bar{u}), h \rangle \geq 0$, et ce, quelque soit h. Donc nous pouvons conclure $\nabla J(\bar{u}) = 0$.

La condition que nous venons de déduire est classique pour l'optimisation sans contraintes. La question suivante immédiate est celle de la suffisance. Hélas, là notre espoir se montre aussitôt vain :

Exemple 3.1.1. *Considérez la fonctionnelle $J(u) = -u^2$ sur $\mathbb{V} = \mathbb{R}$ et observez que $\nabla J(0) = J'(0) = 0$, mais que $\bar{u} = 0$ n'est pas un minimum.*

3.2 La géométrie

En commençant par un exemple, réfléchissons à comment la présence de contraintes modifie notre perception des conditions d'optimalité.

Exemple 3.2.1. *Considérons le problème*

$$\min_{u\in\mathbb{R}}(u-2)^2.$$

Il est évident que la solution optimale (globale) se trouve en $\bar{u} = 2$, où $\nabla J(\bar{u}) = 0$.
Si l'on modifie le problème de la manière suivante :

$$\min_{u\in(-\infty,1]}(u-2)^2,$$

l'on se rend compte que dans la solution $\bar{u} = 1$, nous avons $\nabla J(\bar{u}) \neq 0$. Il convient donc de tenir compte des contraintes, où plutôt de la géometrie des contraintes pour formuler une condition d'optimalité utile.

A base de l'exemple précédente, on observe que ce qui est important n'est pas tant en soi le fait que J puisse descendre en $1 + h$ pour $h > 0$, mais que J monte en $1 - h$ pour $h > 0$. La fonctionnelle croît selon les directions qui permettent de rester dans $K = (-\infty, 1]$. Le comportement de la fonctionnelle dans les directions qui mènent en dehors de l'ensemble K n'est pas très intéressante finalement. En effet, imaginez toute modification de J sur $(1, \infty)$ qui en préserve la nature continûment differentiable et sa spécification sur K ! Le minimum n'en sera pas affecté.

La question se pose alors sur comment formaliser l'idée de directions restant dans l'ensemble K. En particulier il faut avoir en tête que K puisse être un ensemble assez "sauvage". La notion suivante formalise cette idée d'une certaine manière (il en existe d'autres, voir Appendice C.

Définition 3.2.1. *Soit $K \subseteq \mathbb{R}^n$ un ensemble fermé et $\bar{u} \in K$ donné. Le cône tangente (de Bouligand / Severi) de K dans $\bar{u}$ est défini comme suit :*

$$\mathcal{T}_K(\bar{u}) := \left\{ d \in \mathbb{V} \ : \ \exists t_n \downarrow 0, \ \exists u_n \in K, u_n \to \bar{u}, t_n^{-1}(u_n - \bar{u}) \to d \right\}. \tag{3.1}$$

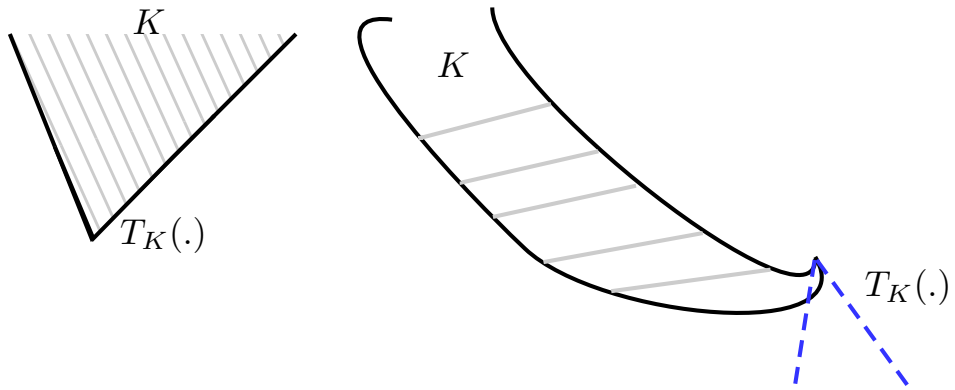


FIGURE 3.1 – Exemples de cônes de directions admissibles

3.3 Conditions d'optimalité

3.3.1 Cas général

Théorème 3.3.1. *Soient K un sous-ensemble non vide de $\mathbb{V}$, u un point de K et J une fonctionnelle de K à valeurs dans $\mathbb{R}$. On suppose que J est Fréchet-différentiable en u. Si u est un point de minimum local de J sur K, on a nécessairement*

$$\langle \nabla J(u), w \rangle \geq 0, \qquad \forall w \in \mathcal{T}_K(u). \tag{3.2}$$

Démonstration. Le fait que u est un minimum local de J sur K veut dire qu'il existe un voisinage U de u, tel que

$$J(u) \leq J(v) \; \forall v \in U \cap K.$$

Pour un $w \in \mathcal{T}_K(u)$ quelconque, on peut trouver par définition, une suite $t_n \downarrow 0$, $u_n \in K$, $u_n \to u$, $t_n^{-1}(u_n - u) \to w$. En particulier pour n suffisamment grand, $u_n \in U$. Par conséquence en posant $w_n := t_n^{-1}(u_n - u)$, nous avons $u + t_n w_n = u_n \in U$ pour n assez grand. Nous savons également que $u_n \in K$, et donc pour n assez grand nous avons $J(u + t_n w_n) \geq J(u)$.

Avec un développement limité nous obtenons :

$$J(u) \leq J(u + t_n w_n) = J(u) + t_n \langle \nabla J(u), w_n \rangle + o(t_n \|w_m\|).$$

Nous pouvons en déduire (en divisant par $t_n > 0$) que

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

ce qui en passant à la limite donne le résultat. □

Notons que si $K = \mathbb{V}$ et $u \in \mathbb{V}$, alors le cône tangent $\mathcal{T}_{\mathbb{V}}(u)$ est l'espace tout entier. Le théorème précédent implique alors que $\langle \nabla J(u), w\rangle \geq 0$, pour tout $w \in \mathbb{V}$. Or si $w \in \mathbb{V}$, alors $-w \in \mathbb{V}$ aussi et donc $-\langle \nabla J(u), w\rangle \geq 0$ aussi. On en déduit que : $\langle \nabla J(u), w\rangle = 0$ pour tout $w \in \mathbb{V}$ et par suite :

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

On peut aussi raisonner de la même manière lorsque le minimum est atteint à l'intérieur de l'ensemble K, en effet pour tout $u \in \overset{\circ}{K}$, le cône tangent $\mathcal{T}_k(u)$ est aussi égal à $\mathbb{V}$.

Nous retrouvons donc notre condition au premier ordre déjà énoncé, mais cette fois-ci comme corollaire d'un cas plus général.

Corollaire 3.3.1. *Si $K = \mathbb{V}$, ou si le minimum u est intérieur à K, alors l'inéquation (3.2) devient*

$$\nabla J(u) = 0. \tag{3.3}$$

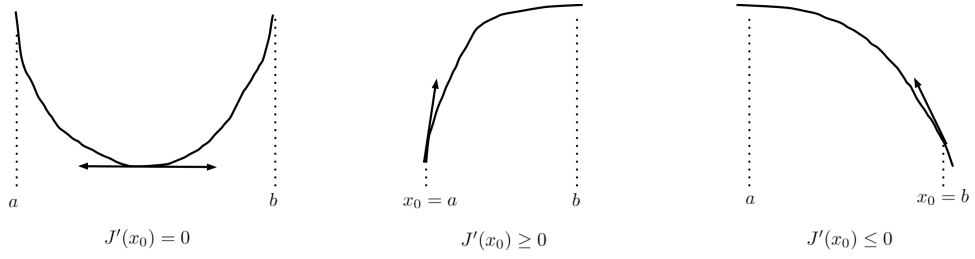
Remarque 3.3.1. *Si $\mathbb{R}^n$ est muni d'une base orthonormale, la condition nécessaire d'existence d'un minimum, dans les conditions du corollaire, peut être écrite sous la forme*

$$\frac{\partial J}{\partial x_1}(u) = \frac{\partial J}{\partial x_2}(u) = \ldots = \frac{\partial J}{\partial x_n}(u) = 0.$$

Il est très important de bien comprendre la distinction entre le résultat du théorème 3.3.1 et celui du corollaire 3.3.1 : Prenons le cas d'un problème simple :

$$\text{Trouver } x_0 \in [a,b],\ J(x_0) = \inf_{y\in[a,b]} J(y).$$

Si $x_0 \in]a,b[$, alors le corollaire implique que $J'(x_0) = 0$. Par contre si $x_0 = a$, alors le gradient n'a aucune raison de s'annuler et le théorème dit simplement que $J'(x_0) \geq 0$. De même, si $x_0 = b$, alors le théorème implique que $J'(x_0) \leq 0$.



Jusqu'ici nous n'avons étudié que la condition nécessaire d'existence d'un minimum, utilisant uniquement la différentielle d'ordre 1. On parle alors de condition du premier ordre. Dans le cas simple où $K = \mathbb{V}$, cette condition traduit la nullité du gradient de J

en le minimum. On peut facilement prouver aussi que cette condition reste nécessaire pour l'existence d'un maximum. Il est donc important de chercher une nouvelle condition d'optimalité pouvant différencier les minima des maxima. Cette condition utilise la différentielle d'ordre 2 et s'énonce comme suit :

Proposition 3.3.1. *On se place à nouveau dans le cadre du corollaire ci-dessus : on suppose que $K = \mathbb{V}$, ou bien que le minimum u est intérieur à K. On suppose en plus que J est deux fois (on considère donc une fonctionnelle Fréchet-différentiable.) différentiable en u. Alors, le minimum u vérifie :*

$$\nabla J(u) = 0, \quad et \quad \left\langle \nabla^2 J(u)h, h \right\rangle \geq 0, \qquad \forall h \in \mathbb{V}. \tag{3.4}$$

Démonstration. (i) Supposons que $K = \mathbb{V}$ et soit u un minimum local de J. Comme J est deux fois différentiable en u, on peut écrire le développement limité de J en u à l'ordre 2. Pour h un élément de $\mathbb{V}$ et $\lambda > 0$ petit, on a, cf. (A.10) :

$$J(u + \lambda\, h) = J(u) + \lambda\, \langle \nabla J(u), h \rangle + \frac{\lambda^2}{2} \left\langle \nabla^2 J(u)h, h \right\rangle + r_2(\lambda\, h),$$

avec $r_2(h)/\|h\|^2 \xrightarrow[h\to 0]{} 0$. D'après le corollaire précédent, $\nabla J(u) = 0$. Et du fait que u est minimum local, il existe $\lambda_0 > 0$ tel que pour tout $\lambda \in (0, \lambda_0)$, $J(u + \lambda h) \geq J(u)$, et par suite,

$$\frac{\lambda^2}{2} \left\langle \nabla^2 J(u)h, h \right\rangle + r_2(\lambda\, h) \geq 0, \text{ soit } \left\langle \nabla^2 J(u)h, h \right\rangle + \frac{2}{\lambda^2} r_2(\lambda\, h) \geq 0.$$

Faisons tendre λ vers 0^+. Le premier terme est indépendant de λ, et $\lim_{\lambda\to 0^+} \frac{2}{\lambda^2} r_2(\lambda\, h) = 0$, d'après le théorème de Taylor-Young A.3.5. Ainsi, on trouve bien $\langle \nabla^2 J(u)h, h \rangle \geq 0$.
(ii) Dans le cas où le minimum de J sur K est atteint en $u \in \overset{\circ}{K}$, alors pour tout $h \in \mathbb{V}$, il existe $\lambda_0 > 0$ tel que pour tout $\lambda \in (0, \lambda_0)$, $u + \lambda h \in K$ et $J(u + \lambda h) \geq J(u)$. Le reste de raisonnement se fait comme en (i). □

Remarque 3.3.2. *Toutes les conditions d'optimalité énoncées jusque là (que ce soit du premier ou du second ordre) ne constituent que des conditions nécessaires mais pas suffisantes. En effet, prenons le cas simple où $\mathbb{V} = \mathbb{R}$ et $J(x) = x^3$. En $x = 0$, les conditions nécessaires du premier et second ordre sont bien vérifiées ($J'(0) = J''(0) = 0$) et pourtant 0 n'est pas un minimum de J !*

Nous allons maintenant énoncer une condition suffisante :

Théorème 3.3.2. *Supposons que $K = \mathbb{V}$. Soit J une fonction deux fois différentiable et soit $u \in \mathbb{V}$. Si*

$$\begin{cases} \nabla J(u) = 0 \\ \exists U \text{ un voisinage de } u \text{ tel que } \left\langle \nabla^2 J(w)h, h \right\rangle \geq 0 \ \forall h \in \mathbb{V}, \forall w \in U \end{cases} \tag{3.5}$$

alors u est un minimum local de J.

Démonstration. Soit $u \in \mathbb{V}$ vérifiant (3.5) et soit $v \in U$. En écrivant le développement limité de J autour de u, on obtient : Il existe $\lambda \in]0,1[$ tel que

$$w = u + \lambda(v-u) \in U, \quad J(v) = J(u) + \langle \nabla J(u), v-u \rangle + \frac{1}{2} \left\langle \nabla^2 J(w)(v-u), v-u \right\rangle .$$

Ce qui avec (3.5) donne

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

Ceci étant vrai pour tout $v \in U$, on conclut que u est un minimum local de J. □

3.3.2 Cas convexe

Lorsque l'ensemble K est convexe nous pouvons associer directement avec K, un autre cône, le cône normal (convexe). Lorsque K est non-convexe, nous pouvons aussi définir des notions de cônes normaux, mais une prudence certaine s'avère alors nécessaire. Pour le lecteur intéressé nous référons à [15]. Pour notre cas, donnons directement la définition, avant de le lier à la notion de cône tangente :

Définition 3.3.1. *Soit $K \subseteq \mathbb{V}$ un ensemble convexe fermé et $\bar{u} \in K$ donné. Le cône normal (convexe) est défini comme suit :*

$$\mathrm{N}_K(\bar{u}) := \{x^* \in \mathbb{V}^* \ : \langle x^*, u - \bar{u} \rangle \leq 0 \ \forall u \in K\} . \tag{3.6}$$

Rappelons également la definition suivante :

Définition 3.3.2. *Soit $K \subseteq \mathbb{V}$ un cône fermé, c.à.d., $x \in K \Rightarrow \lambda x \in K$ pour tout $\lambda > 0$. Le cône polaire (ou cône dual négatif) K^* est défini comme suit :*

$$K^* := \{x^* \in \mathbb{V}^* \ : \langle x^*, x \rangle \leq 0 \ \forall x \in K\} . \tag{3.7}$$

Nous pouvons maintenant énoncer le lien entre le cône tangente et le cône normal, qui en fait sont liés par une relation polaire.

Proposition 3.3.2. *Soit $K \subseteq \mathbb{V}$ un ensemble convexe fermé et $\bar{u} \in K$ donné. Alors nous avons*

$$\mathcal{T}_K(\bar{u})^* = \mathrm{N}_K(\bar{u}). \tag{3.8}$$

Démonstration. Prenons un $w \in \mathcal{T}_K(\bar{u})$ arbitraire et de manière analogue un $x^* \in \mathrm{N}_K(\bar{u})$ arbitraire. On peut trouver par définition, une suite $t_n \downarrow 0$, $u_n \in K$, $u_n \to \bar{u}$,

$t_n^{-1}(u_n - \bar{u}) \to w$. Par définition de x^* et par positivité de t_n, il suit $\langle x^*, t_n^{-1}(u_n - \bar{u})\rangle \leq 0$. En passant à la limite nous obtenons évidemment $\langle x^*, w\rangle \leq 0$. Puisque x^*, w étaient arbitraire nous en concluons $\mathrm{N}_K(\bar{u}) \subseteq \mathcal{T}_K(\bar{u})^*$.

Inversement, soit maintenant $u^* \in \mathcal{T}_K(\bar{u})^*$ donné et arbitraire. Alors pour tout $u \in K$, et $\lambda \in [0,1]$, $\bar{u} + \lambda(u - \bar{u}) \in K$ par convexité. Fixons un $u \in K$ temporairement. Nous pouvons prendre une séquence $\lambda_n \downarrow 0$ et définir $u_n := \bar{u} + \lambda_n(u - \bar{u})$. Il suit $u_n \to \bar{u}$, et $w = (u - \bar{u}) \in \mathcal{T}_K(\bar{u})$ par définition. Donc par définition de u^*, nous obtenons $\langle u^*, w\rangle = \langle u^*, u - \bar{u}\rangle \leq 0$. Mais étant donné que u était en fait arbitraire, nous avons montré $u^* \in \mathrm{N}_K(\bar{u})$. Nous avons donc montré l'inclusion $\mathrm{N}_K(\bar{u}) \supseteq \mathcal{T}_K(\bar{u})^*$ et par cela terminé la preuve. □

Ce résultat est intéressant car il permet aussitôt de révisiter le Théorème 3.3.1 :

Corollaire 3.3.2 (au Théorème 3.3.1)**.** *Soient K un sous-ensemble non vide convexe fermé de $\mathbb{V}$, u un point de K et J une fonctionnelle de K à valeurs dans $\mathbb{R}$. On suppose que J est Fréchet-différentiable en u. Si u est un point de minimum local de J sur K, on a nécessairement*

$$-\nabla J(u) \in \mathrm{N}_K(u) \Leftrightarrow 0 \in \nabla J(u) + \mathrm{N}_K(u). \tag{3.9}$$

Démonstration. En effet, $\langle s, w\rangle \geq 0$ pour tout $w \in \mathcal{T}_K(u)$ implique $-s \in \mathcal{T}_K(u)^* = \mathrm{N}_K(u)$. □

Lorsque l'on se place dans un ensemble convexe, on obtient une condition nécessaire d'existence d'un minimum local, dite **inéquation d'Euler**. Grace à notre travail préparatoire, cette condition est devenu une conséquence triviale de ce qui précède.

Théorème 3.3.3. *Soient K un sous-ensemble convexe non vide de $\mathbb{V}$, u un point de K et J une fonctionnelle de K à valeurs dans $\mathbb{R}$. On suppose que J est différentiable en u. Si u est un point de minimum local de J sur K, on a nécessairement*

$$\langle \nabla J(u), v - u\rangle \geq 0, \qquad \forall v \in K. \tag{3.10}$$

Démonstration. Par définition du cône normal (convexe) et en utilisant le Corollaire 3.3.2, nous obtenons :

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

pour tout $v \in K$. □

Théorème 3.3.4. *Soient K un sous-ensemble convexe de $\mathbb{V}$, u un point de K et J une fonctionnelle convexe et différentiable sur K. Alors u est un point de minimum global de J sur K si, et seulement si,*

$$\langle \nabla J(u), v-u\rangle \geq 0, \qquad \forall v \in K. \tag{3.11}$$

Démonstration. Si u est un point de minimum de J, nous savons déjà que l'inéquation d'Euler (3.11) est vérifiée.
Réciproquement, du fait de la convexité de J, on sait que pour tout $v \in K$, on a :

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

Avec (3.11), on conclut que $J(v) \geq J(u)$ pour tout $v \in K$, et donc u est un point de minimum global de J sur K. □

Exemple 3.3.1 (Projection sur un convexe fermé)**.** *Reprenons l'exemple 2.3.1. On a vu que pour tout $w \in \mathbb{R}^n$, il existe une unique projection $P_k(w) \in K$ de w sur K. Cette projection $P_K(w)$ est solution de $\min_{v\in K} \|w-v\|$. Remarquons d'abord que $P_K(w)$ est aussi solution de*

$$\|w - P_K(w)\|^2 = \min_{v\in K} \|w-v\|^2.$$

On pose $J_1(v) = \|w-v\|^2$. Rappelons que $\nabla J_1(v) = 2(v-w)$. Le théorème 3.3.4 implique que $P_K(w)$ est l'unique élément de K tel que

$$-\frac{1}{2}\langle \nabla J_1(P_K(w)), v - P_K(w)\rangle = \langle w - P_K(w), v - P_K(w)\rangle \leq 0 \quad \forall v \in K \tag{3.12}$$

Nous aurons l'occasion d'utiliser cette caractérisation de la projection dans d'autres sections du cours.

Le théorème 3.3.4 dit que la condition (3.11) caractérise complètement le minimum d'une fonction convexe : c'est une condition nécessaire et suffisante. **Ce résultat n'est plus vrai lorsque J n'est pas convexe.**
Bien évidement dans les situations particulières où $K = \mathbb{V}$ ou u est intérieur à K. On a :

Théorème 3.3.5. *Soient K un sous-ensemble convexe de $\mathbb{V}$, u un point de K et J une fonctionnelle convexe et différentiable de K. Si $K = \mathbb{V}$ (ou si u est intérieur à K), alors u est un point de minimum (global) de J si, et seulement si,*

$$\nabla J(u) = 0. \tag{3.13}$$

La démonstration, immédiate, est laissée au lecteur.

3.4 Moindres carrés linéaires

Nous considérons dans ce chapitre un problème de minimisation, relativement courant en pratique, appelé problème de moindres carrés. Nous nous contentons de considérer le cas particulier des moindres carrés linéaires. Notons qu'un certain nombre d'outils développés dans le cours MA103 [3] permettent de résoudre ce type de problèmes.

3.4.1 Problèmatique d'une application

De prime abord, il est rassurant (!?) de résoudre exactement un problème. En pratique, cependant, on se rend compte que, dans de nombreux cas, il n'existe pas de solution "exacte". C'est souvent le cas lorsque l'on désire réaliser l'opération suivante :

— A partir d'un nombre fini (parfois très grand) de mesures, inférer un comportement (idéalement) valable dans tous les cas, passés, présents ou à venir.

Typiquement, d'une part on dispose d'un modèle abstrait, et d'autre part de données, et l'on souhaite fusionner l'un et l'autre, pour disposer d'une modélisation concrète du phénomène étudié, et/ou d'outils de prédiction. Prenons l'exemple suivant.

Carl Friedrichs Gauss (1777-1855) désirait déterminer la trajectoire de planètes, et notamment celle d'Uranus, découverte à la fin du 18ème siècle. D'après les lois de Képler, si l'on néglige la présence des autres planètes autour du Soleil, Uranus décrit une ellipse. Si l'on suppose *connus* le plan de la trajectoire (écliptique) ainsi que la direction du grand axe, sa trajectoire est une ellipse E dans le plan de l'écliptique, dont l'équation est

$$\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} = 1. \tag{3.14}$$

L'ellipse E est donc caractérisée par quatre paramètres, (x_0, y_0, a, b). Dès que l'on dispose de quatre positions (ou plus) d'Uranus dans le ciel, il est possible de caractériser sa trajectoire elliptique.

Enfin, pour caractériser la trajectoire, il est important d'observer que pour trois mesures ou moins, il existe une infinité de possibilités. Quatre mesures sont idéales, puisque qu'il leur correspond une unique ellipse. A partir de cinq mesures ou plus, il faut *espérer* que tous les points de la trajectoire, à partir du $5^{ème}$, se trouvent sur l'ellipse définie par les quatre premiers ! Cette prise de conscience (existence d'une **surdétermination**) est fondamentale, lorsque l'on résout ce type de problèmes. Afin de répondre à ce problème, Gauss a inventé le principe dit des **moindres carrés** (en 1801). Disposant de K mesures de la position d'Uranus $M_k(x_k, y_k)_{1\leq k\leq K}$, on choisit (x_0, y_0, a, b), ce qui définit une unique ellipse $E = E_{x_0,y_0,a,b}$. A partir de là, on introduit les points $(M'_k)_{1\leq k\leq K}$:

pour chaque valeur de k, M'_k est le point d'intersection de l'ellipse avec la droite passant par M_k et le centre de l'ellipse, le plus proche de M_k, de coordonnées

$$x'_k = p_E(x_k, y_k),\ y'_k = q_E(x_k, y_k), \quad 1 \leq k \leq K. \tag{3.15}$$

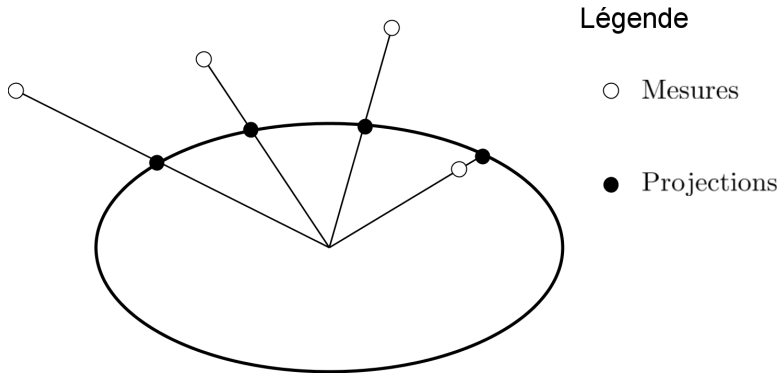


FIGURE 3.2 – Projections sur l'ellipse

Pour mesurer l'erreur commise entre les positions mesurées et leurs projections sur l'ellipse E, on forme la quantité

$$\nu = \sum_{k=1}^{K} M_k M_k'^2. \tag{3.16}$$

Si tous les points de la trajectoire mesurée se trouvent sur l'ellipse E, on obtient $\nu = 0$; dans le cas contraire, $\nu > 0$. Précisons, avant de continuer, que les *données* sont $(x_k, y_k)_{1 \leq m \leq K}$, et que les *inconnues* sont (x_0, y_0, a, b). Comme les nombres $(x'_k, y'_k)_{1 \leq k \leq K}$ sont caractérisés par les relations (3.15), on peut donc introduire la fonctionnelle

$$\nu(x_0, y_0, a, b) = \sum_{k=1}^{K} \{\|x_k - p_E(x_k, y_k)\|^2 + \|y_k - q_E(x_k, y_k)\|^2\}. \tag{3.17}$$

L'idée est de partir d'une première ellipse, puis de la modifier, de façon à diminuer la valeur de ν correspondante, et ainsi de suite... Le but est de *minimiser* la valeur de $\nu(x_0, y_0, a, b)$ sur le quadruplet (x_0, y_0, a, b) décrivant $\mathbb{R}^4$:

$$\begin{aligned} &\text{Trouver } (x_0^{opt}, y_0^{opt}, a^{opt}, b^{opt}) \in \mathbb{R}^4, \\ &\qquad \text{tel que } \nu(x_0^{opt}, y_0^{opt}, a^{opt}, b^{opt}) = \inf_{(x_0, y_0, a, b) \in \mathbb{R}^4} \nu(x_0, y_0, a, b). \end{aligned} \tag{3.18}$$

Idéalement, comme nous l'avons remarqué plus haut, *si* les mesures sont exactes, *et si* la trajectoire est effectivement elliptique dans le plan de l'ecliptique, on détermine une solution telle que

$$\nu(x_0^{opt}, y_0^{opt}, a^{opt}, b^{opt}) = 0.$$

Malheureusement, on sait que toute mesure est approchée, ce qui interdit de trouver un tel résultat. *Heureusement*, ceci n'est pas incompatible avec la résolution du problème (3.18).

Dans la suite, nous nous limiterons à l'étude de modèles-type, pour lesquels la dépendance par rapport aux inconnues est *linéaire*. A des fins illustratives, dans le formalisme adopté ci-dessus, on aurait

$$\left\{ \begin{array}{l} x'_k = \alpha_k x_0 + \beta_k y_0 + \gamma_k a + \delta_k b + f(x_1, y_1, \cdots, x_K, y_K), \\ y'_k = \alpha'_k x_0 + \beta'_k y_0 + \gamma'_k a + \delta'_k b + f'(x_1, y_1, \cdots, x_K, y_K), \end{array} \right. \quad 1 \leq k \leq K,$$

$$\text{soit} \quad \nu(v) = \|Av - b\|^2, \ v \in \mathbb{R}^4, \ A \in \mathbb{R}^{2K \times 4}, \ b \in \mathbb{R}^{2K}. \tag{3.19}$$

On parle alors de **moindres carrés linéaires**.

Remarque 3.4.1. *Pour refermer la parenthèse historique (voir [13] pour plus de détails), mentionnons que Gauss a mené à bien ses calculs (sans ordinateur !). A la suite de quoi, on s'est aperçu qu'au cours du temps la trajectoire elliptique optimale variait... Après avoir éliminé les incertitudes liées aux erreurs de mesure, on en a déduit que la trajectoire n'était pas une ellipse, mais plutôt une perturbation de trajectoire elliptique. L'influence des autres planètes a été prise en compte, mais cela ne résolvait toujours pas la difficulté.* Urbain le Verrier *(1811-1877) a donc eu l'idée de chercher une nouvelle planète, introduisant une nouvelle perturbation, qui validerait le modèle : il a découvert Neptune en 1846.*

3.4.2 Le formalisme abstrait et son étude : pourquoi des carrés ?

Dans la suite, pour A une matrice *non nulle* de $\mathbb{R}^{m \times n}$ et b un vecteur de $\mathbb{R}^m$, on considère la résolution du problème :

$$\min_{v \in \mathbb{R}^n} f(v), \ \text{avec} \ f(v) = \|Av - b\|_m ,$$

où m et n sont deux éléments non nuls quelconques de $\mathbb{N}$, *a priori* distincts. Pour cette raison, on indicera les normes et produits scalaires par $_m$ ou $_n$ si nécessaire, pour éviter les confusions.

On remarque, avant de commencer l'étude proprement dite du problème de minimisation, que f est *convexe*. En effet, on vérifie que pour v et w deux éléments de $\mathbb{R}^n$, et θ dans $]0, 1[$, on a l'inégalité

$$f(\theta v + (1 - \theta)w) \leq \theta f(v) + (1 - \theta) f(w).$$

Comme f est à valeurs positives, il est équivalent de prouver que les carrés sont dans cet ordre. On pose $x = Av - b$ et $y = Aw - b$:

$$\begin{aligned} f(\theta v + (1-\theta)w)^2 &= \|A(\theta v + (1-\theta)w) - b\|^2 \\ &= \|\theta x + (1-\theta)y\|^2 \\ &= \theta^2\|x\|^2 + 2\theta(1-\theta)\langle x, y\rangle + (1-\theta)^2\|y\|^2 \\ &\leq \theta^2\|x\|^2 + 2\theta(1-\theta)\|x\|\,\|y\| + (1-\theta)^2\|y\|^2 \\ &= [\theta\|x\| + (1-\theta)\|y\|]^2 \\ &= [\theta f(v) + (1-\theta)f(w)]^2. \end{aligned}$$

En conséquence, d'après les résultats du chapitre 3, les conditions d'existence de minimum seront *nécessaires et suffisantes*. Comment caractériser le minimum ? C'est l'objet des deux sous-sections ci-dessous...

3.4.3 L'approche directe

En vue d'appliquer les résultats du chapitre 3, calculons le gradient de f, sans toutefois oublier de *vérifier* que f est différentiable.

Allons-y... Soient donc v et h deux éléments de $\mathbb{R}^n$, et θ un réel destiné à tendre vers 0 par valeurs positives.

$$f(v + \theta h) - f(v) = \|x - \theta Ah\| - \|x\|, \text{ avec } x = Av - b.$$

1. On se place pour commencer dans le cas général $x \neq 0$.

$$\begin{aligned} \|x - \theta Ah\| - \|x\| &= \frac{1}{\|x - \theta Ah\| + \|x\|}[\|x - \theta Ah\|^2 - \|x\|^2] \\ &= \frac{1}{\|x - \theta Ah\| + \|x\|}[2\theta\langle x, Ah\rangle_m + \theta^2\|Ah\|^2] \\ &= \frac{1}{\|x - \theta Ah\| + \|x\|}[2\theta\langle A^\mathsf{T}x, h\rangle_n + O(\theta^2)]. \end{aligned}$$

   Par ailleurs, $\|x\| - \|\theta Ah\| \leq \|x - \theta Ah\| \leq \|x\| + \|\theta Ah\|$: on a donc $\|x - \theta Ah\| = \|x\| + O(\theta)$. Ainsi, puisque x est fixé (avant-dernière égalité),

$$\frac{1}{\|x - \theta Ah\| + \|x\|} = \frac{1}{2\|x\| + O(\theta)} = \frac{1}{2\|x\|(1 + O(\theta))} = \frac{1}{2\|x\|}(1 + O(\theta)).$$

   D'où

$$\|x - \theta Ah\| - \|x\| = \theta\frac{\langle A^\mathsf{T}x, h\rangle_n}{\|x\|} + O(\theta^2) = \theta\frac{\langle A^\mathsf{T}Av - A^\mathsf{T}b, h\rangle_n}{\|Av - b\|} + o(\theta).$$

On a donc trouvé

$$\nabla f(v) = \frac{A^\mathsf{T} Av - A^\mathsf{T} b}{\|Av - b\|}. \tag{3.20}$$

NB. On vérifie que f est Fréchet-différentiable selon une procédure similaire.

N.B. : Pour voir que $1/(1 + O(\theta))$“ = ”$1 + O(\theta)$, rappelons que $f(t)$ est $O(t)$ (pour t assez petit) lorsqu'il existe un $C > 0, \delta > 0$ tel que $|f(t)| \leq C\,|t|$ pour tout t avec $|t| \leq \delta$. Alors tout d'abord il existe un $\delta' > 0$ tel que $|1 + f(t)| \geq \frac{1}{2}$ pour tout $|t| \leq \delta'$. Ensuite, pour un tel t assez petit nous avons

$$\begin{aligned}\left|\frac{1}{1+f(t)} - 1\right| &= \left|\frac{1}{1+f(t)} - \frac{1+f(t)}{1+f(t)}\right| = \left|\frac{f(t)}{1+f(t)}\right| \\ &\leq 2\,|f(t)| \leq 2C\,|t|\,,\end{aligned}$$

ce qui permet de conclure.

2. Que se passe-t-il dans le cas particulier $x = 0$? Supposons que f soit différentiable, de différentielle $h \mapsto \langle g, h\rangle$ ($g \in \mathbb{R}^n$). Par définition de la Gateaux-differentiabilité :

$$f(v + \theta h) - f(v) = \theta\|Ah\| = \theta\,\langle g, h\rangle + o(\theta), \quad \forall h \in \mathbb{R}^n.$$

Prenons, pour les deux directions h et $-h$, la même valeur de θ, soit

$$\theta\,\langle g, h\rangle + o(\theta) = \theta\|Ah\| = \theta\|A(-h)\| = \theta\,\langle g, -h\rangle + o(\theta),$$

et divisons par θ, que l'on fait tendre vers 0. Il reste $2\,\langle g, h\rangle = 0$, pour toute direction h de $\mathbb{R}^n$. On infère la nullité de g, ce qui implique finalement

$$\theta\|Ah\| = o(\theta),$$

soit $Ah = 0$ pour tout h, ou encore $A = 0$. Or, on a supposé que A est une matrice non nulle. En conclusion, *f n'est pas différentiable en 0*!

Outre le fait que le calcul n'est pas immédiat, nous sommes confrontés à un problème majeur. f n'est pas différentiable en v_0 si $Av_0 = b$. Mais, si $Av_0 = b$, $f(v_0) = 0$ et v_0 est un point de minimum de f, puisque f est à valeurs positives ! Les résultats du chapitre 3 ne sont donc pas applicables, puisqu'ils requièrent la différentiabilité au point de minimum. Comment remédier à cette difficulté ? C'est l'objet de la sous-section suivante.

3.4.4 Une astuce de calcul

Comme f est à valeurs positives, les minima et points de minimum de f sont identiques à ceux de son carré, f^2 ! On peut donc considérer le problème de minimisation

$$\min_{v\in\mathbb{R}^n} J(v), \text{ avec } J(v) = \|Av-b\|_m^2.$$

On vérifie sans peine que

$$J(v+\theta h) - J(v) = 2\theta \langle Av-b, Ah\rangle_m + \theta^2\|Ah\|_m^2 = 2\theta \left\langle A^\mathsf{T}Av - A^\mathsf{T}b, h\right\rangle_n + o(\theta).$$

Ainsi, J est différentiable en tous points (la Fréchet-différentiabilité est obtenue de même), et l'on a déterminé l'expression suivante du gradient

$$\nabla J(v) = 2A^\mathsf{T}Av - 2A^\mathsf{T}b. \tag{3.21}$$

Cette fois, on peut appliquer les résultats du chapitre 3. Pour commencer, J est convexe, d'après le point (iii) du théorème 2.4.1, puisque

$$\langle \nabla J(v) - \nabla J(u), v-u\rangle = 2\left\langle A^\mathsf{T}A(v-u), v-u\right\rangle_n = 2\|A(v-u)\|_m^2.$$

Qui plus est, on a le résultat ci-dessous :

Théorème 3.4.1. *u est un point de minimum global de J si, et seulement si, u est solution de*

$$A^\mathsf{T}Au = A^\mathsf{T}b. \tag{3.22}$$

Démonstration. Ceci est une simple application du théorème 3.3.5. □

Définition 3.4.1. *L'équation $A^\mathsf{T}Au = A^\mathsf{T}b$ est appelée* **équation normale**.

Il faut faire très attention. Si bien sûr $Au = b$ entraîne (3.22), la réciproque est **fausse** en général...

3.4.5 Existence du point de minimum

On a le

Théorème 3.4.2. *Il existe au moins un point de minimum global.*

Démonstration. Ceci revient à montrer que le système linéaire (3.22) admet toujours au moins une solution. Pour cela, nous allons utiliser la relation $\operatorname{Im} A = (\operatorname{Ker} A^{\mathsf{T}})^{\perp}$, énoncée et démontrée au lemme B.1.2.

Pour tout élément b de $\mathbb{R}^m$, on peut écrire $b = b_0 + b_{\perp}$, avec $b_0 \in \operatorname{Ker} A^{\mathsf{T}}$ et $b_{\perp} \in (\operatorname{Ker} A^{\mathsf{T}})^{\perp}$. Alors, $A^{\mathsf{T}} b = A^{\mathsf{T}} b_{\perp}$, et d'après la relation ci-dessus, il existe un élément u de $\mathbb{R}^n$ tel que $b_{\perp} = Au$. On en déduit finalement, pour ce vecteur u :

$$A^{\mathsf{T}} b = A^{\mathsf{T}} b_{\perp} = A^{\mathsf{T}} A u.$$

□

On peut se servir d'outils différents pour retrouver ce résultat. Nous allons détailler la démarche, car elle est fort instructive, et utile pour la suite du chapitre... La matrice $A^{\mathsf{T}}A$, qui apparaît dans le terme quadratique de J, est une matrice symétrique et positive ; en effet :

$$\begin{aligned} (A^{\mathsf{T}}A)^{\mathsf{T}} &= A^{\mathsf{T}}A, \text{ et} \\ \left\langle A^{\mathsf{T}}Ax, x \right\rangle_n &= \langle Ax, Ax \rangle_m = \|Ax\|_m^2 \geq 0, \quad x \in \mathbb{R}^n. \end{aligned}$$

Par voie de conséquence, il existe $(v_i)_{1 \leq i \leq n}$ une base orthonormale de $\mathbb{R}^n$ de vecteurs propres, de valeurs propres associées $(\lambda_i)_{1 \leq i \leq n}$, appartenant à $\mathbb{R}^+$: $A^{\mathsf{T}}Av_i = \lambda_i v_i$, pour $1 \leq i \leq n$. Dans la suite, on les classe par ordre *décroissant*, et l'on définit q, le cardinal de l'ensemble $\{\lambda_i \ : \ \lambda_i > 0\}$, c'est-à-dire que $q = \operatorname{rg}(A^{\mathsf{T}}A)$. Notons que, puisque A n'est pas la matrice nulle, on a $1 \leq q \leq n$.

Dans l'expression de J, on a également un terme linéaire, de la forme $-2\langle b, Av \rangle_m$. Soient donc les vecteurs de $\mathbb{R}^m$ définis par

$$w_i = \frac{1}{\sqrt{\lambda_i}} A v_i, \quad 1 \leq i \leq q.$$

Pourquoi avoir introduit le facteur $1/\sqrt{\lambda_i}$? Parce que, pour $1 \leq i, j \leq q$, on a la relation

$$\begin{aligned} \langle w_i, w_j \rangle_m &= \left\langle \frac{1}{\sqrt{\lambda_i}} A v_i, \frac{1}{\sqrt{\lambda_j}} A v_j \right\rangle_m \\ &= \frac{1}{\sqrt{\lambda_i \lambda_j}} \left\langle A^{\mathsf{T}} A v_i, v_j \right\rangle_n \\ &= \sqrt{\frac{\lambda_i}{\lambda_j}} \langle v_i, v_j \rangle_n = \delta_{ij}. \end{aligned}$$

En d'autres termes, $(w_i)_{1\leq i\leq q}$ est une famille orthonormale de $\mathbb{R}^m$. NB. Au passage, on vient de prouver que

$$dim[\mathrm{Im}\, A] = dim[Vect(Av_1, \cdots, Av_n)] = dim[Vect(w_1, \cdots, w_q)] = q.$$

Ceci signifie en particulier que $\mathrm{rg}(A)$=$\mathrm{rg}(A^\mathsf{T}A)$ et $q \leq m$.

On complète $(w_i)_{1\leq i\leq q}$, le cas échéant, en une base orthonormale de $\mathbb{R}^m$. On peut alors décomposer le vecteur courant v ainsi que b sur les bases *ad hoc*, soit $v = \sum_{i=1}^n x_i v_i$ et $b = \sum_{i=1}^m b_i w_i$, pour obtenir

$$\begin{aligned} J(v) &= \|Av - b\|_m^2 \\ &= \|\sum_{i=1}^q \sqrt{\lambda_i} x_i w_i - \sum_{i=1}^m b_i w_i\|_m^2 \\ &= \sum_{i=1}^q \left(\sqrt{\lambda_i} x_i - b_i\right)^2 + \sum_{i=q+1}^m {b_i}^2. \end{aligned} \tag{3.23}$$

NB. Dans (3.23), la seconde somme peut être vide (si $q = m$).

Qu'en déduit-on ?

Proposition 3.4.1. *u est un point de minimum de J si, et seulement si,*

$$u = \sum_{i=1}^n x_i^0 v_i, \textit{ avec } x_i^0 = \frac{1}{\sqrt{\lambda_i}} b_i,\ 1 \leq i \leq q, \quad x_i^0 \textit{ quelconques }, q+1 \leq i \leq n. \tag{3.24}$$

De façon équivalente, si on note $u^0 = \sum_{i=1}^q x_i^0 v_i$, u est un point de minimum si, et seulement si,

$$u \in u^0 + Vect(v_{q+1}, \cdots, v_n). \tag{3.25}$$

Par construction (encore une fois !), l'ensemble des points de minimum est non vide...

Exercice 3.4.1. *Vérifier que (3.24) ou (3.25) est équivalent à (3.22).*

Ceci est un bon exemple de la propriété générale suivante. Supposons que, pour un problème posé à l'aide d'une matrice, on puisse prouver que celle-ci est *diagonalisable*. Alors, sous réserve que l'on connaisse ses éléments propres, résoudre le problème initial revient à résoudre un ensemble de problèmes dans $\mathbb{R}$.

Bien évidemment, le *défaut majeur* est qu'en général, il est beaucoup trop coûteux de calculer l'ensemble des éléments propres d'une matrice ! Dans le cas des moindres carrés linéaires, on choisit plutôt de construire des algorithmes numériques directs ou itératifs permettant d' *"inverser"* l'équation normale (c'est-à-dire de calculer un vecteur u solution de (3.22)).

Chapitre 4

Algorithmes pour problèmes sans contraintes : Fonctionnelle quadratique

Dans ce chapitre, nous allons étudier des algorithmes qui permettent de calculer **numériquement** la solution du problème de minimisation,

$$\text{Trouver } u \in \mathbb{R}^n \text{ tel que } J(u) = \min_{v \in \mathbb{R}^n} J(v).$$

Ici, J est la fonctionnelle qui à v associe $J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle$. Nous supposons dans la suite que A est une matrice *symétrique définie-positive* de $\mathbb{R}^{n\times n}$ et b un vecteur quelconque de $\mathbb{R}^n$. Nous avons vu, aux chapitres 2 et 3, que la solution d'un tel problème existe et est unique, et qu'elle vérifie le système linéaire

$$\text{(Problème sans contraintes)} \qquad Au = b. \tag{4.1}$$

A partir de là on peut penser que le minimum peut être obtenu **explicitement** à l'aide d'une résolution exacte à l'aide de la méthode de Cramer.

Remarque 4.0.1 (Formule de Cramer)**.** *Soit $A = [A_1 A_2 \dots A_n]$ une matrice $n \times n$ inversible. La i^e composante x_i de la solution x du système linéaire $Ax = b$ est donnée par :*

$$x_i = \frac{det(A_1, A_2, \dots, A_{i-1}, b, A_{i+1}, \dots, A_n)}{detA}.$$

En réalité, une telle méthode demanderait plusieurs opérations de calcul et est inconcevable pour des grandes dimensions $n \geq 10$. En effet, la méthode de Cramer nécessite le calcul de $n+1$ déterminants qui se calculent chacun en $n!$ multiplications. Ce qui fait

un total de $(n+1)!$ opérations (sans compter les additions). Avec un ordinateur qui réalise 1 milliard de milliards d'opérations par seconde, il faudrait plus environ 6 mois pour résoudre un système linéaire lorsque $n = 25$ ... Pour $n = 30$, il faut des millions d'années !

Il est donc nécessaire d'élaborer des *méthodes numériques* plus rapides. Les algorithmes d'optimisation, que nous allons discuter dans ce chapitre, consistent tous à choisir un **point initial** $u_0 \in \mathbb{R}^n$, puis à construire une suite $(u_k)_{k\geq 1}$. Pour que de telles méthodes soient efficaces, il faut qu'elles possèdent les deux propriétés suivantes :

— La convergence de la suite (u_k) est assurée, quel que soit le vecteur initial.

— La convergence doit être "suffisamment rapide".

Le premier critère admet une interprétation claire, d'un point de vue mathématique. Le sens du second critère est plus flou, et nous essayerons de le préciser dans les sections suivantes.

4.1 Précisions Numériques. Critères associés à la convergence

Tout d'abord, il faut être conscient, lorsque l'on effectue un calcul *numérique*, que la précision est **finie**, à la différence du calcul *formel*, par exemple.

La finitude de la précision découle de la représentation en machine des nombres réels, sous la forme générique (mais en base 2) :

$$\pm a_0, a_1 \cdots a_p \ 10^d, \text{avec } (a_0, \cdots, a_p) \in \{0, \cdots, 9\}^{p+1},\ a_0 \neq 0,\ d \in \{-d_{max}, \cdots, d_{max}\},$$

où p et d_{max} dépendent du microprocesseur qui effectue les calculs. On dit aussi que $p+1$ est le nombre maximal de chiffres significatifs de la représentation en machine, et que $10^{-d_{max}}$ est la précision machine. Cette représentation génère deux difficultés :

1. Tout nombre dont la valeur absolue est plus grande que $10^{d_{max}+1}$ est considéré comme infini, et symétriquement, tout nombre dont la valeur absolue est strictement plus petite que $10^{-d_{max}}$ est considéré comme étant nul ;
2. Les opérations sur ces nombres (addition, extraction de racine, ... etc) sont effectuées en précision finie. Prenons l'exemple de la multiplication : si les deux nombres ont respectivement q et q' chiffres significatifs $(q, q' \in \{1, \cdots, p+1\})$, leur produit possède $q+q'-1$ ou $q+q'$ chiffres significatifs. Dès lors que $q+q'-1 > p+1$, une *troncature* est effectuée lors de la mise en mémoire du résultat (même si le calcul était exact), puisque la représentation de tout nombre comporte au plus $p+1$ chiffres significatifs.

C'est la raison pour laquelle les calculs numériques produisent en général des erreurs d'arrondi...

Par voie de conséquence, et pour revenir à notre problème, il devient difficile d'obtenir un résultat du type $Au-b=0$. En fait même, d'après l'exposé précédent, si l'ordinateur affirme que $Au-b=0$, ceci signifie uniquement que la différence est plus petite que la précision machine.

Par ailleurs, on se contente en général d'une valeur *approchée*, c'est-à-dire à ε près. Quel est le sens mathématique sous-jacent ? Typiquement, si on note $\|.\|$ une norme quelconque, pour $\varepsilon \geq 0$, on cherche v_ε tel que

$$\|Av_\varepsilon - b\| \leq \varepsilon. \tag{4.2}$$

Il est clair que l'ensemble des v_ε qui satisfont à (4.2) n'est pas réduit à un singleton ! Quoiqu'il en soit, à ε près, l'obtention d'un tel v_ε est suffisante... On parle de *convergence numérique*.

Exercice 4.1.1. *Quel est l'ensemble défini par (4.2) ?*

4.1.1 Test d'arrêt.

A la notion de calcul à ε près correspond, par dualité, celle de la précision requise, ce qui permet de déterminer un **critère (ou test) d'arrêt** pour nos méthodes. En effet, pour $\varepsilon \geq 0$ et u_0 donnés, on va effectuer des itérations,

$$\text{Pour } k = 0, 1, \cdots, \quad \text{tant que } \|Au_k - b\| > \varepsilon \quad \text{itérer } u_k \to u_{k+1}. \tag{4.3}$$

(Les itérations sont interrompues pour la première valeur de k telle que $\|Au_k - b\| \leq \varepsilon$.)

4.1.2 Evaluer le coût calcul d'une méthode itérative.

Le premier point important auquel on s'intéresse dans une méthode numérique est le **nombre d'itérations** nécessaire à la validation du critère d'arrêt. Naturellement, on aura tendance à privilégier une méthode nécessitant peu d'itérations. Mais baser une analyse de la qualité d'une méthode itérative sur le nombre d'itérations uniquement est *incorrect*. A titre d'un experiment de pensée, songez à la méthode de Cramer comme une méthode ayant un seul itération !

Un second point, complémentaire du premier, est le **coût d'une itération**. Typiquement, il s'agit du nombre d'opérations nécessaires à la réalisation d'une itération,

c'est-à-dire au calcul de u_{k+1}, connaissant u_k. On obtient une idée du **coût de calcul** en multipliant le nombre d'itérations par le coût d'une itération.

Donnons deux exemples élémentaires d'estimation du nombre d'opérations dans $\mathbb{R}^n$.

1. Le *produit scalaire* de deux vecteurs, qui s'écrit

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

est effectué en n multiplications et $(n-1)$ additions.

Usuellement, on ne conserve que le terme principal, ce qui signifie que l'on considère que le produit scalaire requiert n additions et n multiplications (i.e., $2n$ opérations).

2. La *multiplication matrice-vecteur*, qui s'écrit composante par composante,

$$(Ax)_i = \sum_{j=1}^{n} A_{i,j} x_j, \quad 1 \leq i \leq n,$$

requiert n^2 additions et n^2 multiplications, ce qui laisse à penser qu'un produit matrice-vecteur est équivalent à n produits scalaires... Ceci étant, que se passe-t-il si l'on sait que la matrice A est creuse, c'est-à-dire avec K éléments non nuls par ligne, avec K très petit devant n ? On ne va stocker que les positions, i. e., les paires d'indices (i, j), et les valeurs $A_{i,j}$ non nulles ! Lorsque l'on multiplie A par x, on n'effectue que les multiplications pour lesquelles $A_{i,j} \neq 0$ (et les additions de termes non nuls). On aura donc effectué Kn additions, et autant de multiplications...

Prenons le cas où $K \leq 7$, et la dimension de l'espace est $n = 10^4$ (ce qui est tout à fait envisageable !), on voit que les deux évaluations du coût de calcul donnent

$$2n^2 = 2 \times 10^8 \text{ et } 2Kn = 14 \times 10^4,$$

ou l'équivalent de 10.000 produits scalaires, contre 14.

Une autre façon d'estimer le coût du calcul est de mesurer le **temps de calcul**, par l'intermédiaire d'une horloge. Noter que ce temps de calcul dépend de la machine sur laquelle on effectue le calcul numérique (On raisonne usuellement en opérations flottantes par seconde, ou **FLOPs** = **FL**oating **OP**erations per **s**econd, pour un processeur donné, sans distinguer les opérations entre elles.). Une machine peut (pour simplifier, car il existe d'autres modes de fonctionnement), soit travailler *séquentiellement*, soit *en parallèle*. Dans le premier cas, les opérations sont exécutées l'une après l'autre. Dans

le second cas, la machine est constituée de plusieurs processeurs, qui peuvent alors exécuter simultanément des opérations, et échanger des données entre eux (On suppose que l'algorithme de calcul le permet. Le fait qu'un algorithme soit effectivement exécutable en parallèle, ou *parallélisable*, sort du cadre de ce cours...).

Bref, le temps horloge n'est pas le même sur toutes les machines, alors que le nombre total d'opérations est identique. Ceci dit, le temps horloge est l'unité qui permet le plus facilement de communiquer auprès des "profanes". Tant que la comparaison de différentes approches s'effectue en même mode et sur la même machine, on peut argumenter pour un certain intérêt à regarder le temps CPU.

4.1.3 Stockage mémoire.

Enfin, il peut également être utile de quantifier le **stockage mémoire** requis pour l'exécution d'une méthode / d'un algorithme. Cette question paraît un peu désuet, à l'époque où l'on sature rarement la mémoire d'une machine avec les applications "jouets", même en faisant pas attention du tout. Imaginez toutefois une application de vision en robotique, ou pour simplifier une image est associé à une matrice pleine ayant un certain nombre de pixels. Imaginez le gamme : full HD au 8K nécessitant entre 2 millions et 33 millions réels (matrices 1920×1080 et 7680×4320 respectivement) et un algorithme ayant besoin pour analyser une seconde de vision d'avoir l'intégralité des images en mémoire pour chaque milliseconde. En stockant les réels en double précision, on arrive ainsi à un besoin mémoire entre 259 Go et 4 To. Cette question du mémoire est donc quelque chose à garder en tête, bien que pendant le cours il ne sera pas rencontrée effectivement.

4.1.4 Comparaison d'algorithmes

Plus généralement la question de la comparaison de plusieurs algorithmes est délicate. Il est bien évidemment de bon usage d'effectuer les comparaisons sur une même machine. Quelque part, il devrait aller de soi d'effectuer des tests sur un grand nombre d'instances. Ces instances devraient couvrir un large spectre de conditions et difficultés. Mais concrètement, comment construit-t-on un tel "batch d'instances" ?

- Il est populaire d'utiliser "un générateur" aléatoire d'instances. Dans notre cas, nous n'avons qu'à générer des matrices A (positif défini) et vecteurs b au hasard, n'est-ce pas ? Au delà de la question de comment concrètement faire, quelques critiques :
  - Quelle loi prend on pour les entrées ? Ces choix, confèrent elles une structure

particulière au problème ?
- En pratique les instances ainsi crées sont souvent ou trop facile ou artificiellement dur.
- Imaginez convaincre une personne ayant un problème concret (ici : matrice A et b issu de la pratique avec une structure bien particulière !) du bien fondé de votre approche en mettant en avant le déroulement d'un algorithme sur des instances générées au hasard ?

— Quels sont les aspects du problème le rendant difficile ? Nous verrons sous peu que dans notre cadre bien restreint, la notion de conditionnement joue un rôle. Mais est-t-il si facile d'identifier ces aspects dans des problèmes réelles ayant de multiples aspects ?

— Comment gère-t-on la possibilité pour certains algorithmes d'échouer sur certain problèmes et d'autres sur d'autres instances ?

On réfère le lecteur intéressé à [6] pour une discussion sur ce dernier point et plus généralement la comparaison d'algorithmes.

4.1.5 Observations finales

La discussion de cette section est volontairement restée très générale, et elle peut être vue comme une introduction à l'algorithmique numérique. Ce qu'il faut retenir, c'est qu'il convient d'être prudent lorsque l'on évalue la qualité d'une méthode numérique, car celle-ci résulte habituellement de compromis entre les divers critères et contraintes que nous avons évoqués ci-dessus. Pour ce type de problèmes, il est fort utile d'acquérir de l'expérience, notamment en réalisant des comparaisons entre plusieurs méthodes.

4.2 Conditionnement d'un problème

Nous avons vu, dans la section précédente, qu'il est inévitable d'avoir des erreurs d'arrondi. Ces erreurs peuvent se propager et même s'accumuler d'une itération à l'autre, et aussi **s'amplifier** au cours du calcul. Nous allons maintenant essayer de comprendre la raison de cette amplification. Prenons l'exemple suivant :

$$\begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix} x = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \quad \text{a pour solution } x = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

Si on modifie un tout petit peu le second membre on obtient une solution très différente :

$$\begin{bmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{bmatrix} x = \begin{pmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{pmatrix} \quad \text{a pour solution } x = \begin{pmatrix} 9.2 \\ -12.6 \\ 4.5 \\ -1.1 \end{pmatrix}.$$

Cet exemple numérique montre que de très faibles erreurs sur les données (la matrice ou le vecteur du second membre) peuvent engendrer une grande erreur sur la solution. Pour quantifier cet écart, nous allons introduire la notion de conditionnement d'une matrice.

Définition 4.2.1. *Soit $\|\cdot\|$ une norme matricielle induite (Definition B.1.1). On appelle conditionnement d'une matrice réelle inversible $A \in \mathbb{R}^{n\times n}$, relatif à cette norme, la valeur définie par :*

$$\mathrm{cond}(A) = \|A\| \left\|A^{-1}\right\|.$$

Proposition 4.2.1. *Soit une matrice inversible A. Soit $b \neq 0$ un vecteur non nul.*

1. *Soient x et $x+\delta x$ les solutions respectives des systèmes linéaires $Ax = b$, et $A(x+\delta x) = b+\delta b$, alors*

$$\frac{\|\delta x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|\delta b\|}{\|b\|}. \tag{4.4a}$$

2. *Soient x et $x+\delta x$ les solutions respectives de $Ax = b$, et $(A+\delta A)(x+\delta x) = b$. Alors on a :*

$$\frac{\|\delta x\|}{\|x+\delta x\|} \leq \mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}. \tag{4.4b}$$

Démonstration. Prenons x et $x+\delta x$ solutions de $Ax = b$, et $A(x+\delta x) = b+\delta b$, et remarquons d'abord qu'on a $\delta x = A^{-1}\delta b$. D'où $\|\delta x\| \leq \|A^{-1}\|\|\delta b\|$. D'autre part,on a $\|b\| \leq \|A\|\|x\|$, ou encore (puisqu'on a supposé $b \neq 0$) $\frac{1}{\|x\|} \leq \|A\|\frac{1}{\|b\|}$. On en déduit alors l'inégalité (4.4a). L'inégalité (4.4b) s'obtient par des majorations analogues. □

Remarque 4.2.1. *Noter que le conditionnement d'une matrice est toujours supérieur à 1.*

En effet, remarquons d'abord que pour n'importe quelle norme induite (cf. proposition B.1.1) le conditionnement de la matrice identité I_n est : $\mathrm{cond}(I_n) = \|I_n\|\,\|I_n\| = 1$. *En plus, pour une matrice inversible A et pour n'importe quelle norme induite, on a :*

$$1 = \|I_n\| = \|A\,A^{-1}\| \leq \|A\|\,\|A^{-1}\| = \mathrm{cond}(A).$$

Remarque 4.2.2. *Les inégalités de la proposition précédente donnent des "majorations" du taux de perturbation de la solution en fonction du conditionnement de la matrice et aussi en fonction de la precision sur les données A et b. Une conséquence importante de la proposition précédente est l'observation suivant.*

Si on cherche la solution u du système $Au = b$, et si par une méthode numérique on calcule une ε-solution, i.e. v_ε solution de $\|Av_\varepsilon - b\| \leq \varepsilon$. Alors, même pour un ε assez petit, v_ε **peut être** *assez loin de la solution exacte u. Et plus exactement, la distance $\|v_\varepsilon - u\|$ dépendra du conditionnement de la matrice A.*

Proposition 4.2.2. *Soit A une matrice symétrique définie positive. Soient $\lambda_{\min}$ et $\lambda_{\max}$ respectivement la plus petite et la plus grande valeur propre de A. Le conditionnement de A pour la norme Euclidienne $\|\cdot\|_2$, est*

$$\kappa := \operatorname*{cond}_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

4.2.1 Taux et vitesse de convergence.

Dans le reste de ce chapitre, nous allons nous intéresser à des méthodes numériques dont on étudiera en particulier la convergence. Cette convergence théorique est guarantie sans tenir compte de tous les phénomènes de précisions numériques, amplification des erreurs d'arrondi, et même sans tenir compte du fait qu'on ne cherche à satisfaire que le test d'arrêt (4.3). Pour compléter l'étude, il nous faudra aussi nous intéresser aux taux et vitesse de convergence.

Définition 4.2.2. *Soit une méthode numérique produisant une suite d'itérés $(u_k)_k$. Soit $C > 0$ la plus petite constante telle que : $\|u_{k+1} - u\| \leq C\|u_k - u\|$ pour tout $k \geq 0$. C est appelée taux de convergence.*

On appellera aussi vitesse de convergence la quantité $R := -\ln C$.

Il est clair que si le taux de convergence C est inférieur strictement à 1, alors la méthode sera convergente. De plus, la vitesse de convergence d'une méthode est d'autant plus grande que son taux de convergence C est petit devant 1.

4.3 Méthodes de descente

Nous allons nous intéresser dans cette section aux méthodes dites de **descente**. Le principe de cette méthode est le suivant. Supposons l'itéré u_k connu : on choisit une

direction, dite **de descente** , $d_k \neq 0$, et un **pas de descente** ρ_k. On construit l'itéré u_{k+1} par la formule :

$$u_{k+1} = u_k + \rho_k d_k.$$

Le choix de d_k et ρ_k se fera de manière à assurer que :

$$J(u_{k+1}) < J(u_k).$$

On répétera ce procédé jusqu'à ce que le test d'arrêt (4.3) soit satisfait.

Remarquons, avant de poursuivre que, si u_k est égal à u, la solution cherchée, on espère avoir $d_k = 0$, et que de fait $u_{k+1} = u$ également.

Nous allons voir qu'il y a plusieurs façons de choisir les directions de descente. Pour le pas de descente, on choisit soit un pas fixe ($\rho_k = \rho$) et nous verrons par la suite qu'il existe des résultats théoriques pour guider le choix de l'utilisateur, soit on prend un pas ρ_k optimal, dans le sens que ρ_k réalise le minimum de la fonctionnelle

$$f_k \quad : \quad \rho \mapsto J(u_k + \rho d_k). \tag{4.5}$$

En d'autres termes, on minimise J sur la droite passant par u_k, de direction d_k. Dans le cas qui nous intéresse $J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle$, on remarque que

$$f_k(\rho) = \frac{\rho^2}{2}\langle Ad_k, d_k\rangle + \rho\langle Au_k - b, d_k\rangle + J(u_k).$$

C'est un polynôme de degré 2, avec un coefficient strictement positif pour le terme d'ordre 2. Il existe donc un point de minimum unique de f_k, ρ_k, caractérisé par

$$f_k'(\rho_k) = 0, \quad \text{soit} \quad \rho_k = \frac{\langle b - Au_k, d_k\rangle}{\langle Ad_k, d_k\rangle} = -\frac{\langle \nabla J(u_k), d_k\rangle}{\langle Ad_k, d_k\rangle}. \tag{4.6}$$

Remarque 4.3.1. *C'est parce que nous étudions un problème quadratique qu'il est possible de faire le raisonnement précédent. Dans un cas plus général, il est nécessaire de calculer (formellement ou numériquement) le gradient de J pour déterminer le minimum de la fonction f_k. Les problèmes inhérents à ce type de calcul ne sont pas étudiés ici ; ils ont donné naissance à une riche littérature, et sont entre autres abordés dans [8].*

A partir de là, nous sommes en mesure de décrire quelques méthodes numériques de minimisation.

4.3.1 Relaxation

Pour définir la méthode de **relaxation**, une base orthonormale $(e_i)_{1\leq i\leq n}$ de $\mathbb{R}^n$ étant donnée, on choisit la suite de directions de descente $d_0 = e_1,\ d_1 = e_2, \cdots$; si l'algorithme

n'a pas convergé après n itérations (supposition raisonnable !), on prend $d_n = e_1$, $d_{n+1} = e_2$ et ainsi de suite... Dans cette méthode, on choisira un pas optimal donné par la formule (4.6). Dans ce cas, l'algorithme devient

$$\left\| \begin{array}{l} \text{pour } \ l \geq 0,\ i \in \{1, \cdots, n\} \quad (k = ln + i - 1) \\ \rho_{ln+i-1} = \frac{\langle b - Au_{ln+i-1}, e_i \rangle}{\langle Ae_i, e_i \rangle}, \quad u_{ln+i} = u_{ln+i-1} + \rho_{ln+i-1} e_i. \end{array} \right. \tag{4.7}$$

Entre les deux itérés successifs u_{ln+i} et u_{ln+i-1}, on en déduit que seule la $i^{ème}$ composante diffère.

Comme seule une composante (sur n) évolue, on peut introduire la suite $(\tilde{u}_l)_{l \geq 0}$ telle que

$$\left\{ \begin{array}{ll} \tilde{u}_0 = u_0, & \\ \tilde{u}_1 = u_n, & \text{le résultat des } n \text{ premières itérations,} \\ \tilde{u}_2 = u_{2n}, & \text{le résultat des } n \text{ suivantes, etc.} \\ \quad \vdots & \end{array} \right.$$

Ainsi, toutes les composantes de $\tilde{u}_{l+1}$ sont *a priori* distinctes de celles de $\tilde{u}_l$. De plus, par construction, chaque composante est mise à jour une fois et une seule. Plus précisément, on a vu que la $i^{ème}$ composante est modifiée lorsque l'on considère la direction de descente e_i, ce qui donne, d'après (4.7) :

$$(\tilde{u}_{l+1} - \tilde{u}_l, e_i) = \rho_{ln+i-1}, \text{ et } \|\tilde{u}_{l+1} - \tilde{u}_l\|^2 = \sum_{i=1}^{n} {\rho_{ln+i-1}}^2 = \sum_{i=1}^{n} \|u_{ln+i} - u_{ln+i-1}\|^2. \tag{4.8}$$

Ces expressions seront fort utiles pour démontrer la proposition 4.3.1 ci-dessous. Avant de l'aborder, établissons le lemme :

Lemme 4.3.1. *Soit A une matrice symétrique, et $\lambda_{\min}$ et $\lambda_{\max}$ ses plus petite et plus grande valeurs propres. Alors*

$$\forall v \in \mathbb{R}^n, \qquad \lambda_{\min}\|v\|^2 \leq \langle Av, v \rangle) \leq \lambda_{\max}\|v\|^2 \ ; \tag{4.9}$$

$$\textit{si de plus } A \textit{ est positive,} \qquad \lambda_{\min}\|v\| \leq \|Av\| \leq \lambda_{\max}\|v\|. \tag{4.10}$$

Démonstration. On sait qu'il existe une base orthonormale de vecteurs propres de A ; notons-la $(p_i)_{1 \leq i \leq n}$. On pose $v = \sum_{i=1}^{n} v_i p_i$, et l'on effectue

$$\langle Av, v \rangle = \left\langle \sum_{i=1}^{n} v_i A p_i, \sum_{j=1}^{n} v_j p_j \right\rangle = \left\langle \sum_{i=1}^{n} \lambda_i v_i p_i, \sum_{j=1}^{n} v_j p_j \right\rangle = \sum_{i=1}^{n} \lambda_i {v_i}^2.$$

On en déduit alors (4.9). En effet :

$$\lambda_{\min}\|v\|^2 = \lambda_{\min} \sum_{i=1}^{n} {v_i}^2 \leq \langle Av, v \rangle \leq \lambda_{\max} \sum_{i=1}^{n} {v_i}^2 = \lambda_{\max}\|v\|^2.$$

Comme

$$\|Av\|^2 = \sum_{i=1}^{n} {\lambda_i}^2 {v_i}^2,$$

et on déduit de même (4.10), car

$${\lambda_{\min}}^2 \|v\|^2 = {\lambda_{\min}}^2 \sum_{i=1}^{n} {v_i}^2 \leq \|Av\|^2 \leq {\lambda_{\max}}^2 \sum_{i=1}^{n} {v_i}^2 = {\lambda_{\max}}^2 \|v\|^2.$$

□

Proposition 4.3.1. *Supposons que la matrice A est symétrique définie positive. Alors, la méthode de relaxation est convergente.*

Démonstration. — **Etape 1.** Commençons par borner $\|\tilde{u}_{l+1} - \tilde{u}_l\|$. Pour cela, on remarque que

$$J(u_k) - J(u_{k+1}) = f_k(0) - f_k(\rho_k) = \frac{\langle Ae_k, e_k \rangle}{2} {\rho_k}^2 \geq \frac{\lambda_{\min}}{2} {\rho_k}^2 = \frac{\lambda_{\min}}{2} \|u_k - u_{k+1}\|^2,$$

en se rappelant (4.5), i.e., la définition de f_k.
En conséquence, pour la suite $(\tilde{u}_l)_l$, on arrive à la minoration :

$$\begin{aligned} J(\tilde{u}_l) - J(\tilde{u}_{l+1}) &= J(u_{ln}) - J(u_{l(n+1)}) = \sum_{i=1}^{n} J(u_{ln+i-1}) - J(u_{ln+i}) \\ &\geq \frac{\lambda_{\min}}{2} \sum_{i=1}^{n} \|u_{ln+i-1} - u_{ln+i}\|^2 = \frac{\lambda_{\min}}{2} \|\tilde{u}_{l+1} - \tilde{u}_l\|^2. \end{aligned}$$

Par construction, la suite $(J(\tilde{u}_l))_l$ est décroissante et minorée. En conséquence, la différence de deux termes successifs $|J(\tilde{u}_l) - J(\tilde{u}_{l+1})|$ tend vers 0 lorsque l tend vers l'infini. D'après la majoration ci-dessus, on obtient $\lim_{l \to +\infty} \|\tilde{u}_{l+1} - \tilde{u}_l\| = 0$. De l'imbrication des suites $(u_k)_k$ et $(\tilde{u}_l)_l$, on en déduit également

$$\lim_{l \to +\infty} \|u_{ln+i} - \tilde{u}_l\| = 0, \text{ pour chaque } i \in \{1, \cdots, n\}. \tag{4.11}$$

— **Etape 2.** Convergence de $(\tilde{u}_l)_l$. Reprenons maintenant (4.10), avec $v = \tilde{u}_l - u$, il en résulte

$$\lambda_{\min} \|\tilde{u}_l - u\| \leq \|A(\tilde{u}_l - u)\|. \tag{4.12}$$

Nous allons évaluer le terme de droite :

$$\|A(\tilde{u}_l - u)\|^2 = \sum_{i=1}^{n} (A\tilde{u}_l - b)_i^2 = \sum_{i=1}^{n} \langle A\tilde{u}_l - b, e_i \rangle^2.$$

Revenons aux définitions (4.6)-(4.7), on obtient :

$$0 = f_k'(\rho_k) = \langle \nabla J(u_k + \rho_k d_k), d_k \rangle = \langle \nabla J(u_{k+1}), d_k \rangle = \langle A u_{k+1} - b, d_k \rangle .$$

Pour $k = ln + i - 1$, on trouve $0 = (Au_{ln+i} - b, e_i)$, soit $(b, e_i) = (Au_{ln+i}, e_i)$. Nous pouvons donc transformer l'expression du terme de droite de (4.12) en

$$\left\{ \sum_{i=1}^{n} \langle A(\tilde{u}_l - u_{ln+i}), e_i \rangle^2 \right\}^{1/2} .$$

En utilisant à nouveau l'inégalité (4.10), on arrive à

$$\sum_{i=1}^{n} \langle A(\tilde{u}_l - u_{ln+i}), e_i \rangle^2 \leq \sum_{i=1}^{n} \|A(\tilde{u}_l - u_{ln+i})\|^2 \leq \lambda_{\max}^2 \sum_{i=1}^{n} \|\tilde{u}_l - u_{ln+i}\|^2.$$

Et grâce à (4.12), on conclut que :

$$\|\tilde{u}_l - u\| \leq \frac{\lambda_{\max}}{\lambda_{\min}} \left(\sum_{i=1}^{n} \|\tilde{u}_l - u_{ln+i}\|^2 \right)^{1/2} . \tag{4.13}$$

Lorsque l tend vers l'infini, chaque terme de la somme tend vers 0. Par ailleurs, le nombre de termes de la somme est borné indépendamment de l. On arrive donc finalement à

$$\lim_{l \to +\infty} \|\tilde{u}_l - u\| = 0. \tag{4.14}$$

— **Etape 3.** Convergence de $(u_k)_k$. Nous venons donc de prouver la convergence de $(\tilde{u}_l)_l$ vers u. Bien évidemment, $(u_k)_k$ converge également vers u. En effet,

$$\|u_k - u\| \leq \|u_k - \tilde{u}_l\| + \|\tilde{u}_l - u\|, \text{ avec } l = E(k/n),$$

et (4.11), (4.14) permettent de conclure !

□

Exercice 4.3.1. *Le but de cet exercice est de montrer que l'algorithme de relaxation peut se réecrire sous une forme plus simple. On note $({u_k}^j)_{1 \leq j \leq n}$ les composantes du vecteur u_k.*

1. *Prouver que l'on peut écrire (4.7) sous la forme*

$$A_{i,i} {u_{k+1}}^i = b_i - \sum_{j \neq i} A_{i,j} {u_k}^j, \text{ pour } i \text{ tel que } k = ln + i - 1.$$

2. *On découpe A en trois parties : $A = D - E - F$, avec*

— *la partie diagonale :* $D_{i,i} = A_{i,i}$, $1 \leq i \leq n$, $D_{i,j} = 0$ *sinon ;*
— *la partie triangulaire inférieure :* $E_{i,j} = -A_{i,j}$, $1 \leq j < i \leq n$, $E_{i,j} = 0$ *sinon ;*
— *la partie triangulaire supérieure :* $F_{i,j} = -A_{i,j}$, $1 \leq i < j \leq n$, $F_{i,j} = 0$ *sinon.*

On revient aux itérés $\tilde{u}_l$*, c'est-à-dire ceux dont chaque composante est mise à jour une fois et une seule par itération. Montrer que*

$$(D - E)\tilde{u}_{l+1} = b + F\tilde{u}_l.$$

.

Dans le cas de minimisation quadratique, la méthode de relaxation correspond donc à la méthode itérative de **Gauss-Seidel***, de résolution d'un système linéaire.*
Cette méthode est étudiée en appendice B.3.)

4.3.2 Gradient à pas fixe, à pas optimal

Cette catégorie de méthodes a été conçue à partir de la réponse à la question suivante : dans quelle direction diminue-t-on le plus la valeur d'une fonctionnelle ? Ou, en termes plus mathématiques, si on pose

$$w_\varepsilon = u + \varepsilon d, \text{ avec } d \in \mathbb{R}^n, \ \|d\| = 1, \ \varepsilon > 0,$$

comment maximiser la différence $J(u) - J(w_\varepsilon)$ par rapport au paramètre d ? Pour cela, cf. (A.6), on écrit

$$J(u) - J(w_\varepsilon) = -\varepsilon \langle \nabla J(u), d \rangle + o(\varepsilon).$$

Lorsque ε est petit, la différence se comporte comme $-\varepsilon \langle \nabla J(u), d \rangle$ (si $\nabla J(u) \neq 0$), c'est-à-dire qu'elle est maximale pour

$$d = -\frac{\nabla J(u)}{\|\nabla J(u)\|}.$$

L'opposé de la direction du gradient est une direction *privilégiée.*

Gradient à pas fixe (GPF)

Dans un premier temps, nous allons donc considérer l'algorithme suivant :
— (Initialisation) Soit $u_0 \in \mathbb{R}^n$ un point initial. Soit $k = 0$ et $\rho > 0$ donné.

— (Calcul d'une direction) : Calculons :

$$d_k = -\nabla J(u_k) = b - Au_k$$

— (Mise à jour de l'itéré) : Pour $\rho_k = \rho$, posez :

$$u_{k+1} = u_k + \rho_k d_k.$$

— (Itération) : L'index k devient $k+1$.

Cette méthode est dite **méthode du gradient à pas fixe**.

Proposition 4.3.2. *Supposons que la matrice A est symétrique définie positive. La méthode de gradient à pas fixe est convergente, sous réserve que le pas de descente ρ vérifie :*

$$0 < \rho < \frac{2}{\lambda_{\max}}.$$

(Rappelons que $\lambda_{\max}$ désigne la plus grande valeur propre de A).

Démonstration. Elle est notablement plus simple que celle prouvant la convergence de la méthode de relaxation. Soit u le minimum de J sur $\mathbb{R}^n$. On a :

$$u_{k+1} - u = u_k + \rho(b - Au_k) - u = (I_n - \rho A)u_k + \rho Au - u = (I_n - \rho A)(u_k - u).$$

Nous allons majorer la norme de l'erreur à l'itération $k+1$ en fonction de celle de l'itération k, grâce à la relation ci-dessus, en reprenant la démonstration de (4.10) :

$$\begin{aligned}
(I_n - \rho A)v &= \sum_{i=1}^{n}(I_n - \rho A)v_i p_i = \sum_{i=1}^{n}(1 - \rho\lambda_i)v_i p_i \ ; \\
\|(I_n - \rho A)v\|^2 &= \left\langle \sum_{i=1}^{n}(1 - \rho\lambda_i)v_i p_i, \sum_{j=1}^{n}(1 - \rho\lambda_j)v_j p_j \right\rangle = \sum_{i=1}^{n}(1 - \rho\lambda_i)^2 {v_i}^2 \\
&\leq \max_i (1 - \rho\lambda_i)^2 \sum_{j=1}^{n} {v_j}^2 = \left\{ \max_i |1 - \rho\lambda_i| \right\}^2 \|v\|^2.
\end{aligned}$$

En regroupant les deux résultats, on trouve

$$\|u_{k+1} - u\| \leq \max_i |1 - \rho\lambda_i| \, \|u_k - u\|.$$

Si on note $\gamma_\rho = \max_i |1 - \rho\lambda_i|$, on a obtenu $\|u_{k+1} - u\| \leq \gamma_\rho \|u_k - u\|$. Par récurrence, on en déduit

$$\|u_k - u\| \leq {\gamma_\rho}^k \|u_0 - u\|. \tag{4.15}$$

Si γ_ρ est strictement plus petit que 1, on aura démontré la convergence. C'est ce que nous allons vérifier maintenant.

$$\begin{aligned} & \lambda_{\min} \leq \lambda_i \leq \lambda_{\max},\ 1 \leq i \leq n \\ \Longrightarrow\ & 1 - \rho\lambda_{\min} \geq 1 - \rho\lambda_i \geq 1 - \rho\lambda_{\max},\ 1 \leq i \leq n \\ \Longrightarrow\ & |1 - \rho\lambda_i| \leq \max(|1 - \rho\lambda_{\min}|, |1 - \rho\lambda_{\max}|),\ 1 \leq i \leq n. \end{aligned}$$

Puisque les bornes sur les valeurs propres $\lambda_{\min}$ et $\lambda_{\max}$ sont atteintes,

$$\gamma_\rho = \max(|1 - \rho\lambda_{\min}|, |1 - \rho\lambda_{\max}|). \tag{4.16}$$

Pour conclure, nous majorons γ_ρ, à l'aide des hypothèses sur A (semi-définie positive), et sur ρ :

$$\left\{ \begin{array}{l} 0 < \lambda_{\min} \leq \lambda_{\max} \\ 0 < \rho < \frac{2}{\lambda_{max}} \end{array} \right. \Longrightarrow \left\{ \begin{array}{l} -1 < 1 - \rho\lambda_{\min} < 1 \\ -1 < 1 - \rho\lambda_{\max} < 1 \end{array} \right. .$$

On vient donc de prouver que

$$\gamma_\rho < 1.$$

□

A partir de ce résultat, on constate que, pour appliquer la méthode du gradient à pas fixe, il faut connaître la valeur propre $\lambda_{\max}$ ou, au moins, une estimation de cette dernière.

Dans la preuve de la proposition 4.3.2, nous avons aussi montré que le taux de convergence du gradient à pas fixe est donné par (pour $0 < a \leq \rho \leq b < \frac{2}{\lambda_{\max}}$) :

$$C_{\mathrm{GPF}}(\rho) = \max(1 - \rho\lambda_{\min}; \rho\lambda_{\max} - 1).$$

Ce taux de convergence sera minimal pour une valeur de $\rho_{\min} = \frac{2}{\lambda_{\max}+\lambda_{\min}}$. Dans ce cas, la vitesse "maximale" de convergence est :

$$\boxed{R_{\mathrm{GPF}}(\rho_{\min}) := -ln\left(1 - \tfrac{2}{1+\kappa}\right), \qquad \text{avec} \quad \kappa = \tfrac{\lambda_{\max}}{\lambda_{\min}}.} \tag{4.17}$$

Méthode du gradient à pas optimal (GPO).

Cette méthode consiste à prendre comme direction de descente la direction opposée au gradient, et comme pas de descente le pas optimal donné par (4.6).

— (Initialisation) Soit $u_0 \in \mathbb{R}^n$ un point initial. Soit $k = 0$ et $\rho > 0$ donné.

— (Calcul d'une direction) : Calculons :

$$d_k = -\nabla J(u_k) = b - Au_k$$

— (Mise à jour de l'itéré) : Pour $\rho_k = \frac{\|d_k\|^2}{\langle Ad_k, d_k\rangle}$, posez :

$$u_{k+1} = u_k + \rho_k d_k. \tag{4.18}$$

— (Itération) : L'index k devient $k+1$.

Notons dès maintenant que, d'après (4.6), on a la propriété

$$0 = f_k'(\rho_k) = \langle \nabla J(u_{k+1}), d_k\rangle = -\langle d_{k+1}, d_k\rangle . \tag{4.19}$$

En clair, deux directions *consécutives* de descente sont orthogonales.

Proposition 4.3.3. *Si la matrice A est symétrique définie positive, alors la méthode de gradient à pas optimal est convergente.*

Démonstration. Majorons pour commencer la norme $\|u_k - u\|$:

$$\begin{aligned}\lambda_{\min}\|u_k - u\|^2 &\leq \langle A(u_k - u), u_k - u\rangle = \langle Au_k - b, u_k - u\rangle \\ &= -\langle d_k, u_k - u\rangle \leq \|d_k\|\,\|u_k - u\|.\end{aligned}$$

On infère

$$\|u_k - u\| \leq \frac{1}{\lambda_{\min}}\|d_k\|.$$

Utilisons maintenant l'orthogonalité entre deux directions consécutives de descente :

$$\begin{aligned}\|d_k\|^2 &= \langle d_k - d_{k+1}, d_k\rangle = \langle A(u_{k+1} - u_k), d_k\rangle \leq \|A(u_{k+1} - u_k)\|\,\|d_k\| \\ &\leq \lambda_{\max}\|u_{k+1} - u_k\|\,\|d_k\|.\end{aligned}$$

Ainsi

$$\|d_k\| \leq \lambda_{\max}\|u_{k+1} - u_k\|.$$

On arrive alors à la majoration

$$\|u_k - u\| \leq \frac{\lambda_{\max}}{\lambda_{\min}}\|u_{k+1} - u_k\|. \tag{4.20}$$

Si $d_k = 0$, $u_{k+1} - u_k = 0$ et donc $u_k = u$, la méthode a convergé. Nous pouvons donc supposer $d_k \neq 0$. L'égalité $u_{k+1} - u_k = \rho_k d_k$ implique

$$|\rho_k| = \frac{\|u_{k+1} - u_k\|}{\|d_k\|}.$$

Puis

$$J(u_k) - J(u_{k+1}) = \frac{\langle Ad_k, d_k\rangle}{2}{\rho_k}^2 = \frac{\langle Ad_k, d_k\rangle}{2\|d_k\|^2}\|u_{k+1} - u_k\|^2$$
$$\geq \frac{\lambda_{\min}}{2}\|u_k - u_{k+1}\|^2.$$

Comme $(J(u_k))_k$ est décroissante et minorée (on a supposé l'existence d'un minimum), la différence de deux termes consécutifs tend vers 0. Donc obligatoirement nous avons $\lim_{k\to\infty}\|u_{k+1} - u_k\| = 0$. La convergence de $(u_k)_k$ vers u découle donc de la propriété $\lim_{k\to+\infty}\|u_{k+1} - u_k\| = 0$ en se rappelant (4.20).

Enfin, on peut même combiner les inégalités pour obtenir l'estimation suivante liant directement le progress sur la fonction objectif et la distance à l'optimum.

$$\|u_k - u\| \leq \sqrt{2}\frac{\lambda_{\max}}{\lambda_{\min}^{3/2}}(J(u_{k+1}) - J(u_k))^{1/2}. \tag{4.21}$$

□

Dans le cas de la fonctionnelle quadratique J, on a vu que le calcul de la valeur optimale ρ_k ne présente aucune difficulté. Il peut en être tout autrement dans le cas général.

4.3.3 Gradient conjugué

Nous avons remarqué que pour la méthode du gradient à pas optimal, deux directions successives de descente sont orthogonales. Par contre, il n'y a aucune raison pour que trois (ou plus) directions de descente soient orthogonales entre elles. Ainsi, on peut théoriquement "revenir" dans des directions déjà explorées... Le principe de la **méthode du/des gradient(s) conjugué(s)** est de construire une suite de *directions de recherche* que l'on garde en mémoire, pour essayer d'éviter les retours. Pour cela, si $u_1, \ldots, u_k$ ont déjà été calculés, et si tous les gradients $g_l = \nabla J(u_l)$, $0 \leq l \leq k$ sont non nuls, on cherche u_{k+1} tel que

$$J(u_{k+1}) = \min_{v\in u_k+G_k} J(v), \text{ avec } G_k = \text{Vect}(g_0, g_1, \cdots, g_k).$$

NB. Pour respecter l'esprit des méthodes de gradient, on conserve les directions "optimales", c'est-à-dire les gradients de J aux itérés successifs.

Le principe de la méthode est très attrayant, puisqu'on espère éviter les redondances dans le choix de la direction de descente, i. e., on espère que

$$\dim(G_{k+1}) = \dim(G_k) + 1, \ k = 0, 1, \cdots.$$

Bien sûr, ceci n'est nullement garanti (il faut et il suffit que $g_{k+1} \notin G_k$)... Par ailleurs, la construction de la suite des espaces vectoriels $(G_k)_{k\geq 0}$, et la résolution des problèmes posés sur $u_k + G_k$, semblent très coûteuses, puisqu'on doit gérer des espaces vectoriels dont la dimension peut fort bien devenir comparable à n.

Heureusement, dans le cas présent, et c'est la "magie" de la méthode du gradient conjugué, nous allons vérifier qu'aucun de ces deux problèmes ennuyeux n'en est un !

Tout d'abord, étant donné que u_{k+1} est minimum du problème suivant :

$$J(u_{k+1}) = \min_{v \in u_k + G_k} J(v),$$

alors la condition d'optimalité (Théorème 3.3.3) de u_{k+1} implique que :

$$\begin{cases} u_{k+1} \in u_k + G_k, \\ \langle g_{k+1}, v - u_{k+1} \rangle \geq 0 \quad \forall v \in u_k + G_k \end{cases} \tag{4.22}$$

où on a noté $g_{k+1} := \nabla J(u_{k+1}) = Au_{k+1} - b$.

Par ailleurs, G_k est un (sous-)espace vectoriel et donc $w \in G_k$ implique $-w \in G_k$. Il en suit donc, que pour tout $w \in G_k$ nous avons $\pm w + u_{k+1} \in u_k + G_k$. Donc pour tout $w \in G_k$, le vecteur $w + u_{k+1}$ appartient à l'espace $u_k + G_k$, et d'après (4.22), on a :

$$\begin{aligned} \langle g_{k+1}, w \rangle &= \langle g_{k+1}, (w + u_{k+1}) - u_{k+1} \rangle \geq 0, \\ \langle g_{k+1}, -w \rangle &= \langle g_{k+1}, (-w + u_{k+1}) - u_{k+1} \rangle \geq 0. \end{aligned}$$

On en déduit, que $\langle g_{k+1}, w \rangle = 0$ pour tout $w \in G_k$, ce qui signifie que g_{k+1} est orthogonale à G_k. Etant donné que cet espace contient déjà $\{g_0, ..., g_k\}$ il en suit également :

$$\langle g_{k+1}, g_l \rangle = 0, \ \forall l = 0, ..., k.$$

Par récurrence, nous obtenons également :

$$\langle g_i, g_j \rangle = 0, \ 0 \leq i < j \leq k+1. \tag{4.23}$$

Il devient donc immédiatement clair que la méthode devrait s'arrêter en au plus n itérations car nous pouvons difficilement générer plus que n éléments orthogonaux en $\mathbb{R}^n$! Présentons cet observation de manière formelle :

Proposition 4.3.4. *La méthode du gradient conjugué converge en n itérations au plus.*

Démonstration. Au bout de $n-1$ itérations, si aucun des $\nabla J(u_k) = g_k, 0 \leq k \leq n-1$, ne s'annule, on a construit une famille libre de $\mathbb{R}^n$ à n éléments ; ou, en d'autres termes, une base de $\mathbb{R}^n$! Comme $g_n = \nabla J(u_n)$ est orthogonal à ces n vecteurs, il est nécessairement nul, ce qui signifie que $u_n = u$. □

Etudions maintenant les aspects *pratiques* de l'algorithme, et notamment la gestion des espaces $(G_k)_k$. Commençons par la propriété suivante, qui porte sur les directions de descentes $(\delta_k)_k$, que nous définissons :

$$\delta_k := u_{k+1} - u_k.$$

Lemme 4.3.2. *Les directions $(\delta_k)_k$ sont telles que*

$$\langle A\delta_i, \delta_j \rangle = 0 \tag{4.24}$$

pour tout $i > j$.

Démonstration. Nous commençons par rappeler que $\delta_k = u_{k+1} - u_k \in G_k$ comme résultat de (4.22). Il en est évidemment de même pour $l = 0, ..., k$, i.e., $\delta_l \in G_l$. Par conséquence, δ_l peut s'écrire comme une certaine combinaison linéaire des vecteurs $\{g_0, ..., g_l\}$ qui "spannent" G_l. Il est donc aussi clair que

$$\langle g_{k+1}, d_l \rangle = 0,$$

pour $l \leq k$. En effet, rappelons le, que g_{k+1} est orthogonal à G_k et $G_l \subseteq G_k$ pour $l \leq k$. Toutefois

$$g_{k+1} = Au_{k+1} - b = Au_k - b + A\delta_k = g_k + A\delta_k.$$

Donc pour $l \leq k - 1$,

$$\langle A\delta_k, \delta_l \rangle = \langle g_{k+1} - g_k, \delta_l \rangle = \langle g_{k+1}, \delta_l \rangle - \langle g_k, \delta_l \rangle = 0.$$

□

Définition 4.3.1. *On dit que des directions (non nulles) $(\delta_k)_k$ vérifiant (4.24) sont* **conjuguées** *par rapport à la matrice A.*

Il nous reste dans (4.22) le calcul concret de u_{k+1}. Pour cela, nous allons nous appuyer sur la représentation des directions δ dans la base des espaces G_k. Observons donc que nous pouvons écrire pour tout k :

$$\delta_k = \sum_{l=0}^{k} \beta_l^k g_l, \tag{4.25}$$

où $\beta_k^k \neq 0$.

Pour tout $m \leq k-1$, nous pouvons utiliser le Lemme 4.3.2 et obtenir :

$$\begin{aligned} 0 &= \langle A\delta_k, \delta_m \rangle = \langle \delta_k, \delta_m \rangle = \langle \delta_k, g_{m+1} - g_m \rangle \\ &= \sum_{l=0}^{k} \beta_l^k \langle g_l, g_{m+1} - g_m \rangle = \beta_{m+1}^k \|g_{m+1}\|^2 - \beta_m^k \|g_m\|^2 . \end{aligned}$$

On en déduit donc la relation :

$$\beta_m^k = \beta_{m+1}^k \frac{\|g_{m+1}\|^2}{\|g_m\|^2},$$

avec comme cas particulier $m = k-1$:

$$\beta_m^k = \beta_k^k \frac{\|g_k\|^2}{\|g_m\|^2}.$$

Nous pouvons alors immédiatement conclure que δ_k s'écrit également de la manière suivante :

$$\delta_k = \beta_k^k \left\{ \sum_{l=0}^{k} \frac{\|g_k\|^2}{\|g_l\|^2} g_l \right\}. \tag{4.26}$$

Essayons maintenant d'écrire δ_k comme un increment de δ_{k-1} en utilisant cette dernière expression.

$$\begin{aligned} \delta_k &= \beta_k^k \left\{ \sum_{l=0}^{k} \frac{\|g_k\|^2}{\|g_l\|^2} g_l \right\} = \beta_k^k \left(\frac{\|g_k\|^2}{\|g_{k-1}\|^2} \sum_{l=0}^{k-1} \frac{\|g_{k-1}\|^2}{\|g_l\|^2} g_l + g_k \right) \\ &= \beta_k^k \left(\frac{\|g_k\|^2}{\|g_{k-1}\|^2} \frac{1}{\beta_{k-1}^{k-1}} \delta_{k-1} + g_k \right). \end{aligned} \tag{4.27}$$

En normalisant nous n'aurons pas besoin de représenter "δ_k" dans la base de G_k. En effet, posons comme "direction" :

$$d_k = -\frac{1}{\beta_k^k} \delta_k, \tag{4.28}$$

pour tout $k \geq 0$. Nous déduisons alors des équations (4.26) et (4.27) :

$$\begin{aligned} d_0 &= -g_0 \\ d_k &= \frac{\|g_k\|^2}{\|g_{k-1}\|^2} d_{k-1} - g_k, k \geq 1. \end{aligned} \tag{4.29}$$

Après tout ce travail préparatoire nous pouvons poser formellement l'algorithme :

— (Initialisation) Soit $u_0 \in \mathbb{R}^n$ un point initial. Soit $k = 0$ donné et calculons $g_0 = Au_0 - b$ ainsi que $d_0 = -g_0$. Soit $\varepsilon > 0$ une tolérance d'arrêt.
— (Test d'Arrêt) : Si $\|g_k\| < \varepsilon$ alors u_k est solution et l'algorithme se termine.
— (Mise à jour de l'itéré) : Calculez dans l'ordre

$$\begin{aligned}
\rho_k &= \frac{\|g_k\|^2}{\langle Ad_k, d_k\rangle} \\
u_{k+1} &= u_k + \rho_k d_k \\
g_{k+1} &= \nabla J(u_{k+1}) = Au_{k+1} - b \\
\beta_k &= \frac{\|g_{k+1}\|^2}{\|g_k\|^2} \\
d_{k+1} &= -g_{k+1} + \beta_k d_k
\end{aligned}$$

— (Itération) : L'index k devient $k+1$ et retour au test d'arrêt.

Notre façon d'écrire cet algorithme le rapproche de celle d'écrire un algorithme de descente. L'observation suivante justifie cette manière de voir.

Remarque 4.3.2 (Gradient conjugé comme méthode de descente)**.** *Regardons ce qu'il advient si l'on minimise la fonctionnelle J sur la droite passant par u_k de direction d_k :*

— $\min_{\rho\in\mathbb{R}} J(u_k + \rho d_k) \geq \min_{v\in u_k+G_k} J(v) = J(u_{k+1})$*, puisque* $d_k \in G_k$.
— u_{k+1} *se trouve sur la droite passant par* u_k *de direction* d_k *(*$u_{k+1} = u_k - \beta_k^k d_k$*).*

En d'autres termes, le minimum est bien atteint en $u_{k+1} = u_k + \rho_k d_k$*, avec,*

$$\rho_k = \frac{\langle b - Au_k, d_k\rangle}{\langle Ad_k, d_k\rangle} = -\beta_k^k.$$

Les propriétés fondamentales de l'algorithme ci-dessus sont au nombre de deux :

1. La minimisation est effective sur un sous-espace vectoriel dont la dimension croît à chaque itération. En conséquence, la solution u est calculée en n itérations au plus (aux erreurs de calcul près !).
2. Les directions $(d_k)_k$ sont faciles à calculer, et il suffit d'en conserver deux en mémoire à tout instant de l'algorithme.

Pour accélérer la vitesse de convergence, on peut *préconditionner* le système linéaire, ce qui conduit en pratique à une réduction notable du nombre d'itérations. Plus précisement, Soit A une matrice symétrique définie positive *mal conditionnée* (i.e., $\text{cond}(A) >> 1$) . On cherche le minimum u dans $\mathbb{R}^n$ de la fonctionnelle J définie par $J(v) := \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle$, ou de manière équivalente on cherche à résoudre le système $Au = b$, et supposons qu'il existe une matrice C inversible, et telle que $\text{cond}(CAC^{\mathsf{T}})$

soit petit. Dans ce cas, le système préconditionné consiste à déterminer la solution v de $CAC^{\mathsf{T}}v = Cb$, le minimum u se déduit alors de v par $u = C^{\mathsf{T}}v$.

Cependant, comme indiqué à la section 4.1, il faut veiller à ne pas trop augmenter le coût de calcul par itération... Ce type de considération a donné naissance à une littérature considérable (citons notamment [9, 11]).

4.3.4 Extensions

Il est prouvé dans [4] que les méthodes de relaxation et de gradient à pas optimal, ou à pas fixe, sont applicables dans un espace de Hilbert, sous réserve que la fonctionnelle J vérifie certaines propriétés. En clair, la fonctionnelle J n'est qu'un cas très particulier, mais fort utile, puisqu'elle permet de construire des méthodes de résolution de systèmes linéaires. Précisément, si la fonctionnelle J est $\mathcal{C}^1$ et α-convexe, avec une différentielle Lipschitzienne, les méthodes convergent. Ceci signifie :

— J α-convexe : cf. remarque 2.4.2, points (viii) et (ix).

— dJ lipschitzienne : $\exists M > 0$, $\forall u, v \in \mathbb{V}$, $\|dJ(u) - dJ(v)\| \leq M\|u - v\|$.

Sous ces conditions, bien que la démonstration est plus ardue, nous pouvons retrouver les résultats des propositions 4.3.1 et 4.3.3.

Pour ce qui est de la méthode du gradient conjugué, certaines adaptations sont également possibles, dans le cas d'une fonctionnelle plus générale, sous réserve toutefois de modifications de l'algorithme (cf. [14], et [8] pour une discussion détaillée.)

Si maintenant on considère la résolution d'un système linéaire, dont la matrice n'est pas symétrique, il est *impossible* de conserver à la fois les deux propriétés remarquables de l'algorithme du gradient conjugué, à savoir la convergence en n itérations au plus, associée à l'utilisation de récurrences de taille constante (voir [7]) ! Il faut, au choix, soit conserver toutes les directions précédentes de descente, ce qui accroît notablement le coût calcul, soit ne garder que les p (pour p fixé, petit devant n) dernières directions, et raisonner dans

$$u_k + \mathrm{Vect}(g_k, \cdots, g_{k-p+1}).$$

Nous renvoyons le lecteur intéressé à [16], article dans lequel la méthode GMRES (**G**eneralized **M**inimum **RES**idual algorithm.) a été introduite pour résoudre des systèmes linéaires, de matrice non symétrique.

Chapitre 5

Conditions nécessaires d'optimalité II

5.1 Introduction

Commençons par rappeler le Corollaire 3.3.2. Ce corollaire nous indique que si $\bar{u}$ est un optimum local du problème

$$\min_{u \in K} J(u),$$

pour $J : \mathbb{R}^n \to \mathbb{R}$ Fréchet differentiable et $K \subseteq \mathbb{R}^n$ convexe et fermé, alors

$$0 \in \nabla J(\bar{u}) + \mathrm{N}_K(\bar{u}).$$

D'une certaine façon, on peut se dire que nous en avons fini avec la caractérisation d'une solution en termes de conditions d'optimalité. Toutefois, la condition est peu explicite lorsque K est lui même donné par un jeu d'inégalités. Par exemple considérons une fonction $g : \mathbb{R}^n \to \mathbb{R}^m$, convexe et $K = \{u \in \mathbb{R}^n \ : g(u) \leq 0\}$. Peut-on alors exprimer N_K d'une manière plus simple ? Peut-on lier l'expression de N_K à g, ou ∇g^* (le Jacobien) ?

L'objectif du chapitre est d'avancer sur ce chemin. Commençons par un exemple simple, où le calcul direct est possible :

Lemme 5.1.1. *Soient $a, b \in \mathbb{R}^n, a \leq b$ donnés. Alors pour $\bar{u} \in K = [a, b]$:*

$$\mathrm{N}_K(\bar{u}) := \{\lambda \in \mathbb{R}^n \ : \left\{ \begin{array}{lll} \lambda_i \leq 0 & si & \bar{u}_i = a_i \\ \lambda_i = 0 & si & a_i < \bar{u}_i < b_i \\ \lambda_i \geq 0 & si & \bar{u}_i = b_i, \end{array} \right\}, i = 1, ..., n\}. \tag{5.1}$$

La vérification de ce résultat élémentaire est laissé au lecteur.

Nous pouvons immédiatement utiliser le Lemma 5.1.1 pour dériver des conditions plus explicites pour le problème :

$$\min_{u\in[a,b]} J(u).$$

En effet, pour $\bar{u}$ une solution locale (et donc obligatoirement $\bar{u} \in [a,b]$!) il existe $\lambda \in \mathbb{R}^n$ tel que :

$$\nabla J(\bar{u}) + \lambda = 0,$$

et pour $i = 1, ..., m$: $\lambda_i \geq 0$ si $\bar{u}_i = b_i$, $\lambda_i = 0$ pour $\bar{u}_i \in (a_i, b_i)$ et $\lambda_i \leq 0$ pour $\bar{u}_i = a_i$. Nous pouvons aussi déduire qu'il existe $\lambda_1, \lambda_2 \in \mathbb{R}^n_+$ tel que :

$$\nabla J(\bar{u}) + \lambda_2 - \lambda_1 = 0,$$

et $\lambda_2^\intercal(b - \bar{u}) = 0$, $\lambda_1^\intercal(\bar{u} - a) = 0$ (Réfléchissez bien pourquoi !).

Pour lier le contenu au chapitre, et à guise de mise en garde, on a le résultat suivant (dont la démonstration peut se faire dans un cadre non-convexe plus générale – faisant intervenir d'autres cônes–, dont le scope dépasse donc largement le cadre du cours).

Proposition 5.1.1. *Soit $g : \mathbb{R}^n \to \mathbb{R}^m$ une fonction continûment differentiable et $K = \{u \in \mathbb{R}^n : g(u) \leq 0\}$ un ensemble convexe et fermé. Supposons également que à $\bar{u} \in \mathbb{R}^n$ la condition :*

$$\nabla g(\bar{u})^* \, \mathrm{N}_{(-\infty,0]^m}(g(\bar{u})) = 0 \Longrightarrow \mathrm{N}_{(-\infty,0]^m}(g(\bar{u})) = \{0\}, \tag{5.2}$$

est satisfaite, c.à.d., tout $\lambda \in \mathrm{N}_{(-\infty,0]^m}(g(\bar{u}))$ tel que $\nabla g(\bar{u})^*\lambda = 0$ doit être nulle. Alors à un tel $\bar{u}$ nous avons :

$$\mathrm{N}_K(\bar{u}) = \nabla g(\bar{u})^* \, \mathrm{N}_{(-\infty,0]^m}(g(\bar{u})). \tag{5.3}$$

La condition (5.2) est appelé une "qualification" de contraintes et est nécessaire !

Nous allons voir dans la suite du chapitre, que la situation est bien plus simple lorsque K est un polyèdre.

5.2 Contraintes d'égalité affines

Supposons, ce qui est un cas relativement courant en pratique, que l'ensemble K soit défini par :

$$K = \{v \in \mathbb{R}^n \ : \ C\,v = f\},$$

avec C une matrice $p \times n$, et f un élément de $\mathbb{R}^p$. On suppose $K \neq 0$. Notons que K est aussi l'ensemble des v vérifiant les p contraintes :

$$\begin{cases} C_{11}v_1 + \cdots + C_{1n}v_n = f_1 \\ \vdots \\ C_{p1}v_1 + \cdots + C_{pn}v_n = f_p \end{cases}$$

Soit J une fonctionnelle continue sur $\mathbb{V}$. D'après le théorème 3.3.3, puisque K *est un convexe non vide*, si $u \in K$ est un point de minimum local de J sur K, et si J est différentiable en u, on a nécessairement

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

Or, on a $Cu = f$ et $Cv = f$, donc $C(v - u) = 0$. Cela veut dire $u - v \in \operatorname{Ker} C$. Donc notre inéquation d'Euler équivaut à dire

$$\langle \nabla J(u), d \rangle \geq 0 \ \forall d \in \operatorname{Ker} C,$$

où encore de manière équivalente (car $d \in \operatorname{Ker} C$ ssi $-d \in \operatorname{Ker} C$) :

$$\nabla J(u) \in [\operatorname{Ker} C]^{\perp}.$$

Ceci permet d'affirmer, grâce au résultat $\operatorname{Im} C^{\mathsf{T}} = (\operatorname{Ker} C)^{\perp}$, (voir Lemme B.1.2), que $\nabla J(u)$ est dans l'image de la transposée de C^{T} noté $\operatorname{Im} C^{\mathsf{T}}$:

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

On a donc démontré le résultat suivant, dit de Karush, Kuhn et Tucker (K.K.T.).

Théorème 5.2.1. *Soit C une matrice $p \times n$, et f un élément de $\mathbb{R}^p$, K le sous-espace défini par $K := \{v \in \mathbb{R}^n \ : \ C\,v = f\}$, u un point de K et J une fonctionnelle sur K. On suppose que J est différentiable en u. Si u est un point de minimum local de J sur K, on a nécessairement*

$$\begin{cases} \exists \lambda \in \mathbb{R}^p, \quad \nabla J(u) + C^{\mathsf{T}}\lambda = 0, \\ Cu - f = 0. \end{cases} \tag{5.4}$$

Remarque 5.2.1. *Dans le théorème précédent le vecteur λ est unique si C est surjectif*[1]*. En effet, si λ_1 et λ_2 conviennent, alors $C^{\mathsf{T}}(\lambda_1 - \lambda_2) = 0$ avec C^{T} injectif, donc $\lambda_1 = \lambda_2$.*

1. C est surjectif $\iff$ C^{T} est injectif $\iff \operatorname{rg}(C) = \min(p, n)$.

Afin de conclure cette partie, faisons le lien avec le cône normal. Notre argumentation permet en fait de déduire la caractérisation suivante :

Proposition 5.2.1. *Soit $\bar{u} \in K$ donné. Alors nous avons*

$$\mathrm{N}_K(\bar{u}) = C^{\mathsf{T}}\,\mathrm{N}_{\{0\}}(C\bar{u} - f) = \mathrm{Im}\, C^{\mathsf{T}}.$$

Démonstration. La deuxième égalité est évidente, car en effet $\bar{u} \in K$ si et seulement si $C\bar{u} - f = 0$ et alors $\mathrm{N}_{\{0\}}(C\bar{u} - f) = \mathbb{R}^p$. Nous pouvons alors nous concentrer sur l'autre égalité. Prenons un $\lambda \in \mathbb{R}^p$ arbitraire, alors en se rappelant $v \in K$ implique $v - \bar{u} \in \mathrm{Ker}\, C$, nous obtenons :

$$\left\langle C^{\mathsf{T}}\lambda, v - \bar{u}\right\rangle = \langle \lambda, C(v - \bar{u})\rangle = 0 \leq 0,$$

et ce quelque soit $v \in K$. Donc par définition $C^{\mathsf{T}}\lambda \in \mathrm{N}_K(\bar{u})$, où encore $\mathrm{Im}\, C^{\mathsf{T}} \subseteq \mathrm{N}_K(\bar{u})$. Pour dériver l'inclusion inversée, rappelons nous que $w \in \mathrm{Ker}\, C$ implique $\bar{u} + w \in K$. Donc pour $x^* \in \mathrm{N}_K(\bar{u})$ nous avons :

$$0 \geq \langle x^*, (\bar{u} + w) - \bar{u}\rangle = \langle x^*, w\rangle\,,$$

ce qui avec le fait que aussi $-w \in \mathrm{Ker}\, C$ donne $\langle x^*, w\rangle = 0$ pour tout $w \in \mathrm{Ker}\, C$. Donc $\mathrm{N}_K(\bar{u}) \subseteq \mathrm{Ker}\, C^{\perp} = \mathrm{Im}\, C^{\mathsf{T}}$ en utilisant de nouveau le Lemme B.1.2. □

Ce dernier règle de calcul permet donc immédiatement de dériver le Théorème 5.2.1 du Corollaire 3.3.2.

5.2.1 Application : fonctionnelle quadratique

La fonctionnelle est, dans cette sous-section, quadratique en v : J est définie pour des éléments de $\mathbb{R}^n$, par

$$J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle + c, \tag{5.5}$$

avec A une matrice *symétrique* de $\mathbb{R}^{n\times n}$, b un vecteur de $\mathbb{R}^n$ et c un réel. On considère ses variations sur l'espace affine

$$K = \{v \in \mathbb{R}^n \ :\ C\,v = f\},$$

avec C appartenant à $\mathbb{R}^{p\times n}$, et f un élément de $\mathbb{R}^p$. D'après l'expression (5.4), si u est un point de minimum de J sur K, alors

$$\exists \lambda \in \mathbb{R}^p \text{ tel que } \left\{ \begin{array}{l} A\,u + C^{\mathsf{T}}\lambda = b \\ C\,u = f \end{array} \right. .$$

En d'autres termes,

Corollaire 5.2.1. *Soit J défini par* (5.5) *avec A une matrice symétrique, et soit K le sous-espace non vide $K = \{v \mid Cv = f\}$. Si u est un point de minimum de J sur K, alors il existe $\lambda \in \mathbb{R}^p$ tel que le couple (u, λ) de $\mathbb{R}^n \times \mathbb{R}^p$ soit solution du système linéaire*

$$\left(\begin{array}{cc} A & C^{\mathsf{T}} \\ C & 0 \end{array} \right) \left(\begin{array}{c} u \\ \lambda \end{array} \right) = \left(\begin{array}{c} b \\ f \end{array} \right). \tag{5.6}$$

Enfin, pour en finir ici avec les problèmes avec contraintes d'égalités, l'on suppose cette fois que la matrice A est *symétrique et positive (semi-)défini* (ssi $\langle Av, v\rangle > (\text{semi} : \geq)0$ pour tout $v \in \mathbb{V}$.).

On a alors le

Théorème 5.2.2. *Supposons que A est symétrique positive. le vecteur u est un point de minimum de J sur K si, et seulement si, il existe un élément λ de $\mathbb{R}^p$ tel que le couple (u, λ) soit solution de*

$$\left(\begin{array}{cc} A & C^{\mathsf{T}} \\ C & 0 \end{array} \right) \left(\begin{array}{c} u \\ \lambda \end{array} \right) = \left(\begin{array}{c} b \\ f \end{array} \right). \tag{5.7}$$

Si de plus A est positive définie *et C est surjectif, le système linéaire (5.7) admet une solution unique. En d'autres termes, il existe un point de minimum global de J sur K et un seul.*

Démonstration. Si u est un point de minimum, on applique le corollaire 5.2.1. Par ailleurs J est convexe et la condition est donc également suffisante.

On suppose ici que A est symétrique définie-positive. La matrice $\left(\begin{array}{cc} A & C^{\mathsf{T}} \\ C & 0 \end{array} \right)$ appartient à $\mathbb{R}^{(n+p)\times(n+p)}$. Pour prouver que le système linéaire (5.7) admet une solution unique, il suffit de vérifier que le noyau de l'application linéaire associée est réduit à $\{0\}$. Soit donc un couple (u, λ) de $\mathbb{R}^n \times \mathbb{R}^p$ tel que

$$\left(\begin{array}{cc} A & C^{\mathsf{T}} \\ C & 0 \end{array} \right) \left(\begin{array}{c} u \\ \lambda \end{array} \right) = \left(\begin{array}{c} 0 \\ 0 \end{array} \right).$$

Comme A est inversible, on infère, par implications successives que

— $u = -A^{-1}C^{\mathsf{T}}\lambda$ (première ligne) ;
— $CA^{-1}C^{\mathsf{T}}\lambda = 0$ (seconde ligne) ; $(CA^{-1}C^{\mathsf{T}}\lambda, \lambda)_p = 0$ (produit scalaire par λ) ; $(A^{-1}C^{\mathsf{T}}\lambda, C^{\mathsf{T}}\lambda) = 0$ (transposition) ; Comme A est symétrique définie-positive, A^{-1} l'est également, et ainsi

$$C^{\mathsf{T}}\lambda = 0.$$

On note que l'application linéaire associée à C^{T} va de $\mathbb{R}^p$ dans $\mathbb{R}^n$, et qu'elle est de rang p. Par conséquent, $\dim(Ker(C^{\mathsf{T}})) = 0$, dont on déduit que $\lambda = 0$.
— Finalement, $u = 0$ par retour à la première ligne du système linéaire.

La conclusion suit. □

Exercice 5.2.1. *On reprend $J(v) = \frac{1}{2}\langle Av, v\rangle) - \langle b, v\rangle + c$, définie sur $\mathbb{R}^n$. A quelle(s) condition(s) J est-elle convexe, strictement convexe, α-convexe ?*

Exercice 5.2.2. *Soit J une fonctionnelle α-convexe et différentiable sur $\mathbb{R}^n$. Montrer que J admet un minimum global, et le caractériser.*

5.3 Contraintes d'inégalité affines

On considère dans toute cette section que l'ensemble des contraintes K est donné par :

$$K := \{v \in \mathbb{V}, Cv \leq f\}, \tag{5.8}$$

où C est une matrice $p \times n$. On considère à nouveau aussi une fonctionnelle différentiable J sur $\mathbb{V}$ à valeurs dans $\mathbb{R}$.

Notre première étude sera celle des contraintes "actives" :

Lemme 5.3.1. *Soit $g : \mathbb{R}^n \to \mathbb{R}^p$ une fonction continu et considérons $\bar{K} = \{u \in \mathbb{R}^n \ : g(u) \leq 0\}$ ainsi que pour un $\bar{u} \in \bar{K}$, l'ensemble des indices actives :*

$$I(\bar{u}) := \{j = 1, ..., p \ : g_j(\bar{u}) = 0\}. \tag{5.9}$$

Alors il existe un voisinage U de $\bar{u}$, tel que $I(v) \subseteq I(\bar{u})$ pour tout $v \in U \cap \bar{K}$.

Démonstration. L'argument se fait par continuité. En effet pour chaque $j \in \{1, ..., p\} \setminus I(\bar{u})$, $g_j(\bar{u}) < 0$. Nous pouvons donc définir :

$$\bar{\varepsilon} = \min_{j \in \{1,...,p\} \setminus I(\bar{u})} -g_j(\bar{u})$$

et observer pour les j identifié ci-dessus que $g_j(\bar{u}) \leq -\bar{\varepsilon} < 0$. Par continuité de g nous pouvons obtenir un voisinage U de $\bar{u}$ sur lequel (pour les mêmes j), $g_j(v) \leq -\frac{1}{2}\bar{\varepsilon}$ pour tout $v \in U$. Le résultat suit donc. □

Nous aurons également besoin d'un Théorème des alternatives, le fameux lemme de Farkas.

Lemme 5.3.2 (Farkas)**.** *Pour une matrice C $p \times n$ et un vecteur $f \in \mathbb{R}^p$, il existe bien une solution $u \in \mathbb{R}^n$ de*

$$Cu \geq f, u \geq 0$$

ou il existe une solution $\lambda \in \mathbb{R}^p$ de

$$C^\mathsf{T}\lambda \leq 0, f^\mathsf{T}\lambda > 0, \lambda \geq 0,$$

mais les deux ne peuvent être vrai simultanément.

Corollaire 5.3.1. *Pour une matrice A $p \times n$ et un vecteur $b \in \mathbb{R}^p$, il existe bien une solution $x \in \mathbb{R}^n$ de*

$$Ax \leq b$$

ou il existe une solution $\lambda \in \mathbb{R}^p$ de

$$A^\mathsf{T}\lambda = 0, b^\mathsf{T}\lambda < 0, \lambda \geq 0,$$

mais les deux ne peuvent être vrai simultanément.

Démonstration. On applique le lemme de Farkas avec $C = [-AA]$, $u = [x^+, x^-]$, $f = -b$. Ce qui donne donc l'exclusion entre existence d'un x, t.q. $Ax \leq b$ et l'existence d'un λ avec $C^\mathsf{T}\lambda \leq 0$, i.e., $A^\mathsf{T}\lambda = 0$, $f^\mathsf{T}\lambda > 0$, i.e., $b^\mathsf{T}\lambda < 0$ et enfin $\lambda \geq 0$. □

Une autre conséquence est encore

Corollaire 5.3.2. *Soit $A = \{w \in \mathbb{R}^n,\ Bw \leq 0\}$ où $B \in \mathbb{R}^{m \times n}$, et soit $y \in \mathbb{R}^n$. Alors*

$$\forall d \in A,\ \langle y, d\rangle \geq 0 \quad \Leftrightarrow \quad \exists \lambda \in (\mathbb{R}^+)^m,\ y = -B^\mathsf{T}\lambda.$$

Démonstration. S'il existe un $d \in A$, c.à.d., tel que $Bd \leq 0$ avec $\langle y, d\rangle < 0$, on peut la décomposer en partie positif et négatif pour arriver à l'existence d'un $\mu = (d^+, d^-) \geq 0$ avec $C\mu \leq 0$, $C = [B - B]$, $\langle f, \mu\rangle < 0$ avec $f = (y, -y)$. Cela veut donc dire qu'il ne peut exister une solution λ à $C^\mathsf{T}\lambda \leq f$, i.e., $B^\mathsf{T}\lambda = y$. Etant donné que λ est sans signe, il ne peut pas non plus exister de solution à $y = -B^\mathsf{T}\lambda$.

L'argument peut se lire dans les deux directions – réfléchissez-y – afin de terminer la demonstration. □

Enfin, nous pouvons utiliser le lemme de Farkas pour dériver le résultat suivant :

Corollaire 5.3.3. *Soit $\bar{u} \in K$ donné. Alors nous avons*

$$\mathrm{N}_K(\bar{u}) = C^{\mathsf{T}}\, \mathrm{N}_{(-\infty,0]^p}(C\bar{u} - f)$$

Démonstration. Il nous suffit d'invoquer la Proposition 5.1.1 après avoir vérifié (5.2). En ce qui concerne ce dernier, supposons l'existence d'un $\lambda \in \mathrm{N}_{(-\infty,0]^p}(C\bar{u} - f)$, c.à.d. donc un $\lambda \geq 0$ tel que $C^{\mathsf{T}}\lambda = 0$, mais $\lambda \neq 0$. Cela veut dire qu'il existe un $j \in I(\bar{u})$ (l'ensemble des indices actives), tel que $\lambda_j > 0$. Pour les indices i non-actives $\lambda_i = 0$. Nous allons donc nous restreindre aux lignes actives, la sous matrice $\bar{C}$ obtenu à partir de C en ne gardant que les lignes actives.

Pour $b = -e_j$, nous obtenons alors un $\lambda \geq 0$ avec $\bar{C}^{\mathsf{T}}\lambda = 0$ et $b^{\mathsf{T}}\lambda < 0$. Nous concluons par le Lemme de Farkas, à la non-existence d'une solution à $\bar{C}u \leq b$. Supposons maintenant l'existence d'un $u \in K$ tel que $C_j u < f_j$, C_j faisant référence à la $j$ème ligne. Puis posons $b = e_j(-\frac{1}{2}f_j + \frac{1}{2}C_j u) < 0$. Il en suit pour notre λ mise de côté (rappelons $\lambda > 0$) que $b^{\mathsf{T}}\lambda < 0$. Par le lemme de Farkas, il suit qu'il n'existe aucune solution à $\bar{C}v \leq b$ (dans K). Toutefois, tout $v \in K$ satisfait $\bar{C}(v - \bar{u}) \leq 0$, car $\bar{C}\bar{u} = f$. En insérant u, et en regardant la $j$ème ligne

$$C_j(u - \bar{u}) - b_j = C_j u - f_j - b_j = C_j u - f_j + \frac{1}{2}f_j - \frac{1}{2}C_j u = \frac{1}{2}(C_j u - f_j) < 0,$$

on réalise que $\bar{C}u \leq b$ ce qui est donc une contradiction. Donc :

- Il n'y a aucune contrainte d'inégalité réelle, c.à.d., la matrice contient en fait également une ligne $-C_j u \leq -f_j$. Mais dans ce cas nous pouvons invoquer la Proposition 5.2.1 et nous avons terminé.
- Ou notre hypothèse que (5.2) n'était pas satisfait a conduit à une contradiction, et dans ce cas nous avons également terminé.

□

Notre préparation nous permet enfin d'aboutir à une formulation explicite des conditions d'optimalité

Théorème 5.3.1. *Considérez l'ensemble fermé K défini dans* (5.8) *et supposez que J est une fonction différentiable sur $\mathbb{V}$ dans $\mathbb{R}$. Si $u \in K$ est un minimum local de J sur K alors*

$$\exists \lambda \in \mathbb{R}^p, \quad \nabla J(u) + C^{\mathsf{T}}\lambda = 0, \tag{5.10a}$$

$$\lambda \geq 0, \quad \lambda_i[Cu - f]_i = 0, \tag{5.10b}$$

$$Cu \leq f. \tag{5.10c}$$

De plus, si J est convexe alors (5.10) *est aussi une condition suffisante de minimalité. c.à.d. si u vérifie* (5.10)*, alors u est un minimum global de J sur K.*

Démonstration. Le résultat suit immédiatement du Corollaire 5.3.3 en combinaison avec le Corollaire 3.3.2. Sous convexité, le côté suffisante vient du Théorème 3.3.4. □

preuve directe. Supposons que $u \in K$ est un minimum de J sur K.

— Si l'ensemble des indices actives $I(u)$ est vide, alors u est dans l'intérieur de K et $\nabla J(u) = 0$, on peut alors prendre $\lambda = 0$.

— L'ensemble des indices actives n'est pas vide. Soit alors $\bar{C}$ la matrice crée à partir de C en ne gardant que les lignes actives. Soit alors $K^{\#} := \{w \in \mathbb{V}, \bar{C}w \leq 0\}$. Nous allons prouver d'abord que $\langle \nabla J(u), w\rangle \geq 0$, pour tout $w \in K^{\#}$.
Pour un $w \in K^{\#}$ fixé mais arbitraire et un $\varepsilon > 0$ arbitraire, évidemment $C_j(u + \varepsilon w) = f_j + \varepsilon C_j w \leq f_j$, pour tout $j \in I(u)$. En ce qui concerne $j \notin I(u)$, nous pouvons évidemment trouver un $\bar{\varepsilon} > 0$ assez petit pour que $C_j(u + \varepsilon w) < f_j$ pour $\varepsilon < \bar{\varepsilon}$ par continuité. En conclusion nous pouvons supposer $\bar{\varepsilon} > 0$ assez petit pour que $u + \varepsilon w \in K$ pour tout $\varepsilon < \bar{\varepsilon}$. L'inéquation d'Euler donne alors

$$\langle \nabla J(u), (u + \varepsilon w) - u\rangle \geq 0,$$

i.e., $\varepsilon \langle \nabla J(u), w\rangle \geq 0$, i.e., $\langle \nabla J(u), w\rangle \geq 0$. Ceci permet de conclure car $w \in K^{\#}$ était arbitraire. Nous pouvons maintenant appliquer le lemme de Farkas : Corollaire 5.3.2 pour dériver l'existence d'un $\bar{\lambda} \in \mathbb{R}^{\bar{p}}$, $\bar{\lambda} \geq 0$, tel que

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

Nous pouvons étendre ce $\bar{\lambda}$ vers un $\lambda \in \mathbb{R}^p$ en y ajoutant des zéros pour des composantes non-actives.

□

5.4 Contraintes d'égalité et d'inégalité affines

Dans cette section, on suppose que l'ensemble K est défini par, à la fois, des contraintes d'égalité et d'inégalité affines :

$$K = \{v \in \mathbb{R}^n \mid C_I v \leq f_I \text{ et } C_E v = f_E\}, \tag{5.11}$$

où C_I et C_E sont des matrices respectivement de tailles $p \times n$ et $m \times n$, et $f_I \in \mathbb{R}^p$, $f_E \in \mathbb{R}^m$.

Remarquons que la contrainte $C_E v = f_E$ est équivalente à

$$C_E v \leq f_E \quad \text{et} \quad -C_E v \leq -f_E,$$

et l'ensemble K peut donc être redéfini par :

$$K = \{v \in \mathbb{R}^n,\ Cv \leq f\}, \quad \text{avec } C = \begin{bmatrix} C_E \\ -C_E \\ C_I \end{bmatrix} \quad \text{et} \quad f = \begin{bmatrix} f_E \\ -f_E \\ f_I \end{bmatrix}. \tag{5.12}$$

On peut obtenir immédiatement :

Lemme 5.4.1. *$\bar{u} \in K$ donné. Alors nous avons*

$$\mathrm{N}_K(\bar{u}) = C_I^{\mathsf{T}}\, \mathrm{N}_{(-\infty,0]^p}(C_I\bar{u} - f_I) + \mathrm{Im}\, C_E^{\mathsf{T}}.$$

Démonstration. La caractérisation précédente (5.12) de K et le Corollaire 5.3.3 permettent d'établir ce résultat. Les détails sont laissés au lecteur. □

Théorème 5.4.1. *Considérez l'ensemble fermé K défini dans* (5.11) *et supposons que J est une fonction différentiable sur $\mathbb{V}$ dans $\mathbb{R}$. Si $u \in K$ est un minimum local de J sur K alors*

$$\exists \lambda \in \mathbb{R}^p, \exists \mu \in \mathbb{R}^m, \quad \nabla J(u) + C_I^{\mathsf{T}}\lambda + C_E^{\mathsf{T}}\mu = 0, \tag{5.13a}$$

$$\lambda \geq 0, \quad \lambda_i[C_I u - f_I]_i = 0, \tag{5.13b}$$

$$C_I u \leq f_I, \quad C_E u = f_E. \tag{5.13c}$$

De plus, si J est convexe alors (5.13) *est aussi une condition nécessaire et suffisante de minimalité.*

Démonstration. Le résultat suit immédiatement du Lemme 5.4.1 en combinaison avec le Corollaire 3.3.2. Sous convexité, le côté suffisante vient du Théorème 3.3.4. □

Enfin nous pouvons également obtenir la caractérisation suivante (en écrivant μ comme $\mu^+ - \mu^-$) à partir du Théorème 5.4.1, où encore en combinant la représentation (5.12) de K avec le Théorème 5.3.1. Il en résulte que :

$$\begin{aligned} \exists \begin{pmatrix} \lambda \\ \mu_1 \\ \mu_2 \end{pmatrix} \in \mathbb{R}^p \times \mathbb{R}^m \times \mathbb{R}^m, \qquad & \text{tel que :} \\ \nabla J(u) + C_I^{\mathsf{T}}\lambda + C_E^{\mathsf{T}}(\mu_1 - \mu_2) = 0, & \\ \lambda \geq 0, \quad \mu_1 \geq 0, \quad \mu_2 \geq 0, & \\ C_E u = f_E, \quad C_I u \leq f_I, \quad (\lambda, C_I u - f_I) = 0. & \end{aligned}$$

5.5 Cas général : conditions de Karush-Kuhn-Tucker*

Poursuivons maintenant notre investigation vers le cadre général.

Théorème 5.5.1. *Soit $g : \mathbb{R}^n \to \mathbb{R}^p$ une fonction continûment différentiable (et convexe) et soit $J : \mathbb{R}^n \to \mathbb{R}$ également (Fréchet) différentiable. Posons*

$$K = \{u \in \mathbb{R}^n \ : g(u) \leq 0\}$$

et considérons le problème de minimisation suivant :

$$\min_{u \in K} J(u).$$

Alors si $\bar{u} \in K$ est une solution locale du problème et (5.2) est vérifié, alors il existe un $\lambda \in \mathbb{R}^p$, tel que :

$$0 = \nabla J(\bar{u}) + \sum_{j=1}^{p} \lambda_j \nabla g_j(\bar{u}) \tag{5.14a}$$

$$0 \leq \lambda \tag{5.14b}$$

$$0 \geq g(\bar{u}) \tag{5.14c}$$

$$0 = \lambda^{\mathsf{T}} g(\bar{u}). \tag{5.14d}$$

Démonstration. Cette fois-ci on combine le Corollaire 3.3.2 avec la Proposition 5.1.1. □

Définition 5.5.1. *Les conditions (5.14) sont appelés les conditions de Karush-Kuhn-Tucker.*

Remarque 5.5.1 (Qualification de contraintes*)**.** *La condition (5.2) est induite par des conditions classiques tel que LICQ (Linear Independence Constraint Qualification) et MFCQ (Mangasarian-Fromovitz Constraint Qualification).*
LICQ stipule qu'en $\bar{u}$, les gradients $\{\nabla g_j(\bar{u})\}_{j \in I(\bar{u})}$ sont linéairément independents.
MFCQ stipule l'existence d'un $d \in \mathbb{R}^n$ tel que $\nabla g_j(\bar{u})^{\mathsf{T}} d < 0$ pour tout $j \in I(\bar{u})$.

5.5.1 Le Lagrangien - in a nutshell

Considérons l'idée de se débarrasser tout simplement des contraintes en les pénalisant. Donnons nous alors un paramètre de pénalisation $\mu \geq 0$, $\mu \in \mathbb{R}^p$ et considérons le Lagrangien :

$$L(u, \mu) = J(u) + \mu^{\mathsf{T}} g(u). \tag{5.15}$$

Il est relativement clair que nous avons l'identité suivante :

$$\inf_{u\in\mathbb{R}^n}\sup_{\mu\geq 0} L(u,\mu) = \inf_{u\in K} J(u), \tag{5.16}$$

mais cette identité n'est pas très utile (à quoi bon de devoir résoudre deux problèmes au lieu d'un ?).

En revanche, si l'on se permette d'inverser sup et inf, nous obtenons une structure intéressante. Définissons le "dual Lagrangien" : $\theta : \mathbb{R}^p \to \bar{\mathbb{R}}$:

$$\theta(\mu) = \inf_{u\in\mathbb{R}^n} L(u,\mu). \tag{5.17}$$

Alors, il s'avère que θ est concave en μ ! Nous avons donc récupéré de la convexité, où auparavant il n'y en avait pas nécessairement. Sans surprise, il n'y a pas de "free-lunch" : nous avons payé un coût. Par ailleurs, nous avons toujours

$$\sup_{\mu\in\mathbb{R}^p} \theta(\mu) \leq \inf_{u\in K} J(u), \tag{5.18}$$

et parfois égalité. Un cas classique d'égalité est lorsque l'on considère la programmation linéaire (J affine, g affine !). En régle général, même dans le cas convexe, il peut y avoir un écart. Cet écart est appelé "saut de dualité".

Définition 5.5.2. *On dit que (u,λ) est un point-selle de L sur $\mathbb{R}^n\times\mathbb{R}^p_+$ si et seulement si :*

$$\forall \mu\in\mathbb{R}^p_+ \quad L(u,\mu)\leq L(u,\lambda)\leq L(v,\lambda)\ \forall v\in\mathbb{R}^n \tag{5.19}$$

ainsi que $(u,\lambda)\in\mathbb{R}^n\times\mathbb{R}^p_+$.

Il est intéressant de lier la notion de point selle à celle d'un point KKT.

Proposition 5.5.1. *Soit (u,λ) un point-selle de L sur K alors les conditions* (5.14) *sont satisfaites avec le couple (u,λ).*

Démonstration. Par définition u réalise le minimum sur $\mathbb{R}^n$ de L dans son premier argument. Nous obtenons donc $\nabla_u L(u,\lambda)=0$. Aussi par définition, λ réalise le minimum (maximum) de $-L(u,\mu)$ ($(L(u,\mu))$ sur $\mathbb{R}^p_+$. Cela nous donne

$$0\in -\nabla_\mu L(u,\lambda) + \mathrm{N}_{\mathbb{R}^p_+}(\lambda),$$

où en d'autres mots, il existe un $\nu\leq 0$, $\nu^\mathsf{T}\lambda = 0$ tel que $-g(u)+\nu=0$. Encore en d'autres mots, lorsque $\lambda_i>0$, $g_i(u)=0$ et lorsque $\lambda_i=0$, $g_i(u)\leq 0$. Donc $g(u)\leq 0$, i.e., $u\in K$ et en résumant nous avons également $\lambda^\mathsf{T} g(u)=0$. Les conditions (5.14) sont donc satisfaites. □

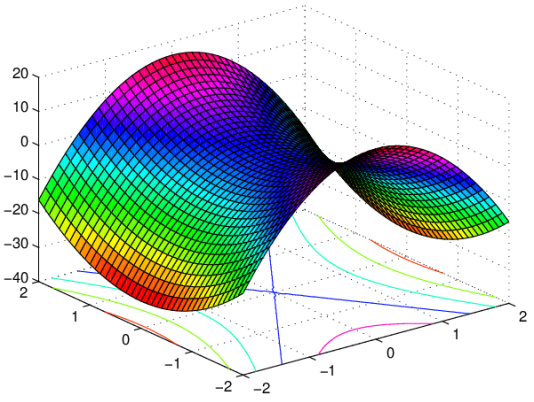


FIGURE 5.1 – Exemple d'un point selle

Chapitre 6

Algorithmes pour problèmes contraints

De façon similaire au cas sans contraintes, les méthodes de résolution des problèmes contraints sont très nombreuses. Nous allons en présenter quelques unes, qui conduisent à des algorithmes numériques simples et utilisables en pratique.

Le problème auquel nous nous intéressons, dans tout ce chapitre, est le suivant :

$$\text{Trouver } u \in K \text{ tel que } J(u) = \min_{v \in K} J(v).$$

Ici, J est la fonctionnelle qui à v associe $J(v) = \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle$, où la matrice A est supposée *symétrique définie-positive* de $\mathbb{R}^{n\times n}$, b un vecteur quelconque de $\mathbb{R}^n$ et K est un fermé convexe de $\mathbb{R}^n$.

6.1 Méthode du gradient projeté.

D'abord insistons sur l'hypothèse que K est un convexe fermé. Du théorème 3.3.4, nous savons que la solution optimale u du problème contraint, vérifie l'inéquation d'Euler suivante :

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

D'autre part, rappelons que pour tout $w \in \mathbb{R}^n$, il existe une unique projection de w sur K, notée $P_K(w) \in K$, solution de (voir Exemple 3.3.1)

$$\|P_K(w) - w\| = \min_{v \in K} \|v - w\|.$$

De plus, pour tout $w \in \mathbb{R}^n$, la projection $P_K(w)$ est **caractérisée** par :

$$\langle P_K(w) - w, P_K(w) - v\rangle \leq 0 \quad \forall v \in K. \tag{6.1}$$

On arrive alors à la proposition suivante :

Proposition 6.1.1. *Soit K un convexe fermé de $\mathbb{R}^n$ et $\rho > 0$. $u \in K$ est solution du problème contraint $J(u) = \min_{v\in K} J(v)$ si et seulement si*

$$u = P_K(u - \rho\nabla J(u)).$$

Démonstration. Du fait que J est convexe, l'inéquation d'Euler constitue une condition nécessaire et suffisante de minimalité. Ainsi, pour $u \in K$ et $\rho > 0$, on a :

$$\begin{aligned} u \text{ est minimum} &\iff \langle \nabla J(u), u - v\rangle \leq 0 \qquad \forall v \in K, \\ &\iff \langle \rho\nabla J(u), u - v\rangle \leq 0 \qquad \forall v \in K, \\ &\iff \left\langle u - \underbrace{\big(u - \rho\nabla J(u)\big)}_{\boldsymbol{w}}, u - v \right\rangle \leq 0 \quad \forall v \in K, \\ &\iff u = P_K\big(u - \rho\nabla J(u)\big). \end{aligned}$$

□

L'algorithme du **gradient projeté** se base sur la propriété ci-dessus et propose de construire une suite $(u_k)_k$ qui converge vers u en tant que point fixe de l'application $v \longmapsto P_K(v - \rho\nabla J(v))$. Plus précisément, l'algorithme s'écrit :

1) **Initialisation** : $u_0 \in \mathbb{R}^n$, un pas $\rho > 0$ et une précision $\eta > 0$.

   On calcule $u_1 = P_K(u_0 - \rho\nabla J(u_0))$.

2) **Tant que** $\|u_k - u_{k-1}\| > \eta$, on définit u_{k+1} par

   $$u_{k+1} = P_K(u_k - \rho\nabla J(u_k)).$$

La convergence de l'algorithme s'appuie sur le fait que une projection P_K est une application contractante :

Lemme 6.1.1. *L'application $w \in \mathbb{R}^n \longmapsto P_K(w)$ est 1-Lipshitz :*

$$\|P_K(w) - P_K(v)\|_2 \leq \|w - v\|_2 \qquad \forall w, v \in \mathbb{R}^n.$$

Remarque 6.1.1. *Noter que dans cet algorithme le test d'arrêt est différent de celui qu'on avait utilisé dans les cas des algorithmes d'optimisation sans contrainte.*

— *On ne peut plus prendre comme test d'arrêt $\|\nabla J(u_k)\| \sim 0$ car pour le problème avec contraintes le minimum u ne satisfait pas forcément "$\nabla J(u) = 0$" !*

— *Le test d'arrêt $\|u_k - u_{k-1}\| < \eta$ est le mieux adapté à l'algorithme de gradient projeté, étant donné qu'on cherche un point fixe. En effet, le test signifie aussi que $\|u_{k-1} - P_K(u_{k-1} - \rho\nabla J(u_{k-1}))\| < \eta$, en d'autre terme u_k est un point fixe à une précision η-près.*

Théorème 6.1.1. *Soit $K \neq \emptyset$ un convexe fermé de $\mathbb{R}^n$, et Soit*

$$J : v \longmapsto \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle ,$$

où $A \in \mathbb{R}^{n\times n}$ est symétrique, définie positive, et $b \in \mathbb{R}^n$. Si $0 < \rho < \frac{2}{\lambda_{\max}(A)}$, alors quel que soit $u_0 \in \mathbb{R}^n$, la suite (u_k) définie par le gradient projeté converge vers le minimum u. ($\lambda_{\max}(A)$ désigne la plus grande valeur propre de A.)

Démonstration. Notons d'abord que $\nabla J(v) = Av - b \ \ \forall v \in K$, et

$$\big\|u_{k+1} - u\big\|_2 \ = \ \Big\|P_K\big(u_k - \rho(Au_k - b)\big) - P_K\big(u - \rho(Au - b)\big)\Big\|_2.$$

A l'aide du lemme précédent, on obtient :

$$\big\|u_{k+1} - u\big\|_2 \leq \Big\|\Big(I - \rho A\Big)(u_k - u)\Big\|_2 \leq \gamma_\rho\big\|u_k - u\big\|_2,$$

où $\gamma_\rho := |\lambda_{\max}(I - \rho A)|$. Pour tout $0 < \rho < \frac{2}{\lambda_{\max}(A)}$, on a $\gamma_\rho < 1$, et donc :

$$\big\|u_k - u\big\|_2 \leq (\gamma_\rho)^k \big\|u_0 - u\big\|_2 \xrightarrow[k\to+\infty]{} 0. \qquad \square$$

□

Remarque 6.1.2. *Le théorème précédent peut être généralisé à une classe plus large de problèmes d'optimisation convexe avec contraintes. Nous renvoyons à [8] pour une discussion plus approfondie.*

D'un point de vue pratique, la projection d'un élément $v \in \mathbb{R}^n$ sur un convexe quelconque peut être très difficile à déterminer. Il existe cependant un cas où cette projection est aisée :

Lemme 6.1.2. *Si $K = \prod_{i=1}^n [a_i, b_i]$. Soit $y = (y_1, \ldots, y_n)^\mathsf{T} \in \mathbb{R}^n$, la projection $x = P_K(y)$ a pour composantes :*

$$x_i = \min\Big(\max(a_i, y_i), b_i\Big) \quad \textit{pour } 1 \leq i \leq n.$$

Démonstration. Laissée à titre d'exercice. □

6.2 Méthode d'Uzawa

Supposons que l'ensemble K est donnée par l'une des formes suivantes

$$K := \{v \in \mathbb{R}^n; Cv = f\}, \text{ ou} \tag{6.2a}$$
$$K := \{v \in \mathbb{R}^n; Cv \leq f\}, \tag{6.2b}$$

où la matrice C est $p \times n$ et $f \in \mathbb{R}^p$.

Soit $u \in K$ le minimum de la fonctionnelle J sur K. Une réecriture générale des théorèmes 5.2.1-5.10 donne : il existe $\lambda \in \mathbb{R}^p$ tel que :

$$\nabla J(u) + C^{\mathsf{T}}\lambda = 0, \tag{6.3a}$$
$$\lambda \in \mathbb{F}, \quad Cu \leq f, \quad (\lambda, Cu - f) = 0, \tag{6.3b}$$
$$\langle \lambda - \mu, Cu - f \rangle \geq 0 \quad \forall \mu \in \mathbb{F}, \tag{6.3c}$$

avec $\mathbb{F} = \mathbb{R}^p$ dans le cas où K est donné par (6.2a), et $\mathbb{F} = (\mathbb{R}^+)^p$ dans le cas (6.2b).

Proposition 6.2.1. *Soit $\lambda \in \mathbb{R}^p$ tel que (u, λ) vérifie (6.3). Pour tout $\rho > 0$, on :*

$$\lambda = P_{\mathbb{F}}(\lambda + \rho(Cu - f)), \tag{6.4}$$

$P_{\mathbb{F}}$ étant la projection de $\mathbb{R}^p$ sur le convexe fermé $\mathbb{F}$ (et rappelons que $\mathbb{F} = \mathbb{R}^p$ dans le cas de contraintes d'égalité, et $\mathbb{F} := (\mathbb{R}^+)^p$ dans le cas de contraintes d'inégalité).

Démonstration. Comme le multiplicateur λ vérifie (6.3c), alors on a aussi

$$\langle \lambda - (\lambda + \rho(Cu - f)), \lambda - \mu \rangle \leq 0 \quad \forall \mu \in \mathbb{F}, \ \forall \rho > 0.$$

Ce qui d'après (6.1) signifie que $\lambda = P_{\mathbb{F}}(\lambda + \rho(Cu - f))$, pour tout $\rho > 0$. □

Tenant compte de la proposition précédente, nous proposons l'algorithme suivant, appelé **Algorithme d'Uzawa**

1. On choisit : une condition initiale $\lambda_0 \in \mathbb{R}^p$, un pas $\rho > 0$ et une précision $\eta > 0$. On calcule u_0 solution de $Au_0 = b - C^{\mathsf{T}}\lambda_0$.
2. **Tant que** $\|\lambda_k - \lambda_{k-1}\| > \eta$ ou $\|u_k - u_{k-1}\| > \eta$, on définit (u_{k+1}, λ_{k+1}) par
$$\left\| \begin{array}{l} \lambda_{k+1} = P_{\mathbb{F}}\Big(\lambda_k + \rho(Cu_k - f)\Big); \\ u_{k+1} \text{ est solution de } Au_{k+1} = b - C^{\mathsf{T}}\lambda_{k+1}. \end{array} \right.$$

Théorème 6.2.1. *Supposons que A est symétrique définie positive. Si $0 < \rho < \frac{2\lambda_{\min}(A)}{\|C\|^2}$, alors quel que soit l'élément initial $\lambda_0 \in \mathbb{R}^p$, le suite (u_k) définie par l'algorithme d'Uzawa converge vers le minimum u.*

Démonstration. Grâce à des calculs directs et au lemme 6.1.1, on obtient :

$$\begin{aligned}
\|\lambda_{k+1}-\lambda\|_2^2 &= \|P_{\mathbb{F}}\Big(\lambda_k+\rho(Cu_k-f)\Big)-P_{\mathbb{F}}\Big(\lambda+\rho(Cu-f)\Big)\|_2^2\\
&\leq \|\lambda_k-\lambda+\rho C(u_k-u)\|_2^2\\
&\leq \|\lambda_k-\lambda\|_2^2+\rho^2\|C\|^2\|u_k-u\|_2^2+2\rho\big(C^{\mathsf{T}}(\lambda_k-\lambda),u_k-u\big)\\
&\leq \|\lambda_k-\lambda\|_2^2+\rho^2\|C\|^2\|u_k-u\|_2^2-2\rho\big(A(u_k-u),u_k-u\big)\\
&\leq \|\lambda_k-\lambda\|_2^2+\Big(\rho^2\|C\|^2-2\rho\lambda_{\min}(A)\Big)\|u_k-u\|_2^2.
\end{aligned}$$

Si $0<\rho<\frac{2\lambda_{\min}(A)}{\|C\|^2}$, en posant : $\beta=2\rho\lambda_{\min}(A)-\rho^2\|C\|^2$, il vient que $\beta>0$ et

$$\|u_k-u\|_2^2\leq\frac{1}{\beta}\Big(\|\lambda_k-\lambda\|_2^2-\|\lambda_{k+1}-\lambda\|_2^2\Big).$$

Ce qui prouve que la suite $\Big(\|\lambda_k-\lambda\|_2^2\Big)_{k\geq 0}$ est décroissante et minorée (évidement par 0). Par conséquent, $\|u_k-u\|_2^2\xrightarrow[k\to+\infty]{}0$. □

6.3 Techniques de pénalisation

Nous allons revenir à l'idée d'éliminer la contrainte $v\in K$ en la "pénalisant". Pour cela, nous introduisons un paramètre $\varepsilon>0$, la fonctionnelle

$$J_\varepsilon(v)=J(v)+\frac{1}{\varepsilon}\|Cv-f\|^2,$$

ainsi que le **problème pénalisé**

$$\text{Trouver } u_\varepsilon\in\mathbb{R}^n \text{ tel que } \tilde{J}(u_\varepsilon)=\min_{v\in\mathbb{R}^n}\tilde{J}_\varepsilon(v). \tag{6.5}$$

Dans la suite, on appelle ψ la fonctionnelle qui à v associe $\|Cv-f\|^2$. On remarque que ψ est à valeurs dans $\mathbb{R}^+$, convexe, continue et telle que, pour tout élément v de K, $\psi(v)=0$. En particulier, pour $\bar{u}$ la solution optimale du problème, $\psi(\bar{u})=0$, ce qui signifie que, pour tout $\varepsilon>0$, $J_\varepsilon(\bar{u})=J(\bar{u})$.

Proposition 6.3.1. *Le problème (6.5) admet une solution unique, pour tout $\varepsilon>0$.*

Démonstration. — **Existence d'une solution.** J_ε est continue. Montrons qu'elle est de plus infinie à l'infini. On écrit

$$\begin{aligned}
J_\varepsilon(v) &= J(v)+\frac{1}{\varepsilon}\psi(v)\geq J(v)=\frac{1}{2}(Av,v)-(b,v)\\
&\geq \frac{\lambda_{min}}{2}\|v\|^2-\|b\|\,\|v\|,\\
&\quad \text{quantité qui tend vers l'infini lorsque } \|v\|\to+\infty.
\end{aligned}$$

— **Unicité de la solution.** J étant strictement convexe (cf. exercice 5.2.1), et $\frac{1}{\varepsilon}\psi$ étant convexe, leur somme J_ε est strictement convexe. En conséquence, le point de minimum est unique, d'après le théorème 2.3.1.

□

Nous allons maintenant prouver que la suite $(u_\varepsilon)_\varepsilon$ possède une propriété très intéressante...

Proposition 6.3.2. *La suite $(u_\varepsilon)_\varepsilon$ converge vers $\bar{u}$, solution du problème*

$$\min_{u\in\mathbb{R}^n} J(u) \text{ s.t. } Cu = f,$$

lorsque ε tend vers 0^+.

Démonstration. — **Etape 1.** Par définition de ψ, $J(u_\varepsilon) \leq J(u_\varepsilon)+\frac{1}{\varepsilon}\psi(u_\varepsilon) = J_\varepsilon(u_\varepsilon)$; or, u_ε réalise le minimum de J_ε sur $\mathbb{R}^n$, donc $J_\varepsilon(u_\varepsilon) \leq J_\varepsilon(\bar{u})$. Enfin, d'après ce que l'on a remarqué plus haut, $J_\varepsilon(\bar{u}) = J(\bar{u})$. Ainsi

$$\forall \varepsilon > 0, \quad J(u_\varepsilon) \leq J(\bar{u}). \tag{6.6}$$

La fonctionnelle J étant infinie à l'infini, nous en déduisons que $(u_\varepsilon)_\varepsilon$ est bornée.

— **Etape 2.** Comme nous nous trouvons dans $\mathbb{R}^n$, il existe une sous-suite extraite $(u_{\varepsilon'})_{\varepsilon'}$ qui converge. Appelons u' sa limite. D'après la continuité de J et la relation (6.6), qui s'applique notamment pour tous les termes de la sous-suite :

$$J(u') = \lim_{\varepsilon'\to 0+} J(u_{\varepsilon'}) \leq J(\bar{u}).$$

Ceci démontrait l'optimalité de u' si toutefois $u' \in K$. C'est donc ce point qui fait l'objet de notre investigation suivante.
Par ailleurs, nous pouvons estimer :

$$0 \leq \psi(u_{\varepsilon'}) = \varepsilon'\{J_{\varepsilon'}(u_{\varepsilon'}) - J(u_{\varepsilon'})\} \leq \varepsilon'\{J_{\varepsilon'}(\bar{u}) - J(u_{\varepsilon'})\} = \varepsilon'\{J(\bar{u}) - J(u_{\varepsilon'})\}.$$

On vient de voir que $(J(u_{\varepsilon'}))_{\varepsilon'}$ admet une limite (égale à $J(u')$), ce qui entraîne que

$$\lim_{\varepsilon'\to 0+} (\varepsilon'\{J(\bar{u}) - J(u_{\varepsilon'})\}) = 0, \text{ et donc } \lim_{\varepsilon'\to 0+} \psi(u_{\varepsilon'}) = 0.$$

Comme ψ est continue : $\psi(u') = 0$, i.e., $u' \in K$. Bien sûr, $\bar{u}$ réalisant le minimum de J sur K, nous donne $J(\bar{u}) \leq J(u')$. On en arrive finalement à l'égalité $J(\bar{u}) = J(u')$, et comme J est strictement convexe, $\bar{u} = u'$.

— **Etape 3.** Pour finir, supposons que $(u_\varepsilon)_\varepsilon$ ne converge pas vers u. Ceci signifie qu'il existe une sous-suite extraite, toujours notée $(u_{\varepsilon'})_{\varepsilon'}$, et $\eta > 0$ tels que $\|u_{\varepsilon'} - \bar{u}\| \geq \eta$, pour tout ε'.
On reprend le raisonnement de l'étape 2 : $(u_{\varepsilon'})_{\varepsilon'}$ étant bornée, on peut en extraire une sous-suite, $(u_{\varepsilon''})_{\varepsilon''}$, qui converge. En poursuivant le même raisonnement (n'oublions pas que, par construction, $(u_{\varepsilon''})_{\varepsilon''}$ est également une sous-suite extraite de $(u_\varepsilon)_\varepsilon$!), on prouve que $(u_{\varepsilon''})_{\varepsilon''}$ converge nécessairement vers $\bar{u}$. Ceci contredit le fait que $\|u_{\varepsilon''} - \bar{u}\| \geq \eta$, pour tout ε''.
En conclusion, toute la suite $(u_\varepsilon)_\varepsilon$ converge vers $\bar{u}$.

□

Pour cette méthode, le problème central est celui du choix d'une *suite de valeurs de* ε, qui permette d'obtenir rapidement une bonne approximation de $\bar{u}$. Par rapidement, on entend sans avoir à résoudre "beaucoup" de problèmes sans contraintes du type (6.5). Notons que J_ε peut être développée sous la forme :

$$\begin{aligned} J_\varepsilon(v) &= \frac{1}{2}\langle Av, v\rangle - \langle b, v\rangle + \frac{1}{\varepsilon}\{\langle C^\mathsf{T}Cv, v\rangle - 2\langle C^\mathsf{T}f, v\rangle + \|f\|^2\} \\ &= \frac{1}{2}\left\langle [A + \frac{2}{\varepsilon}C^\mathsf{T}C]v, v\right\rangle - \left\langle b + \frac{2}{\varepsilon}C^\mathsf{T}f, v\right\rangle + \frac{1}{\varepsilon}\|f\|^2. \end{aligned}$$

La matrice $A + \frac{2}{\varepsilon}C^\mathsf{T}C$ est symétrique définie-positive. Cependant, sa structure interne peut être très différente de celle de A.

6.4 Optimisation par résolution directe des conditions du premier ordre

On se place maintenant dans le cas où $K := \{v \in \mathbb{R}^n; Cv = f\}$, avec C une matrice de $\mathbb{R}^{p\times n}$ de rang p. Le système de minimalité revient, cf. Corollaire 5.2.1, à résoudre le système linéaire, de solution (u, λ) appartenant à $\mathbb{R}^n \times \mathbb{R}^p$:

$$\left\{ \begin{array}{l} Au + C^\mathsf{T}\lambda = b \\ Cu = f \end{array} \right. . \tag{6.7}$$

Nous allons étudier deux méthodes purement algébriques permettant de résoudre (6.7). Dans le système (6.7), l'inconnue qui nous intéresse est u, la valeur λ est "en général" moins utilisé.

6.4.1 Elimination des contraintes - première approche

La première approche consiste à écrire $\lambda : u = A^{-1}(b - C^\mathsf{T}\lambda)$, ce qui implique la relation $CA^{-1}b - CA^{-1}C^\mathsf{T}\lambda = f$. En d'autres termes, λ est la solution de

$$\text{Trouver } \lambda \in \mathbb{R}^p \text{ tel que } CA^{-1}C^\mathsf{T}\lambda = CA^{-1}b - f. \tag{6.8}$$

Si p est très petit devant n, la difficulté est la formation de la matrice $CA^{-1}C^\mathsf{T}$ de $\mathbb{R}^{p\times p}$, et du second membre $CA^{-1}b$ appartenant à $\mathbb{R}^p$. En effet, une fois ceux-ci connus, il est raisonnable de supposer que la résolution de (6.8) sera aisée. Qui plus est, $CA^{-1}C^\mathsf{T}$ est symétrique définie-positive. A partir de là, u est la solution de

$$\text{Trouver } u \in \mathbb{R}^n \text{ tel que } Au = b - C^\mathsf{T}\lambda, \tag{6.9}$$

et l'on en revient aux méthodes de la section précédente. Pour ce qui est de la formation de $CA^{-1}C^\mathsf{T}$, notons que l'on peut écrire

$$CA^{-1}C^\mathsf{T} = CC', \text{ avec } C' = A^{-1}C^\mathsf{T} \in \mathbb{R}^{n\times p}.$$

C' est caractérisée par

$$\text{Trouver } C' \in \mathbb{R}^{n\times p} \text{ telle que } AC' = C^\mathsf{T}. \tag{6.10}$$

Ce système linéaire peut être reformulé *colonne par colonne*. En effet, si on note $(c'_i)_{1\le i\le p}$ les colonnes de C' et $(\underline{c}_i)_{1\le i\le p}$ celles de C^T, (6.10) est équivalent à

$$\text{Pour } i = 1, \cdots, p, \text{ trouver } c'_i \in \mathbb{R}^n \text{ tel que } Ac'_i = \underline{c}_i. \tag{6.11}$$

L'obtention de $CA^{-1}C^\mathsf{T}$ est alors immédiate, par simple multiplication. Pour ce qui est du calcul de $A^{-1}b$, on procède de façon similaire, en résolvant cette fois

$$\text{Trouver } c_{p+1} \in \mathbb{R}^n \text{ tel que } Ac_{p+1} = b, \tag{6.12}$$

puis en construisant $CA^{-1}b$, résultat de la multiplication de C par c_{p+1}.

De cette façon, on a démontré la

Proposition 6.4.1. *On peut ramener le calcul de (u, λ), solution de (6.7), à la résolution de $p + 2$ problèmes de minimiation sans contraintes, de type (4.1).*

Démonstration. Il suffit de résoudre (6.11)-(6.12), soit $p+1$ problèmes, puis (6.8), dont le coût est supposé "faible", et enfin (6.9). □

Cette méthode présente l'avantage d'être complètement compatible avec les algorithmes proposés à la section 4, puisque l'on a uniquement des problèmes sans contraintes à résoudre. En outre, elle est particulièrement indiquée si p est petit... Si p est grand, la même technique n'en reste pas moins valable, sachant que l'étape (6.8) peut devenir prépondérante, et qu'il faut la traiter avec attention.

6.4.2 Elimination des contraintes - deuxième approche

Rappelons le contexte sous un angle un peu différent, c'est-à-dire sans multiplicateur de Lagrange. Le but est de minimiser J sur K, qui est défini par $\{v \in \mathbb{R}^n \ : \ Cv = f\}$. C est une matrice de $\mathbb{R}^{p\times n}$ de rang p. On suppose que l'on peut l'écrire par blocs sous la forme

$$C = \left(\begin{array}{cc} C_{11} & C_{12} \end{array} \right), \ C_{11} \in \mathbb{R}^{p\times p}, \ rg(C_{11}) = p.$$

(Eventuellement après un réarrangement des colonnes.)

$$Cv = f \Longleftrightarrow C_{11}v_1 + C_{12}v_2 = f, \text{avec } v = \left(\begin{array}{c} v_1 \\ v_2 \end{array} \right), \ v_1 \in \mathbb{R}^p, \ v_2 \in \mathbb{R}^{n-p}.$$

D'où

$$v_1 = {C_{11}}^{-1}(f - C_{12}v_2) = g - \underline{C}v_2, \text{ avec } g = {C_{11}}^{-1}f, \ \underline{C} = {C_{11}}^{-1}C_{12}.$$

Nous allons maintenant reécrire $J(v)$ sous la forme $\tilde{J}(v_2)$, *pour tout* $v \in K$.
Le terme *linéaire* :

$$\begin{aligned} -\langle b, v\rangle &= -\langle b_1, v_1\rangle_1 - \langle b_2, v_2\rangle_2 = -\langle b_1, g\rangle_1 + \langle b_1, \underline{C}v_2\rangle_1 - \langle b_2, v_2\rangle_2 . \\ &= \alpha_{lin} + \left\langle \underline{C}^{\mathsf{T}} b_1 - b_2, v_2 \right\rangle_2 , \end{aligned} \tag{6.13}$$

où $\alpha_{lin} = -\langle b_1, g\rangle_1$ est une constante.
Le terme *quadratique* : (les deux blocs diagonaux, A_{11} et A_{22}, sont nécessairement symétriques)

$$Av = \left(\begin{array}{cc} A_{11} & A_{12} \\ A_{12}^{\mathsf{T}} & A_{22} \end{array} \right) \left(\begin{array}{c} v_1 \\ v_2 \end{array} \right) = \left(\begin{array}{c} A_{11}v_1 + A_{12}v_2 \\ A_{12}^{\mathsf{T}}v_1 + A_{22}v_2 \end{array} \right).$$

On en déduit que :

$$\begin{aligned} \frac{1}{2}\langle Av, v\rangle &= \frac{1}{2}\langle A_{11}v_1 + A_{12}v_2, v_1\rangle_1 + \frac{1}{2}\left\langle A_{12}^{\mathsf{T}}v_1 + A_{22}v_2, v_2\right\rangle_2 \\ &= \frac{1}{2}\langle A_{11}v_1, v_1\rangle_1 + \left\langle A_{12}^{\mathsf{T}}v_1, v_2\right\rangle_2 + \frac{1}{2}\langle A_{22}v_2, v_2\rangle_2 . \end{aligned}$$

Examinons le premier terme :

$$\begin{aligned} \frac{1}{2}\langle A_{11}v_1, v_1\rangle_1 &= \frac{1}{2}\langle A_{11}g - A_{11}\underline{C}v_2, g - \underline{C}v_2\rangle_1 \\ &= \alpha_{quad} - \left\langle \underline{C}^{\mathsf{T}}A_{11}g, v_2\right\rangle_2 + \frac{1}{2}\left\langle \underline{C}^{\mathsf{T}}A_{11}\underline{C}v_2, v_2\right\rangle_2 , \ \alpha_{quad} = \frac{1}{2}\langle A_{11}g, g\rangle_1 . \end{aligned}$$

Le second terme :

$$\left\langle A_{12}^{\mathsf{T}}v_1, v_2\right\rangle_2 = \left\langle A_{12}^{\mathsf{T}}g - A_{12}^{\mathsf{T}}\underline{C}v_2, v_2\right\rangle_2 = \left\langle A_{12}^{\mathsf{T}}g, v_2\right\rangle_2 - \left\langle A_{12}^{\mathsf{T}}\underline{C}v_2, v_2\right\rangle_2 .$$

En regroupant le tout, on trouve, avec $\alpha = \alpha_{lin} + \alpha_{quad}$,

$$J(v) = \frac{1}{2}\left\langle \{A_{22} + \underline{C}^{\mathsf{T}} A_{11} \underline{C} - 2A_{12}^{\mathsf{T}} \underline{C}\} v_2, v_2 \right\rangle_2 - \left\langle b_2 + \underline{C}^{\mathsf{T}} A_{11} g - \underline{C}^{\mathsf{T}} b_1 - A_{12}^{\mathsf{T}} g, v_2 \right\rangle_2 + \alpha.$$

Notons que l'on peut rendre le terme quadratique symétrique :

$$\left.\begin{array}{l} J(v) = \tilde{J}(v_2), \text{ avec } \tilde{J}(v_2) = \frac{1}{2}\left\langle \tilde{A}_{22} v_2, v_2 \right\rangle_2 - \left\langle \tilde{b}_2, v_2 \right\rangle_2 + \alpha \\ \tilde{A}_{22} = A_{22} + \underline{C}^{\mathsf{T}} A_{11} \underline{C} - A_{12}^{\mathsf{T}} \underline{C} - \underline{C}^{\mathsf{T}} A_{12}, \quad \tilde{b}_2 = b_2 + \underline{C}^{\mathsf{T}} A_{11} g - \underline{C}^{\mathsf{T}} b_1 - A_{12}^{\mathsf{T}} g. \end{array}\right\} \tag{6.14}$$

On peut donc remplacer $J(v)$ par $\tilde{J}(v_2)$, *pour tout* $v \in K$. Réciproquement, à chaque $v_2 \in \mathbb{R}^{n-p}$, on peut associer un unique $v^\star \in K$, égal à

$$v^\star = \begin{pmatrix} g - \underline{C} v_2 \\ v_2 \end{pmatrix} \in K, \text{ et l'on a } \tilde{J}(v_2) = J(v^\star).$$

Proposition 6.4.2. *Résoudre le problème avec contraintes est équivalent à*

$$\text{Trouver } u_2 \in \mathbb{R}^{n-p} \text{ tel que } \tilde{J}(u_2) = \min_{v_2 \in \mathbb{R}^{n-p}} \tilde{J}(v_2). \tag{6.15}$$

De plus, la matrice $\tilde{A}_{22}$ intervenant dans la fonctionnelle $\tilde{J}$ est symétrique définie-positive, ce qui permet d'utiliser les techniques énoncées auparavant.

Démonstration. Il reste à vérifier que $\tilde{A}_{22}$ est bien symétrique définie-positive. Bien sûr, $\tilde{A}_{22}$ est symétrique par construction. Par ailleurs,

$$\begin{aligned} \left\langle \tilde{A}_{22} v_2, v_2 \right\rangle_2 &= \langle A_{22} v_2, v_2 \rangle_2 + \langle A_{11} \underline{C} v_2, \underline{C} v_2 \rangle_1 - \left\langle A_{12}^{\mathsf{T}} \underline{C} v_2, v_2 \right\rangle_2 - \langle A_{12} v_2, \underline{C} v_2 \rangle_1 \\ &= \langle -A_{11} \underline{C} v_2 + A_{12} v_2, -\underline{C} v_2 \rangle_1 + \left\langle -A_{12}^{\mathsf{T}} \underline{C} v_2 + A_{22} v_2, v_2 \right\rangle_2 \\ &= \left\langle \begin{pmatrix} A_{11} & A_{12} \\ A_{12}^{\mathsf{T}} & A_{22} \end{pmatrix} \begin{pmatrix} -\underline{C} v_2 \\ v_2 \end{pmatrix}, \begin{pmatrix} -\underline{C} v_2 \\ v_2 \end{pmatrix} \right\rangle \\ &= \left\langle A \begin{pmatrix} -\underline{C} v_2 \\ v_2 \end{pmatrix}, \begin{pmatrix} -\underline{C} v_2 \\ v_2 \end{pmatrix} \right\rangle. \end{aligned}$$

Le produit scalaire est strictement positif, sauf si $\begin{pmatrix} -\underline{C} v_2 \\ v_2 \end{pmatrix} = 0$, i. e. $v_2 = 0$. $\tilde{A}_{22}$ est bien symétrique définie-positive. □

Par rapport à la première technique d'éliminations de contraintes, notons que cette méthode est très attractive, puisqu'elle ne requiert pas $p + 2$ résolutions de problèmes sans contraintes de matrice A... Par contre, deux inconvénients potentiels sont à prendre en considération

— Il faut extraire le bloc C_{11} de rang p de la matrice C.

— La structure interne de $\tilde{A}_{22}$ est complètement différente de celle de A.

En particulier, même si A est une matrice creuse, $\tilde{A}_{22}$ peut être une matrice pleine. Le coût d'un produit matrice vecteur (voir l'exemple de la section 4.1) est alors bien plus important lorsque l'on résout (6.15). Ce type de considération doit absolument être examiné, pour évaluer les mérites de la mise en oeuvre numérique.

Annexe A

Quelques rappels de calcul différentiel

Dans ce chapitre, nous rappelons les fondements du calcul différentiel, en adoptant une approche relativement abstraite, qui ne repose que marginalement sur la notion de dérivée partielle.

A.1 Différentiabilité

Soient $\mathbb{V}$ et $\mathbb{F}$ deux espaces vectoriels normés sur $\mathbb{R}$, on note $\mathcal{L}_c(\mathbb{V}, \mathbb{F})$ l'ensemble des applications linéaires et continues de $\mathbb{V}$ dans $\mathbb{F}$.

Remarque A.1.1. *Lorsque la dimension de $\mathbb{V}$ est finie, toutes les applications linéaires sont continues. C'est faux lorsque la dimension de $\mathbb{V}$ infinie !*

Dans la suite, on notera Ω un ouvert de $\mathbb{V}$ contenant u, et f une application de $\Omega \subset \mathbb{V}$ dans $\mathbb{F}$; on dit que f est **continue** en un point $u \in \Omega$ si

$$\forall h \in \mathbb{V} \qquad f(u+h) = f(u) + \varepsilon_0(h), \tag{A.1}$$

où ε_0 est une application de $\mathbb{V}$ dans $\mathbb{F}$ telle que

$$\|\varepsilon_0(h)\|_{\mathbb{F}} \to 0 \quad \text{quand} \quad \|h\|_{\mathbb{V}} \to 0.$$

Alternativement on peut spécifier la définition de continuité comme suit : pour chaque $\varepsilon > 0$, il existe $\delta > 0$ tel que $\|f(u') - f(u)\|_{\mathbb{F}} \leq \varepsilon$, pour tout $u' \in \Omega$ ayant $\|u' - u\|_{\mathbb{V}} < \delta$. Alternativement encore, f est continu en u si et seulement si l'image originelle de chaque voisinage de $f(u)$ (dans $\mathbb{F}$) est ouvert (dans $\mathbb{V}$).

La notation $\forall h \in \mathbb{V}$ sous-entend : pour tout h de $\mathbb{V}$ tel que $u+h$ appartienne à Ω. (En termes plus mathématiques, ceci signifie

$$\forall \epsilon > 0, \quad \exists \eta > 0, \quad \forall v \in \Omega, \quad \|v-u\|_{\mathbb{V}} < \eta \Longrightarrow \|f(v)-f(u)\|_{\mathbb{F}} < \epsilon.)$$

L'expression (A.1) est un **développement limité d'ordre 0** au voisinage de u.

Remarque A.1.2. *(préliminaire) Certaines des définitions de ce chapitre sont données dans le contexte général d'espaces vectoriels normés ; on peut pour simplifier se limiter au cas $\mathbb{V} = \mathbb{R}^n$, $\mathbb{F} = \mathbb{R}^p$, où $n \geq 1$ et $p \geq 1$ sont deux entiers naturels.*

Définition A.1.1. *On dit que l'application f est (Fréchet)* **différentiable** *en un point $u \in \mathbb{V}$ s'il existe g appartenant à $\mathcal{L}_c(\mathbb{V}, \mathbb{F})$, qui vérifie*

$$\forall h \in \mathbb{V} \qquad f(u+h) = f(u) + g(h) + \|h\|\, \varepsilon(h),$$

où ε est une application de $\mathbb{V}$ dans $\mathbb{F}$ telle que

$$\|\varepsilon(h)\|_{\mathbb{F}} \to 0 \quad \textit{quand} \quad \|h\|_{\mathbb{V}} \to 0,$$

i.e.,

$$\lim_{h \to 0} \frac{|f(x+h) - f(x) - \langle g, h\rangle|}{\|h\|} = 0.$$

L'application linéaire continue g est notée $df(u)$, et on l'appelle **différentielle de f en** *u. On note l'action de $df(u)$ sur h*

$$df(u)\cdot h.$$

L'expression ci-dessus correspond à un **développement limité d'ordre 1** *au voisinage de u, de la forme*

$$\forall h \in \mathbb{V} \qquad f(u+h) = f(u) + df(u)\cdot h + \|h\|\, \varepsilon(h). \tag{A.2}$$

Remarque A.1.3. *Le développement* (A.2) *peut se reécrire sous la forme :*

$$f(u+h) = f(u) + df(u)\cdot h + o(h),$$

avec la propriété $\dfrac{\|o(h)\|_{\mathbb{F}}}{\|h\|_{\mathbb{V}}} \to 0$ lorsque $\|h\|_{\mathbb{V}} \to 0$.

Proposition A.1.1. *Si la différentielle de f en u existe, elle est unique.*

Démonstration. L'unicité de la différentielle en u est obtenue de la manière élémentaire suivante. Soient deux applications linéaires continues $df_1(u)$ et $df_2(u)$ satisfaisant la relation (A.2). Alors, pour tout vecteur non nul v, et pour tout réel λ strictement positif suffisamment petit pour que $u + \lambda v$ appartienne à Ω, on a l'égalité

$$df_1(u)\cdot(\lambda v) - df_2(u)\cdot(\lambda v) = \lambda(\varepsilon_1(\lambda v) - \varepsilon_2(\lambda v)).$$

Par linéarité de $df_1(u)$ et $df_2(u)$, on arrive à

$$df_1(u)\cdot v - df_2(u)\cdot v = (\varepsilon_1(\lambda v) - \varepsilon_2(\lambda v)).$$

Si on fait tendre λ vers 0, on obtient que l'application linéaire $df_1(u) - df_2(u)$ est nulle, soit finalement $df_1(u) = df_2(u)$. □

Exercice A.1.1. *1. Soit $f : \mathbb{R} \to \mathbb{R}$ dérivable. Montrer que f est différentiable sur $\mathbb{R}$ et calculer $df(x)$, pour $x \in \mathbb{R}$.*

2. Soit $f : \mathbb{R}^n \to \mathbb{R}^p$, une application affine, $f(u) = Au + b$. Montrer que f est différentiable sur $\mathbb{R}^n$ et calculer $df(u)$, pour $u \in \mathbb{R}^n$.

3. Soit $f : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$, $f(A) = A^2$. Montrer que f est différentiable sur $\mathbb{R}^{n\times n}$ et calculer $df(A)$, pour $A \in \mathbb{R}^{n\times n}$.

4. Soit Ω_n l'ensemble des matrices inversibles de $\mathbb{R}^{n\times n}$, et $f : \Omega_n \to \Omega_n$, définie par $f(A) = A^{-1}$. Pourquoi Ω_n est-il ouvert ? Montrer que f est différentiable sur Ω_n et vérifier que, pour $A \in \Omega_n$,

$$df(A) \cdot H = -A^{-1} H\, A^{-1}.$$

5. Soit $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|$. Montrer que f n'est pas différentiable en $x = 0$, mais qu'elle l'est sur $\mathbb{R}^n \setminus \{0\}$, et calculer $df(x)$ pour $x \neq 0$.

Bien sûr, en toute généralité, toute application différentiable en un point est continue en ce point, et on retrouve la formule de limite de taux de variation ; c'est l'objet de la

Proposition A.1.2. *Si l'application f de $\mathbb{V}$ dans $\mathbb{F}$ est différentiable en u, elle est continue en ce point et*

$$\forall h \in \mathbb{V} \qquad df(u)\cdot h = \lim_{\theta\to 0^+} \frac{f(u+\theta h) - f(u)}{\theta}. \tag{A.3}$$

Démonstration. A partir de la définition de la différentiabilité, en utilisant notamment le fait que la différentielle en u est continue, on tire

$$\|f(u+h) - f(u)\|_{\mathbb{F}} \leq \|df(u)\|\ \|h\|_{\mathbb{V}} + \|h\|_{\mathbb{V}}\ \|\varepsilon(h)\|_{\mathbb{F}}$$

et ainsi

$$\|f(u+h)-f(u)\|_{\mathbb{F}} \to 0 \quad \text{quand} \quad \|h\|_{\mathbb{V}} \to 0.$$

De plus par linéarité de l'application $df(u)$,

$$\forall h \in \mathbb{V},\ \forall \theta > 0 \qquad f(u+\theta h) = f(u) + \theta df(u)\cdot h + \theta\ \|h\|\ \varepsilon(\theta h)$$

et finalement

$$\forall h \in \mathbb{V} \qquad df(u)\cdot h = \lim_{\theta \to 0^+} \frac{f(u+\theta h)-f(u)}{\theta}.$$

□

Définition A.1.2. *On dit que l'application f est* **différentiable** *dans Ω, si elle est différentiable en tout point $u \in \Omega$. Dans ce cas, on peut définir une application df qui à tout point $u \in \Omega$ associe une application linéaire et continue $df(u)$ de $\mathbb{V}$ dans $\mathbb{F}$; on l'appelle* **différentielle** *de f dans Ω. Si la différentielle df est une application continue de $\mathbb{V}$ dans $\mathcal{L}_c(\mathbb{V},\mathbb{F})$, on dit que f est une application* **continûment différentiable**, *ou encore de* **classe $\mathcal{C}^1$**

La définition A.1.1 introduit la notion de différentielle au sens de **Fréchet**. On peut également définir la différentielle de f en u, $h \mapsto df(u)\cdot h$, au sens de **Gateaux**. On parle souvent de **dérivée directionnelle**.

Définition A.1.3. *On dit que l'application f, définie sur un voisinage de u, est différentiable au sens de Gateaux s'il existe $df(u)$ de $\mathcal{L}_c(\mathbb{V},\mathbb{F})$ telle que*

$$\forall h \in \mathbb{V} \qquad df(u)\cdot h = \lim_{\theta \to 0^+} \frac{f(u+\theta h)-f(u)}{\theta}.$$

On peut aussi écrire la définition équivalente, pour chaque h *dans* $\mathbb{V}$

$$f(u+\theta\, h) = f(u) + \theta\, df(u)\cdot h + o(\theta), \qquad \theta \geq 0, \tag{A.4}$$

avec la propriété $\dfrac{\|o(\theta)\|_{\mathbb{F}}}{\theta} \to 0$ *lorsque* $\theta \to 0^+$.

Proposition A.1.3. *Si $\mathbb{V} = \mathbb{R}$, Fréchet-différentiabilité et Gateaux-différentiabilité coïncident.*

Supposons maintenant que $dim(\mathbb{V}) \geq 2$. La différence entre (A.2) écrite avec $\theta\, h$ au lieu de h et (A.4) se trouve dans l'expression du reste

- $\theta\, \|h\| \varepsilon(\theta\, h)$ pour la Fréchet-différentiabilité ;
- $o(\theta)$ pour la Gateaux-différentiabilité.

En d'autres termes, elle est *uniforme* en h pour la première, ce qui n'est pas garanti pour la seconde. De manière plus imagée, la Gateaux-différentiabilité est la différentiabilité le long de toute *droite* passant par u, alors que la Fréchet-différentiabilité correspond à la différentiabilité le long de toute *courbe* passant par u. De façon générale, la proposition A.1.2 montre qu'une application différentiable au sens de Fréchet est toujours différentiable au sens de Gateaux (et les différentielles sont égales !), alors que la réciproque est fausse. D'ailleurs, la Gateaux-différentiabilité n'implique même pas la continuité, comme le montre le contre-exemple qui suit.

Exercice A.1.2. *On se place dans* $\mathbb{V} = \mathbb{R}^2$. *Soient* $q \geq p > 5$ *deux réels. Montrer que la fonctionnelle* f *définie par*

$$f(x,y) = \begin{cases} \dfrac{x^p}{(y-x^2)^2 + x^q} & \text{si } (x,y) \neq (0,0) \\ 0 & \text{si } (x,y) = (0,0) \end{cases} .$$

est différentiable au point $(0,0)$ *au sens de Gateaux, mais qu'elle n'est pas continue en ce point.*

Exercice A.1.3. *Soit encore* $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|$. *Vérifier que* f *n'est pas Gateaux-differentiable en* $x = 0$.

Plaçons nous maintenant dans le cas, important en pratique, où $\mathbb{V} = \mathbb{R}^n$ et $\mathbb{F} = \mathbb{R}$, c'est-à-dire que f est à valeurs réelles. On munit $\mathbb{R}^n$ d'un produit scalaire, noté $\langle \cdot, \cdot \rangle$ dans la suite. Si f est différentiable en u, sa différentielle $df(u)$ est une forme linéaire de $\mathbb{R}^n$ dans $\mathbb{R}$. A cette forme, on peut associer un unique vecteur de $\mathbb{R}^n$, appelé **gradient de** f **en** u, et noté $\nabla f(u)$, tel que

$$\forall h \in \mathbb{V} \qquad df(u)\cdot h = \langle \nabla f(u), h \rangle . \tag{A.5}$$

Dans ce cas particulier, la formule (A.2) prend la forme

$$\forall h \in \mathbb{V} \qquad f(u+h) = f(u) + \langle \nabla f(u), h \rangle + \|h\| \, \varepsilon(h). \tag{A.6}$$

Par exemple, avec J définie en (5.5), qu'obtient-on comme expressions de $dJ(u)\cdot h$ et $\nabla J(u)$ pour u et h éléments de $\mathbb{R}^n$? L'expression (A.4) nous fournit la réponse ; en effet, on a

$$J(u+h) - J(u) = \langle Au - b, h \rangle + \frac{1}{2} \langle Ah, h \rangle .$$

D'après l'inégalité de Cauchy-Schwarz et par définition de la norme matricielle induite par la norme Euclidienne, on trouve

$$| \langle Ah, h \rangle | \leq \|Ah\| \, \|h\| \leq \|A\| \, \|h\|^2 \ ;$$

ainsi, par identification, on trouve que

$$\varepsilon(h) = \frac{1}{2}\frac{\langle Ah, h\rangle}{\|h\|}, \quad et \quad |\varepsilon(h)| \leq \frac{1}{2}\|A\|\,\|h\| \to 0 \; quand \; \|h\| \to 0.$$

On infère immédiatement que

$$dJ(u)\cdot h = \langle Au - b, h\rangle\,, \text{ et } \nabla J(u) = Au - b.$$

Remarque A.1.4. *Lorsque la matrice A n'est pas symétrique, les expressions ci-dessus sont fausses ! En effet, on doit remplacer A par $\frac{1}{2}(A + A^{\mathsf{T}})$.*

Le gradient dépend seulement du produit scalaire. En particulier, il est *indépendant* de la base de l'espace *Euclidien* $\mathbb{R}^n$. Supposons maintenant que $\mathbb{R}^n$ soit muni d'une base *orthonormale* $(e_k)_{1\leq k\leq n}$, et soit $(x_k)_{1\leq k\leq n}$ le système de coordonnées associé : $u = \sum_{k=1}^n x_k\, e_k$.
Dans la base $(e_k)_{1\leq k\leq n}$, le vecteur $\nabla f(u)$ a pour composantes

$$\nabla f(u) = \begin{pmatrix} \partial_1 f(u) \\ \partial_2 f(u) \\ \vdots \\ \partial_n f(u) \end{pmatrix}. \tag{A.7}$$

On note aussi $\dfrac{\partial f}{\partial x_k}(u)$ ses composantes ; $\partial_k f(u)$ est appelée $k^{ème}$ **dérivée partielle** de f en u.

Remarque A.1.5. *pourquoi parle-t-on de dérivée partielle ? La raison en est simple. Si on choisit $h = \theta e_k$ dans (A.6), on arrive à*

$$f(u + \theta e_k) = f(u) + \theta\,\langle \nabla f(u), e_k\rangle + |\theta|\,\varepsilon(\theta e_k) = f(u) + \theta\frac{\partial f}{\partial x_k}(u) + |\theta|\,\varepsilon(\theta e_k).$$

Par ailleurs, modulo un petit abus de notations, on peut réécrire $f(u)$ sous la forme $f(x_1, \cdots, x_n)$. En d'autres termes, $\dfrac{\partial f}{\partial x_k}(u)$ représente la dérivée de f en u dans la direction e_k, ce qui correspond finalement à la dérivée de l'application

$$\theta \mapsto f(x_1, \cdots, x_{k-1}, x_k + \theta, x_{k+1}, \cdots, x_n) \; en \; \theta = 0.$$

Exercice A.1.4. *Vérifier que si f est différentiable en u, elle admet une dérivée partielle par rapport à chaque variable en ce point. Réciproquement, montrer que, si f admet des dérivées partielles sur Ω qui sont continues en u, alors f est différentiable en u et que, de plus, elle est de classe $\mathcal{C}^1$ sur un ouvert contenant u.*

Dans le cas où $\mathbb{F} = \mathbb{R}^p$, $f(u)$ correspond à un vecteur à p composantes

$$f(u) = \begin{pmatrix} f_1(u) \\ f_2(u) \\ \vdots \\ f_p(u) \end{pmatrix}, \quad \text{ou } f(u) = \sum_{l=1}^{p} f_l(u) e'_l,$$

dès lors que l'on a choisi une base $(e'_l)_{1 \leq l \leq p}$ de $\mathbb{F}$. On peut reprendre la construction ci-dessus, et différencier chaque composante de f. La différentielle de f en u (lorsqu'elle existe) peut alors être écrite composante par composante

$$\begin{aligned}
df_1(u)\cdot h &= \langle \nabla f_1(u), h \rangle = \partial_1 f_1(u) h_1 + \partial_2 f_1(u) h_2 + \ldots + \partial_n f_1(u) h_n \\
df_2(u)\cdot h &= \langle \nabla f_2(u), h \rangle = \partial_1 f_2(u) h_1 + \partial_2 f_2(u) h_2 + \ldots + \partial_n f_2(u) h_n \\
\vdots &= \vdots \\
df_p(u)\cdot h &= \langle \nabla f_p(u), h \rangle = \partial_1 f_p(u) h_1 + \partial_2 f_p(u) h_2 + \ldots + \partial_n f_p(u) h_n.
\end{aligned}$$

La matrice associée à $df(u)$ dans les bases $(e_k)_{1 \leq k \leq n}$ et $(e'_l)_{1 \leq l \leq p}$ est appelée **matrice jacobienne de f en** u, et on la note $[df(u)]$:

$$[df(u)] = \begin{pmatrix} \partial_1 f_1(u) & \partial_2 f_1(u) & \ldots & \partial_n f_1(u) \\ \partial_1 f_2(u) & \partial_2 f_2(u) & \ldots & \partial_n f_2(u) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 f_p(u) & \partial_2 f_p(u) & \ldots & \partial_n f_p(u) \end{pmatrix}.$$

Lorsque $n = p$, son déterminant est appelé **jacobien** de f en u, égal à

$$J_{f(u)} = \begin{vmatrix} \partial_1 f_1(u) & \partial_2 f_1(u) & \ldots & \partial_n f_1(u) \\ \partial_1 f_2(u) & \partial_2 f_2(u) & \ldots & \partial_n f_2(u) \\ \vdots & \vdots & \ddots & \vdots \\ \partial_1 f_n(u) & \partial_2 f_n(u) & \ldots & \partial_n f_n(u) \end{vmatrix}.$$

Remarque A.1.6. *Revenons un instant au cas $\mathbb{F} = \mathbb{R}$, c'est-à-dire $p = 1$. $[df(u)]$ est un* vecteur ligne*, égal à $\begin{pmatrix} \partial_1 f(u) & \partial_2 f(u) & \ldots & \partial_n f(u) \end{pmatrix}$. Si on compare cette expression à celle du* vecteur colonne $\nabla f(u)$*, i.e. (A.7), on en déduit que dans ce cas*

$$\nabla f(u) = [df(u)]^{\mathsf{T}}.$$

A.2 Propriétés de la différentielle

Nous démontrons quelques résultats simples, concernant l'addition et la composition d'applications différentiables.

Proposition A.2.1. *Soient f et g deux applications de $\mathbb{V}$ dans $\mathbb{F}$ Fréchet-différentiables en $u \in \mathbb{V}$, alors l'application $f+g$ est Fréchet-différentiable en u et $d(f+g)(u) = df(u) + dg(u)$.*

Démonstration. Des relations

$$f(u+h) = f(u) + df(u)\cdot h + \|h\|\, \varepsilon_f(h)$$

$$g(u+h) = g(u) + dg(u)\cdot h + \|h\|\, \varepsilon_g(h)$$

on tire par addition

$$(f+g)(u+h) = (f+g)(u) + df(u)\cdot h + dg(u)\cdot h + \|h\|(\, \varepsilon_f(h) + \, \varepsilon_g(h)).$$

Comme

$$\|(\varepsilon_f(h) + \varepsilon_g(h))\|_{\mathbb{F}} \to 0 \quad \text{quand} \quad \|h\|_{\mathbb{V}} \to 0$$

on voit que l'application linéaire $d(f+g)(u)$ définie par

$$d(f+g)(u)\cdot h = df(u)\cdot h + dg(u)\cdot h$$

correspond à la définition de la différentiabilité en u de $f+g$. □

Remarque A.2.1. *De la même façon, on peut prouver que la somme de deux applications Gateaux-différentiables en un point est Gateaux-différentiable.*

Proposition A.2.2. *Soit f une application de $\mathbb{V}$ dans $\mathbb{F}$ Fréchet-différentiable en $u \in \mathbb{V}$, et soit g une application de $\mathbb{F}$ dans $\mathbb{G}$ Fréchet-différentiable en $f(u) \in \mathbb{F}$, alors l'application $g \circ f$ est Fréchet-différentiable en u et*

$$d(g\circ f)(u) = dg(f(u)) \circ df(u).$$

Démonstration. Des relations

$$\begin{aligned} f(u+h) &= f(u) + df(u)\cdot h + \|h\|\, \varepsilon_f(h) \\ g(f(u)+h') &= g(f(u)) + dg(f(u))\cdot h' + \|h'\|\, \varepsilon_g(h'), \end{aligned}$$

on tire

$$\begin{aligned} g\circ f(u+h) &= g(f(u+h)) \\ &= g\left(f(u) + df(u)\cdot h + \|h\|\, \varepsilon_f(h)\right) \\ &= g(f(u)+h') \quad \text{avec} \quad h' = df(u)\cdot h + \|h\|\, \varepsilon_f(h) \\ &= g(f(u)) + dg(f(u))\cdot h' + \|h'\|\, \varepsilon_g(h'). \end{aligned}$$

Mais l'application différentielle $dg(f(u))$ est linéaire par définition, d'où

$$dg(f(u))\cdot h' = dg(f(u))\cdot(df(u)\cdot h) + \|h\|\, dg(f(u))\cdot(\varepsilon_f(h)).$$

On arrive alors à l'expression

$$g\circ f(u+h) = g\circ f(u) + \{dg(f(u))\circ df(u)\}\cdot h + \|h\|\, dg(f(u))\cdot(\varepsilon_f(h)) + \|h'\|\, \varepsilon_g(h').$$

Il suffit maintenant de vérifier que les deux termes de droite peuvent être réécrits sous la forme $\|h\|\, \varepsilon_{g\circ f}(h)$, avec $\|\varepsilon_{g\circ f}(h)\| \to 0$ lorsque $\|h\| \to 0$. Or, on a d'une part

$$\|dg(f(u))\cdot(\varepsilon_f(h))\| \leq \|dg(f(u))\|\, \|\varepsilon_f(h))\| \to 0 \quad quand \quad \|h\| \to 0\ ;$$

et d'autre part

$$\|h'\| \leq \|df(u)\|\, \|h\| + \|h\|\, \|\varepsilon_f(h)\| = o(1).$$

On obtient finalement

$$g\circ f(u+h) = g\circ f(u) + \{dg(f(u))\circ df(u)\}\cdot h + \|h\|\, \varepsilon_{g\circ f}(h).$$

□

On a également le résultat suivant, si l'on affaiblit l'hypothèse sur f.

Proposition A.2.3. *Soit f une application de $\mathbb{V}$ dans $\mathbb{F}$ Gateaux-différentiable en $u \in \mathbb{V}$, et soit g une application de $\mathbb{F}$ dans $\mathbb{G}$ Fréchet-différentiable en $f(u) \in \mathbb{F}$, alors l'application $g\circ f$ est Gateaux-différentiable en u et*

$$d(g\circ f)(u) = dg(f(u))\circ df(u).$$

Démonstration. Soit $h \in \mathbb{V}$ donné : $f(u+\theta\, h) = f(u) + \theta\, df(u)\cdot h + o(\theta)$.
Si on note $h'_\theta = \theta\, df(u)\cdot h + o(\theta)$, on a en particulier $\dfrac{\|h'_\theta\|}{\theta}$ borné lorsque $\theta > 0$ est petit.

$$\begin{aligned} g\circ f(u+\theta\, h) - g\circ f(u) &= dg(f(u))\cdot h'_\theta + \|h'_\theta\|\varepsilon_g(h'_\theta) \\ &= \theta\{dg(f(u))\circ(df(u))\}\cdot h + dg(f(u))\cdot o(\theta) + \|h'_\theta\|\varepsilon_g(h'_\theta), \text{d'où} \\ \frac{g\circ f(u+\theta\, h) - g\circ f(u)}{\theta} &= \{dg(f(u))\circ(df(u))\}\cdot h + dg(f(u))\cdot o(1) + \frac{\|h'_\theta\|}{\theta}\varepsilon_g(h'_\theta). \end{aligned}$$

Comme $\|h'_\theta\| \to 0$ lorsque $\theta \to 0^+$, on a $\|\varepsilon_g(h'_\theta)\| \to 0^+$. Ainsi

$$\lim_{\theta\to 0^+} \frac{g\circ f(u+\theta\, h) - g\circ f(u)}{\theta} = [dg(f(u))\circ df(u)]\cdot h.$$

□

Remarque A.2.2. *La Fréchet-différentiabilité de g est nécessaire pour pouvoir considérer la différentielle de la composée. Le résultat sur la composition est faux, si l'on suppose uniquement que g est Gateaux-différentiable, même si f est Fréchet-différentiable.*

Posons $v = f(u)$.

Lorsque $\mathbb{V} = \mathbb{R}^n$, $\mathbb{F} = \mathbb{R}^p$ et $\mathbb{G} = \mathbb{R}^m$, et que chacun de ces trois espaces est muni d'une base orthonormale, $df(u)$ est représentée par une matrice de $\mathbb{R}^{p\times n}$, $dg(v)$ par une matrice de $\mathbb{R}^{m\times p}$ et $d(g\circ f)(u)$ par une matrice de $\mathbb{R}^{m\times n}$. D'après la proposition A.2.2, $[d(g\circ f)(u)]$ est égale au produit des matrices associées à $dg(v)$ et $df(u)$:

$$\begin{aligned}[d(g\circ f)(u)] &= [dg(v)]\,[df(u)] \qquad (A.8)\\ &= \begin{pmatrix} \partial_1 g_1(v) & \partial_2 g_1(v) & \dots & \partial_p g_1(v)\\ \partial_1 g_2(v) & \partial_2 g_2(v) & \dots & \partial_p g_2(v)\\ \dots & \dots & \dots & \dots\\ \partial_1 g_m(v) & \partial_2 g_m(v) & \dots & \partial_p g_m(v) \end{pmatrix} \begin{pmatrix} \partial_1 f_1(u) & \partial_2 f_1(u) & \dots & \partial_n f_1(u)\\ \partial_1 f_2(u) & \partial_2 f_2(u) & \dots & \partial_n f_2(u)\\ \dots & \dots & \dots & \dots\\ \partial_1 f_p(u) & \partial_2 f_p(u) & \dots & \partial_n f_p(u) \end{pmatrix},\end{aligned}$$

que l'on écrit composante par composante

$$\frac{\partial (g\circ f)_i}{\partial x_j}(u) = \sum_{k=1}^{p} \frac{\partial g_i}{\partial x_k}(v)\frac{\partial f_k}{\partial x_j}(u) \qquad 1\le i\le m,\quad 1\le j\le n. \qquad (A.9)$$

Dans le cas où la fonctionnelle g est à valeurs dans $\mathbb{R}$ (soit $m = 1$), on a vu que $[dg(v)] = \nabla g(v)^\mathsf{T}$ (cf. remarque A.1.6) ; $g\circ f$ est également à valeurs dans $\mathbb{R}$, et l'on a de même $[d(g\circ f)(u)] = \nabla(g\circ f)(u)^\mathsf{T}$. En transposant (A.8), on en déduit finalement que

$$\nabla(g\circ f)(u) = [df(u)]^\mathsf{T}\nabla g(v) \qquad avec \quad v = f(u).$$

Exercice A.2.1. *Soit toujours $f : \mathbb{R}^n \to \mathbb{R}$, $f(x) = \|x\|$. En l'écrivant sous la forme $f = s\circ g$, s et g à déterminer, retrouver l'expression de $df(x)$ et de son gradient, lorsque x est non nul.*

A.3 Différentielles d'ordre supérieur et formules de Taylor

Dans cette section, on considère des applications différentiables au sens de Fréchet.

A.3.1 Différentielles d'ordre supérieur

On suppose que f est différentiable en u. Si la différentielle df est elle-même différentiable en u, on définit l'application $d^2f(u)$, appelée **différentielle seconde** de l'application f en u, et on dit que f est deux fois différentiable au point u ; $d^2f(u)$ appartient à $\mathcal{L}_c(\mathbb{V}, \mathcal{L}_c(\mathbb{V}, \mathbb{F}))$. Si la différentielle d^2f est une application continue de $\mathbb{V}$ dans $\mathcal{L}_c(\mathbb{V}, \mathcal{L}_c(\mathbb{V}, \mathbb{F}))$, on dit que f est une application de classe $\mathcal{C}^2$.

Remarque A.3.1. *Pour définir la différentielle seconde* **en** u*, on doit supposer que* f *est différentiable sur un* **voisinage de** u*.*

Remarque A.3.2 (Identification de $\mathcal{L}_c(\mathbb{V}, \mathcal{L}_c(\mathbb{V}, \mathbb{F}))$ à $\mathcal{L}_c(\mathbb{V} \times \mathbb{V}, \mathbb{F})$)**.** *Si* $A \in \mathcal{L}_c(\mathbb{V} \times \mathbb{V}, \mathbb{F})$*, ceci signifie que* A *est bilinéaire en* (x, y)*, et qu'il existe* $C \in \mathbb{R}$ *tel que*

$$\sup_{\|x\|=1, \|y\|=1} \|A(x, y)\| \leq C.$$

(i) Pour $x \in \mathbb{V}$*, soit* $A_x = A(x, \cdot)$ *:* A_x *est linéaire et* $\sup_{\|y\|=1} \|A_x(y)\| = \sup_{\|y\|=1} \|A(x, y)\| \leq C \|x\|$*. Ainsi,* $A_x \in \mathcal{L}_c(\mathbb{V}, \mathbb{F})$*, et* $\|A_x\| \leq C_x$*, avec* $C_x = C \|x\|$*.*
(ii) Soit maintenant $\tilde{A} : x \to A_x$*. Comme* A *est linéaire en sa première variable,* $\tilde{A}$ *est un élément de* $\mathcal{L}(\mathbb{V}, \mathcal{L}_c(\mathbb{V}, \mathbb{F}))$*. Il reste à vérifier la continuité, or*

$$\sup_{\|x\|=1} \|\tilde{A}(x)\| \leq \sup_{\|x\|=1} C_x \leq C.$$

Réciproquement, soit $\tilde{A} \in \mathcal{L}_c(\mathbb{V}, \mathcal{L}_c(\mathbb{V}, \mathbb{F}))$*. On définit* $A : (x, y) \to \tilde{A}(x)(y)$*. Par construction,* A *est bilinéaire de* $\mathbb{V} \times \mathbb{V}$ *dans* $\mathbb{F}$ *et, par ailleurs, comme* $\tilde{A}(x) \in \mathcal{L}_c(\mathbb{V}, \mathbb{F})$ *pour tout* x*,*

$$\sup_{\|x\|=1, \|y\|=1} \|A(x, y)\| = \sup_{\|x\|=1, \|y\|=1} \|\tilde{A}(x)(y)\| \leq \sup_{\|x\|=1} \|\tilde{A}(x)\| \leq \|\tilde{A}\|.$$

On peut *identifier* $\mathcal{L}_c(\mathbb{V}, \mathcal{L}_c(\mathbb{V}, \mathbb{F}))$ à $\mathcal{L}_c(\mathbb{V} \times \mathbb{V}, \mathbb{F})$, et on écrit donc :

$$(d^2f(u) \cdot h) \cdot h' = d^2f(u) \cdot (h, h'), \qquad (h, h') \in \mathbb{V} \times \mathbb{V}.$$

Si $h' = h$, on condense les notations en $d^2f(u) \cdot h^2$.

On rappelle le

Théorème A.3.1. *(de* **Schwarz***) Soit* f *une application deux fois différentiable en* u*. Alors* $d^2f(u)$ *est une application (bilinéaire, continue et) symétrique de* $\mathbb{V} \times \mathbb{V}$ *dans* $\mathbb{F}$*.*

Replaçons-nous maintenant dans le cadre qui nous a permis de définir les dérivées partielles (premières), c'est-à-dire $\mathbb{V} = \mathbb{R}^n$ et $\mathbb{F} = \mathbb{R}$: $d^2f(u)$ est une forme bilinéaire et continue de $\mathbb{R}^{n\times n}$. D'après l'identification ci-dessus, il existe un **unique** élément $\nabla^2 f(u)$ de $\mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ tel que

$$d^2f(u)\cdot(h,h') = \left\langle \nabla^2 f(u)h, h' \right\rangle, \qquad h, h' \in \mathbb{R}^n.$$

Encore une fois, $\nabla^2 f(u)$ est indépendant de la base choisie.

Munissons $\mathbb{R}^n$ d'une base orthonormée ; on peut construire, de la même façon que les dérivées partielles premières, les dérivées partielles secondes, notées $\dfrac{\partial^2 f}{\partial x_k \partial x_l}(u)$ ou $\partial_k\partial_l f(u)$. La différentielle seconde $d^2f(u)$ est formée des $n \times n$ dérivées partielles $\partial_l\partial_k f(u)$ de chacune des composantes $\partial_k f(u)$ du gradient. On se trouve donc dans le cas du calcul de la différentielle d'une application de $\mathbb{R}^n$ dans lui-même, et on représente $d^2f(u)$ par la **matrice Hessienne** :

$$[\nabla^2 f(u)] = \begin{pmatrix} \partial_1\partial_1 f(u) & \partial_2\partial_1 f(u) & \dots & \partial_n\partial_1 f(u) \\ \partial_1\partial_2 f(u) & \partial_2\partial_2 f(u) & \dots & \partial_n\partial_2 f(u) \\ \dots & \dots & \dots & \dots \\ \partial_1\partial_n f(u) & \partial_2\partial_n f(u) & \dots & \partial_n\partial_n f(u) \end{pmatrix}.$$

En particulier, on peut écrire : $d^2f(u)\cdot(h,h') = (\nabla^2 f(u)h, h') = \sum_{i,j=1}^n h_i\, h'_j \partial_i\partial_j f(u)$.

Enfin, on infère immédiatement du théorème de Schwarz le

Corollaire A.3.1. $[\nabla^2 f(u)]$ *est une matrice symétrique.*

Exercice A.3.1. *Soit J_0 définie en (5.5). Calculer $d^2J_0(u)\cdot(h,h')$ et $[\nabla^2 J_0(u)]$ pour u, h et h' éléments de $\mathbb{R}^n$.*

Bien évidemment, il est loisible de définir, par récurrence, les différentielles d'ordre supérieur ($k \geq 3$), à partir de ce qui est écrit ci-dessus :

$$d^k f(u)\cdot(h, h', \cdots, h^{(k-1)}), \qquad h, h', \cdots, h^{(k-1)} \in \mathbb{V}\times\mathbb{V}\times\cdots\times\mathbb{V}.$$

Lorsque tous les arguments sont identiques, on adopte la notation $d^k f(u)\cdot h^k$. Si la différentielle $d^k f$ est une application continue de $\mathbb{V}$ dans l'espace $\mathcal{L}_c(\mathbb{V}\times\mathbb{V}\times\cdots\times\mathbb{V}, \mathbb{F})$, on dit que f est une application de classe $\mathcal{C}^k(\Omega)$.. Pour définir la différentielle d'ordre k **en** u, on doit supposer que f est $k-1$ fois différentiable sur un **voisinage de** u.

A.3.2 Formules de Taylor

Nous énonçons pour finir quelques résultats concernant les formules de Taylor des applications différentiables.

On suppose que l'application f de $\mathbb{V}$ dans $\mathbb{F}$ est k fois différentiable en u, avec $k \geq 0$ (si $k = 0$, ceci signifie simplement que f est continue en u). Pour h suffisamment petit, c'est-à-dire tel que $u + h \in \Omega$, on introduit le **reste de rang k de f en u**

$$r_k(h) = f(u+h) - f(u) - \sum_{m=1}^{k} \frac{1}{m!} d^m f(u) \cdot h^m.$$

En d'autres termes, on écrit le **développement limité d'ordre** k au voisinage de u

$$f(u+h) = f(u) + \sum_{m=1}^{k} d^m f(u) \cdot h^m + r_k(h). \tag{A.10}$$

Remarque A.3.3. *Supposons par exemple que $\mathbb{V} = \mathbb{R}^n$, $\mathbb{F} = \mathbb{R}$ et $k = 2$; dans ce cas particulier*

$$f(u+h) = f(u) + \langle \nabla f(u), h \rangle + \frac{1}{2} \left\langle \nabla^2 f(u) h, h \right\rangle + r_2(h).$$

L'objet des résultats ci-dessous (pour lesquels on indique brièvement l'idée de la démonstration, voir également [10]) est d'estimer le reste $r_k(h)$. Pour u et v deux éléments de $\mathbb{V}$, on appelle $[u, v]$ le segment défini par

$$[u, v] = \{w \in \mathbb{V} \ : \ \exists \lambda \in [0,1], \ w = \lambda u + (1-\lambda) v\}.$$

On suppose ici que f est un peu plus régulière.

Théorème A.3.2. *(***inégalité de Taylor–Lagrange***) Supposons que f soit de classe $\mathcal{C}^k$ sur Ω. On choisit h tel que le segment $[u, u+h]$ soit inclus dans Ω. On suppose de plus que f admet en tout point de $]u, u+h[$ une différentielle d'ordre $k+1$, dont la norme est majorée par M uniformément sur $]u, u+h[$. Alors, le reste r_k vérifie*

$$\|r_k(h)\| \leq \frac{1}{(k+1)!} M \|h\|^{k+1}.$$

Démonstration. On note $\gamma(t) = u + t\,h$ le chemin défini le long du segment $[u, v]$, pour $t \in [0, 1]$, ce qui permet d'introduire la fonction $\mu : t \mapsto f \circ \gamma(t)$.

On applique ensuite l'inégalité de Taylor-Lagrange pour μ, fonction d'une variable réelle. □

Lorsque $k = 0$, l'inégalité précédente est appelée **inégalité des accroissements finis**.

Théorème A.3.3. *(***formule de Taylor–Mac Laurin***) On se place dans le cas d'une fonctionnelle à valeurs numériques, c'est-à-dire que $\mathbb{F} = \mathbb{R}$. Supposons que f soit de classe $\mathcal{C}^k$ sur Ω. On choisit h tel que le segment $[u, u+h]$ soit inclus dans Ω. On suppose de plus que f admet en tout point de $]u, u+h[$ une différentielle d'ordre $k+1$. Alors, il existe $\lambda \in]0, 1[$ tel que*

$$r_k(h) = \frac{1}{(k+1)!} d^{k+1} f(u + \lambda h) \cdot h^{k+1}.$$

Démonstration. On procède comme pour l'inégalité de Taylor-Lagrange. □

Remarque A.3.4. *Le théorème A.3.3 est faux si $\mathbb{F} \neq \mathbb{R}$.*

Et, si f est encore un peu plus régulière, on obtient le

Théorème A.3.4. *(***du reste intégral***) Supposons que f soit de classe $\mathcal{C}^{k+1}$ sur Ω. On choisit h tel que le segment $[u, u+h]$ soit inclus dans Ω. Alors, $r_k(h)$ est égal à*

$$r_k(h) = \int_0^1 \frac{(1-\theta)^k}{k!} d^{k+1} f(u + \theta h) \cdot h^{k+1} \, d\theta.$$

Démonstration. On procède comme pour l'inégalité de Taylor-Lagrange. □

Pour finir, si l'on en revient à la régularité initiale, on a le

Théorème A.3.5. *(***de Taylor–Young***) Soit f une application k fois différentiable en u. Le reste r_k vérifie*

$$\|r_k(h)\| = o(\|h\|^k).$$

Démonstration. La démonstration est faite par récurrence sur k.

Lorsque $k = 1$, on retrouve la définition de la différentiabilité de f en u. Par récurrence, on différencie le reste r_{k+1}, et on utilise la formule des accroissements finis, en notant que la différentielle de $h \mapsto d^m f(u) \cdot h^m$ est

— si $m = 1$: $df(u)$.

— si $m > 1$: $h \mapsto m \, d^m f(u) \cdot h^{m-1}$.

□

Annexe B

Quelques rappels de l'algèbre linéaire

Nous rappelons ici quelques résultats d'algèbre linéaire matricielle. Dans tout le chapitre, on notera par $\mathbb{R}^{n\times n}$ l'espace des matrices carrées réelles et par $\mathbb{C}^{n\times n}$ celui des matrices carrées complexes.

B.1 Normes matricielles

Définition B.1.1. *Soit $\|\cdot\|$ une norme sur $\mathbb{R}^n$. On lui associe une norme matricielle sur $\mathbb{R}^{n\times n}$, dite norme induite de la norme vectorielle, définie par*

$$\|A\| := \sup_{x\in\mathbb{R}^n, \|x\|=1} \|Ax\|.$$

De même, si $\|\cdot\|$ est une norme sur $\mathbb{C}^n$, on définit la norme matricielle sur $\mathbb{C}^{n\times n}$ induite de la norme vectorielle par :

$$\|A\| := \sup_{x\in\mathbb{C}^n, \|x\|=1} \|Ax\|.$$

Tant qu'il n'y a pas de confusion, on note de la même façon les normes vectorielle et matricielle induite.

Il est facile de vérifier le résultat suivant.

Proposition B.1.1. *Soit $\|\cdot\|$ une norme matricielle induite sur $\mathbb{R}^{n\times n}$.*

1. *Pour toute matrice $A \in \mathbb{R}^{n\times n}$, il existe $x_A \in \mathbb{R}^n$ avec $\|x_A\| = 1$ et $\|A\| = \|Ax_A\|$.*

2. *La norme de la matrice identité I_n vaut 1 : $\|I_n\| = 1$.*
3. *Soient $A, B \in \mathbb{R}^{n\times n}$. On a : $\|AB\| \leq \|A\|\,\|B\|$. (Les mêmes propriétés restent valables pour les normes sur $\mathbb{C}^n$.)*

Démonstration. 1. Remarquons que la fonction $x \in \mathbb{R}^n \longmapsto \|Ax\|$ est continue et l'ensemble $S := \{x \in \mathbb{R}^n \mid \|x\| = 1\}$ est un compact. Donc il existe un $x_A \in S$ tel que $\|A\| := \sup_{x\in S} \|Ax\| = \|Ax_A\|$.

2. évident.

3. Conséquence directe de l'inégalité : $\|ABx\| \leq \|A\|\,\|Bx\| \leq \|A\|\,\|B\|\,\|x\|$.

□

Il est possible de définir des normes matricielles qui ne sont pas induites. Prenons par exemple le cas de la norme définie sur $\mathbb{R}^{n\times n}$ par : $\|A\|^2 = \sum_{i,j=1}^n A_{ij}^2$. En effet, on a $\|I_n\| = \sqrt{n}$, ce qui n'est pas possible pour une norme induite.

Dans la suite, nous allons nous intéresser en particulier aux normes matricielles sur $\mathbb{R}^{n\times n}$ induites des normes vectorielles $\|\cdot\|_2$ et $\|\cdot\|_\infty$ (définies pour $x \in \mathbb{R}^n$, par $\|x\|_2^2 = \sum_i x_i^2$ et $\|x\|_\infty = \max_i |x_i|$).

Proposition B.1.2. *La norme matricielle $\|\cdot\|_2$ vérifie :*

$$\|A\|_2 = \|A^\mathsf{T}\|_2 = \sqrt{\lambda_{\max}(A^\mathsf{T}A)} \qquad \forall A \in \mathbb{R}^{n\times n}.$$

De plus, si $A \in \mathbb{R}^{n\times n}$ est symétrique, alors

$$\|A\|_2 = |\lambda_{\max}(A)|.$$

(Ici on a désigné par $\lambda_{\max}(B)$ la plus grande valeur propre de B.)

Démonstration. Remarquons d'abord qu'on a :

$$\|A\|_2^2 = \sup_{x\in\mathbb{R}^n,\ \|x\|=1} \|Ax\|^2 = \sup_{x\in\mathbb{R}^n,\ \|x\|=1} (A^\mathsf{T}Ax, x) = \|A^\mathsf{T}A\|.$$

Comme la matrice $A^\mathsf{T}A$ est symétrique réelle positive, elle est diagonalisable et ses valeurs propres sont positives $\lambda_{\min}(A^\mathsf{T}A) = \lambda_1(A^\mathsf{T}A) \leq \cdots \leq \lambda_n(A^\mathsf{T}A) = \lambda_{\max}(A^\mathsf{T}A)$. Et dans la base formée par les vecteurs propres, on vérifie que :

$$\sup_{x\in\mathbb{R}^n,\ \|x\|=1} \left\langle A^\mathsf{T}Ax, x\right\rangle = \lambda_{\max}(A^\mathsf{T}A).$$

D'où $\|A\|_2^2 = \lambda_{\max}(A^\mathsf{T}A)$. De plus, si A est symétrique positive, alors $A^\mathsf{T}A = A^2$ et $\lambda_{\max}(A^\mathsf{T}A) = \lambda_{\max}(A)^2$ avec $\lambda_{\max}(A) \geq 0$. On en déduit que $\|A\|_2 = \lambda_{\max}(A)$. □

Exercice B.1.1. *Soit A une matrice symétrique inversible. Que vaut la norme $\|A^{-1}\|_2$?*

Proposition B.1.3. *La norme matricielle $\|\cdot\|_\infty$ vérifie :*

$$\|A\|_\infty = \max_i \sum_j |A_{ij}| \qquad \forall A = (A_{ij})_{ij} \in \mathbb{R}^{n\times n}.$$

Démonstration. est laissé à titre d'exercice. □

Définition B.1.2. *Soit A une matrice dans $\mathbb{C}^{n\times n}$. On appelle rayon spectral de A, et on note $\rho(A)$, le maximum des modules des valeurs propres de A.*

Le rayon spectral est défini pour toutes les matrices, même dans le cas où les valeurs propres ne sont pas réelles. dans le cas particulier où A est symétrique, on a le résultat suivant.

Lemme B.1.1. *Si A est une matrice symétrique, alors $\rho(A) = \|A\|_2$.*

Dans le cas général, on peut toujours trouver une norme induite "comparable" au rayon spectral. Plus précisément, on a : (résultat admis)

Proposition B.1.4. *Soit $\|\cdot\|$ une norme induite sur $\mathbb{C}^n$. On a :*

$$\rho(A) \leq \|A\|.$$

Réciproquement, pour toute matrice A et pour tout réel $\varepsilon > 0$, il existe une norme matricielle induite $\|\cdot\|$ sur $\mathbb{C}^{n\times n}$ (qui dépend de A et ε) telle que

$$\|A\| \leq \rho(A) + \varepsilon.$$

On arrive donc à un résultat qui nous sera fort utile :

Proposition B.1.5. *Soit A une matrice de $\mathbb{R}^{n\times n}$. Les trois assertions suivantes sont équivalentes*

1. $\lim_{k\to+\infty} A^k = 0$,
2. $\lim_{k\to+\infty} A^k x = 0$ *pour tout* $x \in \mathbb{R}^n$,
3. $\rho(A) < 1$.

Démonstration. Pour tout $x \in \mathbb{R}^n$, on sait que $\|A^k x\| \leq \|A^k\|\,\|x\|$. On en déduit facilement que (1)$\Longrightarrow$(2). D'autre part, on sait qu'il existe une valeur propre $\lambda \in \mathbb{C}$ et $x \neq 0$ tels que $Ax = \lambda x$ et $|\lambda| = \rho(A)$. Donc si (2) est vrai, alors $\lambda^k x \xrightarrow[k\to+\infty]{} 0$. Ceci implique que $|\lambda| < 1$. On a donc prouvé que (2)$\Longrightarrow$(3).

Finalement, pour montrer que (3)$\Longrightarrow$(1), on utilise la proposition B.1.4. Ainsi, on prend $\varepsilon > 0$ assez petit tel que $\rho(A) + \varepsilon < 1$. On sait alors qu'il existe une norme induite telle que $\|A\| < 1$, et on a :

$$\|A^k\| \leq \|A\|^k \longrightarrow 0 \quad \text{quand } k \to +\infty,$$

ce qui montre (1). □

Lemme B.1.2. *Soit C une matrice de $\mathbb{R}^{p\times n}$, alors* $\operatorname{Im} C^{\mathsf{T}} = (\operatorname{Ker} C)^{\perp}$.

Démonstration. Prouvons pour commencer que $\operatorname{Im} C^{\mathsf{T}} \subset (\operatorname{Ker} C)^{\perp}$. Soit donc x un élément de $\operatorname{Im} C^{\mathsf{T}}$; il existe $v \in \mathbb{R}^p$ tel que $x = C^{\mathsf{T}}v$. Alors, pour tout élément y appartenant à $\operatorname{Ker} C$, on a

$$(x, y)_n = (C^{\mathsf{T}}v, y)_n = (v, Cy)_p = 0.$$

Pour prouver l'égalité entre ces deux sous-espaces vectoriels de $\mathbb{R}^n$, vérifions qu'ils ont même dimension. D'une part, puisque $\operatorname{Ker} C$ et $(\operatorname{Ker} C)^{\perp}$ sont supplémentaires,

$$n = dim[\operatorname{Ker} C] + dim[(\operatorname{Ker} C)^{\perp}].$$

Et, d'autre part, comme $C : \mathbb{R}^n \to \mathbb{R}^p$, d'après le théorème du rang ($\operatorname{rg}(C)$=$\operatorname{rg}(C^{\mathsf{T}})$), on trouve

$$n = dim[\operatorname{Ker} C] + dim[\operatorname{Im} C] = dim[\operatorname{Ker} C] + dim[\operatorname{Im} C^{\mathsf{T}}].$$

On a bien l'égalité entre les dimensions, $dim[(\operatorname{Ker} C)^{\perp}] = dim[\operatorname{Im} C]$, ce qui permet d'arriver à l'égalité annoncée. □

B.2 Décomposition en valeurs singulières

Dans cette partie, nous allons considérer un aspect algébrique lié aux problèmes de moindres carrés linéaires, celui de la factorisation de la matrice A de $\mathbb{R}^{m\times n}$ sous la forme

$$A = W\Sigma V^{\mathsf{T}} \tag{B.1}$$

où W et V sont deux matrices *orthogonales* (appartenant respectivement à $\mathbb{R}^{m\times m}$ et $\mathbb{R}^{n\times n}$), et Σ une matrice *dont les seuls éléments non nuls sont situés sur la diagonale*, de $\mathbb{R}^{m\times n}$.

Remarque B.2.1. *Il est tout à fait possible de reprendre le raisonnement qui suit et de l'appliquer à une matrice de $\mathbb{C}^{m\times n}$. Dans (B.1), W et V sont alors des matrices unitaires.*

Pourquoi la **décomposition en valeurs singulières** de A, (B.1), est-elle liée aux problèmes de moindres carrés étudiés ci-dessus ? Tout simplement parce que V est reliée à la base orthonormale $(v_i)_{1\leq i\leq n}$ de vecteurs propres de $A^{\mathsf{T}}A$, Σ aux valeurs propres $(\lambda_i)_{1\leq i\leq n}$, et W à la base orthonormale $(w_i)_{1\leq i\leq m}$. Ceci est résumé dans le

Théorème B.2.1. *Soit A une matrice de $\mathbb{R}^{m\times n}$. Il existe W et V deux matrices orthogonales de $\mathbb{R}^{m\times m}$ et $\mathbb{R}^{n\times n}$ respectivement, et Σ une matrice dont les seuls éléments non nuls sont situés sur la diagonale, de $\mathbb{R}^{m\times n}$, telles que (B.1) soit satisfaite.*

Démonstration. Par définition des deux bases orthonormales, on a les relations

$$Av_k = \sigma_k w_k, \quad 1 \leq k \leq n,$$

avec $\sigma_k = \sqrt{\lambda_k}$, pour $1 \leq k \leq n$, ce que l'on peut réécrire sous la forme

$$A \begin{pmatrix} \vdots & \vdots & & \vdots \\ v_1 & v_2 & \cdots & v_n \\ \vdots & \vdots & & \vdots \end{pmatrix} = \begin{pmatrix} \vdots & & \vdots & \vdots & & \vdots \\ \sigma_1 w_1 & \cdots & \sigma_q w_q & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \end{pmatrix}.$$

Soit

$$AV = \begin{pmatrix} \vdots & & \vdots & \vdots & & \vdots \\ \sigma_1 w_1 & \cdots & \sigma_q w_q & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \end{pmatrix}, \text{ où l'on a posé } V = \begin{pmatrix} \vdots & \vdots & & \vdots \\ v_1 & v_2 & \cdots & v_n \\ \vdots & \vdots & & \vdots \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

Par construction,V est orthogonale, puisque

$$(V^{\mathsf{T}}V)_{i,j} = \sum_{k=1}^{n} (V^{\mathsf{T}})_{i,k} V_{k,j} = \sum_{k=1}^{n} V_{k,i} V_{k,j} = \sum_{k=1}^{n} (v_i)_k (v_j)_k = \langle v_i, v_j \rangle_n = \delta_{ij}.$$

Si maintenant, on pose

$$W = \begin{pmatrix} \vdots & \vdots & & \vdots \\ w_1 & w_2 & \cdots & w_m \\ \vdots & \vdots & & \vdots \end{pmatrix} \in \mathbb{R}^{m\times m} \text{ et}$$

$$\Sigma \in \mathbb{R}^{m\times n} \text{ telle que } \Sigma_{i,j} = \begin{cases} \sigma_i, \ 1 \leq i,j \leq q, \ i = j \\ 0 \text{ sinon} \end{cases},$$

vérifions que l'on a l'identité

$$W\Sigma = \begin{pmatrix} \vdots & & \vdots & \vdots & & \vdots \\ \sigma_1 w_1 & \cdots & \sigma_q w_q & 0 & \cdots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \end{pmatrix}.$$

En effet,

— pour $1 \leq i \leq m$, $1 \leq j \leq q$: $(W\Sigma)_{i,j} = \sum_{k=1}^{m} W_{i,k}\Sigma_{k,j} = \sum_{k=1}^{m}(w_k)_i\sigma_j\delta_{kj} = \sigma_j(w_j)_i$;
— pour $1 \leq i \leq m$, $q+1 \leq j \leq n$: $(W\Sigma)_{i,j} = \sum_{k=1}^{m} W_{i,k}\Sigma_{k,j} = 0$ (la $j^{ème}$ colonne de Σ est composée de zéros).

Par construction, W est elle aussi orthogonale, et l'on trouve finalement

$$AV = W\Sigma, \text{ soit } A = W\Sigma V^{\mathsf{T}}.$$

□

Définition B.2.1. *On appelle $(\sigma_k)_k$ les* **valeurs singulières** *de A.*

Remarque B.2.2. *Quelle est l'apparence de Σ ? Si on appelle $r = \min(n, m)$ on a*

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma_2 & 0 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & 0 & \sigma_{r-1} & 0 \\ 0 & \cdots & \cdots & 0 & \sigma_r \end{pmatrix} \quad si\ r = n = m\ ;$$

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & & \vdots & \vdots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots & 0 & \cdots & 0 \\ \vdots & & 0 & \sigma_{r-1} & 0 & \vdots & & \vdots \\ 0 & \cdots & \cdots & 0 & \sigma_r & 0 & \cdots & 0 \end{pmatrix} \quad si\ r = m < n\ ;$$

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma_2 & 0 & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & 0 & \sigma_{r-1} & 0 \\ 0 & \cdots & \cdots & 0 & \sigma_r \\ 0 & \cdots & 0 & \cdots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{pmatrix} \quad si\ r = n < m\ ;$$

Bien sûr, on a toujours $q \leq r$...

Pour aller encore un peu de l'avant, démontrons à présent les identités *matricielles* de la proposition ci-dessous. Un vecteur colonne v de $\mathbb{R}^l$ appartient aussi à $\mathbb{R}^{l\times 1}$, et le vecteur ligne v^{T} appartient lui à $\mathbb{R}^{1\times l}$; le symbole $\cdot$ représente la multiplication matricielle.

Proposition B.2.1.

$$A = \sum_{k=1}^{q} \sigma_k w_k \cdot v_k^{\mathsf{T}}, \quad A^{\mathsf{T}}A = \sum_{k=1}^{q} {\sigma_k}^2 v_k \cdot v_k^{\mathsf{T}}.$$

Démonstration. Plutôt que la simple vérification des résultats, construisons les identités, en commençant par la première. De (B.1), on tire, pour $1 \leq i \leq m$, $1 \leq j \leq n$,

$$\begin{aligned}
A_{i,j} &= \sum_{k=1}^{n} (W\Sigma)_{i,k} V_{k,j}^{\mathsf{T}} \\
&= \sum_{k=1}^{q} (W\Sigma)_{i,k} V_{j,k} \text{ (pour } k > q \text{, la } k^{ème} \text{ colonne de } W\Sigma \text{ est composée de zéros)} \\
&= \sum_{k=1}^{q} \sigma_k W_{i,k} V_{j,k} = \sum_{k=1}^{q} \sigma_k (w_k)_i (v_k)_j \\
&= \sum_{k=1}^{q} \sigma_k (w_k)_{i,1} (v_k^{\mathsf{T}})_{1,j} \text{ (on passe des vecteurs aux matrices)} \\
&= \sum_{k=1}^{q} \sigma_k (w_k \cdot v_k^{\mathsf{T}})_{i,j} = (\sum_{k=1}^{q} \sigma_k w_k \cdot v_k^{\mathsf{T}})_{i,j}.
\end{aligned}$$

Pour la seconde identité, on procède de la même façon. Tout d'abord, on remarque que

$$A^{\mathsf{T}}A = V\Sigma^{\mathsf{T}}W^{\mathsf{T}}W\Sigma V^{\mathsf{T}} = VDV^{\mathsf{T}}, \text{ avec } D = \Sigma^{\mathsf{T}}\Sigma = diag({\sigma_i}^2) \in \mathbb{R}^{n \times n}.$$

(Ce qui exprime aussi le fait que $(v_i)_{1 \leq i \leq n}$ est une base orthonormale de vecteurs propres de $A^{\mathsf{T}}A$, de valeurs propres associées $({\sigma_i}^2)_{1 \leq i \leq n}$.)
A partir de là, on obtient, pour $1 \leq i \leq n$, $1 \leq j \leq n$,

$$\begin{aligned}
(A^{\mathsf{T}}A)_{i,j} &= \sum_{k=1}^{n} (VD)_{i,k} V_{k,j}^{\mathsf{T}} \\
&= \sum_{k=1}^{q} (VD)_{i,k} V_{j,k} \text{ (pour } k > q \text{, la } k^{ème} \text{ colonne de } VD \text{ est composée de zéros)} \\
&= \sum_{k=1}^{q} {\sigma_k}^2 V_{i,k} V_{j,k} = \cdots = (\sum_{k=1}^{q} {\sigma_k}^2 v_k \cdot v_k^{\mathsf{T}})_{i,j}.
\end{aligned}$$

□

Avant de vérifier l'utilité pratique des deux identités ci-dessus, introduisons le **pseudo-inverse** de Σ : soit $\Sigma^{\dagger}$ la matrice de $\mathbb{R}^{n \times m}$ définie par

$$(\Sigma^{\dagger})_{i,j} = \begin{cases} \frac{1}{\sigma_i}, \ 1 \leq i,j \leq q, \ i = j \\ 0 \text{ sinon} \end{cases}.$$

On vérifie immédiatement que l'on a

$$\Sigma^\dagger\Sigma = \begin{cases} I_q & \text{si } q = n \\ \begin{pmatrix} I_q & 0 \\ 0 & 0 \end{pmatrix} & \text{si } q < n \end{cases}.$$

A l'aide de la décomposition en valeurs singulières, nous pouvons maintenant définir le pseudo-inverse de A.

Définition B.2.2. *On appelle* **pseudo-inverse** *de la matrice A de $\mathbb{R}^{m\times n}$ la matrice $A^\dagger$ de $\mathbb{R}^{n\times m}$ définie par*

$$A^\dagger = V\Sigma^\dagger W^\mathsf{T}.$$

A partir de là, on établit aisément les identités ci-dessous

Lemme B.2.1.

$$A^\dagger = \sum_{k=1}^{q} \frac{1}{\sigma_k} v_k{\cdot}w_k^\mathsf{T}, \quad AA^\dagger = \sum_{k=1}^{q} w_k{\cdot}w_k^\mathsf{T}, \quad A^\dagger A = \sum_{k=1}^{q} v_k{\cdot}v_k^\mathsf{T}.$$

Démonstration. La démonstration de la première égalité est semblable à celle de la première identité énoncée pour A.

En ce qui concerne la deuxième égalité, on a

$$\begin{aligned} AA^\dagger &= \left(\sum_{k=1}^{q} \sigma_k w_k{\cdot}v_k^\mathsf{T}\right)\cdot\left(\sum_{l=1}^{q} \frac{1}{\sigma_l} v_l{\cdot}w_l^\mathsf{T}\right) = \sum_{k,l=1}^{q} \frac{\sigma_k}{\sigma_l} w_k{\cdot}v_k^\mathsf{T}{\cdot}v_l{\cdot}w_l^\mathsf{T} \\ &= \sum_{k,l=1}^{q} \frac{\sigma_k}{\sigma_l} w_k{\cdot}(v_k, v_l)_n{\cdot}w_l^\mathsf{T} = \sum_{k,l=1}^{q} \delta_{kl}\frac{\sigma_k}{\sigma_l} w_k{\cdot}w_l^\mathsf{T} = \sum_{k=1}^{q} w_k{\cdot}w_k^\mathsf{T}. \end{aligned}$$

La troisième et dernière égalité se démontre à l'identique. □

On peut alors démontrer le résultat élégant ci-dessous.

Théorème B.2.2. *Un point de minimum du problème de moindres carrés linéaires étudié précédemment est $A^\dagger b$.*

Démonstration. On écrit simplement

$$\begin{aligned} A^\dagger b &= \left(\sum_{k=1}^{q} \frac{1}{\sigma_k} v_k{\cdot}w_k^\mathsf{T}\right)\cdot\left(\sum_{i=1}^{m} b_i w_i\right) = \sum_{k=1}^{q} \frac{1}{\sigma_k} v_k{\cdot}\left(\sum_{i=1}^{m} b_i w_k^\mathsf{T}{\cdot}w_i\right) \\ &= \sum_{k=1}^{q} \frac{b_k}{\sigma_k} v_k = \sum_{k=1}^{q} x_k^0 v_k = x^0. \end{aligned}$$

Or, x^0 appartient à l'ensemble des points de minimum, d'après (3.25). □

Examinons, pour conclure ce chapitre, l'expression du pseudo-inverse dans certains cas particuliers.

Proposition B.2.2. — *Si* $rg(A) = n$*, on a la relation* $A^\dagger = (A^\mathsf{T}A)^{-1}A^\mathsf{T}$.
— *Si* $rg(A) = n = m$*, on a la relation* $A^\dagger = A^{-1}$.

Démonstration. Supposons que $\mathrm{rg}(A) = n$. On a vu que le rang de A et celui de $A^\mathsf{T}A$ sont identiques (et égaux à q). Dès lors que $\mathrm{rg}(A^\mathsf{T}A) = n$, on peut inverser cette dernière. Cette constatation étant faite, on a les relations :

$$(A^\mathsf{T}A)A^\dagger \;=\; (A^\mathsf{T}A)\sum_{k=1}^{n}\frac{1}{\sigma_k}v_k{\cdot}w_k^\mathsf{T} = \sum_{k=1}^{n}\sigma_k v_k{\cdot}w_k^\mathsf{T} = A^\mathsf{T}.$$

On a utilisé le fait que les $(v_k)_{1\le k\le n}$ sont les vecteurs propres de $A^\mathsf{T}A$, ainsi que la transposition de la première égalité de la proposition B.2.1. Comme $A^\mathsf{T}A$ est inversible, la première égalité suit.

Supposons que $\mathrm{rg}(A) = n = m$. On se trouve ici dans le cas où A est une matrice inversible de $\mathbb{R}^{n\times n}$. D'après ce que l'on vient de prouver, on déduit

$$A^\dagger = (A^\mathsf{T}A)^{-1}A^\mathsf{T} = A^{-1}(A^\mathsf{T})^{-1}A^\mathsf{T} = A^{-1}.$$

□

B.3 Méthodes itératives de résolution de systèmes

Dans cette section, on va s'intéresser à des méthodes numériques de résolution du système linéaire

$$\nabla J(u) = Au - b = 0.$$

La solution de ce système est le minimum recherché de la fonctionnelle quadratique J sur $\mathbb{R}^n$.

Le principe des méthodes itératives est le suivant :

1. On décompose la matrice A sous la forme : $A = M - N$ avec $\underline{M \text{ inversible}}$
2. Partant de $u_0 \in \mathbb{R}^n$, on construit $(u_k)_k$ par :

$$Mu_{k+1} = Nu_k - b, \qquad i.e., \quad u_{k+1} = M^{-1}Nu_k - M^{-1}b. \tag{B.2}$$

Ces méthodes ne sont intéressantes que si le choix de M rend (B.2) particulièrement facile à résoudre.

Si la suite $(u_k)_k$, définie par la relation de récurrence (B.2), converge vers une limite u, alors par passage à la limite dans (B.2), on obtient :

$$(M - N)u = Au = b.$$

Par conséquent, si $(u_k)_k$ converge, alors sa limite est forcément la solution du système linéaire.

Dans le lemme suivant, nous allons énoncer une condition nécessaire et suffisante pour la convergence d'une méthode itérative à l'aide du rayon spectrale de la matrice $M^{-1}N$ (voir annexe B.1, pour la définition du rayon spectrale).

Lemme B.3.1. *La suite $(u_k)_k$ définie par la méthode itérative* (B.2) *est convergente si et seulement si le rayon spectrale $\rho(M^{-1}N)$ vérifie : $\rho(M^{-1}N) < 1$.*

Démonstration. On a :

$$\begin{aligned} u_{k+1} - u &= (M^{-1}Nu_k + M^{-1}b) - (M^{-1}Nu + M^{-1}b) \\ &= M^{-1}N(u_k - u) \end{aligned}$$

Donc $u_k - u = (M^{-1}N)^k(u_0 - u)$. Par application de la proposition B.1.5, on en déduit que u_k tend vers u, quel que soit $u_0 \in \mathbb{R}^n$, si et seulement si $\rho(M^{-1}N) < 1$. □

Le rayon spectral est souvent difficile à calculer. Cependant, nous avons le résultat général (fort utile) suivant :

Lemme B.3.2. *Soit A une matrice symétrique définie positive. Soit une décomposition de A définie par $A = M - N$ avec M inversible. Si $(M^{\mathsf{T}} + N)$ est aussi définie positive, alors*

$$\rho(M^{-1}N) < 1.$$

Démonstration. Notons d'abord que que si A est symétrique, alors $M^{\mathsf{T}} + N$ l'est aussi et donc ses valeurs propres sont réelles. En effet,

$$(M^{\mathsf{T}} + N)^{\mathsf{T}} = M + N^{\mathsf{T}} = A + N + N^{\mathsf{T}} = A^{\mathsf{T}} + N^{\mathsf{T}} + N = M^{\mathsf{T}} + N.$$

D'autre part, de la proposition B.1.4, on sait que $\rho(M^{-1}N) \leq \|M^{-1}N\|$ pour toute norme induite $\|.\|$. On choisit la norme matricielle induite de la norme vectorielle définie par $\|v\|_A^2 = (Av, v)$ (c'est bien une norme vectorielle, puisque A est symétrique définie

positive). Ainsi, pour $v \in \mathbb{R}^n$, en posant $w = M^{-1}Av$, il vient :

$$\begin{aligned}
\|M^{-1}Nv\|_A^2 &= \left\langle AM^{-1}Nv, v\right\rangle = \langle v-w, A(v-w)\rangle \\
&= \|v\|_A^2 - \langle w, Av\rangle - \langle w, Av\rangle + \langle w, Aw\rangle \\
&= \|v\|_A^2 - \langle w, Mw\rangle - \langle w, Mw\rangle + \langle w, Aw\rangle \\
&= \|v\|_A^2 - \langle w, (M-A)w\rangle - \left\langle M^{\mathsf{T}}w, w\right\rangle \\
&= \|v\|_A^2 - \left\langle w, (M^{\mathsf{T}}+N)w\right\rangle \\
&\leq \|v\|_A^2 - \lambda_{\min}(M^{\mathsf{T}}+N)\|w\|_2^2.
\end{aligned}$$

Comme $M^{\mathsf{T}} + N$ est définie positive, on a $\lambda_{\min}(M^{\mathsf{T}} + N) > 0$. De plus,

$$\begin{aligned}
Av = Mw &\implies \langle Av, v\rangle = \langle Mw, v\rangle \\
&\implies \|v\|_A^2 \leq \|M\|_2 \|w\|_2 \|v\|_2,
\end{aligned}$$

et d'autre part on a $\lambda_{\min}(A)\|v\|_2^2 \leq \|v\|_A^2$ avec $\lambda_{\min}(A) > 0$ puisque A est définie-positive. Finalement, on obtient :

$$\|M^{-1}Nv\|_A^2 \leq \left[1 - \frac{\lambda_{\min}(M^T+N)\lambda_{\min}(A)}{\|M\|_2^2}\right] < 1 \quad \forall v \in \mathbb{R}^n, \text{ vérifiant } \|v\|_A = 1.$$

On en déduit que

$$\rho(M^{-1}N) \leq \|M^{-1}N\|_A < 1.$$

□

Nous allons maintenant donner les exemples les plus classiques de méthodes itératives. Pour celà, notons $D = \text{diag}(A)$ la diagonale de A, $-E = \text{triang}_{\text{inf}}(A)$ la partie triangulaire inférieure de A et par $-F = \text{triang}_{\text{sup}}(A)$ la partie triangulaire supérieure.

B.3.1 Méthode de Jacobi

On appelle la méthode de Jacobi, la méthode itérative associée au choix : $M = D$ et $N = D - A = E + F$. On désigne par matrice de Jacobi la matrice $\mathcal{J} = M^{-1}N = D^{-1}(E + F)$. L'algorithme s'écrit alors :

$$\left\|\begin{array}{l}
1) \text{ On choisit } u_0 \in \mathbb{R}^n, \ \varepsilon > 0 \\
2) \text{ Tant que } \|Au_k - b\| > \varepsilon, \text{ calculer } u_{k+1} = \mathcal{J}u_k + D^{-1}b.
\end{array}\right.$$

Pour que cette méthode soit bien définie, il faut que la matrice D soit inversible (c.a.d. que tous les éléments diagonaux de A soient non nuls).

B.3.2 Méthode de Gauss-Seidel

On appelle la méthode de Gauss-Seidel, la méthode itérative associée au choix : $M = D - E$ et $N = F$. On désigne par matrice de Gauss-Seidel la matrice $\mathcal{G} = M^{-1}N = (D-E)^{-1}F = (A+F)^{-1}F$. L'algorithme s'écrit alors :

$$\left\| \begin{array}{l} 1) \text{ On choisit } u_0 \in \mathbb{R}^n,\ \varepsilon > 0 \\ 2) \text{ Tant que } \|Au_k - b\| > \varepsilon, \text{ calculer } u_{k+1} = \mathcal{G}u_k + (D-E)^{-1}b. \end{array} \right.$$

La méthode de Gauss-Seidel est bien définie si la matrice $D - E$ est inversible, c'est à dire que D est inversible. Notons que $D - E$ est facile à inverser puisque c'est une matrice triangulaire).

Théorème B.3.1.

1. *Si A est symétrique et définie-positive, alors la méthode de Gauss-Seidel converge.*
2. *Si A est à diagonale strictement dominante (i.e. $|A_{ii}| > \sum_{i \neq j} |A_{ij}|$ pour tout $i = 1, \cdots, n$), alors les méthodes Gauss-Seidel et Jacobi convergent*
3. *Si A est tridiagonale, alors $\rho(\mathcal{G}) = \rho(\mathcal{J})^2$.*

Démonstration.

1. Le point (i) s'obtient à partir des lemmes B.3.1 et B.3.2, en remarquant que si A est symétrique définie positive, alors $M^{\mathsf{T}} + N = (D - E)^{\mathsf{T}} + F = D - F + F = D$ est définie positive.
2. Pour prouver le point (ii), notons d'abord que la matrice $\mathcal{J} = M^{-1}N$ vérifie :

$$\mathcal{J}_{ii} = 0; \quad \text{et } \mathcal{J}_{ij} = \frac{A_{ij}}{A_{ii}} \text{ pour } i \neq j.$$

   Si A est à diagonale strictement dominante, alors

$$\|\mathcal{J}\|_\infty = \max_i \sum_j |\mathcal{J}_{ij}| = \max_i \frac{\sum_j |A_{ij}|}{|A_{ii}|} < 1.$$

   On en déduit que $\rho(\mathcal{J}) \leq \|\mathcal{J}\|_\infty < 1$.
3. Pour prouver (iii), nous allons vérifier que lorsque A est tridiagonle, $\lambda \neq 0$ est valeur propre de $\mathcal{J}$ si et seulement si et seulement si λ^2 est valeur propre de $\mathcal{G}$. Pour celà, considérons la matrice $Q(\delta) := \text{diag}(\delta, \delta^2, \cdots, \delta^n)$, avec $\delta \neq 0$. Remarquons d'abord que pour toute matrice tridiagonale de la forme $B = P - L - U$, avec

$$P_{ij} = \begin{cases} B_{ii} & i = j \\ 0 & i \neq j \end{cases}, L_{ij} = \begin{cases} B_{ii-1} & j = i-1 \\ 0 & j \neq i-1 \end{cases}, \text{et } U_{ij} = \begin{cases} B_{ii+1} & j = i+1 \\ 0 & j \neq i+1 \end{cases},$$

on a :

$$Q(\delta)BQ(1/\delta) = P - \delta L - \frac{1}{\delta}U,$$

et

$$\begin{aligned} \det(B) &= \det(P - L - U) = \det(Q(\delta))\det(B)\det(Q(1/\delta)) \\ &= \det(P - \delta L - \frac{1}{\delta}U) \qquad \forall \delta \neq 0. \end{aligned} \tag{B.3}$$

Revenons maintenant à la matrice A. $\lambda \in \mathbb{C}$ est valeur propre de $\mathcal{J}$ ssi λ est racine du polynôme :

$$P_{\mathcal{J}}(\lambda) := \det(\lambda I - D^{-1}(E + F)) = \det(\lambda D - E - F)/\det(D).$$

De même, λ est valeur propre de $\mathcal{G}$ ssi λ est racine du polynôme

$$P_{\mathcal{G}}(\lambda) := \det(\lambda I - (D - E)^{-1}F) = \det(\lambda D - \lambda E - F)/\det(D - E).$$

Or, on a $\det(D) = \det(D - E)$, et d'après (B.3) (en posant $\delta = \lambda$, $P = \lambda^2 D$, $L = \lambda E$, et $U = \lambda F$) on a :

$$\det(\lambda^2 D - \lambda^2 E - F) = \det(\lambda^2 D - \lambda E - \lambda F) = \lambda^n \det(\lambda D - E - F).$$

Ce qui prouve que pour $\lambda \neq 0$,

$$P_{\mathcal{J}}(\lambda) = 0 \text{ ou } P_{\mathcal{J}}(-\lambda) = 0 \Longleftrightarrow P_{\mathcal{G}}(\lambda^2) = 0,$$

en d'autre terme λ ou $-\lambda$ est valeur propre de $\mathcal{J}$ si et seulement si λ^2 est valeur propre de $\mathcal{G}$. Par conséquent, on a bien $\rho^2(\mathcal{J}) = \rho(\mathcal{G})$.

□

Annexe C

Chemins et cônes derivables

Commençons par introduire la notion de **chemin**, et rappeler celle des **tangentes**.

Définition C.1.1. *On appelle chemin réel une fonction dérivable $\gamma : t \mapsto \gamma(t)$ de $\mathbb{R}$ dans $\mathbb{V}$. On appelle tangente au chemin en t_0 la droite passant par $\gamma(t_0)$ et de direction $\gamma'(t_0)$. On appelle chemin une fonction de $[0, \alpha[$ à valeurs dans $\mathbb{V}$, dérivable sur $]0, \alpha[$ et dérivable à droite en 0, avec $\alpha > 0$. Dans ce cas, la tangente en 0 est une demi-droite, passant par $\gamma(0)$, et de direction $\gamma'_d(0)$, c'est-à-dire : $\{w \in \mathbb{V} \ : \ \exists \eta \geq 0, \ w = \gamma(0) + \eta\, \gamma'_d(0)\}$.*

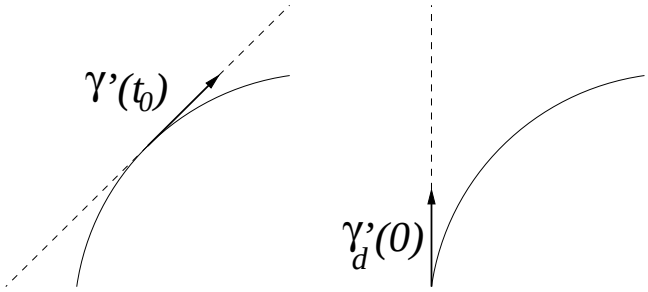


FIGURE C.1 – Tangentes

On rappelle que, lorsque la variable est réelle,

$$\gamma'(t_0) = \lim_{\theta \to 0} \frac{\gamma(t_0 + \theta) - \gamma(t_0)}{\theta}, \text{ et } \gamma'_d(t_0) = \lim_{\theta \to 0^+} \frac{\gamma(t_0 + \theta) - \gamma(t_0)}{\theta}.$$

Proposition C.1.1. *Soit γ un chemin. On a, pour $t_0 \in]0, \alpha[$, avec $\alpha > 0$*

$$d\gamma(t_0) \cdot h = h\, \gamma'(t_0), \qquad \forall h \in \mathbb{R}.$$

Démonstration. Il suffit de comparer la définition A.1.1 à celle de la dérivée usuelle

$$\gamma(t_0 + h) = \gamma(t_0) + h\gamma'(t_0) + o(h), \qquad \forall h \in]-t_0, \alpha - t_0[.$$

On en déduit donc que l'on a l'égalité $d\gamma(t_0)\cdot h = h\,\gamma'(t_0)$, pour h suffisamment petit. Comme $d\gamma(t_0)$ est une application linéaire, l'égalité précédente est vraie pour tout h de $\mathbb{R}$. □

Remarque C.1.1. *La dérivée à droite peut être vue comme une dérivée directionnelle (cf. définition A.1.3) dans le sens positif. En effet,*

$$\gamma'_d(t_0) = \lim_{\theta\to 0^+} \frac{\gamma(t_0+\theta\,(+1)) - \gamma(t_0)}{\theta} = d\gamma(t_0)\cdot(+1).$$

Soit maintenant f une application Fréchet-différentiable de $\mathbb{V}$ dans $\mathbb{F}$ (cf. Annexe). On construit $\mu = f\circ\gamma$, une fonction de la variable réelle à valeurs dans $\mathbb{F}$. Tous les résultats connus s'appliquent sur une telle fonction (théorème des accroissements finis, formules de Taylor, etc.). En particulier μ est dérivable, comme composée d'applications différentiables, et on a

$$\begin{aligned} & d\mu(t_0)\cdot h = df(\gamma(t_0))\cdot(d\gamma(t_0)\cdot h), \qquad \forall h\in\mathbb{R} \\ \Longleftrightarrow\quad & h\,\mu'(t_0) = df(\gamma(t_0))\cdot(h\,\gamma'(t_0)), \qquad \forall h\in\mathbb{R} \\ \Longleftrightarrow\quad & h\,\mu'(t_0) = h\,df(\gamma(t_0))\cdot\gamma'(t_0), \qquad \forall h\in\mathbb{R}, \end{aligned}$$

puisque $df(\gamma(t_0))$ est linéaire, soit finalement

$$\mu'(t_0) = df(\gamma(t_0))\cdot\gamma'(t_0). \tag{C.1}$$

Si J est une fonctionnelle Fréchet-différentiable de $\mathbb{V}=\mathbb{R}^n$ dans $\mathbb{F}=\mathbb{R}$, si on pose $\mu = J\circ\gamma$, on infère que (cf. (A.5))

$$\mu'(t_0) = \langle \nabla J(\gamma(t_0)), \gamma'(t_0)\rangle\,, \tag{C.2}$$

et si enfin $\gamma(t) = u + t\,w$, avec $u, w\in\mathbb{R}^n$, on a $\gamma'(t_0) = w$, ce qui donne

$$\mu'(t_0) = \langle \nabla J(u+t_0\,w), w\rangle\,. \tag{C.3}$$

Proposition C.1.2. *Si J est de classe $\mathcal{C}^2$ et $\mu(t) = J(u+t\,w)$, on a*

$$\mu''(t_0) = \left\langle \nabla^2 J(u+t_0\,w)w, w\right\rangle. \tag{C.4}$$

Démonstration. On écrit

$$\begin{aligned} \mu'(t_0+h) - \mu'(t_0) &= \langle \nabla J(u+t_0\,w+h\,w) - \nabla J(u+t_0\,w), w\rangle \\ &= (h\nabla^2 J(u+t_0\,w)w + \|h\,w\|\varepsilon(h\,w), w) = h\left\langle \nabla^2 J(u+t_0\,w)w, w\right\rangle + o(h), \end{aligned}$$

puisque ∇J est de classe $\mathcal{C}^1$ de $\mathbb{R}^n$ dans $\mathbb{R}$, cf. (A.6). D'où finalement

$$\mu''(t_0) = \lim_{h\to 0}\frac{\mu'(t_0+h)-\mu'(t_0)}{h} = \left\langle \nabla^2 J(u+t_0\,w)w, w\right\rangle.$$

□

Notons que si J est simplement Gateaux-différentiable, on doit se limiter aux chemins inclus dans des droites, c'est-à-dire de la forme $\gamma(t) = u + t\,w$ évoquée ci-dessus.

On considère K un sous-ensemble non vide quelconque de $\mathbb{V}$, et J une fonctionnelle de K dans $\mathbb{R}$. Il est commode d'introduire le **cône des directions admissibles**.

Définition C.1.2. *Soit K un sous-ensemble non vide de $\mathbb{V}$, et v un point de K. On appelle cône (dérivable) des directions admissibles l'ensemble $\mathcal{T}_K(v)$ des tangentes en v aux chemins inclus dans K et commençants en v.*

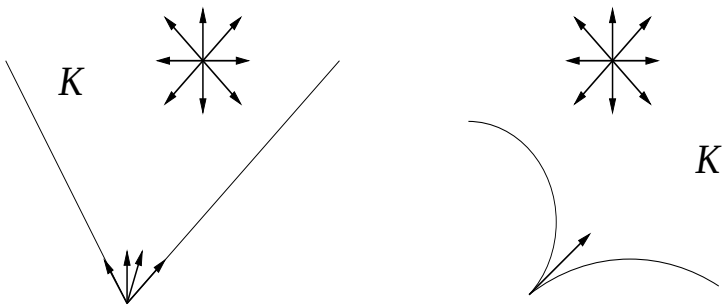


FIGURE C.2 – Exemples de cônes de directions admissibles

En d'autres termes, w appartient à $\mathcal{T}_K(v)$, si, et seulement si, il existe un chemin $\gamma : [0, \alpha[\to K$ $(\alpha > 0)$ tel que $\gamma(0) = v$ et $w = \gamma'_d(0)$. En d'autres termes :

$$w \in \mathcal{T}_K(v) \Longleftrightarrow \exists t_0, \forall t \in (0, t_0), \quad v + tw + o(t) \in K.$$

Exercice C.1.1. *Supposons que $K \neq \emptyset$ est un convexe fermé. Soient $u \in K$ et $w \in \mathbb{V}$ avec $w \neq 0$. Montrer que $w \in \mathcal{T}_K(u)$ si, et seulement si l'une des assertions suivantes est vérifiée.*

(i) il existe une suite (u_k) d'éléments de K, telle que :

$$u_k \xrightarrow[k\to\infty]{} u \quad \text{et} \quad \frac{u_k - u}{\|u_k - u\|} \xrightarrow[k\to\infty]{} \frac{w}{\|w\|}.$$

(ii) il existe une suite $(w_k)_k \subset \mathbb{V}$, et une suite $(\lambda_k)_k$ telles que

$$\lim_k \lambda_k = 0,\ \lim_k w_k = w, \quad \text{et pour tout } k \geq 0,\ \lambda_k > 0 \text{ et } u + \lambda_k w_k \in K.$$

Avec la notion de cône que nous venons de définir nous pouvons révisiter la preuve du Théorème 3.3.1.

Preuve du Théorème 3.3.1 avec notion alternative de cône tangente. Si $w = 0$, l'inégalité est tout simplement une égalité. Si w est un élément (non nul) de $\mathcal{T}_K(u)$, il existe un

chemin $\gamma : [0, \alpha[\to K$ $(\alpha > 0)$ passant par u en $t = 0$ tel que $w = \gamma'_d(0)$. Pour t suffisamment petit, $\gamma(t)$ est proche de u. Ainsi il existe $t_0 > 0$ tel que

$$J \circ \gamma(t) \geq J \circ \gamma(0), \qquad \forall t \in [0, t_0[. \tag{C.5}$$

Or, $J \circ \gamma$ est dérivable à droite en 0 de dérivée $\langle \nabla J(u), w \rangle$: en effet, on peut écrire γ sous la forme $\gamma(t) = u + t\, w + o(t)$, pour $t \in [0, \alpha[$. On a alors

$$J \circ \gamma(t) = J(u + t\, w + o(t)) = J(u + h), \text{ avec } h = t\, w + o(t).$$

Or, $\|h\|$ tend vers 0 lorsque t tend vers 0^+, puisque $\|h\| \leq 2t\|w\|$, pour t suffisamment petit. Par application de (A.6), on trouve alors

$$J \circ \gamma(t) = J(u) + \langle \nabla J(u), t\, w + o(t) \rangle + o(t) = J(u) + t \langle \nabla J(u), w \rangle + o(t).$$

D'où par passage à la limite dans (C.5)

$$\frac{J \circ \gamma(t) - J \circ \gamma(0)}{t} = \langle \nabla J(u), w \rangle + \frac{o(t)}{t} \overset{t \to 0^+}{\longrightarrow} \langle \nabla J(u), w \rangle .$$

Ce qui implique bien que $\langle \nabla J(u), w \rangle \geq 0$. □

Bibliographie

[1] A. Björck. *Solutions of equations in $\mathbb{R}^n$ (Part I) : Least squares methods*. Handbook of numerical analysis, Volume I. North Holland, Amsterdam, 1990.

[2] J.F. Bonnans, J.C. Gilbert, C. Lemaréchal, and C. Sagastizábal. *Numerical Optimization : Theoretical and Practical Aspects*. Springer-Verlag, 2nd edition, 2006.

[3] P. Ciarlet and P. Joly. Introduction au calcul scientifique. Cours MA 103 ENSTA.

[4] P. G. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation*. Masson, Paris, 1982.

[5] A.-S. Bonnet Ben Dhia and M. Lenoir. Outils élémentaires d'analyse pour les équations aux dérivées partielles. Cours MA 102 ENSTA.

[6] E. D. Dolan and J. J. Moré. Benchmarking optimization software with performance profiles. *Mathematical Programming*, 91 :201–213, 2002.

[7] V. Faber and T. Manteuffel. Necessary and sufficient conditions for the existence of a conjugate gradient method. *SIAM Journal on Numerical Analysis*, 21(1) :352–362, 1984.

[8] J.-C. Gilbert. Optimisation différentiable. théorie et algorithmes. Cours AO 201 ENSTA.

[9] G. H. Golub and G. Meurant. *Résolution numérique des grands systèmes linéaires*. Eyrolles, Paris, 1983.

[10] F. Jean. Linéarisation et stabilité des équations différentielles. Cours AO 102 ENSTA.

[11] G. Meurant. *Computer solution of large linear systems*, volume 28 of *Studies in Mathematics and its Applications*. North Holland Amsterdam, 1999.

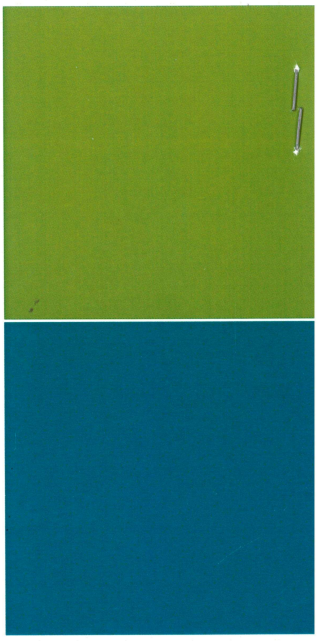
[12] B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation I. Basic Theory.*, volume 330 of *Grundlehren der mathematischen Wissenschaften*. Springer, Heidelberg, 2006.

[13] J. Pérez. Gravitation classique. enseignement thématique d'astrophysique. MAT 40 ENSTA.

[14] E. Polak and G. Ribière. Sur la convergence de la méthode des gradients conjugués. *Revue Française d'Informatique et de Recherche Opérationnelle*, 16(R1), 1969.

[15] R.T. Rockafellar and R. J.-B. Wets. *Variational Analysis*, volume 317 of *Grundlehren der mathematischen Wissenschaften*. Springer Verlag Berlin, 3rd edition, 2009.

[16] Y. Saad and M. H. Schultz. Gmres : a generalized minimum residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(1) :856–869, 1986.



École Nationale Supérieure
de **Techniques Avancées**
828, boulevard des Maréchaux - 91762 Palaiseau Cedex - France
www.ensta-paris.fr

ENSTA

IP PARIS